Supporting Information

Portales-Casamar et al. 10.1073/pnas.1009158107

SI Materials and Methods

The general methods used have been described previously (1). Below are only reported modifications or additions specific to this work.

Regulatory Resolution Score Characterization. ***Identification of genomic sequence boundaries for regulatory resolution scoring.*** For regulatory resolution scoring of each gene in the Ensembl human genome database (version 46) (2), we define the transcription start site (TSS) as the start position of the 5′-most exon annotated for the gene. We then determine the boundaries of the region to be analyzed relative to the TSS as follows. In most cases, the upstream boundary is defined as the start/end position of the upstream gene (depending on the upstream orientation of the gene). If the upstream gene is less than 1 kb from the TSS of the gene of interest, we extend the analysis to introns of the upstream gene located within 10 kb of the TSS. In most cases, the downstream boundary is the end of the gene of interest. However, if the gene is longer than 30 kb from the TSS to the end of the last exon, intronic regions within 30 kb downstream of the TSS are used. Conversely, we include 2 kb of sequence downstream of the TSS for genes shorter than 2 kb in length.

Nonexonic conserved regions. PhastCons scores and PhastCons "conserved elements" computed from comparisons of 17-way vertebrate multispecies alignments (3) were downloaded from the UCSC Genome Browser database (4). Only PhastCons conserved elements that are both 20 bp or longer and nonoverlapping with annotated human mRNAs or Ensembl human gene annotations are retained for analysis. PhastCons conserved elements separated by less than 100 bp are chained together (excluding the intervening regions) and thereafter considered part of a single longer conserved region.

Score definition. We define a raw regulatory resolution score as:

$$raw\ score = \log_{10}\left(\frac{\sum_n l(c-b)}{n^2}\right),$$

where l is the length of the conserved region, c is the "conservation level" of the conserved region (i.e., the mean PhastCons score for the conserved region), b is the baseline conservation level (i.e., the mean PhastCons score for the entire genomic segment analyzed), and n is the number of conserved regions. Thus, for each conserved region, we consider the amount of conserved sequence, how well distinguished the region is from the background, and penalize genes with many conserved regions. After computing the raw score, we normalize it to obtain a value between 0 and 1 using the following formula:

$$normalized\ score = \frac{raw\ score - \min\ raw\ score}{\max\ raw\ score - \min\ raw\ score}.$$

Thus, zero is a gene with little resolution and 1 is highly resolved.

Genomewide distribution of regulatory resolution scores. Regulatory resolution scores were computed for all human genes as reported in the Ensembl annotations (Fig. S2). Of 22,298 genes tested, 2,411 did not contain any conserved PhastCons elements (Fig. S2*B*) and therefore we could not compute regulatory resolution scores for these genes. The distribution of scores is skewed toward zero, with a median score of 0.34 and a mean score of 0.36 (Fig. S2*A*). Genes with up to five conserved regions receive higher scores (Fig. S2*C*), with the top 20th percentile having an average of 2.2 conserved regions per gene. The highest scores were assigned to genes with less than 1,000 bp of conserved nonexonic nucleotides (Fig. S2*D*), with an average of 330 bp of conserved sequence per gene for genes scoring within the top 20th percentile.

Features of genes with the highest, average, and lowest regulatory resolution. The *ADCK5* locus was assigned the highest regulatory resolution score due to the presence of a single 1,277-bp highly conserved region within the upstream intergenic region (Fig. S3*A*). Two smaller conserved regions directly upstream of *ADCK5* in the "17-Way Most Cons" track are excluded from the analysis. The larger of the two overlaps with human mRNAs, wheras the smaller conserved element is only 10 bp long (Fig. S3*A*). The low baseline conservation level across the entire region further contributed to the high score. The *ELOVL3* locus (Fig. S3*B*) is an example of an average gene, receiving the mean score of 0.35. It contains 4 small, conserved nonexonic regions containing a total of 186 bp of sequence within the boundaries of the analysis. The lowest scoring gene, *NR4A3*, features 90 small conserved regions, containing a total of 9,281 bp of nonexonic sequence, which are distributed across the entire locus (Fig. S3*C*). The majority of the *NR4A3* locus is conserved and the conservation profile reveals few insights into the location of potential regulatory regions for targeted promoter construct design.

Manual promoter curation. Promoters for 100 genes were manually assessed on the basis of a number of gene features, including: (*i*) the location of the transcription start points; (*ii*) the boundaries of analysis, i.e., the amount of noncoding sequence to be analyzed upstream and downstream of the gene of interest; and (*iii*) the number and qualitative conservation level of conserved regions located proximal to the TSS within the defined boundaries. The genes were ranked from 1 to 5 on the basis of the curators' perception of their suitability for MiniPromoter (MiniP) design, "1" being a gene not suitable and "5" a very good candidate.

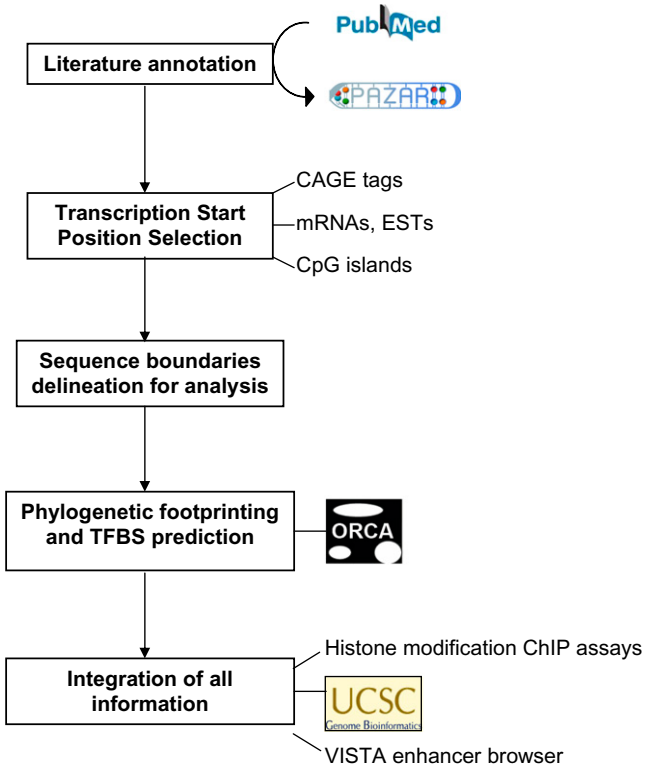
MiniPromoter Design. The MiniP design pipeline is represented below and includes the following resources. PubMed (http://www.ncbi.nlm.nih.gov/pubmed), PAZAR (http://www.pazar.info), the UCSC genome browser (http://genome.ucsc.edu), ORCAtk (http://www.cisreg.ca/cgi-bin/ORCAtk/orca), the VISTA enhancer browser (http://enhancer.lbl.gov), and histone modification ChIP (chromatin immunoprecipitation) assays performed on mouse and human cortex.

