Supporting Information

Colquitt et al. 10.1073/pnas.1302759110

SI Materials and Methods

Sequencing Alignment and Initial Processing. Read mapping was performed using Bowtie2 (default parameters, mm9). PCR duplicates and reads with mapping quality (MAPQ) < 30 were removed. Tracks were generated consisting of read counts binned into 25 bp normalized by the number of millions of reads and scaled to reflect the average number of reads per 1 kb to obtain reads per million (RPM) [i.e., 1E6 * 1E3/(window size * total number of reads)]. Except for differential peak analysis, subsequent analyses used tracks with values averaged across biological replicates, where available. For visualization, tracks were smoothed using a 125-bp moving average window. RNA-seq mapping was performed using Tophat2 [with no gene transfer format (GTF) file]. Reads with MAPQ < 30 were removed.

Genome-Wide Correlations. Pairwise Pearson correlation values were computed as in Wang (1). Read data binned into 1,000-bp windows were divided into groups of 100, correlation values were computed within these groups, and these values were averaged to obtain a genome-wide correlation value.

Differential Peak Analysis. Regions with differential enrichment of 5-hydroxymethylcytosine (5hmC) and 5-methylcytosine (5mC) were identified using HOMER (2). Tag directories (GC-normalized) for each sample and replicate were created independently from processed BAM files. Differential regions were called within either globose basal cells (GBCs) or mature olfactory sensory neurons (mOSNs) using horizontal basal cells (HBCs) as input, "histone" presets, peak size of 1,000 bp, and at least threefold enrichment over input (findPeaks -style histone -size 1000 -F 3). Control and Tet3-tg 5hmC differential peaks were similarly identified. Differential peak sets from each replicate were intersected (requiring an overlap of at least 50% of peak size) to create common peak sets for feature intersection analysis. Genomic positions for gene elements (1–3 kb upstream, 1 kb upstream, CGI, CDS, 3′ UTR, 5′ UTR) were obtained from the mm9 University of California at Santa Cruz (UCSC) refGene table. Intergenic sites ("intergenic") were determined using the complement of refGene regions extended by 5 kb. Repetitive elements were subtracted from these regions using the mm9 rmsk table. Conserved intergenic regions ("intergenic conserved") were obtained by intersecting this intergenic set with the vertebrate conserved elements (phastConsElements30way).

Aligned Feature Profiles. For aligned gene profiles, refGene position information was obtained from the UCSC genome browser, and 50 200-bp windows were generated upstream and downstream and spaced evenly throughout gene bodies. Transcription start site (TSS) profiles were generated using 25-bp windows extending from the 5′ most annotated TSS for each gene. For each profile, the mean of a given aligned position was computed (excluding values at extreme 1%), and 95% confidence intervals were calculated through bootstrap resampling for 1,000 iterations using the R package boot. Significant differences at a given aligned position between two samples were computed by permutation as follows. Bins of five positions were made, and average profile values were computed in these bins. For a given position and set of two samples, the differences between the mean values from 1,000 label randomizations were computed. The number of differences from the randomized data exceeding the true difference was used as *P* values. False discovery rates were computed from these *P* values using the Benjamini–Hochberg method.

Analysis of 5hmC/5mC and Transcription. Fragments per kilobase per millions of reads (FPKM) values were computed for each sample using Cufflinks (3) with assembly to the iGenomes Ensembl GTF file and multiread correction. Aligned gene profiles were grouped into quartiles by associated FPKM values. Transcriptional indices were generated by grouping FPKM-ranked genes into 100 groups (of 143 genes each). Guidance molecules are defined here as genes with these prefixes: *Efn*, *Eph*, *Dscam*, *Nphs*, *Kirrel, Nrp*, *Plxn*, *Sema*, *Pcdh*, *Ctcn*, *Robo*, and *Slit.*

