Supplementary tables

Stage	z-value for H3K27me3 enrichments
Embryonic stem cell	4.2
Neural progenitor	5.7
Terminal neuron	1.3

Supplementary Table 1: For each of the three stages we compared mean and variance of H3K27me3 levels in all REST target promoters (i.e. those promoters with a predicted number of binding sites larger than 0.75) and all non-target promoters and calculated a z-value for the difference. The positive numbers in all three stages indicate that REST target promoters show more H3K27me3 than non-target promoters at all stages.

Binding site type	Percentage
Canonical	55
Non-canonical spacing	15
Only half-site	5
No site predicted	25

Supplementary Table 2: REST binding site predictions for REST ChIP-seq data: The majority (55%) of REST peaks contain a match to the canonical Rest motif, which represents a 16-fold enrichment of this motif at binding peaks relative to its occurrence at promoters (data not shown). An additional 15% of peaks contain sites with non-canonical spacing between the half sites of the motif, and 5% contain only one half site in agreement with previous observations (Johnson et al. 2007). About one quarter of the REST peaks lack a computationally identifiable REST binding site.

Promoter class	REST target	Non-target
High-CpG	405 (1.8%)	12886 (56.9%)
Low-CpG	93 (0.4%)	9255 (40.1%)

Supplementary Table 3: Numbers (and corresponding percentages) of promoters in high- and low-CpG classes that are either targeted by REST (i.e. with a REST binding peak within 2kb of the TSS) or that are not a REST target. REST predominantly targets high-CpG promoters, as high-CpG promoters are roughly 3 times more likely to be REST targets than low-CpG promoters.

H3K27me3 region class	REST target	Non-target
Proximal	351 (1.9%)	5250 (28.7%)
Distal	199 (1.1%)	12496(68.3%)

Supplementary Table 4: Numbers (and corresponding percentages) of H3K27me3-enriched regions that are either proximal or distal to a TSS, and that are either targeted by REST or not (Non-target). REST predominantly targets proximal regions with H3K27me3, as proximal regions are 4 times more likely to be a REST target than distal regions.

Motif	Correlation coefficient r	z-value (high-CpG)	z-value (low-CpG)
Strong anticorrelation:			
Ciz	-0.996	2.14	0.60
Rreb1	-0.995	2.12	2.87
Rest	**-0.992**	**5.08**	**2.00**
Ubx	-0.986	2.12	0.12
Gcnf	-0.932	2.23	1.01
Err	-0.905	2.15	0.45
Blimp1	-0.810	3.08	1.50
Weak (anti)correlation:			
Sp1	-0.568	5.33	1.28
Irf7	-0.540	2.06	0.11
Snail	-0.473	4.56	1.56
Tfap2a	-0.453	0.51	2.34
FoxD3	-0.295	3.80	2.17
Maz	-0.187	4.59	3.05
Sf1	-0.085	3.56	0.49
Tef	-0.082	2.86	0.55
Prrx2	-0.078	2.77	4.08
Hmgiy	0.168	2.52	0.48
Myf	0.207	3.81	2.54
Gata4	0.290	2.11	0.77
Sox5	0.591	0.73	2.00
Broad complex 4	0.592	2.62	0.54
Strong correlation:			
bZip/Creb	0.696	2.46	0.16
Foxl1	0.782	2.44	1.92
Klf4	0.905	4.97	0.30
Fac1	0.952	2.49	0.01
FoxP1	0.984	1.21	2.88
Zeb1	0.995	3.52	0.47

Supplementary Table 5: Shown are all motifs for which Epi-MARA predicted a z-value bigger than 2 (for either the high-CpG or low-CpG regions). For each motif the correlation coefficient r of the high-CpG and low-CpG profiles is shown in the second column. The z-scores for high-CpG and low-CpG activities are shown in the third and fourth columns. Whereas REST shows a strong negative correlation, other TFs such as ZEB1 and Klf4 show a strong positive correlation. Thus, correlation coefficients are TF specific.

