Supplementary material

Supplementary Figure Legends

Figure S1. Broad H3K4me3 stretches are present in different cell types and organisms, but are independent of signal intensity, promoter architecture, gene length and genomic location. (Related to Figure 1).

A) Skewness of H3K4me3 domain breadth distribution across eukaryotic species. Non-parametric skewness was calculated for 202 datasets from nine different species. Positive skew was observed for all studied datasets.

B) Venn Diagram of top 5% broadest H3K4me3 domain-marked genes in three mouse tissues reveals different genomic patterns for extreme H3K4me3 breadth between cell types. Overlaps were computed in a gene-centric manner using genes annotated to top 5% broadest H3K4me3 domains in mouse datasets liver, bone marrow and C2C12-derived myotubes (**Table S1**). See also Jaccard index heatmaps in **Figure 2C** and **S2D.**

C-D) Scatterplots of intensity of H3K4me3 peaks as a function of H3K4me3 breadth quantile in H1 human ESCs (B) and C2C12-derived myotubes (C). Intensity was measured as the normalized number of mapped ChIP reads per base pair for each peak determined by MACS2. Values were normalized to the sequencing depth of the ChIP sample and to the mapped read density of the cognate input sample at the same genomic locus. Note that above the 20^{th} percentile of breadth, there is no strong correlation between peak intensity and peak breadth ($R \sim 0.15$).

E) H3K4me3 domain breadth is similarly called by a variety of peak calling algorithms. Spearman rank correlation of corresponding peak breadth called by MACS2, HOMER, QuEST, CCAT and SICER. Significance of correlation was always $p < 2.2x10^{-308}$.

F) Circular genome plot showing the genomic distribution of the top 5% broadest H3K4me3 domains in H1 hESCs.

G) Boxplots representing the gene lengths in kb associated with various H3K4me3 domain breadths in H1 hESCs and C2C12-derived myotubes.

H-I) The top 5% broadest H3K4me3 domains are not associated with the presence of more used TSSs per gene, as seen using RNA-seq datasets (G) or PolII binding sites per marked gene (H). Significance of association in χ-square tests was > 0.05 in all cases.

J) Genes marked by the top 5% broadest H3K4me3 domains are expressed within a similar range as those marked by 0-95% broad H3K4me3 domains. FPKM from RNA-seq analysis in Cufflinks normalized to the median expression in each cell. No significant differences (n.s.) were observed between the top 5% broadest H3K4me3 domains and the rest of H3K4me3 domains ($p > 0.05$ in a Wilcoxon test). Experiment-derived NPC RNA-seq data is reported as the last pair in the boxplot (see **Figure 6E**).

Figure S2. The broadest H3K4me3 domains enrich for cell type/tissue-specific genes. (Related to Figure 2).

A-B) Example top enrichments from the GREAT annotation tool (McLean et al., 2010) for human CD34 primary cells, Brain Mid-frontal lobe tissue, and Skeletal muscle (A) or mouse Bone marrow, Cortex and Heart tissues (B), using the top 5% broadest H3K4me3 domains as input and the whole H3K4me3 domain set as background. p-values were derived from the annotation tool and reflect the Benjamini-Hochberg correction for multiple hypothesis testing.

C) The top 5% broadest H3K4me3 domains enrich for stem cell regulators regardless of H3K4me3 tag count at proximal promoters (-300bp, +300bp with respect to TSS). Roughly signal-matched control sets were randomly drawn 10,000 times from 0-95% broad H3K4me3 genes to assess the baseline enrichment expected from genes with similar promoter signal to the top 5% broadest H3K4me3 domains. Reported p-value applies to both H1 hESCs and mESCs and was obtained in non-parametric one-sample Wilcoxon tests comparing the 10,000 control enrichment values to the corresponding observed enrichment of genes marked by the top 5% broadest H3K4me3 domains.

D) Hierarchical clustering of mouse tissues/cells based on the Jaccard similarity index that measures the similarity in the genomic distribution of the top 5% broadest H3K4me3 domains.

E) Measure of cluster tightness (Silhouette index) from different sets of H3K4me3 domains in mouse tissues.

F-G) Measure of deviation from the 'gold standard' biological categorization (variation of information) from the Jaccard similarity index clustering for human (F) and mouse (G) tissues.

H-I) Hierarchical clustering of fibroblasts, ESCs, and iPSCs, based on Jaccard Index similarity in the genomic distribution of the top 5% broadest H3K4me3 domains. Clustering based on the top 5% set of peaks from each cell type in human (H) and in mouse (I). Note that the partially reprogrammed iPSC clone clusters apart from ESCs and fibroblasts in the mouse datasets.

J) Relevant functional enrichments of dynamic H3K4me3 domains (with >2-fold change in H3K4me3 breadth) in three differentiation paradigms. Enrichment statistics were

obtained using the complete set of H3K4me3 peaks as a background with the GREAT annotation tool.

K) Top 5% broadest H3K4me3 domains and 'super enhancers' (Whyte et al., 2013) capture different gene sets and genomic intervals in mESCs. Super enhancers genomic coordinates were obtained from the supplemental material of Whyte et al, 2013. Left panel: Venn Diagrams of genes marked by top 5% broadest H3K4me3 domains and super enhancers. Right panel: Venn Diagrams of genomic intervals defined by the top 5% broadest H3K4me3 domains and super enhancers.

L) Table summarizing the characteristics of the top 5% broadest H3K4me3 domains and super enhancers in mESCs. Ability of each signature to capture known regulators in mESCs is assessed using statistical measures: precision, sensitivity, F1 score, and enrichment statistics (Fisher's exact test) and the list of regulators from **Table S2**.

M) Comparison of clustering quality between top 5% broadest H3K4me3 domains and super enhancers in human tissues. For a fair comparison, clustering was assessed on the 13 tissues and cell types for which we could obtain H3K4me3 data and H3K27ac data. Super enhancers were mapped using the ROSE software on H3K27ac ChIP-seqs, as described in (Hnisz et al., 2013). Measure of the tightness of lineage clusters (Silhouette index) and of deviation from the 'gold standard' biological categorization (Variation of information) from the Jaccard similarity index clustering top 5% broad H3K4me3 domains or super enhancers. Note that because the set of tissues that are used here in this panel is different from that used in **Figure 2D** and **S2F**, the absolute values for these measures is different.

Figure S3. Use of the top 5% broadest H3K4me3 domains as a discovery tool. (Related to Figure 3).

A) Spearman rank correlations of H3K4me3 domain breadths. Upper panel: Spearman rank correlations in independent datasets of NPCs (NPC1, NPC2 and Ramos et al, 2013) and an NPC niche dataset. Lower panel: Spearman rank correlations between NPCs and unrelated cell types (quiescent hair follicle stem cells qHFSC, mESCs, C2C12 myotubes).

B) The top 5% broadest H3K4me3 domains enrich for neural progenitor cell regulators regardless of H3K4me3 tag count at proximal promoters (-300bp, +300bp with respect to TSS). Roughly signal-matched control sets were randomly drawn 10,000 times from 0-95% broad H3K4me3 genes to assess the baseline enrichment expected from genes with similar promoter signal as the top 5% broadest H3K4me3 domains. p-value in one-sample Wilcoxon tests comparing the 10,000 control enrichment values to the corresponding observed enrichment of genes marked by the top 5% broadest H3K4me3 domains.

C) Confirmation of knock-down of endogenous RNA levels after infection by cognate shRNA lentiviruses by RT-qPCR. Mean + SEM of RNA levels normalized to endogenous housekeeping gene *Gapdh* and corresponding control in at least two independent experiments, except for *Sox2* #1,3 and *Sall3* #3, *Sall1* #4.

D) Effect of individual hairpins targeting candidate genes marked by the broadest H3K4me3 domains in NPC proliferation. Normalized MTT optical density of replicates with respect to the cognate empty vector controls is shown. Mean + SD of replicates compiled from at least two independent experiments, except for *Zfp213* #1. Genes with

hashed blue bars are top 5% broadest H3K4me3 genes whose role in NPCs was discovered while this study was in preparation (Agoston et al., 2014; Ninkovic et al., 2013).

E) Knock-down of select genes marked by the broadest H3K4me3 domains impairs EdU incorporation in DNA. Mean + SD of percentage of EdU positive cells relative to the empty vector was calculated from > 1,000 cells from two independent experiments.

F) Effect of individual hairpins targeting candidate genes marked by the broadest H3K4me3 domains in NPC neurogenesis. Percentage of DCX positive cells normalized with respect to the cognate empty vector controls is displayed. Mean + SEM of replicates compiled from 2-4 independent experiments, except for *Otx1* #2. Over-expression of known NPC regulator *Ascl1* increases neurogenesis, as expected (Guillemot, 1999).

G) Possible use of the Buffer Domains database (http://bddb.stanford.edu).

Figure S4. Machine learning models reveal that the top 5% broadest H3K4me3 domains represent a distinct biochemical entity. (Related to Figure 4).

A) Detailed scheme of the classification models. H3K4me3 profiles were obtained in 13 benchmark cell lines.Random sampling of 5% sets of H3K4me3 domains from 0-95% of the H3K4me3 breadth distribution matched to the corresponding top 5% broadest domains from each cell line was performed 100 times, and co-occurrence of each H3K4me3 domain with other features (e.g. transcription factor binding, other histone modifications, DNA methylation, etc.) was determined after correcting for potential domain length bias. Classification algorithms (Random Forest, K-Nearest Neighbor [KNN], and Support Vector Machines [SVM] with either linear or Gaussian kernels)

were used to build predictive models. Then, these models were used to predict, based on the co-occurrence of features, whether H3K4me3 domains that were not used for training belonged to the top 5% broadest domains or to the rest of the breadth distribution. Accuracy of the prediction determined by comparing 'predicted' top 5% broadest vs. non top 5% H3K4me3 domains to their actual H3K4me3 breadth status.

B) Accuracy of classification models using 4 machine learning algorithms. Models are trained using 100 repetitions of the classification task with equal number of class xexamples (i.e. top 5% broadest vs. random 0-95% broad H3K4me3 domains).

C) Accuracy of classification models using Random Forest Algorithm, which are trained using either 100 repetitions of the top 5% broadest H3K4me3 domains vs. rest of distribution or randomly drawn sets of 0-95% broad H3K4me3 domains vs. other 0-95% broad H3K4me3 examples (non-overlapping domains). Note that the random vs. random classification, which is not expected to have any true signal, displays accuracy levels around the expectations for random classifiers (50% accuracy).

D) Accuracy of classification model using Random Forest Algorithm, which was trained using either 100 repetitions of the top 5% most intense H3K4me3 domains vs. rest of the distribution or randomly drawn sets of H3K4me3 domains examples (non-overlapping domains).

E) Significant contributors (features) whose values when disrupted produce more than a 1% decrease on average in the accuracy of the 100 models in each paradigm (using Random Forest). The number of inputted features in each cell line or organism is reported between parentheses.

F) Heatmaps of ChIP-seq intensity of features of the top 5% broadest H3K4me3 domain signature in H1 hESCs around TSSs, ranked by the breadth of associated H3K4me3 domains.

G) Differential binding of tissue-specific transcription factors and negative controls to top 5% broadest vs. 0-95% broad H3K4me3 domains. p-values of enrichment estimated using a simulated null distribution generated by 10,000 random samplings of 0-95% broad H3K4me3 domains whose breadth was adjusted to mimic the breadth distribution of the top 5% broadest H3K4me3 domains (to avoid length bias for intersections). ***: $p < 1x10^{-4}$; n.s.: non significant. The rightmost group shows enrichment of other chromatin binding proteins or nuclear domains that are not expected to be enriched for binding to the top 5% broadest H3K4me3 domains in mESCs (negative controls). LADs: Lamina-Associated Domains (UCSC Genome Browser track).

Figure S5. The top 5% broadest H3K4me3 domains are associated with regulators of PolII elongation and increased chromatin accessibility. (Related to Figure 5).

A) Heatmaps of ChIP-seq intensity of components of the elongation/pause-release machinery around TSSs in mESCs, ranked by the breadth of associated H3K4me3 domains. P-TEFb: Positive Transcription Elongation Factor b; NELF: Negative ELongation Factor; DSIF: DRB-Sensitivity Inducing factor; PAF: Polymerase Associated Factor; SEC: Super Elongation Complex.

B,D) Differential binding of components of the elongation machinery to top 5% broadest vs. 0-95% broad H3K4me3 domains in HCT-116 (B) or 293T (D) cells. p-values of enrichment estimated using a simulated null distribution generated by 10,000 random

samplings of 0-95% broad H3K4me3 domains whose breadth was adjusted to mimic the top 5% broadest H3K4me3 domain breadth distributions (see Extended Experimental Procedures).

