CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors

Aarathi Sugathan[a,b,c,1], Marta Biagioli[a,c,1], Christelle Golzio[d,1], Serkan Erdin[a,b,1], Ian Blumenthal[a,b], Poornima Manavalan[a], Ashok Ragavendran[a,b], Harrison Brand[a,b,c], Diane Lucente[a], Judith Miles[e,f,g], Steven D. Sheridan[a,b,c], Alexei Stortchevoi[a,b], Manolis Kellis[h,i], Stephen J. Haggarty[a,b,c,i], Nicholas Katsanis[d,j], James F. Gusella[a,i,k], and Michael E. Talkowski[a,b,c,i,2]

[a]Molecular Neurogenetics Unit and [b]Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114; Departments of [c]Neurology and [k]Genetics, Harvard Medical School, Boston, MA 02115; [d]Center for Human Disease Modeling and [j]Department of Cell Biology, Duke University, Durham, NC 27710; Departments of [e]Pediatrics, [f]Medical Genetics, and [g]Pathology, The Thompson Center for Autism and Neurodevelopmental Disorders, University of Missouri Hospitals and Clinics, Columbia, MO 65201; [h]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139; and [i]Broad Institute of M.I.T. and Harvard, Cambridge, MA 02142

Edited* by Patricia K. Donahoe, Massachusetts General Hospital, Boston, MA, and approved September 12, 2014 (received for review March 24, 2014)

Truncating mutations of chromodomain helicase DNA-binding protein 8 (*CHD8*), and of many other genes with diverse functions, are strong-effect risk factors for autism spectrum disorder (ASD), suggesting multiple mechanisms of pathogenesis. We explored the transcriptional networks that *CHD8* regulates in neural progenitor cells (NPCs) by reducing its expression and then integrating transcriptome sequencing (RNA sequencing) with genome-wide CHD8 binding (ChIP sequencing). Suppressing *CHD8* to levels comparable with the loss of a single allele caused altered expression of 1,756 genes, 64.9% of which were up-regulated. CHD8 showed widespread binding to chromatin, with 7,324 replicated sites that marked 5,658 genes. Integration of these data suggests that a limited array of direct regulatory effects of *CHD8* produced a much larger network of secondary expression changes. Genes indirectly down-regulated (i.e., without CHD8-binding sites) reflect pathways involved in brain development, including synapse formation, neuron differentiation, cell adhesion, and axon guidance, whereas CHD8-bound genes are strongly associated with chromatin modification and transcriptional regulation. Genes associated with ASD were strongly enriched among indirectly down-regulated loci ($P < 10^{-8}$) and CHD8-bound genes ($P = 0.0043$), which align with previously identified coexpression modules during fetal development. We also find an intriguing enrichment of cancer-related gene sets among CHD8-bound genes ($P < 10^{-10}$). In vivo suppression of *chd8* in zebrafish produced macrocephaly comparable to that of humans with inactivating mutations. These data indicate that heterozygous disruption of *CHD8* precipitates a network of gene-expression changes involved in neurodevelopmental pathways in which many ASD-associated genes may converge on shared mechanisms of pathogenesis.

CHD8 | NPCs | RNA-seq | ChIP-seq | autism

The genetic architecture of autism spectrum disorder (ASD) is complex and heterogeneous. A wave of recent discoveries has identified individual genes that contribute to ASD when they suffer heterozygous inactivation by coding mutation, copy number variation, or balanced chromosomal abnormalities (1–7). Many of these genes fit neatly into current biological models of ASD involving altered synaptic structure and glutamatergic neurotransmission, but others have been surprising, with a less ready biological interpretation, including genes involved in chromatin modification, DNA methylation, cell adhesion, and global transcriptional regulation. This diversity of genes predisposing to ASD suggests either that there are many pathways that independently can result in the autism phenotype or that functionally distinct ASD-risk genes can trigger consequences that converge on a limited number of shared pathways of ASD pathogenesis. Because now experimental tools are available to reduce gene expression specifically in human neural progenitor cells (NPCs), which can mimic the impact of functional hemizygosity, we have explored this question by investigating the functional genomic consequences of suppressing chromodomain helicase DNA-binding protein 8 (*CHD8*), a particularly penetrant ASD gene.

CHD8 is an ATP-dependent chromatin remodeler of the SNF2 family (8). *CHD8* was identified as one of the genes in the minimal region of overlap of de novo 14q11.2 microdeletions in two children with developmental delay and cognitive impairment (9). We previously detected direct disruption of *CHD8* by a de novo balanced translocation, with concomitantly reduced mRNA expression, in a patient diagnosed with ASD, intellectual disability, obsessive–compulsive disorder, precocious puberty, macrocephaly, and mild facial dysmorphism (7). No other clinical abnormalities were observed in a recent follow-up examination. Concurrent

Significance

Truncating mutation of chromodomain helicase DNA-binding protein 8 (*CHD8*) represents one of the strongest known risk factors for autism spectrum disorder (ASD). We mimicked the effects of such heterozygous loss-of-function mutations in neural progenitor cells and integrated RNA sequencing with genome-wide delineation of CHD8 binding. Our results reveal that the molecular mechanism by which *CHD8* alters neurodevelopmental pathways may involve both direct and indirect effects, the latter involving down-regulation following *CHD8* suppression. We also find that *chd8* suppression in zebrafish results in macrocephaly, consistent with observations in patients harboring loss-of-function mutations. We show that reduced expression of *CHD8* impacts a variety of other functionally distinct ASD-associated genes, suggesting that the diverse functions of ASD risk factors may constitute multiple means of triggering a smaller number of final common pathways.

Author contributions: M.B., C.G., N.K., J.F.G., and M.E.T. designed research; A. Sugathan, C.G., S.E., I.B., P.M., A.R., D.L., J.M., and A. Stortchevoi performed research; S.D.S., M.K., and S.J.H. contributed new reagents/analytic tools; A. Sugathan, S.E., I.B., A.R., and H.B. analyzed data; C.G., P.M., and A. Stortchevoi performed experiments; D.L. and J.M. provided clinical phenotype information; and A. Sugathan, M.B., C.G., J.F.G., and M.E.T. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Data deposition: The sequences reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database (accession no. GSE61492).

[1]A. Sugathan, M.B., C.G., and S.E. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: talkowski@chgr.mgh.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1405266111/-/DCSupplemental.

exome sequencing and targeted mutation screening studies have now confirmed unambiguously that de novo truncating mutations of *CHD8* are among the strongest individual risk factors for ASD (1–4, 10). Interestingly, *CHD8* alterations also have been described as somatic events in gastric, colorectal, skin, and glioblastoma multiforme cancers (11–14). Despite its evident importance, little is known from previous studies about the cellular and molecular consequences of disrupting a single copy of *CHD8* and the regulatory connection between this chromatin-remodeling enzyme and the critical pathways associated with either neurodevelopment or cancer. Therefore we sought to determine the effects of perturbing the network of genes regulated by *CHD8* in early neural development by suppressing its expression in human induced pluripotent stem cell (iPSC)-derived NPCs. We integrated transcriptome sequencing (RNA-seq), to evaluate the consequences of *CHD8* suppression on global gene expression, with delineation of the genome-wide distribution of CHD8-binding sites using ChIP sequencing (ChIP-seq). Our findings indicate that *CHD8* regulates many functionally distinct genes associated with ASD and members of pathways important to neurodevelopment, suggesting that apparently diverse genetic lesions actually converge on shared pathways of ASD pathogenesis.

Results

Generation and Characterization of Stable *CHD8* Knockdown NPCs. Fig. 1*A* provides an overview of the integrative functional genomic approach. To mimic the ∼50% reduction in expression of *CHD8* expected to result from heterozygous inactivating mutation [and actually observed in lymphoblasts from our index translocation case (7)], we used lentiviral delivery of shRNAs into a cell type more relevant to ASD, a previously characterized human iPSC-derived NPC line from a control individual, GM8330-8 (15). We used six independent shRNAs targeting *CHD8* coding sequences to ensure a high number of biological replicates and two controls designed against the coding sequence of GFP and bacterial β-galactosidase (LacZ), respectively. All experiments were performed in duplicate, and, in addition, independent infection in each of two batches was carried out for one *CHD8* hairpin (sh6 and sh6_2), one GFP hairpin (GFP and GFP_2), and one LacZ hairpin (LacZ and LacZ_2) (Fig. 1). We performed RNA-seq on all lines using our previously published customization of the strand-specific dUTP method (Fig. 1 *B* and *C*) (16) and also carried out Western blotting using three independent commercially available antibodies (Fig. 1*D*). The knockdown of *CHD8* RNA ranged from 38–69% across lines (Fig. 1*C*). The degree of CHD8 suppression did not affect NPC morphology or the expression levels of the neural ectodermal markers paired box 6 (*PAX6*), sex determining region Y-box 1 (*SOX1*), and Musashi homolog 1 (*MSI1*) in comparison with nontargeting controls (GFP and LacZ) (Fig. S1*A* and Dataset S1).

Transcriptional Consequences of *CHD8* Suppression in NPCs. We generated an average of 40.6 million reads per line by strand-specific RNA-seq to monitor changes in genome-wide gene expression (Dataset S2*B*). All libraries contained synthetic RNA spike-ins, which we used to determine empirical transcript detection thresholds (Fig. S2 *B* and *C*); transcripts from 15,903 genes were detectable above thresholds in all lines. We performed analysis of differential expression incorporating batch and treatment as factors in a regression model (17). Overall, 1,756 genes were differentially expressed as a consequence of *CHD8* suppression (nominal $P < 0.05$), 369 of which were significant at Benjamini–Hochberg $q < 0.05$ (see Fig. S2*D* for the full range of differentially expressed genes from $q < 0.1$ to $q < 0.0001$). Many more genes were up-regulated than down-regulated following *CHD8* suppression (1,140 vs. 616). Pathway and gene ontology (GO) term enrichment analysis revealed a striking difference in the nature of the pathways represented by up-regulated and down-regulated genes (Fig. 2 and Dataset S3). The former, larger set was associated with the terms "chondroitin sulfate biosynthesis," "cytoplasmic sequestering of protein," and "RING-type zinc fingers" (Fig. 2 and Dataset S3), whereas the smaller latter group was associated with terms related to neural development and function including "cell adhesion," "neuron differentiation," "synapse," "ion transport," "axon guidance," "cadherin signaling pathway," and "protocadherin gene family" (Fig. 2 and Dataset S3). Weighted gene coexpression network analysis (WGCNA) (18) clustered all 15,903 genes into 21 modules and identified four modules that had very high correlation with *CHD8* expression ($r > 0.7$ and $P < 1 \times 10^{-4}$) (see Figs. S3 and S4 for all coexpression modules and protein–protein interactions and *SI Materials and Methods* for complete details). Consistent with the pathways associated with down-regulated genes, "cell adhesion" (in addition to "Wnt signaling" and "cell projection") was among the most enriched annotation terms for coexpression modules with genes whose expression decreased in correlation with *CHD8* suppression (Fig. S3).

These analyses provide a direct link between the chromatin modifier *CHD8* and regulation of genes of critical importance to neural development in humans. Many of the strongest individual effects were detected among genes involved in neuronal function or synaptic regulation [e.g., laminin alpha 4 (*LAMA4*), $P = 5.95 \times 10^{-13}$; neural cell adhesion molecule 1 (*NCAM1*), $P = 2.99 \times 10^{-10}$; *LRRC4B*, $P = 9.17 \times 10^{-10}$; *TIMP3*, $P = 2.65 \times 10^{-10}$; multiple EGF-like-domains 10 (*MEGF10*), $P = 1.1 \times 10^{-8}$; discs large homolog 2 (*DLG2*), $P = 1.19 \times 10^{-8}$; *SLIT1*, $P = 1.38 \times 10^{-8}$, to list a few], including genes previously implicated in ASD risk [e.g., sodium channel, voltage-gated, type II, alpha subunit (*SCN2A*), $P = 3.85 \times 10^{-9}$; methyl-CpG binding domain protein 3 (*MBD3*), $P = 4.9 \times 10^{-8}$; SH3 and multiple ankyrin repeat domains 3 (*SHANK3*), $P = 2.4 \times 10^{-4}$]. The majority of the most significant genes involved in neural development were down-regulated. Consequently, we examined the entire set of down-regulated genes and found it to be strongly enriched for 628 genes associated with ASD [$P = 3.25 \times 10^{-8}$, odds ratio (OR) = 2.78] as defined by the SFARI gene 2.0 (19) (574 genes) and AutismKB (20) (171genes; 117 overlap with SFARI) databases, both of which have established varying levels of support for an association with ASD of each gene included (see *SI Materials and Methods* for details on supporting evidence for these gene lists).

