Supporting Information

Sugathan et al. 10.1073/pnas.1405266111

SI Materials and Methods

Cell Culture. GM8330-8 was cultured with neural expansion medium [70% (vol/vol) DMEM (Invitrogen), 30% (vol/vol) HAMS-F12 (Mediatech) supplemented with 2% (vol/vol) B27 supplement (Invitrogen) and 1% penicillin-streptomycin-glutamine] on culture plates coated with poly-L-ornithine (20 μg/mL; Sigma) and laminin (5 μg/mL; Sigma). NPC medium was supplemented with basic fibroblast growth factor (bFGF) (20 ng/mL; R&D Systems), EGF (20 ng/mL, Sigma), and heparin (5 μg/mL; Sigma).

Generation of Stable *CHD8* Knockdowns. For efficient transduction, GM8330-8 was cultured in six-well plates to about 80–90% confluency. Then 20 μL of the designated virus stock (10^7–10^8) was added dropwise to each well, and puromycin selection was started 48 h post transduction.

ChIP. Approximately 60 million cells were fixed with 1% formaldehyde, washed with ice-cold PBS, harvested, pelleted, and resuspended in SDS lysis buffer [50 mM Tris·HCl (pH 8.1), 1% SDS, 10 mM EDTA]. Samples were sonicated with a Bioruptor sonicator (Diagenode), and sheared chromatin was diluted 10-fold in ChIP dilution buffer [16.7 mM Tris·HCl (pH 8.1), 167 mM NaCl, 0.01% SDS, 1.1% (vol/vol) Triton X-100, 1.2 mM EDTA]. After a control aliquot was removed (INPUT), the sample was incubated at 4 °C overnight with anti-CHD8 antibodies (Novus Biological NB100-60417, NB100-60418, and Bethyl A301-224A) and anti-CHD7 antibody (Bethyl A301-223A). Complexes precipitated with Dynabeads Protein A beads were washed sequentially with low-salt [20 mM Tris·HCl (pH 8.1), 150 mM NaCl, 0.1% SDS, 1% Triton X-100, 2 mM EDTA], high-salt [20 mM Tris·HCl (pH 8.1), 500 mM NaCl, 0.1%SDS, 1% Triton X-100, 2 mM EDTA], LiCl [10 mM Tris·HCl (pH 8.1), 0.25 M LiCl, 1% Nonidet P-40, 1% sodium deoxycholate,1 mM EDTA], and Tris·EDTA (TE) [10 mM Tris·HCl (pH 8.0), 1 mM EDTA] wash buffers. Immunoprecipitated chromatin was eluted in elution buffer (TE plus 1% SDS, 150 mM NaCl, 5 mM DTT), de-crosslinked at 65 °C for 8 h (or overnight), and treated with proteinase K (Roche).

Protein Extraction and Western Blotting. Protein extracts were prepared from PBS-washed cell pellets. RIPA buffer [50 mM Tris (pH 7.5– 8.0), 150 mM NaCl, 1% Triton X-100, 0.5% sodium deoxycholate, 0.1% SDS, 5 mM EDTA, 10 mM NaF, 1× protease inhibitor (Roche mixture 25×), Halt phosphatase inhibitor 1× (Thermo Scientific)] was used to lyse the cells. Fifty micrograms of protein extract was subjected to 3–8% (wt/vol) Tris-Acetate SDS/PAGE. After electrophoresis, the proteins were transferred to PVDF membrane (Immobilon-P; Millipore). The following primary antibodies and concentrations were used: Novus Biological NB100-60417 (1:500), NB100-60418 (1:500), Bethyl A301-224A (1:500), and Millipore HSP90 (1:2,000). After extensive washes, the blots were incubated with HRP-conjugated secondary antibodies. The membranes then were processed using an ECL chemiluminescence substrate kit (Perkin-Elmer) and were exposed to autoradiography.

RNA Sequencing. All shRNA infections, harvesting, library preparation, and sequencing were performed in two batches. RNA-seq libraries were prepared using a customized version of the originally published, strand-specific dUTP method (1, 2). In brief, library production was performed in a 96-well format using the total RNA isolated using a TRIzol/chloroform extraction and was quality monitored on an Agilent Tape Station. We then selected mRNA on magnetic oligo(dT) beads, treated with DNase and proceeded to second strand synthesis using random hexamer in presence of actinomycin D to reduce spurious reverse transcription. A second strand was generated with dUTP replacing dTTP; then the second strand was removed to retain the strandedness of the original transcripts. Standard Illumina paired-end library preparation was performed with barcoded adaptors. A uracil-specific exonuclease was used to remove the dUTP-marked strands, followed by minimal PCR amplification and quantification. Each library also included 1 μL of a 1:10 dilution of ERCC RNA Control Spike-Ins (Ambion) that were added from one of two mixes, each containing the same 92 synthetic RNA standards of known concentration and sequence. These synthetic RNAs cover a 10^6 range of concentration, as well as varying in length and GC content to allow validation of dose response and the fidelity of the procedure in downstream analyses (3).

Raw sequence data were quality checked using fastQC (4) version 0.10.0 and reads containing Ns or bases with map quality less than 20 were filtered (Dataset S2*B*). After alignment to the human genome, the bedTools version 2.17.0 (5) command multibamcov was used to calculate read coverage for each library at all Ensembl genes (GrCH37, build 71). Outlier samples [one control LacZ library with low yield and low data quality (LacZb; Fig. S1*B* and Dataset S2*B*) and both technical replicates for sh3, which had low correlation between the two technical replicates (Fig. S1*B*)] were removed. In multidimensional scaling (MDS) plots (Fig. S8*A*) without batch correction, samples are separated by batch in the first and second dimensions. After batch correction using the removeBatchEffect function in edgeR (version 3.4.2) (6), clustering of samples showed a clear separation of control samples and knockdown samples (Fig. S8 *B* and *C*). Genes with fewer than four mapped reads (chosen based on analysis of ERCC spike-ins; Fig. S2 *B* and *C*) in any of the samples were filtered out, leaving 15,903 genes. For differential expression analysis comparing all knockdown samples with all control samples, samples from the hairpin sh1 were also excluded because of the low level of knockdown of *CHD8* (Dataset S2*B*). Sample clustering after batch correction showed that this knockdown clustered with the control samples (Fig. S8*B*). Of the 15,903 genes, 15,896 genes successfully converged in generalized linear models (GLM) fitting, and these 15,896 genes, after excluding *CHD8* itself, were used as the background for DAVID functional annotation and disease gene enrichments shown in Figs. 2 and 4 and Datasets S3, S6, S7, and S8. For enrichment analysis using ToppGene, the entire human genome background was used. The fragments per kilobase per million reads (FPKMs) shown in Fig. 1*B* and Figs. S1*A* and S5 *C* and *D* were generated using Cufflinks (version 2.0.2) (7).

ChIP Sequencing. Library preparation was carried out using the NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs catalog no. E73705), according to the manufacturer's instructions. Libraries were sequenced using Illumina paired-end 50-cycle sequencing on a HiSEq 2000 (Dataset S10*A*). Peak detection was carried out after filtering out multiply mapped reads, reads not in proper pairs, and duplicate read pairs. We obtained 64–86 million reads, and subsampling using the –diag option in MACS 1.4.2 showed that, using 30% of reads, ~80% of total peaks were detected for all three libraries (Fig. S5*A*). The numbers of detected peaks were 18,688, 15,694, and 15,535, and pairwise overlaps between the three sets of peaks ranged from 60–75% (Dataset S10*B*). In contrast, overlaps in peaks between

CHD8 antibodies and the CHD7 antibody ranged from 15–39%, and the 7,324 replicated CHD8 peaks and 7,917 CHD7 peaks had an overlap of only 12%. To ensure that we were working with the highest-confidence peak list, we used only peaks that were detected by all three antibodies. To do so, the peak lists for each CHD8 antibody at $q < 0.05$ were concatenated and then merged into a single list of 27,056 merged peaks. The individual peak lists then were compared with the merged peak list to determine the overlaps, and 7,324 peaks were identified that intersected a peak from all of the three individual peak lists (Fig. S5*B*).

Recognizing the limitations of comparing FDRs across libraries (8), we also applied an irreproducible discovery rate (IDR) (9) to assess the reproducibility of our three CHD8 ChIP-seq datasets, following the pipeline at sites.google.com/site/anshulkundaje/projects/idr last updated on July 7, 2013. IDR determines how many peaks (ranked by *P* value) are reproducible between replicates (in our case, antibodies) and also are reproducible between pseudoreplicates (generated by randomly separating the libraries into two samples). Self-consistency thresholds for the three CHD8 antibodies were within a factor of 2, and the original replicate threshold and pooled pseudoreplicate threshold also were within a factor of 2, which meet the recommended cutoffs for reproducibility. This finding justified our decision to combine peak lists from the three different antibodies. Furthermore, by following the IDR pipeline, we generated conservative and optimal lists of 10,333 and 14,051 peaks, respectively, and observed the same patterns of ASD and cancer-related gene enrichments as obtained using the set of 7,324 peaks meeting $q < 0.05$ for all three antibodies: strongest SFARI/AutismKB ASD enrichment among unbound, down-regulated genes ($P < 3 \times 10^{-8}$), Willsey ASD enrichment only among all CHD8-bound or nonregulated CHD8-bound genes ($P < 2 \times 10^{-3}$), and strongest cancer enrichment among bound, nonregulated genes ($P < 5 \times 10^{-9}$).

