Supporting Information

Goff et al. 10.1073/pnas.1411263112

SI Materials and Methods

Mice. LncRNA KO mice were generated in collaboration with Regeneron Pharmaceuticals by replacing the selected lncRNA gene with a *lacZ* reporter cassette as previously described (1). Briefly, targeting vectors were constructed using the VelociGene technology (2), and targeted mouse ES cell clones then were introduced into an eight-cell-stage mouse embryo using the VelociMouse method (3). To remove the *loxP*-flanked neomycin resistance gene included in the lncRNA targeting constructs for the initial selection, each of the original C57BL/6J backcrossed mutant strains (1) was crossed once with the cre-recombinase strain B6.C-Tg(CMV-cre)1Cgn/J, and subsequently backcrossed once with C57BL/6J. For each strain, mice free of both the neomycin-resistance and cre-recombinase genes were selected for colony expansion. The strains (N3 for all mutant strains except *Kantr*, which is N2.5) were maintained by heterozygous breeding, and mutant mice were identified by genotyping for loss of the lncRNA allele and gain of the *lacZ* cassette (Transnetyx, Inc.). Heterozygous mice at E14.5 and at 2–3 mo of age (adult) were used to determine the brain expression pattern of each lncRNA (knocked-in *lacZ* reporter gene). E14.5 and adult KO and WT littermate mice were used to determine the effects of lncRNA deletion on global gene expression (RNA-seq). Mice were killed either by CO_2 inhalation followed by cervical dislocation or by transcardial perfusion-fixation performed under Avertin anesthesia. Mice were housed under controlled pathogen-free conditions at Harvard University's Biological Research Infrastructure, and all procedures were carried out in accordance with National Institutes of Health guidelines for the care and use of laboratory animals and were approved by the Harvard University Committee on the Use of Animals in Research and Teaching.

Tissue Collection and Processing. For β-gal expression, E14.5 whole brains were harvested from embryos fixed by immersion in 4% (vol/vol) PFA at 4 °C overnight. Tissue samples were collected before fixation for genotyping (Transnetyx, Inc.). Adult whole brains were harvested from heterozygous mice after fixation by transcardial perfusion with 4% (vol/vol) PFA and were postfixed in 4% (vol/vol) PFA at 4 °C for 12 h. Coronal sections (40-μm thick) were cut on a vibrating microtome and mounted on VistaVision HistoBond SuperFrost Plus slides (VWR). For RNA isolation, E14.5 whole brains were harvested and immediately homogenized in TRIzol (Life Technologies, 1 mL per brain). Whole brains from adult KO and WT mice were harvested immediately after mice were killed, snap frozen in liquid nitrogen, and stored at −80 °C. Frozen whole brains were pulverized in liquid nitrogen and homogenized in TRIzol (5 mL per brain). One-milliliter aliquots of the TRIzol lysates were used for RNA isolation.

β-Gal Staining and Immunostaining. Brain-wide expression of the *lacZ* reporter gene was assessed in all mutant strains by histochemical detection of β-gal (X-gal staining) ($n = 2$). Staining was performed on coronal brain sections, whole brains, or whole embryos as previously described (1). Postfixed stained whole embryos were stored in 70% (vol/vol) ethanol before brain dissection and sectioning. Sequential sections, obtained at every 80 μm for E14.5 brains and at every 240 μm for adult brains, were imaged at 5× and 10× magnification using a Zeiss Axio Scan.Z1, a Nikon 90i microscope equipped with a Retiga Exi camera (QIMAGING), or a Zeiss LSM700 confocal microscope. Immunohistochemistry for β-gal and immunostaining for the interneuron marker TH was performed using standard methods. Primary antibodies and dilutions were chicken anti–β-gal (1:500; CGAL-45; ICL) and rabbit anti-TH (1:5,000; AB152; Chemicon). Appropriate secondary antibodies were from the Molecular Probes Alexa series. Images were acquired and processed with the Zen (Zeiss) or Volocity analysis software v4.0.1 (Improvision).

mRNA-Seq Library Preparation and Sequencing. RNA-seq was performed for 2–4 KO adult and 2–4 KO E14.5 embryos (Dataset S3) of each mutant strain, obtained from two or more litters per strain, and a total of 14 WT adult and 14 WT embryos, balanced across the different strains and litters. Whole-brain total RNA isolation and mRNA-seq library construction (TruSeq RNA Sample Preparation Kit v2; Illumina) were performed as previously described (1). The libraries were prepared using 500 ng of total RNA as input, with the exception of the E14.5 Peril libraries (250 ng), and a 10-cycle PCR enrichment to minimize PCR artifacts. KO and WT samples from different strains and age groups were processed within each library preparation to dilute any operator or batch biases. The indexed libraries were sequenced in pools of six, each pool including KO and WT samples from different strains and ages, on the Illumina HiSeq 2500 platform using the rapid-full flow cell, paired-end, 50-bp read-length sequencing protocol (NWL Bauer Core, Harvard University FAS Center for Systems Biology).