Based upon the pathways associated with up-regulated genes and the reported association of *CHD8* with multiple cancers, we performed similar analyses using a large list of 5,873 cancer-associated genes ("TCGA cancer") compiled from a variety of studies by The Cancer Genome Atlas (TCGA) Gene Ranker (*Materials and Methods*). We found enrichment of these TCGA-defined genes among the up-regulated loci ($P = 9.82 \times 10^{-7}$, OR = 1.37). Notably, these results were specific, because there was no enrichment of ASD genes in the up-regulated set ($P = 0.79$) and no enrichment of TCGA cancer loci in the down-regulated set ($P = 0.51$), prompting us to investigate further the molecular mechanisms driving these differences.

Genome-Wide Targets of CHD8 Binding. Because the genome-wide binding sites of CHD8 had not been delineated previously, we performed ChIP-seq in control cells using three independent CHD8 antibodies to distinguish indirect from potentially direct consequences of changing CHD8 levels. Very deep sequencing was performed for all antibodies (61–84 million reads per antibody) and input (88 million reads). All three CHD8 antibodies yielded similar genomic distributions that were highly enriched for peaks at promoter regions (Fig. 3 *A*–*C*). Overall, we identified widespread binding of CHD8 throughout the genome, with 7,324 sites that were replicated by all three antibodies at a Benjamini–Hochberg q value < 0.05 using Model-Based Analysis of ChIP-Seq 2

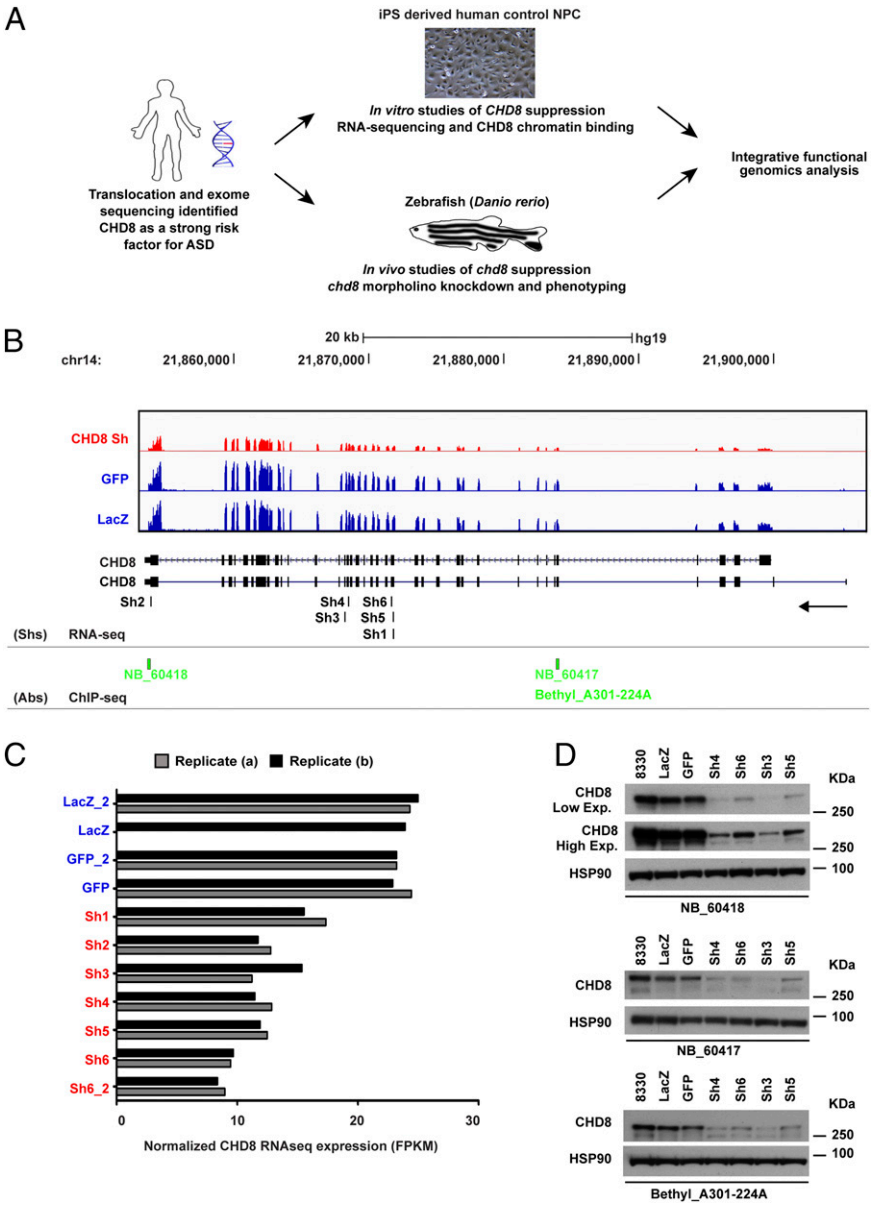


Fig. 1. Generation and characterization of human NPC lines with stable *CHD8* knockdown. (*A*) Schematic representation of the study design and experimental flowchart presented in the article. Following the identification of *CHD8* as a strong risk factor for ASD, we characterized the transcriptional effects of *CHD8* knockdown in human control NPCs and the genome-wide binding targets of CHD8. In parallel, we analyzed the in vivo phenotypes associated with *chd8* suppression in zebrafish. The functional genomics output emerging from integrated analyses of these datasets is discussed in this paper. (*B*) *CHD8* expression levels measured from RNA-seq are shown for control NPCs (blue) and stable CHD8 knock-down (KD) clones (red). Pooled tracks for all samples in each condition are presented. Track height is proportional to total library size. The locations of the different shRNA sequences used (Sh1–Sh6) are indicated at the bottom of the graph (RNA-seq row), and the epitope regions of the different CHD8 antibodies used in the ChIP-seq studies are indicated in green (ChIP-seq row). (*C*) Normalized expression levels of *CHD8* transcript are plotted for technical (Replicate a) and biological (Replicate b) replicates as normalized expression values for convenience. FPKM, fragments per kilobase per million reads. Reduced *CHD8* expression was observed in all knockdown clones. LacZb is excluded; it was removed because of insufficient reads (*SI Materials and Methods*). (*D*) Western blotting analysis of CHD8 protein levels in *CHD8* stable knockdown clones for each of the three antibodies used (NB_60417, NB_60418, and Bethyl A301-224A). Two different isoforms of CHD8 protein (~270 kDa and ~290 kDa) were observed in the control lines (8330, GFP, and LacZ), and down-regulation of protein was observed only for *CHD8* knockdown clones (Sh3, Sh4, Sh5, and Sh6). Comparable amounts of total protein were used for different samples, and HSP90 was used as loading control.

(MACS2) (Fig. S5*B*). We focused all subsequent integration with expression-related analyses on this stringently defined set of replicated sites. Overlaying the CHD8-binding sites with genome-wide chromatin states from www.broadinstitute.org/~anshul/projects/roadmap/segmentations/models/coreMarks/parallel/set2/final/ (accessed May 28, 2014) in an ES cell-derived neural progenitor line generated by the Roadmap Epigenomics consortium (21), we found that CHD8-binding sites are localized predominantly to genomic regions marked by histone H3 trimethyl Lys4 (H3K4me3), signifying active transcription start sites (TSSs), with 83% of CHD8-binding sites being in an active TSS state, as compared with only 1% of the whole genome (Fig. 3*D*). CHD8-binding sites also are enriched to a lesser extent (twofold) for enhancer status (characterized by the presence of H3K4me1), comprising 4% of sites as compared with 2% of the whole genome (Fig. 3*D*). In our expression dataset, 5,658 genes

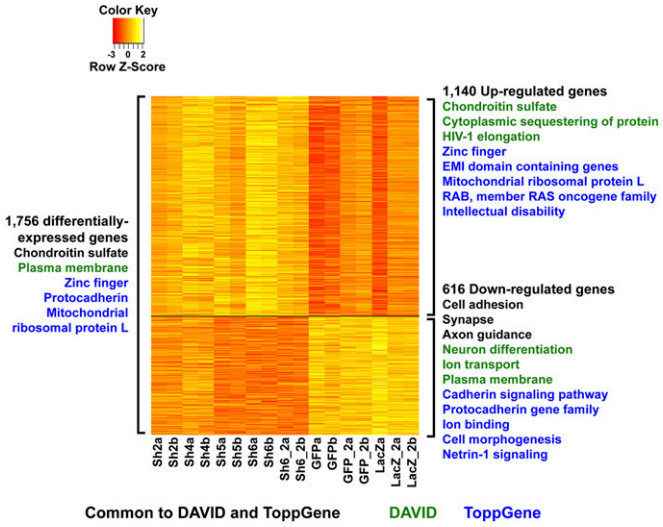


Fig. 2. Differentially expressed genes and associated annotation terms. The heatmap shows gene expression in $\log_2$ cpm after batch correction for the 1,756 differentially expressed genes, with genes down-regulated by CHD8 knockdown (616 genes) on the bottom and genes up-regulated following CHD8 suppression (1,140 genes) on the top. Values have been centered and scaled for each row. Each row represents a single gene. Statistically significant functional annotation and pathway terms identified using DAVID (FDR < 5%) and ToppGene (Bonferroni-corrected $P < 0.05$) for all 1,756 differentially expressed genes are listed on the left. On the right of the heatmap, significant terms are provided for down- and up-regulated genes separately. Similar terms have been condensed and summarized for simplicity in this figure; the full list of associated terms and P values is provided in Dataset S3. The most significant terms for up-regulated genes were "chondroitin sulfate biosynthesis" ($P = 2.55 \times 10^{-6}$) and "mitochondrial ribosomal protein L genes" ($P = 2.28 \times 10^{-6}$); for down-regulated genes the most significant terms were "plasma membrane" ($P = 4.31 \times 10^{-11}$), "protocadherin genes" ($P = 1.16 \times 10^{-10}$), "calcium ion binding" ($P = 1.35 \times 10^{-7}$), and "single organismal cell–cell adhesion" ($P = 4.78 \times 10^{-7}$). P values for synapse, neuron differentiation, and axon guidance among down-regulated genes ranged from 2.55×10^{-3} to 2.81×10^{-5} (see Dataset S3 for complete results).

contained at least one CHD8-binding site within 10 kb of the TSS, and the proportion of genes that have a CHD8-binding site increased with increasing baseline gene expression (Fig. S5*C*). Analysis of the set of all genes with CHD8-binding sites using GREAT (22) yielded many enriched functional annotations, but the top GO molecular functions were related to transcriptional regulation, and enriched terms from pathway databases included "p53 pathway," "Hedgehog signaling pathway," and "cell cycle" (Dataset S4). In this system, 522 of the genes with CHD8-binding sites were differentially expressed, representing only 9.2% of all CHD8-bound genes and 29.7% of all differentially expressed genes, indicating that the majority of the gene-expression changes that we detected upon *CHD8* suppression are likely to be caused by indirect regulatory effects. A higher proportion of up-regulated genes than down-regulated genes are bound by CHD8 (*SI Materials and Methods* and Fig. S6 *A*–*C*).

To explore sequence motifs at the CHD8-binding sites, we carried out de novo motif analyses separately for the sets of binding sites detected by each of the three antibodies. We found that a motif matching that for CCCTC-binding factor (CTCF) binding was the most significant motif that was replicated between antibodies (*SI Materials and Methods* and Dataset S5), as is consistent with CTCF being a previously known interactor of CHD8 (23). A motif matching the binding site for yin yang 1 (YY1), a ubiquitous transcription factor that interacts with CTCF (24), also was discovered, as the second most significant hit (Dataset S5). Other enriched motifs that may represent coactivators or corepressors with CHD8 are listed in Dataset S5.