The endogenous promoters of genes are identified using genome annotations for 5′ cap analysis gene expression (CAGE) tags (5), transcripts (mRNAs, ESTs), and CpG islands (6). The boundaries for analysis are defined similarly to the regulatory resolution score analysis (see above), except that if one of the neighboring genes has an expression pattern similar to the gene of interest, the boundaries are extended to include the surrounding sequences of this additional gene. In a few cases, the GENSAT project had generated and tested BAC mice for the gene of interest and the expression pattern reported matched the endogenous expression pattern (7). In such cases, the BAC sequence defined the boundaries for regulatory sequence analysis.

Phylogenetic footprinting and transcription factor binding site (TFBS) prediction were performed using the ORCA toolkit (8) and the following steps:

Retrieval and alignment of human and mouse orthologous sequences within the defined boundaries.

Computation of the noncoding conserved regions above a user-defined threshold (ranging from 50 to 85% identity in our analyses).



Prediction of TFBS in those conserved regions for the transcription factors that have been described to be relevant for the expression of this specific gene or for expression in the brain region of interest in general. The TF binding models were extracted from the JASPAR database (9) or custom-generated from the PAZAR database on the basis of the manually curated "Pleiades Genes" project (10).

Hprt Targeting Vectors and MiniPromoters. The *Hprt* targeting vectors used in this study are pEMS1306 (EGFP reporter, ref. 1), pEMS1313 (lacZ reporter), and pEMS1302 (EGFP/cre reporter). The pEMS1313 and pEMS1302 fragments from the multiple cloning site (MCS) to the end of the reporter gene were synthesized by GeneArt and cloned into the *Hprt* targeting plasmid pJDH8A/246b (11) using EcoRI restriction sites.

MiniPs typically comprised up to four distinct genomic segments joined by fusion PCR. Each genomic segment was first PCR amplified independently using AccuPrime *Pfx* DNA Polymerase (Invitrogen), PCR primers (Integrated DNA Technologies), and BAC DNA template (10 pg to 200 ng). PCR primers for the outermost 5′ and 3′ segments were tailed with the appropriate restriction sites to allow for cloning. For MiniPs with two segments or more, PCR products of upstream segments were 3′ tailed with 18-bp linkers homologous to the first 18 bp of the adjacent downstream genomic segment. Reaction conditions were 0.25 Unit enzyme, 1× AccuPrime *Pfx* reaction mix, 1.0 μM each primer mix in a 20-μL volume. A 2-min denaturation at 95 °C was followed by 30 cycles of 95 °C for 15 s, 30 s (at T_m corresponding to primer pair), and 68 °C for 90 s, plus a final extension at 68 °C for 10 min. The PCR was run on a 1% low melting point agarose gel, visualized using SYBR Green (Invitrogen), excised, and recovered from the gel using QIAquick gel extraction kit (Qiagen). Reaction products were eluted using 30 μL of Ultrapure water (Gibco Invitrogen) and then quantified using the NanoDrop (Thermo Fisher Scientific). For MiniPs with multiple elements, fusion PCR was performed as above, but using 2.0 μL of gel purified first round reaction products (10 pg to 200 ng). Additional binary fusions were executed as above until the full length was obtained. A subset of nine MiniPs was generated by direct synthesis at GeneArt.

The final MiniPs were cloned into one of our *Hprt* targeting vectors and sequence verified with primers located every 300 bp along the construct on both strands. All discrepancies between the designed and constructed sequences were inspected using the UCSC Genome Browser annotations (hg18) (4). We tolerated discrepancies if they were known polymorphisms, located in a nonconserved region [PhastCons Vertebrate Multiz Alignment and Conservation (17 species) score below 0.7] or if analysis did not show any further regulatory implication. We rejected any sequence with insertion or deletion bigger than 10 bp.

Knockin Immediately 5′ of the Hprt Locus. The mEMS1204 [B6129F1-*Gt(ROSA26)Sor*^{tm1Sor}/+, *Hprt*$^{b\text{-}m3}$/Y], mEMS1202 [B6129F1-*Gt(ROSA26)Sor*$^{+/+}$, *Hprt*$^{b\text{-}m3}$/Y], and E14TG2A (12) ESC lines were electroporated with constructs built in pEMS1302, pEMS1306, or pEMS1313, respectively. Clones were maintained under HAT selection for 3–4 d of expansion in 96-well plates and then transferred to 2 × 24 wells and cultured in HT media. Once cells reached confluence, both wells were frozen in HT-freeze media and stored in liquid nitrogen (LN_2).

PCR Analysis of Genomic DNA. Vector NTI software (Invitrogen) was used to design PCR assays for the different constructs. MiniP-specific PCR genotyping assays are available on the http://www.pleiades.org website.

In Vitro Neural Differentiation. Neural differentiation of ESCs was conducted as previously described (13), with the following modifications. Once confluent, ESCs were trypsinized and seeded in duplicate wells onto confluent MS-5 feeder layers at 500 cells/cm^2 for seven time points. Total cell RNA was extracted with the RNeasy Plus mini kit (Qiagen) and used in RT-PCR analysis in both +RT and −RT conditions, using the OneStep RT-PCR kit (Qiagen) according to manufacturer's instructions. Ple53-EGFP immunohistochemistry was performed on day 11 of differentiation. Cells were washed once with 1× PBS and fixed using 4% paraformaldehyde in PBS for 15 min at room temperature. Cells were then blocked using Image-iT FX signal enhancer (Invitrogen) reagent and subsequently incubated with 1:1,000 rabbit polyclonal anti-GFP antibody (Abcam) followed by 1:1,000 Alexa-488 secondary anti-rabbit antibody (Invitrogen). Cells were imaged on a Zeiss Axiovert 200M microscope at ×20 with the FITC filter set. Ple88-lacZ staining was performed as outlined at http://openwetware.org/wiki/LacZ_staining_of_cells, on day 14 of differentiation. Brightfield images were taken with the ×10 objective on an Olympus Bx61 microscope.