C,E) The set of top 5% broadest H3K4me3 domains maximally enriches for components of the elongation machinery in HCT-116 (C) or 293T (E) cells. Enrichments expressed as a percentage of the maximal enrichment that can be observed for differential binding (see panels **B,D**).

F) Mean ChIP-seq enrichment of PolII in 293T cells. TSS: transcription start site; TTS: transcription termination site.

G) Normalized PolII ChIP-seq density over the proximal promoter and gene region. Indicated p-values of significance from the top 5% broadest H3K4me3 domain associated-genes were calculated using one-sided one-sample Wilcoxon tests against the expected genome-wide value obtained by the mean of 10,000 random samplings (red dashed line). Comparisons of top 5% broadest H3K4me3 domains against the rest of the distribution also significant in Wilcoxon tests ($9.6x10^{-10} < p < 5.4x10^{-3}$) (continued from **Figure 5D**).

H) Mean ChIP-seq enrichment of initiating PolII (Ser5P) in mESCs. TSS: transcription start site; TTS: transcription termination site.

I) Normalized initiating PolII (Ser5P) ChIP-seq density over the proximal promoter. p-values of significance for the top 5% broadest H3K4me3 domain associated-genes calculated using one-sided one-sample Wilcoxon tests against the expected genome-wide value obtained by the mean of 10,000 random samplings (red dashed line). Comparisons

of top 5% broadest H3K4me3 domains against the rest of the distribution also significant in Wilcoxon tests ($1.0x10^{-19} < p < 1.3x10^{-7}$)

J) Normalized PolII- Ser2P ChIP-seq density over gene bodies. p-values of significance from the top 5% broadest H3K4me3 domain associated-genes calculated using one-sided one sample Wilcoxon tests against the expected genome-wide value obtained by the mean of 10,000 random samplings (red dashed line). Comparisons of top 5% broadest H3K4me3 domains against the rest of the distribution also significant in Wilcoxon tests ($7.3x10^{-23} < p < 2.6x10^{-5}$) (continued from **Figure 5H**).

K,L) Broad H3K4me3 domains mark genes with increased chromatin accessibility at promoters as assessed by DNAse-seq (K) or ATAC-seq (L). Normalized number of nick/insertions sites in the -300,+300bp region around TSSs. p-values of significance for the top 5% broadest H3K4me3 domain associated-genes calculated using one-sided one-sample Wilcoxon tests against the expected genome-wide value obtained by the mean of 10,000 random samplings (red dashed line). Comparisons of top 5% broadest H3K4me3 domains against the rest of the distribution also significant in Wilcoxon tests (DNAse-seq: $5.5x10^{-194} < p < 4.0x10^{-2}$; ATAC-seq: $p = 3.5x10^{-38}$).

Figure S6. H3K4me3 breadth is associated with increased transcriptional consistency. (Related to Figure 6).

A) Significance for lower transcriptional variability in single-cell RNA-seq datasets against the expected transcriptome-wide value expressed as -log10(p-value) in one-sided Wilcoxon tests.

B) Transcriptional variability at the cell population level (steady state mRNA) (Continued from **Figure 6C**). p-values in one-sample Wilcoxon tests between the top 5% broadest H3K4me3 domains associated genes and the expected value from transcriptome-wide samplings obtained by the mean of 10,000 random samplings (red dashed line). Comparisons of top 5% broadest H3K4me3 domains against the rest of the distribution also significant in Wilcoxon tests ($7.1x10^{-174} < p < 2.9x10^{-3}$).

C) Transcriptional variability at the cell population level (steady state mRNA) after excluding genes marked by bivalent/poised domains (H3K4me3 domains with any overlap with H3K27me3 domains). p-values in one-sample Wilcoxon tests between the top 5% broadest H3K4me3 domains associated genes and the expected value from transcriptome-wide samplings obtained by the mean of 10,000 random samplings (red dashed line). Comparisons of top 5% broadest H3K4me3 domains against the rest of the distribution also significant in Wilcoxon tests ($1.8x10^{-10} < p < 2.3x10^{-9}$).

D) Transcriptional variability at the cell population level (steady state mRNA) after matching of H3K4me3 tag count at proximal promoters (-300bp, +300bp with respect to TSS). Roughly signal-matched control sets were randomly drawn 10,000 times from 0-95% broad H3K4me3 genes to assess the baseline scaled variance expected from genes with similar promoter signal as the top 5% broadest H3K4me3 domains. Mean scaled variance in control sets reported alongside to mean scaled variance of top 5% broadest H3K4me3 domain associated genes. p-values in one-sample Wilcoxon tests comparing the 10,000 control scaled variance values to the corresponding scaled variance of genes marked by the top 5% broadest H3K4me3 domains.

E) Significance for lower transcriptional variability in nascent RNA GRO-seq datasets against the expected transcriptome-wide value expressed as -log10(p-value) in one-sided Wilcoxon tests.

Figure S7. Effect of H3K4me3 regulators on H3K4me3 breadth and transcriptional consistency. (Related to Figure 7).

A) Differential binding of subunits of the COMPASS/Trithorax/Trithorax-related complex and of JARID1 H3K4me3 demethylases to top 5% broadest vs. 0-95% broad H3K4me3 domains. p-values of enrichment estimated using a simulated null distribution generated by 10,000 random samplings of typical domains whose length was adjusted to mimic the length distributions of the top 5% broadest domains.

B) Western Blot analysis of NPCs treated with Control (empty vector), *Luciferase* shRNA or two independent *Wdr5* shRNA lentiviral constructs after 24h of infection (see also **Figure 7B**). The construct that was used for all genome-wide studies is shRNA #2, denoted elsewhere as "*Wdr5* shRNA".

C) Knock-down of *Wdr5* mRNA in NPCs after 24h of shRNA lentiviral infection measured by RT-qPCR in 2 independent experiments (related to the ChIP-seq samples).

D) *Wdr5* knock-down does not significantly affect NPC viability at 24h post infection. Mean + SD of percentage viable cells (Propidium Iodide negative in Flow Cytometry) from 6 replicates infections in 2 independent experiments. n.s.: not significant in a Wilcoxon test against empty vector control.

E,F) Coverage histograms of H3K4me3 enriched regions in *Wdr5* knock-down samples by down-sampled sets of reads from empty vector control H3K4me3 ChIP-seq in NPCs

ChIP-seq replicates 1 (E) and 2 (F). Coverage histogram of the H3K4me3 ChIP-seq upon *Wdr5* knock-down in black. Graphical matching of coverage histograms (so that the coverage of regions is equal or less in control compared to *Wdr5* knock-down) yields 65% of original depth for replicate 1 and 45% of original depth for replicate 2.

G) Mean loss of H3K4me3 breadth from 2 biological H3K4me3 ChIP-seqs replicates in NPCs following *Wdr5* knock-down. Amount of H3K4me3 breadth lost upon *Wdr5* knock-down normalized to the original breadth of the corresponding marked region in control. The red line denotes the expected value of percentage breadth lost genome-wide (average of the whole experiment).

H) Volcano plot of empty vector versus *Wdr5* knock-down RNA-seq experiments (biological triplicates). The tuxedo suite (Trapnell et al., 2009; Trapnell et al., 2010) was used to estimate mRNA expression levels and differential expression p-values and FDR.

I) Knock-down of *Jarid1b* mRNA in mESCs after 48h of shRNA lentiviral infection from Affymetrix microarray data (Schmitz et al., 2011).

J) Coverage histograms of H3K4me3 enriched regions in control samples at matched levels of signal-to-noise ratio (down-sampling higher coverage control sample to 40% of depth) H3K4me3 mESCs ChIP-seq.

K) Volcano plot of scramble versus *Jarid1b* knock-down Affymetrix array experiments (Schmitz et al., 2011) (biological triplicates). SAM (Tusher et al., 2001) was used to compute differential expression raw p-values and FDR.

L) H3K4me3 breadth remodeling upon *Jarid1b* knock-down and gain of transcriptional consistency in mESCs. Upper panel: Significance for lower variance against the expected transcriptome-wide value obtained by the mean of 10,000 random samplings (expressed

as -log10 (p-value) in one-sided Wilcoxon test). Lower panel: Significance for decreased variability of genes gaining H3K4me3 breadth against genes of the same original H3K4me3 quantile with constant breadth expressed as -log10(p-value) in one-sided Wilcoxon test. Red dashed line: $p = 0.05$.

Inventory of Supplementary files

Table S1. Accession numbers and dataset identifications of all ChIP-seq or ChIP-chip datasets that were used in this study. (related to **Figure 1, 2, 4, 5, 6, 7, S1, S2, S4, S5, S6** and **S7**).

Table S2. List of genes (Entrez Gene ID) for the 'stem cell regulators' curated from the literature or identified by RNAi screens in ESCs and for NPC regulators curated from the literature. (related to **Figure 2, 3, S2** and **S3**).

Table S3. Known reprogramming and identity factors. (related to **Figure 2E**).

Table S4. Annotated H3K4me3 domains in adult mouse NPC datasets. (related to **Figure 3** and **S3**).

Table S5. Example annotated top 5% broadest H3K4me3 domains in hESCs, skeletal muscle, mESCs, cortex and C2C12 myotubes. (related to **Figure 1, 2, S1, S2** and **S3**).

Table S6. Accession numbers of RNA-seq, microarray, and GRO-seq datasets that were used in this study. (related to **Figure 1, 6, 7, S1, S6** and **S7**).

Table S7. Genes whose expression level was affected by 24h of WDR5 knock-down in adult NPCs. (related to **Figure 7** and **S7**).

Extended experimental procedures

ChIP-seq and ChIP-chip data analysis pipeline

Publicly available ChIP-seq and ChIP-chip datasets were obtained from the following repositories: ENCODE (http://genome.ucsc.edu/ENCODE/) (Dunham et al., 2012), Roadmap Epigenomics (http://www.roadmapepigenomics.org/) (Hawkins et al., 2010; Zhu et al., 2013), GEO (http://www.ncbi.nlm.nih.gov/gds), ArrayExpress/EBI (http://www.ebi.ac.uk/arrayexpress/) or SRA (http://www.ncbi.nlm.nih.gov/sra). Corresponding accession numbers and references for all analyzed datasets are available in **Table S1**. For ChIP-seq datasets, sequence reads were quality filtered to retain only those with a minimum Phred score of 15 over 85% of the read length. Datasets with small number of reads after quality filtering were discarded from further analyses (< 7 million reads). Reads were mapped to corresponding reference genome builds (hg19, mm9, tair10, ce6, dm3, saCer3, xenTro2) using bowtie0.12.7 (Langmead et al., 2009) software. ChIP-seq peaks were called using the MACS2.08 software (Feng et al., 2012; Zhang et al., 2008) with default settings, adding the "--broad option" for histone marks. When available, input datasets were used during peak calling as controls. Datasets with too few peaks were discarded as low quality (< 1000 peaks). For comparison purposes, we also used three other peak calling algorithms (HOMER (Heinz et al., 2010), CCAT (Xu et al., 2010a), QuEST (Valouev et al., 2008) and SICER (Zang et al., 2009)) to detect H3K4me3 domains from the ENCODE H1 hESCs dataset.

C. elegans early embryo and *D. melanogaster* S2 cells ChIP-Chip datasets (and corresponding RNA-seq datasets) were obtained from ModENCODE data coordination

center (http://www.modencode.org/) (Muers, 2011). The worm data comes from Roche Nimblegen tiling arrays, and ChIP-Chip peaks were called using the MA2C algorithm (Song et al., 2007). Genomic coordinates of the called peaks were lifted over to the ce6 genome build. For S2 cells, we used called peak files - using the MAT peak caller (Johnson et al., 2006) - from the Affymetrix tiling array datasets generated by the ModENCODE consortium.

All statistically significantly enriched regions (aka peaks) obtained from ChIP-seq and ChIP-chip datasets were annotated to genes using the HOMER suite (Heinz et al., 2010). HOMER attributes a peak to the gene with the closest transcription start site. H3K4me3 peaks that were longer than the 95th percent of all H3K4me3 peaks were termed as top 5% broadest H3K4me3 domains. A MySQL database was created to catalogue all genes detected in association to top 5% broadest H3K4me3 domains in human, mouse and *C. elegans* samples. This database has been made directly available and searchable online at http://bddb.stanford.edu (see **Table S5** and **Data S1**).

