

Uniform, optimal signal processing of mapped deep-sequencing data

Vibhor Kumar¹, Masafumi Muratani², Nirmala Arul Rayan¹, Petra Kraus^{3,4}, Thomas Lufkin^{3,4}, Huck Hui Ng^{2,3} & Shyam Prabhakar¹

Despite their apparent diversity, many problems in the analysis of high-throughput sequencing data are merely special cases of two general problems, signal detection and signal estimation. Here we adapt formally optimal solutions from signal processing theory to analyze signals of DNA sequence reads mapped to a genome. We describe DFilter, a detection algorithm that identifies regulatory features in ChIP-seq, DNase-seq and FAIRE-seq data more accurately than assay-specific algorithms. We also describe EFilter, an estimation algorithm that accurately predicts mRNA levels from as few as 1–2 histone profiles ($R \sim 0.9$). Notably, the presence of regulatory motifs in promoters correlates more with histone modifications than with mRNA levels, suggesting that histone profiles are more predictive of *cis*-regulatory mechanisms. We show by applying DFilter and EFilter to embryonic forebrain ChIP-seq data that regulatory protein identification and functional annotation are feasible despite tissue heterogeneity. The mathematical formalism underlying our tools facilitates integrative analysis of data from virtually any sequencing-based functional profile.

High-throughput DNA sequencing has transformed genomics by enabling genome-wide equivalents of functional assays that were traditionally done on individual loci. The technology is now used to obtain genome-wide maps of histone modification, transcription factor binding, chromatin openness, gene expression, DNA methylation, replication origins and mRNA binding by microRNAs and ribosomes, to name but a few of the applications in a rapidly growing list^{1–3}. At the same time, a plethora of bioinformatic tools has been developed for processing sequence-tag signals from the various assays. For example, distinct, assay-specific tools have been designed for analyzing transcription factor ChIP-seq (TF ChIP-seq), histone ChIP-seq, DNase-seq and formaldehyde-assisted isolation of regulatory elements (FAIRE)-seq data^{1,4–9}. Examples of specialized algorithms include blind deconvolution for TF ChIP-seq analysis⁸ and nonlocal means for detecting RNA

polymerase II (Pol II)-enriched genomic segments¹⁰. However, the specialized nature of most existing analytical methods makes it difficult to compare, integrate or uniformly analyze data from multiple sources, particularly when they differ in signal patterns and noise characteristics. Moreover, many of the current tools are based on heuristics, rather than formal mathematical techniques for optimizing accuracy.

Here, we organize diverse sequence-tag data analysis problems into two major classes of problems, signal detection and signal estimation, and demonstrate that they can be solved by adapting ‘uniform’ and ‘formally optimal’ techniques from the signal processing literature. Signal detection is a generalization of peak detection, in that it accommodates signals of arbitrary shape (e.g., the peak-valley-peak signature of some histone modifications at promoters). We implemented a linear detection filter, known as a Hotelling observer¹¹, that provides mathematically optimal detection accuracy, as measured by the receiver operating characteristic–area under the curve (ROC-AUC). This approach, which uses finite-width linear filtering, is equivalent to computing a weighted average of the signal (binned tag profile) in a sliding window spanning a fixed number of bins, that is, windowed smoothing. The objective of the Hotelling detection filter is to maximize the difference between filter outputs at true-positive regions and noise regions. More precisely, the Hotelling detection filter maximizes the ratio of the mean of this difference to its s.d. (that is, the *z*-score). Dfilter, our algorithm based on this approach, was benchmarked on histone ChIP-seq, TF ChIP-seq, DNase-seq and FAIRE-seq data from multiple cell lines. Unique features of the method were exploited to assess and compare the predictive power of DNase-seq and FAIRE-seq, both individually and in combination.

Signal estimation is the problem of estimating a continuous hidden variable (e.g., expression level, translation rate or binding strength) from observations of other correlated variables. Our algorithm, EFilter, estimates the unknown signal using a linear least-squares approach, and incorporates two additional features to facilitate the removal of bias and the use of cell types for training that are distinct from the target cell type. We benchmarked EFilter on estimating mRNA levels from histone ChIP-seq data, and used it to obtain insights into the relationship between transcription factor binding, chromatin state and gene expression. Finally, we generated histone ChIP-seq profiles from e11.5 mouse forebrain and analyzed the data using DFilter and EFilter, to assess the utility of chromatin profiling in a complex, heterogeneous tissue.

¹Computational and Systems Biology, Genome Institute of Singapore, Singapore.

²Gene Regulation Laboratory, Genome Institute of Singapore, Singapore. ³Stem Cell and Developmental Biology, Genome Institute of Singapore, Singapore.

⁴Department of Biology, Clarkson University, Potsdam, New York, USA.

Correspondence should be addressed to S.P. (prabhakars@gis.a-star.edu.sg).

Received 23 November 2012; accepted 29 April 2013; published online 16 June 2013; doi:10.1038/nbt.2596

RESULTS

Signal detection using DFilter

We illustrate DFilter by applying it to the analysis of ChIP-seq profiles from a human cell line of trimethylated lysine 4 on histone H3 (H3K4me3), a chromatin mark enriched at active promoters. The DFilter algorithm involves the following steps. (i) Construct a genome-wide profile by binning the input set of mapped sequence tags (Fig. 1a, blue curve). (ii) Normalize the tag-count profile using the local tag density in a user-supplied control (for example, non-ChIP genomic DNA; Online Methods). (iii) Construct an approximate training set of positive regions (active promoters, in this case) using windowed averaging. In this example, the initial training set consisted of the 500 RefSeq-annotated transcription start sites (TSSs) having the largest ChIP-seq tag count within a 3-kbp window. (iv) Design the Hotelling detection filter, which is a function of the mean and covariance of the tag profiles at the training regions¹¹ (Supplementary Fig. 1). (v) Smooth the normalized ChIP-seq tag profile using the Hotelling detection filter (Fig. 1a, black curve). (vi) Detect active promoters by thresholding the smoothed profile at a level corresponding to a desired *P*-value cutoff (Online Methods).

We systematically evaluated the accuracy of DFilter in detecting active promoters using previously published H3K4me3 ChIP-seq data from three cell types (CD4⁺ T cells, K562 and GM12878)^{12,13}. We also analyzed the accuracy of the widely used peak-finding algorithms MACS⁵, FindPeaks¹⁴, F-Seq⁶, ZINBA⁷ and SICER¹⁵ on the same data. True positives were defined as RefSeq promoters with above-median gene expression. Expression was calculated using RNA-seq data from the respective cell types^{16,17}. Of the six methods, DFilter consistently displayed the highest accuracy (Fig. 1b). The performance gap between DFilter and the other methods was greatest for

CD4⁺ T cells, perhaps because control data were not available for this data set. DFilter also displayed the highest accuracy in overall promoter detection, regardless of expression (Supplementary Fig. 2). The tradeoff between precision (positive predictive value) and recall (sensitivity) indicates that, in all cases, there is an inflection point in the precision-recall curve, where precision drops sharply as sensitivity increases (Fig. 1c and Supplementary Fig. 2c). This point, which lies at ~3–7% false-discovery rate in our examples, is a potential threshold for annotating novel promoters. In addition to detecting focal signals, such as H3K4me3 at promoters, DFilter also provided the highest accuracy for detecting broad regions of trimethylated histone H3 lysine 36 (H3K36me3) enrichment, which characterize expressed gene bodies (Fig. 1d).

Histone modification signals are typically fuzzy, spanning at least 1–2 kbp, whereas the ChIP-seq signal of a transcription factor binding site forms a sharp peak ~200 bp wide. Consequently, the performance of detection algorithms could vary between the two scenarios. We tested DFilter on five TF ChIP-seq data sets from three cell lines^{4,18,19}. DFilter generated its own training set of positive regions by scanning the genome for windows of high tag density. For comparison, we evaluated Quest⁴ and MACS⁵, two methodologically distinct algorithms that show high accuracy on TF ChIP-seq benchmarks⁹. We calculated the fraction of peaks that contained a high-affinity (high-scoring) sequence motif for the target transcription factor, and used this fraction as an approximate measure of precision. By this measure, the precision of DFilter was comparable to, or greater than, that of the other two methods at all detection thresholds in all five data sets (Fig. 2a).

Next, we asked whether the same algorithm could also detect regions of open chromatin based on DNase-seq and FAIRE-seq signals.