Immunohistochemistry and Histochemistry. The mice were anesthetized with avertin and perfused transcardially with 4% PFA (paraformadehyde) in 0.1 M phosphate buffer (pH 7.4) for 15 min. brains were dissected and postfixed in the same fixation solution for 2 h and transferred into 25% sucrose-PBS overnight. Each brain was sectioned in a cryostat and 20-μm sagittal sections were collected. EGFP expression was detected with anti-GFP using the Vectastain Elite ABC kit (Vector Labs) and DAB, as a brown chromogen, following the manufacturer's directions. Expression of the beta-galatosidase (lacZ) or the EGFP/cre fusion protein [following recombination of the *Gt(ROSA)26Sor*^{tm1Sor} locus (14)] was detected with X-gal (5-Bromo-4-chloro-3-indolyl-β-d-galactopyranoside) staining as previously described (15). High-resolution serial images of brightfield material were acquired using a Nikon Optiphot-2 microscope with a LEP motorized stage connected to a Dell Precision 390 computer equipped with hardware and software from

MicroBrightField Images were captured and tilled using MBF Neurolucida Virtual Slice v8.2.3.0.

Double-label immunofluorescence for colocalization of EGFP and endogenous proteins was performed as previously described (16). Either native EGFP fluorescence (nGFP) or anti-GFP detection with an Alexa-488 secondary antibody was combined with a second primary antisera and detection with a Cy3 or Alexa-555 secondary antibody. Costaining of lacZ activity and tyrosine hydroxylase or NeuN was performed sequentially as previously described (15). Primary antibodies used for these studies include: rabbit anti-DCX (1:500, Cell Signaling), rabbit anti-orexin (1:500, Millipore), mouse anti-GFAP (1:1,000, Millipore), mouse anti-S100 (1:1,000, Abcam), mouse anti-NeuN (1:500, Chemicon, mouse anti-TH (1:3,000, Chemicon), mouse anti-RIP (1:500, Chemicon). Secondary antibodies include: goat anti-rabbit Alexa-488 [1:500, Molecular Probes (Invitrogen)], goat anti-rabbit-Cy3 (1:500, Jackson ImmunoResearch Laboratories), goat anti-mouse Alexa-555 (1:500, Molecular Probes), goat anti-mouse Alexa-488 (1:500, Molecular Probes), donkey anti-goat-Cy3 (1:500, Jackson ImmunoResearch Laboratories). Sections were counterstained with TOTO3 (2 μM, Molecular Probes) and mounted with anti-fade reagent FluroSave Reagent (Calbiochem). Detection of double immunofluorescence was performed using a confocal laser-scanning microscope (CLSM, BioRad). Whole-mount X-gal histochemistry was performed on 4% PFA fixed embryos (E10.5, E11.5) or dissected brains (E15.5, P0.5) following a similar protocol described above after preincubation of the tissue in 0.1 M PBS containing 0.3% Triton X-100. Stained embryos and brains were photographed, cryosectioned, and counterstained with neutral red for localization of lacZ expressing cells.

Histochemistry of 1-mm Brain Slices. Mice were perfused with 4% PFA, and postfixed 2–4 h as described above. The brains were then removed from 4% PFA and immediately sectioned. The dissected brains were placed ventral side up into the adult mouse coronal or sagittal, Rodent Brain Matrix (ASI Instruments). Slices were sectioned lateral through medial to lateral (for sagittal) or rostral to caudal (for coronal) using single-edge razor blades (Electron Microscopy Sciences). All slices from one brain were placed into one well of a 12-well plate containing 1× PBS (Invitrogen), or 0.1% sodium azide (in 1× PBS) until staining. The staining was performed with 3–5 mL of X-gal staining solution (25 mg/mL X-gal, 1 M MgCl2, 50 mM potassium ferri-cyanide, 50 mM potassium ferrocyanide, and 1× PBS to volume) per well in a 24-well plate. The plate was wrapped in aluminum foil and incubated at 37 °C for 10–16 h. Subsequently, the sections were transferred into PBS, examined under a dissecting microscope, and photographed using a CoolSNAP-Procf color camera (Media Cybernetics) mounted on a Leica MZ12.5 stereomicroscope (Leica Microsystems) and Image-Pro Express v.4.5.1.3 software (Media Cybernetics).

Regulatory Element Predictions in OLIG1 Enhancer Sequences.

Identification of "most conserved" aligned sequences in OLIG1 construct sequences. The genomic coordinates for each of the conserved regions constituting the tested *OLIG1* MiniPs were retrieved using the BLAT sequence search tool at the UCSC browser against the Human March 2006 assembly (4). The genomic coordinates were used as input to the UCSC table retrieval function to extract the human sequence alignment in the 17-way multiple mammalian-species and 28-way placental mammals "most conserved" alignments (17-way and 28-way, respectively, in Table S1) and each aligned human sequence (with gaps) was stored in FASTA format.

Transcription factor binding site (TFBS) predictions in "most conserved" sequences. Each of the conserved regions making up a MiniP was subjected to a TFBS prediction analysis. A PERL script was developed using the TFBS PERL module (17) and the JASPAR CORE database (9) (supplemented with additional model annotations for Glia-related TFBS: POU2F1; EGR1; EGR2; EGR3; EGR4; POU3F1; NKX2-2; NKX2-5) to evaluate vertebrate TFBS models across the most conserved sequence elements of each region using profile score threshold levels of 75% and 80% (75% and 80%, respectively, in Table S1). TFBS predictions were written to BED formatted files for each analyzed alignment.

Analysis of in vitro oligodendrocyte gene expression profile data. To identify TF candidates that could be directing *OLIG1* regulation in oligodendrocyte cells, we analyzed an in vitro 8-d time point oligodendrocyte differentiation dataset produced by Dugas et al. (18). This dataset is composed of recorded gene expression profiles across a timescale of differentiated, purified, rat cortical oligodendrocyte progenitor cells (OPC) using the Affymetrix RG_U34-A, RG_U34-B, and RG_U34-C chips. A total of 96 Affymetrix CEL files (8 time points × 4 biological replicates × 3 chips) were obtained from J. C. Dugas. We developed R code (http://www.R-project.org) and used the Bioconductor packages (19) to perform a robust multichip analysis (RMA) (20) on each chip dataset to obtain a probe-level summarization. All pairwise experiments were subjected to a two-sample *T* test with a random variance model (21) implemented in the BRB-array software (http://linus.nci.nih.gov/~brb). The Rat Affymetrix chip probes were mapped to Entrez rat genes using Bioconductor packages. The rat Entrez Genes were mapped to mouse Entrez Genes (where possible) using Homologene (22). A set of mouse TF genes (23) was mapped to the rat Affymetrix probes. PERL software was written to convert all HTML-formatted expression analysis results to text files and extract and report all significantly (P-value $\leq$ 0.001), differentially expressed genes across the pairwise expression profiles and mapped TF genes in this set were identified.