Differential Expression Analysis. To define sets of genes with developmentally differential expression, the expression of the status of each gene was determined using normal mixture model-based clustering via the R package *mclust*. Cufflinks-generated log2 (FPKM+1) values were fit independently for each sample to a two-Gaussian mixture model with equal variance [*R* command: Mclust(values, $G = 2$, modelName = "E")]. The resulting classification was used to generate "specific" and "common" gene expression sets. To identify genes with differential expression, we followed the approach used in Katz (4), which applied a two-sided point null hypothesis test. For each gene, *g*, in our data, *d*, we computed $dg = Ag - Bg$, where *Ag* and *Bg* are the FPKM of *g* in the two samples *A* and *B*. The null hypothesis *H0* states that $dg = 0$ and the alternative hypothesis *Hg* states that $di ! = 0$. To select between the two hypotheses, we computed the Bayes Factor $(BF) = p(D|H1)p(H1)/p(D|H0)p(H0)$. Specifically, we applied a Kernel Density Estimator to compute the distribution $p(d|D, H1)$ and estimated each Bayes Factor as $BFg \cong 1/p(dg = 0|H1, D)$ [the Savage–Dickey density ratio with a prior density $p(dg = 0 \mid H1) = 1$].

Principal Component Analysis. Gene-body 5hmC levels were sampled for each gene using 50 200-bp windows spaced from TSS to TES (genes less than 10 kb long were excluded). Principal component analysis was performed on the resulting matrices using the R function *prcomp*. The original data matrix was then ordered by principal component 1 or 2, averaged by every 50 genes, and plotted as heatmaps.

1. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359.
2. Wang Z, et al. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 40(7):897–903.
3. 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.
4. 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.
5. 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.

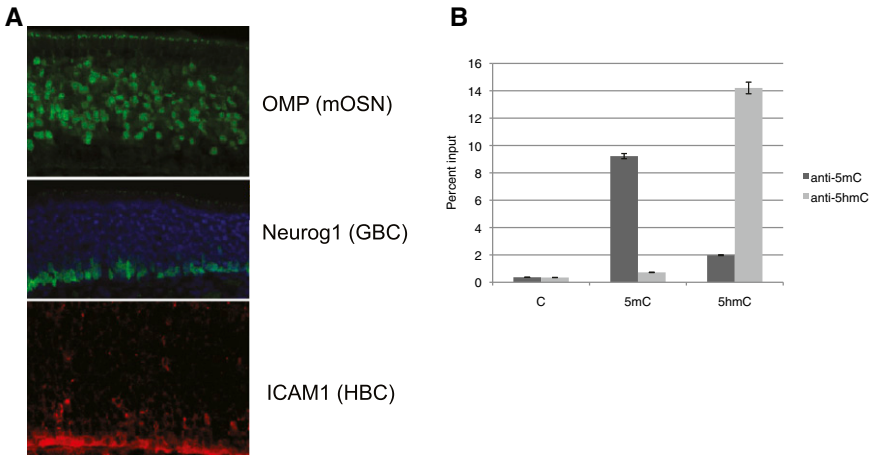

Fig. S1. (*A*) Coronal sections of adult main olfactory epithelium (MOE) from *OMP*-ires-GFP, *Neurog1*-GFP, and C57BL/6 mice stained with anti-ICAM1-phycoerythrin. (*B*) DNA immunoprecipitation (DIP)-quantitative PCR (qPCR) was performed using a 200-bp linear template containing unmodified cytosines, 5mC, or 5hmC. Error bars are SEM of three qPCR replicates.