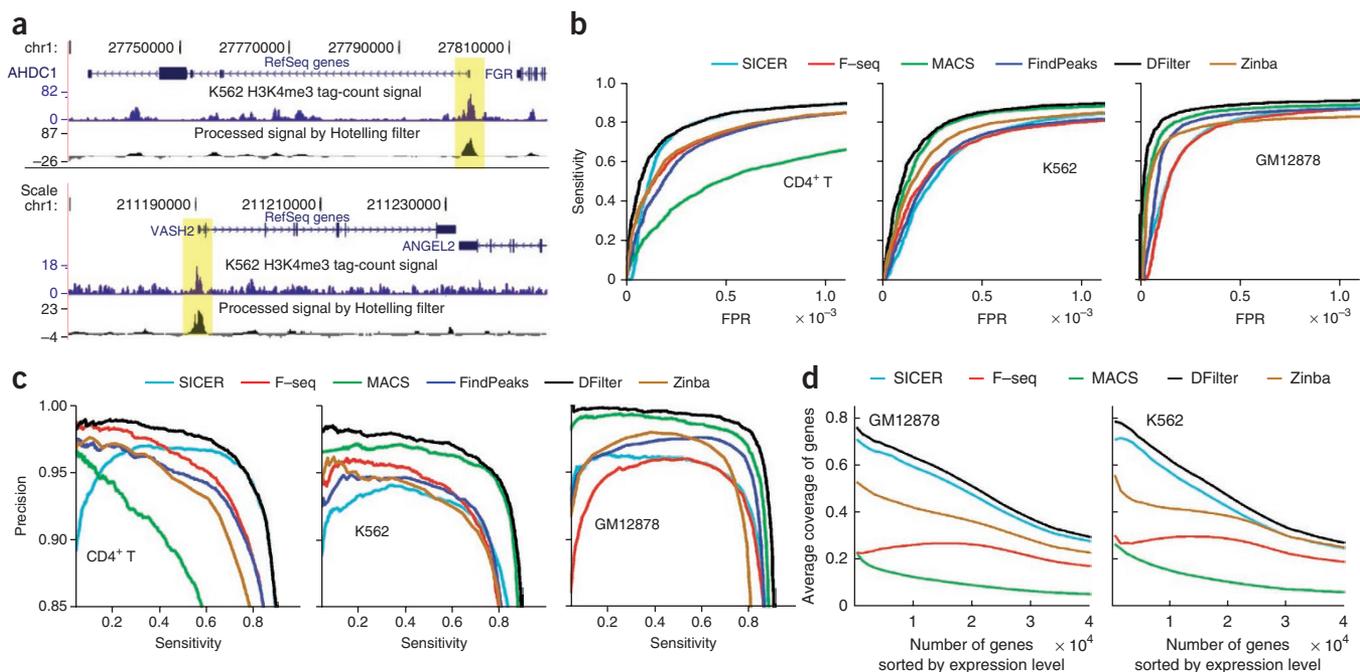


Figure 1 Signal detection by DFilter using histone modification ChIP-seq data. **(a)** Illustration of DFilter output: promoter detection in K562 cells. The filter is designed to match the shape of the histone modification profile at active promoters and have zero response to regions characterized by unvarying tag-count density. Filter output is higher at peaks of the desired shape than at peaks of incorrect shape. **(b)** Performance (ROC curve) of DFilter and five other algorithms at active promoter detection using H3K4me3 ChIP-seq in CD4⁺ T, K562 and GM12878 cells. Positive set: RefSeq promoters with expression above median. Negative set: 2-kbp non-CpG regions that are at least 2.5 kbp away from RefSeq and UCSC gene TSSs. FPR: false-positive rate (1-specificity). **(c)** Corresponding precision-recall plot. **(d)** Coverage of active genes by H3K36me3-enriched segments. For uniformity, each algorithm was allowed to cover at most 10% of the genome with enriched blocks.

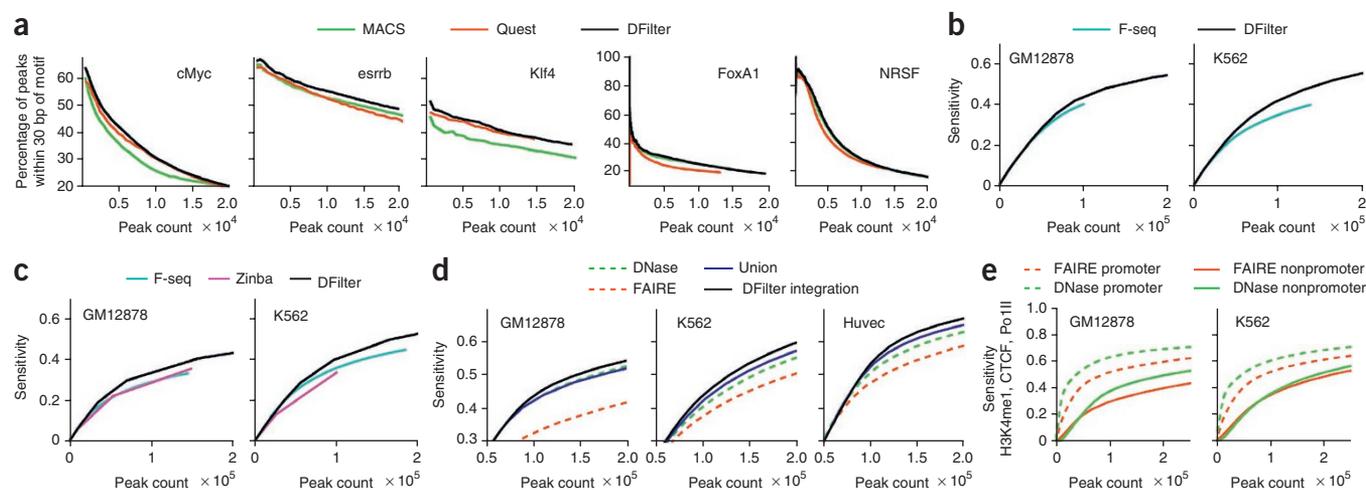


Figure 2 Performance of DFilter on other data types. **(a)** TF ChIP-seq peak detection. Accuracy is measured as the fraction of peaks that contain a sequence motif for the corresponding transcription factor within the central 60 bp, as detection threshold is varied. ChIP-seq data for cMyc, Esrrb and Klf4 are from mouse ES cells. FoxA1 data are from human MCF7 cells, and NRSF data are from human Jurkat cells. **(b)** Sensitivity of DNase-seq peak detection relative to a *cis*-regulatory benchmark (fraction of benchmark regions detected). The benchmark set is defined as the union of ENCODE ChIP-seq peak calls for PolII, H3K4me1 and CTCF. **(c)** Sensitivity of FAIRE-seq peak detection relative to the *cis*-regulatory benchmark. **(d)** Sensitivity of integrative open chromatin detection using DNase-seq and FAIRE-seq data in combination. Dashed lines, shown for reference, represent the sensitivity of DFilter when only one data type is used. Blue solid lines: Fisher's method for combining DNase-seq and FAIRE-seq *P*-values. Solid black lines: DFilter integrative peak detection using DNase-seq and FAIRE-seq simultaneously. **(e)** Comparison of DNase-seq and FAIRE-seq sensitivity at promoter and nonpromoter regions. DFilter peak calls, all available sequence tags used.

For comparison, we evaluated two other methods, F-Seq⁶ and ZINBA⁷, that were designed for open chromatin detection. DNase-seq and FAIRE-seq data from 'Tier 1' human cell lines of the ENCODE project (<http://genome.ucsc.edu/encode/cellTypes.html>) were used in this analysis³. As before, DFilter generated its own training set by scanning the genome for windows of high tag density. Owing to the lack of an independent benchmark set of open chromatin regions, we quantified detection accuracy based on overlap with putative regulatory elements (ENCODE peak calls for CTCF, Pol II and monomethylated histone H3 lysine 4 (H3K4me1); Online Methods). We also created a benchmark consisting of acetylated histone H3 lysine 27 (H3K27ac) peaks, which represent active promoter and enhancer regions¹³. DFilter peaks consistently showed the greatest overlap with externally defined regulatory regions (Fig. 2b,c and Supplementary Fig. 3), suggesting that the Hotelling approach provides uniformly high accuracy across a broad range of signal detection problems.

So far, we have detected signals in individual data sets. However, the mathematical formalism underlying DFilter is also applicable to integrative signal detection from multiple data sets. To assess the benefits of integrative detection, we used DFilter to detect open chromatin regions in GM12878, K562 and human umbilical vein endothelial cells (Huvec) cells using both DNase-seq and FAIRE-seq data simultaneously. We found that DFilter detected open chromatin signals more accurately from the combined data set than from DNase-seq or FAIRE-seq alone (Fig. 2d). A similar result was obtained previously³, although in that case the two data types were integrated by first calling peaks individually and then combining the resulting *P* values using Fisher's combined probability test³. In our analysis, this two-step method was less successful than direct integrative analysis using DFilter (Fig. 2d). This is perhaps because DFilter accounts for the correlation between DNase-seq and FAIRE-seq signals, whereas Fisher's method assumes statistical independence.

It is possible that DNase-seq and FAIRE-seq measure subtly different aspects of the chromatin state. For example, the two assays appear

to differ in their sensitivity at promoter regions³. However, a direct quantitative comparison on a comprehensive benchmark data set has not been performed. We partitioned the benchmark set of regulatory elements into promoter and nonpromoter subsets, and repeated the analysis in Figure 2d. As expected, DNase-seq provided substantially higher sensitivity at promoter regions (Fig. 2e). However, even at nonpromoter regulatory elements, the sensitivity of DNase-seq was similar to or greater than that of FAIRE-seq. Notably, FAIRE-seq benefited from two- to threefold greater sequencing depth, relative to DNase-seq (Supplementary Table 1). When tag counts were equalized by random downsampling, DNase-seq showed greater sensitivity at most thresholds, both at promoters and at nonpromoter regulatory regions (Supplementary Figs. 4 and 5). Thus, FAIRE-seq could potentially require greater sequencing depth than DNase-seq to achieve comparable sensitivity at nonpromoter regulatory elements.