Signal profiles of ChIP-seq and gene length calculations

Meta-gene profiles for PolII datasets, which is defined as the average profile of PolII ChIP-seq signal on all genes normalized to the same length (to a 'meta-gene'), were computed using the normalized tag counts over genomes and the CEAS software (Shin et al., 2009). To calculate the length of genes associated with the top 5% broadest H3K4me3 domains, gene definitions from Ensembl.v69 were used.

Skewness analysis

Non-parametric skewness of H3K4me3 peak length distributions was calculated as $Skewness = \frac{Median - Mean}{SD}$. In a symmetric distribution, skewness is 0. A positive value indicates right skew.

Functional annotations of H3K4me3 peaks

Functional term enrichment analysis was conducted by comparing genes marked by broad H3K4me3 domains to all H3K4me3 domains as background using GREAT (McLean et al., 2010) and DAVID (Huang da et al., 2009a, b; Huang da et al., 2009c) portals. For GREAT analysis, genomic coordinates of peaks (in the form of bed files) we used. Whereas for DAVID analysis, gene lists annotated by the HOMER software were used.

Functional enrichments of gene sets associated with increasing breadth of H3K4me3 domains

To assess functional enrichments with respect to the breadth of H3K4me3 domains, we used either a sliding window approach comparing genes sets associated with 5% of peaks, sliding on decreasing length per one peak basis (**Figure 2A**), or a 5% quantile binning approach based on sorted H3K4me3 domain breadth (**Figure 2B**). Enrichments

were computed for each set of genes associated with a different subset of H3K4me3 domains using one-tailed Fisher exact tests. Following gene lists were used for enrichment analysis. As 'stem cells' gene-set for mouse or human embryonic stem cells, we used the combined list of validated ‘stem cell regulators’ – genes involved in ESC pluripotency and self-renewal curated from experimental work (Xu et al., 2010b) or unbiased genome-wide screens (Abujarour et al., 2010; Chia et al., 2010; Hu et al., 2009; Westerman et al., 2011; Zhang et al., 2006) (reported in **Table S2**). For the human brain data, we compiled the list of genes annotated with GO terms; GO:0030182 (neuron differentiation), GO:0007409 (axonogenesis), GO:0048812 (neuron projection morphogenesis), GO:0048667 (cell morphogenesis involved in neuron differentiation), GO:0048666 (neuron development), GO:0031175 (neuron projection development), and GO:0007411 (axon guidance). For the human skeletal muscle data, we compiled the list of genes annotated with GO terms; GO:0061061 (muscle structure development), GO:0007525 (somatic muscle development), GO:0030016 (myofibril), GO:0006941 (striated muscle contraction), GO:0006936 (muscle contraction), GO:0003012 (muscle system process), GO:0007517 (muscle organ development). The NPC regulator list was generated based on literature curation in addition to genes associated with the GO terms listed in **Table S2**. Interestingly, 10 out of 13 known reprogramming factors to directly convert fibroblasts into induced-neural stem cells (iNSCs) (Han et al., 2012; Lujan et al., 2012; Sheng et al., 2012; Thier et al., 2012) were marked by the broadest H3K4me3 domains in NPCs. For the mouse heart data, we compiled the list of genes annotated with GO terms; GO:0001944 (vasculature development), GO:0001568 (blood vessel development), GO:0048514 (blood vessel morphogenesis), GO:0001525 (angiogenesis),

GO:0007507 (heart development), GO:0048738 (cardiac muscle tissue development), GO:0003007 (heart morphogenesis), GO:0048738 (cardiac muscle tissue development). For genes associated with embryonic development of *Caenorhabditis elegans*, we compiled all genes associated with GO:009790 (embryo development). For the *Arabidopsis thaliana* data, we compiled the list of genes annotated with GO term; GO:0015979 (Photosynthesis).

Cell identity signature comparison

To compare the top 5% broadest H3K4me3 domain signature to the super enhancer signature (Whyte et al., 2013), genes marked by each signature were identified and compared to the known stem cell regulators in mESCs. Statistical measures of performance, *i.e.* precision, sensitivity, and F1 score were used. Precision quantifies the fraction of known regulators among genes that are marked by the signature. Sensitivity quantifies the fraction of stem cell genes captured by the signature among all known stem cell regulators. F1 score measures the harmonic mean of precision and sensitivity. For each signature, enrichment statistics for known stem cell regulators (from **Table S2**) were calculated using Fisher's exact test.

Gene set enrichment analysis (GSEA)

Gene set enrichment analysis (GSEA) were conducted on pre-ranked gene lists generated by ranking genes with respect to the breadth of their H3K4me3 domains. As 'stem cell'

gene-set for embryonic stem cells, we used the combined list of published curated list and unbiased RNAi screens listed in **Table 2.** Enrichment statistics were calculated using the 'classic' method. Nominal p-values were calculated based on 10,000 permutations. Note that a p-value of zero indicates that the p-value is smaller than 1/(number of permutations), which we reported as $p < 1x10^{-4}$.

Tissue identity clustering

To measure similarity between tissues in terms of shared subsets of H3K4me3 domains, we used the Jaccard similarity index (obtained by dividing the number of shared elements between two sets by the number of all unique elements in their union), which captures the proportion of H3K4me3 peaks that are shared between two datasets. Using all H3K4me3 datasets in human (and mouse) cells/tissues, we constructed a symmetric similarity matrix based on the presence of intersecting regions using bedtoolsv2.16.1 (Quinlan and Hall, 2010). This distance matrix was clustered using 'absolute' correlation with hierarchical clustering (complete linkage) algorithm from the 'pheatmap' R package. To quantify the quality of clustering, we used two separate measures, silhouette index and variation of information as defined in the 'fpc' R package. Silhouette index measures cluster tightness (higher values imply tighter clustering). Variation of information measures the distance between a hierarchical clustering output and the gold standard categorization of known biological lineages.

Analysis of effect from H3K4me3 signal at promoters or along significant peaks

H3K4me3 ChIP-seq reads were mapped to the appropriate genome assembly. For analyses requiring matching ChIP-seq signal at proximal promoters, pile-up of ChIP-seq reads mapped within the proximal promoter (-300bp;+300bp with respect to TSS) was computed as the coverage in the specified intervals using the default settings of coverageBed function of bedtools software and annotations from the cognate assembly downloaded from UCSC genome browser. We defined three levels of signal for this analysis: low (0-25th percentiles of signal distribution), medium (25-75th percentiles of signal distribution) and high (75-100th percentiles of signal distribution). We then obtained 10,000 random draws roughly matched in signal to the promoters associated with the top 5% broadest H3K4me3 domains (same number of "high" signal promoters, same number of "medium" signal promoters, and same number of "low" signal promoters). We then compared the enrichments obtained from these random samplings to the true enrichment associated with the top 5% broadest H3K4me3 domains using a one-sided one-sample Wilcoxon text. For transcriptional variability, we obtained the mean scaled variance of each sampling and compared the simulated mean scaled variances to the observed mean scaled variance associated with the top 5% broadest H3K4me3 domains.

For analyses on the ChIP-seq signal intensity at enriched regions, pile-up of ChIP-seq reads mapped within the significant MACS2 peak was computed as the coverage in the specified intervals using the default settings of coverageBed function of bedtools suite, and normalized to the input signal in the same interval. Then, signal was normalized to peak breadth to obtain tags per bp of peak.

Mouse NPC Cultures

Adult (3-4 month-old) C57BL/6 (NPC1 dataset) or FVBN (NPC2 dataset) mouse NPCs were isolated as previously described (Palmer et al., 1997; Pastrana et al., 2009). Briefly, the subventricular zone was finely microdissected and chopped into ice-cold PBS. Tissue chunks were digested by 30-40 min incubation in HBSS (Invitrogen) with 1 U/ml DispaseII (Roche), 250 U/ml DNaseI (Sigma) and 2.5 U/ml Papain (Worthington) at 37^{o}C. Following mechanical tituration, cells were purified by sequential 25% and 65% Percoll (Amersham) gradients. NPCs were plated at a density of $<10^5$ cells/cm^2 as non-adherent spheres in NPC media (Neurobasal A (Invitrogen) medium supplemented with 1% penicillin/streptomycin/glutamine (Invitrogen), 2% B27 (Invitrogen) and 20 ng/ml each of FGF2 (Peprotec) and EGF (Peprotec)). Cells were routinely passaged using Accutase enzyme (Stem Cell Technologies, 07920).

H3K4me3 ChIP-seq in NPCs and NPC niches

ChIP experiments on mouse NPC cultures were performed as previously described (Webb et al., 2013). Briefly, NPC neurospheres (passage 2-3) were dissociated 12 hours prior to collection. 1,000,000-1,500,000 cells were crosslinked with 1% formaldehyde for 9 min at room temperature and the reaction was quenched with 0.125M glycine for 5 min at room temperature. For ChIP experiments on the NPC niche, SVZs were microdissected from 3 month-old C57BL/6 mice. SVZs were pooled and minced manually prior to

crosslinking, and 150 mg tissue was used per ChIP. Cells/Tissue were washed twice with cold 1X PBS and resuspended in 700 □l of SDS lysis buffer (50 mM Tris-Hcl pH7.5, 10 □M EDTA, 1% SDS in PBS pH7.4). Chromatin was sheared with a Vibra-Cell Sonicator VC130 (Sonics) 7 times for 30 sec at 60% amplitude and diluted 1:5 fold in RIPA buffer (1% NP-40, 0.5% sodium deoxycholate in PBS pH 7.4). Chromatin was immunoprecipitated with 5 µl of H3K4me3 antibody (Active Motif antibody 39159). Libraries were generated according to Illumina instructions and PCR amplified for 18-19 cycles. Library quality was assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies). 34-36 bp reads were generated on an Illumina Genome Analyzer II and subsequently analyzed with our standardized ChIP-seq data analysis pipeline.

Lentiviral constructs for gene knock-down in primary NPCs

shRNA hairpin sequences were cloned into either H1-FUGW (Fasano et al., 2007) or were purchased in the lentiviral vector backbone (PLKO.1) (Sigma Aldrich). H1-FUGW contained a GFP expression cassette, allowing for visualization of infected cells. PLKO.1 plasmid includes a puromycin resistance cassette, allowing for selection of infected cells at 0.5µg/ml puromycin (Invivogen). In all experiments, data were normalized to their respective empty vector control. For over-expression of *Ascl1*, constructs were obtained in TetO-FUW backbone (see Webb et al., 2013). See below for the shRNA hairpin sequences.

H1-FUGW	Target Sequence	Primer 5' to 3'	Targeted region
Sall3 #1	GAACTCTGCAACCTTTAAA	GAACTCTGCAACCTTTAAATTCAA GAGATTTAAAGGTTGCAGAGTTCT TTTTTGT	3UTR
Sall3 #2	GGCTCTCATTAATACTTAA	GGCTCTCATTAATACTTAATTCAA GAGATTAAGTATTAATGAGAGCCT TTTTTGT	3UTR
Sall1 #1	GCAAATACGTCACCAAATA	GCAAATACGTCACCAAATATTCAA GAGATATTTGGTGACGTATTTGCT TTTTTGT	CDS
Sall1 #2	GTATGTTGTTCAACCTCTA	GTATGTTGTTCAACCTCTATTCAA GAGATAGAGGTTGAACAACATAC TTTTTTGT	3UTR
Meis1 #1	GAAGCCTCCTTACATTAAA	GAAGCCTCCTTACATTAAATTCAA GAGATTTAATGTAAGGAGGCTTCT TTTTTGT	3UTR
Meis1 #2	GTGAACAATTGGTTTATTA	GTGAACAATTGGTTTATTATTCAA GAGATAATAAACCAATTGTTCACT TTTTTGT	CDS
Bmi1 #1	ATATGGATGTTAAGTGGAA	ATATGGATGTTAAGTGGAATTCAA GAGATTCCACTTAACATCCATATT TTTTTGT	3UTR
p53 #1	GTACTCTCCTCCCCTCAAT	GTACTCTCCTCCCCTCAATTTCAA GAGAAT TGAGGGGAGGAGAGTACTTTTTTG T	CDS
Sp5 #1	GGAGCTTTGTGGATTCAAA	GGAGCTTTGTGGATTCAAATTCAA GAGATTTGAATCCACAAAGCTCCT TTTTTGT	3UTR
6530411M01Rik #1	CAAAGCAGCTTGAAGTTAA	CAAAGCAGCTTGAAGTTAATTCAA GAGATTAACTTCAAGCTGCTTTGT TTTTTGT	CDS

