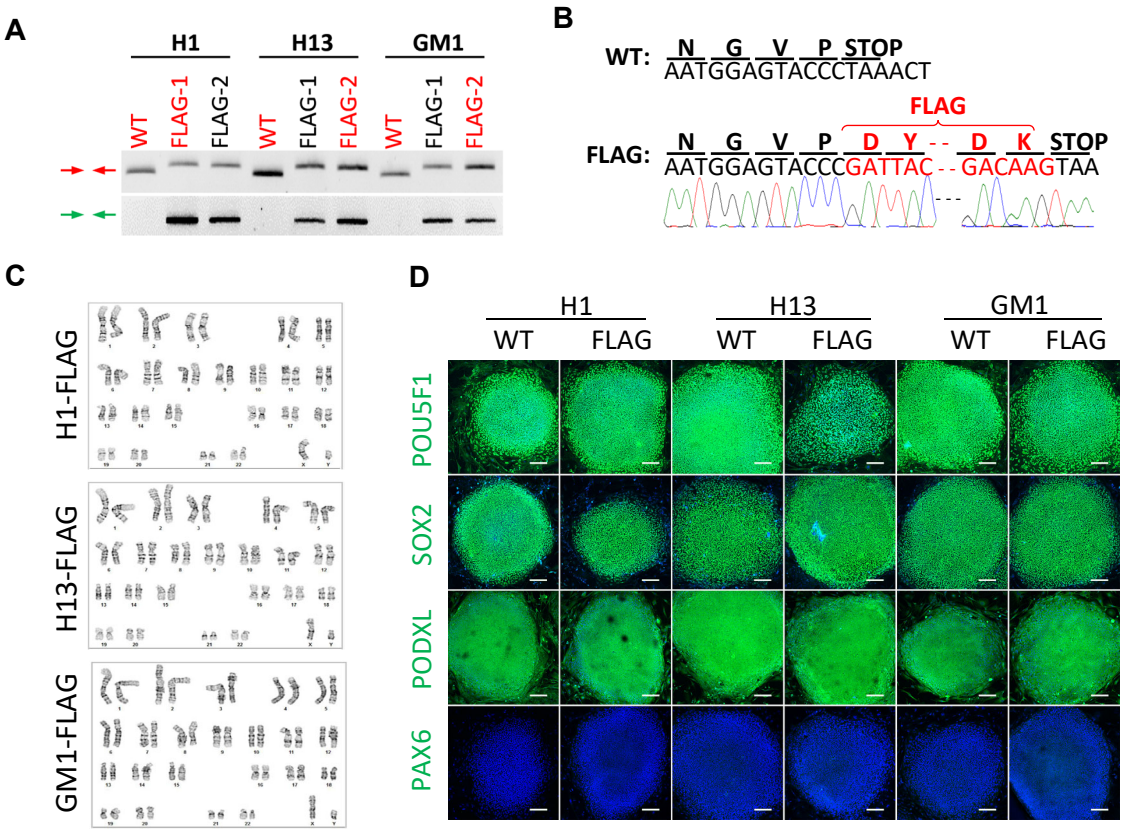
Supplemental Information

Identification and integrative analyses of cell type specific FMR1 regulated molecular networks in human neurodevelopment

Meng Li, Junha Shin, Ryan D Risgaard, Molly Parries, Jianyi Wang, Deborah A Chasman, Shuang Liu, Sushmita Roy, Anita Bhattacharyya, and Xinyu Zhao

Table of Content

Supplemental figure S1

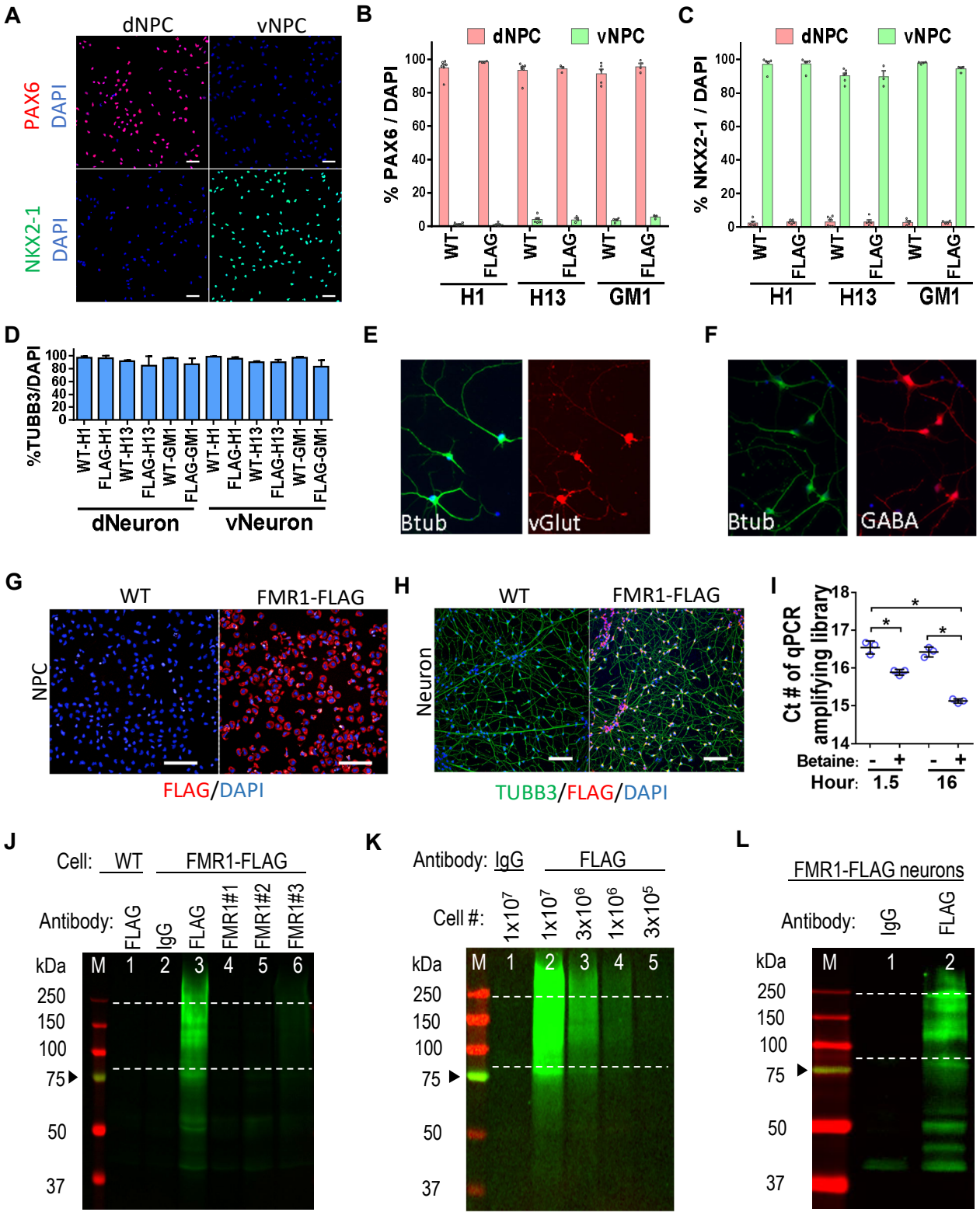


E

	position (hg19)	sequence	genic	testing result
1	chr4:-184674478	TAATTTACTCCATTCAAGAGAAG	no	Good
2	chr12:-80218129	TGGGGTAGTTCATTCATGAGAAG	no	Good
3	chr15:-53901519	AAGGGTACACCATTCACTAGGGG	no	Good
4	chr7:-1364038	CTGGGTACCCCATTCAGGAGGAG	no	Good
5	chr10:+51825245	TGGGCTACTCTATTCAGGAGAAG	no	Good

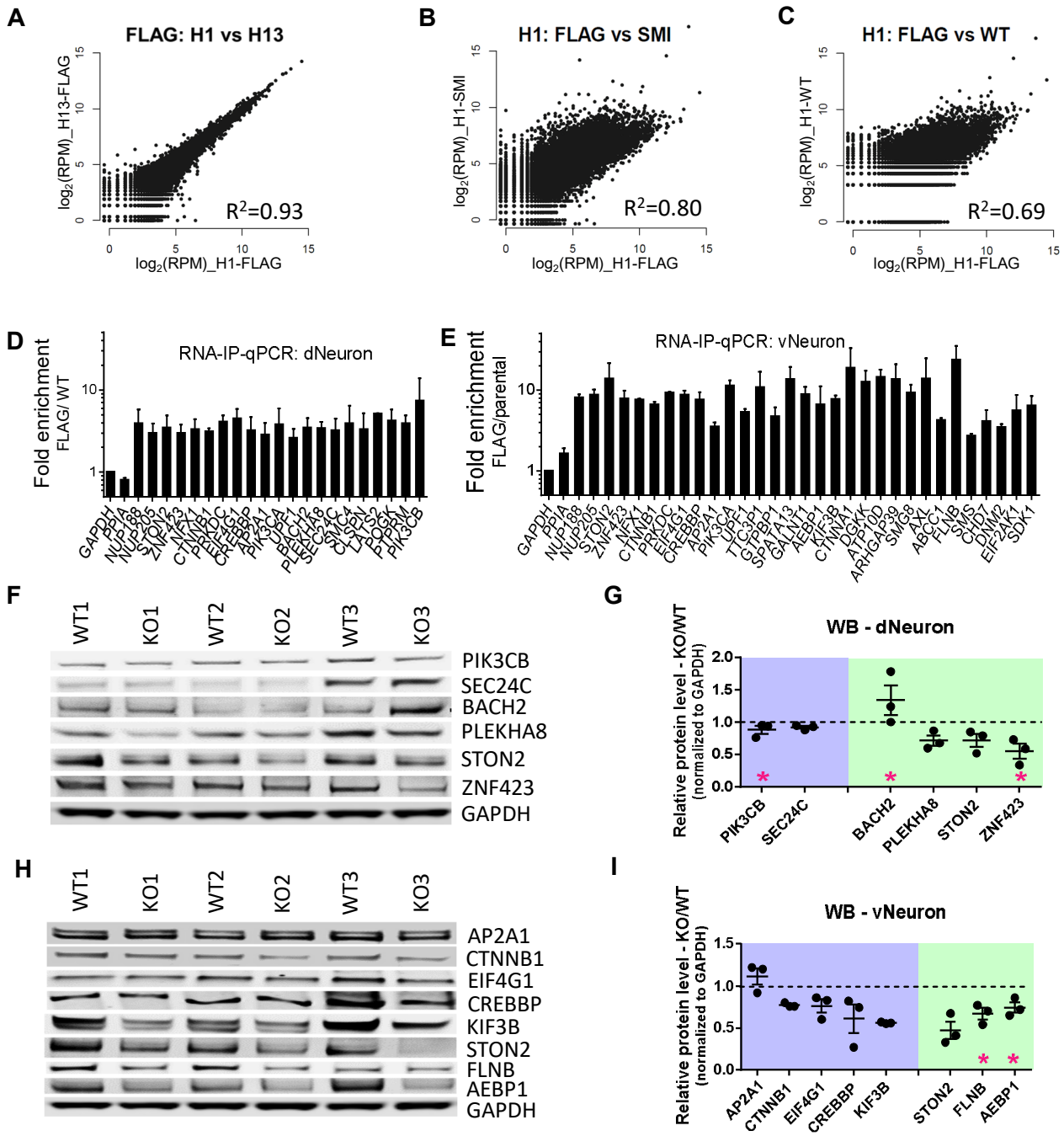
Supplemental figure S1. Characterization of FMR1-FLAG hPSCs. (A) Genotyping indicating successful insertion of FLAG sequence. Primers (red and green arrows) are the same as in Figure 1B. **(B)** Sanger sequencing confirming insertion of FLAG to the C-terminus of FMR1. **(C)** G-banding confirming normal karyotypes of the FMR1-FLAG hPSCs. **(D)** Immunofluorescence of WT and FLAG hPSCs for stem cell markers, OCT4 (POU5F1), SOX2, TRA-1-81 (PODXL), and a neural stem cell marker, PAX6 (scale bar 200 μm). **(E)** No off-target mutations detected in the FMR1-FLAG hPSCs by Sanger sequencing of top 5 predicted off-target sites.