Signal estimation using EFilter

Although signal detection is a major goal of sequencing-based functional assays, not all analysis tasks fit the detection paradigm. Consider, for example, the problem of using histone ChIP-seq data to predict gene expression²⁰. This is a 'signal estimation' problem, in which the observable variables (data) are used to estimate the value of a continuous hidden variable. As before, we asked, which is the best linear estimator? The answer is provided by standard linear regression theory, and relies, as before, on the covariance matrix of the tag-count signal (Online Methods). We developed the EFilter program to apply this technique to tag profiles of ChIP-seq data.

We first tested a naive version of EFilter by estimating mRNA levels from ChIP-seq data on five histone marks (H3K27ac, monomethylated histone H4 lysine 20 (H4K20me1), H3K4me3, H3K36me3 and acetylated H3 lysine 9 (H3K9ac)) from four cell lines (lymphoblastoid cell line GM12878, liver carcinoma line HepG2, Huvec and normal human epidermal keratinocyte line NHEK)^{13,17}. These histone marks were selected because they were the top five in terms of predictive

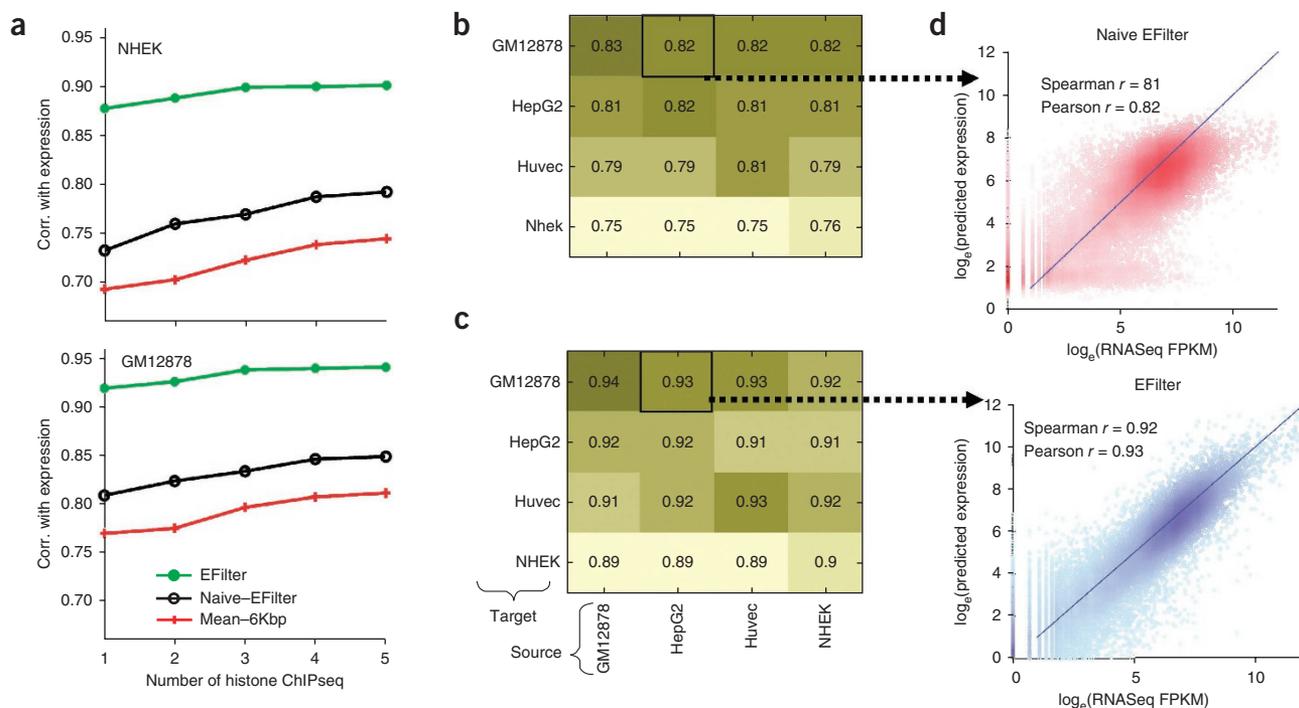


Figure 3 Gene expression prediction by EFilter using promoter histone ChIP-seq signals. Expression was predicted using up to five histone ChIP-seq data sets (in order of inclusion: H3K27ac, H4K20me1, K3K4me3, H3K36me3 and H3K9ac), and also measured as RNA-seq FPKM using ENCODE data. Prediction accuracy was quantified as the Pearson correlation coefficient between predicted and measured log-transformed expression levels. (a) Three linear regression approaches were used to predict expression. Red lines: prediction based on total promoter tag count in each ChIP-seq data set²⁰; black lines: naive EFilter; green lines: EFilter final version, which uses gene-specific residual errors between predicted and measured expression in three other cell lines. (b) Naive-EFilter expression prediction using the first three histone marks: 16 combinations of four source and four target cell lines. Each number in the grid represents the correlation between predicted and measured expression in the indicated target cell line (row) when expression of rank-invariant genes is imputed from the source cell line (column). (c) Same as in b, this time using the EFilter proper. (d) Illustration of prediction accuracy. Source: HepG2; target: GM12878.

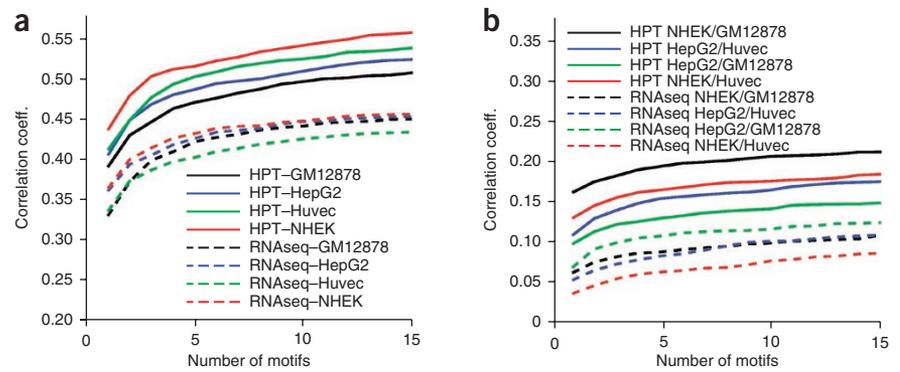
power (data not shown). We used ChIP-seq tag counts in 200-bp bins from 2 kbp upstream to 4 kbp downstream of RefSeq TSSs as predictor variables. Expression was calculated as log-transformed RNA-seq fragments per exonic kbp per million (log-FPKM (fragments per kbp per million))²¹. For comparison, we used an existing linear regression method²⁰ and predicted mRNA levels based on the total tag count of each ChIP-seq data set within the 6-kbp window around TSS. Training and test data were derived from the same cell line, using twofold cross-validation. Regardless of how many ChIP-seq data sets were used as input, the naive version of EFilter provided more accurate predictions (Fig. 3a, black and red curves).

In practice, one would like to predict mRNA levels in cell types for which genome-wide expression data are unavailable. However, the above implementation of EFilter required prior knowledge of expression levels in the target cell type, for training. We therefore devised a technique of imputing expression levels across cell types. Imputation works by assuming that genes with similar promoter histone modification levels in two cell types will also have similar levels of expression. We first identify a 'source' cell type for which mRNA levels and histone ChIP-seq are both available. Genes in the source and target cell types are ranked by total promoter ChIP-seq tag count, and the expression of rank-invariant genes is then imputed from the source to the target cell type (Online Methods). Imputed expression levels are then used to train the regression model in the target cell type. We tested this approach using three histone modifications (H3K27ac, H4K20me1 and H3K4me3) and all 16 possible combinations of four source and four target cell lines. Notably, the correlations between

predicted and measured expression were only marginally smaller when EFilter used imputed, rather than actual, expression values for training (Fig. 3b). Thus, EFilter can be used to predict expression levels *de novo* in any cell type for which histone ChIP-seq data are available.

The naive version of EFilter described above assumes a constant relationship between chromatin state and mRNA level across all genes. However, inaccurate gene models, variation in mRNA degradation rates, common technical biases and variations in the enzymatic activity of distinct PolII complexes could all cause EFilter to systematically over- or underestimate the expression of any individual gene. Such systematic biases would in general be consistent across cell types, and we reasoned that it should therefore be possible for EFilter to learn the aggregate bias from training cell types and then cancel it out in the test cell type. First, we ran the naive EFilter on individual cell types, with both source and target being the same, and calculated the residual error between predicted and measured mRNA level. Then, we devised a 'full' version of EFilter (Supplementary Fig. 6), which takes as input not only binned tag counts, but also gene-specific residual errors from three nontarget cell lines. This expanded set of predictor variables is then used to estimate mRNA levels in the target cell line, using the same linear regression approach as before. The full-fledged EFilter algorithm, which incorporates residual correction, showed markedly improved performance, yielding correlation coefficients in the range of 0.89–0.93 for all RefSeq genes (Fig. 3a, green curves; Fig. 3c,d) and 0.77–0.87 for genes with fivefold differential expression between source and

Figure 4 Transcription factor-binding motifs at promoters are more correlated with histone-predicted expression (HPT) than with expression measured by RNA-seq. HPT was calculated for each RefSeq gene by EFilter using H3K27ac and H4K20me1 ChIP-seq, with no residual correction. (a) Solid lines: HPT was modeled as a linear function of occurrence of TRANSFAC and JASPAR motifs (presence = 1, absence = 0) in promoter regions. Accuracy of the motif-based model of HPT (Pearson correlation) is shown for the top k motifs, as k is varied from 1 to 15. Dashed lines: motifs used to model RNA-seq FPKM. (b) Same as in a, except that motif occurrence was used to predict expression fold change, rather than absolute expression. To demonstrate generality, we calculated HPT using H3K36me3 and H3K20me1 ChIP-seq data in this case.



target cell lines (Supplementary Fig. 7). This improvement in accuracy derives from the fact that the combined effect of gene-specific biological and technical biases is largely consistent across multiple cell types ($R \sim 0.60\text{--}0.70$; Supplementary Fig. 8).