PLKO.1	TRC Clone #	Target Sequence	Targeted region
2610017I09Rik #1	TRCN0000179216	GAATAACTGCCATGGAAGGAT	NA
2610017I09Rik #2	TRCN0000179454	GCCTTCATCAAGTGGTATGAA	NA
Bahcc1 #1	TRCN0000177458	CGTATCTCTTACCTCTGTTTA	CDS
Bahcc1 #2	TRCN0000181431	CGGACTTCAAGATCCAGTGTA	CDS
Bahcc1 #3	TRCN0000181541	GAAGCGAAGCAAACTGGGAAA	CDS
Fam72a #1	TRCN0000177185	GCGATTTCAAATCAATGACTT	3'UTR
Fam72a #2	TRCN0000176612	CCTGTGTTGCAAATTCTGTAA	CDS
Gtl3 #1	TRCN0000108570	GCACGCAACAAACGGAAGAAT	3UTR
Gtl3 #2	TRCN0000108571	CCGTATCCGAAGGGTTTACTT	CDS
Irs1 #1	TRCN0000105881	CGAGACGAACACTTTGCCATT	CDS
Irs1 #2	TRCN0000105882	CCCAGGAGAATATGTGAATAT	CDS
Irs2 #1	TRCN0000055108	CGAGTACATCAACATTGACTT	CDS
Irs2 #2	TRCN0000055109	CCCGAACCTCAATAACAACAA	CDS
Luciferase	NA	CACTCTGATTGACAAATA	NA
Nfib #1	TRCN0000012089	CCATTTATTGAGGCACTTCTT	CDS
Nfib #2	TRCN0000012090	CCTTCCAGCTACTTCTCTCAT	CDS
Nr2f1 #1	TRCN0000026160	GAGCAGTTTCAACTGGCCTTA	CDS
Nr2f1 #2	TRCN0000026161	GCTACCTGTCTGGCTACATTT	CDS
Otx1 #1	TRCN0000085323	GCGTCCAAGAAACAGAACTTT	3UTR
Otx1 #2	TRCN0000085324	CCGTATCTAGCTCTGCTTCTT	CDS
Meis2 #1	TRCN0000075588	GCACCATAAGTAGGATTCTAT	3UTR
Meis2 #2	TRCN0000075589	CCACGAACTATGTGATAACTT	CDS
Sall1 #3	TRCN0000098342	GCTGCGCTGAATTCTTTGAAT	CDS
Sall1 #4	TRCN0000098343	GCACTATCTGTGGAAGAGCAT	CDS
Sall3 #3	TRCN0000097901	CGCGAGGTTCATTGAGGATAA	CDS
Setd1b #1	TRCN0000095449	CCCATCCTCTTCAGGGTTAAT	3'UTR
Setd1b #2	TRCN0000095450	CCAGTGAAAGTTCTGGATCTT	CDS
Sox2 #1	TRCN0000424718	AGGAGCACCCGGATTATAAAT	CDS
Sox2 #2	TRCN0000420955	ACCAATCCCATCCAAATTAAC	3'UTR
Sox2 #3	TRCN0000416106	CAAAGAGATACAAGGGAATTG	3'UTR
Sox2 #4	(This study cloning)	GAAGGAGCACCCGGATTAT	CDS
Spry4 #1	TRCN0000065934	CCACTCACCATCTTACCCATT	CDS
Spry4 #2	TRCN0000065935	GCCCGCTGTGACCAGGATATT	CDS
Sp5 #2	TRCN0000084558	CGGACTTTGTACAGGTTATTT	3UTR
Sp5 #3	TRCN0000084559	CCCGTCGGACTTTGCACAGTA	CDS
Trnc18 #1	TRCN0000181383	CAGGCCCTGTTCACAGATATT	CDS
Trnc18 #2	TRCN0000256218	TCACGACTCCTCATCTGATTT	CDS
Srgap1 #1	TRCN0000106110	CCACTGCAGAACTCCAGAAAT	3'UTR
Srgap1 #2	TRCN0000106111	GCTGGCTCTAGGTTTCCATAT	CDS
Wdr5 #1	TRCN0000034415	GCAGCGTTAGAGAACGACAAA	CDS
Wdr5 #2	TRCN0000034416	GCCGTTCATTTCAACCGTGAT	CDS
Wdr74 #1	TRCN0000124370	CCCTTATCACATGTGTGGATT	CDS
Wdr74 #2	TRCN0000124371	CCCAACCAAGTACCCTCAGAA	CDS
Zfp110 #1	TRCN0000095460	GCCCTGATTCAATCCCTCTAT	CDS
Zfp110 #2	TRCN0000095461	GCAGCCCAATACACGTTCAAA	CDS
Zfp110 #3	TRCN0000329579	GACACATGCCTTACCTGTAAA	CDS
Zfp213 #1	TRCN0000239964	TACAGAATCAGTCGTTGAAAG	CDS
Zfp213 #2	TRCN0000239968	TTTGAGGATCACGTGATATTT	3'UTR
Zfr #1	TRCN0000085277	CTACCCAAACAGCTTGCTGTT	CDS
Zfr #2	TRCN0000085276	CCACAAATGAATCAGCGCTTT	CDS

Lentivirus production

HEK293T cells were plated at a density 50,000-80,000 cells/cm^2 in DMEM (Gibco) supplemented with penicillin/streptomycin/glutamine (Invitrogen), 100mM sodium pyruvate (Invitrogen) and 10% fetal bovine serum. For production of H1-shRNA-FUGW virus, HEK293Ts were transfected with 10 µg of H1-FUGW shRNA plasmid; 5 µg each of helper plasmids pCMV-dR8.91 and HCMVG. For production of shRNA-PLKO.1 virus, HEK293Ts were transfected with 10 µg of pLKO.1 plasmid, 7.5 µg of helper plasmid psPax2 and 2.5 µg of helper plasmid pMD2G. After 12-16 hours, plates were rinsed with 1X PBS and media was replaced with NPC media (Neurobasal A (Invitrogen) medium supplemented with penicillin/streptomycin/glutamine (Invitrogen), 2% B27 (Invitrogen)). NPC media containing virus was collected 48 hours post transfected and filtered with 0.45 µm low protein binding membranes (Millex-HV, Millipore). For MTT and neurogenesis assays, virus was produced as aforementioned in 6 well plates by transfection of 1.25 µg of shRNA plasmid, 0.94 µg of psPax2, and 0.32 µg of pMD2G.

Verification of knock-down by RT-qPCR

RNA was extracted using RNAeasy Kit (Qiagen) according to the manufacturer's protocol and cDNA was produced using the High Capacity cDNA Reverse Transcription Kit (Ambion). Real-time PCR experiments were performed using iQ SYBR green Supermix (BioRad) and the BioRad C1000 Thermal Cycler.

Primer Sequences used were:

2610017L09Rik-Forward: ATGGCCGTTCCTAACTTTGAAG;

2610017L09Rik-Reverse: GGTCTTATCCGCCTTACAGTCC;

Bahcc1-Forward: GCGTACCCCAGATTTTCGGG;

Bahcc1-Reverse: GAAACGATGTTGCCCATAGAGAA;

Fam72a- Forward: TTTCAAAGACCGATGCGTATCC;

Fam72a- Reverse: CTATGTCAGTATCAGCCAGCAAA**;**

Meis1-Forward: TTGCTTCAGGTCCGGTAGAC;

Meis1-Reverse: TGCCTACTCCATCCATACCC;

Otx1-Forward; ATGTCTTACCTCAAACAACCCCC;

Otx1-Reverse; GTAGCGAGTCTTTGCGAACAG;

p53-Forward: GCCATGGCCATCTACAAGAA;

p53-reverse: CTCGGGTGGCTCATAAGGTA;

Sall1-Forward: TGATGTTTGAGCCAGCATGT;

Sall1-Reverse: GCAGCTCTTTTTATGGAGCA;

Sall3- Forward: TGCTGTTCCTGAGCAGAGAG;

Sall3-Reverse: GCCGTTCACTTCCATTTTGA;

Sox2-Forward: AAGGGTTCTTGCTGGGTTTT;

Sox2-Reverse: AGACCACGAAAACGGTCTTG.

Wdr5-Forward: ATGGGCAGGCAAAGTCTTGAG;

Wdr5-Reverse: TTTGAAGATTTGGGACGTGAGTT;

Zfp213-Forward: AGGAGAGGTGTTGGATGGCT;

Zfp213-Reverse: GAATGGGGAGATACGACCCAG;

Zfr-Forward; GCGACCGGCAACTACTTTG;

Zfr-Reverse; GATGGGAATAGGCTACACCCG;

Gapdh-Forward: TGTGTCCGTCGTGGATCTGA;

Gapdh-Reverse: TTGCTGTTGAAGTCGCAGGAG.

For knockdown of non-coding RNA *2610017L09Rik*, SYBR green results were confirmed using gene specific Taqman assays purchased from Life Technologies: Gapdh (Mm99999915_g1) and 2610017L09Rik (Mm00806301_m1).

NPC proliferation assay

Primary NPCs (passages lower than 6) were plated at a density of 8000 cells/cm^2 in 96 well plates pre-coated with 50 □g/ml of poly-D-lysine (Sigma-Aldrich) and 5 □g/ml of laminin (Invitrogen). Five hours post-plating, adherent cells were transduced with a 30% dilution of lentiviral supernatant in fresh NPC proliferation media. After 24 hours, media was replaced with fresh media containing 0.5 □g/ml of puromycin (Invivogen) and media was changed every 2 days. Four days after infection, the resulting number of cells was quantified by 3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) (Behar et al., 2012) (Vybrant MTT Cell Proliferation Assay, Molecular Probes), according to the manufacturer's protocol. Optical density readings at 570 nm were taken using the Tecan Infinite 200 Pro 96 well plate reader.

Neurosphere assay

Primary NPCs (passages 2-3) were plated adherently at a density of 30,000 cells/cm^2 on wells pre-coated with 50 mg/ml poly-D-lysine and infected with 50% lentiviral supernatant media in fresh NPC proliferation media for 24 hours. After 24 hours, media was removed and replaced with fresh NPC proliferation media (Neurobasal A (Invitrogen) medium supplemented with penicillin/streptomycin/glutamine (Invitrogen), 2% B27 (Invitrogen), 20ng/ml each of FGF2 (Peprotec) and EGF (Peprotec)). Seventy-two hours post infection, cells were removed from plates using Accutase (Stem Cell Technologies) and re-plated in triplicate wells at low density (1 cell/ml). After 6 days, the number of GFP positive spheres (>40 μm diameter) was counted and the percentage of cells forming spheres out of 1000 was calculated and normalized to the empty vector control. Values are reported as the normalized percentage of sphere-forming cells per well for replicates coming from at least two independent experiments (*i.e.* 6 or more values).

EdU incorporation assay

Adult/postnatal primary NPCs (passage 2-4) were plated at a density of 10,000 cells/cm^2 on nitric acid treated coverslips pre-treated with 50 mg/ml of poly-D-lysine (Sigma-Aldrich) and infected with 30% lentiviral supernatant in fresh NPC proliferation media. After 24 hours, media was replaced with fresh NPC proliferation media. At 48 hours post infection, cells were incubated with 5 μM EdU for 1 hour and fixed with 4% paraformaldehyde/2% sucrose for 10 minutes at room temperature. EdU detection was

carried out as per the manufacturer's protocol (Click_IT EdU Alexa-Fluor 594 Imaging Kit, Invitrogen). Images were taken at 20 X magnification. The percent of EdU positive cells were quantified in a blinded manner for >1000 cells per condition, from two or three independent experiments.

Neurogenesis assay

Primary NPCs (passages 2-3) were isolated from the micro-dissected subventricular zone of postnatal mice (postnatal day 7-10). Cells were plated in triplicate at a density of 12,000 cells/cm^2 in 96 well plates pre-coated with 50 μg/ml poly-D-lysine (Sigma-Aldrich) and 10 ug/ml Laminin (Gibco). 24 hours post-plating, cells were infected with 30% lentiviral supernatant in fresh NPC media. After 24 hours, media was removed and replaced with fresh NPC proliferation media (Neurobasal A (Invitrogen) medium supplemented with penicillin/streptomycin/glutamine (Invitrogen), 2% B27 (Invitrogen), 20ng/ml each of FGF2 (Peprotec) and EGF (Peprotec), and 0.5 μg/ml of puromycin (Invivogen)). 72 hours post infection, cells were switched to NPC differentiation media (Neurobasal A (Invitrogen) medium supplemented with penicillin/streptomycin/glutamine (Invitrogen), 2% B27 (Invitrogen), 0.5% Fetal Bovine Serum (Gibco), and 0.5 μg/ml of puromycin (Invivogen)). Cells were fed NPC differentiation media every other day for a total of 4 days. Test cells expressing shRNAs against candidate genes were also plated at a higher density (30,000 cells/cm^2) to roughly account for observed differences in proliferation as measured by MTT (**Figure 3F, S3D**) and infected with a proportional amount of viral supernatant. Results were only

compared for cells with roughly matched density to control at the onset of differentiation. After four days of differentiation, cells were fixed with 4% paraformaldehyde for 10 min at room temperature. Cells were stained with the following antibodies: 1:150 dilution of Doublecortin (Santa Cruz sc-8066), 1:350 of GFAP (Calbiochem 2.2 B10, 345860), and 1:1000 dilution of DAPI (1 mg/ml). For Fam72a, cells differentiated for 14 days were stained with 1:1000 dilution of Tuj1 antibody (Covance MRB-435P). To quantify the percentage of neurons, cells were imaged two pictures/well were taken using a Zeiss AxioVision Scope. The number of neurons were counted in a blinded manner and the relative percent of neurons was normalized to the cognate empty vector control. Values are reported as the normalized percentage of DCX+ cells per well for replicates coming from 2-4 independent experiments (*i.e.* 6 or more values).