Epi-MARA		**Motif Enrichment (ES)**		**Motif Enrichment (NP)**		**Motif Enrichment (TN)**	
Name	**z-value**	**Name**	**z-value**	**Name**	**z-value**	**Name**	**z-value**
Sp1	6.2	Blimp1	6.4	Sp1	18.7	Tbp	4.1
Snail	4.5	Klf4	5.3	Maz	14.5	Snail	3.8
Nrf1	4.1	Tbp	4.2	Ap2	9.3	Smad	3.6
Nr2f1/Hnf4	2.8	Myf	3.9	Mazr	9.0	Myf	3.1
Zeb1	2.7	Rreb1	3.8	Snail	8.9	Tbx5	2.8
Rest	2.5	Lmo2	3.7	Tfap2a	8.3	Zeb1	2.6
Staf	2.3	Smad	3.4	Hif1	8.1	Lmo2	2.5
Tcf1	2.3	Tbx5	3.2	Mtf1	7.9	Areb6	2.5
Arnt/Ahr	2.3	T	3.0	E2f	7.9	Blimp1	2.5
Prrx2	2.2	Tead	3.0	Hic1	7.9	Klf4	2.3
bZip/Creb	2.1	Snail	2.9	Lmo2	7.8	Hand1/E47	2.2
Mtf1	2.1	Gli	2.9	Arnt/Ahr	7.0	Nf-E2	2.0
		:		:		:	
		Rest	2.9	Rest	5.9	Rest	1.4

Supplementary Table 6: The table shows the Epi-MARA results (for reference, leftmost column) as well as the top most enriched motifs at each of the stages. We compared binding site densities of each motif at the promoters that were enriched for H3K27me3 at that stage, versus all other promoters that were not enriched for H3K27me3. The results show some overlap, i.e. Sp1 motifs are highly enriched at the NP stage, Snail motifs are enriched at all three stages, and also the Rest motif is enriched at all three stages. Notice, however, that both the REST and Snail motifs that we validate here are much lower in the list of enriched motifs.

Observables	**H3K27me3 levels (ChIP-chip)**	**Transcript levels (chip)**
Fraction of explained variance	8.0%	6.9%

Supplementary Table 7: For each promoter p we calculate the variance of both its expression levels (as measured by micro-array probes for the associated transcripts and its H3K27me3 levels (as measured by ChIP-chip). By comparing the observed expression and H3K27me3 levels, with the levels predicted by MARA (Suzuki et al. 2009) and Epi-MARA using the linear model with the fitted motif activities A_{ms}, we can calculate what fraction of the total variance, i.e. summed over all promoters, is explained by the linear model. The table shows that the linear model captures a similar fraction of variance of both the H3K27me3 and transcript level dynamics across the differentiation.

Supplementary Figure Legends

Supplementary Figure 1: Epi-MARA predicts transcription factor activities that explain dynamics in H3K27me3 levels during neuronal differentiation: Depicted are the normalized activity profiles of the top 30 motifs (blue lines, with standard errors indicated) with their respective z-values. The three time points correspond to the embryonic stem cell (ES), neuronal progenitor (NP), and terminal neuron (TN) stage. Sequence logos of each of the motifs and the transcription factors thought to bind to them are shown as insets.

Supplementary Figure 2: MARA expression analysis predicts the activities of transcription factors driving gene expression dynamics during neuronal differentiation: Normalized activity profiles of the nine most significant motifs that explain changes in gene expression during the differentiation process (red lines, with standard errors indicated). The three time points correspond to the ES, NP and TN stage. Sequence logos of each of the motifs and the transcription factors thought to bind to them are shown as insets.

Supplementary Figure 3: REST expression and binding during differentiation: **a)** Protein levels of REST as detected by Western Blot in extracts from the ES, NP and TN stages in wildtype as well as RESTko background (upper panel). Tubulin serves as a loading control (lower panel). **b)** Distribution of the distance between REST binding peaks and the nearest transcription start site (TSS). Genes with REST binding within +/- 2000 bp of the TSS were classified as REST targets (cut-off indicated by dashed red line). This resulted in a total of 380 target genes in ES cells and 284 in progenitors, with a 96% overlap. **c)** Quantitative PCR of REST ChIPs at the ES and NP stage confirms all 10 tested sites of REST binding as identified by ChIP-seq. Enrichments are normalized to a negative control (*Hprt*). RESTko cells show no signal confirming the specificity of the antibody. Shown are mean enrichments. Error bars show the standard deviation of three biological replicates.