Read Alignment, Quantitation, and RNA-Seq Analysis. Sequencing reads were aligned to mm10 using Tophat2 (4) with the following additional parameters: –no-coverage-search –max-multihits 10 -p 8. No reference transcriptome was used during alignment, but Tophat2 was not prevented from identifying novel splice junctions. Each aligned sample was quantified against a modified version of the GENCODE (version M2) transcriptome as a reference (5). Each binary alignment map (bam) was then quantified using Cuffquant (6) against the reference transcriptome with the following additional parameters: –no-update-check -p 8. Pairwise differential analysis was performed on each combination of KO and WT conditions using Cuffdiff v2.2.1 (6) with the following additional parameters: -p 8.

To standardize the analysis of each differential comparison, we created a report template that integrated components from several R/Bioconductor (7) packages including cummeRbund (8), gene set analysis (GSA) (9), clusterProfiler (10), and limma (11). The analysis template is made available as Dataset S7, and individual KO versus WT reports are collected in Dataset S6. Gene set analysis (GSA) was performed on the test-stat-rank–ordered list of genes for each comparison using the GSA R/Bioconductor package (9) and compared against gene sets derived from the curated collections at MSigDB (12) including Reactome (c2.cp.reactome.v4.0), Biocarta (c2.cp.biocarta.v4.0), Kyoto Encyclopedia of Genes and Genomes (KEGG) (c2.cp.kegg.v4.0), Transcription Factor Targets (c3.tft.v4.0), Oncogenic signatures (c6.all.v4.0), and Immunologic signatures (c7.all.v4.0). GO, Reactome, and KEGG enrichment analysis was additionally performed using the clusterProfiler R/Bioconductor package on the lists of significantly differentially expressed genes as detailed in Dataset S7.

Digital Genotyping. Genotypes were confirmed after sequencing by isoform-level expression plots for each lncRNA and the *lacZ* reporter in each replicate. Gene tracks were constructed from indexed bam files after alignment of sequencing reads with

Tophat2 to mm10. These tracks validated that any isoform-level expression of an lncRNA in a KO was caused by either a repetitive region-based mapping error or incomplete targeting of the gene body (Dataset S6).

Differential Analysis. Differential analysis was performed for each strain at both embryonic and adult time points using Cuffdiff2 on two to four replicates from each KO strain against a set of 14 WT whole brains (Dataset S7). Each replicate WT bam file represents an individual mouse. An all-sample differential analysis was also performed in which every strain was compared with WT for the purpose of global analysis.

***Cis* Region Analysis.** The number of genes differentially regulated within a 4-Mb window of each targeted lncRNA was gathered and compared with the number of differentially regulated genes in 10,000 randomly selected size-matched windows across the same dataset to generate a bootstrapped *P* value of the likelihood of finding that number of differentially expressed genes by chance. A P value < 0.05 was used as the significance threshold.

1. Sauvageau M, et al. (2013) Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife* 2:e01749.
2. Valenzuela DM, et al. (2003) High-throughput engineering of the mouse genome coupled with high-resolution expression analysis. *Nat Biotechnol* 21(6):652–659.
3. Poueymirou WT, et al. (2007) F0 generation mice fully derived from gene-targeted embryonic stem cells allowing immediate phenotypic analyses. *Nat Biotechnol* 25: 91–99.
4. Kim D, et al. (2013) TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14(4):R36.
5. Harrow J, et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* 22(9):1760–1774.
6. Trapnell C, et al. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 31(1):46–53.
7. Gentleman RC, et al. (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol* 5(10):R80.
8. Trapnell C, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7(3):562–578.
9. Efron B, Tibshirani R (2010) GSA: Gene set analysis. R package version 1.03. Available at CRAN.R-project.org/package=GSA. Accessed February 1, 2014.
10. Yu G, Wang L-G, Han Y, He Q-Y (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16(5):284–287.
11. Smyth G (2005) Limma: linear models for microarray data in *Bioinformatics and Computational Biology Solutions* (Springer, New York).
12. Subramanian A, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102(43):15545–15550.