Recently, additional methods have been developed to predict expression from histone ChIP-seq^{22–25}. These methods were used to predict RNA-seq-based expression levels, using a large number of histone ChIP-seq data sets as input. The maximum accuracy achieved by these methods on K562 data was $R = 0.84$. In contrast, using only three K562 histone marks and the same tenfold cross-validation approach, EFilter achieved $R = 0.93$.

The correlation between gene expression and transcription factor motif occurrence at promoters is often used to infer the identity of master regulatory factors²⁶. We hypothesized that promoter histone modification profiles might show even greater correlation with transcription factor motif occurrence because transcription factor binding could be mechanistically ‘closer’ to chromatin state than to gene expression. To test this hypothesis, we estimated transcription factor presence or absence at promoters, based on thresholded scores of motifs from the TRANSFAC and JASPAR^{27,28} databases (above threshold = 1, below threshold = 0, for transcription factor binding). We used these transcription factor-binding estimates to predict expression in four cell types by multiple linear regression (Online Methods). Notably, in all four cell types, motif-based transcription factor binding estimates were more correlated with histone-predicted

transcript levels (HPT) than with actual mRNA levels measured by RNA-seq (Fig. 4a). This conclusion was robust to alterations in the choice of histone modification (Supplementary Fig. 9), and also valid when motif occurrence was used to predict fold change, rather than absolute expression (Fig. 4b). Note that, in these tests, we calculated HPT without residual correction. Incorporation of residuals renders HPT more similar to mRNA level (Fig. 3c), but less correlated with motif occurrence in promoters (data not shown). These results consistently imply that the occurrence of transcription factor motifs in promoters is more correlated to chromatin states than to mRNA level. Consequently, if one is interested in inferring the identity of key regulatory transcription factors, chromatin profiling could be more predictive than expression profiling.

Application to a complex tissue: developing mouse forebrain

The analyses described thus far have used as input functional data from pure cell lines or purified cell populations. However, analysis of *in vivo* gene regulation in heterogeneous tissue samples is more challenging because the genomic signals of individual cell types are diluted in the cellular mixture, resulting in a lower overall signal-to-noise ratio. We hypothesized that DFilter would provide greater benefit in such circumstances because the proportion of ‘marginal’ signals that could be mistaken for noise by suboptimal algorithms would be greater.

To test our bioinformatic methods on noisier data, and also more generally to assess the utility of chromatin profiling in a complex

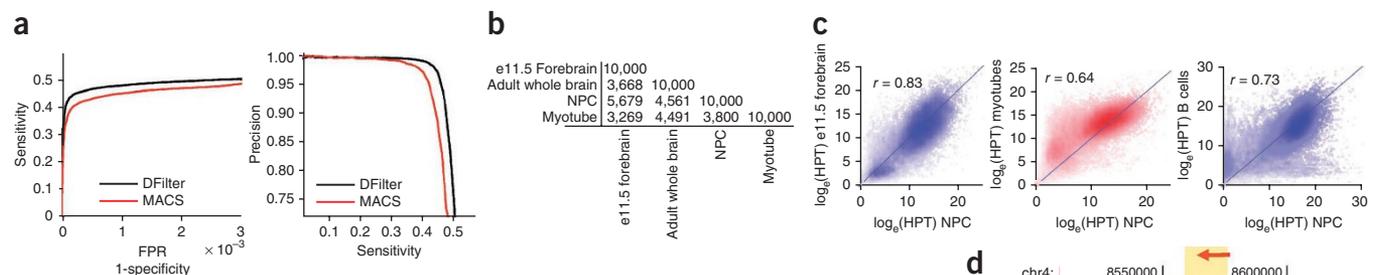
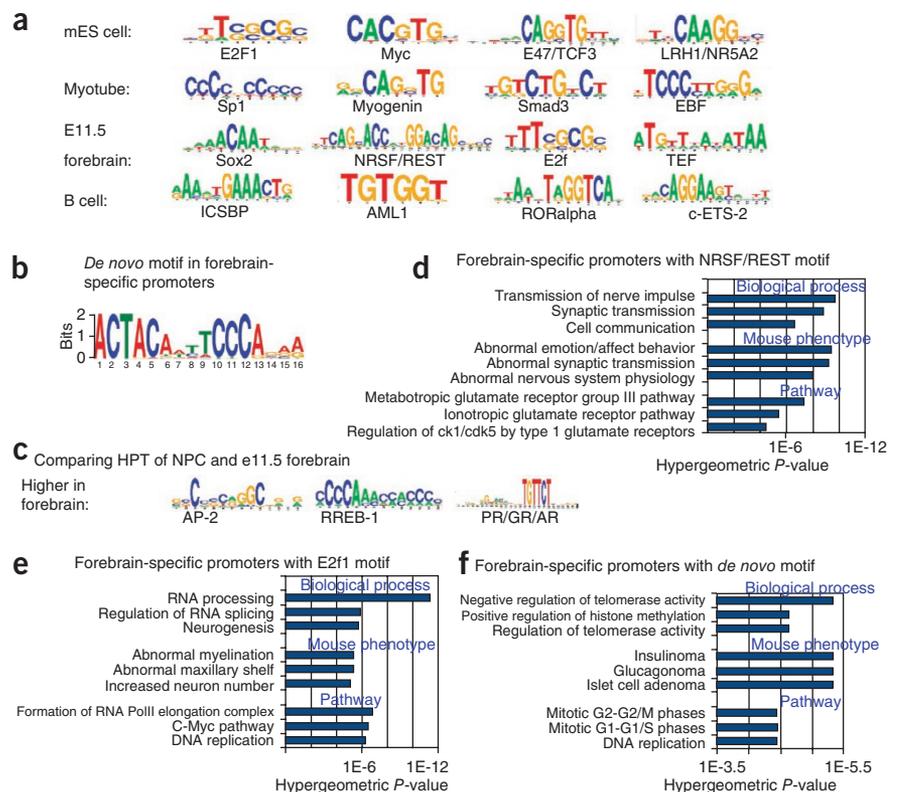


Figure 5 *In vivo* functional profiling in a heterogeneous tissue (e11.5 mouse forebrain).

(a) Accuracy of RefSeq promoter prediction using peak calls from forebrain H3K4me3 ChIP-seq data. FPR: false-positive rate. (b) Number of overlapping H3K4me3 peaks among the top 10,000 peak calls (by DFilter) from four sample types. NPCs and e11.5 forebrain showed the greatest overlap, suggesting that the functional signature of neural progenitors in the embryonic forebrain can be discerned, to some extent, despite tissue heterogeneity. (c) Comparison of predicted expression (HPT) between NPCs and other samples. Expression was predicted using H3K4me3 and H3K36me3 ChIP-seq data. Compared with other samples, NPCs had the highest correlation with e11.5 forebrain. (d) Visualization on UCSC browser (<http://genome.ucsc.edu/>) of previously unreported gene discovered using e11.5 forebrain ChIP-seq profiles. Forebrain expression of this gene was validated using qPCR (Supplementary Fig. 11).

Figure 6 Transcription factor inference from ChIP-seq on pure and heterogeneous samples. Expression was predicted in e11.5 forebrain and seven cell lines by EFilter based on H3K4me3 and H3K36me3 ChIP-seq data. Expression levels of rank-invariant genes were imputed from ES-cell RNA-seq data. Expression fold-change relative to the median in the eight samples was modeled using motif (from TRANSFAC and JASPAR^{27,28}) occurrence at promoters, to predict the most relevant transcription factors. (a) The top four predictive transcription factor motifs are shown for mouse e11.5 forebrain, mouse ES (mES) cells, myotube cells and human B cells. (b) The *de novo* motif detected in promoters of the top 3,000 genes with forebrain-specific predicted expression (HPT). (c) Three motifs most predictive of HPT upregulation in forebrain relative to NPC. The third motif belongs to steroid hormone receptors (NR3C1-4), namely, progesterone receptor (PR), glucocorticoid receptor (GR) and androgen receptor (AR). (d–f) Enriched annotations of genes with forebrain-specific HPT that contain a specific motif (REST, E2f or *de novo* motif) within the promoter.



developing tissue, we carried out H3K4me3 and H3K36me3 ChIP-seq on pooled, micro-dissected mouse forebrain at embryonic day 11.5 (e11.5). We first used the H3K4me3 data to test the accuracy of forebrain promoter prediction and found a substantial performance gap between DFilter and MACS (Fig. 5a). DFilter provided similar, or even greater, improvements when tested on data from other heterogeneous tissues²⁹ (Supplementary Fig. 10). Notably, the difference in accuracy between DFilter and MACS was greater than previously observed in tests on *in vitro* data from pure cell populations (Fig. 1b,c). These results suggest that the need for optimal signal processing is greater when genomic signals are noisy or diluted by cellular heterogeneity.