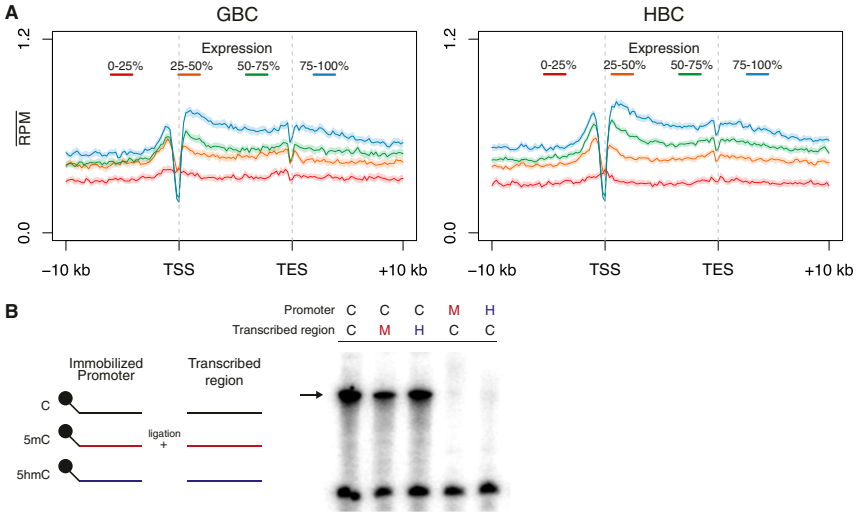

Fig. S2. (*A*) Aligned gene profiles of GBC and HBC 5hmC levels split by gene expression within each cell type. Error is 95% bootstrap confidence interval. (*B*) In vitro transcription assay testing the effects of 5mC and 5hmC within either the CMV promoter or a 200-bp transcribed fragment. (*Left*) Schematic of unmodified and modified template construction. (*Right*) Digital autoradiograph of transcribed product (arrow) with modified templates. C, unmodified; M, 5mC; H, 5hmC.