Integrative computational models (classification)

Classification models for each cell type from examples of top 5% broadest and 0-95% broad H3K4me3 domains were built. The broadest 5% of all H3K4me3 domains were labeled as *broad* domains for classification purposes. To represent the rest of the H3K4me3 domain breadth distribution, we randomly sampled H3K4me3 domains from the rest of the domains (excluding top 5% broadest H3K4me3 domains) and labeled them as reference *'non broad'* domains. To avoid biases in the classification output due to the imbalanced number of top 5% broadest and non top 5% broadest H3K4me3 domains, at each sampling we generated a list of control H3K4me3 domains that are equal in number to the 5% broadest H3K4me3 domains. We repeated the sampling procedure 100 times to

eliminate random grouping biases. Co-occurrence of other datasets with H3K4me3 domains was obtained by intersecting coordinates of the H3K4me3 domains with corresponding genomic and epigenomic datasets (a.k.a. the classification features). To correct for biases in intersections due to the difference in the breadth of H3K4me3 domains, we symmetrically extended the length of non-broad domains by mimicking the length distribution of broad domains before interval intersections. We built classification models using four different classification algorithms as implemented in R packages 'e1071' (Support Vector Machines with linear and radial kernels; SVM), 'caret' (k-nearest neighbor) and 'randomForest' (Random Forest). 10-fold cross validation accuracies for SVMs and k-nearest neighbor classifiers, and out of bag prediction accuracies for Random Forest runs were used to estimate the performance of the models. To rank features for their contributions to the models' accuracies, we used the mean decrease in the classification accuracy obtained from all Random Forest models. For each cell type, we identified important features that decrease the classification accuracy at least 1% on average.

Gradual classification

Gradual classification results were obtained by comparing consecutive 5% windows of H3K4me3 domains in H1 hESCs and mESCs on the basis of their breadth to the top 5% broadest domains. To minimize biases due to the difference in the length of the domains, the breadth of domains was extended to mimic the breadth distribution of the top 5% broadest domains in each bin. Random Forest algorithm was used for the classification as

explained above. Classification accuracy represents our ability to discriminate any of these 5% windows from the top 5% broadest domains.

Using integrative models for prediction

To identify all H3K4me3 domains that have the broad H3K4me3 domain signature, at each repeat, we predicted the class label (i.e. 5% broadest vs. non 5% broadest) of all H3K4me3 domains that were not used to train the individual classification trees. Domains predicted as "broad" in 99 out of 100 runs were considered robust 'broad' domain predictions.

Enrichment for binding of transcription factors or chromatin regulators at H3K4me3 domains

To assess enrichments for specific protein binding to the top 5% broadest H3K4me3 domains in comparison to the rest of the H3K4me3 domain breadth distribution, or to specific other H3K4me3 domains breadth quantiles, we accounted for the potential impact of differences in H3K4me3 domain length on genomic region intersections. For this analysis, we obtained 10,000 random samples from non top 5% broadest H3K4me3 domains, where each sample is equal in number to the 5% broadest H3K4me3 domains. We then adjusted the randomly chosen domain lengths to mimic the observed length distribution of the top 5% broadest domains. Using these random samples, we computed a null distribution for genomic intersection ratios for each feature with H3K4me3

domains using the bedtools software (version 2.16). Then, the intersection ratio was calculated for top 5% broadest H3K4me3 domains for each of these features and the null distribution was used to estimate the p-value of enrichment for broad H3K4me3 domains. For the quantile-based analysis, to account for length bias, domains in all quantiles (including the top 5% broadest) were symmetrically extended to the maximal breadth observed in the dataset.

Analysis of PolII accumulation at promoters marked by different H3K4me3 breadth

Sequencing reads were mapped to the appropriate genome assembly. Pile-up of ChIP-seq reads mapped within the proximal promoter (-300bp;+300bp with respect to TSS), gene body (+300bp from TSS;TTS), or total gene region (-300bp from TSS;TSS) were computed as the coverage in the specified intervals using the default settings of coverageBed function of bedtools software and annotations from the cognate assembly downloaded from UCSC genome browser. Traveling ratios were computed as (tags per bp in proximal promoter)/(tags per bp in gene body). Promoter was defined as -300bp,+300bp region from TSS and gene body as +300bp from TSS to TTS. Statistical differences were measured using a one-sided one sample Wilcoxon test compared to expected values from the whole genome (average of 10,000 random samplings of the same number of genes as the set coated by top 5% broadest H3K4me3 domains).

Analysis of chromatin accessibility data at promoters by DNAse-seq or ATAC-seq

Sequencing reads were mapped to the appropriate genome assembly. To reflect the actual portion of DNA that was accessible to DNAseI or Tn5 transposons, each read was then represented solely as its 5'-most coordinate. Gene coordinates were obtained from the UCSC Genome Browser, and promoter regions were considered as spanning 300bp upstream and 300bp downstream of mapped transcription start sites. The number of nick or insertions sites within the proximal promoter (-300bp;+300bp with respect to TSS) were computed as the coverage in the specified intervals using the default settings of coverageBed function of bedtools software and annotations from the cognate assembly downloaded from UCSC genome browser. Statistical differences were measured using a one-sided one sample Wilcoxon test compared to expected values from the whole genome (average of 10,000 random samplings of the same number of genes as the set coated by top 5% broadest H3K4me3 domains).

mRNA expression quantification by microarray and RNA-seq

Microarray data were analyzed using R BioConductor packages (Gentleman et al., 2004), i.e. 'affy' (Gautier et al., 2004) and 'preprocessCore'. In all cases, the RMA method was used to normalize expression levels across samples before subsequent analyses. Expression levels of transcripts were computed from RNAseq datasets by using Tophat1.3.2 (Trapnell et al., 2009) and cufflinks v2.0.2 (Trapnell et al., 2010) softwares. To detect used Transcriptional Start Sites in a sample, start sites of expressed transcripts reconstructed by cufflinks were used. For GRO-seq analyses, quality filtered reads were

mapped to genomes with bowtie0.12.7 software, and the analyzeRNA.pl script from HOMER (Heinz et al., 2010) suite was used to count and normalize the read counts per genes. Differential expression analysis was conducted using the cuffdiff module of the cufflinks software for RNA-seq datasets (Trapnell et al., 2010), and SAM for microarray datasets (Tusher et al., 2001) in the 'siggenes' R package.

Transcriptional Variability

Transcriptional variability was assessed at the single cell or population level using microarray or RNA-seq datasets with three or more replicates generated in conditions similar to matching existing H3K4me3 ChIP-seq datasets. To eliminate biases in the magnitude of variance due to differences in absolute expression levels, gene expression across replicates was scaled to the maximum observed expression level of the gene, and variance of these values is reported per gene (akin to the "coefficient of variation"). Statistical differences were measured using a one-sided one sample Wilcoxon test compared to expected values from the whole transcriptome (average of 10,000 random samplings of the same number of genes as the set coated by top 5% broadest H3K4me3 domains).

Analysis of transcriptional variability after exclusion of bivalent domains

We obtained the genomic coordinates of non-bivalent H3K4me3 domains by excluding all H3K4me3 domains that had any overlap (> 1bp of overlap) with H3K27me3 marked

regions in the same cell type. The set of remaining H3K4me3 domains that had no overlap with H3K27me3 regions were then processed with the usual pipeline to measure transcriptional variability.

RNA-seq in NPCs

RNA from passage 2 adult NPCs was extracted using RNAeasy Kit (Qiagen) according to the manufacturer's protocol. 150 ng of total RNA was used to construct strand specific libraries using the Encore Complete RNA-seq Kit (Nugen Technology #0333), according to the manufacturer's protocol.

Lentiviral based knock-down of *Wdr5*

Primary NPCs (passage 4) were plated at a density of 30,000 cells/^{cm2} in 100 mm (ChIP-Seq), 6 well dishes (protein), or 12 well dishes (RNA) pre-coated with 50 □g/ml of poly-D-lysine (Sigma-Aldrich) and 5 □g/ml of laminin (Invitrogen). Five hours post-plating, adherent cells were transduced with a 40% dilution of lentiviral supernatant in fresh NPC proliferation media for 24 hours. Cells were collected for processing 24 hours post infection.

Western Blots

300,000 NPCs were lysed in 150 ul of ice-cold RIPA buffer (50 mM Tris-HCL pH 7.5, 150 mM NaCl, 2 mM EDTA, 1% NP-40, 0.1% SDS supplemented with 1mM aprotinin and phenylmethylsulfonyl fluoride) and incubated for 10 minutes on ice. Cells were sonicated 3 times for 30 seconds using the Virsonic 600 (VirTis) and centrifuged to remove debris. Following addition of sample buffer (0.0945 M Tris-HCl [pH 6.8], 9.43% glycerol, 2.36% w/v SDS, and 5% β-mercaptomethanol), samples were resolved on 15% SDS-page gels, transferred onto nitrocellulose membranes, and incubated with the following primary antibodies: H3K4me3 (Millipore 07-743), H3K36me3 (Abcam 9050), H3K9me3 (Abcam 8898), H3K27me3 (Millipore 07-449), Total Histone H3 (Abcam 1791), β-actin (Novus Biologicals NB600-501), Wdr5 (Abcam 22512), and PPP1R12C (Banko et al., 2011)). Membranes were incubated with HRP-conjugated anti-mouse or anti-rabbit secondary antibodies (Calbiochem) and visualized using enhanced chemiluminescence (Amersham ECL, GE Healthcare).

Viability Assessment following *Wdr5* knock-down

Primary NPCs transduced as adherent cultures with a 40% dilution of lentiviral supernatant in fresh NPC were collected 24 hours post-infection (see above). Briefly, cells were collected following incubation for 5 minutes in Accutase (Stem Cell Techologies) and resuspended in HBSS buffer (HBSS (Invitrogen), 1% penicillin/streptomycin/glutamine (Invitrogen), 0.1% Bovine Serum Albumin (Sigma-Aldrich)). Propidium Iodide was added at a final concentration of 1 mg/ml and cells were analyzed using a Becton-Dickson LSR II Flow Cytometer.

Computational adjustment of H3K4me3 profiles upon H3K4me3 machinery knock-down

Because H3K4me3 levels were globally down upon WDR5 knock-down in NPCs (or globally up upon JARID1B knock-down in mESCs), the signal-to-noise ratios of these ChIP-seq datasets is necessarily different than that of empty vector infected matched control ChIP-seqs. The signal-to-noise ratio in a ChIP-seq dataset is a crucial parameter in the ability of the peak callers to call significant regions and their boundaries. Thus, an increase in background signal or a lower signal-to-noise ratio would necessarily induce the peaks callers to call more conservative shorter regions even at a matched global sequencing depth. This is very similar to the loss of power of peak callers that occurs when calling peaks from samples at lower depth of sequencing (Chen et al., 2012; Mendoza-Parra et al., 2013).