Evaluation of TFBS predictions. TFBS predictions in the positively expressed MiniP construct (Ple151) and the MiniP constructs that had no reporter gene expression (Ple148 and Ple150) were compared using a PERL script to identify the TFBS predictions that were unique to the expressed MiniP. These unique TFBS predictions were then compared against the expression profile analysis results (Table S1).

Prioritization of candidate TFBS. The compiled TFBS predictions and expression data analyses were reviewed to rank the TFBS candidates. TFBS predictions that were unique to the positive Ple151 construct with differential gene expression and correlated literature evidence support were reported (Table S2).

1. Yang GS, et al. (2009) Next generation tools for high-throughput promoter and expression analysis employing single-copy knock-ins at the Hprt1 locus. *Genomics* 93:196–204.
2. Flicek P, et al. (2008) Ensembl 2008. *Nucleic Acids Res* 36 (Database issue):D707–D714.
3. Siepel A, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15:1034–1050.
4. Karolchik D, et al. (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* 36 (Database issue):D773–D779.
5. Shiraki T, et al. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci USA* 100:15776–15781.
6. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. *J Mol Biol* 196:261–282.
7. Gong S, et al. (2003) A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* 425:917–925.
8. Portales-Casamar E, et al. (2009) The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res* 37 (Database issue):D54–D60.
9. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32 (Database issue):D91–D94.
10. Portales-Casamar E, et al. (2007) PAZAR: A framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biol* 8:R207.
11. Heaney JD, Rettew AN, Bronson SK (2004) Tissue-specific expression of a BAC transgene targeted to the Hprt locus in mouse embryonic stem cells. *Genomics* 83: 1072–1082.
12. Hooper M, Hardy K, Handyside A, Hunter S, Monk M (1987) HPRT-deficient (Lesch-Nyhan) mouse embryos derived from germline colonization by cultured cells. *Nature* 326:292–295.

13. Barberi T, et al. (2003) Neural subtype specification of fertilization and nuclear transfer embryonic stem cells and application in parkinsonian mice. *Nat Biotechnol* 21:1200–1207.
14. Soriano P (1999) Generalized lacZ expression with the ROSA26 Cre reporter strain. *Nat Genet* 21:70–71.
15. Reiner A, et al. (2007) R6/2 neurons with intranuclear inclusions survive for prolonged periods in the brains of chimeric mice. *J Comp Neurol* 505:603–629.
16. Liu L, et al. (2007) A transgenic mouse class-III beta tubulin reporter using yellow fluorescent protein. *Genesis* 45:560–569.
17. Lenhard B, Wasserman WW (2002) TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* 18:1135–1136.
18. Dugas JC, Tai YC, Speed TP, Ngai J, Barres BA (2006) Functional genomic analysis of oligodendrocyte differentiation. *J Neurosci* 26:10967–10983.
19. Gentleman RC, et al. (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol* 5:R80.
20. Irizarry RA, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249–264.
21. Wright GW, Simon RM (2003) A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* 19:2448–2455.
22. Wheeler DL, et al. (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 36 (Database issue):D13–D21.
23. Fulton DL, et al. (2009) TFCat: The curated catalog of mouse and human transcription factors. *Genome Biol* 10:R29.

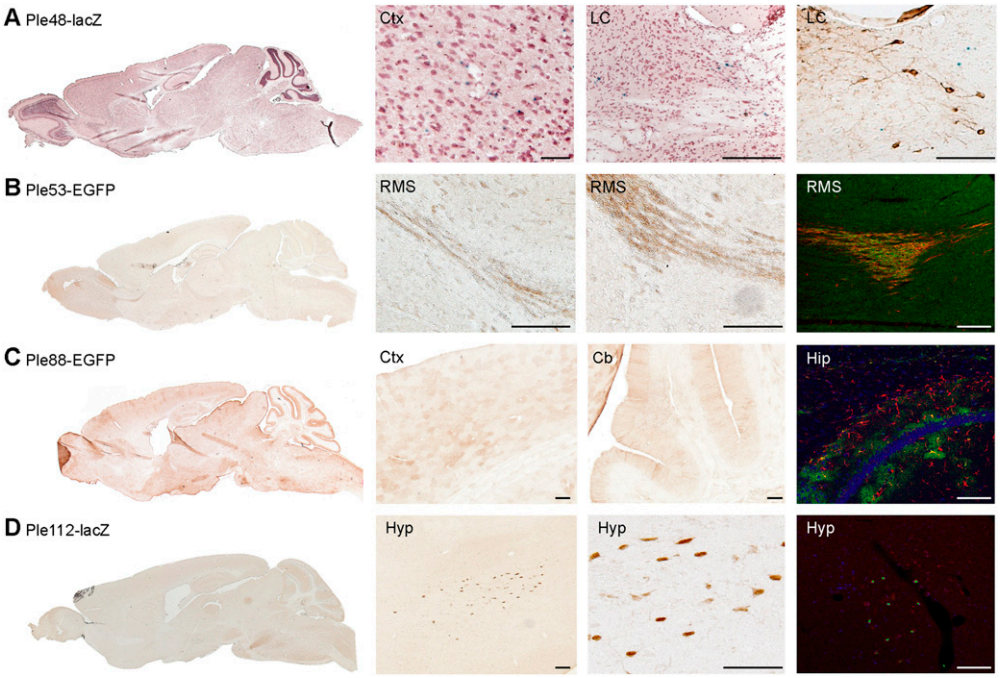
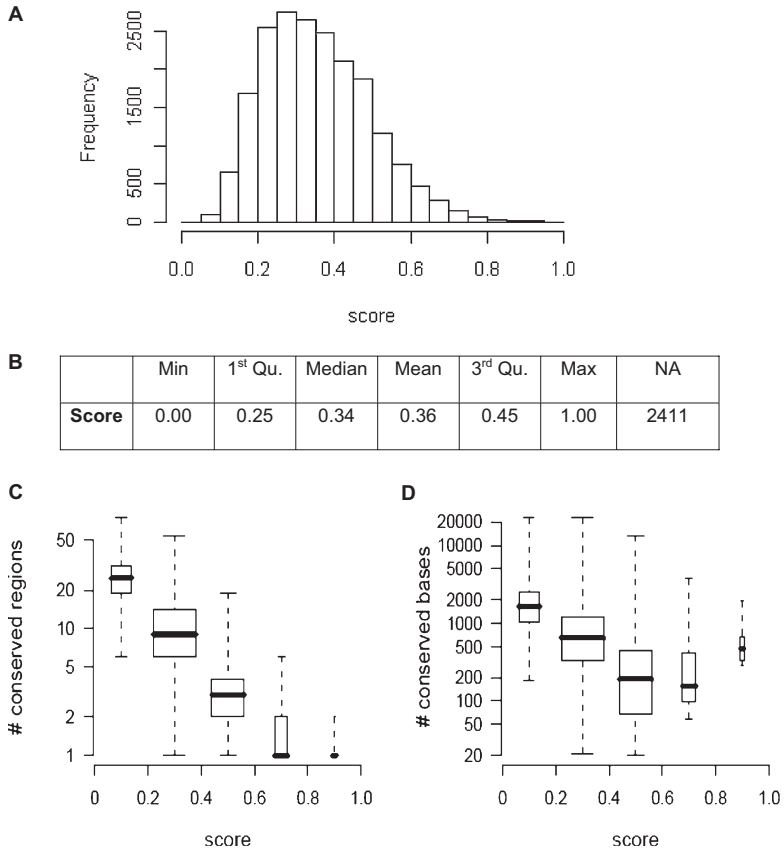