For the chromatin states at CHD8-binding sites shown in Fig. 3*D*, we obtained genome segmentations by 15 chromatin states accessed from www.broadinstitute.org/~anshul/projects/roadmap/segmentations/models/coreMarks/parallel/set2/final/ (accessed on May 28, 2014), based on five histone modifications (H3K4me3, H3K4me1, H3K36me3, H3K9me3, and H3K27me3), for an ES cell-derived neural progenitor line from the NIH Roadmap Epigenomics consortium (http://nihroadmap.nih.gov/epigenomics/) (10). For each CHD8-binding site that overlapped genome segments assigned to multiple chromatin states, the one state that covered the largest fraction of the peak region was considered the chromatin state at that peak. For the whole genome, coverage for a particular chromatin state was calculated as the number of base pairs assigned that state.

De novo motif discovery using Homer (11) was carried out separately for each of the three CHD8 antibody peak lists to ensure that discovered motifs were replicable. All de novo motifs from each peak list were compared with de novo motifs from the other two peak lists using STAMP (12) to identify motifs that were discovered in more than one peak list. A de novo motif was considered to be discovered in two peak lists if the two motifs were reciprocally each other's best matches and at least one of those comparisons met E-value < 1e-5. In this way, nine de novo motifs were identified as being discovered in more than one peak list and are listed in Dataset S5. The known motif library provided as part of the Homer package was used to identify the predicted binding factor represented by each de novo motif. The best-matching known motif for each de novo motif also is listed in Dataset S5.

We used binding and expression target analysis (BETA) (13), which assesses the regulatory potential of a transcription factor by ranking up- and down-regulated genes by the distances to all binding sites within 100 kb and comparing this distribution to that of expressed but nonregulated genes using a one-tailed Kolmogorov–Smirnov test. For peaks generated by each of the three CHD8 antibodies, regulatory potential is significantly greater ($P < 3 \times 10^{-4}$) for up-regulated genes compared with background but not for down-regulated genes (Fig. S6 *A*–*C*). Regulatory potential for CHD7 is statistically significant for genes down-regulated by CHD8 ($P = 1 \times 10^{-4}$) but not for up-regulated genes (Fig. S6*D*).

Disease-Associated Gene Sets. Comprehensive ASD gene sets were obtained from SFARI Gene 2.0 (574 genes) (https://gene.sfari.org/autdb/Welcome.do) (14) and AutismKB (171 genes) (http://autismkb.cbi.pku.edu.cn) (15) databases on Dec 26, 2013. The union of these two datasets constitutes a list of 628 unique ASD genes that were used in this study. SFARI scores genes from 1 (high confidence) to 6 (not supported), assigning the best scores to genetic evidence in humans, with a separate score, "S," for syndromic genes. The full set of 574 SFARI ASD genes included many proposed genes with evidence levels of 5 (hypothesized but untested; 68 genes) or 6 (not supported; 23 genes) as well as 248 genes not assigned an evidence score. AutismKB assigns genes a weighted score based on type of evidence and number of studies, with highest weight given to the GWAS and lowest weight to expression studies. It uses a total score of 9, the minimal score of a benchmark dataset of high-confidence genes from highly accessed review articles, as the threshold score; for our ASD gene set we included only AutismKB genes that had a score of at least 16 along with syndromic genes, their "recommended Autism gene list." Ten of the low-scoring genes in the SFARI list and 71 unscored genes were also in the AutismKB list. Because many low-scoring and unscored SFARI genes were not in the AutismKB list, a reduced ASD gene set (235 genes) also was obtained from SFARI Gene 2.0 following the criteria described in ref. 16, using genes scored as syndromic (S) and evidence levels 1–4 (high confidence to minimal evidence). We also confirmed that the same patterns of enrichment were observed when the 23 level-6 genes were excluded from the SFARI dataset: strong enrichment among unbound, down-regulated genes ($P = 1.41 \times 10^{-9}$) and nominal enrichment among CHD8-bound, nonregulated genes ($P = 0.029$).

The comprehensive cancer gene list (5,873 genes) was obtained from http://cbio.mskcc.org/tcga-generanker by combining 39 gene lists and ranking them based on genes' representation in those lists. We used genes with rank ≥1. To investigate hypothesis-driven subsets, we tested a reduced cancer gene list (224 genes) from the Lawrence et al. study of somatic mutation sequencing (17), and the COSMIC cancer gene census (18) data, which includes a manually curated list of 513 genes with mutations that were causally implicated in cancer (http://cancer.sanger.ac.uk/cancergenome/projects/census/). A list of 669 genes associated with the "abnormality of skull size" (HP:0000240) phenotype was obtained from the ToppGene suite (19). FMRP targets (842 genes), intellectual disability genes (401 genes), and schizophrenia-associated genes (186 genes) were obtained from previous publications (16, 20, 21). For attention deficit hyperactivity disorder (http://adhd.psych.ac.cn) (22), bipolar disorder (http://bdgene.psych.ac.cn) (23), and major depressive disorder (http://mdd.psych.ac.cn) (24), we selected subsets (38, 96, and 94 genes, respectively) that met statistical significance in at least two studies, based on the criteria used by the database authors. GWAS gene sets (Dataset S9) were obtained from the NHGRI GWAS catalog (25). Disease gene-set enrichment was assessed using a one-tailed Fisher's exact test, with the 15,896 genes that converged in GLM-fitting by DESeq, excluding *CHD8*, used as the background. To calculate permutation *P* values for disease gene enrichments, we randomly sampled the same number of genes as in a given disease gene set 10,000 times and further assessed enrichments of these random sets using one-tailed Fisher's exact test. The fraction of enrichment *P* values that are equal to or smaller than the original *P* value is reported as a permutation P value for a given disease gene enrichment in a gene list with a given condition.

We compared the overlap between CHD8- and CHD7-binding sites and polymerase II-binding regions, obtained from ENCODE (26) for ES cell-derived neurons (Gene Expression Omnibus sample accession no. GSM1010803) (Dataset S10*B*). Because CHD8 binding is enriched at polymerase II-binding regions (Dataset S10*B*) and CHD8 binding frequency increases with gene expression (Fig. S5*C*), as are consistent with the previously reported association between CHD8 and RNA polymerase II (27), we asked whether the strong enrichment of cancer- and skull size-related gene sets among the nonregulated, CHD8-bound genes is simply a consequence of CHD8 binding at highly expressed genes. Unlike the SFARI/AutismKB ASD genes, which have a range of expression similar to that of non-ASD genes in our dataset (Fig. S5*D*), cancer- and skull size-related genes had higher expression levels in controls than in genes not in those gene sets: For gene-expression bins of $\log_2$ FPKM ~3 and higher, the fraction of genes in the gene set exceeds the fraction not in the gene set (Fig. S5*E*). Therefore we tested enrichment for these gene sets among a set of 2,849 highly expressed [$\log_2$(FPKM)>3], nonregulated genes that are not bound by CHD8. Only the large, most inclusive TCGA gene set was enriched in this group ($P = 1.93 \times 10^{-9}$); the P values for the other cancer gene sets, ASD gene sets (including the Willsey et al. set), and skull-size gene set ranged from 0.08–0.83.

Coexpressed Gene Modules and Integration with Differential Expression and CHD8 Binding. Unlike differential expression, because coexpression analysis did not involve grouping of knockdown samples and control samples separately, the weakest hairpin (sh1) was included. Gene expressions in the form of log cpm for all 15,903 thresholded genes were TMM (trimmed mean of M values) normalized using the edgeR package (6), and the removeBatchEffect function was used to control for batch effects. Adjacency and topological overlap matrices for gene similarity were based on signed correlation, because our pathway and the ASD gene-set enrichments described above revealed that directionality of regulation by CHD8 is an important distinguishing feature between classes of genes. Modules with correlation >0.8 were merged, and at least 200 genes were required per module when merging. Genes that did not belong to a module were assigned the color gray. After merging, each gene was reassigned to the module it matched best. Within each module, hub genes were defined as genes in the top 10% by intramodular connectivity. To assess whether the modules are coexpressed to a greater degree than expected by chance, for each module we randomly sampled 10,000 gene sets of the same size and compared the sum of correlations, as described in ref. 16. All except the gray module were significant ($P < 1 \times 10^{-4}$). From the 21 modules, we selected four modules that had very high correlation between the module eigengene—the representative expression profile of the genes in the module—and *CHD8* expression: one module of up-regulated genes and three modules of down-regulated genes (Fig. S3*C*). Genes within each of the four CHD8-correlated modules that had a gene significance P value <0.05 were considered CHD8-coexpressed genes and were tested for functional enrichments using DAVID (Fig. S3*C*). The 2,028 CHD8-coexpressed genes in the four CHD8-correlated modules shown in Fig. S3*C* included 33% (586) of the 1,756 genes identified as differentially expressed using DESeq. Thus, these complementary approaches do not mirror each other in detecting potential expression network effects of *CHD8* suppression, because the coexpression network analysis implicates an additional 1,024 genes as showing expression closely correlated with *CHD8* but insufficiently strong to meet significance thresholds for differential expression. However, as is consistent with pathways enriched among down-regulated genes shown in Fig. 2, modules with genes whose suppression decreased in correlation with *CHD8* showed enrichment for terms including “cell adhesion,” “WNT signaling,” and “cell projection” (Fig. S3*C*). Among the 699 genes involved in the top 0.5% of coexpression networks, the majority (83%) are genes up-regulated by CHD8 knockdown. Similarly, 72% of the DE genes involved in PPI interactions (Fig. S4*A*) were up-regulated genes, suggesting that genes repressed by CHD8 are more likely to function together in a network.

