Predicting cell-type–specific gene expression from regions of open chromatin

Anirudh Natarajan, Galip Gürkan Yardimci, Nathan C. Sheffield, et al.

Genome Res. 2012 22: 1711-1722
Access the most recent version at doi:10.1101/gr.135129.111

Method

Predicting cell-type–specific gene expression from regions of open chromatin

Anirudh Natarajan,[1] Galip Gürkan Yardımcı,[1] Nathan C. Sheffield,[1] Gregory E. Crawford,[2,3,5] and Uwe Ohler[2,4,5]

[1]*Program in Computational Biology and Bioinformatics, Duke University, Durham, North Carolina 27708, USA;* [2]*Institute for Genome Sciences & Policy, Duke University, Durham, North Carolina 27708, USA;* [3]*Department of Pediatrics, Division of Medical Genetics, Duke University, Durham, North Carolina 27708, USA;* [4]*Department of Biostatistics & Bioinformatics, Duke University, Durham, North Carolina 27708, USA*

Complex patterns of cell-type–specific gene expression are thought to be achieved by combinatorial binding of transcription factors (TFs) to sequence elements in regulatory regions. Predicting cell-type–specific expression in mammals has been hindered by the oftentimes unknown location of distal regulatory regions. To alleviate this bottleneck, we used DNase-seq data from 19 diverse human cell types to identify proximal and distal regulatory elements at genome-wide scale. Matched expression data allowed us to separate genes into classes of cell-type–specific up-regulated, down-regulated, and constitutively expressed genes. CG dinucleotide content and DNA accessibility in the promoters of these three classes of genes displayed substantial differences, highlighting the importance of including these aspects in modeling gene expression. We associated DNase I hypersensitive sites (DHSs) with genes, and trained classifiers for different expression patterns. TF sequence motif matches in DHSs provided a strong performance improvement in predicting gene expression over the typical baseline approach of using proximal promoter sequences. In particular, we achieved competitive performance when discriminating up-regulated genes from different cell types or genes up- and down-regulated under the same conditions. We identified previously known and new candidate cell-type–specific regulators. The models generated testable predictions of activating or repressive functions of regulators. DNase I footprints for these regulators were indicative of their direct binding to DNA. In summary, we successfully used information of open chromatin obtained by a single assay, DNase-seq, to address the problem of predicting cell-type–specific gene expression in mammalian organisms directly from regulatory sequence.

[Supplemental material is available for this article.]

Decades of research on gene regulatory mechanisms has provided a rich framework with which we can explain gene expression. At the transcriptional level, this regulation is achieved by complex interactions between the DNA sequence and transcription factors (TFs), as well as nucleosomes, histone tail modifications, and DNA methylation. In particular, TFs have long been recognized as playing a fundamental role in gene regulation. A good example of the primacy of TFs in orchestrating programs of gene expression is demonstrated by the ability of ectopically expressed TFs to reprogram fibroblasts into induced pluripotent stem cells (Takahashi and Yamanaka 2006; Yu et al. 2007).

TFs influence gene expression by binding to *cis*-regulatory elements, typically between 6 and 20 bp, that are present in the proximal promoter or in distal regulatory regions (Vavouri and Elgar 2005). It has been proposed that the specific combinations of transcription factor binding sites (TFBSs) make it possible to define highly specific expression patterns. Elaborate patterns of gene expression have been shown to be controlled in a spatial, temporal, and cell-type–specific fashion. In contrast, many housekeeping genes have expression patterns that exhibit very little variation across most conditions or cell types. Understanding the extent to which groups of regulatory factors can achieve cell-type–specific gene expression and how this is encoded in the genome has long been a key question in biology (Britten and Davidson 1969).

Genome-wide techniques, such as chromatin immunoprecipitation followed by microarrays or sequencing (ChIP-chip and ChIP-seq), have been instrumental in identifying precise TFBSs that can then be used to predict gene expression. For example, ChIP data for 12 key TFs in embryonic stem (ES) cells were used to predict both absolute and relative expression values with high accuracy (Chen et al. 2008; Ouyang et al. 2009). While impressive, it is important to note the difficulty in procuring this kind of data across a wide variety of cell types. First, in order to conduct ChIP, one needs a high-quality antibody or tagged protein, which is not always available for the TF(s) of interest. Second, TFs have to be assayed individually, which requires many independent ChIP experiments to identify combinatorial patterns of TF binding. Finally, for this method to succeed, one must have a good understanding of the cell type in question to know which TFs to analyze. As a result, for most cell types, there is not enough information available on the binding profiles of TFs to predict cell-type–specific gene expression. Therefore, developing predictive models of gene expression without relying on ChIP would facilitate our understanding of transcriptional regulation.