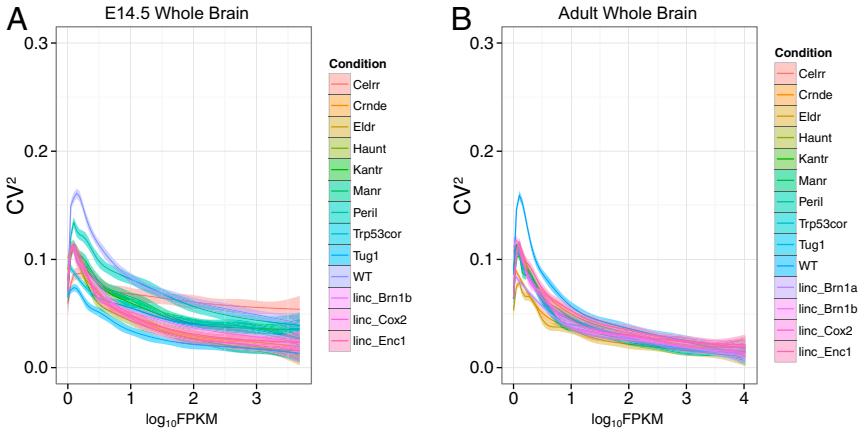

Fig. S1. RNA-Seq cross-replicate variability. A smoothed spline is fit over the squared coefficient of variation (CV^2) of expression estimates (FPKM) across the range of expression values (FPKM) for each strain condition at E14.5 (*A*) and adult (*B*) stages. Low CV^2 values indicate a high degree of agreement between individual replicates for a given condition. The error ribbon shows the 95% confidence interval for the CV^2.