Distinct Characteristics of *CHD8* Regulation Among ASD and Cancer Gene Sets. Integration of CHD8-binding sites with the differentially expressed genes gave further insight into the gene enrichments initially noted and into the diverse functions of genes associated with ASD. It was the set of down-regulated genes that lack CHD8-binding sites (i.e., genes down-regulated indirectly by *CHD8* suppression) that was enriched for neurodevelopmental, cadherin/cell adhesion, and axon guidance pathways (Datasets S6–S8; the most significant genes are listed in Table 1). Among these indirectly down-regulated genes there also was a strong enrichment for the broad set of ASD-associated genes as defined by SFARI and AutismKB ($P = 1.09 \times 10^{-9}$, OR = 3.39), which includes genes discovered from previous genetic studies as well as genes investigated based on prior neurobiological hypotheses of ASD. This gene set also was nominally significant for enrichment among genes with CHD8-binding sites (Fig. 4*A*). Integration of CHD8-binding sites also provided further specificity of the TCGA cancer-associated gene-set enrichment. These genes were highly enriched among all genes with CHD8-binding sites ($P = 1.55 \times 10^{-58}$, OR = 1.80), a result that was significant regardless of whether the genes were differentially expressed. This same gene set was not enriched significantly among either up-regulated or down-regulated genes that did not possess CHD8-binding sites, indicating that potential for direct regulation by CHD8 was the primary driver of the TCGA enrichments (Dataset S8*A*).

We further scrutinized these findings by testing for enrichment of genes from additional ASD and cancer gene sets that were compiled using different criteria. We first evaluated a smaller list of SFARI ASD genes curated using the same criteria applied by Parikshak et al. (25) [235 genes; SFARI gene score = syndromic (S) or evidence level 1 ("high confidence") to 4 ("minimal evidence")]. Like the larger ASD gene set, this restricted set was enriched among down-regulated ASD genes lacking CHD8-binding sites ($P = 2.26 \times 10^{-2}$, OR = 2.18; Dataset S8*A*) and was not associated with CHD8-bound genes ($P = 0.16$; Dataset S8*A*). When we performed analyses with a narrowly defined set of genes associated with ASD based on harboring at least a single de novo loss-of-function (LoF) mutation from exome sequencing studies [131 genes were tested by combining nine statistically significant "high-confidence ASD" genes with another 122 "probable ASD" genes defined by Willsey et al. (26)], we observed enrichment among CHD8-bound genes ($P = 4.34 \times 10^{-3}$, OR = 1.69) but not among down-regulated genes (Fig. 4*A*), in contrast with the SFARI/AutismKB results. When we interrogated the genes and pathways underlying the differences among these ASD gene sets, we found a consistent pattern: The genes identified from de novo LoF mutation were enriched for CHD8 binding and consistently were associated with chromatin modification and transcriptional regulation (Dataset S8*B*), whereas the SFARI/AutismKB dataset was enriched for genes without CHD8-binding sites that were down-regulated following CHD8 suppression, and these genes were associated with cell adhesion and neurotransmitter/axon-related pathways (Dataset S8*B*).

These disparate enrichment patterns for ASD-associated genes based on the mode of discovery and pathways implicated also shed light on the patterns observed in network analyses performed by Parikshak et al. (25) and Willsey et al. (26), who generated coexpression networks across brain regions and development from BrainSpan (www.brainspan.org). Of the five coexpression modules found to be enriched for ASD genes in the study by Parikshak et al., two (M2 and M3) consisted of genes expressed early in fetal neocortical development and were enriched for transcriptional and chromatin regulators and genes harboring rare de novo LoF mutations from previous exome

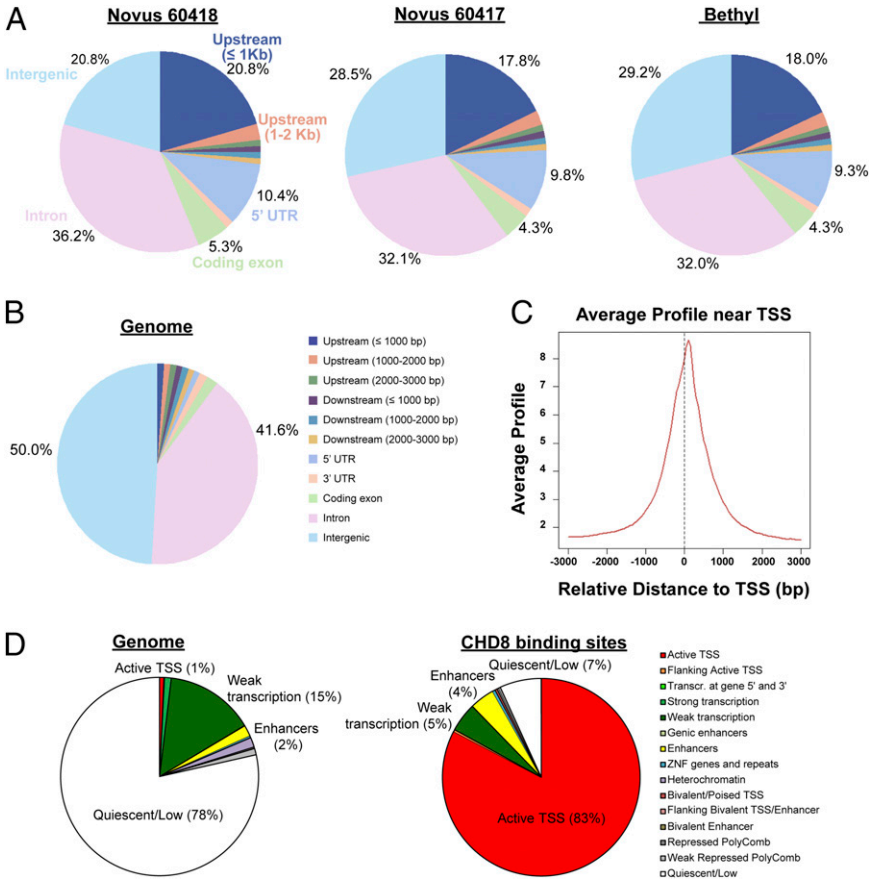


Fig. 3. Distribution of ChIP-seq peaks from three CHD8 antibodies. (*A*) Genomic distribution of sequence peaks captured by each of the three antibodies, compared with the whole genome. Upstream regions are defined as regions upstream of the TSS; the 5′ UTR is the region between the TSS and the coding start site. (*B*) Whole-genome distribution of the genomic features in *A*. "Intergenic" refers to anything that does not fall into any of the preceding categories in the legend shown on the right. (*C*) ChIP-seq read density relative to TSSs for one representative antibody (Novus 60417). We found 7,324 peaks that were detected by all three antibodies. These peaks were mapped to 5,658 genes. (*D*) Distribution of chromatin states identified by the Roadmap Epigenomics consortium (21) in an ES cell-derived NPC for the whole genome (*Left*) and the 7,324 CHD8-binding sites detected by all three antibodies (*Right*).

studies. Three modules (M13, M16, and M17), consisting of genes expressed in late fetal to early postnatal stages, were enriched for synaptic proteins and SFARI ASD genes. We observed a clear separation in the overlap between our gene categories defined by *CHD8* regulation and these five modules: Down-regulated genes without CHD8-binding sites were enriched for genes belonging to M13, M16, and M17 but not M2 or M3, whereas CHD8-bound genes were enriched for genes belonging to M2 and M3 but not M13, M16, or M17 (Fig. 4*B*). We also obtained concordant patterns with the complementary study carried out by Willsey et al. (26), in which three coexpression modules were found to be enriched for de novo LoF ASD genes. All three of those modules were enriched among the set of all CHD8-bound genes in our study (Fig. 4*B*) but not the indirectly down-regulated genes.

In cancer, the TCGA gene set is both large and broadly defined, so we similarly turned to two gene sets compiled using narrower criteria: one much smaller set of 224 genes based upon somatic point mutations in 21 tumor types from a recent publication from Lawrence et al. (13) and a second manually curated Catalog of Somatic Mutations in Cancer (COSMIC) cancer gene census of 513 genes based on causally implicated mutations (27). Both smaller cancer gene sets also were highly enriched among genes with CHD8-binding sites and most strongly among the subset of CHD8 targets that do not show differential expression in NPCs because of CHD8 knockdown ($P < 1.9 \times 10^{-11}$ for all three gene sets; Fig. 4*A* and Dataset S8*A*). Notably, when we considered the 302 genes present in the SFARI/AutismKB ASD and TCGA cancer genes sets (the two most broadly defined gene sets), we found the same pattern of enrichment as in the SFARI/AutismKB genes overall, with enrichment only among down-regulated and unbound genes ($P = 0.023$, OR = 2.01; Dataset S8*C*).

Given the significance of the findings in ASD and cancer gene sets, we sought to establish their specificity. We tested all the 184 gene lists available from other complex human diseases and traits obtained from the National Human Genome Research Institute (NHGRI) Genome-Wide Association Study (GWAS) Catalog (28) (minimum gene-set size = 10 genes). No results approached the significance levels of the ASD and cancer enrichments, because the most significant result was human trait-related enrichment "height" ($P = 4.46 \times 10^{-6}$ among the set of CHD8-bound genes; Dataset S9). In comparison, the most significant ASD and cancer enrichments were 1.01×10^{-9} and 1.55×10^{-58}, respectively (Fig. 4*A*). However, we did find enrichment of curated gene sets associated with schizophrenia (29) and bipolar disorder (30) among genes down-regulated but not bound by CHD8, following the same pattern as the SFARI/AutismKB ASD genes (2.36×10^{-2} and 1.70×10^{-3}, OR = 2.45 and 3.28, respectively). Notably, we did not find enrichment of the genes in proximity to the 108 common polymorphisms that were significant genome-wide in a recently published study of schizophrenia (31). Genes associated with intellectual disability (25) also were enriched among down-regulated genes but, like the ASD genes defined by truncating mutations, were enriched more significantly among

Table 1. Differentially expressed and CHD8-bound genes associated with ASD and neurodevelopmental pathways

Gene	ASD list	FC, knockdown/control	*P* value
Selected genes that are indirectly regulated by CHD8*			
LAMA4		−4.44	5.95×10^{-13}
TIMP3		−3.23	2.65×10^{-10}
KCNJ10	S/A	−4.95	3.44×10^{-10}
SCN2A	Both	−7.31	3.85×10^{-9}
SLIT1		−4.88	1.38×10^{-8}
MBD3	S/A	2.72	4.90×10^{-8}
BAI1		−3.25	2.40×10^{-7}
SYTL4		−3.45	2.79×10^{-7}
GPX1	S/A	1.99	2.58×10^{-5}
SOX9		−2.08	3.33×10^{-5}
HS3ST5	S/A	6.91	4.05×10^{-5}
ACSBG1		−2.16	6.88×10^{-5}
SHANK3	S/A	1.87	2.39×10^{-4}
EFHD1		−2.87	2.66×10^{-4}
LIFR		−1.78	3.02×10^{-4}
TESK2		−2.41	4.03×10^{-4}
HYDIN	S/A	−3.30	7.09×10^{-4}
PLXNA2		−2.20	8.01×10^{-4}
ANO5	W	−1.93	9.12×10^{-4}
RIMS3	S/A	−1.73	9.64×10^{-4}
KANK1	S/A	−1.74	9.95×10^{-4}
Selected ASD-associated genes that are bound by CHD8†			
CHD8‡	Both	−2.56	3.62×10^{-9}
NFKBIL1‡	W	2.02	7.61×10^{-5}
CBX4‡	W	1.85	3.79×10^{-4}
TCF3	W	1.69	1.17×10^{-3}
SDC2	S/A	1.79	1.20×10^{-3}
RAI1	S/A	1.69	1.57×10^{-3}
SLITRK5	S/A	1.69	2.03×10^{-3}
PTEN	S/A	−1.81	3.00×10^{-3}
ARHGAP2	S/A	2.67	5.33×10^{-3}
OGT	S/A	−1.49	0.0111
RPS6KA2	S/A	−1.51	0.0114
TRIO	S/A	1.44	0.0424
ADK	S/A	−1.40	0.0430
LZTS2	S/A	1.39	0.0432
CBS	S/A	1.38	0.0437
POGZ	Both	−1.30	0.100
ARID1B	Both	−1.20	0.264
MBD5	Both	−1.19	0.307
TRIP12	Both	−1.17	0.337
ADNP	Both	−1.13	0.436
SETD2	Both	−1.05	0.757
SYNGAP1	Both	1.04	0.818
CUL3	Both	1.01	0.951
SUV420H	Both	1.00	0.978

Both, genes in both the SFARI/AutismKB and Willsey et al. (26) lists; FC, fold change of gene in shRNA knockdowns compared with controls: A positive fold-change corresponds to up-regulation when CHD8 is knocked down; a negative-fold change corresponds to down-regulation (*P* values for differential expression are listed for each gene); S/A, genes in the ASD gene list from SFARI and AutismKB, as described in *SI Materials and Methods*; W, genes in the Willsey et al. (26) pASD or hcASD gene list. See Dataset S1 for all detectable genes and associated data.