A more widely applicable alternative to ChIP is to use known cognate binding preferences for TFs determined from assays such as SELEX, ChIP-seq, ChIP-chip, and protein binding microarrays (PBMs) (Stormo and Zhao 2010) to find TFBSs in putative regulatory regions. However, without knowing the location of distal

[5]**Corresponding authors**
E-mail greg.crawford@duke.edu
E-mail uwe.ohler@duke.edu
Article and supplemental material are at http://www.genome.org/cgi/doi/10.1101/gr.135129.111. Freely available online through the *Genome Research* Open Access option.

regulatory regions, most studies using this method focus exclusively on TFBS identified in proximal promoter sequences (Das et al. 2006; Ramsey et al. 2008; Sinha et al. 2008; Suzuki et al. 2009). Using these sequence features has revealed, for example, a crucial CG content difference between cell-type–specific and constitutively expressed genes in mammalian organisms (Yamashita et al. 2005; Carninci et al. 2006). However, these approaches have frequently struggled to distinguish between more specific patterns, such as predicting cell-type–specific expression across many cell types. A comprehensive understanding of cell-type–specific expression will require identification of both proximal promoter and distal regulatory elements. While comparative genomics has been successfully used to pinpoint functionally relevant regions, recent reports have stressed the complexity of evolution in functional noncoding regions and the resulting frequent lack of sequence conservation (Ludwig et al. 2005; Odom et al. 2007; Blow et al. 2010).

For more than three decades, mapping DNase I hypersensitive sites (DHSs) has been used to identify the location of many types of active gene regulatory elements (Wu and Gilbert 1981). DNase I is an enzyme that preferentially digests DNA in regions of low nucleosome occupancy, i.e., regions of open or accessible chromatin. DHSs have been found to be well correlated with genomic features such as transcription start sites (TSSs), distal enhancers, insulators, TFBSs, and active histone marks (Heintzman et al. 2007, 2009; Boyle et al. 2008a). A recent study profiling open chromatin in seven cell types in a genome-wide fashion using DNase-seq highlighted that open chromatin regions are similar across functionally related cell types and that cell-type–specific regions are distal to TSSs, and identified groups of DHSs that show coordinated nucleosome depletion (Song et al. 2011). Other studies have indicated that DNase-seq data can be used to identify TFBSs at single-nucleotide resolution (Hesselberth et al. 2009; Boyle et al. 2011; Pique-Regi et al. 2011).

In this study, we use DNase-seq data across 19 diverse human cell lines to define proximal and distal regulatory regions and to quantify the contribution of sequence features in DHSs to specify different patterns of cell-type–specific gene expression. Using expression data from the same 19 cell types, we define classes of up-regulated, down-regulated, and constitutively expressed genes, which show distinct patterns of chromatin accessibility. We then build predictive models specifically for these different expression classes, by using the binding site matches that map within DHSs. Crucially, these models dramatically improve on baseline models of proximal promoter regions and specifically control for the impact of promoter CG content on classifier performance.

Our results demonstrate the crucial role for sequence features in open chromatin regions for determining expression patterns and its usefulness for building predictive regulatory models. We confirm many known regulatory interactions and identify novel putative positive and negative regulators of gene expression. We also reveal the presence of DNase footprints for specific TFs that are identified as predictive in our model indicating direct binding to DNA. Our work provides a general and easily extensible framework to address questions related to gene regulation in vertebrates.

Results

DHSs have different properties depending on their genomic location

As part of the ENCODE Project, DNase-seq has been performed in several human cell lines representing a wide variety of tissue types. Aligned reads were used to define DNase I hypersensitive sites (DHSs) (for details, see Methods). Of these, we selected 19 cell lines to represent a broad and largely unrelated variety of cell types. These include DNase-seq data from a recent study across seven cell lines (Song et al. 2011). In each of the 19 cell lines we used, DHS regions cover ~2% of the genome (Supplemental Table 1). This indicates that a large proportion of *cis*-elements likely to be involved in establishing the expression patterns in each cell line only comprise a small fraction of the genome. Such regions may encode specific activation patterns of genes, but also include insulators that can define target relationships. A hallmark of insulators is the presence of binding sites for the CCCTC binding factor (CTCF). Across the nine cell types for which CTCF ChIP-seq data are available, ~28% of DHSs overlapped CTCF bound sites (Supplemental Table 5), in agreement with recent work (Song et al. 2011).