Supplementary Figure 4: Comparison of the REST binding sites identified in our study with those identified by Johnson *et al.* (Johnson et al. 2008). We first gathered all regions that were identified by *Johnson et al.* as REST binding regions at either the ES or NP stage. For each of these regions, and separately for both the ES and NP stages, we then calculated a REST enrichment z-value according to both the data of *Johnson et al.* and our data. The Venn diagrams show the number of these regions that pass a z-value cut-off of 2 for both the ES (upper panel) and NP (lower panel) stages. On the left (red) are regions that only pass the cut-off according to the data of of *Johnson et al.*, on the right (green) regions that only pass the cut-off according to our ChIP-seq data and in the middle the regions that pass the cut-off in both data-sets. Note that while both studies use ES cells as starting point distinct differentiation protocols are used leading to different neuronal populations, which explains larger variation in the differentiated state.

Supplementary Figure 5: **a)** H3K27me3 signal peaks downstream of transcription start sites: Spatial distribution of H3K27me3 signal relative to TSS. All H3K27me3 ChIP-seq samples were pooled and the total read density was plotted relative to position around TSS (±14000bp). This reveals a H3K27me3 peak in the first 1000bps downstream of TSS. **b)** H3K27me3 and SUZ12 levels peak around REST binding sites. Shown is the normalized average read density of SUZ12 (red) and H3K27me3 (blue) in ES cells (Pasini et al. 2010) and neuronal progenitors. **c)** Distributions of the absolute H3K27me3 levels (log ChIP-seq reads per million averaged across the three stages) at all high-CpG (red) and low-CpG regions (blue) that are significantly enriched for H3K27me3.

Supplementary Figure 6: REST knockout ES cells form neurons similar to wildtype ES cells Part I: Marker proteins show similar staining patterns in immuno-cytochemistry in REST knockout and wildtype (WT) cells: REST wildtype (RESTwt), heterozygous (RESThet) and homozygous knockout (RESTko) neuronal progenitors and terminal neurons were fixed and stained for several marker proteins specific for the neuronal progenitor stage (Tohyama et al. 1992; Heins et al. 2002)(PAX6 (top panel) and NESTIN (middle panel)) and terminal neuron stage (Menezes and Luskin 1994) (TUJ1 (bottom panel)), respectively. The cells shown are representative for the population.

Supplementary Figure 7: REST knockout ES cells form neurons similar to wildtype ES cells Part II: **a)** Pairwise correlations of all gene expression microarrays. Shown are the normalized mean expression levels of three biological replicates each. RESTko cells show the highest correlation to each corresponding REST wildtype stage illustrating that the effect of REST knockout on gene expression is small relative to the changes across the differentiation. **b)** Principal component analysis of the gene expression profiles shows that RESTko cells cluster with the corresponding wildtype stage. **c)** Volcano plots depict, for each gene, the fold-change in gene expression in RESTko vs RESTwt cells and the corresponding adjusted p-value for all three stages of differentiation. REST target genes are colored in blue. At the TN stage only very few genes significantly change expression. At the ES and NP stages the number of significantly affected genes is also relatively small and is dominated by direct REST target genes that are upregulated in the RESTko cells.

Supplementary Figure 8: Single gene validation of REST targets that lose H3K27 methylation in RESTko at the NP stage: Quantitative PCR of H3K27me3 ChIPs at the ES and NP stage confirms the loss of H3K27me3 in RESTko cells. Enrichments are normalized to a positive control (*Evx2*). All three genes (*Chrnb2*, *Xkr7* and *Celsr3*) already loosing H3K27me3 in RESTko ES ChIP-seq data could be validated as well as all genes loosing H3K27me3 in RESTko NPs ChIP-seq data (*Chrnb2*, *Xkr7*, *Stmn3*, *Gabrb3*, *Pgbd5*, *Celsr3*, *Stmn2*, *Bdnf*) compared to control regions (*Cpne9* and *HoxD12*), which did not change. *Hprt* serves as a negative control. Enrichments show the mean of three biological replicates. Error bars indicate standard deviation.