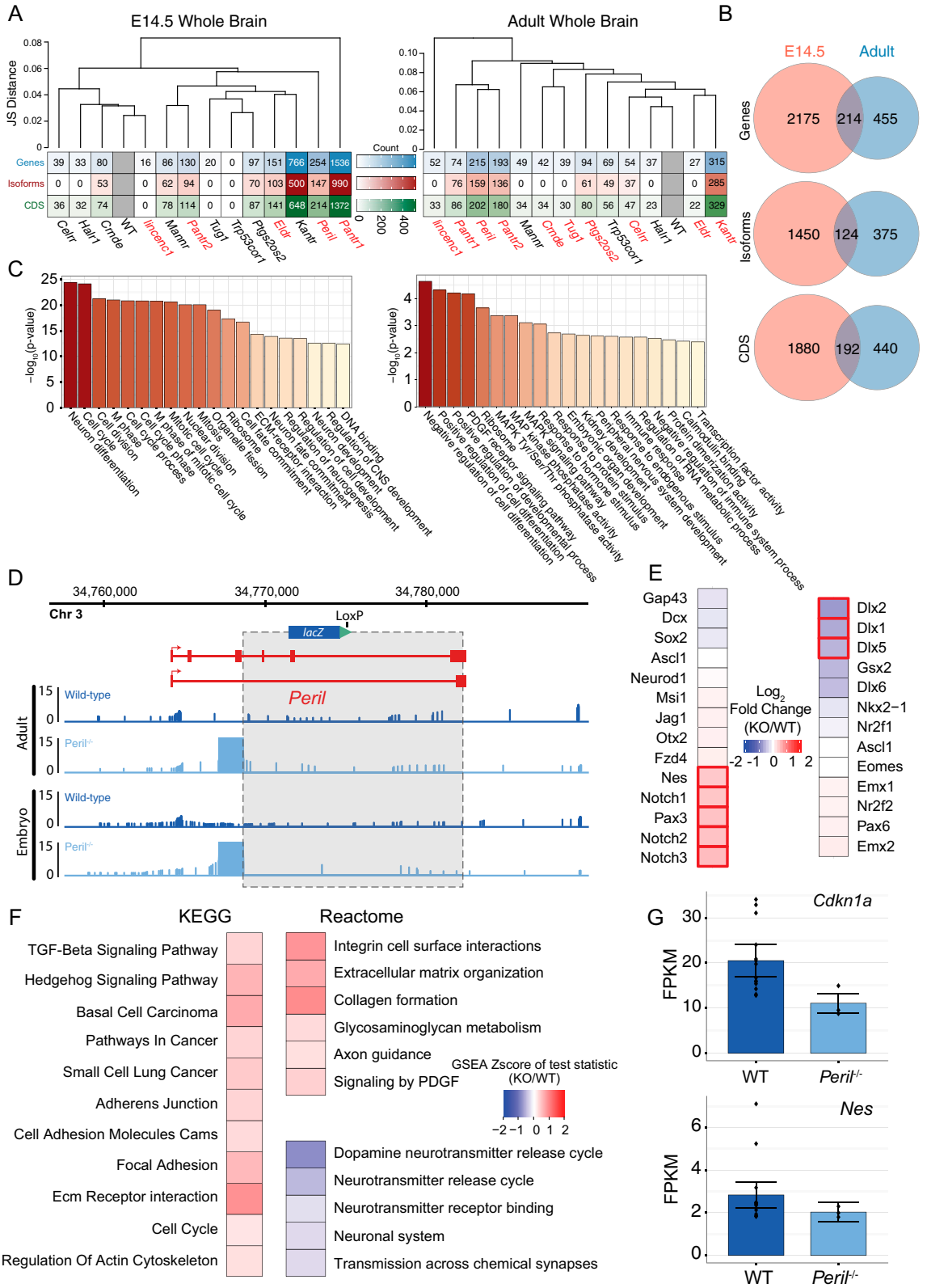

Fig. S2. RNA-sequencing highlights transcriptional changes in developing and adult lncRNA mutant brains. (*A*) Differentially expressed genes, isoforms, and CDS between WT and KO whole brains at E14.5 or adult time points. The dendrogram represents hierarchical clustering of conditions by Jensen–Shannon distance using the normalized expression profiles of the universe of differentially expressed genes across all pairwise comparisons. WT is included to provide a benchmark condition against which global transcriptional changes in the KO strains can be evaluated. (*B*) Total number of differentially expressed genes, isoforms, or CDS at either E14.5 or adult and the overlap between these two sets. (*C*) GO analysis of the universe of significantly differentially regulated genes in any strain at any time point. (*D*) *Peril* locus, transcript structure, and representative read density profiles from WT and KO embryonic and adult brains. (*E*) Progenitor marker enrichment or depletion by *Peril*$^{-/-}$ embryonic RNA-seq. Red outline denotes significance. (*F*) Select GSA results from *Peril* embryonic RNA-seq analysis. (*G*) Expression plots for *Cdkn1a* or *Nestin* in WT and *Peril*$^{-/-}$ adult brain RNA-seq.