Supplemental figure S2



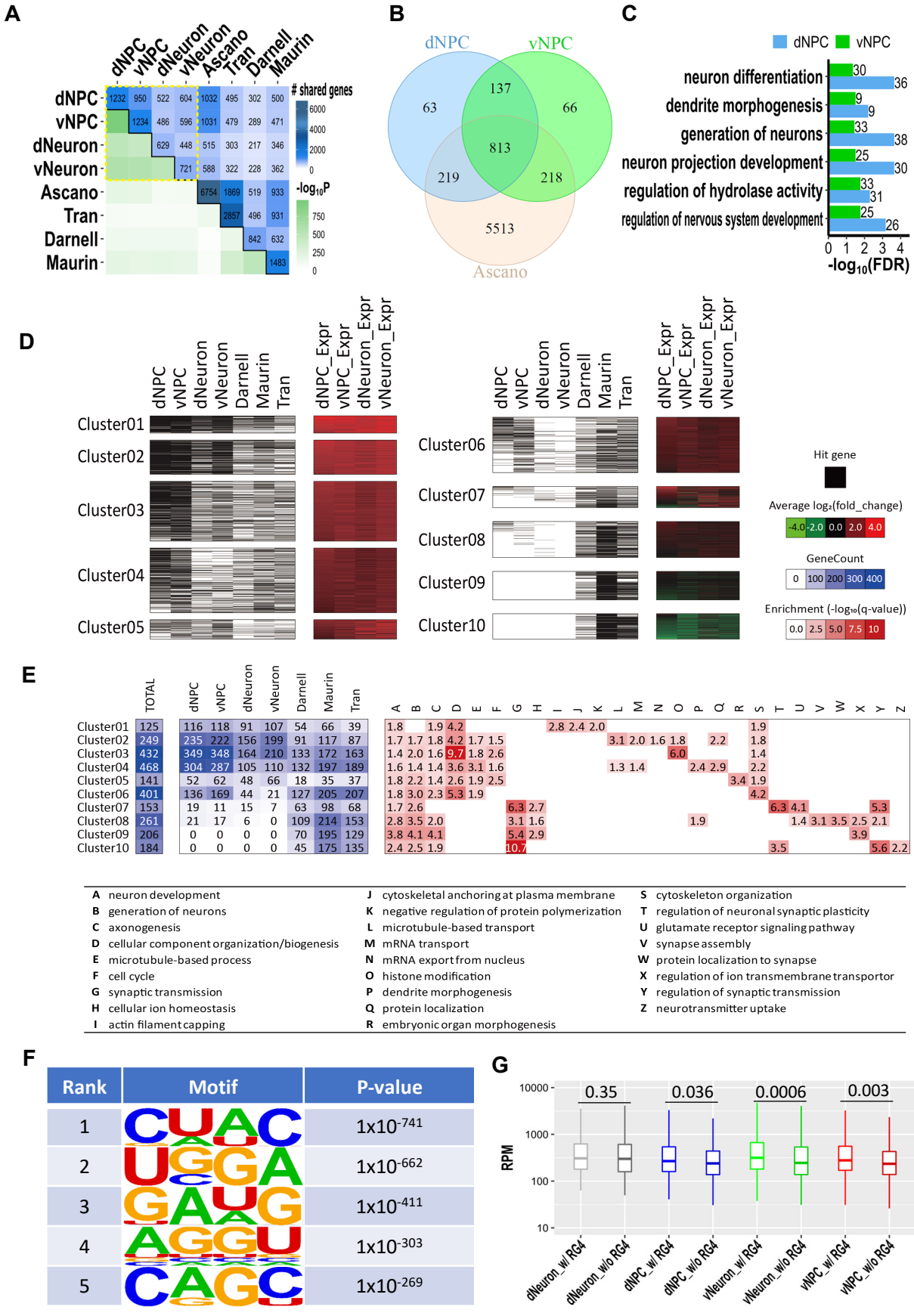
Supplemental figure S2. Neural differentiation and CLIP with neural cells. (A) Representative immunofluorescence of NPCs with dorsal marker, PAX6, and MGE marker, NKX2-1 (scale bar, 50 μm). **(B-C)** Quantification of percentage of PAX6-positive **(B)** and NKX2-1-positive **(C)** dNPCs and vNPCs (mean ± SE). Dots represent different batches of neural differentiation. **(D)** Quantification of percentage of TUBB3-positive neurons (n=2 batches; mean ± SD). **(E)** Representative immunofluorescence of dorsal neurons with neuron marker TUBB3 (Btub) and glutamatergic neuron marker VGLUT1 (SLC17A7). **(F)** Representative immunofluorescence of MGE neurons with neuron marker TUBB3 (Btub) and GABAergic neuron marker GABA. **(G)** Representative immunofluorescence of NPCs with FLAG antibody (scale bar, 100 μm). **(H)** Representative immunofluorescence of neurons with TUBB3 and FLAG (scale bar, 100 μm). **(I)** qPCR quantification of CLIP cDNAs with various circularization conditions. Data are shown as mean ± SE (n=3, dots). Star indicates $p<0.05$. **(J)** Optimizing CLIP conditions by using FLAG or control WT dNPCs and several different antibodies (antibody information in Supplemental information) (M: protein ladder). Isolated RNAs were visualized by dye in green. Arrowhead indicates expected size of FMR1. Dash lines indicate size selection of RNAs. **(K)** Optimizing CLIP conditions with different numbers of dNPCs. M indicates protein ladder. **(L)** Representative image of CLIP with neurons derived from FMR1-FLAG hPSCs.

Supplemental figure S3



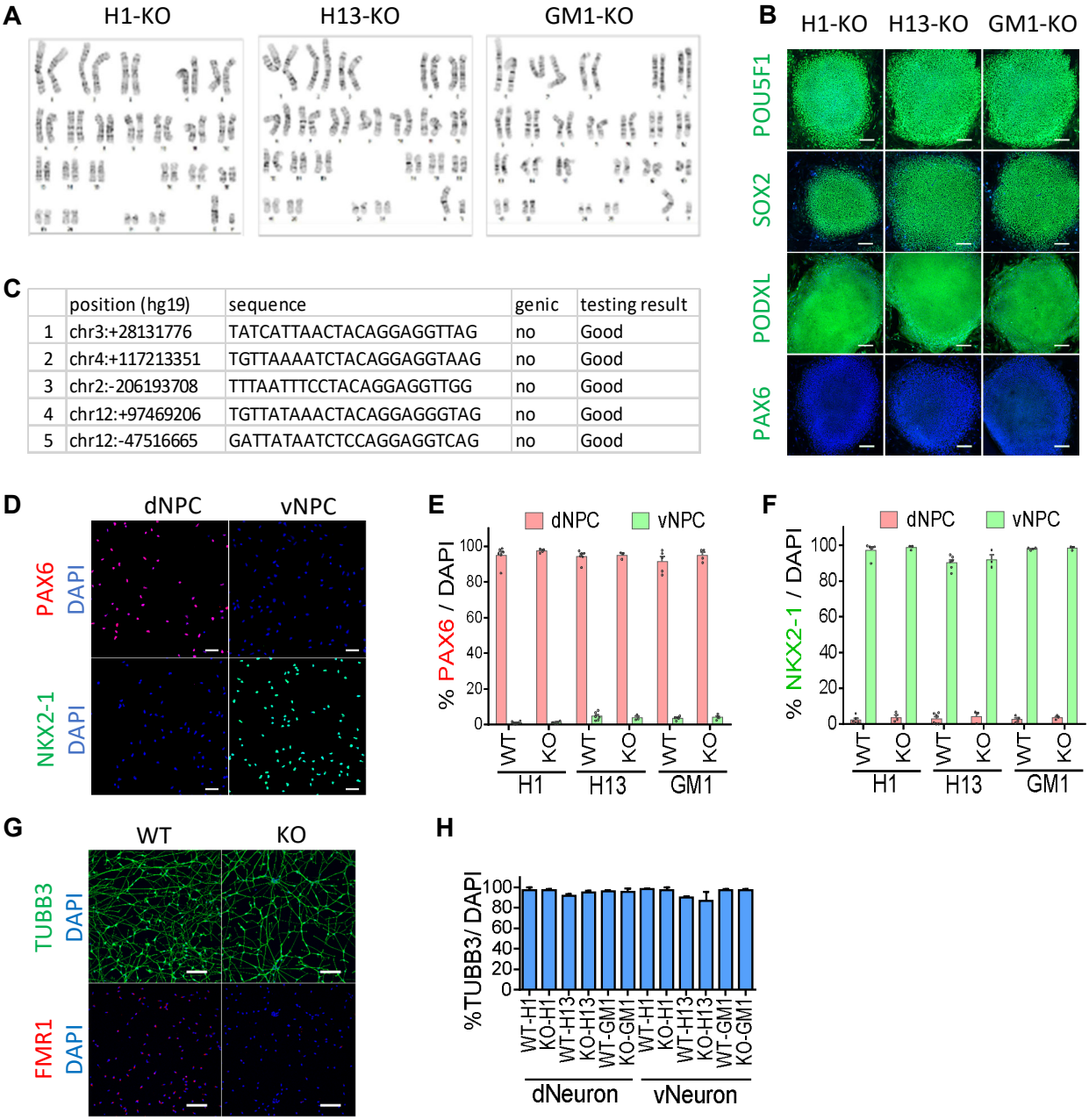
Supplemental figure S3. CLIP analysis and validation of CLIP targets. (A-C) Pearson correlation between CLIP-seq data (dNPC): **(A)** FLAG-H1 and FLAG-H13; **(B)** FLAG-H1 and SMI-H1; **(C)** FLAG-H1 and WT-H1. **(D-E)** Validation of FMR1 targets in dNeuron **(D)** and vNeuron **(E)** using RIP-qPCR. Data shows fold enrichment of RNAs in FLAG cells over parental cells (mean ± SE). **(F-H)** Western blot of FMR1 targets in dNeuron **(F)** and vNeuron **(H)** and quantification **(G** and **I)**. Blue indicates hubs and green are non-hub genes. Red star marks targets not identified in previous mouse and human brain tissues.

Supplemental figure S4



Supplemental figure S4. Comparison of FMR1 targets with published FMR1 targets and characterization of CLIP targets. (A) Heat map of number of overlapping genes and p-values (hypergeometric test) between FMR1 targets identified in the four human neural cell types (dNPC, vNPC, dNeuron, vNeuron) in this study as well as other human (Ascano and Tran) and mouse samples (Darnell and Maurin). FMR1 identified in this study is highlighted by the yellow dot frame. **(B)** Venn diagram showing overlap of FMR1 targets in dividing cells: dNPC, vNPC, and HEK293 cells. **(C)** Bar plots showing FMR1 targets distinct in either dNPC (200 genes) or vNPC (213 genes) from HEK293 cells are enriched in genes involved in neural development. Numbers at the end of bars indicate number of genes represented in the GO terms. **(D)** Gaussian mixture model clusters of 2,620 genes comprising both FMR1 targets identified from our study and targets previously reported from mouse (Darnell and Maurin) and human (Tran) studies. Genes are clustered based on their $\log_2$(Fold change) values of FLAG group versus control groups in CLIP-seq. **(E)** Number of genes and FMR1 targets identified in our 4 cell types and previous studies in each cluster (blue heat map). Shown also are selected biological processes associated with the cluster (red heat map). Terms were selected based on their significance (P-value) and relevance to neuronal processes. **(F)** Top 5 consensus motifs within FMR1-binding sites identified by Homor. **(G)** Boxplot showing the RPM (reads per million) distribution of FMR1 targets with or without RG4 (RNA G-quadruplex). P values were calculated by KS test. Box plot: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range.

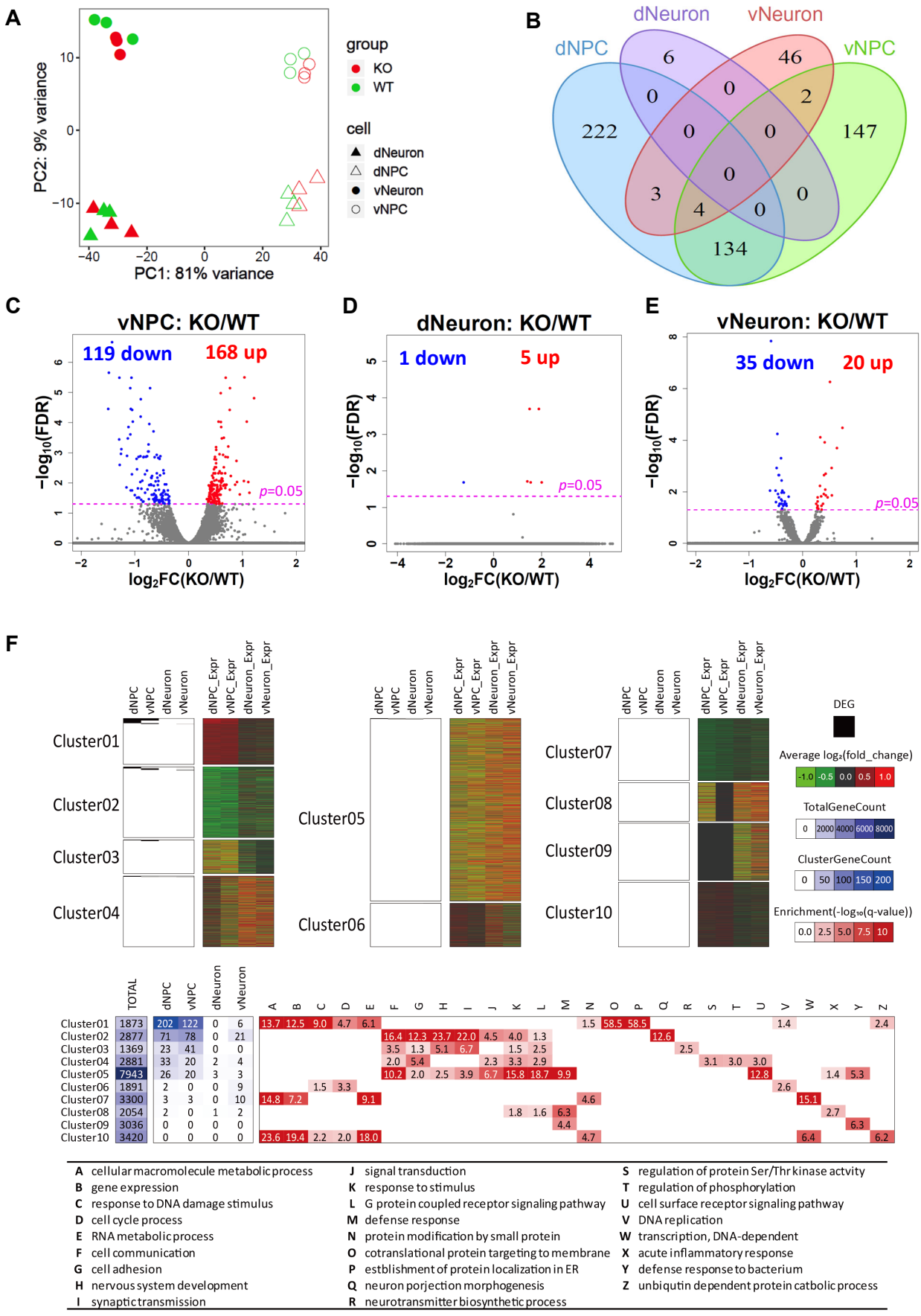
Supplemental figure S5



	position (hg19)	sequence	genic	testing result
1	chr3:+28131776	TATCATTAACTACAGGAGGTTAG	no	Good
2	chr4:+117213351	TGTTAAAATCTACAGGAGGTAAG	no	Good
3	chr2:-206193708	TTTAATTTCCTACAGGAGGTTGG	no	Good
4	chr12:+97469206	TGTTATAAACTACAGGAGGGTAG	no	Good
5	chr12:-47516665	GATTATAATCTCCAGGAGGTCAG	no	Good