*Genes that are strongly differentially expressed following *CHD8* suppression (Benjamin–Hochberg $q < 0.05$) and lack CHD8-binding (i.e., that are indirectly regulated by CHD8). Genes shown are either ASD-associated genes or down-regulated genes associated with cell adhesion or neurodevelopmental pathways.

†ASD-associated genes in which CHD8-binding sites were detected by all three antibodies ($q < 0.05$). Genes either were differentially expressed with a CHD8-binding site or were bound by CHD8 and found in both ASD gene sets described.

‡These genes met $q < 0.05$ for differential expression. FC, fold change of gene in shRNA knockdowns compared with controls. A positive fold-change corresponds to up-regulation when CHD8 is knocked down; a negative-fold change corresponds to down-regulation. P values for differential expression are listed for each gene.

CHD8-bound genes (Dataset S8*A*). Only "targets of fragile X mental retardation protein 1 (FMRP)," a gene set defined by molecular analysis (high-throughput sequencing together with UV-crosslinking and immunoprecipitation, HITS-CLIP) rather than by disease association (32), was enriched among up-regulated genes (as well as among CHD8-bound genes) (Dataset S8*A*). For CHD8-bound genes, one of the most significant human phenotypes from the Human Phenotype Ontology (HPO) database was "abnormality of skull size" ($P = 1.04 \times 10^{-23}$, OR = 2.49), following "abnormality of the cerebrum" and "abnormality of the forebrain" (Datasets S4 and S8*A*). All these are consistent with the clinical phenotype of our index case with translocation interrupting *CHD8*, who exhibits ASD, intellectual disability, and macrocephaly, and with the phenotypes of other subjects heterozygous for inactivating *CHD8* mutation.

In Vivo Analysis of *CHD8* in Zebrafish Embryos. Given that virtually all patients reported with truncating mutations in *CHD8* have a macrocephalic phenotype (33), we asked whether suppression of *chd8* might lead to increased head size in *Danio rerio* (zebrafish) embryos by acting on early neurogenesis. We have shown previously that head-size evaluations in zebrafish embryos can serve as a surrogate for the evaluation of candidate genes for neurocognitive traits (34). Therefore we suppressed the sole zebrafish ortholog of *CHD8* and evaluated the gross morphometric and cell-specific characteristics of morphants.

Using reciprocal BLAST, we first identified a single zebrafish *CHD8* ortholog (*chd8* on chromosome 2; 62% amino acid identity). Next, we designed two splice-blocking morpholinos (sb-MO), targeting the splice donor site of exons 7 and 8 respectively, which we injected into embryos at the one- to two-cell stage. Masked quantitative scoring of embryos at 4.5 d postfertilization (dpf) injected with *chd8* MO1 showed a reproducible macrocephaly phenotype (12% increase in morphants compared with controls, $P < 0.0001$) (Fig. 5 *A–C*). This phenotype was paralleled by the efficiency of splice blocking of the two sb-MOs, which led to the retention of introns 7 and 8, respectively, and the presence of a premature stop codon, as established by RT-PCR and Sanger sequencing (Fig. S7 *A–E*). We further characterized the *chd8* transcripts by quantitative PCR and RNA-seq to confirm the impact of the morpholino on *chd8* transcription (*SI Materials and Methods* and Fig. S7 *F–H*). A scrambled morpholino induced no phenotypes (34). Finally, macrocephaly was unlikely to be driven by overall developmental delay; morphants had a normal appearance with regard to their pigment cells, there was no apparent pathology in other external organs, such as the heart or the swim bladder, and their body length was indistinguishable from that of control embryos from the same clutch.

To probe further the underlying cause(s) of the macrocephalic phenotype, we stained embryos at 2 dpf with an anti-HuC/D (human neuronal protein Hu antigen, a marker for newborn neurons). We selected this time point because it precedes the development of macrocephaly and therefore allowed us to evaluate the forebrain before the appearance of gross anatomical defects. We observed a striking increase in HuC/D expression, which appeared ectopic in 65% of embryos injected with *chd8* MO1, as compared with controls (Fig. 5 *D–F*). Next, we stained the embryos with a phospho-histone H3 antibody, which is an M-phase marker, and quantitatively scored the number of proliferating cells in *chd8* morphants and controls. We counted an average of 408 p-histone $H3^+$ cells for controls compared with 518 p-histone $H3^+$ cells in embryos injected with *chd8* MO1 ($P = 0.0018$; Fig. 5 *G–I*), indicating that the macrocephaly phenotype is likely to be caused by disturbed neuronal proliferation at early developmental stages.

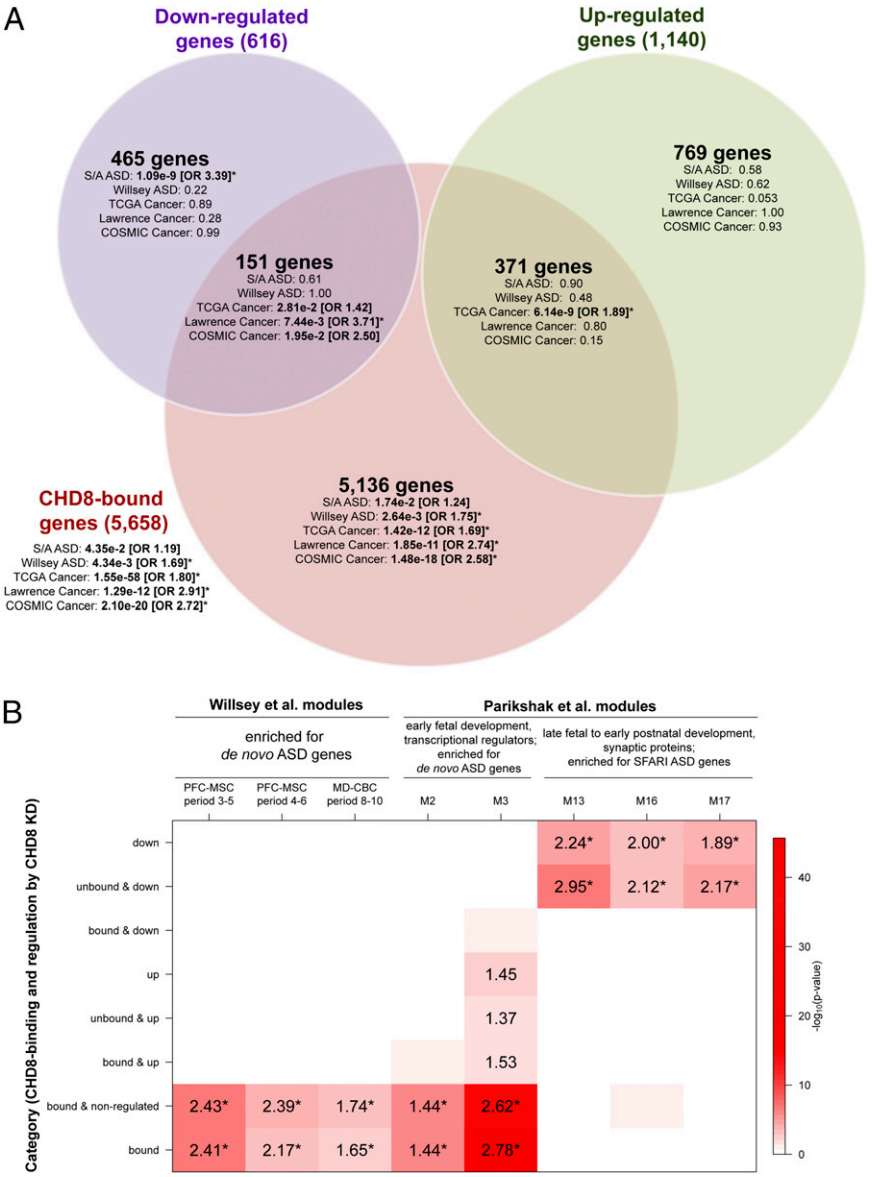


Fig. 4. Enrichments for ASD and cancer gene sets and published BrainSpan coexpression networks among CHD8-regulated and CHD8-bound genes. (*A*) Gene-set enrichments are shown for the sets of genes that are both differentially expressed and bound by CHD8, differentially expressed only, or CHD8-bound only, shown as a Venn diagram. Enrichments for the set of all genes bound by CHD8, independent of differential expression, are shown outside the Venn diagram. For each set of genes in the Venn diagram, enrichment *P* values are shown for five disease gene lists. ORs are shown for enrichments that met $P < 0.05$, with asterisks indicating enrichments that met $q < 0.05$. Disease gene lists were obtained from SFARI and AutismKB (S/A ASD), de novo LoF mutations in ASD are from Willsey et al. (26) (Willsey ASD), TCGA gene ranker (TCGA Cancer), the pan-cancer exome sequencing study by Lawrence et al. (13) (Lawrence Cancer), and the Wellcome trust ("COSMIC Cancer"), as described in *SI Materials and Methods*. All gene-set analyses and Benjamini–Hochberg-corrected *P* values are provided in Dataset S8*A*. (*B*) For each set of CHD8-regulated or CHD8-bound genes, enrichments are shown for overlap with BrainSpan coexpression modules generated by Willsey et al. (26) and Parikshak et al. (25) that had been found to be enriched for ASD genes. The red shading in each cell corresponds to the $\log_{10}$ *P* value for enrichment, as shown in the color scale on the right. The number in each cell is the OR for enrichment, shown only if the enrichment met $P < 0.05$. Enrichments that met $q < 0.05$ are indicated by an asterisk next to the OR. Names of coexpression modules are as reported in the respective publications.

Discussion

CHD8 is a prominent example of several genes, including other chromodomain helicases (*CHD7*, *CHD3, CHD2*), histone demethylases (*ARID1B*, *KDM6B, KDM6A*), methylases (*MLL5*, *EHMT1, METTL2B*), and methylated DNA-binding proteins (*MBD5, MBD3*), that have implicated the disruption of chromatin regulation as a precipitating factor in ASD. However, regulation of chromatin is but one of the many cellular processes that has been proposed by genetic and biological studies in ASD. Investigation of ASD pathogenesis also has focused on RNA surveillance, cell adhesion, synaptic proteins, glutamate neurotransmission, ion transport, and other functions, suggesting that ASD may involve related phenotypes caused by quite different pathogenic mechanisms. Previous studies have sought insight from curated protein–protein interaction (PPI) databases to connect the regulatory networks associated with genes involved in these biological functions (3). Recently, weighted gene coexpression network analysis (18) has been used in elegant studies to explore directly the coexpression of genes that drive neurodevelopment, measured either by microarray or RNA-seq, in human brains and

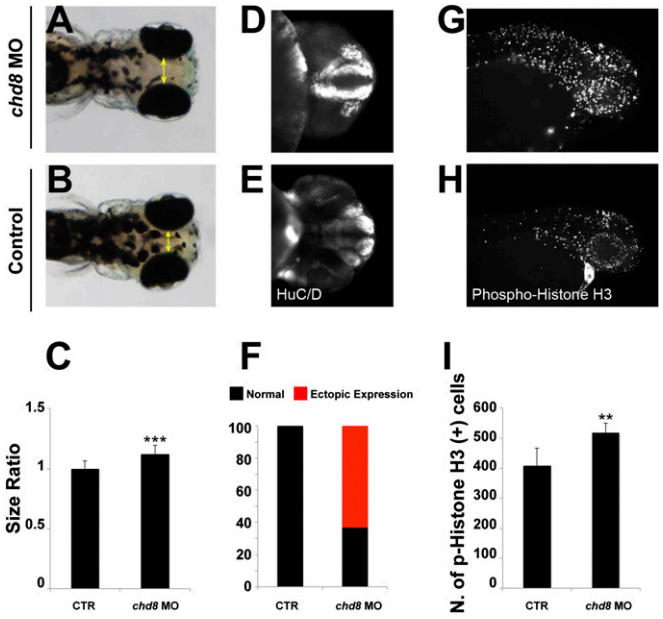


Fig. 5. Injection of *chd8* MO leads to macrocephaly, ectopic expression of HuC/D, and increased numbers of proliferating cells. (*A* and *B*) Representative images show dorsal views of an embryo injected with *chd8* MO (*A*) and a sham-injected control (*B*). (*C*) Quantification of macrocephaly was performed in embryo batches by measuring the distance across the convex tip of the eye cups (yellow arrows) at 4.5 dpf ($n = 70$ embryos; repeated three times). The macrocephaly phenotype represents a 12% increase compared with controls. $***P < 0.0001$ (Student *t* test). (*D* and *E*) Suppression of *chd8* leads to increased ectopic expression of HuC/D at 2 dpf. Representative images (with HuC/D-antibody staining) show the ventral views of an embryo injected with *chd8* MO and a sham-injected control. HuC/D levels in the anterior forebrain of the embryos injected with the *chd8* MO are significantly higher than in controls. (*F*) Percentage of embryos with normal (black) or ectopic (red) HuC/D protein levels in the anterior forebrain in embryo batches injected with *chd8* MO exhibit an ectopic expression of HuC/D compared with controls. (*G* and *H*) Phospho-histone H3 staining for proliferating cells in the zebrafish brain at 2 dpf. Representative images (with p-histone H3-antibody staining) show the lateral views of an embryo injected with *chd8* MO and a sham-injected control. (*I*) Quantification of p-histone H3–positive cells from control embryos or embryos injected with *chd8* MO ($n = 20$ embryos per group). Data are presented as the mean ± SEM. $**P = 0.0018$ (two-tailed *t* test comparisons between MO-injected and controls).

lymphoblastoid cell lines (25, 35). These studies have revealed coexpression modules containing multiple ASD genes and provided insights into developmental timing and regional specificity of transcriptional coexpression signatures (25, 26, 35). However, such correlation studies are not designed to examine directly the cause-and-effect relationships involving ASD mutations and their consequences. Therefore, we sought to examine the role of *CHD8* by establishing the functional genomic effects in human NPCs of reducing its expression to a level comparable to that expected from the heterozygous inactivating mutations seen in ASD.