Based on their genomic location, DHSs were divided into exclusive classes as follows. We first identified a set of TSS DHSs as those that overlapped the transcription start sites (TSSs) of genes based on RefSeq hg19 annotation (Fig. 1A). Other DHSs were designated as Gene Body DHSs if they overlapped exons or introns, and as Intergenic DHSs if they did not overlap any genes. The median size of all DHSs was ~300 bp, with the TSS DHS set as outlier with a median size of ~1 kb (Fig. 1B). The larger size of TSS DHSs may reflect the presence of larger and more stable complexes such as the pre-initiation complex (PIC) near the TSS of genes.

The normalized CG dinucleotide content of Gene Body and Intergenic DHSs showed a median of 0.28 and 0.26, respectively

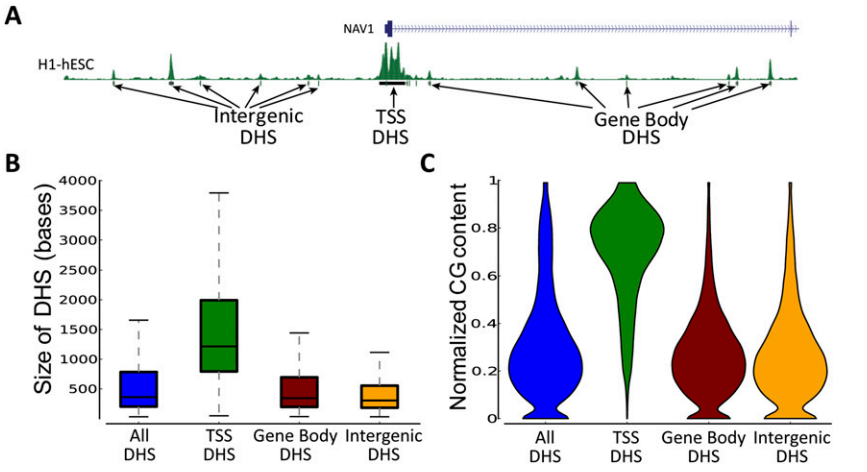


Figure 1. Properties of DHS based on genomic location. (*A*) DHSs that are intergenic and those that are overlapping the TSS and gene body were classified as Intergenic, TSS, and Gene Body DHSs, respectively (Chr1: 201,566,484–201,683,121). (*B*) Sizes of different DHSs for the Chorion cell line. Data from only one cell line were used to avoid multiple counting of ubiquitous DHSs. Other cell lines show similar trends. Outliers are not plotted. (*C*) Violin plot showing normalized CG content for different DHSs in the Chorion cell line. The subset of DHSs with a normalized CG content of zero is comparatively small (median of 128 bp).

(Fig. 1C). For TSS DHSs, the normalized CG content showed a unimodal distribution with its mode at ~0.8, with a heavy tail of several DHSs with CG content below 0.6.

A large proportion of TSSs are found in regions of accessible chromatin

To understand how regions of open chromatin vary between cell types, we inspected the degree to which DHSs were shared in the 19 cell types. A DHS was classified as being specific to a cell line if it was only present in a single cell type or overlapped <50% of its length with a DHS from any of the other 18 cell types (Fig. 2A). Across all DHSs, ~14% were specific to a single cell line (Fig. 2B). Intergenic DHSs showed the highest percentage of being cell-type–specific (~17%). Conversely, TSS DHSs were largely not cell-type–specific with <1% being open specifically in a single cell type. Despite the broad panel of cell lines that vary in expression, the chromatin state at the TSS of these genes was open and largely invariant across multiple cell lines. This is in agreement with a recent study analyzing a subset of the cell types used here (Song et al. 2011).

We determined the normalized CG content in the proximal promoter region of the gene, defining the proximal promoter as −900 to +100 bp around the TSS. If a gene had multiple TSSs, the average of the normalized CG content from each TSS was used. There was a steady positive trend in the number of cell lines in which a DHS overlapped a TSS and the CG content around the TSS (Fig. 2C). Previous studies have reported that gene expression can be predicted from the CG content in the proximal promoter region (Yamashita et al. 2005; Carninci et al. 2006; Zhu et al. 2008). Our result indicates that higher levels of CG dinucleotide content, and thus more frequent presence of CpG islands, are positively correlated with, and could be functioning to preserve, an open chromatin state surrounding the TSS. There were fewer genes with a TSS open in only one cell type (976 genes) and many with an open TSS across all 19 cell types (8393) (Supplemental Table 2).