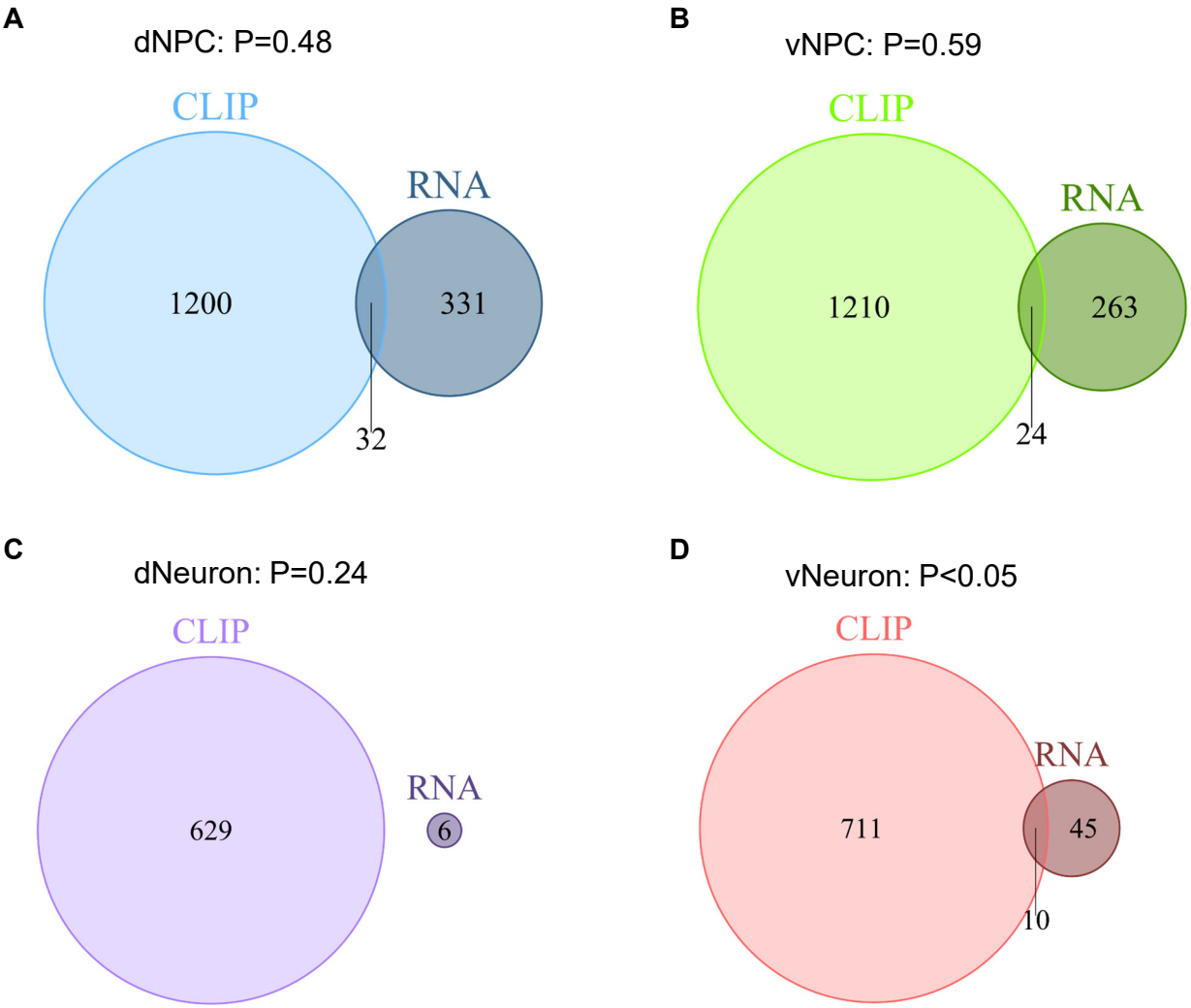
Supplemental figure S5. Characterization of FMR1-KO hPSCs and neural differentiation. (A) G-banding showing normal karyotypes of FMR1-KO hPSCs. **(B)** Immunofluorescent staining of FMR1-KO hPSCs for stem cell markers , OCT4 (POU5F1), SOX2, TRA-1-81 (PODXL), and a neural stem cell marker, PAX6 (scale bar 200 μm). **(C)** No off-target mutations detected in the FMR1-FLAG hPSCs by Sanger sequencing of top 5 predicted off-target sites. **(D)** Representative immunofluorescence of NPCs with dorsal marker, PAX6, and MGE marker, NKX2-1 (scale bar, 50 μm). **(E-F)** Quantification of percentage of PAX6-positive and NKX2-1-positive dNPCs and vNPCs (n=3-4 batches; mean ± SE). Dots represent different batches of neural differentiation. WT cells are the same as in Supplemental Figure S2A. **(G)** Representative immunofluorescence images of neurons for neuron marker, TUBB3, and FMR1 (scale bar, 100 μm). **(H)** Quantification of percentage of TUBB3-positive neurons (n=2 batches; mean ± SE).

Supplemental figure S6



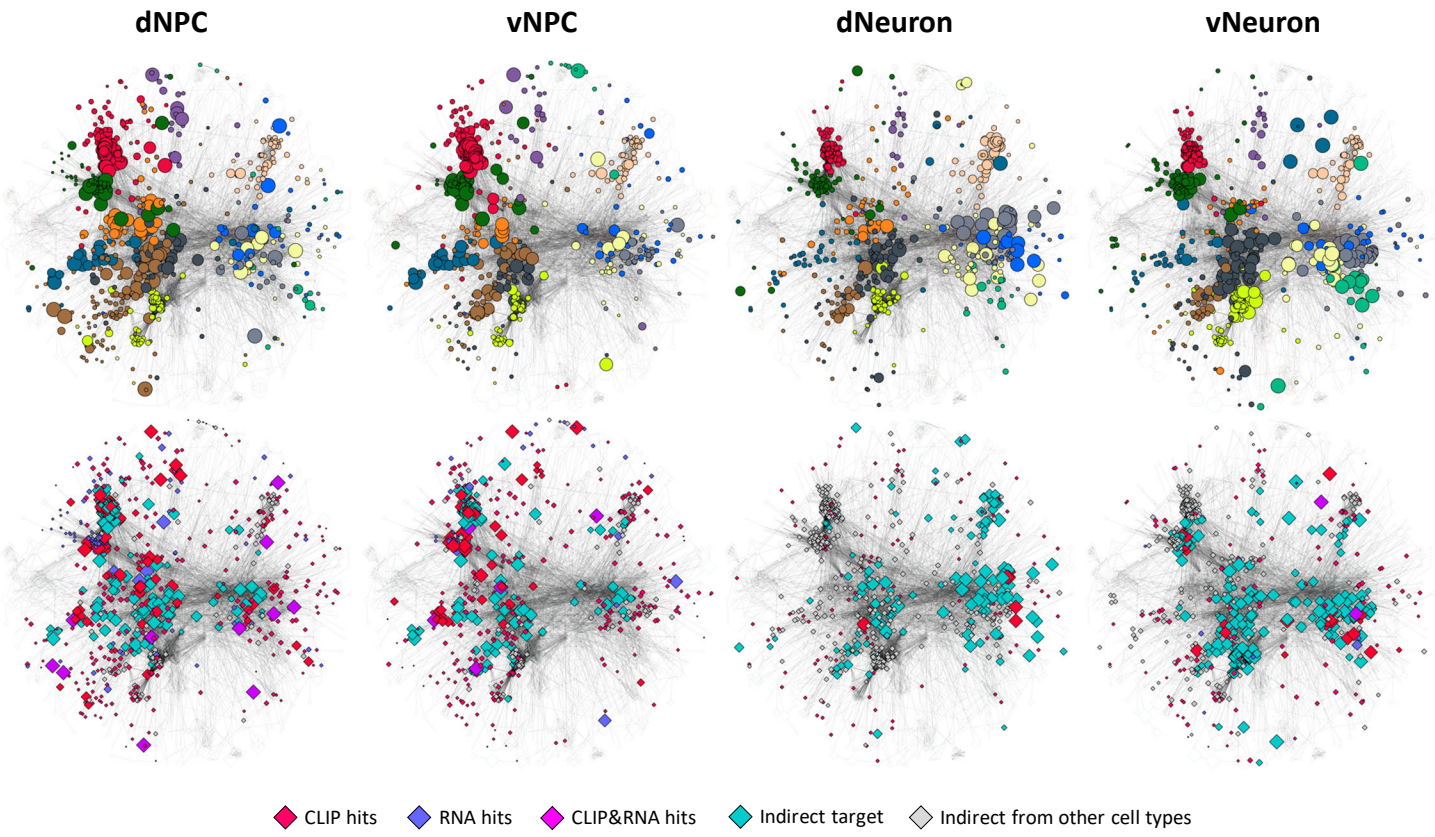
Supplemental figure S6. Analysis of RNA-seq data of WT and KO neural cells. **(A)** PCA analysis of RNA-seq data from all four cell types. **(B)** Venn diagram of DEGs of the four cell types. **(C-E)** Volcano plot of gene expression in WT and FMR1-KO vNPC **(C)**, dNeuron **(D)**, and vNeuron **(E)**. **(F)** Gaussian mixture model clusters of 30,644 genes clustered using log fold change in expression in FMR1-KO versus WT in the four cell types. For each cluster, the RNA-seq differentially expressed genes (DEGs) are shown on the side (white black heat map). Below are the number of genes in each cluster and DEGs in each our 4 cell types (blue heat map). Shown also are selected biological processes enriched in each cluster with the red intensity proportional to -$\log_{10}$(*q* value) (red heat map). Terms are selected based on their significance and relevance to neuronal processes.

Supplemental figure S7



Supplemental figure S7. Overlap between FMR1 targets and DEGs. (A-D) Venn diagrams showing overlap between FMR1 targets and DEGs in the four cell types: dNPCs **(A)**, vNPCs **(B)**, dNeurons **(C)**, vNeurons **(D)**. P values were calculated by hypergeometric test.

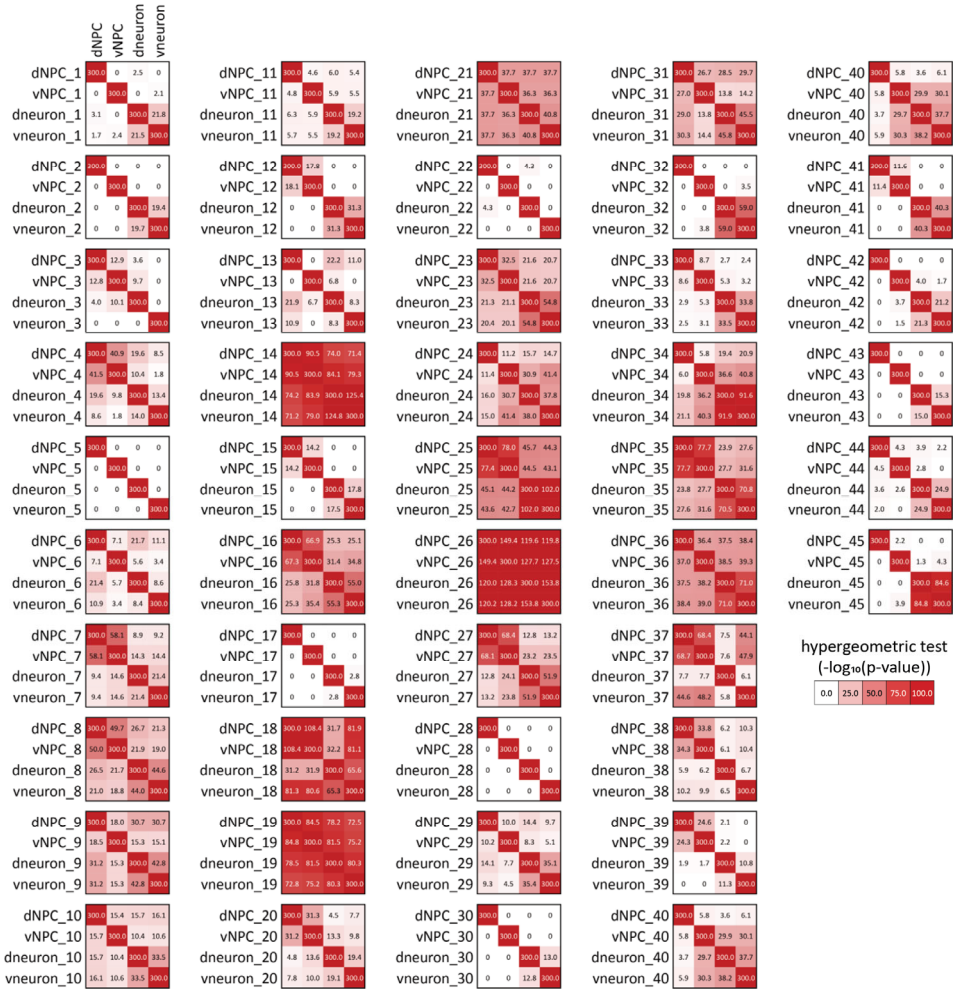
Supplemental figure S8



Supplemental figure S8. Cell type specific network clusters of CLIP and RNA-seq obtained from integrative network clustering. The top four networks are the same as in Fig 4. The bottom four networks have the same layout as the top four, with the node color depicting whether the node was in our RNA-seq or CLIP-seq hit set. The hit nodes are depicted as diamonds and the size of the node is proportional to the diffusion score (similar to Fig 4).