Inferring the identity of master transcription factors from histone-predicted transcript levels could be feasible even in ChIP-seq data from heterogeneous samples. Such transcription factor inference would be most effective if the heterogeneous-tissue chromatin signature was correlated with the signatures of the major constituent cell types. As neural progenitor cells (NPCs) are common in developing forebrain³⁰, we asked if the H3K4me3 signal of NPCs³¹ could be detected in our forebrain ChIP-seq data. We examined H3K4me3 ChIP-seq data from whole adult mouse brain³² (in which NPCs are less common)³⁰, and also myotube cells³³ (in which no NPCs are expected), as controls. We hypothesized that NPC promoters would show high overlap with the e11.5 forebrain promoter set, and only moderate overlap with adult brain promoters, and indeed this was the case (Fig. 5b). The overlap with myotube promoters was the smallest, and presumably dominated by ubiquitously expressed genes. We further evaluated signal correlation by using EFilter to calculate gene expression (HPT) in e11.5 forebrain, NPCs and myotube cells (Fig. 5c). Again, we found that NPC gene expression showed substantially higher correlation with embryonic forebrain than with the other samples. Overall, these results suggest that, despite signal mixing in the heterogeneous forebrain sample, it should be possible to infer transcription factors regulating the major forebrain cell types.

As a quality check before using HPT to infer e11.5 forebrain transcription factors, we tested the qualitative concordance between HPT

and forebrain mRNA levels of nine genes (RT-PCR; Supplementary Table 2). Four genes with extremely low HPT (more than tenfold below median) were found to have no expression in forebrain, whereas five genes with high HPT (more than fivefold above median) showed high expression (Supplementary Fig. 11). The latter included one novel gene (Fig. 5d and Supplementary Fig. 11) and two poorly characterized loci associated with neurodevelopmental disorders^{34,35}. Having qualitatively validated forebrain HPT, we used EFilter to predict mRNA levels in seven additional cell types (NPC, embryonic stem (ES) cell, mouse embryonic fibroblast, B-cell, myoblast, myotube and 3T3-L1 pre-adipocyte), based on H3K4me3 and H3K36me3 ChIP-seq data from previous studies^{31,33,36,37}. Sample-specific gene expression was estimated as the log-ratio relative to the median of the eight samples. We then performed greedy multiple linear regression to identify the four transcription factor-binding motifs from TRANSFAC and JASPAR^{27,28} that best explained sample-specific expression (Fig. 4b). As expected, this analysis revealed known regulators of the relatively pure cell types (Fig. 6a and Supplementary Table 3). Four known regulators of forebrain development, Sox2, Rest, E2f and Tef, were also accurately identified. Moreover, *de novo* motif detection at forebrain-specific promoters revealed an 'extended ZNF143 motif' (ACTACAAnnTCCCAAnRR; Fig. 6b) that was previously found at ZNF143, RBP-J and Notch1 binding sites in human T cells³⁸. This result suggests a novel molecular mechanism for Notch signaling in forebrain development, potentially mediated by ZNF143 and RBP-J binding at the detected motif. Thus, HPT-based inference of master transcription factors is effective in pure cell lines and also in complex tissues *in vivo*.

As NPCs are a major constituent of e11.5 forebrain, it is possible that their chromatin signal might obscure the signals of other forebrain cell types. We therefore asked whether master regulators of non-NP forebrain cells could be inferred from the *in vivo* ChIP-seq data. To this end, we regressed the log ratio of forebrain (*in vivo*) and NPC

(*in vitro*) HPT against TRANSFAC and JASPAR motif occurrence at RefSeq gene promoters. The three motifs most predictive of upregulation in forebrain relative to NPCs (Fig. 6c) were AP-2, RREB1 and steroid hormone receptor (NR3C1-4). AP-2 is not expressed in NPCs, but plays a key role in neural crest development³⁹. RREB1 and NR3C1 (glucocorticoid receptor) are both upregulated (fold change >1.5) in e11.5 forebrain relative to NPCs, based on HPT estimates. These results suggest that non-NPC regulatory signals are also discernible in the forebrain ChIP-seq data.

To explore the functional roles of the inferred transcription factors, we defined their regulatory targets in forebrain as promoter sequences that contained a motif match and showed forebrain-specific HPT. Using the GREAT⁴⁰ tool, we found that synaptic transmission and glutamate receptor pathway genes were enriched in the target gene set for Nrsf (Fig. 6d), which is consistent with the known role of Nrsf in brain development⁴¹. E2f targets were also enriched for functional annotations matching the known neurodevelopmental functions of the E2f family⁴² (Fig. 6e). Notably, Notch1 targets that were predicted by using the *de novo* motif highlighted control of telomerase activity and cell cycle, both functions previously linked to Notch signaling^{43,44} (Fig. 6f). Thus, *in vivo* chromatin profiles indicate not merely the identity but also the functional role of key transcription factors in the target tissue, despite cellular heterogeneity.

DISCUSSION

The data analysis problems associated with sequence-tag profiles are not unique to bioinformatics. Rather, they are special cases of more general problems that have long been studied in the field of signal processing. We have shown that a single ROC-AUC optimizing algorithm (Dfilter) can detect functional signals in tag profiles from diverse assays (e.g., histone ChIP-seq, TF ChIP-seq, DNase-seq and FAIRE-seq), and outperform more specialized algorithms in each case. Importantly, Dfilter extends seamlessly to optimal integrative analysis of multiple data sets, as we demonstrated by combining DNase-seq and FAIRE-seq data to detect open chromatin. Dfilter should also be applicable to other signal detection problems, such as Sono-seq, CLIP-seq and ChIP-exo^{1,45,46}.

The generality of Dfilter allowed unbiased comparison of DNase-seq and FAIRE-seq signals on a benchmark set of *cis*-regulatory regions. Given equal sequencing depth, DNase-seq showed greater sensitivity at a broad range of regulatory sites, though the difference was most pronounced at promoters. Of course, this must be balanced against the relative simplicity of the FAIRE-seq protocol, which can yield comparable sensitivity at nonpromoter regions, given greater sequencing depth.

Our signal estimation method, Efilter, consistently outperformed the sum-of-tags approach²⁰ at predicting mRNA levels from histone ChIP-seq data. Efilter learns systematic gene-specific effects that are correlated across different cell lines, and suffers minimal loss of accuracy (1–2%) when trained on data from unrelated cell types. Efilter achieved correlation coefficients as high as $R = 0.93$ between predicted and measured expression, indicating that gene expression is highly correlated with chromatin state once gene-specific effects are taken into account. Two to three histone ChIP-seq data sets were sufficient to achieve near-maximal correlation with mRNA levels, and even a single suitably chosen data set (H3K27ac) could provide accurate predictions.

There is a vast literature^{26,47} on using transcriptomic data in conjunction with promoter motif analysis to infer the identity of key upstream regulatory factors (that is, transcription factors). These methods are based on the assumption that expression levels are

correlated with the occurrence of relevant transcription factor-binding motifs in the promoter region. We observed that this correlation was even stronger when we replaced actual mRNA levels with chromatin-based expression predictions. As motif occurrence influences both histone marks and expression through transcription factor–DNA binding, our results suggest that transcription factor binding to promoters is more tightly correlated with chromatin state than with mRNA level. Consequently, motif-based regulatory transcription factor inference should ideally be done using histone-predicted, rather than actual, expression.

Tissue heterogeneity is a common cause for concern for *in vivo* sequencing-based functional studies. However, our results demonstrate that ChIP-seq can be used to identify key regulatory transcription factors even in heterogeneous samples. This is also supported by a ChIP-seq study of tissues from e14.5 and adult mouse embryos²⁹. Furthermore, transcription factor functions can also be inferred from the same data, through analysis of enriched target-gene annotations. Notably, we uncovered an ACTACAnnTCCAnRR motif in the promoters of mouse forebrain-specific genes, which may mediate the known effects of Notch signaling on telomerase activity and cell cycle control during brain development^{43,44}. Although the exact degree of heterogeneity in e11.5 forebrain (or in e14.5 anatomical structures) is not known, it is possible that other samples could be even more heterogeneous, and therefore refractory to the methods described here. In such cases, cell sorting might be needed before functional analysis. Moreover, our approach does not directly address the issue of determining the cell-type specificity of the transcription factors inferred to be active in heterogeneous samples, which remains a challenge. Nevertheless, our results demonstrate that ChIP-seq on heterogeneous tissues can yield substantial mechanistic insights.

METHODS

Methods and any associated references are available in the [online version of the paper](#).