Supplementary Figure 9: **a)** Comparison of SUZ12 levels between WT and RESTko cells in H3K27me3-enriched regions at the NP stage. Shown are the distributions of the normalized difference in SUZ12 levels (represented as a z-statistic, see Methods) in WT versus RESTko for non-target regions (black line) and for REST targets in either low-CpG (blue line) or high-CpG (red line) regions. Few low-CpG REST targets significantly change, whereas a considerable fraction of high-CpG REST targets show evidence of losing SUZ12 in the RESTko cells. **b)** Absence of REST reduces the localization of SUZ12 at REST peaks. Shown is the normalized average read-density of SUZ12 in WT and RESTko neuronal progenitors. **c)** Comparison of the effects on SUZ12 and H3K27me3 levels of REST knockout at REST target regions. For each REST target the normalized difference (z-statistic) between WT and RESTko levels at the NP stage for SUZ12 (x-axis) and H3K27me3 (y-axis) are shown. High-CpG regions are shown in the left panel and low-CpG regions in the right panel. Significant correlation is observed between loss of H3K27me3 and loss of SUZ12 cells for high-CpG REST targets in RESTko cells ($r=0.42$, p-value < 2.2e-16). A very weak but still significant anti-correlation (i.e. gain of H3K27me3 and loss SUZ12 in the RESTko cells) is observed for low-CpG REST targets ($r=-0.28$, p-value < 0.001).

Supplementary Figure 10: Independent changes in H3K27me3 levels and gene expression levels at many REST targets. **a)** Pairwise comparison of changes in the significance of H3K27me3 levels (z-value wildtype minus RESTko, horizontal axis) and changes in transcription (log fold-change RESTko minus wildtype, vertical axis) at both the ES (top panel) and NP stages (bottom panel). The horizontal and vertical lines correspond to a z-value of 1 and a log fold-change of 1 (i.e. two-fold upregulation). In general there is only a weak correlation between the amount of H3K27me3 loss and transcriptional up-regulation. At the ES stage many REST targets are transcriptionally up-regulated without showing a loss of H3K27me3. At the NP stage a significant fraction (33%) of REST targets that significantly lose H3K27me3 (z>1) are not significantly up-regulated. **b)** Distribution of expression log fold-changes under RESTko for REST targets that significantly lose H3K27me3 (z>1, red lines) and REST targets that do not significantly lose H3K27me3 (z<1, blue lines), both at the ES (top panel) and NP (bottom panel) stages. As expected there is an overall association between loss of H3K27me3 and transcriptional up-regulation, especially at the NP stage.

Supplementary Figure 11: A REST binding site is essential for REST binding. Transgenic wildtype promoters show strong REST binding, but no or weak binding at the four REST mutant sequences. Levels were measured at, from left to right in each panel, the inserted region, the corresponding endogenous locus, a positive control, and a negative control region. All REST levels are scaled to that of the endogenous region and error-bars show the standard error of three biological replicates. Note that, of all mutant promoter fragments, the *Pgbd5* promoter shows the clearest evidence of residual REST binding.

Supplementary Figure 12 Frequencies of predicted binding sites around transcription start sites for Epi-MARA's top 9 predicted motifs. Sp1, Nrf1, Staf and Arnt/Ahr show a strong binding preference around 50 bps upstream of TSS, while Snail, Zeb1, and Rest motifs are mostly found downstream of TSS. Nr2f1/Hnf4 and Tcf1 show much less pronounced positional preferences.

Supplementary Figure 13: Overview of the methods used in this study. Shown are the steps in the prediction of TFs that regulate H3K27me3 at promoters or genome-wide regions as well as key experiments for subsequent validation. Each panel corresponds to a subsection of the methods (corresponding subsections and page numbers are indicated).

Supplementary Figure 14: a) Reverse-cumulative distribution of the z-statistic for enrichment of ChIP-seq reads from the REST IPs relative to background for 1 kb windows genome-wide. The distribution clearly shows two regimes with a second tail at z-statistics larger than approximately 3. The vertical line (z=3.1) shows the cut-off that we chose for considering a window significantly enriched for REST binding. The cut-off was chosen to ensure good sensitivity in the identification of REST binding regions. **b)**

Reverse-cumulative distribution of the number of background reads per 1 kb window genome-wide. We observe that the distribution drops steeply up to approximately 20 reads per window, after which it shows a long tail with some windows showing over 100 reads. We remove these genomic regions with aberrantly high background counts (vertical line). **c)** Reverse-cumulative distribution of the z-statistic for enrichment of ChIP-seq reads from the H3K27me3 IPs relative to background for 2 kb windows genome-wide. Again two (and maybe even three) regimes in the distribution are clearly evident and we chose a cut-off of z=4.0 (vertical line) to identify windows significantly enriched for H3K27me3. **d)** Reverse-cumulative distribution of the number of background reads per 2 kb window genome-wide. We observe that the distribution drops steeply up to approximately 100 reads per window, after which the distribution shows a long tail with some windows showing over 500 reads. We remove these genomic regions with aberrantly high background counts (vertical line).