Supplemental figure S9

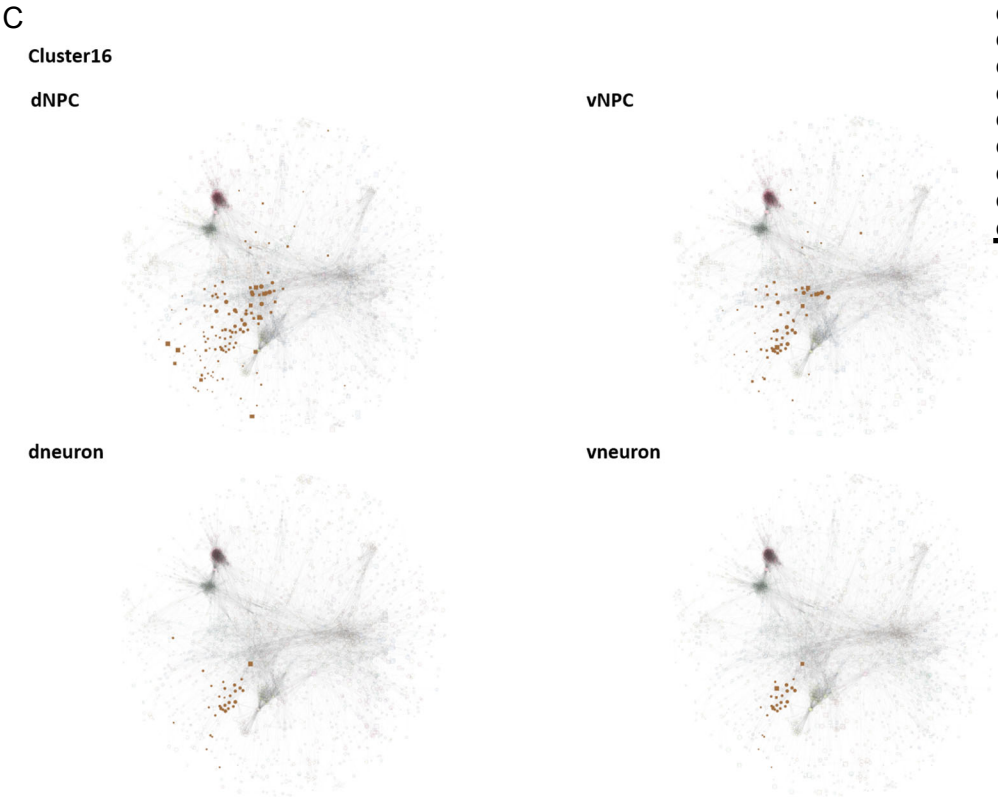
A



B

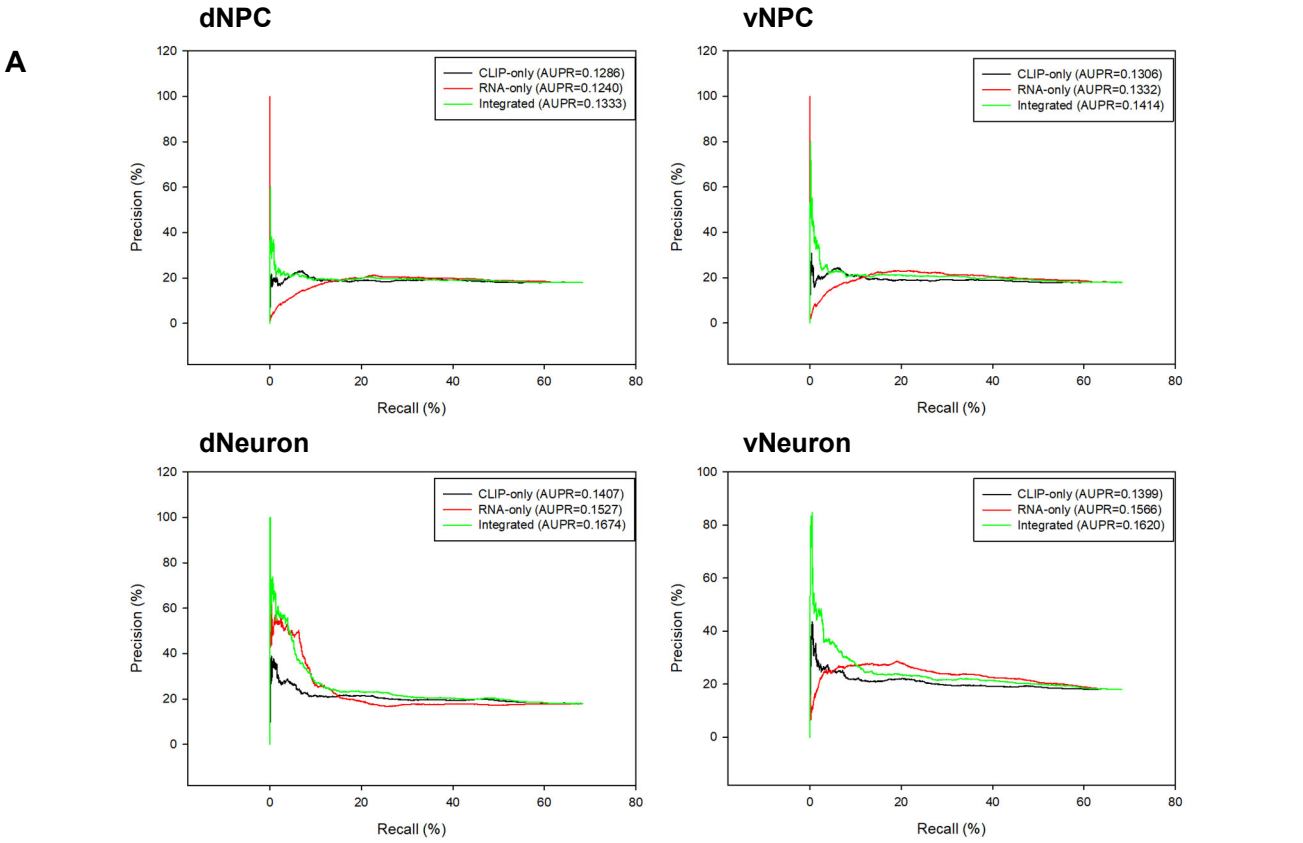
Cluster ID	Observed patterns
Cluster01	neuron-common
Cluster02	neuron-common
Cluster03	Individual
Cluster04	NPC-common, neuron-common
Cluster05	Individual
Cluster06	Common
Cluster07	common
Cluster08	NPC-common
Cluster09	common
Cluster10	common
Cluster11	dNPC-specific, neuron-common
Cluster12	neuron-common
Cluster13	vneuron-specific, dorsal-common
Cluster14	common
Cluster15	NPC-common, neuron-common
Cluster16	dNPC-specific
Cluster17	individual
Cluster18	common
Cluster19	common
Cluster20	NPC-common, neuron-common
Cluster21	common
Cluster22	individual
Cluster23	neuron-common
Cluster24	ventral-common
Cluster25	common
Cluster26	common
Cluster27	NPC-common, neuron-common
Cluster28	individual
Cluster29	neuron-common
Cluster30	neuron-common
Cluster31	common
Cluster32	neuron-common
Cluster33	NPC-common, neuron-common
Cluster34	neuron-common
Cluster35	common
Cluster36	common
Cluster37	dneuron-specific
Cluster38	NPC-common
Cluster39	NPC-common, dneuron-specific
Cluster40	common
Cluster41	NPC-common, neuron-common
Cluster42	individual
Cluster43	individual
Cluster44	individual
Cluster45	vNPC-specific, neuron-common

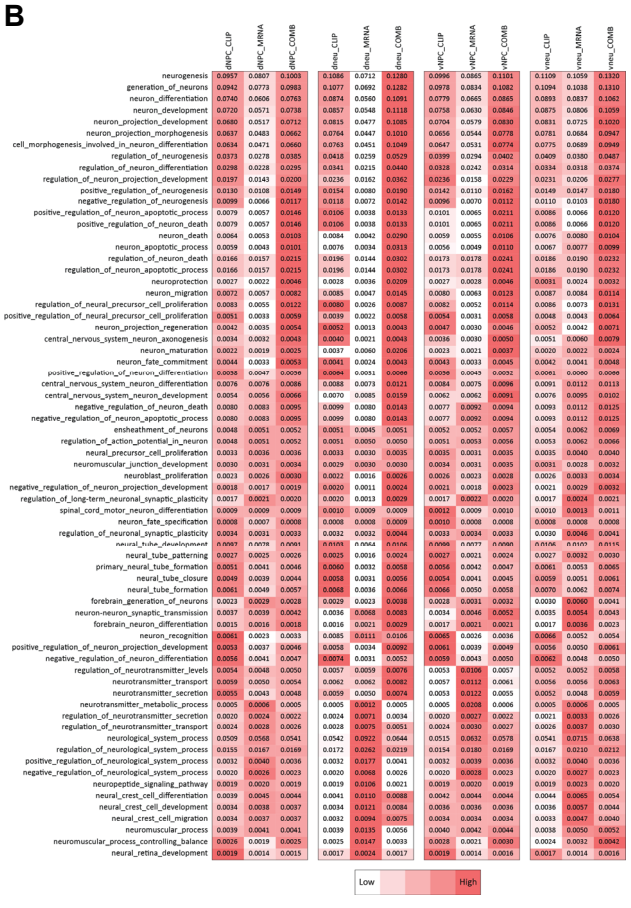
C

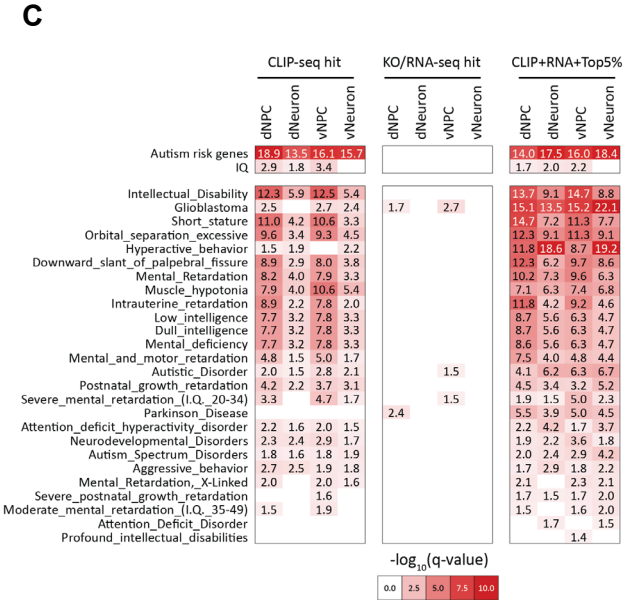


Supplemental figure S9. Similarity of clusters inferred using our multi-task graph clustering approach. (A) Each heat map shows the similarity of gene content in the mapped cluster across cell types for each of the 45 clusters identified from multi-task graph clustering. The overlap is assessed by using a hypergeometric test. Numbers indicate the average of $-\log_{10}(p \text{ value})$ of two hypergeometric tests, each treat one of the cell types as a background. These heatmaps together with the position of the clusters in the cell-type specific network was used to assign a pattern to each cluster ID (B). **(B)** Patterns assigned to each cluster based on the heatmap and their position of the clusters on the network. For example, “all common” is used when gene sets of each cell type for one cluster are all similar to each other. “NPC-common” is used when gene sets are similar specifically in the NPC cell lines. “specific” is used for a cell line A, when the cluster is similar in all other three cell lines but A, e.g., dNPC-specific means that the cluster is similar in vNPC, vneuron and dneuron, but is different for dNPC. “individual” is used when there is no significant overlap between any of the clusters across the cell lines. **(C)** Cluster16 depicted across all four cell types. Cluster16 exhibits a similar structure and organization in all but dNPC and therefore is called dNPC-specific.