To detect regions that were becoming sharper or broader without interference from changes in peak intensity, we computationally accounted for the handicap created by different intensities of H3K4me3 upon *Wdr5* knock down in NPCs (or higher intensities following *Jarid1b* knock-down in mESCs). For this purpose, we matched signal-to-noise ratios of paired control H3K4me3 ChIP-seq samples and knock-down H3K4me3 ChIP-seq samples over the minimal regions that are called by MACS2 in the sample with the least power (i.e. *Wdr5* in the case of *Wdr5* knock-down, or control in the case of *Jarid1b* knock-down). To achieve this, we randomly down-sampled sequencing data from the empty vector control H3K4me3 ChIP-seq libraries gradually until achieving similar

global histogram coverage of these regions (fold coverage per bp). This procedure matches the "height" of the peaks from the peak caller's point of view. Then, after graphically determining the down-sampling rate that allows the coverage histogram of each higher sensitivity H3K4me3 ChIP-seq sample to be equal or lower than that of the matched lower sensitivity H3K4me3 ChIP-seq sample (e.g., *Wdr5* knock-down sample). This stringent thresholding guarantees that any loss of breadth observed is only due to a change in breadth of the enriched region and not of a difference in H3K4me3 intensity. We graphically determined these parameters to be at 60% of original depth for replicate 1, and at 45% of original depth for replicate 2 of our control ChIP-seqs. In addition, to limit the effect of variations in input depth, we also matched the effective depth of Input samples for control or knock-down by down-sampling the "deeper" sample to match the number of mapped reads to that of the lower depth matched input.

While increasing the rate of false negatives by reducing our power to detect shortening (i.e. this method will likely induce us to miss a number of regions that lose breadth), this method would also drastically reduce the rate of false positive regions (i.e. regions that didn't lose significant breadth but lost significant signal intensity, which the peak caller lost power to call as "enriched above background").

Analysis of breadth remodeling and transcriptional variability upon *Wdr5*/*Jarid1b* knock-down

To test whether changes in H3K4me3 breadth following *Wdr5* knock-down in NPCs were matching changes in transcriptional variability, we estimated transcriptional variability as explained above using the scaled variance of expression from triplicate RNA-seq samples of independently infected primary cultures (empty vector or *Wdr5* knock-down). We studied the expression characteristics of genes marked by H3K4me3 domains whose remodeling (or lack thereof) was consistent between our ChIP-seq replicates. Namely, we defined as "maintained H3K4me3 domains" domains whose breadth varied less than 5% between empty vector or *Wdr5* knock-down H3K4me3 ChIP-seq in both replicates after computational adjustment of depth (see above). Conversely, we defined as "reduced H3K4me3 domains" domains that lost over 50% of their breadth between empty vector and *Wdr5* knock-down H3K4me3 ChIP-seq in both replicates. We then tested for effect on losing breadth on changes of transcriptional variability levels by comparing the ratios of scaled variance in *Wdr5* knock-down RNA-seq samples versus empty vector RNA-seq samples between the genes marked by "maintained" or "reduced" H3K4me3 domains within each H3K4me3 breadth quantiles using one-sided Wilcoxon tests.

Supplementary References

Abujarour, R., Efe, J., and Ding, S. (2010). Genome-wide gain-of-function screen identifies novel regulators of pluripotency. Stem Cells *28*, 1487-1497.

Agoston, Z., Heine, P., Brill, M.S., Grebbin, B.M., Hau, A.C., Kallenborn-Gerhardt, W., Schramm, J., Gotz, M., and Schulte, D. (2014). Meis2 is a Pax6 co-factor in neurogenesis and dopaminergic periglomerular fate specification in the adult olfactory bulb. Development *141*, 28-38.

Banko, M.R., Allen, J.J., Schaffer, B.E., Wilker, E.W., Tsou, P., White, J.L., Villen, J., Wang, B., Kim, S.R., Sakamoto, K., *et al.* (2011). Chemical genetic screen for AMPKalpha2 substrates uncovers a network of proteins involved in mitosis. Mol Cell *44*, 878-892.

Behar, R.Z., Bahl, V., Wang, Y., Weng, J.H., Lin, S.C., and Talbot, P. (2012). Adaptation of stem cells to 96-well plate assays: use of human embryonic and mouse neural stem cells in the MTT assay. Curr Protoc Stem Cell Biol *Chapter 1*, Unit 1C 13.

Chen, Y., Negre, N., Li, Q., Mieczkowska, J.O., Slattery, M., Liu, T., Zhang, Y., Kim, T.K., He, H.H., Zieba, J., *et al.* (2012). Systematic evaluation of factors influencing ChIP-seq fidelity. Nat Methods *9*, 609-614.

Chia, N.Y., Chan, Y.S., Feng, B., Lu, X., Orlov, Y.L., Moreau, D., Kumar, P., Yang, L., Jiang, J., Lau, M.S., *et al.* (2010). A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature *468*, 316-320.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., *et al.* (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Fasano, C.A., Dimos, J.T., Ivanova, N.B., Lowry, N., Lemischka, I.R., and Temple, S. (2007). shRNA knockdown of Bmi-1 reveals a critical role for p21-Rb pathway in NSC self-renewal during development. Cell Stem Cell *1*, 87-99.

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nat Protoc *7*, 1728-1740.

Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics *20*, 307-315.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol *5*, R80.

Guillemot, F. (1999). Vertebrate bHLH genes and the determination of neuronal fates. Exp Cell Res *253*, 357-364.

Han, D.W., Tapia, N., Hermann, A., Hemmer, K., Hoing, S., Arauzo-Bravo, M.J., Zaehres, H., Wu, G., Frank, S., Moritz, S., *et al.* (2012). Direct reprogramming of fibroblasts into neural stem cells by defined factors. Cell Stem Cell *10*, 465-472.

Hawkins, R.D., Hon, G.C., Lee, L.K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L.E., Kuan, S., Luu, Y., Klugman, S., *et al.* (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell *6*, 479-491.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934-947.

Hu, G., Kim, J., Xu, Q., Leng, Y., Orkin, S.H., and Elledge, S.J. (2009). A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. Genes Dev *23*, 837-848.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res *37*, 1-13.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.

Huang da, W., Sherman, B.T., Zheng, X., Yang, J., Imamichi, T., Stephens, R., and Lempicki, R.A. (2009c). Extracting biological meaning from large gene lists with DAVID. Curr Protoc Bioinformatics *Chapter 13*, Unit 13 11.

Johnson, W.E., Li, W., Meyer, C.A., Gottardo, R., Carroll, J.S., Brown, M., and Liu, X.S. (2006). Model-based analysis of tiling-arrays for ChIP-chip. Proc Natl Acad Sci U S A *103*, 12457-12462.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Lujan, E., Chanda, S., Ahlenius, H., Sudhof, T.C., and Wernig, M. (2012). Direct conversion of mouse fibroblasts to self-renewing, tripotent neural precursor cells. Proc Natl Acad Sci U S A *109*, 2527-2532.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol *28*, 495-501.

Mendoza-Parra, M.A., Nowicka, M., Van Gool, W., and Gronemeyer, H. (2013). Characterising ChIP-seq binding patterns by model-based peak shape deconvolution. BMC Genomics *14*, 834.

Muers, M. (2011). Functional genomics: the modENCODE guide to the genome. Nat Rev Genet *12*, 80.

Ninkovic, J., Steiner-Mezzadri, A., Jawerka, M., Akinci, U., Masserdotti, G., Petricca, S., Fischer, J., von Holst, A., Beckers, J., Lie, C.D., *et al.* (2013). The BAF complex interacts with Pax6 in adult neural progenitors to establish a neurogenic cross-regulatory transcriptional network. Cell Stem Cell *13*, 403-418.

Palmer, T.D., Takahashi, J., and Gage, F.H. (1997). The adult rat hippocampus contains primordial neural stem cells. Mol Cell Neurosci *8*, 389-404.

Pastrana, E., Cheng, L.C., and Doetsch, F. (2009). Simultaneous prospective purification of adult subventricular zone neural stem cells and their progeny. Proc Natl Acad Sci U S A *106*, 6387-6392.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Schmitz, S.U., Albert, M., Malatesta, M., Morey, L., Johansen, J.V., Bak, M., Tommerup, N., Abarrategui, I., and Helin, K. (2011). Jarid1b targets genes regulating development and is involved in neural differentiation. Embo J *30*, 4586-4600.

Sheng, C., Zheng, Q., Wu, J., Xu, Z., Wang, L., Li, W., Zhang, H., Zhao, X.Y., Liu, L., Wang, Z., *et al.* (2012). Direct reprogramming of Sertoli cells into multipotent neural stem cells by defined factors. Cell Res *22*, 208-218.

Shin, H., Liu, T., Manrai, A.K., and Liu, X.S. (2009). CEAS: cis-regulatory element annotation system. Bioinformatics *25*, 2605-2606.

Song, J.S., Johnson, W.E., Zhu, X., Zhang, X., Li, W., Manrai, A.K., Liu, J.S., Chen, R., and Liu, X.S. (2007). Model-based analysis of two-color arrays (MA2C). Genome Biol *8*, R178.

Thier, M., Worsdorfer, P., Lakes, Y.B., Gorris, R., Herms, S., Opitz, T., Seiferling, D., Quandel, T., Hoffmann, P., Nothen, M.M., *et al.* (2012). Direct conversion of fibroblasts into stably expandable neural stem cells. Cell Stem Cell *10*, 473-479.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A *98*, 5116-5121.

Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods *5*, 829-834.

Webb, A.E., Pollina, E.A., Vierbuchen, T., Urban, N., Ucar, D., Leeman, D.S., Martynoga, B., Sewak, M., Rando, T.A., Guillemot, F., *et al.* (2013). FOXO3 Shares Common Targets with ASCL1 Genome-wide and Inhibits ASCL1-Dependent Neurogenesis. Cell Rep *4*, 477-491.

Westerman, B.A., Braat, A.K., Taub, N., Potman, M., Vissers, J.H., Blom, M., Verhoeven, E., Stoop, H., Gillis, A., Velds, A., *et al.* (2011). A genome-wide RNAi screen in mouse embryonic stem cells identifies Mp1 as a key mediator of differentiation. J Exp Med *208*, 2675-2689.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307-319.

Xu, H., Handoko, L., Wei, X., Ye, C., Sheng, J., Wei, C.L., Lin, F., and Sung, W.K. (2010a). A signal-noise model for significance analysis of ChIP-seq with negative control. Bioinformatics *26*, 1199-1204.

Xu, H., Schaniel, C., Lemischka, I.R., and Ma'ayan, A. (2010b). Toward a complete in silico, multi-layered embryonic stem cell regulatory network. Wiley Interdiscip Rev Syst Biol Med *2*, 708-733.

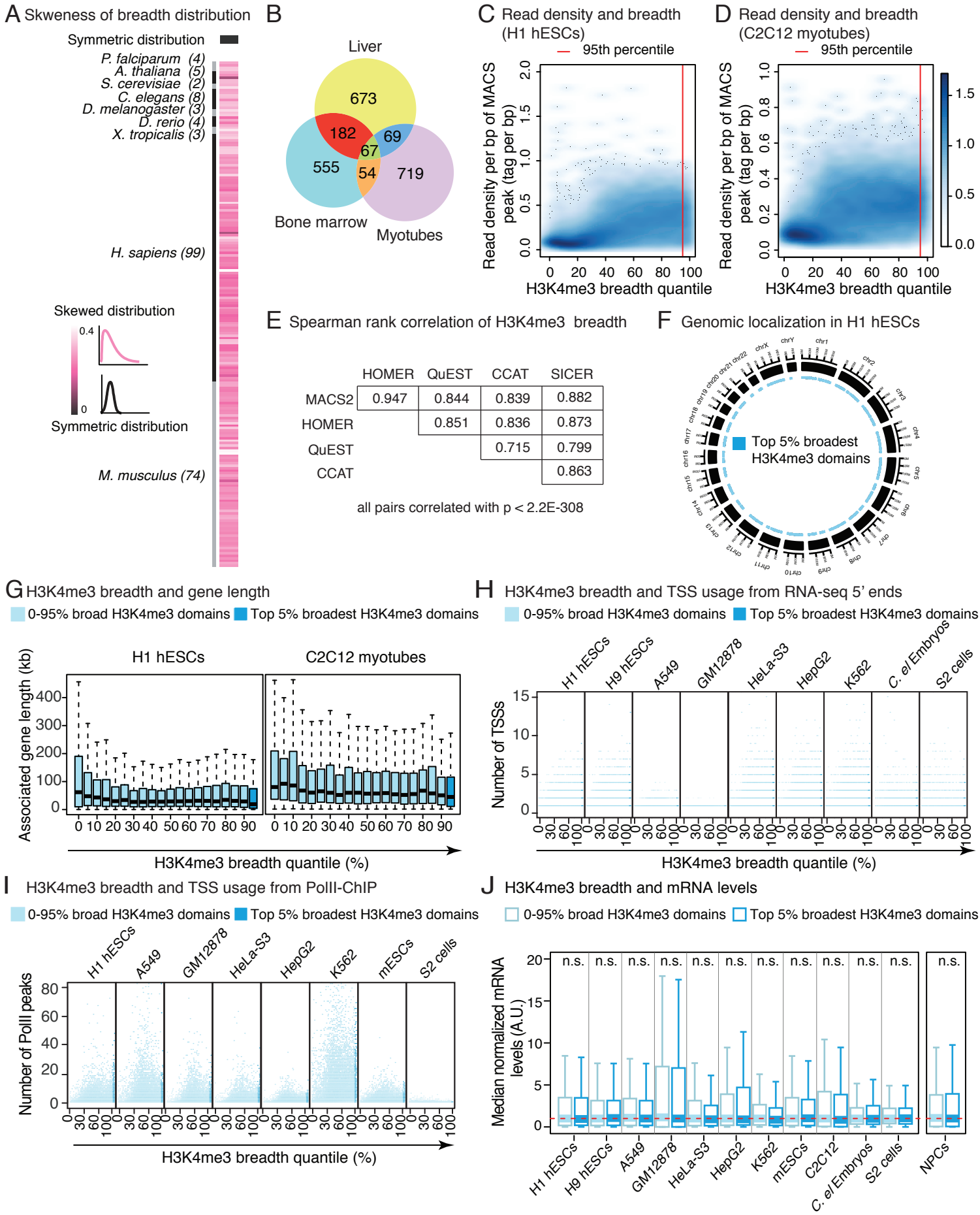
Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics *25*, 1952-1958.