Whole-Mount Immunostaining on Zebrafish Embryos. Whole-mount immunostaining with either HuC/D (postmitotic neurons) or phospho-histone H3 (an M-phase marker) was performed to investigate neuronal development and head-size regulation at a cellular level. Embryos were fixed in 4% (vol/vol) paraformaldehyde overnight and stored in 100% methanol at −20 °C. After rehydration in PBS, paraformaldehyde-fixed embryos were washed in immunofluorescence (IF) buffer (0.1% Tween-20 and 1% BSA in 1× PBS) for 10 min at room temperature. The embryos were incubated in the blocking buffer [10% (vol/vol) FBS and 1% BSA in 1× PBS] for 1 h at room temperature. After two washes in IF buffer for 10 min each, embryos were incubated in the first antibody solution, 1:750 anti-histone H3 (ser10)-R (sc-8656-R; Santa Cruz) or 1:1,000 anti-HuC/D (A21271; Invitrogen), in blocking solution overnight at 4 °C. After two washes in IF buffer for 10 min each, embryos were incubated in the secondary antibody solution, 1:1,000 Alexa Fluor donkey anti-rabbit IgG and Alexa Fluor goat anti-mouse IgG (A21207 and A11001; Invitrogen), in blocking solution for 1 h at room temperature. Staining was quantified by counting positive cells in defined regions of the head and with ImageJ software. All experiments were repeated three times, and a Student t test (for head-size measurements or p-histone H3 staining) or a χ^2 test (for HuC/D staining) was used to determine the significance of the morphant phenotype.

RNA-Seq of *chd8* Transcript in Zebrafish. To confirm the suppression of *chd8* in the MO samples, we first performed targeted quantitative PCR at multiple sites and identified replicable increased expression of *chd8* transcript. Therefore we sought to characterize the transcript architecture of *chd8* MO and wild-type zebrafish fully using RNA-seq. Analysis of split reads using MISO (28) showed inclusion of a portion of intronic sequence between exons 7 and 8 in the MO-treated samples, corresponding to the MO-binding site (Fig. S7 *F* and *G*). This misspliced isoform produces a frameshift and a premature stop codon in this aberrant transcript. A single splice junction differentiated this abnormal isoform from the endogenous transcript. Comparison of normalized expression using all split reads crossing this junction revealed that the increased expression was limited to the aberrantly spliced transcript, which introduced a premature stop, and that the normally spliced product was actually reduced in the MO-treated samples as compared with WT (see Fig. S7*H* for complete splicing architecture). The average change in exon junction expression between MO-treated and WT samples across the gene was 4.16:1; however, the fold change in the properly spliced exon 7–8 junction was ~0.67:1, suggesting a decrease in the expression of normal *chd8* in the zebrafish treated with MO, consistent with the expected result. All results predicted by the RNA-seq data were confirmed by PCR and Sanger sequencing.

1. Levin JZ, et al. (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* 7(9):709–715.

2. Parkhomchuk D, et al. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* 37(18):e123.

3. Jiang L, et al. (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Res* 21(9):1543–1551.
4. Andrews S (2010) Fastqc. A quality control tool for high throughput sequence data. 2010. Available at *www.bioinformatics.babraham.ac.uk/projects/fastqc/*. Accessed September 29, 2014.
5. Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.
6. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139–140.
7. Trapnell C, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5):511–515.
8. Landt SG, et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 22(9):1813–1831.
9. Li Q, Brown BB, Huang H, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* 5(3):1752–1779.
10. Bernstein BE, et al. (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 28(10):1045–1048.
11. Heinz S, et al. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38(4):576–589.
12. Mahony S, Benos PV (2007) STAMP: A web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* 35(Web Server issue):W253–W258.
13. Wang S, et al. (2013) Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat Protoc* 8(12):2502–2515.
14. Abrahams BS, et al. (2013) SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism* 4(1):36.
15. Xu LM, et al. (2012) AutismKB: An evidence-based knowledgebase of autism genetics. *Nucleic Acids Res* 40(Database issue):D1016–D1022.
16. Parikshak NN, et al. (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 155(5):1008–1021.
17. Lawrence MS, et al. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505(7484):495–501.
18. Futreal PA, et al. (2004) A census of human cancer genes. *Nat Rev Cancer* 4(3):177–183.
19. Chen J, Bardes EE, Aronow BJ, Jegga AG (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* 37(Web Server issue):W305–W311.
20. Darnell JC, et al. (2011) FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* 146(2):247–261.
21. Ayalew M, et al. (2012) Convergent functional genomics of schizophrenia: From comprehensive understanding to genetic risk prediction. *Mol Psychiatry* 17(9): 887–905.
22. Zhang L, et al. (2012) ADHDgene: A genetic database for attention deficit hyperactivity disorder. *Nucleic Acids Res* 40(Database issue):D1003–D1009.
23. Chang SH, et al. (2013) BDgene: A genetic database for bipolar disorder and its overlap with schizophrenia and major depressive disorder. *Biol Psychiatry* 74(10):727–733.
24. Guo L, et al. (2012) MK4MDD: A multi-level knowledge base and analysis platform for major depressive disorder. *PLoS ONE* 7(10):e46335.
25. Hindorff LA, et al. A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies. Accessed January 2014.
26. Consortium EP, et al.; ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.
27. Rodríguez-Paredes M, Ceballos-Chávez M, Esteller M, García-Domínguez M, Reyes JC (2009) The chromatin remodeling factor CHD8 interacts with elongating RNA polymerase II and controls expression of the cyclin E2 gene. *Nucleic Acids Res* 37(8): 2449–2460.
28. Katz Y, Wang ET, Airoldi EM, Burge CB (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* 7(12):1009–1015.

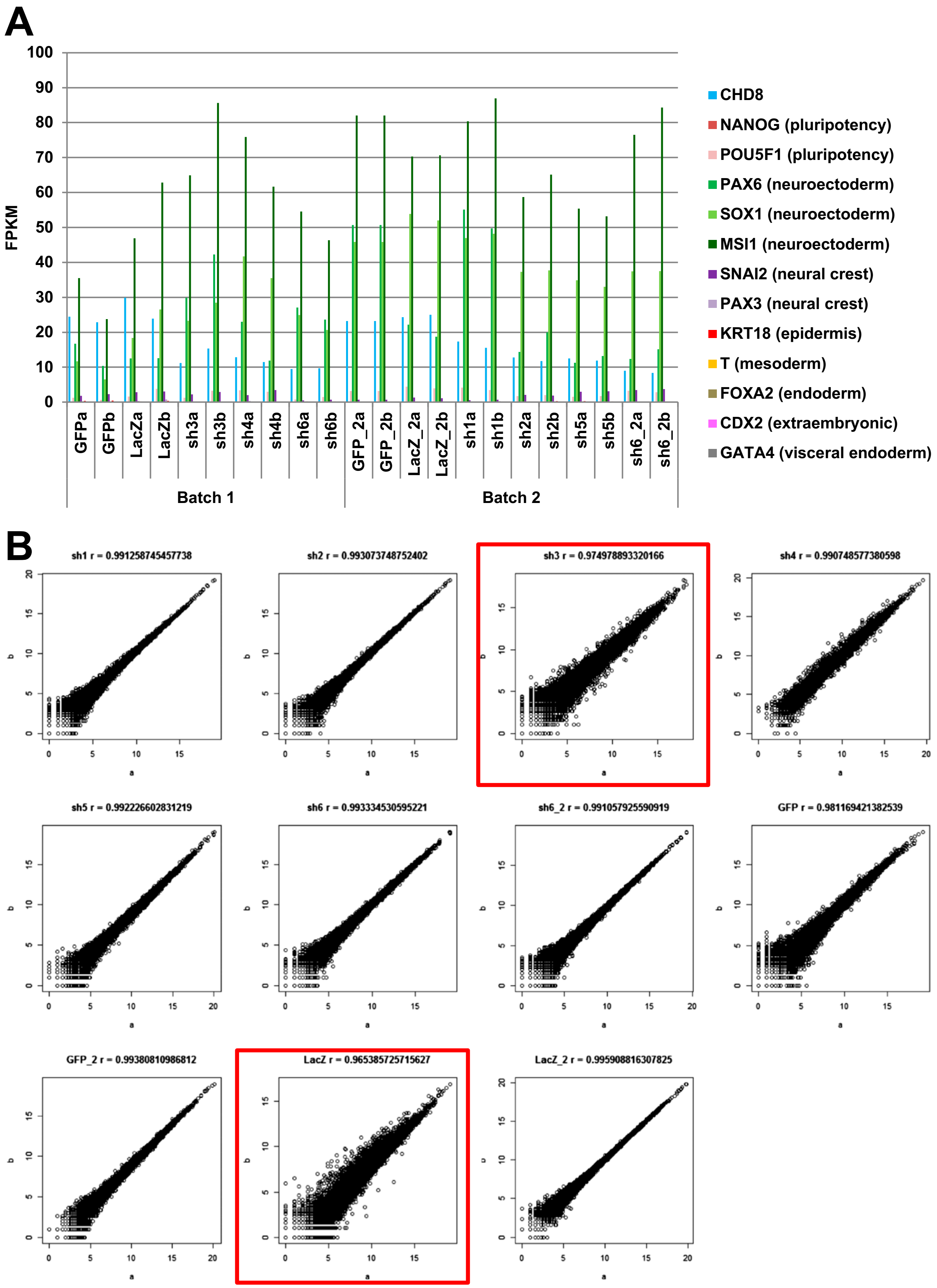


Fig. S1. Data quality of RNA-seq libraries. (*A*) Gene expression in knockdown and control NPCs for marker genes for stages of differentiation. Shown are FPKMs for 11 marker genes, along with the *CHD8* gene (blue) for comparison. All three marker genes for neuroectoderm (*MSI1, SOX1,* and *PAX6*), shown in shades of green, are highly expressed in all samples of both batches. Of the other genes, only two, *POU5F1* (pluripotency) and *SNAI2* (neural crest), met our