Supplementary Figure 15 Shown are on the x-axis the total H3K27me3 levels (sum of wildtype and RESTko signal) and on the y-axis the average H3K27me3 fold-changes between wildtype and RESTko with standard errors (black dots with error-bars) for all non REST target regions separated into high-CpG (left) and low-CpG (right) regions and for both the ES (top) and NP (bottom) stages.

Supplementary Methods

Epi-MARA: Epi-MARA models the dynamics of epigenetic marks in terms of predicted TFBSs in regulatory regions genome-wide, building on the Motif Activity Response Analysis that we developed previously (Suzuki et al. 2009).

Transcription factor binding site predictions for promoters: For the Epi-MARA analysis of the ChIP-chip data, we selected all promoters for which we had H3K27me3 ChIP-chip measurements. Proximal promoter regions were constructed by taking the transcription start sites obtained from UCSC (RefSeq IDs) and extending them by +/- 500 bps. We have established in previous computational analyses that most functional TFBSs occur in these areas (Balwierz et al. 2009). For each proximal promoter region we used the UCSC pairwise alignment to extract orthologous sequences from mouse, human, rhesus macaque, dog, cow, horse, and opossum. We then used T-Coffee (Notredame et al. 2000) to create multiple alignments of the orthologous proximal promoter sequences. Using databases of experimentally determined binding sites (Wingender et al. 1996; Vlieghe et al. 2006), we curated a set of 207 mammalian regulatory motifs (position specific weight matrices) representing the binding specificities of approximately 350 mammalian TFs. The curation methodology was described previously (Suzuki et al. 2009) and involves removal of mutually redundant motifs and associating motifs with their respective binding factors by comparing the protein sequences of DNA binding domains of transcription factors.

To predict transcription factor binding sites on our multiple alignments we use our MotEvo algorithm (van Nimwegen 2007). MotEvo is a Bayesian probabilistic method that treats the alignments as a mixture of columns that are evolving neutrally, segments that are binding sites for one of the motifs that are evolving under the constraints set by the requirement that the sequence segments remain their affinity for the cognate TF, and segments that are under purifying selection of unknown function. Besides the multiple alignments and position specific weight matrices, MotEvo also takes the phylogenetic tree of the species as input, which we obtained by comparing the third positions of fourfold degenerate codons of orthologous proteins. MotEvo then assigns, to each position in the multiple alignments, a posterior probability that a binding site for each of the motifs in our collection occurs at this position. Finally, MotEvo estimates, separately for each motif, the distribution of binding site occurrences as a function of position relative to TSS, and updates the posterior probabilities of predicted sites by taking these positional preferences into account. Finally, we summarize the TFBS predictions by a matrix with components N_{pm}, denoting the sum of the posterior probabilities of all binding sites for motif m in promoter p.

Quantifying H3K27me3 levels: For the ChIP-chip data, each probe measures the ratio of immunoprecipitated fragments and untreated (Input) fragments. The logarithms of these ratios are normalized using standard procedures for oligonucleotide micro-arrays. The H3K27me3 signal at a given promoter is obtained by averaging the log-ratios of the probes intersecting the promoter. We additionally average the signal over biological replicate experiments.

For the ChIP-seq data, we quantity the H3K27me3 at a promoters by collecting reads that overlap a 4 kb region centered at the transcription start site. The 4kb length was chosen based on our analysis of H3K27me3 regions genome-wide. We find that the majority of H3K27me3 enriched regions are 3-4 kb in length (**Supplementary Fig. 4a**). Thus, by summing reads over a 4kb region, we generally collect reads from the entire H3K27me3 enriched region (thereby reducing fluctuations) while on the other hand not diluting the signal by including flanking regions that are not enriched for H3K27me3. To reduce fluctuations for regions that have low read counts we add a pseudo-count to the read count of each 4Kb region. The size of the pseudo-count was chosen to be the average number of reads per region in the background (Input) sample. Finally, we normalize read-counts by the total number of reads in the sample and take the logarithms to obtain the final quantification M_{pt} of the occurrence of the epigenetic mark by at promoter p at time t. In addition, we average the level M_{pt} over available biological replicates.