This study shows that *CHD8* has a broad impact on the regulation of gene expression and reveals an intriguing contrast in the nature of the pathways altered by its suppression based on the directionality and direct/indirect nature of the effect. First, we find that *CHD8* mutation plays an indirect role in down-regulating gene expression in pathways involved in neurodevelopment, supporting a role for chromodomain helicases in neuronal differentiation (36–38). This mechanism connects *CHD8* to many other ASD-associated genes and canonical pathways thought to act in ASD pathogenesis (36–38). Mediators of this regulatory effect could represent high-priority targets for probing probe this mechanism further. Several genes involved in neurodevelopmental pathways are among the most significantly affected by *CHD8* suppression, most notably *SCN2A*, *DLG2*, *SHANK3*, and a number of cell-adhesion genes (*LAMA4*, *NCAM1*, *MEGF10*) ($P < 1.5 \times 10^{-8}$ for all genes), most of which are down-regulated. These observations suggest that *CHD8* mutation may precipitate abnormal neurodevelopment through its indirect regulatory effect on a network of neurodevelopmental genes, many of which are associated with ASD.

These data also show that the enrichments of various gene sets associated with ASD among genes regulated by CHD8 are sensitive to the underlying molecular and physiological functions of the genes. ASD-associated genes that are annotated as functioning in neuronal development are enriched more significantly among genes that are indirectly down-regulated following *CHD8* suppression, a finding derived exclusively from the SFARI and AutismKb gene sets, which include genes implicated in a range of genetic and neurobiological studies. Indeed, analysis of the combined SFARI/AutismKB gene set, irrespective of CHD8 regulation, reveals strongest association with pathways involved in synaptic transmission, cell–cell signaling, neuron differentiation, and neuronal development. On the other hand, the overall gene set defined by the presence of a de novo LoF mutation is associated more significantly with pathways involved in chromatin modification and protein methyltransferase activity, with less significant enrichment for pathways such as axon guidance, and is enriched among genes with CHD8-binding sites, regardless of expression. Overall, the nature of the ASD genes discovered to date indicates that pathogenesis may be precipitated by quite different triggers, but our data show that reduced CHD8 function is a trigger that produces altered expression of a number of other functionally distinct ASD genes, suggesting that these ASD genes may converge on common final pathways. The respective enrichments of CHD8-bound genes and down-regulated, unbound genes among the modules of chromatin/transcriptional regulators expressed in early fetal development (modules M2 and M3 in ref. 25) and of synaptic proteins expressed later in development (M13, M16, and M17 in ref. 25) further suggest that CHD8 may influence abnormal neurodevelopment by both direct and indirect molecular mechanisms. These data indicate that early in fetal development *CHD8* mutation may alter the functioning of transcriptional regulators sensitive to direct regulation by CHD8, whereas later in fetal development indirect mechanisms may regulate genes important for synaptic function, perhaps in conjunction with some of the early transcriptional/chromatin regulators.

Finally, we note an enrichment of cancer-associated loci among genes with CHD8-binding sites, regardless of whether these genes are differentially expressed in these NPCs. *CHD8* has been implicated in multiple cancers in several studies (11–14), most recently a pan-cancer deep sequencing study from Lawrence et al. (13), but the mechanistic link between *CHD8* and cancer pathways is unclear. There is strong enrichment among all genes with CHD8 ChIP-determined binding sites for genes in each of the three cancer datasets that we evaluated, indicating that CHD8 has a direct role in their transcriptional regulation. That most such genes in this NPC system did not show altered expression caused by *CHD8* suppression suggests that cell specificity and other cooperating factors may be important in determining CHD8 regulation of particular cancer pathways in specific tumor types. Notably, in a study published during review of this paper, Bernier et al. (39) performed extensive phenotyping of 15 individuals harboring truncating CHD8 mutations and found that both of the subjects that were assessed after the age of 40 years developed tumors, including one subject who was diagnosed with rectum carcinoma at age 42 and died from complications of metastases. Clearly, the impact of CHD8 on the development of primary tumors and metastatic disease warrants further exploration.

The integration of ChIP-sequencing with transcriptome sequencing enabled a search for binding motifs that propose

cofactors with the potential to act in concert with *CHD8*. Motif analysis implicates YY1, a cofactor of CTCF, CTCF itself, and other factors that may act as coactivators or corepressors with CHD8. However, these analyses have limitations, because the cofactors implicated by motif analysis are necessarily restricted by the available database of known motifs, and in many cases they do not distinguish between family members that recognize similar sequences. Also, this study was performed in a single cell type, albeit one that is highly relevant to neural development. Additional studies performed directly on relevant brain tissue and on peripheral cell types would be of interest, and such studies are in progress. For example, studies of differentiating neurons or cells of origin of specific tumor types are likely to reveal additional functions of CHD8, including differences in gene expression that reflect cell type and stage-specific direct regulatory effects and consequent differences in the networks of indirect effects. In particular, investigation of other cell types could elucidate the regulatory function of CHD8 for the many disease-associated genes with CHD8-binding sites that did not show altered expression in NPCs as the result of CHD8 suppression.

In conclusion, these studies identify a strong association between *CHD8* and ASD pathogenesis and also support a role for the gene in cancer formation through a distinct set of genes. The connection uncovered between *CHD8* and a network of diverse ASD-associated genes supports the hope that targeting therapeutic intervention to a limited number of shared pathways of pathogenesis eventually could provide effective treatment for ASD individuals with quite different genetic defects. It also points to the need to identify the direct target(s) of CHD8 that mediates this indirect effect as one or more additional players in the ASD transcriptional network.

Materials and Methods

Cell Culture, Viral Transduction, and Stable Cell Line Generation. Human control NPCs GM8330-8 were kindly provided by S.J.H. They originated from a control patient (not an affected individual) and were derived from iPSC clones through a neuronal differentiation protocol as described in ref. 15. High-efficiency pLKO.1 HIV-based lentiviral vectors carrying six different shRNAs targeting *CHD8* (Dataset S2*A*) as well as against nontargeting controls (Sh against GFP and LacZ) were developed by the RNAi consortium (TRC-Hs1.0, Human) at the Broad Institute (Cambridge, MA). Further details are provided in *SI Materials and Methods*.

Functional Genomic Studies. ChIP was performed as previously described (40) in control NPCs infected with the GFP hairpin. Three independent anti-CHD8 antibodies were used (Novus Biological NB100-60417 and NB100-60418 and Bethyl A301-224A) (see Fig. 1*B* for epitope location). Complexes were precipitated with Dynabeads Protein A beads (Invitrogen), and immunoprecipitated chromatin was eluted in elution buffer de-crosslinked at 65 °C for 8 h (or overnight) and treated with proteinase K (Roche). DNA was purified by extracting with phenol and chloroform and precipitating in ethanol, followed by library preparation for Illumina HiSEq 2000 sequencing. For RNA-seq, libraries were prepared using a customized version of the originally published, strand-specific dUTP method (41, 42). Libraries were generated for all shRNA knockdowns and controls. Libraries were multiplexed, pooled, and sequenced on multiple lanes of an Illumina HiSeq 2000 to a targeted depth generating an average of 40 M paired-end 50-cycle reads for each sample (average final depth ~45 M total reads). See *SI Materials and Methods* for detailed procedures.

Computational Analyses and Statistical Methods. RNA-seq data were aligned to the human genome (GrCH37, Ensembl build 71) using Gsnap (43) version 2012-07-207. Only reads with unique alignments were retained, and only genes that met the threshold for detection in all the samples were included in analyses [more than three reads, as determined by analysis of External RNA Controls Consortium (ERCC) spike-ins as described in ref. 16] (Fig. S2 *B* and *C*), resulting in 15,903 genes. For differential expression analysis, sh1 was excluded because of the low level of knockdown of *CHD8*. Differential expression analysis was carried out using a two-factor model that incorporated batch effects using differential expression sequencing (DESeq) (17) version 1.12.1 (*SI Materials and Methods* and Fig. S8 *B* and *C*). Functional and pathway enrichments were assessed using DAVID (44) and ToppGene (45). Only functional/pathway enrichments meeting a false-discovery rate (FDR) < 5% (DAVID) or a Bonferroni-corrected *P* value < 0.05 (ToppGene) are presented. Disease gene-set enrichments were assessed using Fisher's exact test, and all results, including Benjamini–Hochberg (46) corrected *P* values, are provided in Dataset S8*A*. The complete disease gene sets used are described in *SI Materials and Methods*. ChIP-seq libraries were aligned to GrCH37 (Ensembl build 71) using Burrows–Wheeler Alignment (BWA) version 0.7.5a (47). CHD8-binding peaks were detected separately for each antibody using MACS2 (version 2.0.10.2013.9.13) (48) with a cutoff of a Benjamini–Hochberg corrected *P* value < 0.05. We used only peaks that were detected by at all three antibodies (7,324 peaks; Fig. S5*B*). We used CEAS (49) to obtain genomic distributions of peaks relative to the hg19 refGene track, as shown in Fig. 3, and GREAT (22) to map peaks to genes, allowing peaks up to 10 kb from a gene's TSS in either direction to be mapped to that gene, and to determine enriched GO terms and pathways (Dataset S4). We obtained genome segmentations by chromatin state, 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/) (21). De novo motif analysis was performed using Homer version 4.2 (50) on the peak list for each antibody separately, which obtained the nine replicated motifs listed in Dataset S5. The full peak length was used, and sequences were masked for repetitive sequences.

WGCNA (18) was performed using signed correlation on all 15,903 genes. Module–trait relationships (correlation between the module eigengene and the trait of interest) and gene significance (correlation between gene expression and the trait of interest) were computed for each module and each gene, respectively, using *CHD8* expression level as the trait. All analyses are provided in *SI Materials and Methods*, and all 21 modules are shown in Fig. S3. PPI network analyses were performed for differentially expressed genes using DAPPLE (51), which assesses significant interactions for given genes based on permutation statistics over the manually curated InWeb database (52). To visualize the network we used Cytoscape 3.0.2 (53) and subnetworks generated using the reactomeFI plugin (54). Based on the topological overlap matrix (TOM), 699 genes were involved in the top 0.5% of coexpression interactions. The PPI subnetwork with the largest overlap with CHD8-coexpression hub genes, defined as genes in the top 10% of genes in each of the four *CHD8*-correlated modules by intramodular connectivity, is shown in Fig. S4*B*.