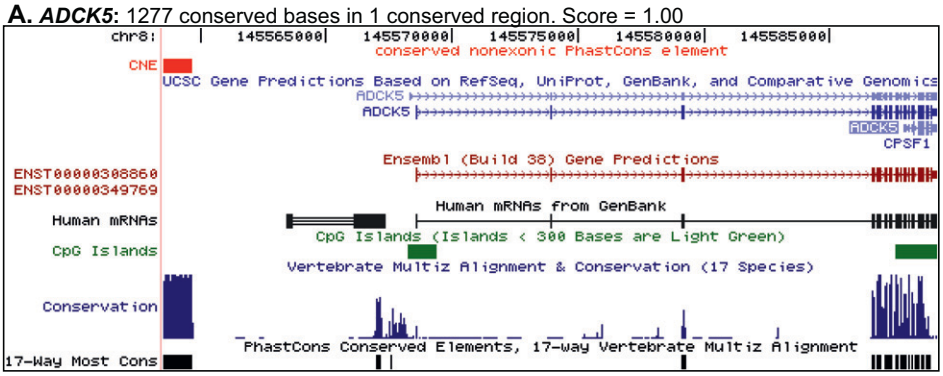
Fig. S1. Targeting of previously characterized MiniPromoters to the *Hprt* locus validates the Pleiades Promoter Project approach. Neurohistological analysis of four control strains carrying previously characterized random-insertion constructs knocked in at the *Hprt* locus. EGFP is detected using anti-GFP immunocytochemistry (brown signal in brightfield images) and lacZ is detected using X-gal histochemistry (blue signal counterstained with neutral red in brightfield images). Cb, cerebellum; Ctx, cortex; Hip, hippocampus; Hyp, hypothalamus; LC, locus coeruleus; RMS, rostral migratory stream. (*A*) Ple48-lacZ expression (previously characterized construct based on *DBH* regulatory regions) is enriched in noradrenergic cells in the locus coeruleus and the adrenal gland but also present in other regions such as the cortex. The last image shows no costaining of beta-galactosidase activity (blue) with tyrosine hydroxylase (brown) in the locus coeruleus. (*B*) Ple53-EGFP (previously characterized construct based on *DCX* regulatory regions) expression is observed in multiple regions of the brain as seen on the whole brain image with enrichment in the olfactory bulb and the rostral migratory stream. The last image shows costaining (yellow) of EGFP (green) with the endogenous Dcx protein (red). (*C*) Ple88-EGFP (previously characterized construct based on *GFAP* regulatory regions) is expressed in astrocytes throughout the brain. The last image shows costaining (yellow) of EGFP (green) with the endogenous Gfap protein (red); the nuclear counterstain is TOTO3 (blue). (*D*) Ple112-EGFP (previously characterized construct based on *HCRT* regulatory regions) is specifically expressed in a cluster of hypothalamic cells. The last image shows costaining of EGFP (green) with the endogenous Hcrt protein (red); the nuclear counterstain is TOTO3 (blue). (Scale bars, 100 μm.)



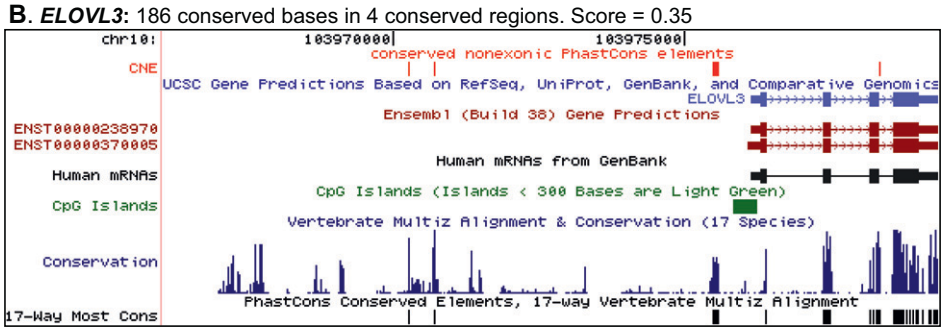
	Min	1st Qu.	Median	Mean	3rd Qu.	Max	NA
Score	0.00	0.25	0.34	0.36	0.45	1.00	2411

Fig. S2. Genomewide distribution of regulatory resolution scores. (*A*) Histogram of scores. (*B*) Summary statistics showing the score by quartiles (Qu.), as well as the median and mean score. NA, not able to score. (*C*) Boxplot showing the distribution of the number of conserved regions by score intervals. (*D*) Boxplot showing the distribution of the number of conserved bases by score intervals. The boxes in both boxplots are drawn with widths proportional to the square roots of the number of observations in the groups.