Fitting the linear model: We then model the chromatin mark levels across time in terms of the predicted binding sites using the following linear model:

$$M_{pt} = noise + c_p + k_t + \sum_m N_{pm} \cdot A_{mt}, \qquad (1)$$

where c_p is the basal level of the chromatin mark, k_t is a constant term accounting for the total expression at time t, and A_{mt} is the unknown activity of motif m at time point t. We assume the deviations between model and measured level M_{pt} (i.e. the `noise' term in the above formula) is Gaussian distributed with the same, but unknown standard deviation σ for each time points. The likelihood of the chromatin mark signal then becomes

$$P(M|A,c,k,N,\sigma) \propto \prod_{pt} \frac{1}{\sigma} exp\left\{-\frac{\left(\sum_m N_{pm}\cdot A_{mt}-M_{pt}-c_p-k_t\right)^2}{2\sigma^2}\right\}. \quad (2)$$

We first set the constants c_p and k_t to their maximal likelihood values, i.e. we maximize (2) with respect to c_p and k_t. This results in the following expression

$$P(M\,|\,A,N,\sigma) \propto \prod_{p,t} \exp\left[-\frac{\left(\sum_m \tilde{N}_{pm}\tilde{A}_{mt}-\tilde{M}_{pt}\right)^2}{2\sigma^2}\right],$$

where $\tilde{N}_{pm} = N_{pm} - \left\langle N_m \right\rangle$ is the normalized matrix of site counts, i.e. where the average number of sites per promoter $\left\langle N_m \right\rangle$ has been subtracted for each motif *m*, $\tilde{A}_{mt} = A_{mt} - \left\langle A_m \right\rangle$ is normalized motif activity, i.e. where the average motif activity $\left\langle A_m \right\rangle$ of motif *m* across the time course has been subtracted, and $\tilde{M}$ is the row and column normalized matrix of chromatin mark levels, i.e. where row and column averages have been subtracted from the matrix *M* such that all rows and columns of $\tilde{M}$ sum to zero.

To avoid over-fitting of the motif activities A_{mt}, we assign a Gaussian prior on each motif activity, i.e.:

$$P(\tilde{A}_{mt}|\,\lambda) \propto \prod_{p} \exp\left[-\frac{\lambda}{2\sigma^2}\left(\tilde{A}_{mt}\right)^2\right].$$

Combining this prior with the likelihood (2) we obtain the following posterior probability distribution over the motif activities

$$P(A\,|\,M,N,\sigma) \propto \prod_{p,t} \frac{1}{\sigma} \exp\left[-\frac{\lambda\sum_m\left(\tilde{A}_{mt}\right)^2+\left(\sum_m \tilde{N}_{pm}\tilde{A}_{mt}-\tilde{M}_{pt}\right)^2}{2\sigma^2}\right].$$

In this expression we can analytically integrate over the unknown standard-deviation *σ* to obtain

$$P(A\,|\,M,N,\sigma) \propto \prod_t \exp\left[-\frac{P\sum_{m,\tilde{m}}\left(\tilde{A}_{mt}-\tilde{A}^*_{mt}\right)W_{m\tilde{m}}\left(\tilde{A}_{\tilde{m}t}-\tilde{A}^*_{\tilde{m}t}\right)}{2\chi_t^2}\right],$$

where *P* is the total number of promoters, the matrix *W* is given by $W_{m\tilde{m}} = \sum_p \left(\tilde{N}_{pm}\tilde{N}_{p\tilde{m}} + \lambda\right)$, the $\tilde{A}^*_{mt}$ are the optimal motif activities, and the chi-squared deviation between model and measurements at time point *t* is given by $\chi_t^2 = \sum_p \left(\tilde{M}_{pt} - \sum_m \tilde{N}_{pm}\tilde{A}_{mt}\right)^2$. The fitted motif activities are determined by singular variance decomposition of the matrix N_{pm}. Note that the resulting fitted activities will depend on the parameter *λ* of the Gaussian prior, i.e. the larger *λ* the smaller the inferred activities. In Epi-MARA, the variable *λ* of the Gaussian prior is chosen by a cross-validation procedure. We randomly select 80% of all promoters as a training set on which we fit the motif activities, and then evaluate the deviation between model and observed levels by M_{pt} on the `test set' of the remaining 20% of promoters. The variable *λ* is chosen so as to minimize the error on the test set.