Supplemental figure S10

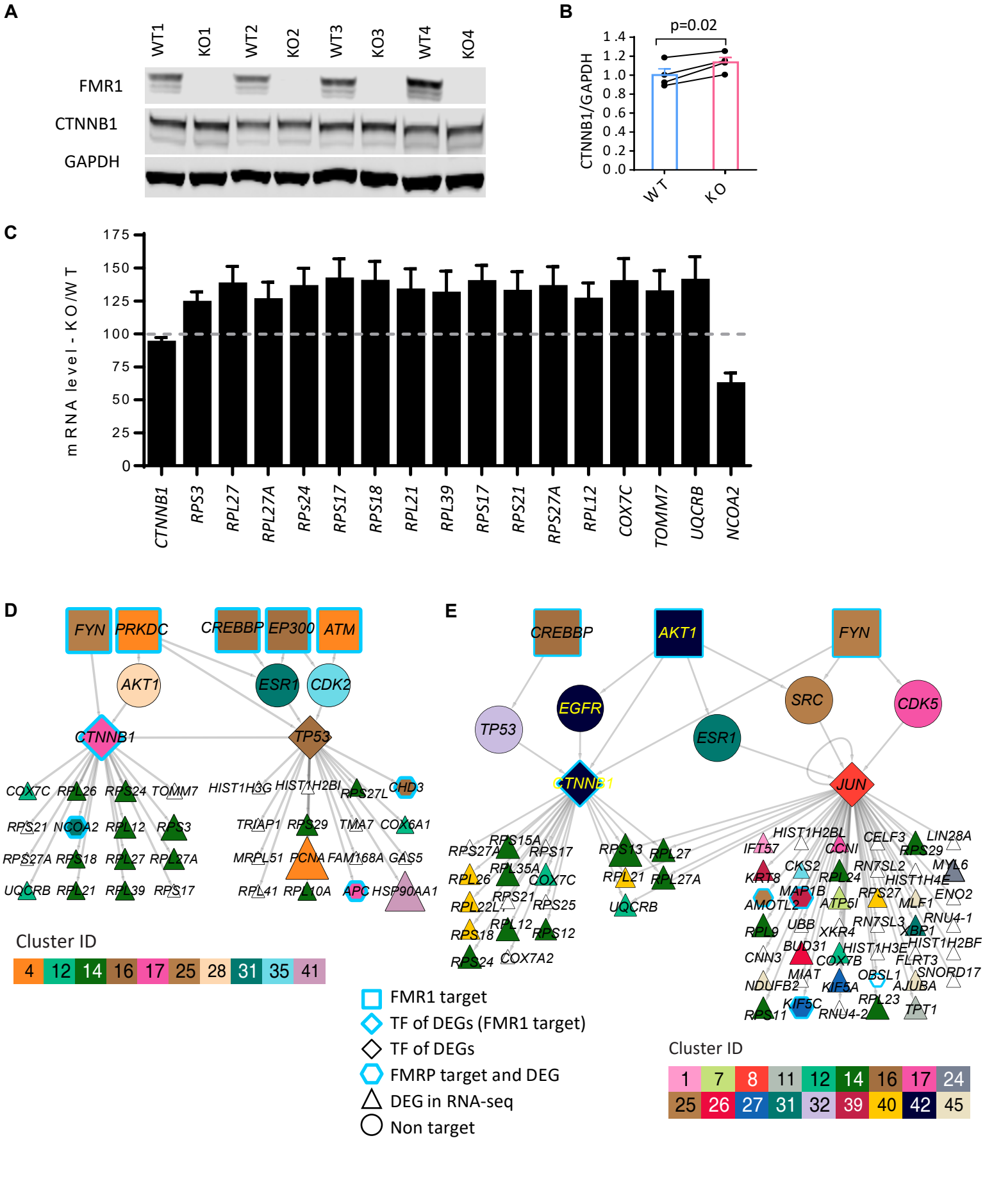






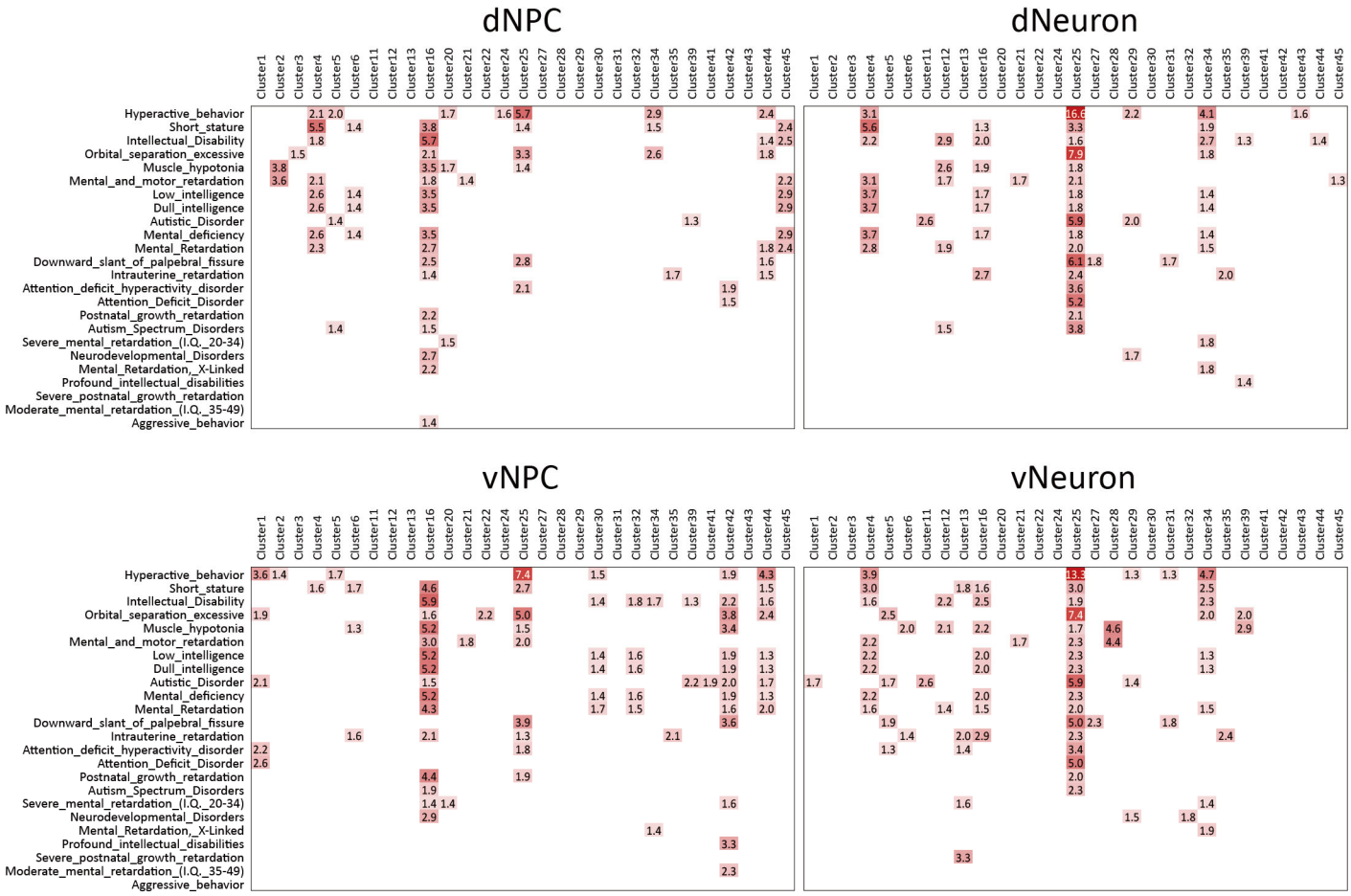
Supplemental figure S10. Evaluating the effect of integrating CLIP-seq and RNA-seq measurements for prioritizing genes. (A) Precision-recall curves and Area Under the Precision-Recall curve (AUPR) assessing the recovery of the union set of 2,227 genes annotated with neuronal functions, using prioritized gene lists from the integrated input (both CLIP-seq and RNA-seq), from CLIP-seq only input and RNA-seq only. **(B)** Table of AUPR values for 69 selected neuronal GO processes with sufficient number of genes. Four sets of columns are shown, each set corresponding to one of the 4 different cell types. Each set of columns corresponds to the 3 different inputs; CLIP-seq alone, RNA-seq alone and Integrated. The red intensity is proportional to the AUPR value, higher is better. **(C)** Expanded version of Figure 6 “Disease and phenotype enrichment of FMR1 targets and prioritized genes in human neural cells” to include the enrichment of genes prioritized using RNA-seq-only input genes.

Supplemental figure S11



Supplemental figure S11. Regulation of DEGs by FMR1 through direct targets. **(A)** Western blot analysis of CTNNB1 in WT and FMR1-KO dNPC. Numbers indicate batches of neural differentiation. **(B)** Quantification of **(A)** (mean ± SE; two tail t-test). Lines indicate paired WT and FMR1-KO cells. **(C)** qPCR quantification of fold changes of mRNAs in FMR1-KO vs WT dNPC (n=3; mean ± SE). **(D-E)** Additional examples of regulatory paths from FMR1 direct targets and DEGs. Shown are two pathways from dNPC (**D**) and vNPC (**E**). Node shapes depict FMR1 targets (cyan border), TFs (diamond node) and DEGs (triangles). Node sizes and colors are corresponding to diffusion score and cluster assignment, respectively.

Supplemental figure S12



Supplemental figure S12. Enrichment of neuronal diseases and symptoms in network clusters. Shown are the enriched neuronal diseases and symptoms from the DisGENET database in the network-based clusters for the four cell types, dNPC, vNPC, dNeuron and vNeuron. Terms shown are selected based on significance and relevance to neuronal cells. The intensity of red is proportional to -$\log_{10}$(q value). Enrichment of disease terms and symptoms were tested using FDR corrected hypergeometric test.

Supplemental methods

Culture of human pluripotent stem cells

Human embryonic cell (hESC) line H1 (WA01) and H13 (WA13) were obtained from WiCell (Madison, WI). GM1 (GM00498-4) is an iPSC line that was generated from fibroblasts from an apparently healthy 3 year old male obtained from Coriell (GM00498). Reprogramming was done using the Yamanaka factors (POU5F1, SOX2, KLF4, MYC) and retroviral vectors as described previously. Pluripotent stem cells were cultured on MEF feeder layers (WiCell) with a daily change of hESC medium of DMEM/F12 (Thermo Fisher Scientific), 20% knockout serum replacement (KSR, Thermo Fisher Scientific), 0.1 mM 2-mercaptoethanol (Sigma-Aldrich), 1x L-Glutamine (Thermo Fisher Scientific), 6 ng/ml FGF-2 (Waisman Biomanufacturing). Cells were passaged using 6 U/ml of dispase (Thermo Fisher Scientific) in hESC medium, washed and replated at a dilution of 1:5 to 1:10. G-banding was performed by WiCell Cytogenetics Services (Madison, WI), as described previously (Li et al. 2017).

Plasmids

Primers are listed in Supplemental Table S7. Plasmids expressing SgCas9 and sgRNA (pFMR1-sgE17A) for generating FMR1-FLAG hPSCs has been published previously (Li et al. 2017). Plasmid expressing SgCas9 and sgRNA (pFMR1-sgE3B) for generating FMR1-KO hPSCs was cloned by inserting DNA oligos, sgE3B-F and sgE3B-R (see Table below), to vector lentiCRISPR (Addgene # 49535; http://www.addgene.org/49535/) as described previously (Li et al. 2017). To construct a donor plasmid (pFMR1-donor FLAG) for inserting FLAG tag at C-terminus of FMR1, 5' and 3' homology arms (1072 bp and 909 bp, respectively) were amplified from genomic DNA of H1 hESCs using primers HA-L-F plus HA-L-R and HA-R-F plus HA-R-R, respectively. Sequence encoding FLAG tag was harbored in the 5' tails of primers HA-L-R and HA-R-F. The homology arms were inserted into vector (Addgene# 31938) digested by enzymes BamH I and Not I (New England Biolabs) using NEBuilder HiFi DNA Assembly Cloning Kit (New England Biolabs) following manufacturer's manual.

Generation of seamless FMR1-FLAG and FMR1-KO gene edited hPSC lines using a modified CRISPR-Cas9 gene editing procedure with transient puromycin selection

To generate FMR1-FLAG hPSC lines, H1, H13, or GM1 hPSCs were dissociated to single cells with TrypLE express (Thermo Fisher Scientific) and washed. 2-4 × 10^6 cells were electroporated (Gene Pulser Xcell, Bio-Rad; 250 V, 500 μF, 4 mm cuvette, infinite resistance) using 15 μg of pFMR1-sgE17A plasmid and 15 μg of donor plasmid pFMR1-donor FLAG (see Supplemental methods for plasmid construction). Because the Cas9-sgRNA plasmid carries a puromycin-resistant gene, cells were transiently selected with puromycin (Thermo Fisher Scientific; 0.5 μg/ml 48-72 hours after electroporation and 0.25 μg/ml 72-96 hours afterwards). About two weeks later, colonies were picked for expansion and screening for FLAG integration. FMR1-FLAG hPSCs were identified by PCR genotyping (primer pairs Red and Green Supplemental methods) and precise integration of FLAG was confirmed by Sanger sequencing (primers Red).

FMR1-KO hPSC lines were generated using similar protocol for generating FMR1-FLAG hPSCs but plasmid pFMR1-sgE3B (see Supplemental methods for plasmid construction) was used instead in electroporation of H1 and GM1 hPSCs. An additional donor ssODN (ssODN-FMR1 KO, 1nmol; see Supplemental methods for sequence) was also used in electroporation of H13 hPSCs. FMR1-KO hPSC clones were identified by PCR amplification followed by Sanger sequencing (primers FMR1 KO-F and FMR1 KO-R; see Supplemental Table S7 for sequences).

Neural differentiation

Neural induction was carried out using a dual SMAD method (Chambers et al. 2009) with modifications. In brief, 5 days after hPSCs were passaged onto MEFs, neural differentiation were induced by switching hESC medium to neural induction medium (NIM) of DMEM/F12:Neurobasal 1:1, 1x N2, 1x L-Glutamine, 1x Anti-Anti (GIBCO), 10 μM SB432542 (Selleck), 100 nM LDN193189 (Selleck), and 5 μM XAV-939 (Selleck). Cells were cultured in NIM for 9 days with a daily medium change. Cells were then dissociated with TrypLE and re-plated 1:1 on Matrigel-coated plates in neural progenitor cell (NPC) medium (Neurobasal medium, 1x GlutaMAX (Thermo Fisher Scientific), 1x N2, 0.5x B27 without vitamin A (Thermo Fisher Scientific), 1x Anti-Anti) supplemented with 10 μM ROCK inhibitor (Y-27632 dihydrochloride, Tocris). For dorsal/ventral patterning, cells were treated with either 5 μM cyclopamine or 10ng/ml SHH plus 1 μM purmorphamine (Maroof et al. 2013). Cells were patterned for 7 days with daily change of NPC medium. The NPCs were either collected for experiments (dNPC/vNPC) or re-plated for differentiation of neurons (dNeuron/vNeuron). For neuronal differentiation, NPCs were dissociated with TrypLE and re-plated on Matrigel-coated plates at 1:6 to 1:15 in NDM medium (NPC medium plus 200 μM ascorbic acid (Sigma-Aldrich), 1 μM cAMP (Sigma-Aldrich), 10 ng/ml BDNF (Peprotech), and 10 ng/ml GDNF (Peprotech)) supplemented with 10 μM ROCK inhibitor and 0.1 μM Compound E (Calbiochem). Half NDM medium was changed twice every week. Neurons were collected after 1 week unless otherwise indicated.

Antibodies used in this study

Antibody	Distributer	Cat #	Dilution
FMR1#1	Thermo Fisher Scientific	MA5-15499	1:1000 (WB/ IF); 10ug (CLIP)
FMR1#2	Santa Cruz	SC-21247	10ug (CLIP)
FMR1#3	Millipore	MAB2160	10ug (CLIP)
FLAG	Sigma-Aldrich	F1804	1:1000 (WB/ IF); 10ug (CLIP)
GAPDH	Thermo Fisher Scientific	TAB1001	1:2500
GAPDH	Thermo Fisher Scientific	MA5-15738	1:5000
PAX6	Biolegend	901301	1:500
NKX2-1	Abcam	76013	1:1000
TUBB3	Promega	G712A	1:2000
GABA	Sigma-Aldrich	A2052	1:5000
vGlut1 (SLC17A7)	Synaptic Systems	135-303	1:5000
OCT4 (POU5F1)	Santa Cruz	sc-5279	1:1000
SOX2	R&D System	MAB2018	1:1000
TRA1-81 (PODXL)	Millipore	MAB4381	1:1000
PIK3CB	abclonal	A0982	1:500
SEC24C	abclonal	A10797	1:3000
BACH2	bethyl	A305-502A-T	1:500
PLEKHA8	Proteintech	15410-1-AP	1:1000
STON2	NeuroMab	73-337	1:30
ZNF423	bethyl	A304-017A-T	1:500
AP2A1	abclonal	A6863	1:2000
CTNNB1	Millipore	06-734	1:1000
EIF4G1	abclonal	A6086	1:2000

CREBBP	abclonal	A17096	1:4000
KIF3B	abclonal	A15754	1:2000
FLNB	abclonal	A2481	1:1000
AEBP1	Santa Cruz	sc-271374	1:1000

Western blot
Primary antibodies used in this study are listed in the antibody table above. Cells were lysed in RIPA buffer (50 mM Tris, pH8.0, 150 mM NaCl, 1% NP-40, 0.1% SDS) supplemented with protease inhibitor (Sigma-Aldrich). After centrifugation for 15 min at 4 degrees, supernatants were quantified by Protein Assay Dye Reagent Concentrate (Bio-Rad). 20ug of total proteins were resolved by SDS-PAGE, transferred to nitrocellulose membrane, blocked with 5% nonfat milk, and probed with primary antibodies. Secondary antibodies conjugated with near infrared fluorescent dyes (IRDye 800CW or IRDye 680LT, LI-COR) were used at dilution of 1:10,000 for visualizing protein bands with an Odyssey Imager (LI-COR).