A. *ADCK5*: 1277 conserved bases in 1 conserved region. Score = 1.00



B. *ELOVL3*: 186 conserved bases in 4 conserved regions. Score = 0.35



C. *NR4A3*: 9281 conserved bases in 90 conserved regions. Score = 0.00

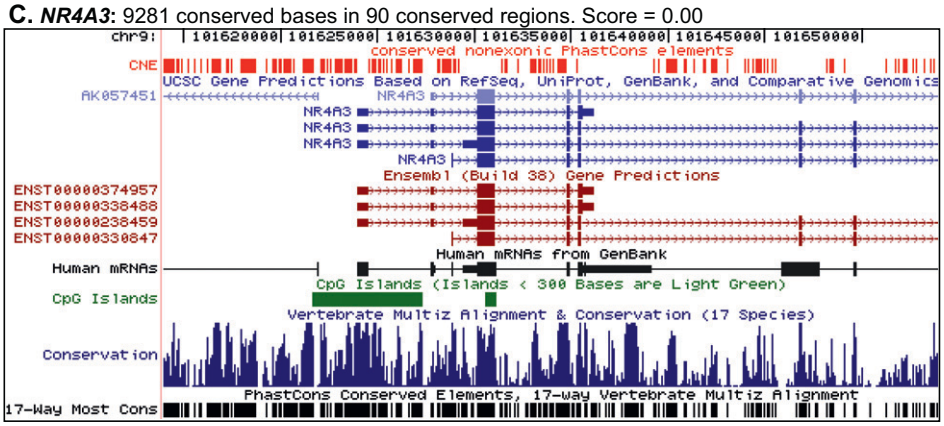


Fig. S3. Genes with (*A*) the highest, (*B*) average, and (*C*) the lowest regulatory resolution scores. Each screenshot from the UCSC Genome Browser (http://genome.ucsc.edu; NCBI Build 36.1) displays: conserved nonexonic (CNE) PhastCons elements used in the analysis; UCSC gene predictions based on RefSeq, GenBank and UniProt data; transcripts for Ensembl genes based on mRNA and protein evidence; a dense display of human mRNAs from GenBank; CpG islands (≥50% GC content, ≥200 bp in length, and an observed CG to expected CG ratio ≥ 0.6); evolutionary conservation in 17 vertebrates based on Multiz alignments and PhastCons scores; and predictions of conserved elements produced by the PhastCons program (17-way Most Cons).

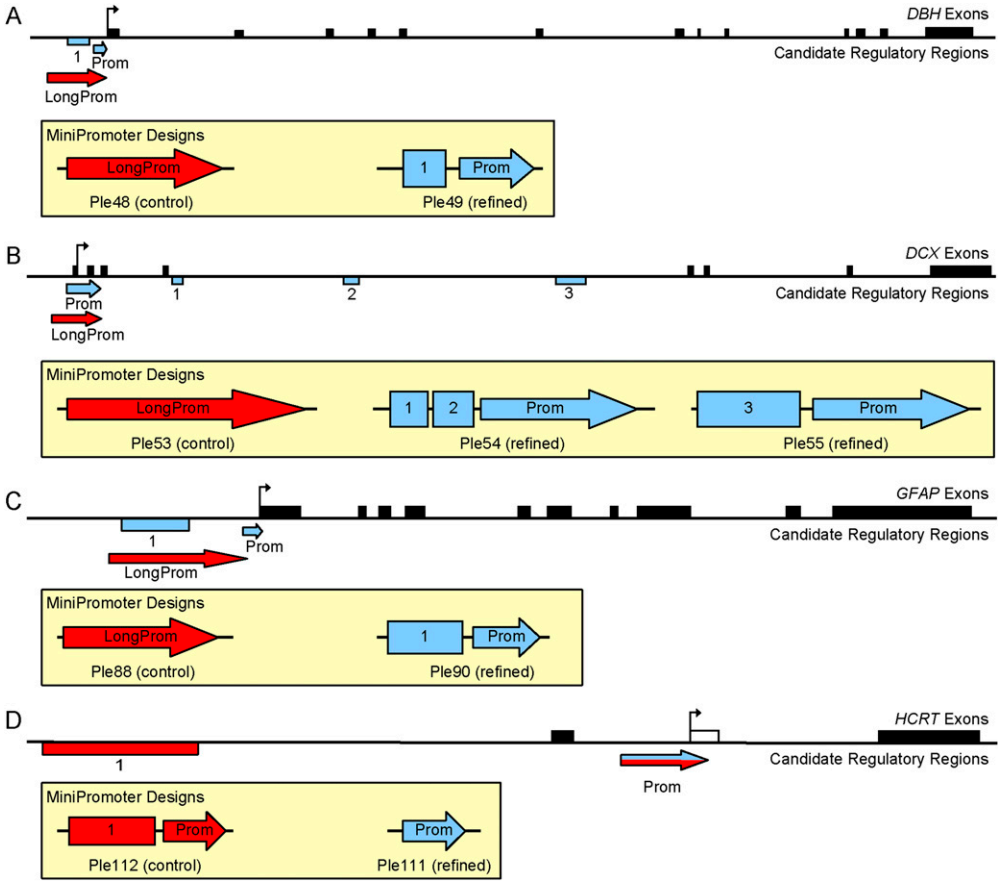


Fig. S4. Comparison of the Pleiades refined MiniPromoters with the previously characterized constructs. The diagrams represent the human genomic sequences upstream and within (*A*) *DBH*, ~30 kb displayed, (*B*) *DCX*, ~130 kb displayed, (*C*) *GFAP*, ~15 kb displayed, and (*D*) *HCRT*, ~5 kb displayed. Black boxes indicate the exons, black arrows for the transcription start sites, red features for the regulatory regions contained in the control promoters, and blue features for the regulatory regions contained in the refined MiniPromoters. The yellow boxes outline the MiniPromoter designs analyzed.