Finally, once *λ* has been determined, we infer both the maximal posterior activities A^*_{mt} and their standard-errors σ_{mt} from the multi-variant Gaussian posterior, in particular $\sigma^2_{mt} = \frac{\left(W^{-1}\right)_{mm}\chi_t^2}{P}$. To rank motifs by their importance in explaining variations in the levels M_{pt} we use a score similar to a z-statistic. The z-score z_m of motif *m* is quantified as an average squared z-value of the activity across conditions, i.e.

$$z_m = \sqrt{\frac{1}{T}\sum_t\left(\frac{\tilde{A}_{mt}}{\sigma_{nt}}\right)^2}\ , \qquad (3)$$

where *T* is the number of time points in the data-set. Note that our z-scores are meant to rank the importance of motifs and cannot be used to assess the statistical significance of motif activities. To assess the statistical significance of a given z-score one can employ a standard randomization test. Using the same input data and site predictions, we randomize the association between promoters and site counts and run Epi-MARA, noting the resulting z-scores for all motifs. In this way we estimated that, for the H3K27me3 data we analyze, the probability of obtaining a z-score of 2.52 by chance is roughly p=5*10^{-6}.

Epi-MARA on H3K27me3 regions genome-wide: In order to run Epi-MARA on all H3K27me3 regions genome-wide we need to quantify H3K27me3 levels at each region across the time points. Different H3K27me3 regions have different sizes and to normalize H3K27me3 levels across all regions we first identified for each H3K27me3 region the 4kb bps window with the highest H3K27me3 levels. The H3K27me3 levels M_{rt} for each 4 kb region r at each time point t were then obtained in the same way as for the promoters, i.e. we sum ChIP-seq reads intersecting the region, add a pseudo-count which corresponds to the average background level, divide by the total number of reads in the sample, and take the logarithm.

To obtain predicted site-counts N_{rm} for each region r, we first obtained multiple alignments for the entire 4kb regions as described above for the promoters. We then ran MotEvo to predict TFBSs for all motifs in the entire 4kb regions. To take into account the distinct base compositions for high-CpG and low-CpG regions, we used separate background models in MotEvo for the high-CpG and low-CpG regions by separately determining the frequencies of A, C, G, and T nucleotides in the high-CpG and low-CpG regions, resulting in:

H3K27me3 class	A=T frequency	C=G frequency
high-CpG	0.197	0.303
low-CpG	0.269	0.231

Instead of taking the predicted binding sites from the entire 4kb regions for the site counts N_{rm} , we wanted to focus in on a 1kb window in each 4kb region that is most likely to contain the most relevant binding sites. We reasoned that the 1kb window with the highest overall density of predicted binding sites (which is typically the window that also shows the highest sequence conservation across mammals) is most likely to contain the relevant TFBSs and we determined N_{rm} using only the predicted TFBSs from this 1kb window.

Given that high-CpG and low-CpG H3K27me3 regions show distinct dynamics of H3K27me3 levels in general, we want to consider the possibility that the occurrences of TFBSs for a given motif m may have different effects at high-CpG and low-CpG regions. Thus we run Epi-MARA on all regions allowing for separate motif activities at high-CpG and low-CpG regions for all motifs. To infer motif activities separately for high- and low-CpG regions we treat, for each motif $m,$ sites within low-CpG regions and sites within high-CpG regions as if they derived from two separate motifs, effectively doubling the number of motifs for which we infer activities. To do this we replace that original matrix of site counts with a new matrix N_{rm} that looks as follows:

$$N_{rm} = \begin{pmatrix} N_{rm}^{high} & 0 \\ 0 & N_{rm}^{low} \end{pmatrix}, \qquad (4)$$

where $N_{rm}^{\{high,low\}}$ is the matrix of the number of expected sites using the predicted TFBSs for high-CpG and low-CpG regions, respectively. We then infer the activities in the same way as for the promoters.