Morpholino, Immunostaining, and Embryo Manipulations. Zebrafish embryos were raised and maintained as previously described (55). Splice-blocking MOs against *chd8* (*chd8*-MO1, 5′- GAGAATGGAATCATAACTTACTTGA-3′, and *chd8*-MO2, 5′- GCAAATGTGCAAGCAAGTAACACCT-3′) were obtained from Gene Tools, LLC. We injected 10 ng of *chd8*-MO1 and *chd8*-MO2 into wild-type zebrafish embryos at the one- to two-cell stage. Suppression of endogenous message was shown by PCR amplification of cDNA reverse transcribed from extracted total mRNA (primers are available upon request). All experiments shown in this study were performed using *chd8*-MO1 and replicated with *chd8*-MO2. Injected embryos were either fixed at 2 dpf for immunostaining or fixed at 4.5 dpf for head-size measurement; the distance across the convex tips of the eye cups was measured and compared with an age-matched control group from the same clutch. Further details on methods in zebrafish, including whole-mount immunostaining and characterization of the *chd8* transcript by RNA-seq, are given in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Dr. Anshul Kundaje of Stanford University and the NIH Roadmap Epigenomics Consortium (nihroadmap.nih.gov/epigenomics/) for providing chromatin state data. This research was supported by the Simons Foundation for Autism Research, the Nancy Lurie Marks Family Foundation, NIH Grants MH095867, MH095088, and GM061354, the March of Dimes, Charles Hood Foundation, the Brain and Behavioral Research Foundation, the Autism Genetic Resource Exchange, Autism Speaks, and Pitt–Hopkins Research Foundation. N.K. is a Distinguished Bromley Professor.

1. Neale BM, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485(7397):242–245.
2. O'Roak BJ, et al. (2012) Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* 338(6114):1619–1622.
3. O'Roak BJ, et al. (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485(7397):246–250.
4. Sanders SJ, et al. (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485(7397):237–241.

5. Talkowski ME, et al. (2011) Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. *Am J Hum Genet* 88(4):469–481.
6. Talkowski ME, et al. (2011) Assessment of 2q23.1 microdeletion syndrome implicates MBD5 as a single causal locus of intellectual disability, epilepsy, and autism spectrum disorder. *Am J Hum Genet* 89(4):551–563.
7. Talkowski ME, et al. (2012) Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell* 149(3):525–537.
8. Flaus A, Martin DM, Barton GJ, Owen-Hughes T (2006) Identification of multiple distinct Snf2 subfamilies with conserved structural motifs. *Nucleic Acids Res* 34(10): 2887–2905.
9. Zahir F, et al. (2007) Novel deletions of 14q11.2 associated with developmental delay, cognitive impairment and similar minor anomalies in three children. *J Med Genet* 44(9):556–561.
10. Iossifov I, et al. (2012) De novo gene disruptions in children on the autistic spectrum. *Neuron* 74(2):285–299.
11. Kim MS, Chung NG, Kang MR, Yoo NJ, Lee SH (2011) Genetic and expressional alterations of CHD genes in gastric and colorectal cancers. *Histopathology* 58(5): 660–668.
12. Tahara T, et al. (2014) Colorectal carcinomas with CpG island methylator phenotype 1 frequently contain mutations in chromatin regulators. *Gastroenterology* 146(2): 530–538 e535.
13. Lawrence MS, et al. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505(7484):495–501.
14. Sawada G, et al. (2013) CHD8 is an independent prognostic indicator that regulates Wnt/β-catenin signaling and the cell cycle in gastric cancer. *Oncol Rep* 30(3): 1137–1142.
15. Sheridan SD, et al. (2011) Epigenetic characterization of the FMR1 gene and aberrant neurodevelopment in human induced pluripotent stem cell models of fragile X syndrome. *PLoS ONE* 6(10):e26203.
16. 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.
17. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.
18. Langfelder P, Horvath S (2008) WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559.
19. Abrahams BS, et al. (2013) SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism* 4(1):36.
20. Xu LM, et al. (2012) AutismKB: An evidence-based knowledgebase of autism genetics. *Nucleic Acids Res* 40(Database issue):D1016–D1022.
21. Bernstein BE, et al. (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 28(10):1045–1048.
22. McLean CY, et al. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 28(5):495–501.
23. Ishihara K, Oshimura M, Nakao M (2006) CTCF-dependent chromatin insulator is linked to epigenetic remodeling. *Mol Cell* 23(5):733–742.
24. Schwalie PC, et al. (2013) Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. *Genome Biol* 14(12): R148.
25. Parikshak NN, et al. (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 155(5):1008–1021.
26. Willsey AJ, et al. (2013) Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* 155(5):997–1007.
27. Futreal PA, et al. (2004) A census of human cancer genes. *Nat Rev Cancer* 4(3): 177–183.
28. Hindorff LA, et al. A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies. Accessed January 2014.
29. Ayalew M, et al. (2012) Convergent functional genomics of schizophrenia: From comprehensive understanding to genetic risk prediction. *Mol Psychiatry* 17(9):887–905.
30. 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.
31. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511(7510): 421–427.
32. Darnell JC, et al. (2011) FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* 146(2):247–261.
33. Krumm N, O'Roak BJ, Shendure J, Eichler EE (2014) A de novo convergence of autism genetics and molecular neuroscience. *Trends Neurosci* 37(2):95–105.
34. Golzio C, et al. (2012) KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature* 485(7398):363–367.
35. Voineagu I, et al. (2011) Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474(7351):380–384.
36. Feng W, Liu HK (2013) Epigenetic regulation of neuronal fate determination: The role of CHD7. *Cell Cycle* 12(24):3707–3708.
37. Potts RC, et al. (2011) CHD5, a brain-specific paralog of Mi2 chromatin remodeling enzymes, regulates expression of neuronal genes. *PLoS ONE* 6(9):e24515.
38. Ronan JL, Wu W, Crabtree GR (2013) From neural development to cognition: Unexpected roles for chromatin. *Nat Rev Genet* 14(5):347–359.
39. Bernier R, et al. (2014) Disruptive CHD8 mutations define a subtype of autism early in development. *Cell* 158(2):263–276.
40. Bernstein BE, et al. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 120(2):169–181.
41. Levin JZ, et al. (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* 7(9):709–715.
42. Parkhomchuk D, et al. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* 37(18):e123.
43. Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26(7):873–881.
44. Huang da W et al. (2009) Extracting biological meaning from large gene lists with DAVID. *Curr Protoc Bioinformatics* Chapter 13: Unit 13-11.
45. 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.
46. Klipper-Aurbach Y, et al. (1995) Mathematical formulae for the prediction of the residual beta cell function during the first two years of disease in children and adolescents with insulin-dependent diabetes mellitus. *Med Hypotheses* 45(5):486–490.
47. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
48. Zhang Y, et al. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9(9): R137.
49. Shin H, Liu T, Manrai AK, Liu XS (2009) CEAS: cis-regulatory element annotation system. *Bioinformatics* 25(19):2605–2606.
50. 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.
51. Rossin EJ, et al.; International Inflammatory Bowel Disease Genetics Constortium (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* 7(1):e1001273.
52. Lage K, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 25(3):309–316.
53. Shannon P, et al. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* 13(11):2498–2504.
54. Wu G, Stein L (2012) A network module-based method for identifying cancer prognostic signatures. *Genome Biol* 13(12):R112.
55. Westerfield M (1995) *The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish (Danio rerio)* (Univ of Oregon Press, Eugene, OR), 3rd Ed.

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 co-expression 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.

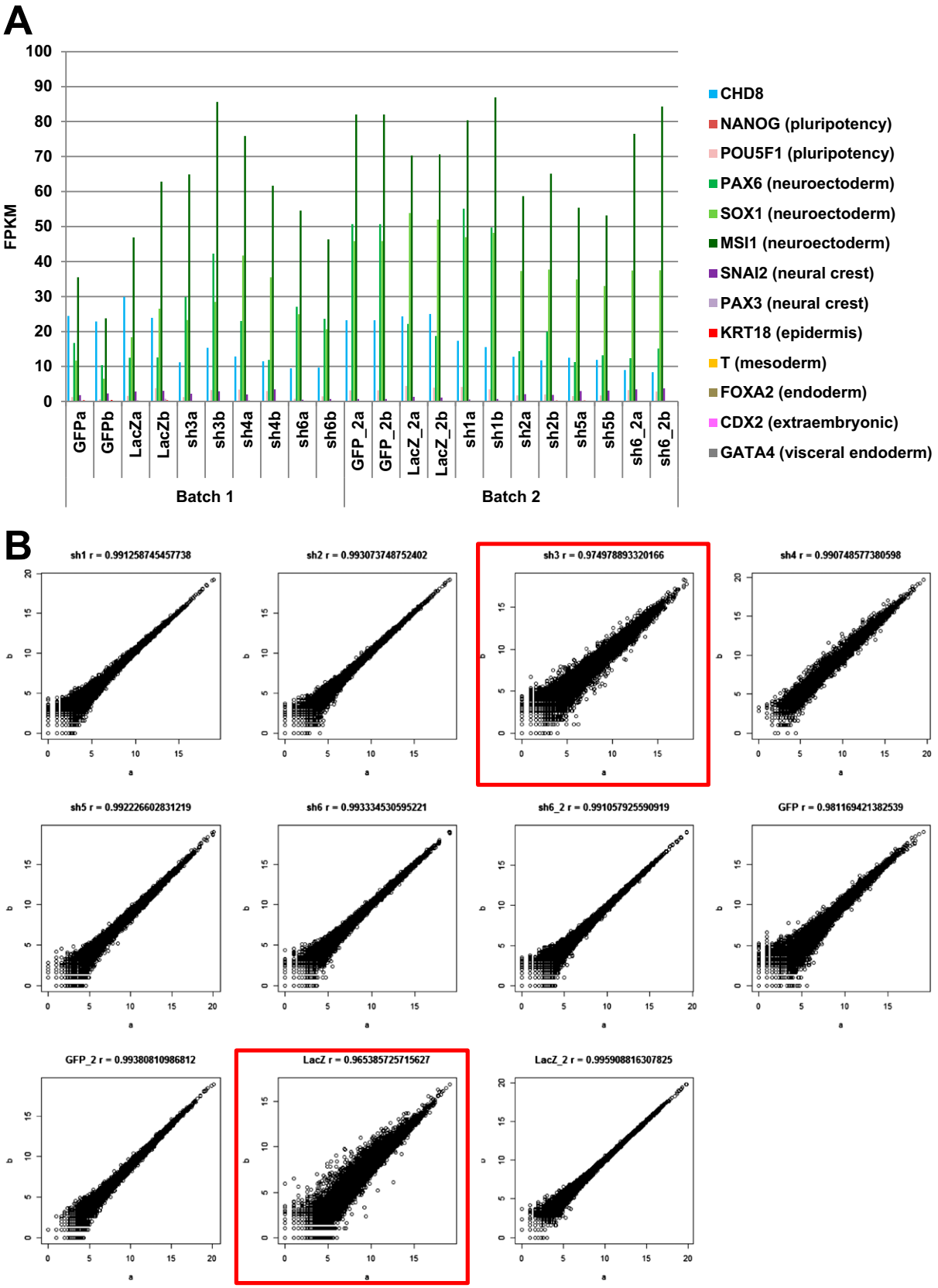


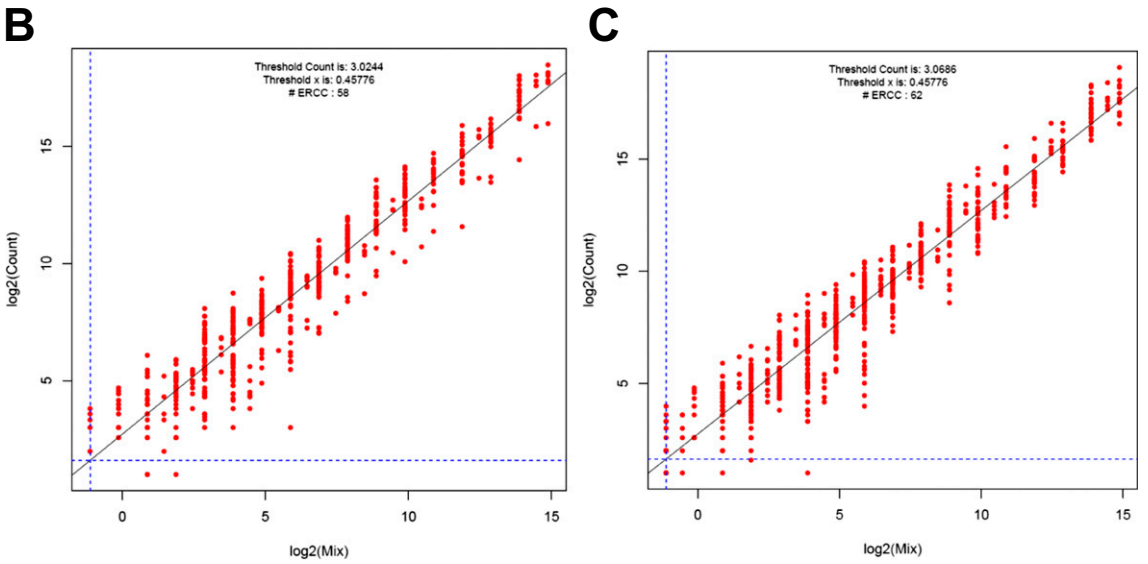
Fig. S1. Data quality of RNA-seq libraries. (*A*) Gene expression in knockdown and control NPCs for marker genes for stages of differentiation. Shown are FPKMs for 11 marker genes, along with the *CHD8* gene (blue) for comparison. All three marker genes for neuroectoderm (*MSI1*, *SOX1*, and *PAX6*), 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 (batch1)	sh4b (batch1)	sh6a (batch1)	sh6b (batch1)	GFPa (batch1)	GFPb (batch1)	LacZa (batch1)	sh6_2a (batch2)	sh6_2b (batch2)	sh2a (batch2)	sh2b (batch2)	sh5a (batch2)	sh5b (batch2)	GFP_2a (batch2)	GFP_2b (batch2)	LacZ_2a (batch2)	LacZ_2b (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	