Legend continued on following page

thresholds for detection (at least four reads per sample), but both genes have substantially lower expression than the neuroectoderm markers and *CHD8*. (*B*) Scatter plots for gene expression in counts per million in each sample, comparing technical replicates a (*x* axis) and b (y axis). The Pearson correlation is shown at the top of each plot. Red squares indicate samples that were discarded because of low data quality and/or low correlation between technical replicates. Gene expression is shown in $\log_2$(counts per million).

A

		sh4a	sh4b	sh6a	sh6b	GFPa	GFPb	LacZa	sh6_2a	sh6_2b	sh2a	sh2b	sh5a	sh5b	GFP_2a	GFP_2b	LacZ_2a	LacZ_2b
		(batch1)	(batch1)	(batch1)	(batch1)	(batch1)	(batch1)	(batch1)	(batch2)	(batch2)	(batch2)	(batch2)	(batch2)	(batch2)	(batch2)	(batch2)	(batch2)	(batch2)
sh4a	(batch1)		0.991	0.957	0.955	0.937	0.935	0.947	0.972	0.972	0.958	0.963	0.958	0.959	0.972	0.973	0.975	0.975
sh4a	(batch1)	0.991		0.939	0.939	0.939	0.938	0.947	0.979	0.977	0.952	0.955	0.957	0.958	0.958	0.959	0.968	0.968
sh4b	(batch1)	0.957	0.939		0.995	0.933	0.940	0.931	0.955	0.954	0.972	0.976	0.973	0.971	0.964	0.966	0.961	0.960
sh6a	(batch1)	0.955	0.939	0.995		0.935	0.945	0.933	0.956	0.954	0.974	0.977	0.975	0.974	0.964	0.966	0.960	0.959
sh6b	(batch1)	0.937	0.939	0.933	0.935		0.985	0.972	0.953	0.955	0.973	0.969	0.967	0.973	0.967	0.963	0.962	0.963
GFPa	(batch1)	0.935	0.938	0.940	0.945	0.985		0.970	0.955	0.956	0.979	0.974	0.974	0.978	0.964	0.959	0.961	0.962
GFPb	(batch1)	0.947	0.947	0.931	0.933	0.972	0.970		0.956	0.959	0.963	0.960	0.955	0.964	0.966	0.961	0.976	0.977
LacZa	(batch1)	0.972	0.979	0.955	0.956	0.953	0.955	0.956		0.994	0.957	0.958	0.973	0.974	0.951	0.951	0.962	0.964
sh6_2a	(batch2)	0.972	0.977	0.954	0.954	0.955	0.956	0.959	0.994		0.957	0.957	0.970	0.973	0.953	0.951	0.964	0.966
sh6_2b	(batch2)	0.958	0.952	0.972	0.974	0.973	0.979	0.963	0.957	0.957		0.996	0.981	0.983	0.974	0.972	0.965	0.966
sh2a	(batch2)	0.963	0.955	0.976	0.977	0.969	0.974	0.960	0.958	0.957	0.996		0.982	0.982	0.975	0.974	0.966	0.967
sh2b	(batch2)	0.958	0.957	0.973	0.975	0.967	0.974	0.955	0.973	0.970	0.981	0.982		0.994	0.957	0.959	0.959	0.960
sh5a	(batch2)	0.959	0.958	0.971	0.974	0.973	0.978	0.964	0.974	0.973	0.983	0.982	0.994		0.963	0.961	0.964	0.965
sh5b	(batch2)	0.972	0.958	0.964	0.964	0.967	0.964	0.966	0.951	0.953	0.974	0.975	0.957	0.963		0.995	0.978	0.977
GFP_2a	(batch2)	0.973	0.959	0.966	0.966	0.963	0.959	0.961	0.951	0.951	0.972	0.974	0.959	0.961	0.995		0.976	0.975
GFP_2b	(batch2)	0.975	0.968	0.961	0.960	0.962	0.961	0.976	0.962	0.964	0.965	0.966	0.959	0.964	0.978	0.976		0.997
LacZ_2a	(batch2)	0.975	0.968	0.960	0.959	0.963	0.962	0.977	0.964	0.966	0.966	0.967	0.960	0.965	0.977	0.975	0.997	

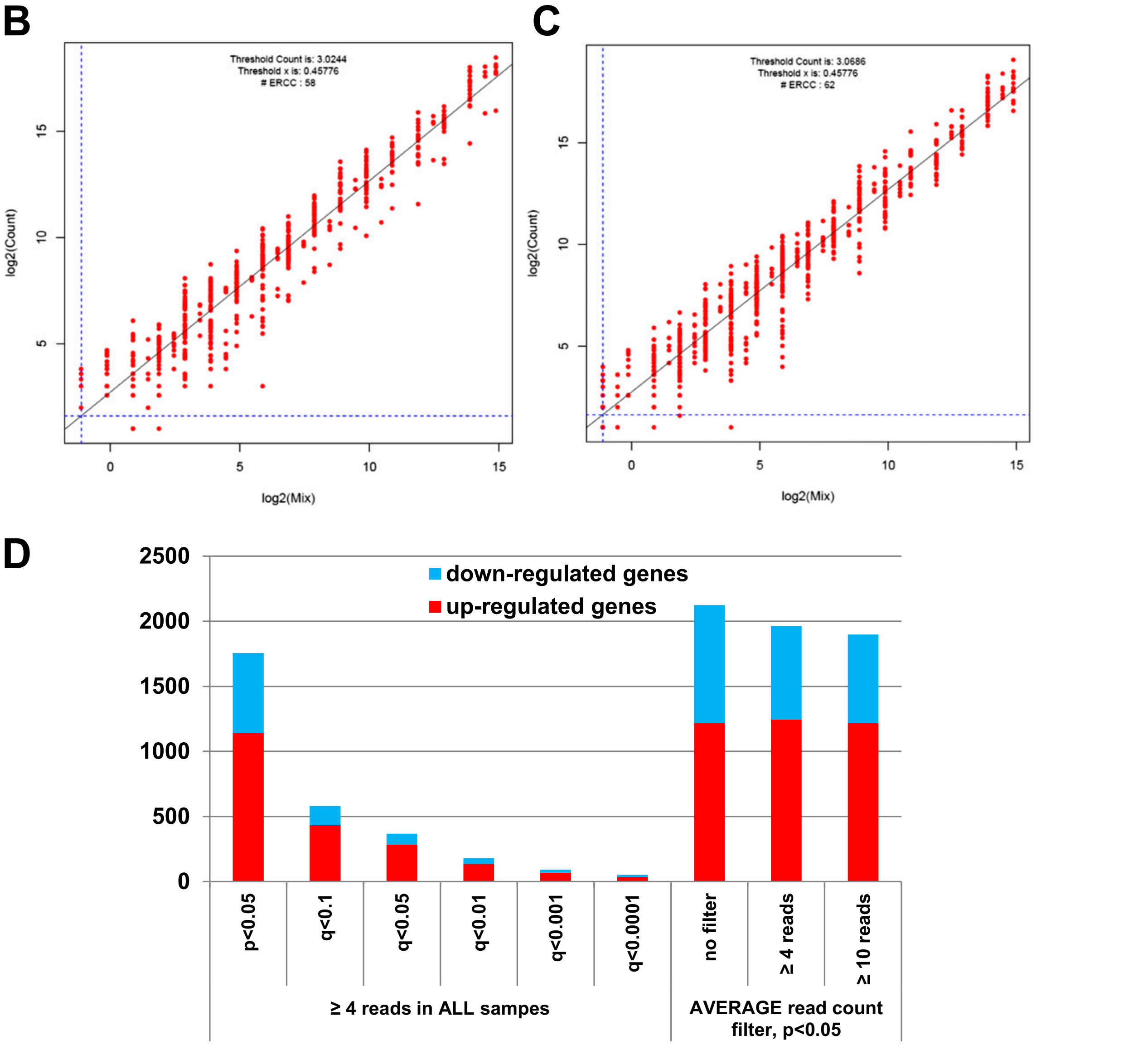
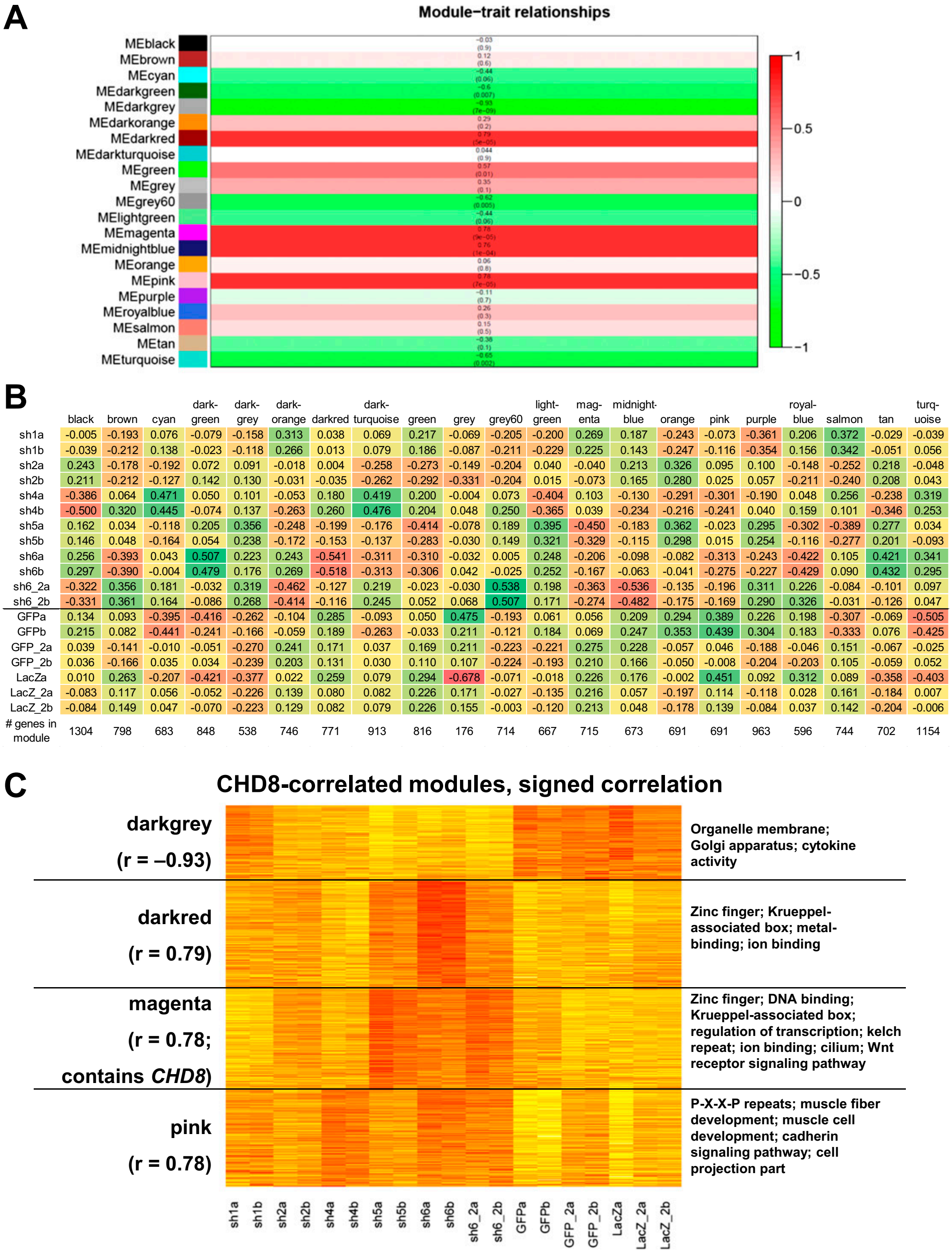