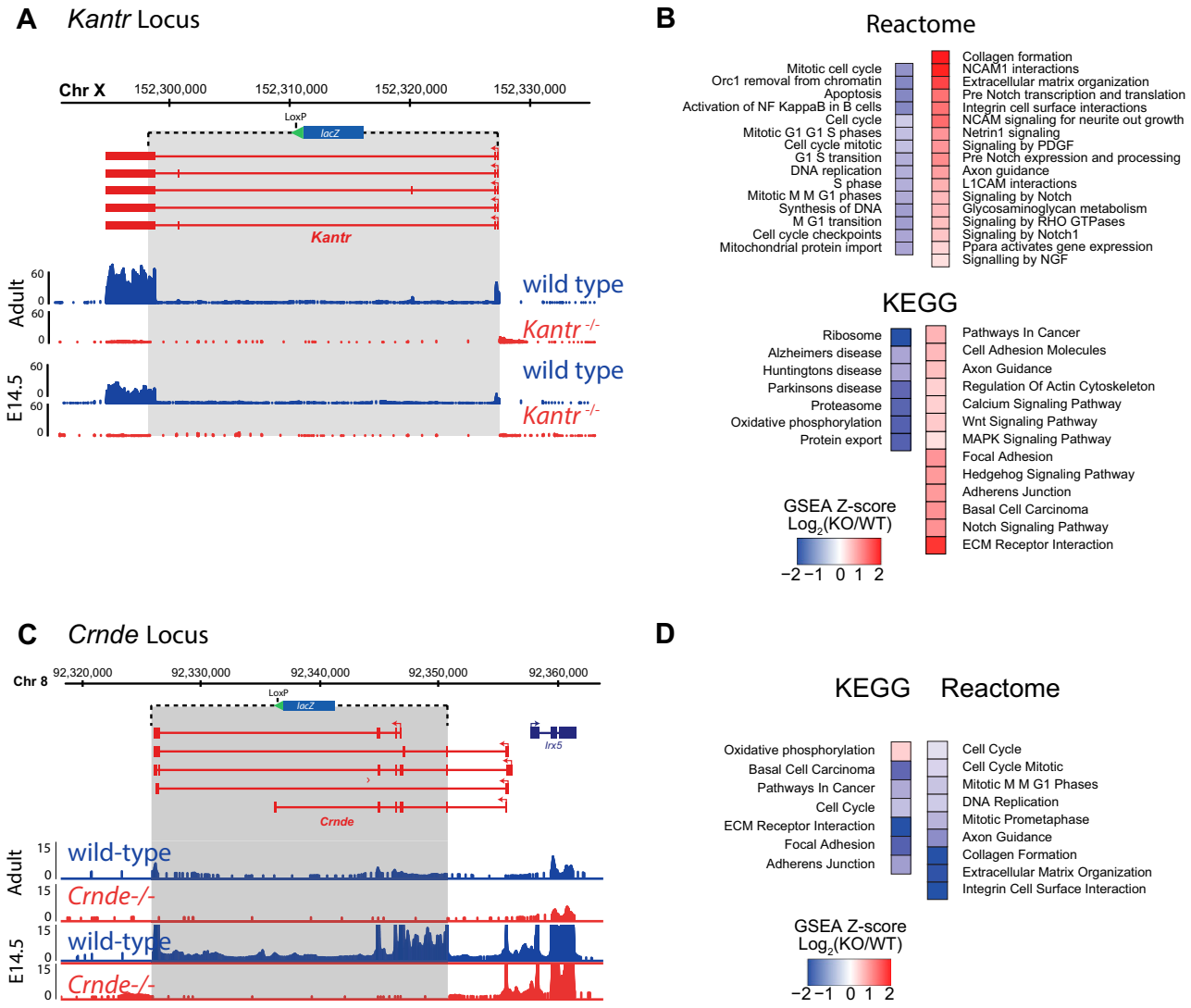

Fig. S3. RNA-seq strain summaries for two lncRNA loci. (*A*) *Kantr* locus, transcript structure, and representative read density profiles from WT and KO embryonic or adult bams. The KO region is highlighted in blue. (*B*) Selected *Kantr* embryonic Reactome or KEGG GSA results. (*C*) *Crnde* locus, transcript structure, and representative read density profiles from WT and KO embryonic or adult bams. The KO region is highlighted in blue. (*D*) Selected Reactome or KEGG GSA results from the *Crnde* embryonic differential analysis.