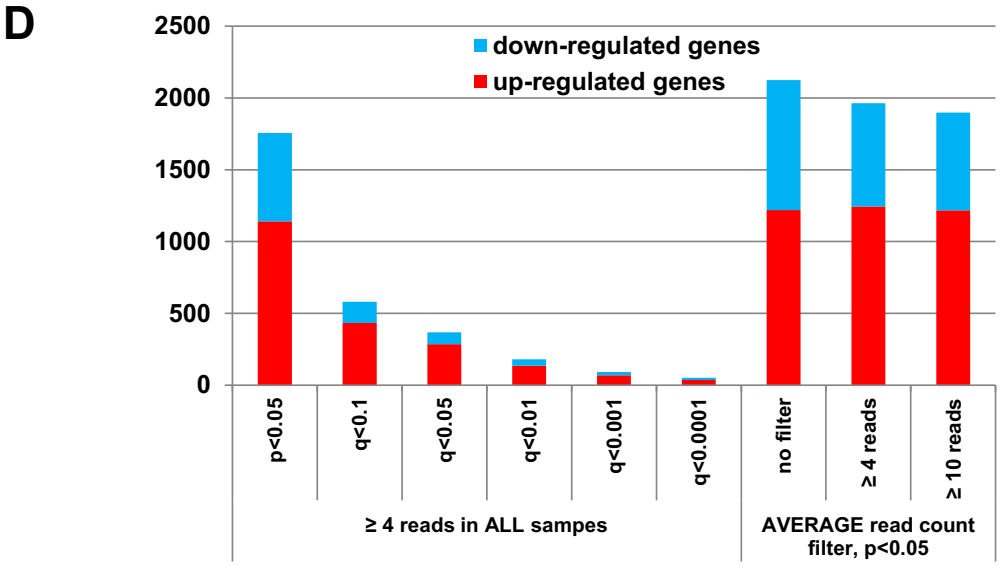


Fig. S2. RNA-seq sample correlations, spike-in analysis, and differentially expressed genes after removal of outlier samples. (*A*) Pearson correlations between all RNA-seq samples, after batch correction. Deeper green represents higher correlation. (*B* and *C*) ERCC spike-in analysis for RNA-seq samples. 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.

A

B

	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

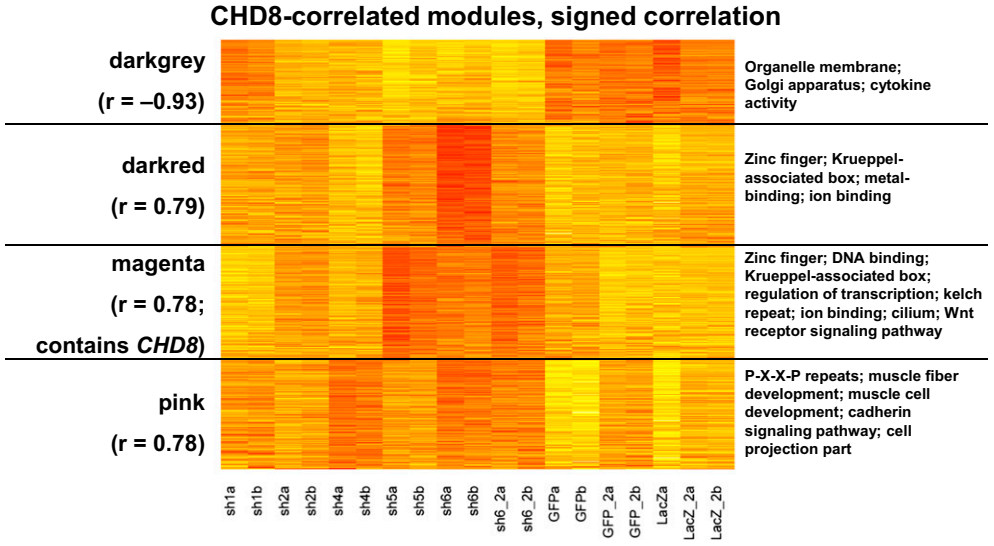
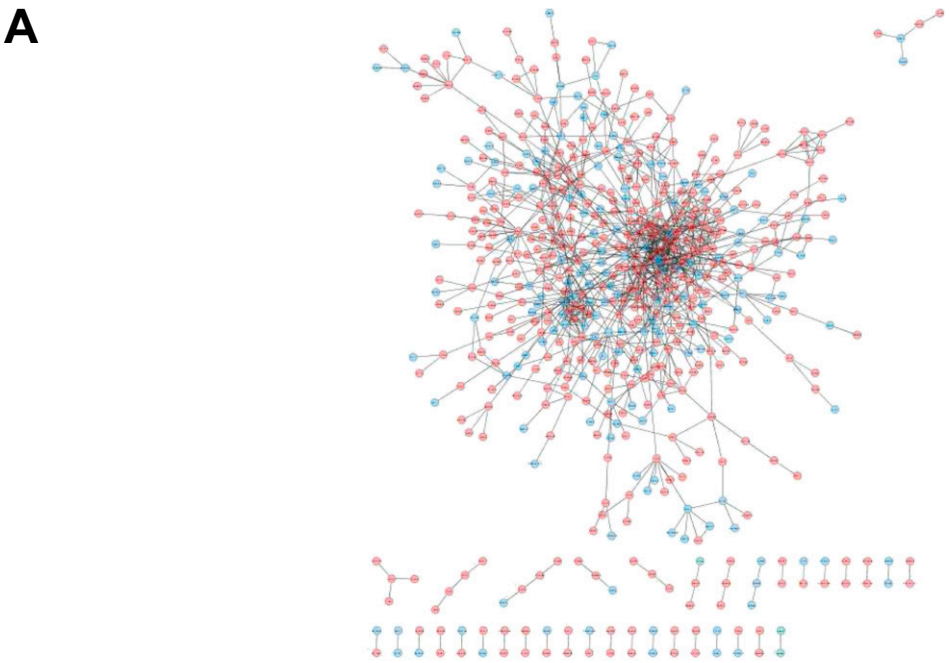
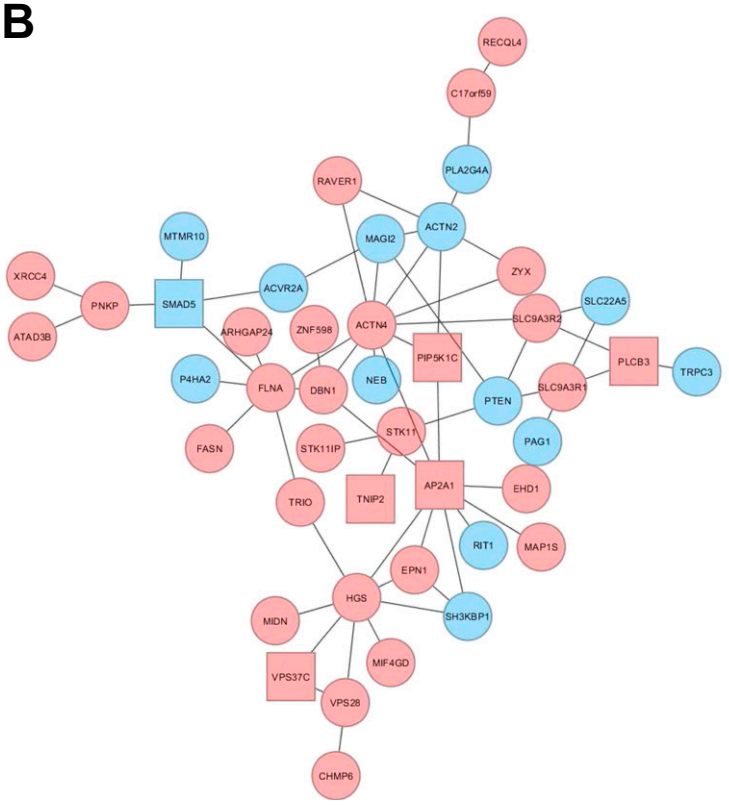


Fig. S3. RNA-seq coexpression modules. (*A* and *B*) Twenty-one modules of coexpressed genes using signed correlation. (*A*) Module–trait relationships. In this analysis, *CHD8* expression in each of the samples was used as the trait. The number shown for each module corresponds to the correlation between 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.

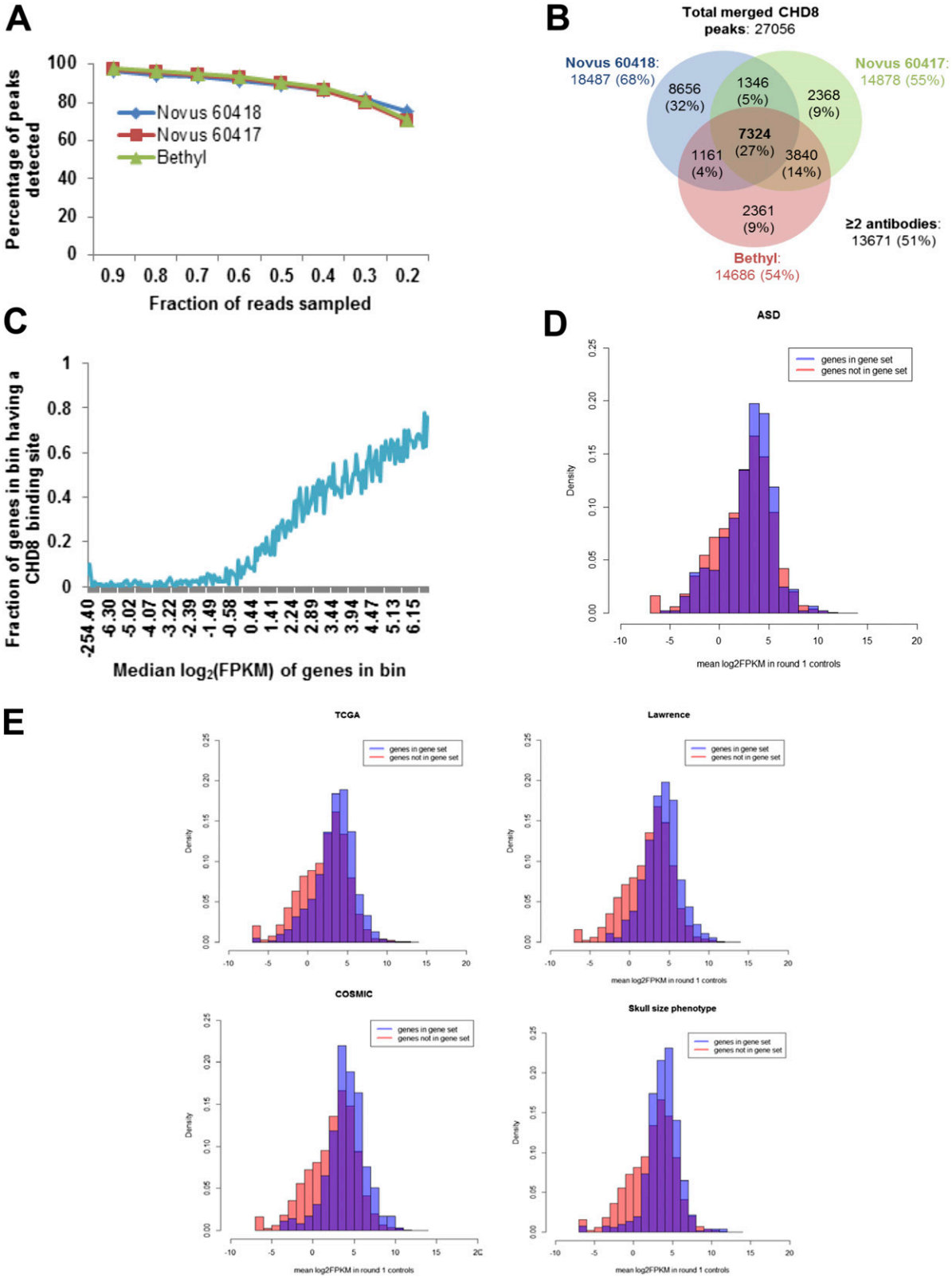


Fig. S5. ChIP-seq data for CHD8 binding in control NPCs. (*A*) Saturation plot for ChIP-seq data. For each antibody, the saturation curve shows the fraction of total peaks for that antibody that is detected (*y* axis) if the ChIP library is down-sampled by the fraction on the *x* axis. Saturation curves 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.