Fig. S2. RNA-seq sample correlations, spike-in analysis, and differentially expressed genes after removal of outlier samples. (*A*) Pearson correlations between all RNA-seq samples, after batch correction. Deeper green represents higher correlation. (*B* and *C*) ERCC spike-in analysis for RNA-seq samples. Analysis of spike-in RNA-seq data was carried out as described in ref. 1. Briefly, observed counts (*y* axis) are regressed on known concentrations (*x* axis) for the ERCC transcripts, pooled across all samples in batch 1 (*B*) and batch 2 (*C*). The detection was set at the average count at the lowest concentration detected among all samples, predicted from the regression equation. In both batches, the detection threshold was more three reads, which we rounded up to four reads. The number of ERCC

Legend continued on following page

transcripts that met the detection threshold also is indicated at the top of each plot. (*D*) Numbers of differentially expressed genes at different significance levels identified using DESeq. The *q* values were generated by Benjamini–Hochberg correction of DESeq *P* values.

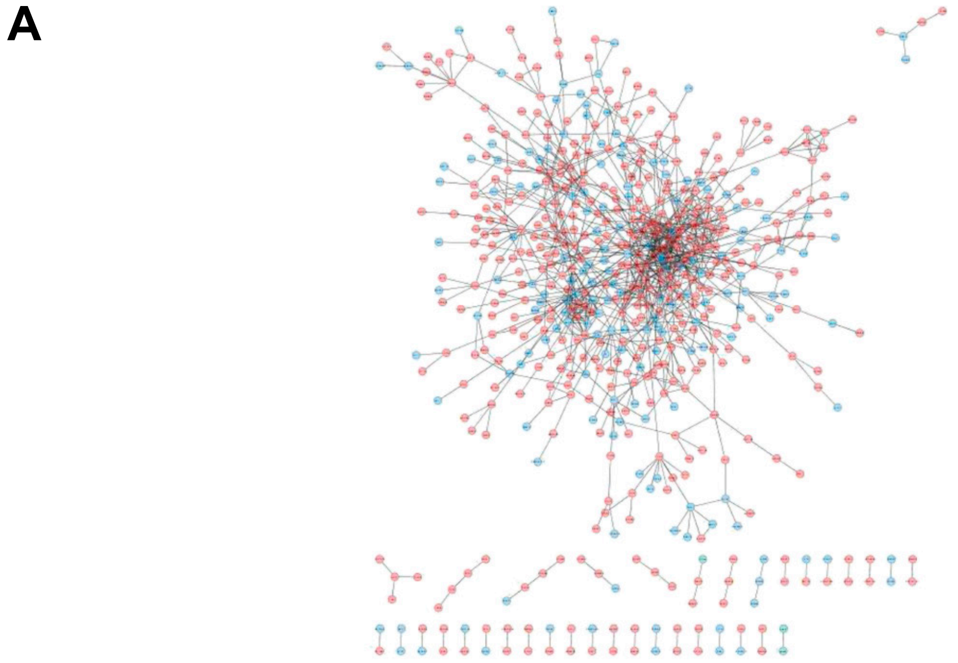
1. Blumenthal I, et al. (2014) Transcriptional consequences of 16p11.2 deletion and duplication in mouse cortex and multiplex autism families. *Am J Hum Genet* 94(6):870–883.