Accession codes. ChIP-seq data from e11.5 mouse forebrain, GEO: GSE34073). The Efilter and Dfilter tools are available at <http://collaborations.gis.a-star.edu.sg/~cmb6/kumarv1/dfilter/> and **Supplementary Data File**.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This study was supported by funds from the Agency for Science, Technology and Research (A*STAR), Singapore including grant JCOAG03_FG02_2009 from Joint Council Office of A*STAR. The authors would like to thank N. Clarke and N. Nagarajan for valuable feedback and comments on the manuscript.

AUTHOR CONTRIBUTIONS

V.K. and S.P. designed the study, analyzed the results and wrote the manuscript. V.K. developed the computational algorithms and performed bioinformatic analyses. N.A.R. and M.M. contributed to the drafting of the manuscript. M.M., N.A.R. and P.K. performed wet-lab experiments on *in vivo* samples. T.L. and H.H.N. guided the design and execution of wet-lab experiments.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Hawkins, R.D., Hon, G.C. & Ren, B. Next-generation genomics: an integrative approach. *Nat. Rev. Genet.* **11**, 476–486 (2010).
- Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129–141 (2010).

3. Song, L. *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* **21**, 1757–1767 (2011).
4. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* **5**, 829–834 (2008).
5. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
6. Boyle, A.P., Guinney, J., Crawford, G.E. & Furey, T.S. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**, 2537–2538 (2008).
7. Rashid, N.U., Giresi, P.G., Ibrahim, J.G., Sun, W. & Lieb, J.D. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol.* **12**, R67 (2011).
8. Lun, D.S., Sherrid, A., Weiner, B., Sherman, D.R. & Galagan, J.E. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. *Genome Biol.* **10**, R142 (2009).
9. Wilbanks, E.G. & Facciotti, M.T. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* **5**, e11471 (2010).
10. Han, Z. *et al.* A signal processing approach for enriched region detection in RNA polymerase II ChIP-seq data. *BMC Bioinformatics* **13** (suppl. 2), S2 (2012).
11. Hotelling, H. The generalization of student's ratio. *Ann. Math. Stat.* **2**, 360–378 (1931).
12. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
13. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
14. Fejes, A.P. *et al.* FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* **24**, 1729–1730 (2008).
15. Zang, C. *et al.* A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952–1958 (2009).
16. Chepelev, I., Wei, G., Tang, Q. & Zhao, K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res.* **37**, e106 (2009).
17. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
18. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
19. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
20. Karlic, R., Chung, H.R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. USA* **107**, 2926–2931 (2010).
21. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
22. Cheng, C. *et al.* A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol.* **12**, R15 (2011).
23. Dong, X. *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* **13**, R53 (2012).
24. Cheng, C. *et al.* Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.* **22**, 1658–1667 (2012).
25. Wang, C. *et al.* Computational inference of mRNA stability from histone modification and transcriptome profiles. *Nucleic Acids Res.* **40**, 6414–6423 (2012).
26. Segal, E., Friedman, N., Kaminski, N., Regev, A. & Koller, D. From signatures to models: understanding cancer using microarrays. *Nat. Genet.* **37**, S38–S45 (2005).
27. Wingender, E., Dietze, P., Karas, H. & Knuppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24**, 238–241 (1996).
28. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94 (2004).
29. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).
30. Kawaguchi, A. *et al.* Nestin-EGFP transgenic mice: visualization of the self-renewal and multipotency of CNS stem cells. *Mol. Cell Neurosci.* **17**, 259–273 (2001).
31. Mikkelsen, T.S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
32. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).
33. Asp, P. *et al.* Genome-wide remodeling of the epigenetic landscape during myogenic differentiation. *Proc. Natl. Acad. Sci. USA* **108**, E149–E158 (2011).
34. Anthoni, H. *et al.* A locus on 2p12 containing the co-regulated MRPL19 and C2ORF3 genes is associated to dyslexia. *Hum. Mol. Genet.* **16**, 667–677 (2007).
35. Need, A.C. *et al.* A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB. *Hum. Mol. Genet.* **18**, 4650–4661 (2009).
36. Daniel, J.A. *et al.* PTIP promotes chromatin changes critical for immunoglobulin class switch recombination. *Science* **329**, 917–923 (2010).
37. Mikkelsen, T.S. *et al.* Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156–169 (2010).
38. Wang, H. *et al.* Genome-wide analysis reveals conserved and divergent features of Notch1/RBPJ binding in human and murine T-lymphoblastic leukemia cells. *Proc. Natl. Acad. Sci. USA* **108**, 14908–14913 (2011).
39. Gammill, L.S. & Bronner-Fraser, M. Neural crest specification: migrating into genomics. *Nat. Rev. Neurosci.* **4**, 795–805 (2003).
40. McLean, C.Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
41. Su, X., Kameoka, S., Lentz, S. & Majumder, S. Activation of REST/NRSF target genes in neural stem cells is sufficient to cause neuronal differentiation. *Mol. Cell Biol.* **24**, 8018–8025 (2004).
42. Swiss, V.A. & Casaccia, P. Cell-context specific role of the E2F/Rb pathway in development and disease. *Glia* **58**, 377–390 (2010).
43. Ferron, S.R. *et al.* Telomere shortening in neural stem cells disrupts neuronal differentiation and neurogenesis. *J. Neurosci.* **29**, 14394–14407 (2009).
44. Hitoshi, S. *et al.* Notch pathway molecules are essential for the maintenance, but not the generation, of mammalian neural stem cells. *Genes Dev.* **16**, 846–858 (2002).
45. Auerbach, R.K. *et al.* Mapping accessible chromatin regions using Sono-Seq. *Proc. Natl. Acad. Sci. USA* **106**, 14926–14931 (2009).
46. Rhee, H.S. & Pugh, B.F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–1419 (2011).
47. Ochoa-Espinosa, A. & Small, S. Developmental mechanisms and cis-regulatory codes. *Curr. Opin. Genet. Dev.* **16**, 165–170 (2006).

ONLINE METHODS

Constructing and normalizing the tag profile signal. To reduce computational memory requirements, we segmented the genome into nonoverlapping n -bp bins of equal width, and the whole-genome sequence tag profile was defined as the number of tags in each genomic bin. This binning approach has no substantial impact on interpretability, provided the bin size remains smaller than the characteristic length scale of the signal of interest. We selected a bin size of 200 bp for EFilter and 100 bp for DFilter, with a second-pass refined bin size of 5 bp when DFilter was used to analyze TF ChIP-seq data. On the assumption that the sequenced DNA fragments were ~200 bp in length, sequence tags were shifted by 100 bp in the direction of their alignment to the reference genome, and then counted in the bin containing the shifted tag start site. A more tailored approach that estimates mean fragment length from the data and shifts tags by half the estimated length was used at a bin size of 5 bp. Multiple duplicate tags with the same orientation and starting base-pair position were replaced by a single tag.

Whenever a control data set was available, the tag-count signal of interest was divided by the corresponding control signal to minimize artifactual peaks caused by repetitive sequences and corrected for GC content biases. In order for this normalization approach to be robust, the control signal must be smoothed over a length scale that includes a sufficient number of tags. We set this minimum tag number at 20. For each genomic bin, the control tag density was estimated within a 1-kbp window centered on the bin, if the 1-kbp window contained at least 20 tags. If not, a 5-kbp window was used, or a 10-kbp window, if even the 5-kbp window was insufficient to accumulate 20 tags. When even this was insufficient, the control tag density was set to a pseudo-count value of 20 tags per 10 kbp. Thus, the smallest value of the denominator during control normalization would be 0.2 tags/bin (assuming 100-bp bins). Control data sets were used for transcription factor and histone ChIP-seq analysis, where available. No control data were available for DNase-seq and FAIRE-seq libraries.

ROC-AUC maximizing detection filter (DFilter). DFilter takes as input a set of sequence tags mapped to a reference genome. Based on the genomic distribution of tags, the algorithm classifies individual n -base-pair bins as positive (signal) or negative (noise) regions. For example, if FAIRE-seq data was supplied as input, DFilter would be used to detect open chromatin regions in the genome. For efficiency, DFilter implements linear finite-impulse-response detection, that is, a windowed linear filter h of user-specified width, followed by the standard thresholding step. If the user does not specify a filter width, DFilter learns the filter width from an average signal profile of a set of positives. The problem of designing a linear detector that maximizes accuracy, as defined by the ROC-AUC, has a well-known solution in the field of signal processing¹¹. The solution relies on an interesting property of the ROC-AUC, namely that it equals the probability of observing a higher detector output at a signal region than at a noise region. Under the Gaussian-noise approximation, maximizing the latter probability is equivalent to maximizing a z -score, the square of which can be thought of as the square root of the signal-to-noise ratio of the detector. The Hotelling observer¹¹, also known as the pre-whitening matched filter, is the filter that maximizes this z -score, and thus maximizes ROC-AUC. DFilter first estimates the requisite means and covariance matrices of the input sequence tag data, calculates the form of the optimal detection filter (Hotelling observer) and then applies this filter to the tag profile. The filtered profile can then be thresholded to detect genomic regions of interest, such as promoters.