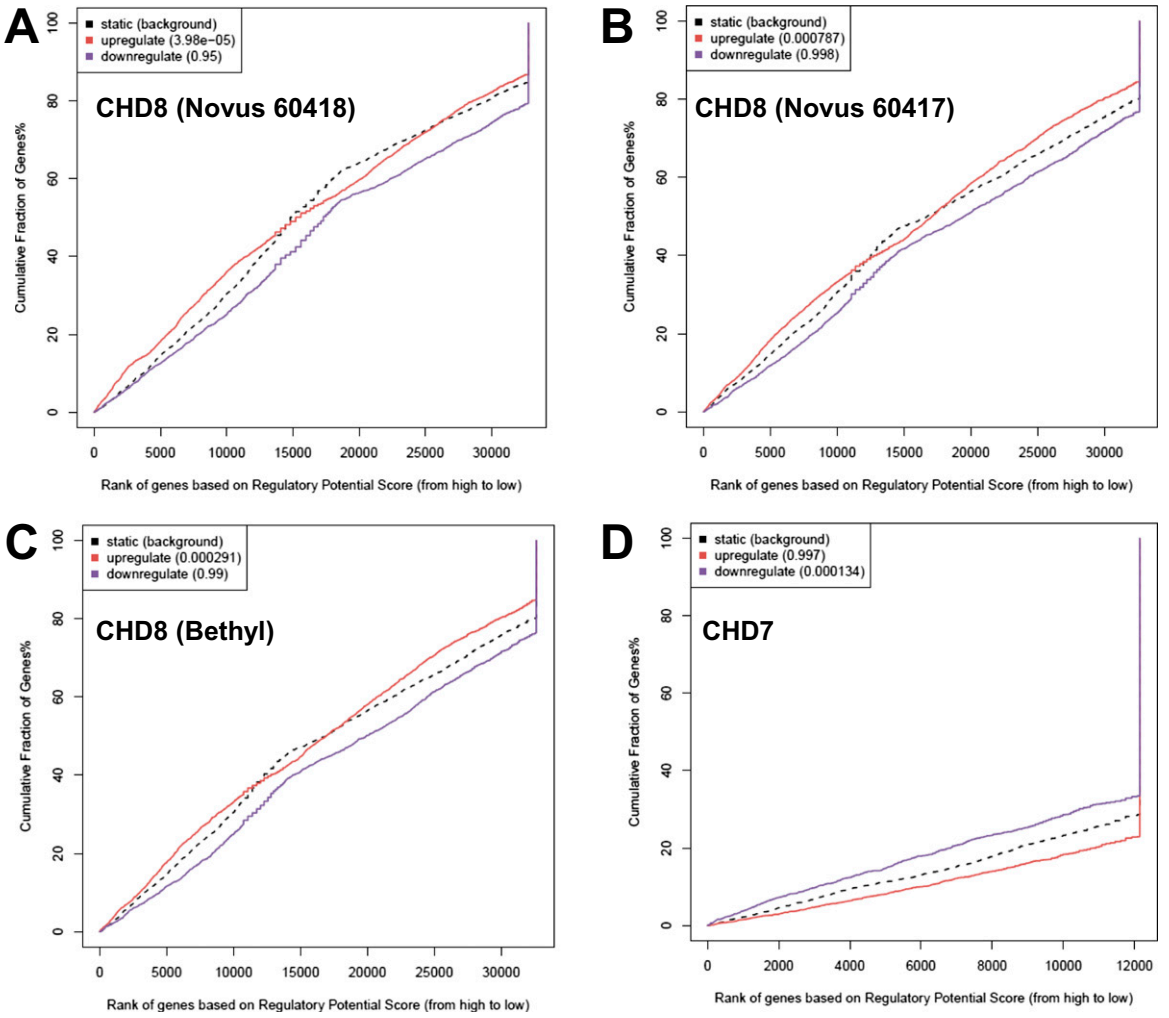


Fig. S6. Regulatory potential for CHD8- and CHD7-binding sites. Regulatory potential for the set of peaks identified by each of the three CHD8 antibodies (*A*–*C*) and for the CHD7 antibody (*D*) generated using BETA (13). Genes up-regulated by CHD8 knockdown are shown in red; down-regulated genes are 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.

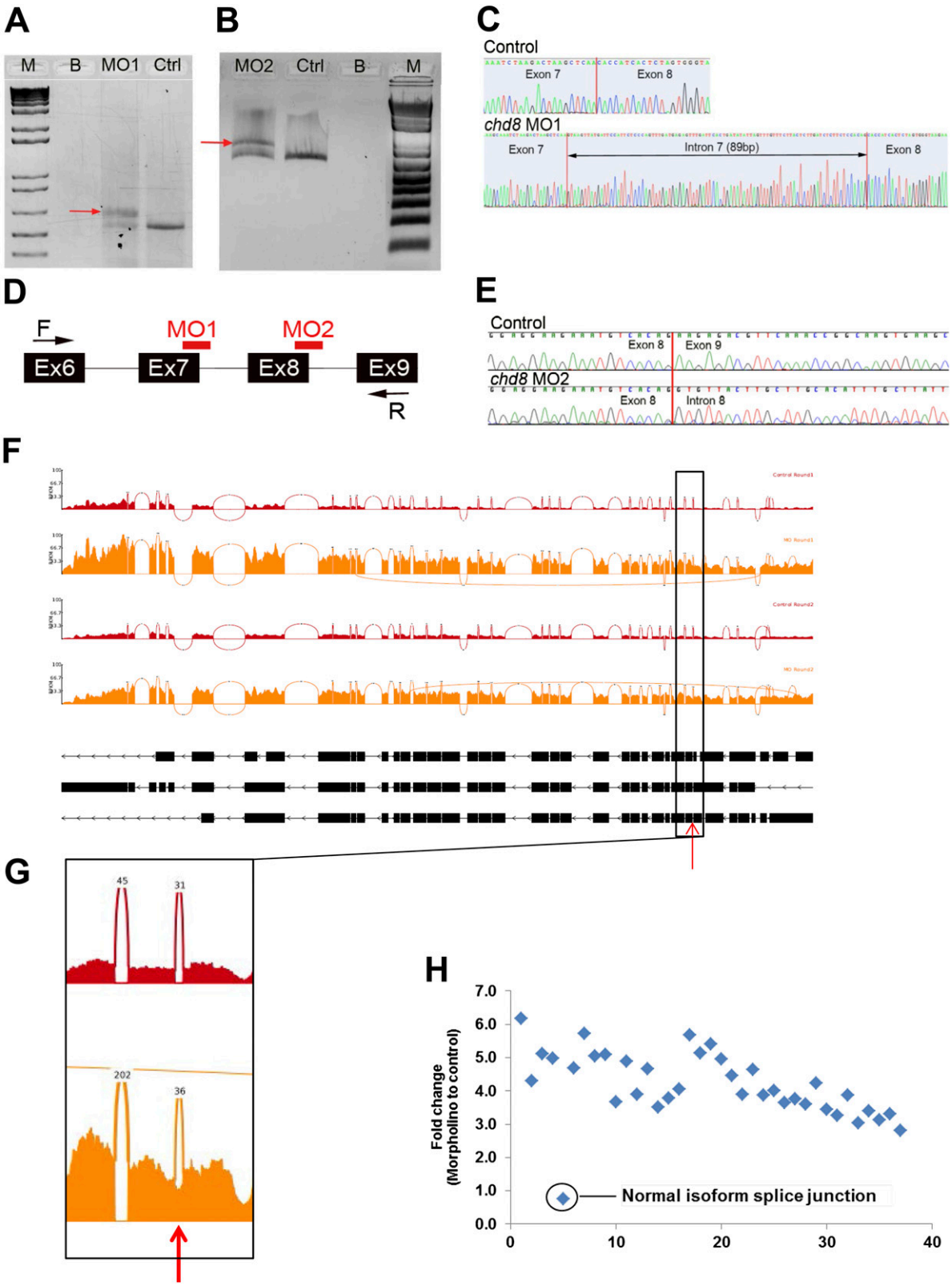


Fig. S7. *chd8* MOs efficiently disrupt the splicing of its zebrafish endogenous message. (*A* and *B*) Injection of *chd8* splice-blocking morpholinos *chd8*-MO1 (*A*) and *chd8*-MO2 (*B*) (10 ng each) results in abnormal splicing as shown by PCR amplification of cDNA reverse transcribed from extract 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.

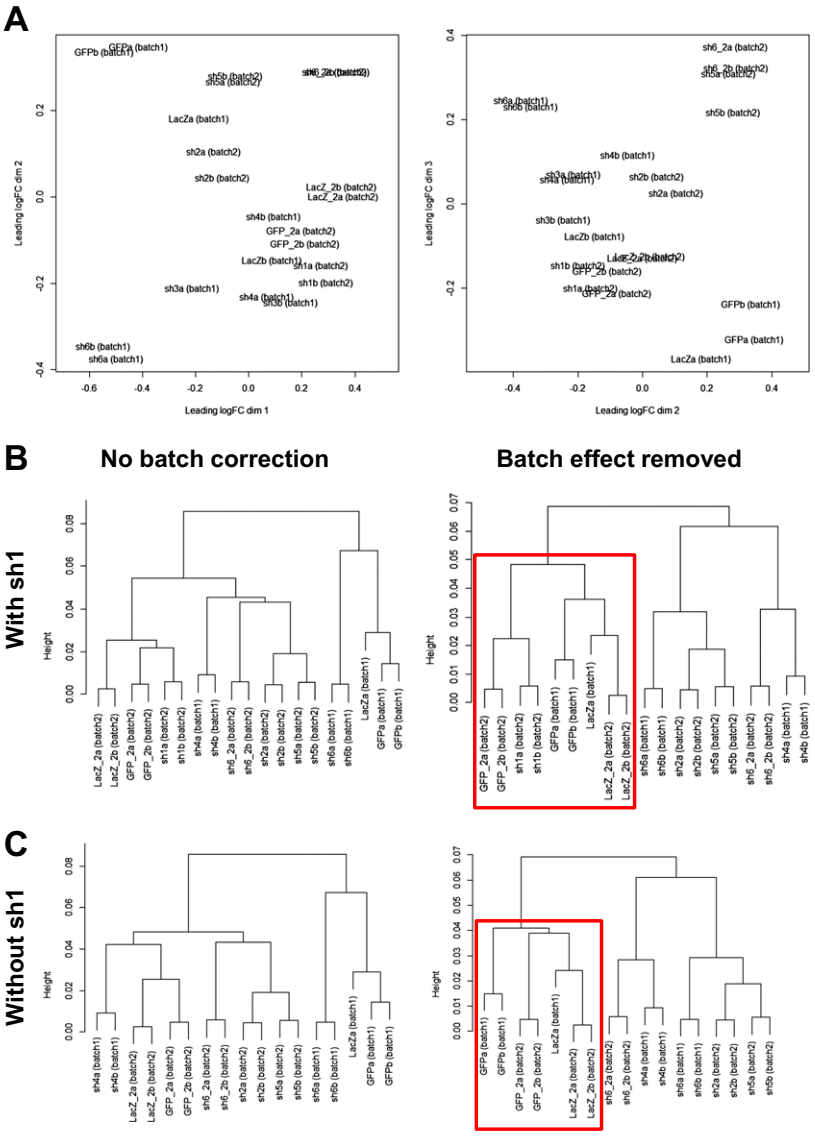


Fig. S8. Batch effect in RNA-seq libraries. (*A*) MDS plots for RNA-seq samples. (*Left*) Dimensions 1 and 2. (*Right*) Dimensions 2 and 3. Plots were generated from TMM-normalized log cpms, using the plotMDS function in the edgeR package. (*B* and *C*) RNA-seq sample clustering after removal of LacZb 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)