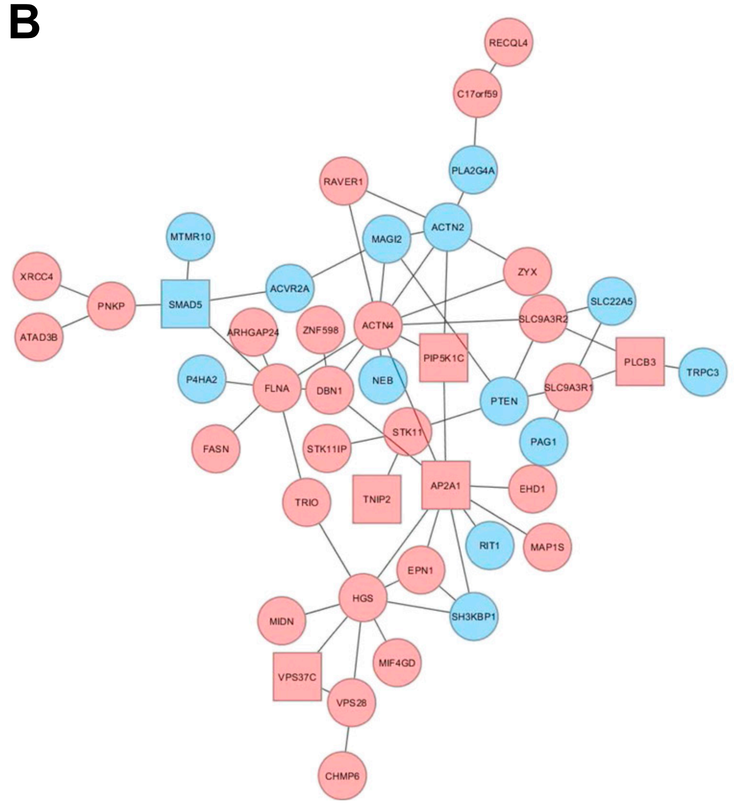


	black	brown	cyan	dark-green	dark-grey	dark-orange	darkred	dark-turquoise	green	grey	grey60	light-green	mag-enta	midnight-blue	orange	pink	purple	royal-blue	salmon	tan	turq-uoise
sh1a	-0.005	-0.193	0.076	-0.079	-0.158	0.313	0.038	0.069	0.217	-0.069	-0.205	-0.200	0.269	0.187	-0.243	-0.073	-0.361	0.206	0.372	-0.029	-0.039
sh1b	-0.039	-0.212	0.138	-0.023	-0.118	0.266	0.013	0.079	0.186	-0.087	-0.211	-0.229	0.225	0.143	-0.247	-0.116	-0.354	0.156	0.342	-0.051	0.056
sh2a	0.243	-0.178	-0.192	0.072	0.091	-0.018	0.004	-0.258	-0.273	-0.149	-0.204	0.040	-0.040	0.213	0.326	0.095	0.100	-0.148	-0.252	0.218	-0.048
sh2b	0.211	-0.212	-0.127	0.142	0.130	-0.031	-0.035	-0.262	-0.292	-0.331	-0.204	0.015	-0.073	0.165	0.280	0.025	0.057	-0.211	-0.240	0.208	0.043
sh4a	-0.386	0.064	0.471	0.050	0.101	-0.053	0.180	0.419	0.200	-0.004	0.073	-0.404	0.103	-0.130	-0.291	-0.301	-0.190	0.048	0.256	-0.238	0.319
sh4b	-0.500	0.320	0.445	-0.074	0.137	-0.263	0.260	0.476	0.204	0.048	0.250	-0.365	0.039	-0.234	-0.216	-0.241	0.040	0.159	0.101	-0.346	0.253
sh5a	0.162	0.034	-0.118	0.205	0.356	-0.248	-0.199	-0.176	-0.414	-0.078	0.189	0.395	-0.450	-0.183	0.362	-0.023	0.295	-0.302	-0.389	0.277	0.034
sh5b	0.146	0.048	-0.164	0.054	0.238	-0.172	-0.153	-0.137	-0.283	-0.030	0.149	0.321	-0.329	-0.115	0.298	0.015	0.254	-0.116	-0.277	0.201	-0.093
sh6a	0.256	-0.393	0.043	0.507	0.223	0.243	-0.541	-0.311	-0.310	-0.032	0.005	0.248	-0.206	-0.098	-0.082	-0.313	-0.243	-0.422	0.105	0.421	0.341
sh6b	0.297	-0.390	-0.004	0.479	0.176	0.269	-0.518	-0.313	-0.306	0.042	-0.025	0.252	-0.167	-0.063	-0.041	-0.275	-0.227	-0.429	0.090	0.432	0.295
sh6_2a	-0.322	0.356	0.181	-0.032	0.319	-0.462	-0.127	0.219	-0.023	-0.030	0.538	0.198	-0.363	-0.536	-0.135	-0.196	0.311	0.226	-0.084	-0.101	0.097
sh6_2b	-0.331	0.361	0.164	-0.086	0.268	-0.414	-0.116	0.245	0.052	0.068	0.507	0.171	-0.274	-0.482	-0.175	-0.169	0.290	0.326	-0.031	-0.126	0.047
GFPa	0.134	0.093	-0.395	-0.416	-0.262	-0.104	0.285	-0.093	0.050	0.475	-0.193	0.061	0.056	0.209	0.294	0.389	0.226	0.198	-0.307	-0.069	-0.505
GFPb	0.215	0.082	-0.441	-0.241	-0.166	-0.059	0.189	-0.263	-0.033	0.211	-0.121	0.184	0.069	0.247	0.353	0.439	0.304	0.183	-0.333	0.076	-0.425
GFP_2a	0.039	-0.141	-0.010	-0.051	-0.270	0.241	0.171	0.037	0.169	0.211	-0.223	-0.221	0.275	0.228	-0.057	0.046	-0.188	-0.046	0.151	-0.067	-0.025
GFP_2b	0.036	-0.166	0.035	0.034	-0.239	0.203	0.131	0.030	0.110	0.107	-0.224	-0.193	0.210	0.166	-0.050	-0.008	-0.204	-0.203	0.105	-0.059	0.052
LacZa	0.010	0.263	-0.207	-0.421	-0.377	0.022	0.259	0.079	0.294	-0.678	-0.071	-0.018	0.226	0.176	-0.002	0.451	0.092	0.312	0.089	-0.358	-0.403
LacZ_2a	-0.083	0.117	0.056	-0.052	-0.226	0.139	0.080	0.082	0.226	0.171	-0.027	-0.135	0.216	0.057	-0.197	0.114	-0.118	0.028	0.161	-0.184	0.007
LacZ_2b	-0.084	0.149	0.047	-0.070	-0.223	0.129	0.082	0.079	0.226	0.155	-0.003	-0.120	0.213	0.048	-0.178	0.139	-0.084	0.037	0.142	-0.204	-0.006
# genes in module	1304	798	683	848	538	746	771	913	816	176	714	667	715	673	691	691	963	596	744	702	1154

C
CHD8-correlated modules, signed correlation
darkgrey
(r = –0.93)
Organelle membrane; Golgi apparatus; cytokine activity
darkred
(r = 0.79)
Zinc finger; Krueppel-associated box; metal-binding; ion binding
magenta
(r = 0.78;
contains CHD8)
Zinc finger; DNA binding; Krueppel-associated box; regulation of transcription; kelch repeat; ion binding; cilium; Wnt receptor signaling pathway
pink
(r = 0.78)
P-X-X-P repeats; muscle fiber development; muscle cell development; cadherin signaling pathway; cell projection part
sh1a
sh1b
sh2a
sh2b
sh4a
sh4b
sh5a
sh5b
sh6a
sh6b
sh6_2a
sh6_2b
GFPa
GFPb
GFP_2a
GFP_2b
LacZa
LacZ_2a
LacZ_2b

Fig. S3. RNA-seq coexpression modules. (*A* and *B*) Twenty-one modules of coexpressed genes using signed correlation. (*A*) Module–trait relationships. In this analysis, *CHD8* expression in each of the samples was used as the trait. The number shown for each module corresponds to the correlation between the module eigengene and *CHD8* expression; the *P* value for correlation is given in parentheses. (*B*) Module eigengenes for all 21 modules. The eigengene of a module is the representative expression profile for all genes in the module. Green represents high expression, and red indicates low expression. (*C*) Heatmap of expression for CHD8-correlated modules. Numbers in parentheses indicate the correlation between the module eigengene and gene expression of the *CHD8* gene. Enriched pathways/functional annotations using DAVID (FDR < 5%) for the genes in each module that are correlated with *CHD8* with *P* < 0.05 are shown on the right.





Term	Count	PValue	FDR (%)
hsa04144:Endocytosis	9	4.55E-06	0.00
GO:0019898~extrinsic to membrane	10	1.90E-05	0.02
cytoplasm	22	9.66E-05	0.11
GO:0015629~actin cytoskeleton	7	1.35E-04	0.16
GO:0005886~plasma membrane	19	2.16E-04	0.25
REACT_9417:Signaling by EGFR	5	5.23E-04	0.36
phosphoprotein	33	7.85E-04	0.90
GO:0019904~protein domain specific binding	7	9.64E-04	1.22
GO:0044459~plasma membrane part	14	1.05E-03	1.23
mutagenesis site	15	1.06E-03	1.47
GO:0005829~cytosol	13	1.30E-03	1.52
IPR002017:Spectrin repeat	3	1.77E-03	2.10
domain:Actin-binding	3	1.61E-03	2.24
repeat:Spectrin 4	3	1.61E-03	2.24
GO:0016197~endosome transport	4	1.69E-03	2.50
repeat:Spectrin 3	3	1.81E-03	2.51
IPR001589:Actinin-type, actin-binding, conserved site	3	2.21E-03	2.62
GO:0042641~actomyosin	3	2.49E-03	2.89
repeat:Spectrin 1	3	2.24E-03	3.09
repeat:Spectrin 2	3	2.24E-03	3.09
hsa04510:Focal adhesion	6	3.41E-03	3.17
IPR018159:Spectrin/alpha-actinin	3	3.24E-03	3.82
compositionally biased region:Pro-rich	9	2.80E-03	3.85
domain:CH 1	3	2.96E-03	4.07
domain:CH 2	3	2.96E-03	4.07
endosome	5	3.60E-03	4.09
GO:0003779~actin binding	6	3.67E-03	4.57
SM00150:SPEC	3	5.10E-03	4.67
actin-binding	5	4.13E-03	4.68
GO:0005925~focal adhesion	4	4.25E-03	4.89
actin binding	3	4.37E-03	4.94

Fig. S4. Network of protein–protein interactions between DE genes. Down-regulated genes are shown in blue; up-regulated genes are shown in red. (*A*) PPI network of all DE genes. (*B*, *Left*) Subnetwork in the PPI network that is enriched for CHD8-coexpressed hub genes (defined as genes in the top 10% of each of the four CHD8-correlated modules by intramodular connectivity; these genes are shown as squares in this figure). (*Right*) The table shows DAVID enrichments (FDR < 5%) for genes in this subnetwork.