Immunofluorescence
Primary antibodies used in this study are listed in the antibody table above. Cells were fixed in 4% PFA for 10 min at room temperature. Then cells were washed with PBS and blocked with PBST (PBS containing 0.2% Triton X-100) plus 10% normal goat or donkey serum (Sigma-Aldrich), followed by incubation with primary antibodies diluted in PBST plus 5% normal serum for overnight at 4°C. Cells were then washed 3 x 5 min with PBS. Alexa Fluor secondary antibodies (Thermo Fisher Scientific) were diluted in PBST plus 5% serum and incubated with cells for 1 hour at room temperature. Cells were washed 2 x 5 min with PBS and counterstained with DAPI. Cells were then washed 2 x 5 min with PBS. Cells were scanned and quantified using Operetta Hi-content imaging system (PerkinElmer). Confocal images were collected with Nikon A1 confocal microscope.

RNA-seq
NPCs and neurons derived from control and FMR1-KO hPSCs of H1, H13, and GM1 lines were harvested in TRIzol (Thermo Fisher Scientific). Total RNA was isolated following manufacturer's manual. 1 ug of total RNA per sample was depleted of ribosomal RNAs and converted to strand-specific cDNA libraries using the TruSeq Stranded Total RNA library prep kit with RiboZero depletion (Illumina). Pool of barcoded libraries were submitted for single-end 100bp sequencing on Illumina HiSeq 2500 platform at the University of Wisconsin Biotechnology Center DNA Sequencing facility. Approximately 20 million total reads per library were obtained.

RNA-seq mapping and differential expression analysis
The human genome sequence (GRCh38.primary_assembly.genome.fa) and annotation file (gencode.v27.chr_patch_hapl_scaff.basic.annotation.gtf) were downloaded from GENCODE Release 27 (GRCh38.p10). The genome sequence and annotation file were used for generating genome indices by STAR (version 2.5.3a) using parameters "STAR --runMode genomeGenerate --genomeDir ./ --genomeFastaFiles ./GRCh38.primary_assembly.genome.fa --sjdbGTFfile ./gencode.v27.chr_patch_hapl_scaff.basic.annotation.gtf --sjdbOverhang 99". The demultiplexed RNA-seq FASTQ files were mapped to the genome indices by STAR using the parameters "STAR --runMode alignReads --genomeDir GRCh38 --readFilesIn FASTQ --clip3pAdapterSeq AGATCGGAAG --outFilterMultimapNmax 1 --outFilterMultimapScoreRange 1 --outFilterMismatchNoverLmax 0.06 --alignIntronMin 20 --alignIntronMax 1000000 --alignSJDBoverhangMin 1 --outFilterIntronMotifs RemoveNoncanonicalUnannotated --quantMode GeneCounts --outFilterType BySJout --outFilterScoreMin 10 --outSAMattrRGline ID:foo --

alignEndsType EndToEnd". Number of reads uniquely mapped to each gene were counted by STAR using the above parameter "--quantMode GeneCounts". Since the libraries were stranded (dUTP method), only reads mapped to the reverse strand were counted. Differentially expressed genes were analyzed using the R (v3.4.3) package DESeq2 (v1.16.1). Cell lines H1, H13, and GM1 were treated as biological replicates for the analysis and paired analysis was performed by defining the design formula of DESeq2 as "~ cell_line + genotype". FDR < 0.05 and no fold-change cutoff were set to identify differentially expressed genes.

Constructing 3'-adaptor

3'-adaptor for CLIP was constructed as described (Zarnegar et al. 2016). A custom synthesized oligonucleotide (5'-OH-AGATCGGAAGAGCACACGTCTGAAAAAAAAAA/iAzideN/AAAAAAAAAAAA/3Bio/-3') (IDTDNA) was converted to 3'-adaptor with 5' adenylation and internal IRdye label.

CLIP

Primers and oligos are listed in Supplemental Table S7. CLIP method was adapted from recently published irCLIP and eCLIP methods (Van Nostrand et al. 2016; Zarnegar et al. 2016). NPCs and neurons were gently rinsed with ice-cold PBS then crosslinked with 254nm UV-C at 0.35 J/cm^2. Following crosslinking, cells were collected in ice-cold PBS by cell scraping (NPCs) or gentle blowing-off (neurons) and centrifuged. Approximately 10-50 mg of cell pellet were lysed in 1ml of lysis buffer (50 mM Tris-HCl, pH 7.4; 100 mM NaCl; 1% Igepal CA630; 0.1% SDS; 0.5% Na-deoxycholate) supplemented with 5 µL Protease inhibitor cocktail III (EMD Millipore) and 5 µL Murine RNase inhibitor (New England BioLabs). After 15 min incubation on ice, cell lysates were sonicated using Bioruptor (Diagenode) at "High" setting for 6 cycles of 30 s on / 90 s off. Sonicated lysates were mixed with 2ul Turbo DNase (Thermo Fisher Scientific) and 30U RNase I (Thermo Fisher Scientific), mixed immediately, incubated in a Thermomixer at 1200 rpm at 37 °C for 5 min, and placed on ice for 5 min. Lysates were clarified by centrifugation at 20,000g for 20 min at 4 °C. The soluble fractions were transferred to fresh tubes and a 10 µL aliquot of the "FLAG" sample was transferred to a new tube as "SMI" control. The rest of the clarified lysates were added to antibody-conjugated Protein G Dynabeads (15 µg antibody and 100 µL beads per sample conjugated in lysis buffer with rotation at room temperature for 1 h then washed twice with lysis buffer) and rotated end-to-end for 2 h at 4 °C. Beads were then sequentially washed at 4 °C for 5 min each step with buffers: twice with 1ml high-salt wash buffer (50 mM Tris-HCl, pH 7.4; 1 M NaCl; 1 mM EDTA; 1% Igepal CA630; 0.1% SDS; 0.5% Na-deoxycholate), once with 500 µL wash buffer (20 mM Tris-HCl, pH 7.4; 10 mM $MgCl_2$; 0.2% Tween-20), and once with 200 µL PNK wash buffer (70 mM Tris-HCl, pH 6.5; 10 mM $MgCl_2$). Washed beads were treated with PNK in 200 ul PNK reaction (70 mM Tris-HCl, pH 6.5; 10 mM $MgCl_2$; 1 mM DTT; 4 µL T4 PNK (NEB); 4 µL Turbo DNase) incubated at 37 °C for 20 min in a thermomixer at 1200 rpm. PNK treated beads were washed at 4 °C twice with 500 µL high-salt wash buffer, twice with 500 µL wash buffer, and once with 100 µL RNA Ligase buffer (50 mM Tris-HCl, pH 7.4; 10 mM $MgCl_2$). On-bead 3'-adaptor ligation was performed by gently resuspending the washed beads in 30 µL ligation reaction (50 mM Tris-HCl, pH 7.4; 10 mM $MgCl_2$; 0.8 µL DMSO; 9 µL 50% PEG 8000; 0.4 µL Murine RNase Inhibitor; 10 pmol 3'-adaptor; 2.5 µL T4 RNA ligase 1, high concentration, NEB) and rotating end-to-end overnight in dark at room temperature. The next day, the beads were washed at 4 °C once with high-salt wash buffer and twice with wash buffer.

The washed beads were resuspended in 20 µL 1x SDS-PAGE loading buffer (1x NuPAGE LDS Sample buffer and 1x NuPAGE Sample Reducing Agent, Thermo Fisher Scientific) and the 10 µL SMI were mixed with 10 µL 2x SDS-PAGE loading buffer. The samples were denatured for 10 min in a thermomixer at 1200 rpm, 70 °C. The denatured samples were magnetically separated and supernatants were resolved in NuPAGE Novex 4-12% Bis-Tris Protein gels running in ice-cold NuPAGE MOPS SDS Running buffer at 200 V for 1 h. The resolved samples were transferred

to nitrocellulose membrane at 400 mA for 1.5 h in ice-cold NuPAGE Transfer buffer with 10% methanol and 0.1% SDS. RNAs on the nitrocellulose membrane were visualized by scanning the membrane using an Odyssey Imager. Membranes were placed on wet filter paper and 90-200 kDa region was excised with the guidance of prestained protein standards (BIO-RAD) between samples. Excised membranes were cut into 1-2 mm slices and transferred to tubes with 200 uL proteinase K solution (100 mM Tris-HCl, pH 7.4; 50 mM NaCl; 10 mM EDTA; 20 µL Proteinase K, Thermo Fisher Scientific). Membranes were incubated in a thermomixer for 20 min at 1200 rpm, 37 °C. 200 µL urea solution (100 mM Tris-HCl, pH 7.4; 50 mM NaCl; 10 mM EDTA; 7 M urea) was added to each tube and further incubated for 20 min at 1200 rpm, 37 °C. 400 µL acid phenol/chloroform/isoamyl alcohol (pH 6.5) was added to each tube, mixed well by shaking, and then incubated in a thermomixer for 5 min at 1200 rpm, 37 °C. The tubes were centrifuged briefly and liquids were transferred to Phase Lock Gel Heavy tubes (5Prime). Phase lock tubes were centrifuged at 13000g for 2 min at room temperature. 400 µL chloroform was added to the phase lock tubes, gently inverted 10 times, and centrifuged again. Aqueous phase was transferred to fresh Eppendorf tubes and mixed with 45 µL 3M NaAc (pH 5.2), 2 µL linear acrylamide (Thermo Fisher Scientific), and 1.1 mL EtOH. RNA was chilled overnight at -80°C, precipitated by centrifugation, washed with 80% ice-cold EtOH, air-dried, and resuspended in 5 µL water.

Processing SMI RNAs

Purified SMI RNAs were mixed with 100 µL PNK solution (70 mM Tris-HCl, pH 6.5; 10 mM $MgCl_2$; 1 mM DTT; 1 µL Murine RNase Inhibitor; 2 µL T4 PNK; 2 µL Turbo DNase) and incubated in a thermomixer for 20 min at 1200 rpm, 37 °C. RNAs were cleaned up using MyOne Silane beads (Thermo Fisher Scientific). RNAs resuspended in 5 ul water were then mixed with 1 µL DMSO and 0.5 µL of 10 µM 3'-adaptor and denatured by incubating at 70 °C for 2 min followed by immediately placing on ice for 2 min. Denatured RNAs were mixed with 13 µL ligation mix (50 mM Tris-HCl, pH 7.4; 10 mM $MgCl_2$; 1 mM DTT; 0.2 µL Murine RNase Inhibitor; 8 µL 50% PEG 8000; 1.3 µL T4 RNA Ligase, high concentration) and incubated for 2 h at room temperature. Ligated SMI RNAs were purified again with MyOne Silane beads and resuspended in 5 µL water.

CLIP library preparation

Primers and oligos are listed in Supplemental Table S7. RNAs ligated with 3'-adaptors were mixed with 1 µL of 1 µM CLIP-RT primer and denatured by incubating at 70 °C for 2 min followed by immediately placing on ice for 2 min. Denatured RNAs were mixed with 8.4 µL water, 4 µL of 5x RT buffer (100 mM Tris, pH 7.5; 50 mM $MgCl_2$; 250 mM NaCl; 25 mM DTT; 0.1% Tween-20) and 0.3 µL Murine RNase Inhibitor, and 0.3 µL TIGRT-III Enzyme (InGex) and were incubated at room temperature for 30 min. Then 1 µL of 25 mM dNTP was mixed with the RT reaction and incubated at 60 °C for 2 h. RNA-cDNA hybrids were cleaned up by incubating with 5 µL MyOne Streptavidin C1 beads (Thermo Fisher Scientific) with rotation at 4 °C for 30 min followed by washing 2x 5 min with high-stringency buffer (20 mM Tris, pH 7.5; 120 mM NaCl; 25 mM KCl; 5 mM EDTA; 1% Trition-X100; 1% Na-deoxycholate) and 2x 5 min with PBS. Beads were resuspended in 7 µL of cDNA elution buffer (2.25 µL water; 1 µL of 1 µM Elute-R; 3 µL of 5 M Betaine; 0.75 µL of 50 mM $MnCl_2$) and placed in a thermocycler with the following program: 95 °C 5 min, 75 °C 1min, ramp -0.1°C/s to 60°C hold for 10 min. Tubes were mixed with 8 µL of circularization solution (6 µL water, 1.5 µL CircLigase II ssDNA Ligase buffer, 0.5 µL CircLigase II ssDNA Ligase; Epicentre) and incubated in a thermocycler at 60°C for 16 hours. Circularized cDNAs were mixed with 1 µL Exonuclease I (Thermo Fisher Scientific) and incubated 37°C 30 min then 80°C 15 min to denature. 32 µL of Ampure XP beads and 32 µL of EtOH were added to the cDNAs and incubated for 10 min with mixing by pipetting every 5 min. Beads were placed on a magnetic stand for 10 min and washed twice with 80% EtOH, air-dried, resuspended in 14 µL water for 5 min, and 11 µL supernatant were transferred to new tubes. Purified cDNA were mixed with 0.25 µL of 5x SYBR Green I (Thermo Fisher Scientific), 0.5 µL of ROX Reference Dye (Thermo Fisher Scientific), 0.5

µL each of 25 µM primers (PCR1-F and PCR1-R), and 12.5 µL Q5 Hot Start High-Fidelity 2X Master Mix (NEB) and cDNA libraries were amplified in a thermocycler with the program: 98°C 30 s; cycles of 98°C 10 s, 64°C 30 s, 72°C 30 s. Samples were removed after qPCR finished exponential phase and were cleaned up by adding 2 volumes of AmpureXP beads and 1 volume of EtOH. Barcoded adaptors were added to the purified amplicons by a second round of PCR by adding Q5 Hot Start High-Fidelity 2X Master Mix, primer seqF, and barcoded primer seqR (CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCC GATCT; NNNNNN represents barcode). The second round of PCR was performed in a thermocycler with the following program: 98°C 30 s; 3 cycles of 98°C 10 s, 64°C 30 s, 72°C 30 s; 72°C 2 min; 4°C hold. Barcoded libraries were purified using 1.5 volume of AmpureXP beads. Purified libraries were submitted for single-end 100bp sequencing on Illumina HiSeq 2500 platform at the University of Wisconsin Biotechnology Center DNA Sequencing facility.