To include the real binding data we have for the TF REST in the Epi-MARA analysis, we first transform the matrix $\mathrm{N_{rm}}$ into a binary matrix, that is to say all entries that are bigger than 0.2 (number of expected sites) are substituted by 1, otherwise by 0. The entries $\mathrm{N_{r,m=REST}}$ are then replaced by the binding data by putting a 1 if the H3K27me3 region p is a REST target (as defined earlier), and 0 if p is a non-target.

Determining H3K27me3 enriched regions genome-wide: To analyze H3K27me3 dynamics across the time course genome-wide, it is essentially that we obtain a reference set of regions for which to calculate H3K27me3 levels at each time point. To this end we decided to identify, genome-wide, all regions that are significantly enriched for H3K27me3 when considering all stages of the differentiation. We proceeded as follows. For each 2 Kb window on the genome, we calculate the fractions f_t of ChIP-seq reads from the samples at each time point t that map to the region in question as well as the fraction f_b of reads from the background sample that maps to the region. Assuming Poissonian noise in the fractions f_t , the variance of the fraction f_t is given by $V_t = \frac{f_t}{N_t}$, with N_t the total number of reads in the ChIP-seq sample at time t. The average fraction across the time course is simply given by $f = \frac{1}{T}\sum_t f_t$ and the variance associated with this fraction is $V = \frac{1}{T^2}\sum_t V_t$. Using this we obtain the following z-value for the overall enrichment at the region: $z = \frac{f - f_b}{\sqrt{V + \frac{f_b}{N_b}}}$. We plot the reverse-cumulative distribution of these z-values for all genomic windows of 2kb wide and chose a cut-off z=4.0 to select significantly enriched regions (**Supplementary Fig. 13**).

References:

Balwierz, P.J., Carninci, P., Daub, C.O., Kawai, J., Hayashizaki, Y., Van Belle, W., Beisel, C., and van Nimwegen, E. 2009. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol* **10**(7): R79.

Heins, N., Malatesta, P., Cecconi, F., Nakafuku, M., Tucker, K.L., Hack, M.A., Chapouton, P., Barde, Y.A., and Gotz, M. 2002. Glial cells generate neurons: the role of the transcription factor Pax6. *Nat Neurosci* **5**(4): 308-315.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**(5830): 1497-1502.

Johnson, R., Teh, C.H., Kunarso, G., Wong, K.Y., Srinivasan, G., Cooper, M.L., Volta, M., Chan, S.S., Lipovich, L., Pollard, S.M., Karuturi, R.K., Wei, C.L. et al. 2008. REST regulates distinct transcriptional networks in embryonic and neural stem cells. *PLoS Biol* **6**(10): e256.

Menezes, J.R. and Luskin, M.B. 1994. Expression of neuron-specific tubulin defines a novel population in the proliferative layers of the developing telencephalon. *J Neurosci* **14**(9): 5399-5416.

Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**(1): 205-217.

Pasini, D., Cloos, P.A., Walfridsson, J., Olsson, L., Bukowski, J.P., Johansen, J.V., Bak, M., Tommerup, N., Rappsilber, J., and Helin, K. 2010. JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. *Nature* **464**(7286): 306-310.

Suzuki, H. Forrest, A.R. van Nimwegen, E. Daub, C.O. Balwierz, P.J. Irvine, K.M. Lassmann, T. Ravasi, T. Hasegawa, Y. de Hoon, M.J. et al. 2009. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet* **41**(5): 553-562.

Tohyama, T., Lee, V.M., Rorke, L.B., Marvin, M., McKay, R.D., and Trojanowski, J.Q. 1992. Nestin expression in embryonic human neuroepithelium and in human neuroepithelial tumor cells. *Lab Invest* **66**(3): 303-313.

van Nimwegen, E. 2007. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics* **8 Suppl 6**: S4.

Vlieghe, D., Sandelin, A., De Bleser, P.J., Vleminckx, K., Wasserman, W.W., van Roy, F., and Lenhard, B. 2006. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res* **34**(Database issue): D95-97.

Wingender, E., Dietze, P., Karas, H., and Knuppel, R. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24**(1): 238-241.