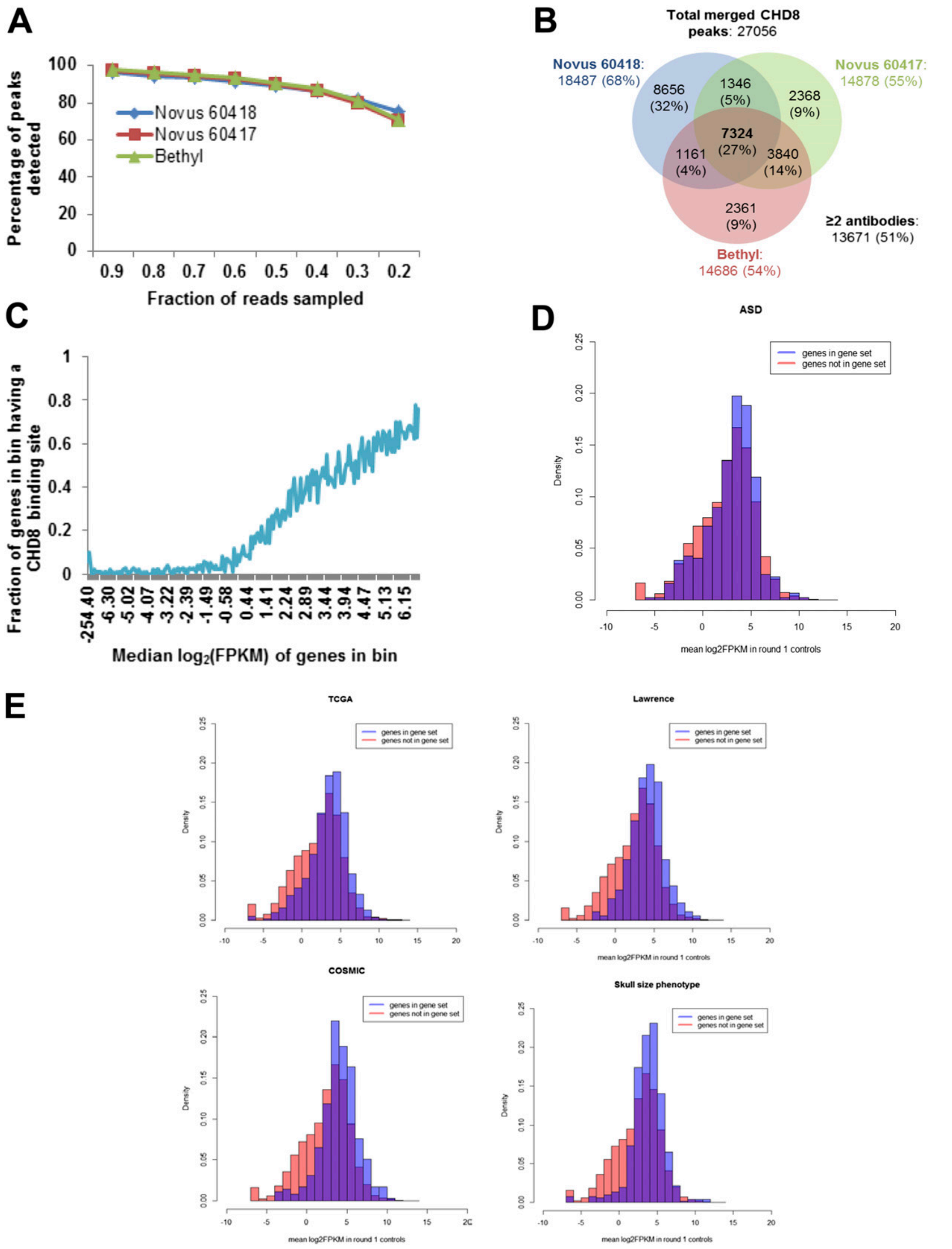


Fig. S5. ChIP-seq data for CHD8 binding in control NPCs. (*A*) Saturation plot for ChIP-seq data. For each antibody, the saturation curve shows the fraction of total peaks for that antibody that is detected (*y* axis) if the ChIP library is down-sampled by the fraction on the *x* axis. Saturation curves were obtained using MACS version 1.4.2. (*B*) Venn diagram of overlaps between merged peaks and lists of individual peak for CHD8. To obtain merged peaks, the union of all three peak lists was generated by concatenating them, and then overlapping peaks were merged into a single peak. The list of merged peak list then was compared with the lists of individual peak to get the overlaps shown in the Venn diagram. (*C*) Fraction of genes, ranked by expression level in controls (mean $\log_2$ FPKM), that have at least one CHD8-binding site. Each bin consists of 100 genes. The *x*-axis labels for each bin are median $\log_2$ FPKM for genes in the bin. (*D* and *E*) histograms of gene expression in controls (mean $\log_2$ FPKM) for genes that are in the gene set (blue) or not in the gene set (pink). The blue and pink bars are overlaid on top of each other, so overlapping sections appear purple.

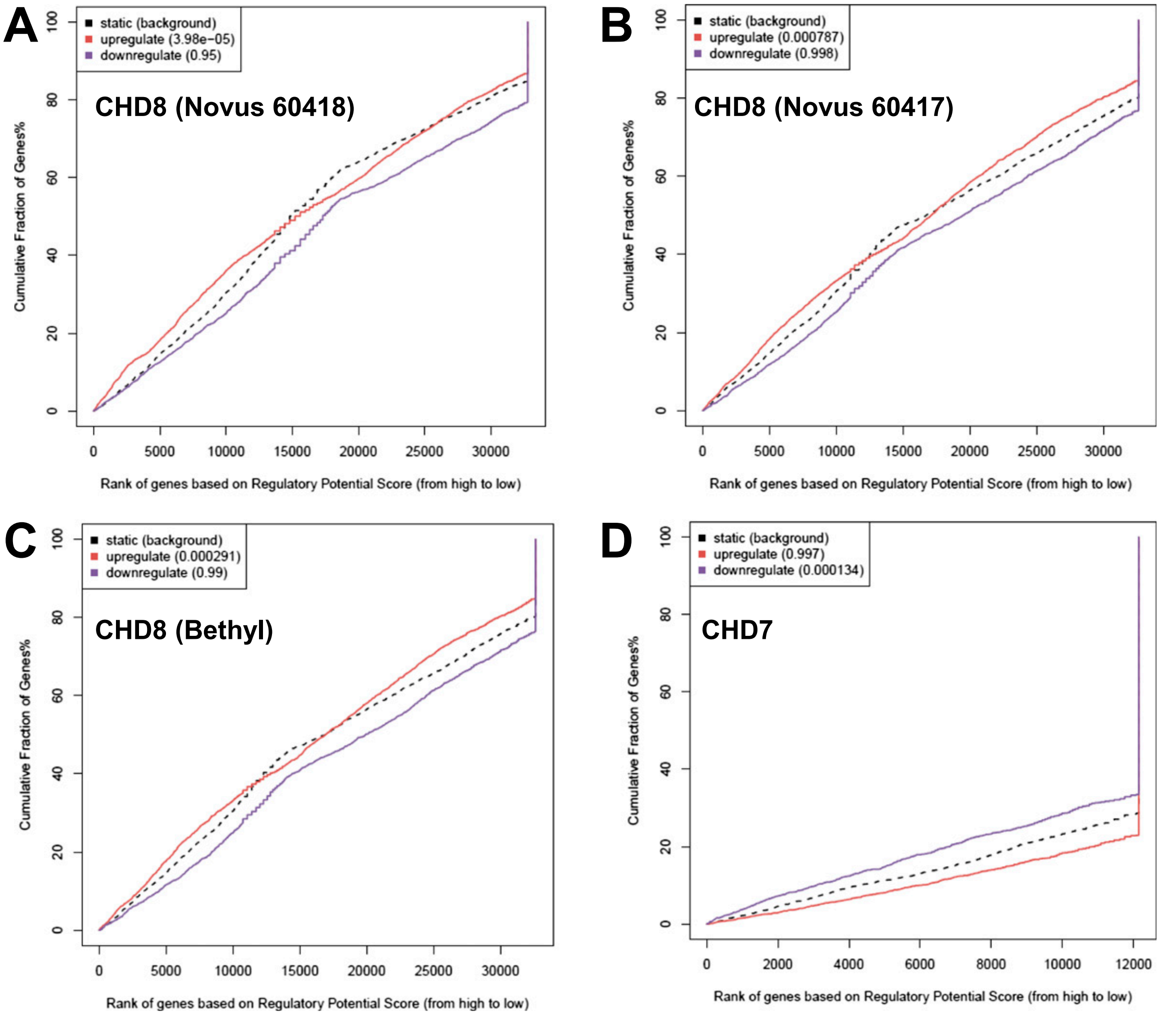


Fig. S6. Regulatory potential for CHD8- and CHD7-binding sites. Regulatory potential for the set of peaks identified by each of the three CHD8 antibodies (*A*–*C*) and for the CHD7 antibody (*D*) generated using BETA (13). Genes up-regulated by CHD8 knockdown are shown in red; down-regulated genes are shown in purple. The top left corner of each panel lists *P* values (one-tailed Kolmogorov–Smirnov test) for significance of regulatory potential for up- and down-regulated genes compared with background.