Analysis of CLIP-seq data

The demultiplexed FASTQ files were first quality filtered using fastq_quality_filter of FASTX Toolkit (v0.0.14) with parameters "-q25 -p80". Adapter sequences were then removed from reads using cutadapt (v1.7.1) with parameters "-n 3 -e 0.1 -O 5 -q 6 -m 20 -a AGATCGGAAGAGCACACG". PCR duplicates were removed using fastx_collapser of FASTX Toolkit. Finally 5' adapter sequences were trimmed from reads using fastx_trimmer of FASTX Toolkit with parameters "-f 13 -l 94". Processed reads were mapped to the genome indices above using STAR (v2.5.3a) with parameters "STAR --runMode alignReads --runThreadN 8 --genomeLoad LoadAndKeep --genomeDir $genome --readFilesIn $file --outSAMunmapped None --outFilterMultimapNmax 1 --outFilterMultimapScoreRange 1 --outFilterMismatchNmax 2 --outFilterMismatchNoverLmax 0.1 --alignIntronMin 20 --alignIntronMax 1000000 --alignSJDBoverhangMin 1 --outFilterIntronMotifs RemoveNoncanonicalUnannotated --outSAMtype SAM --quantMode GeneCounts --outFileNamePrefix $file_ --outSAMattributes All --outFilterType BySJout --outSAMattrRGline ID:foo --alignEndsType EndToEnd". Number of reads uniquely mapped to each gene were counted by STAR using the above parameter "--quantMode GeneCounts". Only reads mapped to the forward strand were counted. Enriched genes in FLAG sample over WT sample or SMI sample were analyzed using the R package DESeq2. Cell lines H1, H13, and GM1 were treated as biological replicates for the analysis and paired analysis was performed by defining the design formula of DESeq2 as "~ cell_line + sample". FMR1 targets were defined as significantly enriched in FLAG samples over both WT sample and SMI samples: padj(FLAG/WT) <0.05 & $\log_2$FC(FLAG/WT)>0 & padj(FLAG/SMI) <0.05 & $\log_2$FC(FLAG/SMI)>0. Actual cutoffs of fold-change were in Supplemental Table S1.

The CLIP-seq data were also analyzed using CLIPper (Van Nostrand et al. 2016)and PureCLIP (Krakau et al. 2017)(Supplemental Table S1D). CLIPper pipeline was adapted from the ENCODE eCLIP-seq processing pipeline v2.0. CLIPper peaks normalized over SMI and WT inputs by IDR analysis. The peaks with fold change > 2 and p values <0.05 in at least two replicates were used to define CLIPper lists of FMR1 target genes. For PureCLIP analysis, binding sites identified in at least two replicates by PureCLIP were used to generate PureCLIP lists of FMR1 target genes.

Motif analysis

UV-crosslinking step of CLIP results in information of protein binding sites, defined as RT-stops (Konig et al. 2010). Bam files of uniquely mapped reads were converted to bed files using bedtools. The bed files were then converted to RT-stops (1 basepair) and intersected with a reference bed file of genes of FMR1 targets. The resultant twelve files (4 cell types x 3 replicates of FLAG group) were merged and frequencies were counted using bedtools merge. High-confident binding sites with 24 counts or more were selected for motif analysis using Homer with parameters as follows: findMotifsGenome.pl hg38r -rna -size -25,25 -len 4.

RNA immunoprecipitation and qPCR

Primers are listed in Supplemental Table S7. RNA-IP was performed as described (Li et al. 2016). Briefly, cell pellets of WT and FMR1-FLAG NPCs were harvested and homogenized in 1 ml of ice-cold lysis buffer [10 mM Hepes (pH 7.4), 200 mM NaCl, 30 mM EDTA, and 0.5% Triton X-100] with Ribonuclease inhibitors (Roche) and 2× complete protease inhibitors (Boehringer-Mannheim). Cell lysates were centrifuged at 20000g for 20 min at 4°C. Supernatant was incubated with a FLAG antibody (ThermoFisher) for 2 hours at 4°C then with washed protein G dynabeads at 4°C for additional 2 hours. After three washes with lysis buffer, the immunoprecipitate was resuspended in TRIzol (ThermoFisher) for total RNA isolation following manufacturer's manual. 0.2-1ug of total RNA was converted to cDNA in a 20ul reaction using Transcriptor First Strand cDNA Synthesis Kit (Roche) and oligo dT primer. qPCR was performed using iTaq Universal SYBR Green Supermix (Bio-Rad) and StepOnePlus Real-Time PCR System (Thermo Fisher Scientific).

Plots

PCA was analyzed using the R package DESeq2. Reads coverage on 5'UTR, CDS, and 3'UTR of genes were generated using geneBody_coverage2.py in RSeQC (v2.6.4). Venn diagrams were generated using the R package VennDiagram (v1.6.17). Scatter plots were generated using the basic plot function in R.

GMM clustering of CLIP-seq and RNA-seq data:

To identify the gene clusters based on the experimentally measured values from CLIP-seq and RNA-seq, we represented each gene by a 4-dimensional vector denoting the average $-\log_2$(fold change) values of replicate measurements from each of the 4 cell types. For CLIP-seq data, we used 2,620 genes for clustering, comprising both targets identified from our experiments and targets previously reported from mouse studies. For the RNA-seq data, we clustered 30,644 genes, which included all genes measured in the four cell types together with targets identified from Drosophila. We included all genes for the RNA-seq data because there were relatively few targets identified using RNA-seq.

We applied multi-variate Gaussian mixture models to cluster (Trevor et al. 2009), where the mean vector of each mixture component was 4-dimensional and the covariance matrix was diagonal. We used k=10 and k=20 as the number of clusters and used Silhouette index with squared Euclidean distance to determine the final k. The Silhouette index is used to assess the quality of clustering and ranges from -1 to 1, with more negative the number the worse the clustering. For both CLIP-seq and RNA-seq data k=10 had a better Silhouette index (CLIP-seq: -0.0716, RNA-seq: -0.2833) than k=20 (CLIP-seq: -0.1142, RNA-seq: -0.3076). Therefore, we performed all downstream analysis with k=10 clusters.

Integrative network-based clustering analysis to identify cell type specific networks from CLIP-seq and RNA-seq data.

To define cell type-specific subnetworks using the CLIP-seq data, we developed a novel graph-based clustering algorithm which consists of two steps: (1) graph diffusion to define cell type specific networks, (2) multi-task graph clustering.

Defining cell-type specific network structure**:** To define cell-type specific networks, we integrated the FMR1 targets identified in a particular cell type in a two-step graph diffusion approach. This

approach requires a skeleton background network and a query set of genes. First, we applied network node diffusion using the FRMP targets identified in a cell type as query nodes. Second, we used the diffused node values to carry out another diffusion, tracking the diffusion weights on each edge of the graph. The node diffusion process ranks all other genes in the network based on their global connectivity to the input set thus providing a measure of influence of the input nodes on all other nodes of the network. The global connectivity is in turn measured using a graph diffusion kernel. We use the regularized Laplacian kernel (Smola and Kondor 2003), which has been used previously for network based ranking of genes (Kohler et al. 2008) and for semi-supervised classification tasks (Fouss et al. 2012). Briefly, the regularized Laplacian kernel is defined as $K = (1 + \lambda L)^{-1}$. Here, L is the symmetric normalized Laplacian and is defined as $L = I - D^{\frac{1}{2}}AD^{\frac{1}{2}}$, where A is the adjacency matrix of the background network G, and D is a diagonal matrix giving the degree of each node. Next, we encode the query gene set using a binary indicator vector q, which has one element per gene in G. Query genes have value 1, and non-query genes have value 0. Finally, the diffusion score for a gene n_i is computed as $K_i q$, where K_i is the i^{th} row. The kernel has a parameter λ, which specifies the width of the kernel. We selected λ from candidate values {0.1, 1, 5, 10} using leave-one-out cross-validation. We scored each possible value of λ as the area under the precision-recall curve that was generated as follows: we ranked each gene according to its diffusion score when that gene was not included in the query set; that is, its entry in *q* was set to 0. At a diffusion score *s*, *precision* is computed as the fraction of genes with score ≥*s* that are in the query set, and recall is computed as the fraction of the genes in the query set that have score ≥*s*. The settings of λ for the different gene sets were as follows: 1 for CLIP and 5 for RNA-seq in dorsal NPC, 5 for CLIP and 0.1 for RNA-seq in ventral NPC, 1 for CLIP and 10 for RNA-seq in dorsal neuron, 5 for CLIP and 10 for RNA-seq in ventral neuron. Once we have the diffused values from CLIP-seq hits and RNA-seq hits for each cell type, we transformed the diffused values into percentile ranks of all nodes and combined two rankings (from CLIP and RNA-seq) into an average rank.

After the first node diffusion, we carried out edge diffusion, which enabled us to get cell type specific weighted graphs where the edge weight between node i and j was proportional to the influence of node i on j. For this step we used the insulated heat diffusion kernel (Vandin et al. 2011; Leiserson et al. 2015) for estimating the effect of one node on its neighbor nodes based on their global connectivity. Briefly, the insulated heat kernel is defined as $K = \beta(I - (1 - \beta)W)^{-1}$. W is a transition matrix which is defined as $W = A \cdot D^{-1}$, where A is the adjacency matrix of the background network G, and D is a diagonal matrix with entries the degree of each node of graph G. The kernel has a parameter β, which specifies the retention rate of the kernel, was set as β =0.5 in this study. We encode the query gene set as a diagonal matrix D_q with entries vector *q* of diffused node values of previous step. The final diffused matrix is $H = K \cdot D_q + (K \cdot D_q)^{-1}$, which is a symmetric matrix because our gene network is an undirected graph. We then converted the weighted adjacency matrix into a diffusion-state distance (DSD) matrix (Cao et al. 2014), P, which entails computing the Euclidean distance between every pair of vertices and using the distance as the value of entry $P(i, j)$. Finally, the distance matrix, P, is converted into a similarity matrix *S* via a Gaussian kernel defined as $S = \exp(-P^2/2{\sigma_P}^2)$, where σ_P is a standard deviation of P.

For both diffusion steps, as the background we use the largest connected component of STRING network (Szklarczyk et al. 2011) a total of 8,076 genes and 85,721 interactions. At the end of these two diffusion steps we obtain a cell type specific weighted network for each cell type where the weights correspond to the similarity values calculated in the last step.

Multi-task graph clustering: To identify subnetworks in each of the cell types individually, we applied a multi-task graph-based clustering algorithm, *Arboretum-HiC*, that we previously developed to cluster Hi-C matrices (Fotuhi Siahpirani et al. 2016). This algorithm finds the network

clusters in multiple cell types by simultaneously applying spectral clustering (von Luxburg 2007) to each cell type specific network while incorporating the relatedness of the cell types. To obtain the relatedness, we generated a 4 × 4 matrix specifying the pairwise Jaccard coefficients between the sets of FMR1 targets from each of four cell types and applied hierarchical clustering to draw a dendrogram. The tree put both NPCs close together and neurons close together. A key property of this multi-task learning framework is that there is a mapping of clusters from one cell type to another. Therefore, cluster i in one cell type corresponds to cluster i in another cell type. For spectral clustering of each cell type-specific network, we used eigenvector matrices of the regularized graph Laplacian, $L_\tau = I - D_\tau^{\frac{1}{2}} A D_\tau^{\frac{1}{2}}$, where A is the adjacency matrix and D_τ is a regularized diagonal matrix defined as $D_\tau = D + \mu_D$, where μ_D is a mean of D (Xu et al. 2018) (Zhang and Rohe 2018). A in turn corresponded to the similarity matrix derived from the DSD matrix spanning nodes including the original CLIP and RNA-seq targets and the union of the top 5% of the non-input set of nodes from each of the four cell types, comprising a gene set of 1,984 genes and 17,540 interactions. We further filtered this set to exclude genes that were singletons or part of smaller connected components resulting in a set of 1,810 genes and 17,490 interactions. We used the rule of thumb to choose the number of clusters as k=45, rounded up to the nearest multiple of 5. The output of Arboretum-HiC was 45 clusters in each of the four cell types. Network images were generated using Cytoscape v3.5.1 (Shannon et al. 2003).

Pattern annotation of the clusters

We annotated each cluster with a pattern based on the conservation properties of the clusters across the four cell types (Supplemental Figure S9). Specifically, we defined common and cell type specific patterns for each cluster ID, using Jaccard coefficient as well as a hypergeometric test to assess the significance of overlap between the clusters from each cell type. A “common” cluster was that which had significant overlap between all four cell types. A NPC/neuron-common or dorsal/ventral-common cluster was defined as one which had significant overlap between NPC/neuron or dorsal/ventral but not all four. For example, “NPC-common” cluster was defined as that which had significant overlap between dorsal and ventral NPCs, and likewise, a “dorsal-common” clusters had overlap between dorsal NPC and dorsal neuron, etc. A cell type-specific cluster (e.g. dNPC-specific) was defined as one with significantly different genes compared to the other three cell types which in turn were significantly overlapping (e.g., Cluster 37, which is dneuron-specific in Supplemental Figure S9B). Finally, “individual” cluster was one where the cluster from each cell type were distinct from each other with little or no overlap between them (e.g., Cluster 5 in Supplemental Figure S9B).