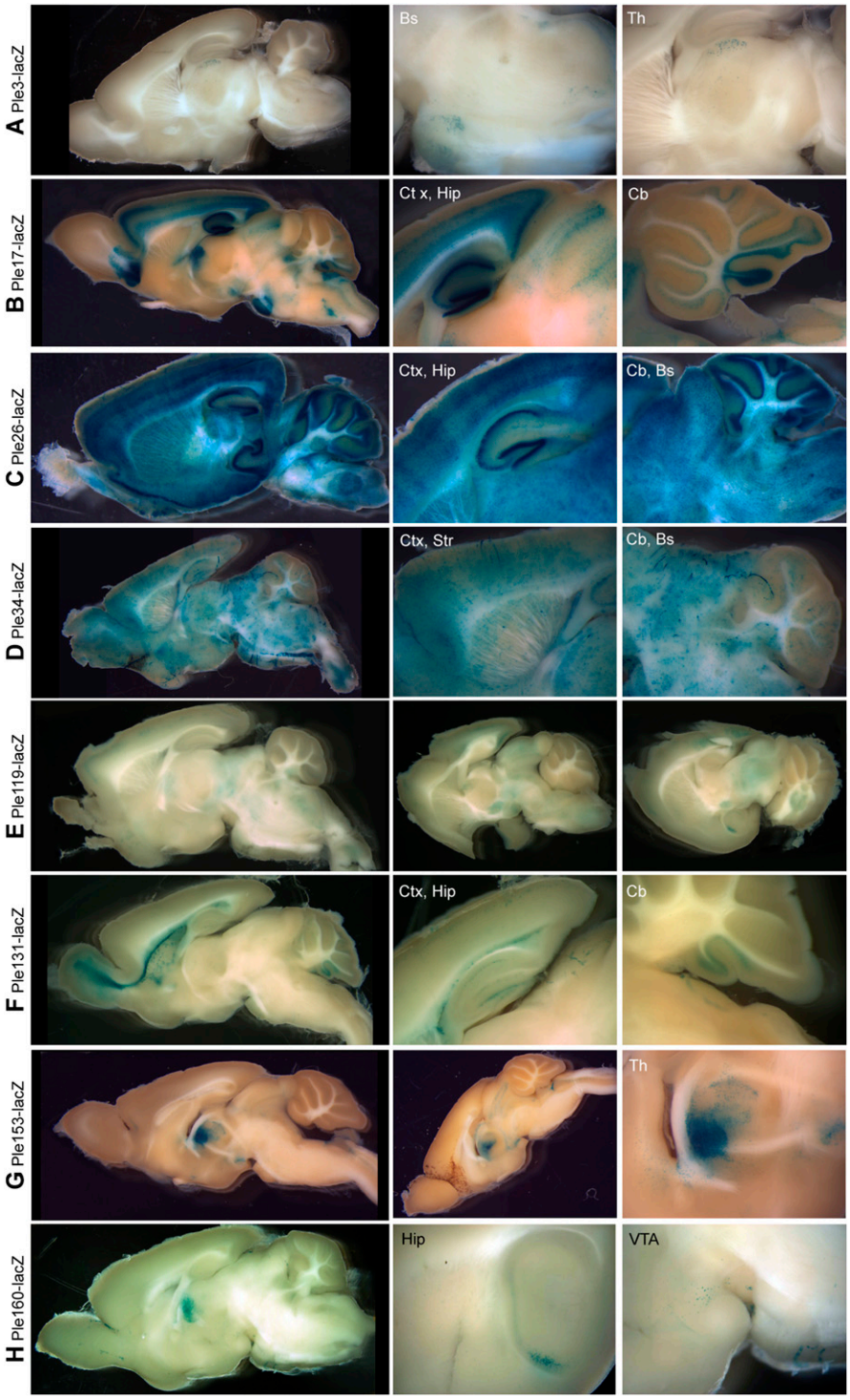


Fig. S5. (Continued)

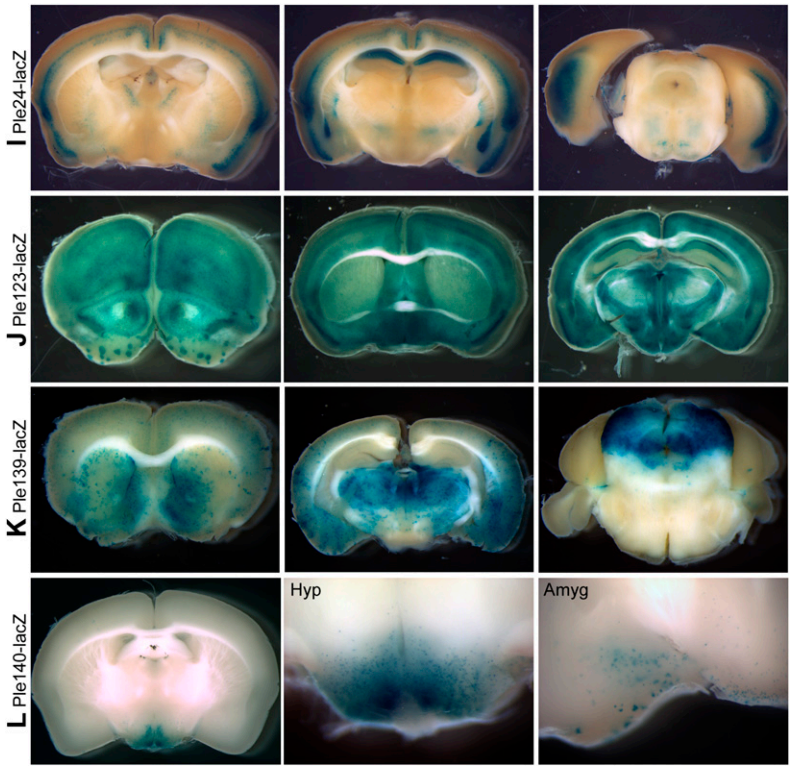


Fig. S5. Selected novel MiniPromoter expression patterns assessed in 1-mm brain slices. (*A*) Ple3-lacZ (*ADORA2A* RRs) expression is observed in the dorsal lateral geniculate and brainstem region. (*B*) Ple17-lacZ (*C8ORF46* RRs) expression is strong in the deep cortex, hippocampus, and posterior lobe of cerebellum. (*C*) Ple26-lacZ (*CCL27* RRs) expression is strong throughout the brain except for the olfactory region and thalamus. (*D*) Ple34-lacZ (*CLDN5* RRs) expression is present in, or around, blood vessels throughout the brain. (*E*) Ple119-lacZ (*HTR1A* RRs) staining is relatively sparse but nicely localized in the ventral thalamic/posterior hypothalamic territories, cortex layer IV, hippocampus area CA1c, and retrosplenial cortex. (*F*) Ple131-lacZ (*MKI67* RRs) expression is strong surrounding the ventricles, in the RMS, and the dentate gyrus. (*G*) Ple153-lacZ (*OXT* RRs) expression focused in the anterior thalamic territory. (*H*) Ple160-lacZ (*PITX3* RRs) expression is strong in the anterior thalamus, ventral-lateral hippocampus, and present in the VTA region. (*I*) Ple24-lacZ (*CCKBR* RRs) expression is enriched in the cortex, basal lateral amygdala, hippocampal pyramidal cells and in the red nucleus. (*J*) Ple123-lacZ (*ICMT* RRs) expression is strong throughout the brain. (*K*) Ple139-lacZ (*NR2E1* RRs) presents a very regional staining, heavy up through dorsal midbrain, then virtually absent going more ventral and posterior. (*L*) Ple140-lacZ (*NR2E1* RRs) expression is strong in the hypothalamus and present in the amygdala.