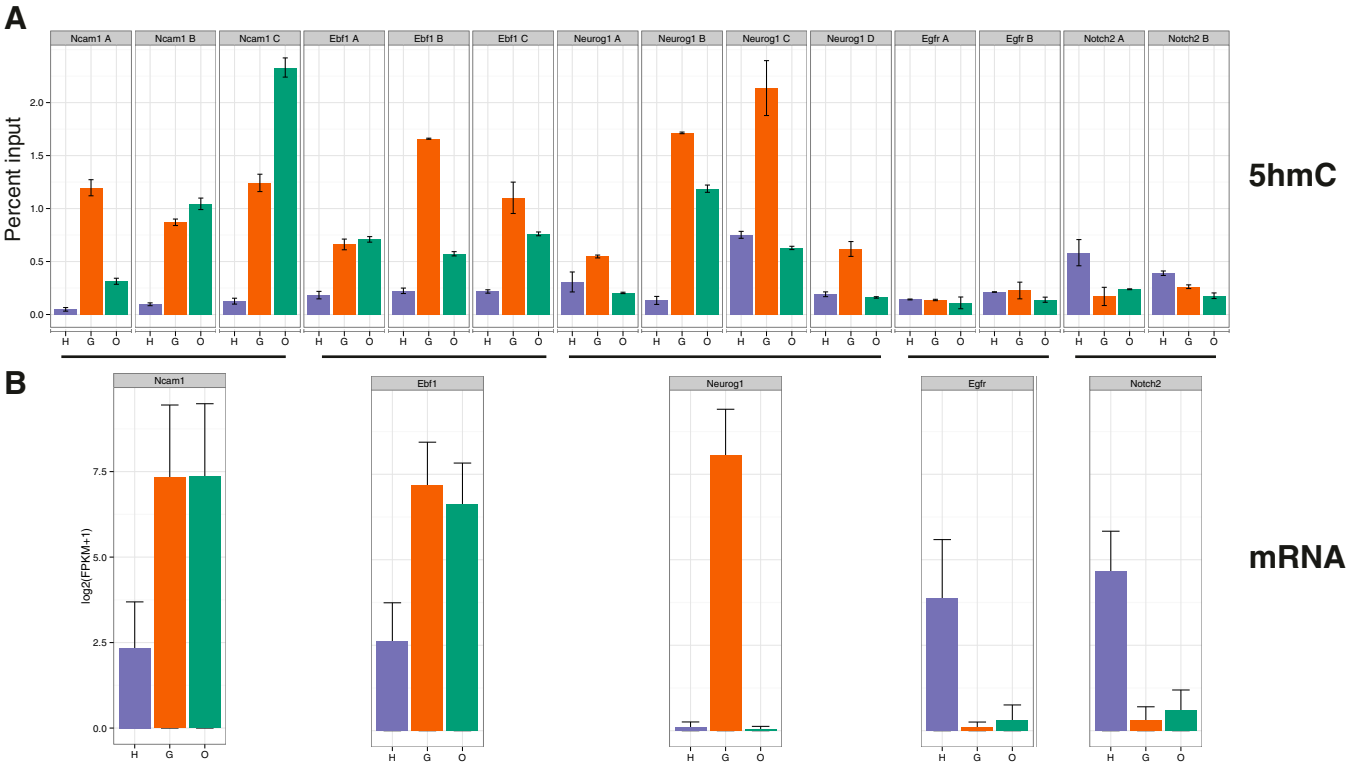

Fig. S3. (*A*) 5hmC and 5mC DIP-qPCR analysis of five differentially expressed genes. Primer sets amplify regions located within the gene body and, in the case of Neurog1 A-C sets, flanking the gene body. Cell types are HBC (H), GBC (G), and mOSN (O). Error bars are the range of qPCR duplicates. (*B*) log2(FPKM + 1) values from sorted-cell mRNA-seq datasets. Error bars are Cufflinks-generated confidence values. For each, the lower confidence bar is equal to 0 and has been omitted for clarity.

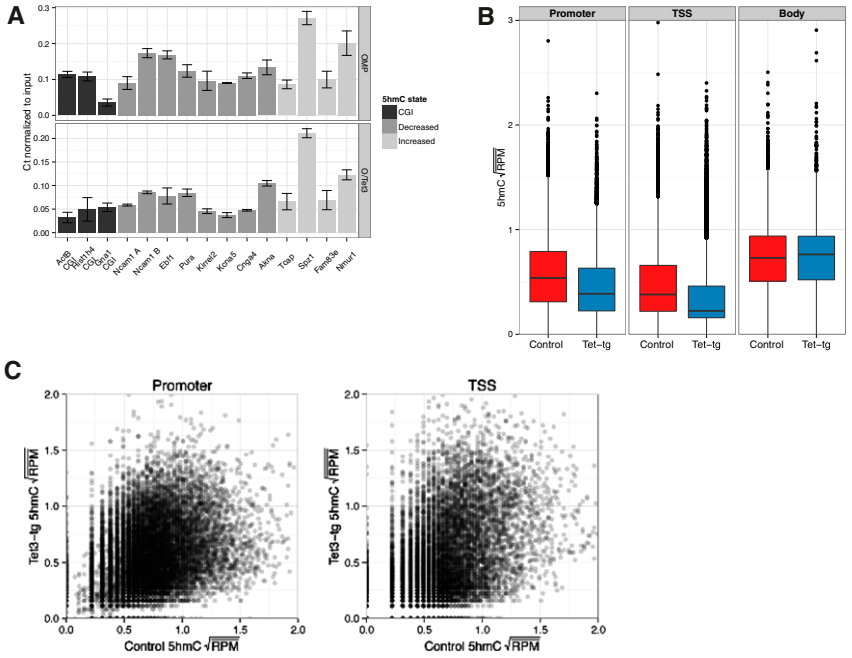

Fig. S4. (*A*) DIP-qPCR analysis of 5fC levels in Tet3-tg and control mOSNs. CGI, CpG island. Error bars are the range of qPCR replicates. (*B*) Distributions of 5hmC levels in control and Tet3-tg mONSs in promoters (−1 kb to TSS), TSSs (−500 bp to +500 bp flanking TSS), and gene bodies (TSS to TES). (*C*) Control versus Tet3-tg square-root mean RPM values in promoters and TSSs.

Dataset S1. Sequencing metadata

Dataset S1

Dataset S2. Developmentally regulated genes used in Fig. 3 *B–F*

Dataset S2

Dataset S3. Mean square-root 5hmC RPM values over gene bodies in control and Tet3-tg mOSNs

Dataset S3

Dataset S4. Log2(FPKM + 1) values from controls and Tet3-tg mOSN rRNA-depleted RNA-seq

Dataset S4