Enrichment analysis of the clusters

To interpret the clusters we tested them for enrichment of Gene Ontology biological processes using FDR corrected hypergeometric test. We used an FDR<0.05 to select enriched terms. We manually inspected the terms and selected terms that were representative by simultaneously considering (1) higher significance in enrichment (i.e. lower FDR), (2) correspondence to the pattern we observed in the pattern annotation of the clusters and (3) relationship to neurological processes. We also tested these gene sets in each cluster for enrichment in genes associated with different diseases from the DisGeNET database (Piñero et al. 2017) using the same FDR based hypergeometric test.

Statistical significance test for differential expressed genes (DEG) assigned in Gaussian mixture model (GMM) clusters of FMR1-KO cell line (Supplemental Figure S6F)

To test the statistical significance of up- or down-regulation signature among the four cell types in each of the 10 GMM RNA-seq clusters, we applied a two-sided *t*-test to test if the means of the gene expression levels of genes in a cluster from one cell type is different from another cell type. The differences between NPC cells (dorsal and ventral) and neurons (dorsal and ventral) are much more significant (P < 5.0E-5) than the expression levels of genes among the two NPCs and two neurons in all clusters except only dNPC and dNeuron of cluster 6.

Comparison of multi-task graph clustering results based on perturbed input hit genes and randomized networks (Supplementary Table S5A)

To investigate the stability of the inferred clusters as a function of a noisy input gene set, we generated 5 input gene sets in which 20% of the input genes were randomly selected and hidden ("5-CV"). We performed node diffusion and edge diffusion followed by multi-task graph clustering using the same hyperparameters as the original input set (k=45). Next we compared the clusters obtained from these 5-CV sets to those from the original input set by using a Jaccard coefficient based overlap score. Briefly let i be a cluster from one of the 5-CV clusterings. We matched cluster i to a cluster j from the original input set such that the Jaccard coefficient between cluster i and j is maximal. We took the average of these 45 Jaccard coefficients. We repeated the procedure considering a cluster i from the original input set and finding its best matching cluster from the 5-CV set and took an average of these coefficients. The final similarity was the average of these two averages.

To assess the contribution of the CLIP-seq and RNA-seq datasets for defining the clusters, we repeated our diffusion and clustering steps using the CLIP-seq alone hits and the RNA-seq alone hits. We compared the resulting clusters with those from the original input set again using the Jaccard coefficient based overlap.

To assess the significance of our clusters compared to random clusters as well as to study the utility of integrating RNA-seq with CLIP-seq data, we randomized the network given as input to the multi-task clustering by permuting the node labels but keeping the structure the same. We applied graph clustering on the randomized network and compared these clusters to those from the input set. We found a substantially low overlap in clusters obtained after randomization compared to those from the input set suggesting that the network structure greatly determines the grouping of the nodes. However, the similarity of the 5-CV sets to the clusters from the original input set was much higher than random suggesting our analysis robust to perturbations in the input set. The overlap of clusters derived from CLIP only or RNA-seq only was lower than the 5-CV set, but significantly higher than the random sets. This suggests that both RNA-seq and CLIP-seq are important for our clusters.

Assessing the ranking of genes identified using network information flow method. (Supplementary Table S5B, Supplementary Figure S10)

To assess the significance of the ranking of genes identified from network information flow, we generated 10 random gene sets as follows: for each of the four cell types, we selected as many genes as there were in the CLIP-seq set or RNA-seq set and then took a union of these gene sets. We performed network diffusion on this set to obtain ranked lists of genes. In addition, we

also evaluated the rankings obtained from the five 5-CV input gene sets (described above), using CLIP hits alone and using RNA-seq hits alone.

We took top 1% genes from the prioritized list of each cell type in each input and took the union to produce prioritized gene sets of 195-393 genes. We next compared these gene sets with the union of the top 1% genes from our original input set. We considered all prioritized genes including the input set as well as the novel set of genes that were in the top 1% but not in the input set. Statistical significance of overlap between gene sets from the original and random gene sets is obtained based on the hypergeometric test. Prioritized gene sets from the 5-CV input sets had the most significant overlap with the set obtained from the original input set followed by CLIP-seq and RNA-seq alone. The overlap between prioritized genes from random input genes and those from the original input set was least significant. This suggests that the prioritized genes using our approach are substantially different from random, leverage both RNA-seq and CLIP-seq signals and are robust to small perturbations to the input set.

To further evaluate the effect of integrating both CLIP-seq target genes and KO RNA-seq DEGs, we assessed the recovery of genes annotated with neuronal processes in Gene Ontology based on the rankings obtained from the CLIP-seq, RNA-seq or by combining both. We collected 260 gene sets from Gene Ontology Biological Processes which contain “neuro” or “neural” in their annotation, spanning 2,227 genes. We benchmarked the rankings by computing an AUPR on two different sets: (a) using the union of all 260 terms resulting in a total of 2,227 genes (Supplemental Figure S10A) and computing a single AUPR for these genes, (b) using genes in one of the 69 process terms which contain at least 20 annotated genes and computing an Area Under the Precision Recall curve (AUPR) for each term (Supplemental Figure S10B). We tested the performance by plotting precision-recall curve by 8,076 prioritized gene and calculating AUPR.

Integer linear programming for defining regulatory paths from CLIP-seq hits to RNA-seq hits (Fig 5C, Supplemental Fig. S11C-D, Supplemental Table 5C).

To identify possible mechanistic paths that explain the relationships between the CLIP-seq hits and the RNA-seq hits we hypothesized that FMR1 impacts the transcriptome by making post-transcriptional changes to regulatory proteins such as transcription factors (TFs) and signaling proteins, which are in turn regulating the mRNA levels of genes identified as differentially expressed in RNA-seq experiments. To identify candidate TFs of RNA-seq DEGs, first we collected human neuron-related TFs from published transcriptional regulatory networks (TRNs) inferred from brain-specific RNA-seq data from the Allen brain atlas (Pearl et al. 2019) and from RNA-seq data from neuroepithelial stem cells during nervous system development (Chasman et al. 2019). We tested the enrichment of targets of TFs (determined by these networks) in the cell type-specific DEGs using an FDR corrected hypergeometric test. At an FDR<0.05, we found 269 TFs enriched in dorsal NPC DEGs, 112 TFs enriched in ventral NPC, 1 TF enriched in dorsal neuron and 3 TFs enriched in ventral neuron.

To define a minimal set of intermediate nodes, we used an integer linear programming (ILP) based optimization approach. These approaches are useful for linking a set of source nodes to sink nodes using a minimal set of intermediate nodes from an input skeleton network. Our skeleton network comprised edges from the transcriptional regulatory networks described above and from the STRING database. Since STRING network is an undirected network, we considered every edges of STRING network as bidirectional, so that we can get the candidate network comprises 2,866 nodes and 910,241 edges.

Next, we defined a set of candidate regulatory paths starting from a CLIP-seq target that is a TF or signaling protein as the source and a TF that is enriched in the DEGs as sink. All nodes in the path had to be a signaling or TF protein and the paths included between 0, 1 or 2 intermediate nodes. To select a minimal set of paths, we formulated a mixed integer linear programming-based (ILP) approach similar to that proposed by Chasman et al. (Chasman et al. 2014). The components of this approach are (1) a set of candidate paths and node scores, (2) a set of constraints that describe valid linear regulatory paths, and (3) an objective function that optimizes for a globally minimal (parsimonious) network that consists of high-scoring nodes.
We used the network diffusion scores from integrated scores of both CLIP- and RNA-seq hit inputs as node scores, thereby focusing on genes likely relevant to FMR1.

Integer linear program constraints. The input to our ILP approach is represented as a graph of nodes $\mathcal{N}$ (gene products), edges $\mathcal{E}$ (interactions), and candidate paths $\mathcal{P}$. We use the notation $\mathcal{E}(p)$ to denote the edges in a path *p*, $\mathcal{N}(e)$ to denote the pair of nodes in an edge *e*, etc. We assign a binary variable to each network element (node, edge, and path) to represent whether the element is included in the subnetwork: y (nodes), x (edges), and σ (paths). Node scores from the diffusion analysis are denoted $s(n)$. The following constraints define a valid subnetwork:

Domain	Constraint	Explanation
For all $p \in \mathcal{P}$, $e \in \mathcal{E}(p)$	$\sigma_p \leq x_e$	To include a path, we must include all of its edges $\mathcal{E}(p)$
For all $e \in \mathcal{E}$	$x_e \leq \sum_{p \in \mathcal{P}(e)} \sigma_p$	An included edge must be in an included path
For all $e \in \mathcal{E}, n \in \mathcal{N}(e)$	$x_e \leq y_n$	An edge can only be included if both of its nodes are included
For all $n \in \mathcal{N}$	$y_n \leq \sum_{e \in \mathcal{E}(n)} x_e$	An included node must be part of an included edge
For all $n \in$ (CLIP-seq hits ∪ RNA-seq hits)	$y_n = 1$	All CLIP-seq and RNA-seq hits must be included

Integer linear program optimization. We optimize three objective functions in serial. After each step we add a new constraint. The third optimization is for recordkeeping and ensures that all possible paths are chosen.

Step	Objective	Explanation
Solve:	$\text{MIN\_NODES} = \min \sum_{n \in \mathcal{N}} y_n$	Find the minimum feasible number of nodes subject to the constraints.
Add constraint:	$\sum_{n \in \mathcal{N}} y_n = \text{MIN\_NODES}$	Set node count.

Solve:	$\hat{y} = \underset{y}{\operatorname{argmax}} \sum_{n \in \mathcal{N}} s(n) y_n$	Find nodes with maximum total node score.
Add constraint:	$y_n = \widehat{y_n}$	Fix node assignments.
Solve:	$\hat{\sigma} = \underset{\sigma}{\operatorname{argmax}} \sum_{p \in \mathcal{P}} \sigma_p$	Find all possible paths between the included nodes.

We modeled the ILP with GAMS rev.240 and solved it using IBM ILOG CPLEX 12.5. Once the ILP finished we obtained a set of prioritized paths. We further selected paths for visualization based on their overall diffusion score, defined by the mean of the score.

Supplemental References

References

Cao M, Pietras CM, Feng X, Doroschak KJ, Schaffner T, Park J, Zhang H, Cowen LJ, Hescott BJ. 2014. New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics* **30**: i219-227.

Chambers SM, Fasano CA, Papapetrou EP, Tomishima M, Sadelain M, Studer L. 2009. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nature biotechnology* **27**: 275-280.

Chasman D, Ho YH, Berry DB, Nemec CM, MacGilvray ME, Hose J, Merrill AE, Lee MV, Will JL, Coon JJ et al. 2014. Pathway connectivity and signaling coordination in the yeast stress-activated signaling network. *Mol Syst Biol* **10**: 759.

Chasman D, Iyer N, Fotuhi Siahpirani A, Estevez Silva M, Lippmann E, McIntosh B, Probasco MD, Jiang P, Stewart R, Thomson JA et al. 2019. Inferring Regulatory Programs Governing Region Specificity of Neuroepithelial Stem Cells during Early Hindbrain and Spinal Cord Development. *Cell Syst* **9**: 167-186 e112.

Fotuhi Siahpirani A, Ay F, Roy S. 2016. A multi-task graph-clustering approach for chromosome conformation capture data sets identifies conserved modules of chromosomal interactions. *Genome biology* **17**: 114.

Fouss F, Francoisse K, Yen L, Pirotte A, Saerens M. 2012. An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification. *Neural Networks* **31**: 53-72.

Kohler S, Bauer S, Horn D, Robinson PN. 2008. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* **82**: 949-958.

Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**: 909-915.

Krakau S, Richard H, Marsico A. 2017. PureCLIP: Capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data. *bioRxiv*.

Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M et al. 2015. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* **47**: 106-114.

Li M, Zhao H, Ananiev GE, Musser MT, Ness KH, Maglaque DL, Saha K, Bhattacharyya A, Zhao X. 2017. Establishment of Reporter Lines for Detecting Fragile X Mental Retardation (FMR1) Gene Reactivation in Human Neural Cells. *STEM CELLS* **35**: 158-169.

Li Y, Stockton ME, Bhuiyan I, Eisinger BE, Gao Y, Miller JL, Bhattacharyya A, Zhao X. 2016. MDM2 inhibition rescues neurogenic and cognitive deficits in a mouse model of fragile X syndrome. *Science Translational Medicine* **8**: 336ra361-336ra361.

Maroof AM, Keros S, Tyson JA, Ying SW, Ganat YM, Merkle FT, Liu B, Goulburn A, Stanley EG, Elefanty AG et al. 2013. Directed differentiation and functional maturation of cortical interneurons from human embryonic stem cells. *Cell stem cell* **12**: 559-572.

Pearl JR, Colantuoni C, Bergey DE, Funk CC, Shannon P, Basu B, Casella AM, Oshone RT, Hood L, Price ND et al. 2019. Genome-Scale Transcriptional Regulatory Network Models of Psychiatric and Neurodegenerative Disorders. *Cell Syst* **8**: 122-135 e127.

Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI. 2017. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research* **45**: D833-D839.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498-2504.

Smola AJ, Kondor R. 2003. Kernels and regularization on graphs. In *Learning theory and kernel machines*, pp. 144-158. Springer.

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P et al. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research* **39**: D561-568.

Trevor H, Robert T, JH F. 2009. The elements of statistical learning: data mining, inference, and prediction. New York, NY: Springer.

Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K et al. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Meth* **13**: 508-514.

Vandin F, Upfal E, Raphael BJ. 2011. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology* **18**: 507-522.

von Luxburg U. 2007. A tutorial on spectral clustering. *Statistics and Computing* **17**: 395-416.

Xu B, Zhang Y, Zhan S, Wang X, Zhang H, Meng X, Ge W. 2018. Proteomic Profiling of Brain and Testis Reveals the Diverse Changes in Ribosomal Proteins in fmr1 Knockout Mice. *Neuroscience* **371**: 469-483.

Zarnegar BJ, Flynn RA, Shen Y, Do BT, Chang HY, Khavari PA. 2016. irCLIP platform for efficient characterization of protein-RNA interactions. *Nat Meth* **13**: 489-492.

Zhang Y, Rohe K. 2018. Understanding Regularized Spectral Clustering via Graph Conductance. *arXiv* arXiv:1806.01468.