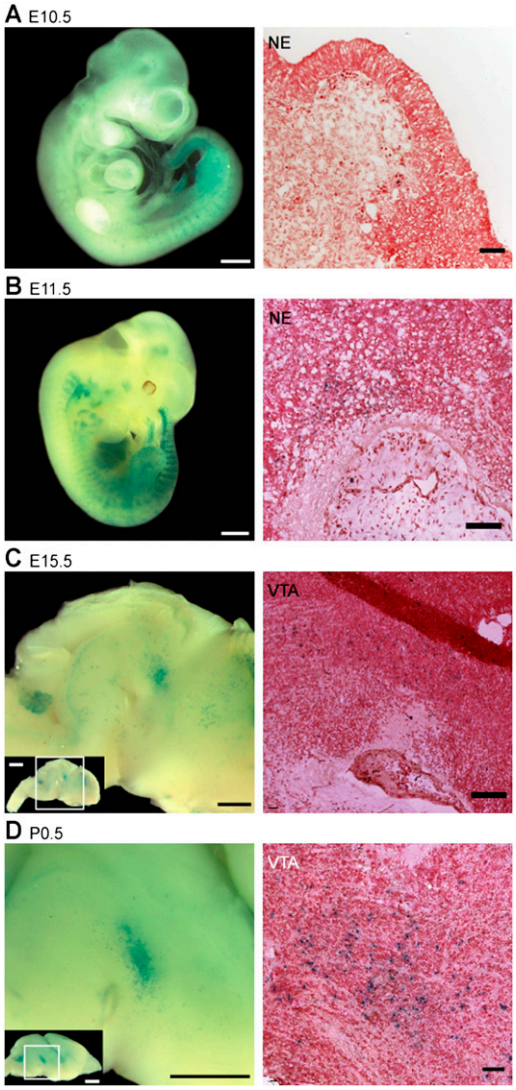


Fig. S6. MiniPromoters as tools to study developmental expression patterns. X-gal staining was performed in whole-mount (w) and sagittal sections (s) in Ple162-EGFP/cre mice (based on *PITX3* regulatory regions) across development from embryonic day (E) (*A*) 10.5, (*B*) 11.5, (*C*) 15.5, and (*D*) postnatal day (P) 0.5. NE, neuroepithelium; VTA, ventral tegmental area. [Scale bars, 50 μm (Cw, As, Bs, Ds), 100 μm (Cs), 500 μm (Aw), 750 μm (Dw), and 1,000 μm (Bw, Cw *Inset*, Dw *Inset*).]

Table S1. List of predicted TFBS that were unique to the positive Ple151 construct

TF predictions	TFs differentially expressed (U34A-C chips)	17-way 80%	17-way 75%	28-way 80%	28-way 75%
AR	No		✓		✓
CREB1	No	✓			
DDIT3 (JP: Ddit3-Cebpa)	No			✓	
E2F1	Probe mapping not available	✓	✓	✓	✓
EGR1	Yes—Down-regulated from OPCs over time and up-regulated at D7-D9				✓
EGR4	No		✓		✓
ELK1	No			✓	
ELK4	Probe mapping not available	✓		✓	
EVI1	No		✓		
FOS	Yes—FOSL1 (FRA-1) expression is down-regulated between OPC and D2	✓			
FOXD1	No	✓		✓	
FOXF2	Probe mapping not available			✓	
GABPA	Probe mapping not available			✓	✓
GLI1	Probe mapping not available		✓		
HINF (JP: MIZF)	Probe mapping not available	✓		✓	
HLF	No	✓		✓	
IRF2	No		✓	✓	✓
MAX	No	✓			
NFKB1	Yes—Down-regulated between OPC and D2				✓
NFYA	No			✓	
NKX2-2	Probe mapping not available			✓	
NKX3-1	Probe mapping not available			✓	
NKX3-2 (JP: Bapx1)	Probe mapping not available			✓	
NR3C1	Yes—Down-regulated between D9 and acute OLs	✓	✓	✓	✓
PAX4	No		✓		✓
PAX5	Probe mapping not available	✓		✓	✓
POU2F1	No	✓		✓	✓
RORA (JP: RORA_2)	Probe mapping not available	✓	✓	✓	✓
RREB1	No		✓		✓
RXRA (JP: RXRA-VDR)	No		✓		✓
STAT1	Yes—Up-regulated between OPC and D2		✓		✓
T brachyury	Probe mapping not available	✓	✓		
TEAD1	Probe mapping not available	✓		✓	
ZNF143 (JP: Staf)	Probe mapping not available		✓		✓
ZNF423 (JP: Roax)	No			✓	

JP, Jaspar profile name,included if different from TF HUGO gene name; OPC, oligodendrocyte progenitor cells; D2, day 2 time point in the Dugas et al. dataset; D9, day 9 time point in the Dugas et al. dataset; Ols, oligodendrocytes.

Table S2. Predicted TFBS candidates with differential gene expression and literature evidence support

TF predictions	TFs differentially expressed (U34A-C chips)	Literature evidence
EGR1	Yes—EGR1 (KROX-24) expression is down-regulated after OPC time point	EGR1 (KROX-24) may be involved in the initial oligodendrocyte differentiation primary response (20)
FOS	Yes—FOSL1 (FRA-1) expression is down-regulated between OPC and D2	The AP-1 family of TFs may play a role in oligodendrocyte differentiation (21)

OPC, oligodendrocyte progenitor cells; D2, day 2 time point in the Dugas et al. dataset.

Other Supporting Information Files

Dataset S1 (XLS)