Given the observed signal vectors at positive regions, x_p , and at negative regions, x_n , with variances R_p and R_n , respectively, the z -score of the positives relative to negatives can be written as

$$Z = \frac{E(hx_p - hx_n)}{\sqrt{(h^H R_p h + h^H R_n h)/2}} = \frac{E(hx_p) - E(hx_n)}{\sqrt{(h^H R_p h + h^H R_n h)/2}}$$

where h is the linear filter and the superscript H represents the Hermitian transpose. Here, h is a row vector and x_p and x_n are column vectors, and each vector spans a user-specified number of genomic bins. The term z -score corresponds to signal-to-noise ratio for the measure of detection. The optimal linear

detection-filter that maximizes the above z -score, and therefore maximizes ROC-AUC under the Gaussian noise approximation, can be derived as

$$h = K^{-1}(E(x_p) - E(x_n))$$

where K is the mean of covariances of observations at positive and negative regions such that $K = (R_p + R_n)/2$.

Designed in this manner, the linear filter h is known as the Hotelling observer, or pre-whitening matched filter^{11,48}. Ideally one would desire an infinite amount of data in order to calculate the exact covariance matrix K . Its estimation using a limited amount of data is a classical signal processing problem, conventionally solved with help of information-theoretic dimension reduction criteria like AIC, MDL or EDC^{49,50}. In our tests, the EDC2 criterion⁵⁰ was the most robust (data not shown), and we therefore used it to estimate the reduced-dimension covariance matrix, and then to invert the matrix.

When multiple signals (next-generation sequencing data sets) are to be integrated and used jointly for signal detection, the optimal filter can be straightforwardly derived by conceptually concatenating the individual data vectors and estimating their joint co-variance matrix. The concatenated signal can be written as $X = (x_1; x_2)$, where x_1 and x_2 denote signal vectors from the two data sets that cover a given genomic region. The ROC-AUC maximizing filter h_c for the combined signal can be derived exactly as before, so that

$$h_c = K^{-1}(E\{X_p\} - E\{X_n\})$$

where K is again the mean of the covariance matrices of the observed concatenated signals at positive (X_p) and negative (X_n) regions. Note that this concatenation approach does not entail any loss of generality or optimality, as it merely replaces a “two-dimensional” (that is, two-row) coefficient matrix with a single-row matrix having twice as many columns.

When used to detect asymmetric, directional features such as promoters, DFilter was trained on positive regions aligned in the same direction, resulting in an asymmetric filter shape. This asymmetric filter was applied to chromatin profiles in both directions, and the larger of the two filter outputs was assigned to each genomic bin. Some chromatin profiles (e.g., H3K4me3) are characterized by focal peaks at regulatory elements and also, occasionally, broad plateaus of enrichment over gene bodies. Copy-number-amplified regions in karyotypically aberrant cell lines such as K562 could also result in broad plateaus of elevated signal. In such cases, when detecting focal peaks at regulatory elements, it is desirable to filter out the broad plateaus, since they could create artifactual peaks. This can be done by applying a pre-filter to the data that suppresses broad signal features (that is, a ‘high-pass’ filter) before further processing by DFilter. The same outcome can also be achieved by modifying the filter coefficients of DFilter. The simplest modification that achieves this goal is to shift all of the filter coefficients down until their sum becomes zero. We set the sum of the DFilter filter coefficients to zero when detecting peaks in H3K4me3, DNase-seq and FAIRE-seq data. In contrast, for TF ChIP-seq and H3K36me3, filters were used as is, without setting the coefficient sum to zero.

When positive regions are not defined by the user, DFilter smoothens the tag profile using a 1-kbp smoothing window, and prioritizes all bins with a smoothed score greater than two s.d. above the mean as potential positives. Sets of prioritized bins within 1 kbp of each other are merged to define candidate regions, and the top 5% of such regions by smoothed peak height are discarded as outliers. The top 1,000 regions among the remaining candidates are defined as positives, aligned by their central bin and trimmed to the user-specified width. The corresponding 1,000 unsmoothed tag-count vectors of fixed length are used to define x_p and R_p . Negative regions of fixed width are chosen at random from the genome, based on closeness to the background signal level, and used to define x_n and R_n .

After defining positive and negative training sets, DFilter calculates the coefficients of the Hotelling detection filter and then applies the filter to the genome-wide binned tag-count signal. The filtered tag-count signal is then thresholded to detect significant genomic regions. The coordinates of predicted regions are chosen so as to include all contiguous bins that score above the threshold, and regions separated only by a short gap are merged

(see “Details of data analysis” below). The score of a region is defined as the highest value of the filtered signal in the underlying bins. Given sufficient sequencing depth, the central limit theorem states that the scores of noise regions will closely follow a Gaussian distribution. In practice, we have found that this is indeed a reasonable assumption. We therefore fit a Gaussian curve to the central region of the observed distribution of bin scores in the genome, and use the fitted distribution to estimate the *P* values of predicted positive regions. **Supplementary Table 4** provides an overview of computational time taken by DFilter in comparison to other methods.

Mathematical aspects of linear estimation method (EFilter). We first describe the simplest case, in which EFilter uses a single sequence tag profile (e.g., ChIP-seq) to predict a continuous-valued parameter (gene expression, defined as log-RPKM). If there are *n* genes in the genome, and x_i is the column vector of binned ChIP-seq tag counts within a genomic window containing the TSS of gene *i*, then the data matrix can be written as $X = [x_1, x_2, \dots, x_n]$. If the genomic window used to estimate gene expression extends from bin $-m1$ to bin $m2$ relative to the TSS, then the EFilter linear regression coefficients can be written as the row vector

$$F = [f(-m1), f(-m1 + 1), \dots, f(m2)]$$

and the row vector *Y* representing the log-expression values of the *n* genes can be written as

$$Y = FX + v$$

where *v* is the error vector. It can be shown that the value of *F* that minimizes the sum of squared prediction errors $v^H v$ is given by

$$F = YX^H (XX^H)^{-1}$$

As in the case of DFilter, we need to estimate the covariance matrix XX^H and its inverse from the limited number of samples available in the training data, and as before, we use the EDC2 (ref. 50) criterion to derive robust matrix estimates. Again, as in the case of DFilter, multiple predictive data sets (ChIP-seq data on 3 different histone modifications, for example) can straightforwardly be incorporated into the signal estimation formalism by concatenating the multiple data vectors for a single gene and then using the same equation as above to calculate the optimal concatenated regression coefficients. The EFilter and Dfilter tools are available at <http://collaborations.gis.a-star.edu.sg/~cmb6/kumarv1/dfilter/> and **Supplementary Data File**.

Embryonic mouse forebrain ChIP-seq library preparation. Mouse embryonic forebrain tissues were microdissected and pooled from timed pregnant matings for e11.5 CD1 (Hsd:ICR(CD-1)) embryos, with e0.5 being the morning a vaginal plug was detected. All animal procedures were done according to the Singapore A*STAR Biopolis Biological Resource Center (BRC) Institutional Animal Care and Use Committee (IACUC Protocol No: 100560). Dissected tissues were mechanically dissociated by passing through a cell strainer and fixed in 1% formaldehyde (diluted from 37% stock solution, Sigma, F8775) simultaneously following pre-treatment with 0.125% Trypsin/Versene solution. 36 forebrain pieces were used for each ChIP-seq library preparation. Cells were resuspended in SDS buffer and sonicated 12 times for 30 s with 30 s intervals with Bioruptor water bath sonicator (Diagenode). Following sonication, samples were diluted with IP dilution buffer and incubated with 50 μ l affinity resin coupled with 10 μ l anti-H3K4me3 antibodies (Upstate, Cat# 07-473, Lot# 32497). Following washing steps, chromatin was reverse-crosslinked for purification of DNA.

H3K4me3 ChIPed DNA and 0.5% input DNA samples were amplified for 8 cycles with GenomePlex Single Cell Whole Genome Amplification Kit (WGA4, Sigma) using universal primer linked to BpmI restriction site. Amplified DNA samples were digested with BpmI to remove universal primer. After quantification of DNA with Quant-iT PicoGreen sdDNA Assay Kit (Invitrogen, P7589), 12 ng of the DNA sample was directly used for Illumina sequencing adaptor ligation. Illumina sequencing was performed using GAII platform in GIS. In order to validate reproducibility of our ChIP-seq protocol, ChIP-seq was

performed for two biological replicates for H3K4me3 (**Supplementary Fig. 11**). H3K36me3 data were obtained similarly using 15 μ l of anti-H3K36me3 antibody (Abcam, Cat# Ab9050, Lot# 707981). As the DNA yield of H3K36me3 ChIP was much higher, the WGA amplification step was not required, and ChIP DNA was directly used for Illumina sample preparation.

RT-PCR based expression validation. Total RNA was isolated from dissected mouse e11.5 forebrain tissues using Trizol (Invitrogen) according to the manufacturer’s instructions. Random primed cDNA synthesis was performed on 2 μ g of total RNA using Multiscribe reverse transcriptase (Applied Biosystems) in accordance with the manufacturer’s protocol. qRT-PCRs were performed on a 7900HT ABI platform using 2X SYBR green master mix (ABI). Analysis of relative mRNAs expression levels was performed by the $2^{-\Delta\Delta CT}$ method with beta-actin as the internal control for normalization. Primers are listed in **Supplementary Table 2**.