Zhang, J.Z., Gao, W., Yang, H.B., Zhang, B., Zhu, Z.Y., and Xue, Y.F. (2006). Screening for genes essential for mouse embryonic stem cell self-renewal using a subtractive RNA interference library. Stem Cells *24*, 2661-2668.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

Zhu, J., Adli, M., Zou, J.Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., De Jager, P.L., *et al.* (2013). Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell *152*, 642-654.

Supplemental Figure



	HOMER	QuEST	CCAT	SICER
MACS2	0.947	0.844	0.839	0.882
HOMER		0.851	0.836	0.873
QuEST			0.715	0.799
CCAT				0.863

all pairs correlated with $p < 2.2E\text{-}308$

Supplemental Figure

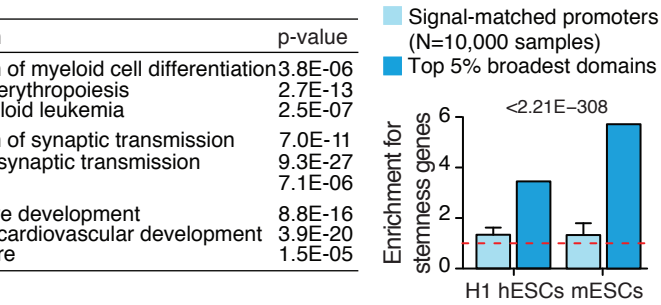
A

Human tissue	Annotation	p-value
CD34 primary cells	Hemopoiesis	9.6E-10
	Abnormal T-cell differentiation	1.3E-18
	Myeloid leukemia	1.2E-10
Brain (Mid frontal lobe)	Axonogenesis	1.7E-07
	Abnormal cerebrum morphology	4.6E-11
	Neuroblastoma	3.9E-06
Skeletal muscle	Muscle organ development	1.5E-12
	Abnormal muscle fiber morphology	1.2E-15
	Nemaline myopathy	1.5E-05

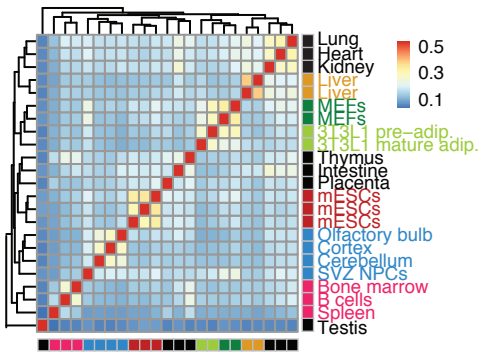
B

Mouse tissue	Annotation	p-value
Bone marrow	Regulation of myeloid cell differentiation	3.8E-06
	Abnormal erythropoiesis	2.7E-13
	Acute myeloid leukemia	2.5E-07
Cortex	Regulation of synaptic transmission	7.0E-11
	Abnormal synaptic transmission	9.3E-27
	Epilepsy	7.1E-06
Heart	Vasculature development	8.8E-16
	Abnormal cardiovascular development	3.9E-20
	Heart failure	1.5E-05

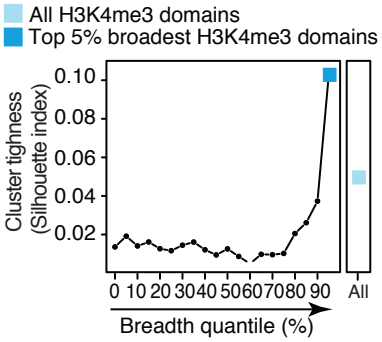
C



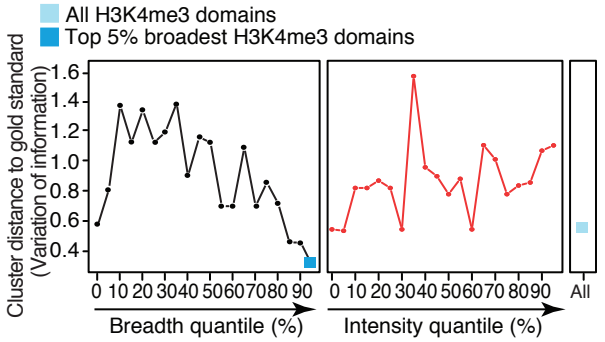
D Clustering of mouse tissues



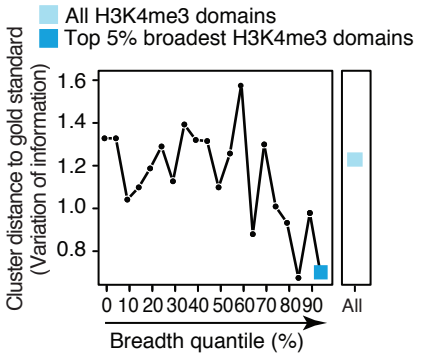
E Clustering quality (Mouse)



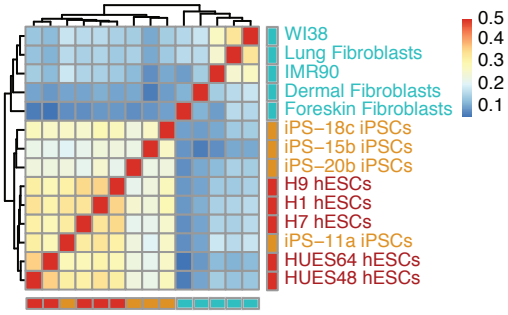
F Deviation from gold standard (Human)



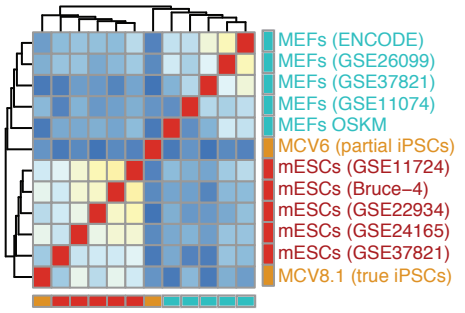
G Deviation from gold standard (Mouse)



H Clustering of human iPSCs



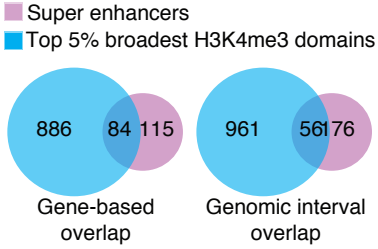
I Clustering of mouse iPSCs



J Functional enrichment associated to remodeled top 5% broadest H3K4me3 domains during differentiation

Differentiation paradigm	Annotation of gained top 5% broadest domains	p-value	Annotation of lost top 5% broadest domains	p-value
3T3L1 adipogenesis	Brown fat cell differentiation	3.4E-13	Growth factor activity	1.3E-09
	Abnormal lipid homeostasis	5.7E-16	Regulation of cell proliferation	5.3E-11
	PPAR signaling pathway	2.1E-12	Cell fate commitment	5.8E-10
C2C12 myogenesis	Contractile fiber	1.6E-08	Chromatin assembly	1.1E-05
	Abnormal skeletal muscle morphology	6.0E-07	DNA replication	1.5E-04
	Genes involved in Striated Muscle Contraction	1.5E-07	Kinetochore	8.7E-04
BM-MSC chondrogenesis	Extracellular matrix structural constituent	1.8E-08	Regulation of cell growth	2.6E-07
	Skeletal system development	3.5E-08	Cytokinesis	2.1E-06
	Abnormal cartilage morphology	2.0E-06	Cleavage furrow	1.5E-04

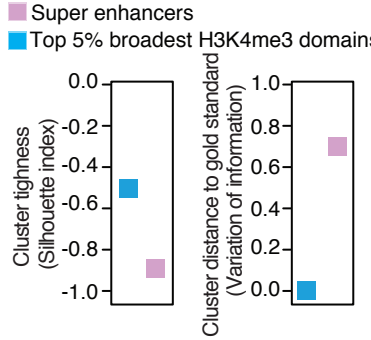
K mESCs



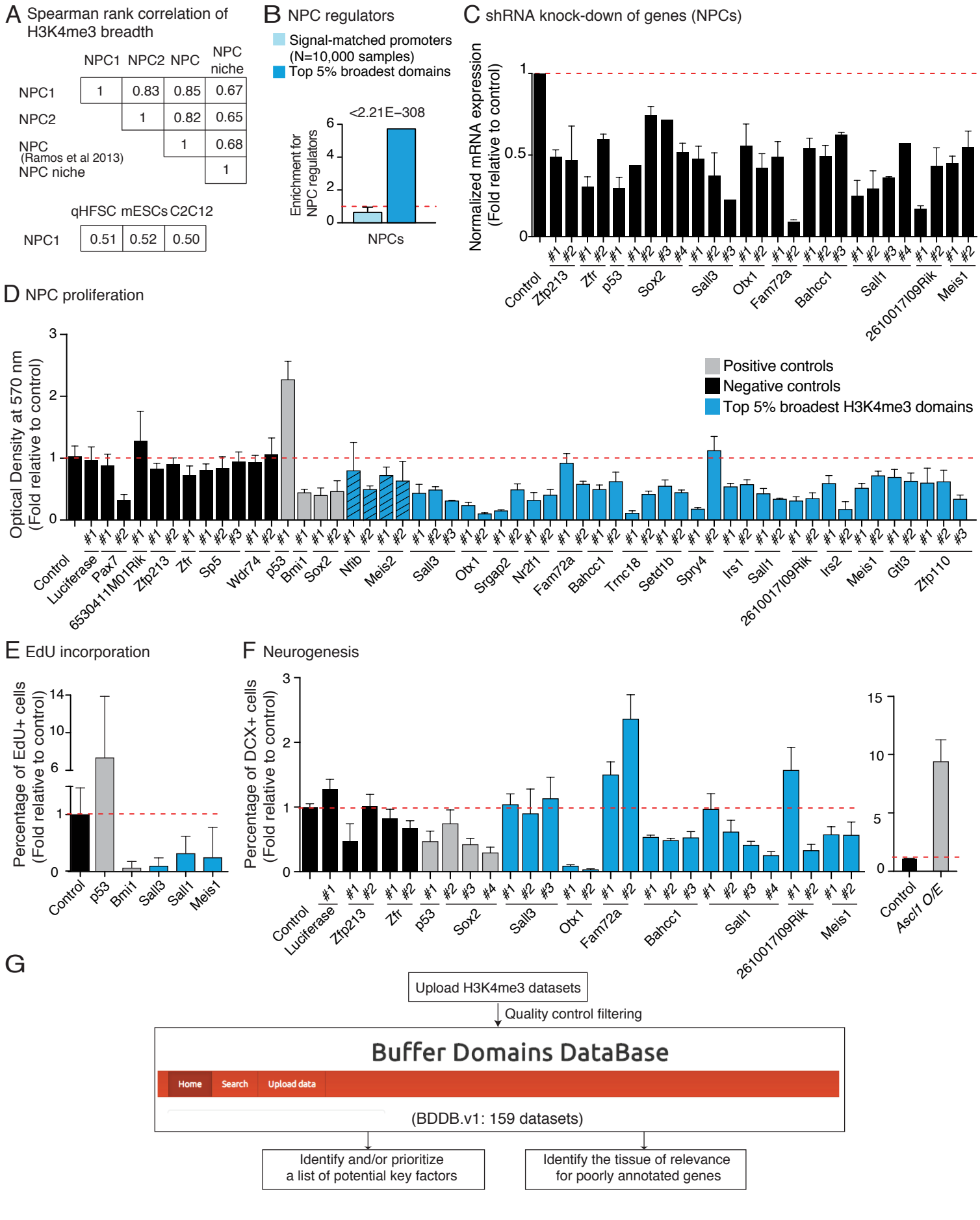
L

Signature type	Top 5% broadest H3K4me3 domains	Super enhancers
Profiled molecules/marks	H3K4me3	Oct4, Sox2, Nanog,Med1 (H3K27ac)
Genes marked in mESCs	972	210
Precision	65/972 (0.07)	**23/210 (0.11)**
Sensitivity	**65/330 (0.20)**	23/330 (0.07)
F1 score (sensitivity x precision)	**0.10**	0.09
Enrichment p-value	**8.58E-18**	3.57E-13
Ranking (GSEA p-value)	**Yes (p < 1E-4)**	Yes/p=0.68