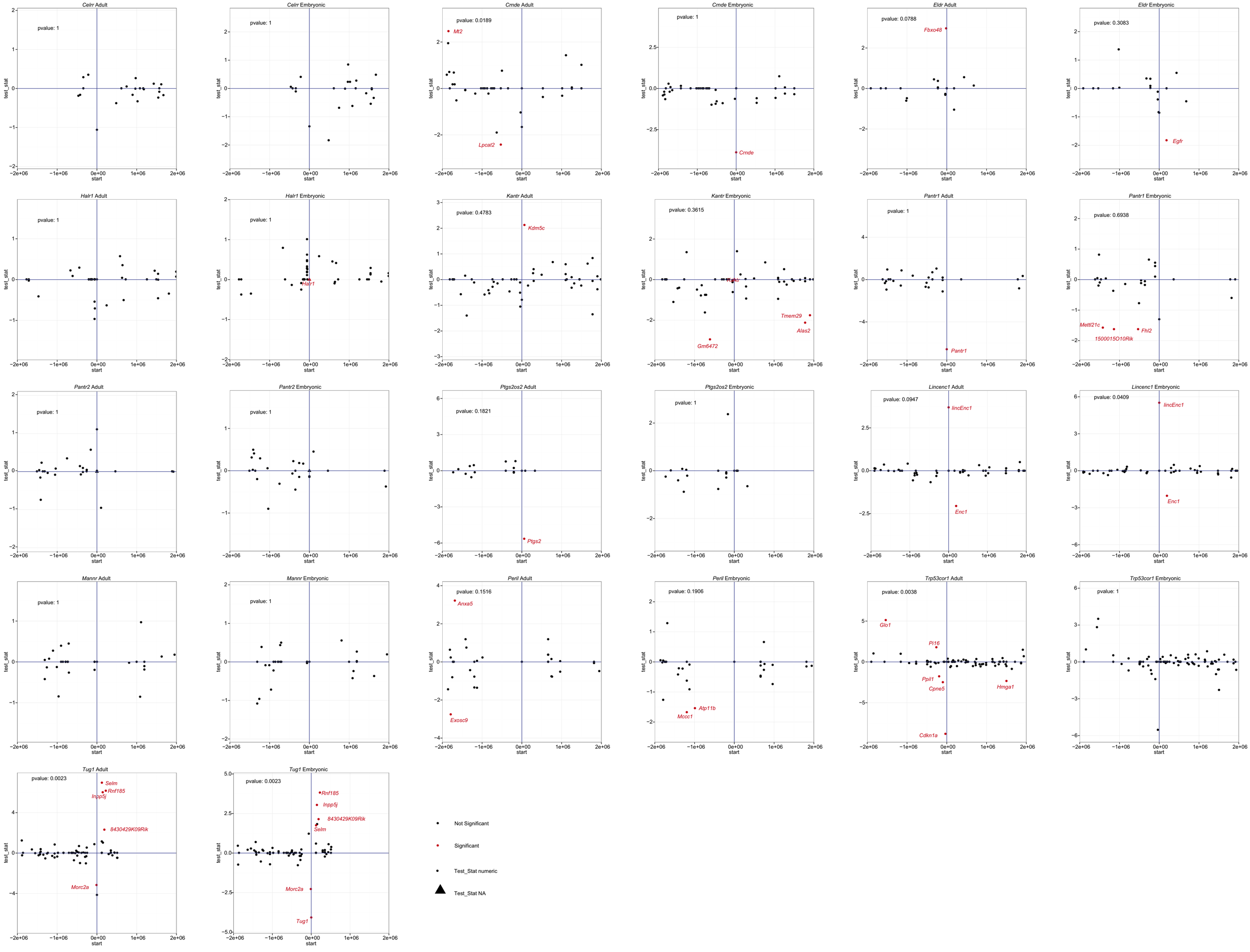

Fig. S4. *Cis* region plots for each lncRNA at each time point. The *x* axis depicts genomic distance (in base pairs) of the neighbor gene TSS from the lncRNA TSS. The *y* axis depicts test statistic (differential analysis by Cuffdiff2). Significantly differentially regulated genes are labeled in red. *P* value for *cis*-regional effect (bootstrap $n = 10{,}000$) (see *Materials and Methods*).

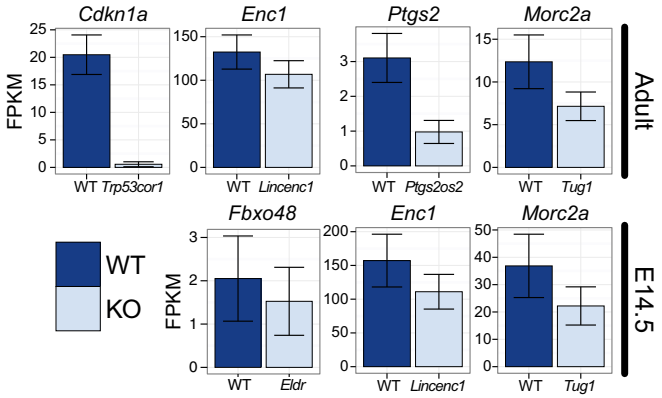

Fig. S5. Expression of neighboring protein-coding genes. Expression bar plots of the significantly differentially regulated closest protein-coding neighbors in their respective WT and lincRNA KO conditions.