Data sources and genome assemblies. All of the human sequence tags analyzed here were mapped to the hg18 assembly of the human genome. CD4⁺ T cell ChIP-seq data were derived from previous work¹². ChIP-seq data and peak calls for K562, GM12878, NHEK, Huvec and HepG2 cells were produced by an ENCODE consortium group¹³ and downloaded via the UCSC genome browser (<http://genome.ucsc.edu/>). DNase-seq and FAIRE-seq analysis was based on previous sequence³, including matched PolII ChIP-seq peak data. Analysis of CTCF peak coverage by FAIRE-seq and DNase-seq was based on ENCODE CTCF peak calls from Broad Institute ChIP-seq data¹³. CD4⁺ T cell RNA-seq data were from previous work¹⁶, whereas RNA-seq data from GM12878, K562, Huvec, HePG2 and NHEK cells generated by Caltech group of ENCODE consortium¹⁷, were downloaded from the UCSC browser.

ChIP-seq data from mouse ES, MEF, NP cells and adult brain^{31,32} were downloaded from the UCSC browser. The mouse ES cell TF ChIP-seq datasets were derived from previous work¹⁸. The downloaded tag locations for H3K4me3 and H3K36me3 ChIP-seq from myotube, myoblast and 3T3 cells^{33,37} were mapped from the mm9 mouse genome assembly to mm8. RNA-seq data for mouse ES cells were adapted⁵¹. The H3K4me3 ChIP-seq data from mouse lung, heart and spleen were adapted²⁹.

Details of data analysis. DFilter trains itself for each data set separately. It does not use the same filter for all cell types or data types as the noise level or signal pattern can vary for different data sets. While processing H3K4me3 ChIP-seq data, DFilter used a Hotelling detection filter (with mean as zero) of width 100 bins (10 kbp); however, for H3K36me3 ChIP-seq it used a filter of 30 bins (3 kbp) without making its mean as zero. Hence a non-zero-mean filter of width 25–40 bins (2.5 kbp–4 kbp) could be suitable for detecting wide pattern in ChIP-seq datasets such as for H3K36me3 and H3K27me3 histone modifications. When processing DNase-seq and FAIRE-seq data, the filter (with mean as zero) width was 50 bins (5 kbp, 1 bin = 100 bp). For TF ChIP-seq analysis, DFilter initially used a bin width of 100 bp, and a filter width of 20 bins (2 kbp) in the first round of detection. Then in the second pass, the bin width was narrowed to 5 bp for improved resolution, and a second Hotelling detection filter of width 40 bins (200 bp) was applied. In the second pass DFilter also re-estimates the length of fragments so that the detected peak centers are found with high accuracy. For promoter detection, peaks in the filter output whose edges were within 1 kbp of each other were merged, whereas for FAIRE-seq and DNase-seq analysis, predicted open chromatin regions were merged only if they lay within 400 bp of each other. Input control data were used for H3K4me3 ChIP-seq data sets except for CD4⁺ T cells. Similarly input control was used for TF ChIP-seq and H3K36me3 ChIP-seq analysis. No input control was used for DNase-seq and FAIRE-seq data sets.

For detecting peaks in H3K4me3 ChIP-seq data, F-seq (ref. 6) was used with default parameters. MACS⁵ was used with *mfold* = 10, and peaks were ranked by *P* value. SICER¹⁵ was used with *window size* = 200 bp, *gap size* = 400 bp and *FDR* = 0.01. Findpeaks¹⁴ was used with *dist_type* = 1 200 (triangular distribution) and *eff_size* = 0.8, with the remaining options at their default settings. For peak-detection in TF ChIP-seq, QuEST⁴ was used with default parameters, except that the detection threshold was lowered in order to call a larger number of peaks. For DNase-seq and FAIRE-seq analysis, F-seq and ZINBA peaks

were downloaded from the UCSC browser. For histone ChIP-seq data analysis ZINBA⁷ was used using the pipeline function provided by authors.

For **Figure 1b–c** the RefSeq promoters whose genes had RNA-seq based FPKM level above median were defined as active. For measuring performance of promoter detection, those predicted sites were considered positives that were within 400 bp of RefSeq gene TSS. All peaks wider than 20 kbp were truncated to a width of 20 kbp. In order to determine the false-positive rate for ROC analysis, negative (nonpromoter regions) were defined as 2-kbp regions that lay at least 2.5 kbp from the nearest RefSeq TSS, UCSC KnownGene TSS or CpG island.

The performance test for H3K36me3 peaks was done by checking the average coverage of genes according to their expression level. In order to avoid bias due to different peak widths, for every method only that many top-ranking peaks were allowed which covered up to 10% of bins in the total genome.

To define an independent benchmark set of positive regions for evaluating open chromatin detection, the union of peaks from H3K4me1, CTCF and PpII ChIP-seq was used. Alternatively, we used peaks in H3K27ac ChIP-seq as positive regions. In order to avoid biasing the benchmark set in favor of DFilter, we used externally generated peak calls for these data sets. ENCODE peaks in H3K4me1, H3K27ac and CTCF ChIP-seq called in previous work¹³ were downloaded from the UCSC browser. Similarly, PolII peaks called using F-seq³ were also downloaded from the same source. A positive region from the benchmark set was counted as detected if it lay within 400 bp of the center (summit) of a FAIRE-seq or DNase-seq peak called by one of the algorithms.

The HPTs for seven mouse cell lines and e11.5 mouse forebrain were calculated using ChIP-seq data of H3K36me3 and H3K4me3 using histone-rank-invariant genes and source as ES cells.

Training and test sets for evaluation of EFilter. Paired-end 2 × 75 bp RNA-seq data generated by the Caltech ENCODE group were downloaded from the ENCODE website (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/>). Expression levels of RefSeq genes were quantified as FPKM using Cufflinks²¹ with the options `-frag-bias-correct` and `-multi-read-correct`. The FPKM values of transcripts having the same strand and TSS were added together. The signal of ChIP-seq tag-counts in 200-bp bins lying between 2 kbp (m1) upstream to 4 kbp (m2) downstream of RefSeq TSSs were used by EFilter to model expression. Gene expression (FPKM) and binned ChIP-seq tag-counts were log-transformed while doing regression. When RNA-seq and ChIP-seq data were from the same cell line, all RefSeq genes were used in the training set and also in the test set. However when RNA-seq data were not from the same cell line as ChIP-seq, only rank-invariant (RI) genes were used in the training set, while the test set still comprised all genes. In generating the results shown in **Figure 3b,c**, three ChIP-seq data sets (H3K4me3, H4K20me1 and H3K27ac) were used to calculate HPT. However, RI genes were defined only on the basis of H3K4me3 and H3K27ac, the two most predictive histone marks. RefSeq genes were ranked by tag count in 3 kbp around the TSS, and genes whose rank difference between source and target cell line was <1,000 for both marks were defined as RI genes.

Motif-based expression prediction. Promoters were defined as 1-kbp regions on either side of RefSeq TSSs, and scanned for the presence or absence of all TF-binding motifs from the TRANSFAC and JASPAR databases^{27,28}. A binding motif was said to be present if the promoter contained at least one DNA *n*-mer with an above-threshold log-odds score. In order to set the score threshold for a motif, we randomly generated 10,000 *n*-mers from the probability distribution (motif-PWM) of bound sites, and chose a cutoff at the 85th percentile, that is, at the score of the 8,500th *n*-mer. Thus for each motif, occurrence score at every promoter was converted to either 1 or 0.

A greedy linear regression method was used to select the set of *k* promoter motifs that best predicted the log-expression level of RefSeq genes, with *k* varied from 1 to 15. Promoter motifs were used to predict expression in four cell types: GM12878, HepG2, Huvec and NHEK. Two-fold cross-validation was performed by randomly choosing half of the gene set for training and the other half for test. For the genes in the test set, we calculated the Pearson correlation between input log-expression values and the motif-based predictions (**Fig. 4** and **Supplementary Fig. 9**).

Other details of analysis. Transcription factor motifs in ChIP-seq peak regions were detected using MOODS⁵². The motif thresholds were defined based on $P < 0.0001$. Then for **Figure 2a**, accuracy of peak detection was defined based on fraction of peaks that contained motif for the corresponding transcription factor within 30 bp of peak center (summit).

The *de novo* motif detection using the top 3,000 forebrain-specific promoters was done using MEME⁵³.

To find enriched functional terms for forebrain-specific promoters GREAT was used in hypergeometric mode. For this purpose positives were locations of promoter having a transcription factor motif and forebrain-specific HPT, and the background set consisted of all promoters that had forebrain-specific HPT. This highlighted the specific role of transcription factor target genes in forebrain.

48. Caucci, L., Barrett, H.H. & Rodriguez, J.J. Spatio-temporal Hotelling observer for signal detection from image sequences. *Opt. Express* **17**, 10946–10958 (2009).
49. Rissanen, J. Modeling by shortest data description. *Automatica* **14**, 465–471 (1978).
50. Zhao, L.C., Krishnaiah, P.R. & Bai, Z.D. On detection of the number of signals in presence of white noise. *J. Multivariate Anal.* **20**, 1–25 (1986).
51. Guttman, M. *et al.* *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).
52. Korhonen, J., Martinmaki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181–3182 (2009).
53. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* (AAAI Press, Menlo Park, California, 1994).