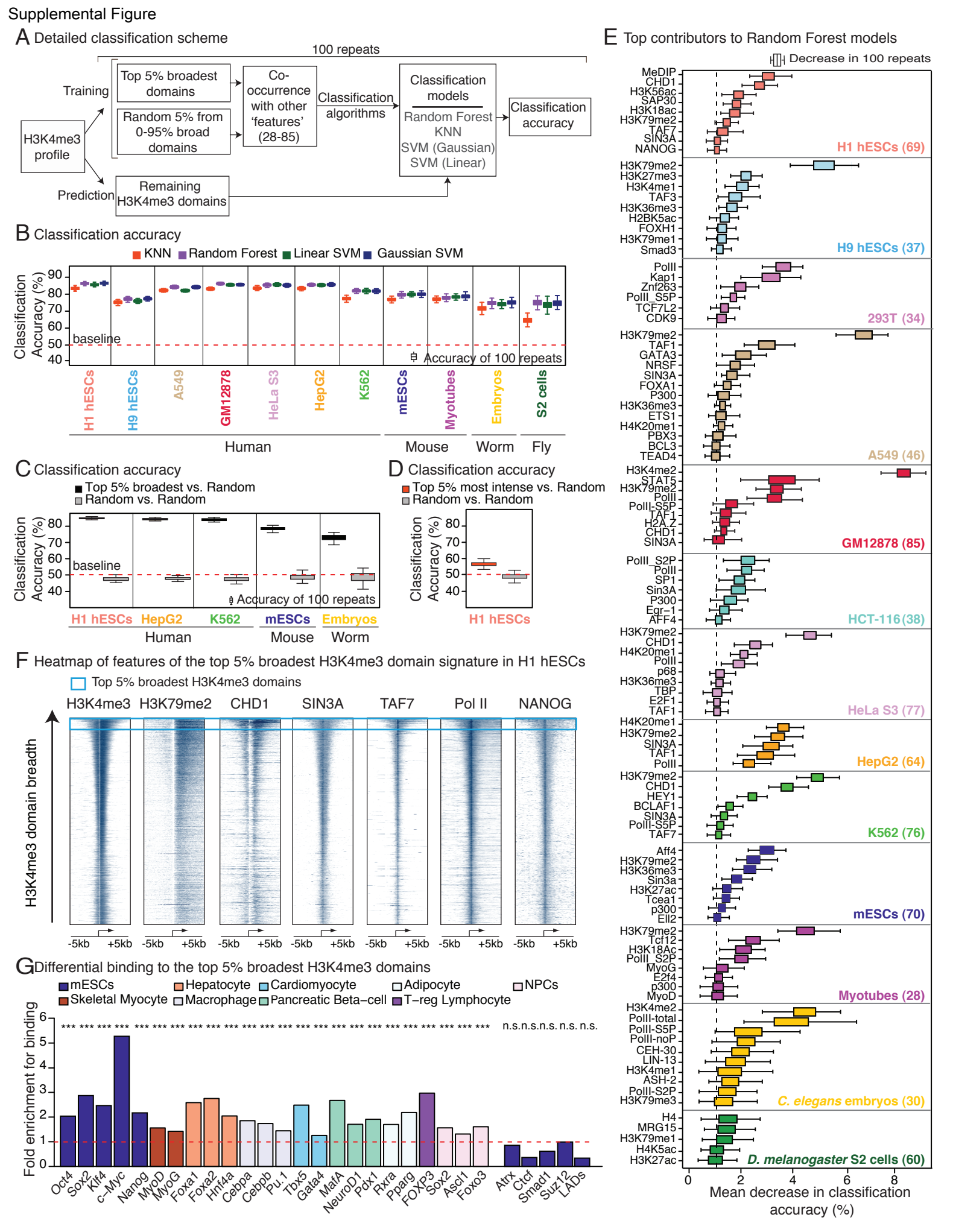
M Clustering quality comparison



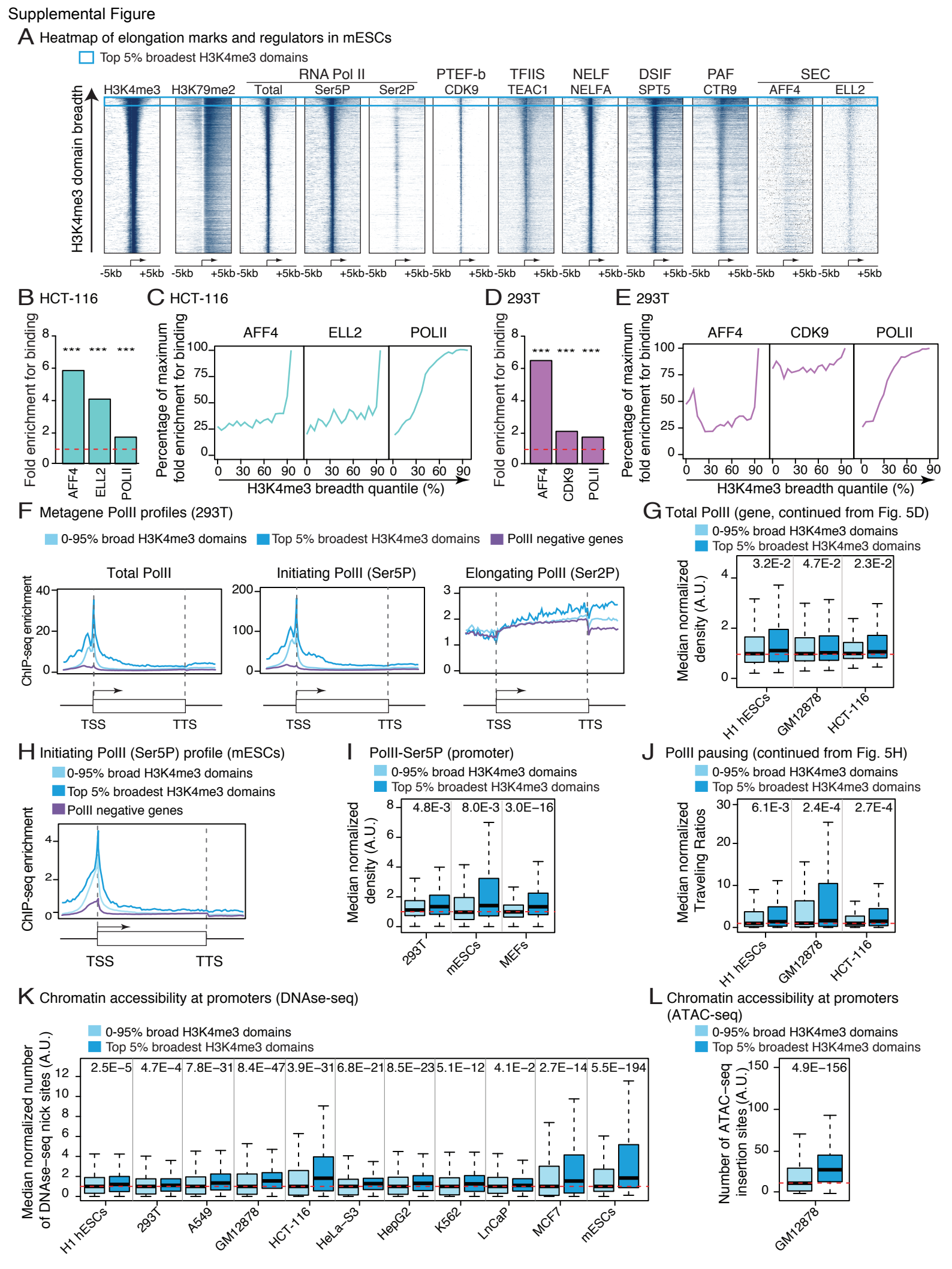
Supplemental Figure



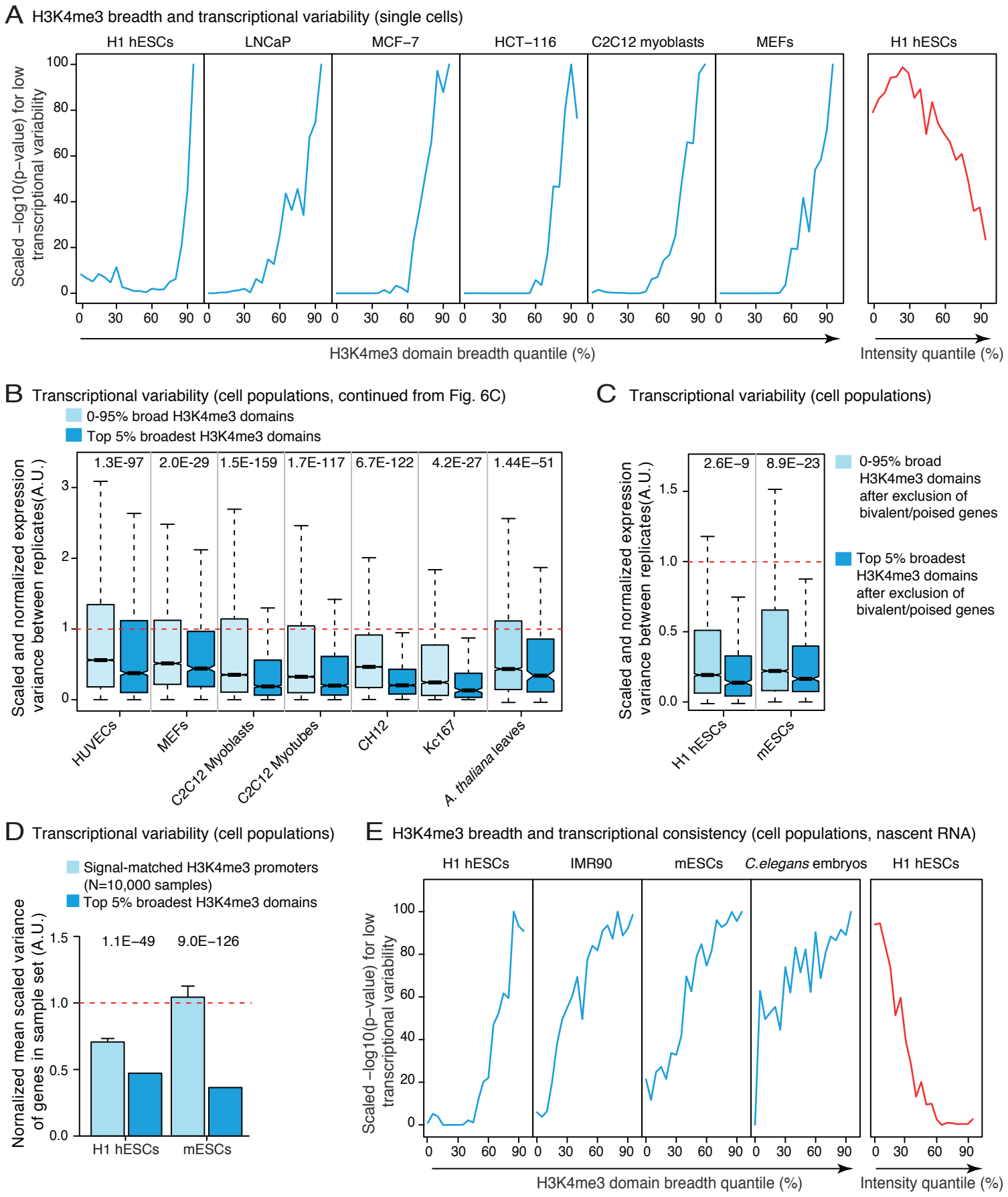
Supplemental Figure



Supplemental Figure



Supplemental Figure



Supplemental Figure