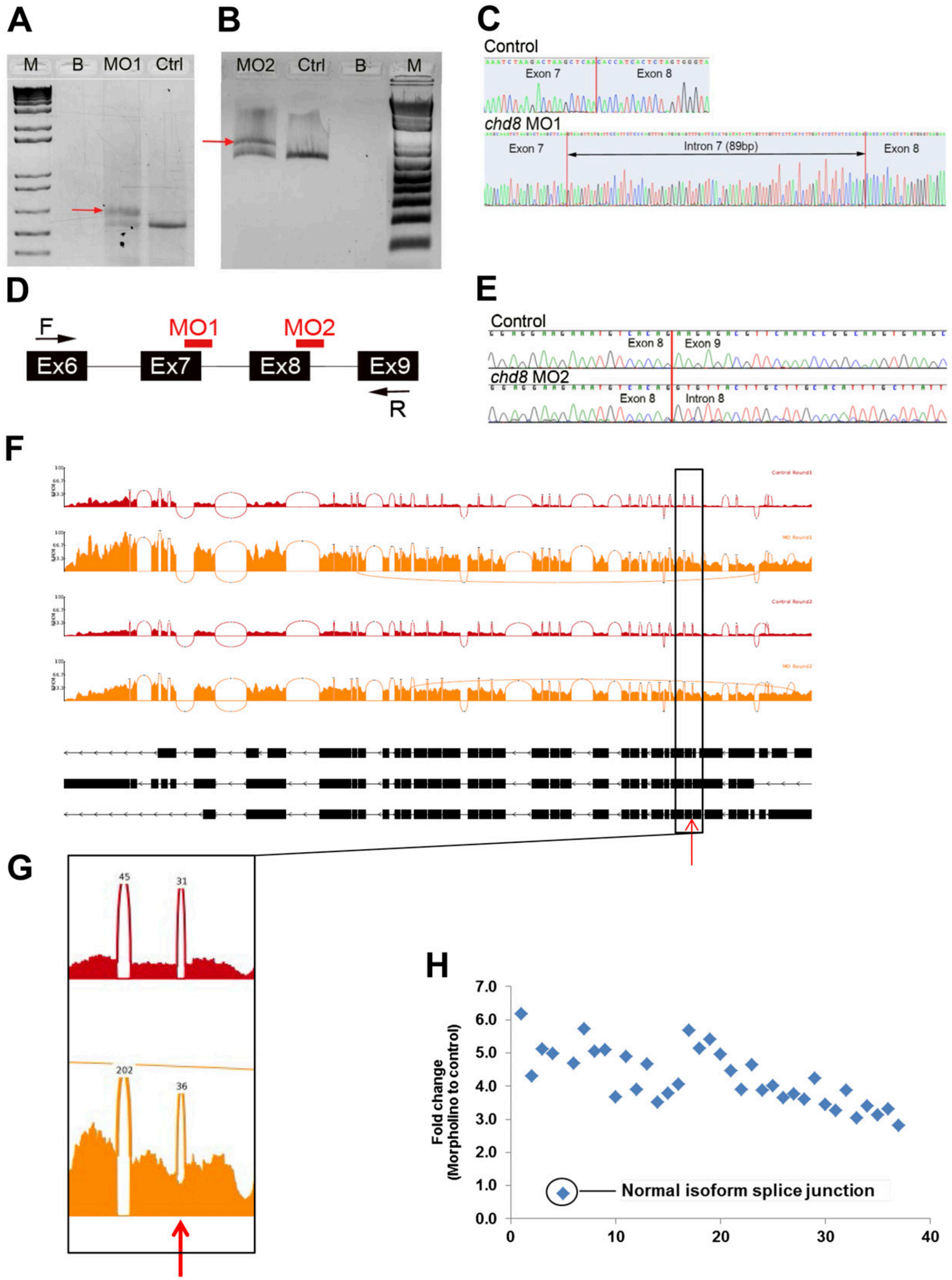


Fig. S7. *chd8* MOs efficiently disrupt the splicing of its zebrafish endogenous message. (*A* and *B*) Injection of *chd8* splice-blocking morpholinos *chd8*-MO1 (*A*) and *chd8*-MO2 (*B*) (10 ng each) results in abnormal splicing as shown by PCR amplification of cDNA reverse transcribed from extract total mRNA. M, 1-kb plus ladder; B, PCR blank; MO1, *chd8* MO1-injected; MO2, *chd8* MO2-injected; Ctrl, sham-injected. Red arrows indicate abnormal longer transcript. (*C*) Electropherograms showing normal splicing in controls and inclusion of intron 7 in embryos injected with *chd8*-MO1. (*D*) The *chd8*-MO1 and *chd8*-MO2 targeting splice sites are complementary to the seventh and eighth exon–intron boundary respectively. (*E*) Electropherograms showing normal splicing in controls and inclusion of intron 8 in embryos injected with *chd8*-MO2. Sequencing of the abnormal longer transcript (red arrows in *A* and *B*) confirms that the natural

Legend continued on following page

splicing sites are disrupted by the MOs and that full intronic sequences (intron 7 or 8) are included in morphants, leading to the appearance of a stop codon 4 bp after the end of exon 7 for *chd8*-MO1 and two consecutive stop codons 33 bp after the end of exon 8 for *chd8*-MO2. (*F*–*H*) RNA-seq of the *chd8* gene. (*F*) Sashimi plot showing split-read support for each exon–exon junction in the *Chd8* gene. The red arrow indicates an intron that is absent in the controls but with reads present in the MO-treated samples. All samples for each treatment in each round were pooled. Controls are shown in red and MO-treated samples in orange. (*G*) Sashimi plot zoomed in on the retained intron, for the round 2 samples. Similar numbers of split reads support the splice event in both controls and MO-treated samples (31 reads and 36 reads, respectively), but in MO-treated samples reads continue into the intron. (*H*) Fold changes in split reads covering each splice junction in MO-treated samples vs. control. All reads except one are more highly expressed in MO-treated than in control samples; the splice junction representing the retained intron, which is spliced out only in the normal isoform, is ~75% underexpressed in MO-treated samples as compared with controls.

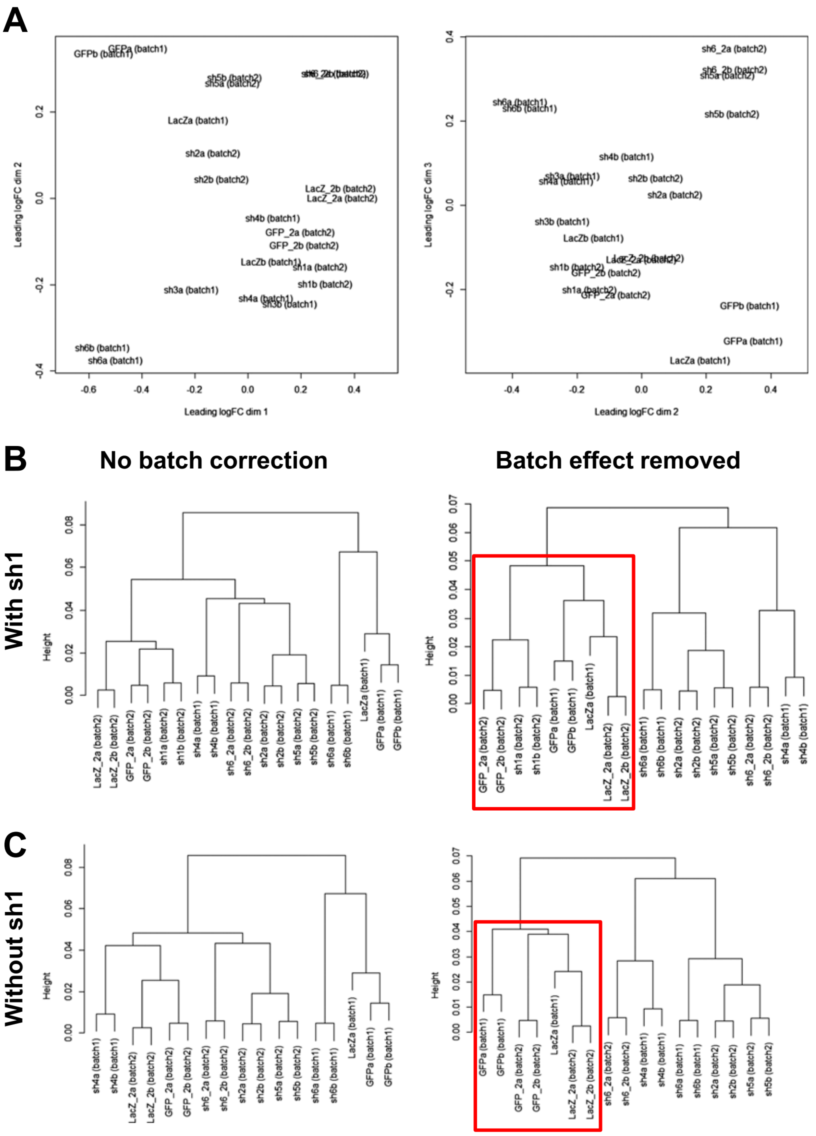


Fig. S8. Batch effect in RNA-seq libraries. (*A*) MDS plots for RNA-seq samples. (*Left*) Dimensions 1 and 2. (*Right*) Dimensions 2 and 3. Plots were generated from TMM-normalized log cpms, using the plotMDS function in the edgeR package. (*B* and *C*) RNA-seq sample clustering after removal of LacZb and sh3, before and after batch correction. Samples are clustered by correlation in gene expression (log cpm) for all genes. Batch correction was performed using the removeBatchEffect function in the edgeR package. The red box highlights the subcluster of control samples obtained when the batch effect is removed. (*B*) Clustering including sh1, which had the lowest level of knockdown and clusters with the controls. (*C*) Clustering excluding sh1.

Other Supporting Information Files

Datasets S1–S10 (XLSX)