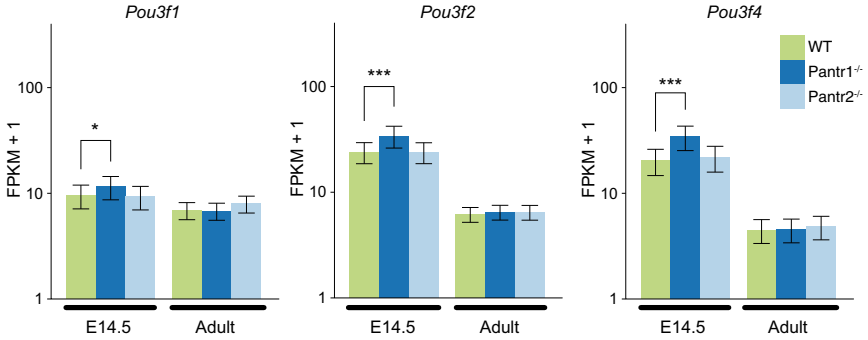

Fig. S6. *Pou3f1*, *Pou3f2*, and *Pou3f4* expression. RNA-seq expression estimates (average from triplicates) for *Pou3f1*, *Pou3f2*, and *Pou3f4* in E14.5 and adult brains collected from WT, *Pantr1*$^{-/-}$, and *Pantr2*$^{-/-}$ mutant mice. *$P < 0.05$, ***$P < 0.001$, Student's *t* test.

Table S1. lncRNA genomic and targeted deletion coordinates (mm10)

MGI symbol	Aliases	Genomic coordinates, mm10	Strand	Targeted deletion coordinates, mm10	RNA-seq expression, E14.5		LacZ staining, E14.5 sections	RNA-seq expression, adult		LacZ staining, adult sections
					lncRNA, FPKM	LacZ, FPKM		lncRNA, FPKM	LacZ, FPKM	
Pantr1	*linc-Brn1a, Pou3f3os, 2610017I09Rik*	chr1:42648200–42694825	–	chr1:42648176–42694815	159.09	43.59	+	61.29	15.52	+
Pantr2	*linc-Brn1b, 2610207O16Rik*	chr1:42707061–42713454	–	chr1:42707143–42713698	2.35	9.26	+	1.24	6.02	+
Ptgs2os2	*linc-Cox2, Gm26687*	chr1:150159043–150164948	–	chr1:150159043–150164948	0.01	0.76	ND	0.05	6.62	+
Eldr	*Fabl, 2810442I21Rik*	chr11:16934709–16951282	–	chr11:16934419–16951083	4.37	2.74	+	0.19	2.67	+
lincenc1	-	chr13:97455803–97482628	–	chr13:97455710–97482249	0.10	29.06	+	0.03	14.55	+
Mannr	*AK147070*	chr3:29891017–29924191	+	chr3:29891188–29923147	0.02	0.88	ND	0.03	0.24	ND
Halr1	*Haunt, linc-Hoxa1, Gm15055*	chr6:52102947–52113684	+	chr6:52106776–52115377	1.75	4.37	ND	0.03	0.41	ND
Celrr	*Celr, B230209K01Rik*	chr1:121087405–121120975	+	chr1:121087772–121137464	0.39	4.06	ND	1.17	1.52	+
Crnde	*linc-Irx5, 4933436C20Rik*	chr8:92326031–92356120	–	chr8:92325920–92350749	12.65	11.69	ND	1.80	1.81	+
Kantr	*Spasm, 2900056M20Rik*	chrX:152294824–152327493	–	chrX:152298544–152327475	7.52	0.64	ND	14.11	2.28	+
Trp53cor1	*linc-p21, Gm16197*	chr17:29057474–29079126	–	chr17:29057474–29079078	0.11	0.31	ND	0.19	0.07	ND
Peril	*Peril, linc-Sox2*	chr3:34764169–34782346	+	chr3:34767849–34782292	0.02	24.57	+	0.01	10.35	+
Tug1	-	chr11:3639785–3648814	–	chr11:3639794–3648758	43.79	4.90	ND	23.86	6.67	+

FPKM, fragments per kilobase of exon per million fragments mapped; MGI, Mouse Genome Informatics Database; ND, no data.

Dataset S1. *LacZ* samples mastersheet

Dataset S1

Dataset S2. Collection of adult brain *lacZ* images for each mutant strain

Dataset S2

Available at rinnlab.rc.fas.harvard.edu/BrainMap/Goff_Suppl_File_S2.pdf.

Dataset S3. RNA-seq samples mastersheet

Dataset S3

Dataset S4. RNA-seq alignment statistics

Dataset S4

Dataset S5. Reference genome file

Dataset S5

Dataset S6. Collection of RNA-seq summary reports for each mutant strain at each time point

Dataset S6

Reports include quality control, digital genotyping, locus visualization, differential gene and isoform analysis, GSA, GO, and *cis*-region analysis.

Dataset S7. R Markdown template file of analysis, methods, and algorithms for individual strain KO vs. WT RNA-sequencing differential gene-expression reports

Dataset S7