Cell-type–specific expressed genes show differing patterns of accessible chromatin at their TSS

Gene expression data for the 19 cell lines were generated using Affymetrix exon arrays. Expression values for each gene were transformed to *Z*-scores across all of the cell lines. Genes with large positive or negative *Z*-score values thus showed a larger deviation from the mean expression across cell types. The *Z*-score transformed expression values were used to select subsets of genes with specific expression patterns (Fig. 3A–C; Supplemental Table 3). Up-regulated genes, exemplified by *GCM1* (Fig. 3A), had a particularly high expression in one cell line, but expression close to the mean in the other cell lines. To identify genes exhibiting this type of expression pattern, we sorted the *Z*-score expression for the genes in each cell line. The top 200 genes in this sorted list were classified as being up-regulated in that cell type (UR genes). Down-regulated genes exhibit low expression levels in one cell type but are otherwise constitutively expressed in other cell lines (Fig. 3B; Thorrez et al. 2011). We classified the last 200 genes in the sorted *Z*-score expression list as being the cell-type–specific down-regulated genes (DR genes). Constitutively expressed genes (Fig. 3C) were identified by filtering all genes that were not in UR and DR gene sets in any cell line and had absolute expression *Z*-score values < 1.7 in all cell lines. Using this cutoff, 168 genes displayed a pattern of constant expression levels across all cell lines.

To address how up-regulated genes are expressed in one particular cell type, we grouped UR genes from all other cell types and denoted this group as UR-Other genes (Fig. 3A). We imposed the additional constraint that such genes would show an expression *Z*-score < 0 in the cell type of consideration, i.e., had expression below its mean expression. As an example, *GCM1* (Fig. 3A) was highly expressed in the first cell type and in none of the others shown. It was therefore grouped into the UR class for the first cell type and into the UR-Other class in each of the other cell types. Similarly, genes denoted

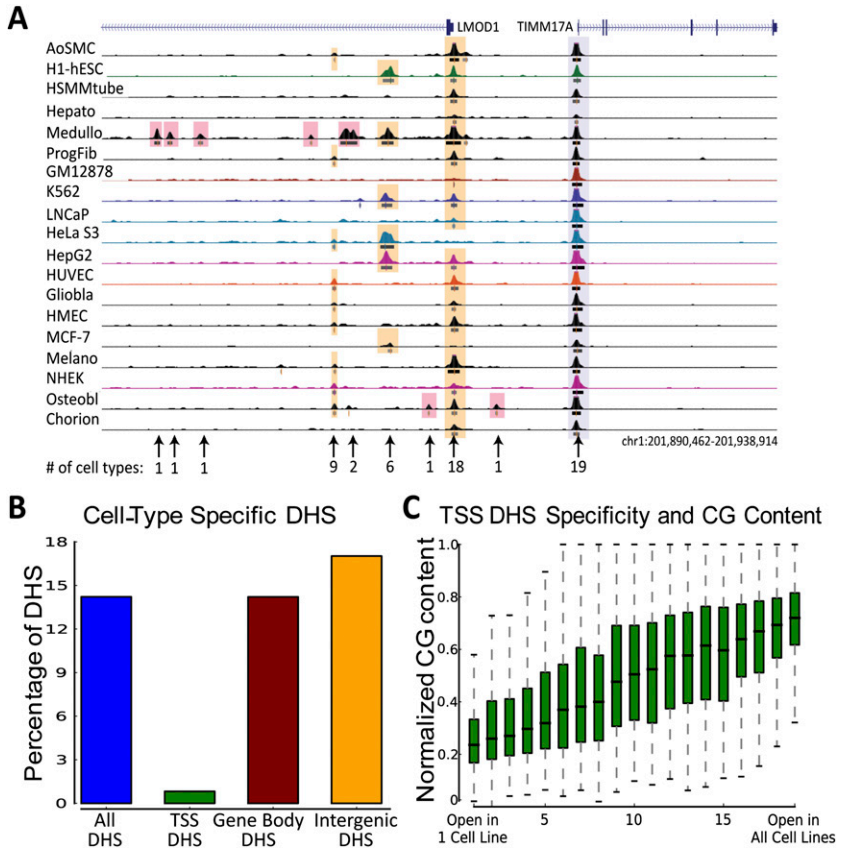


Figure 2. Cell-type specificity of hypersensitive regions. (*A*) Example (Chr1: 201,890,462–201,938,914) showing cell-type–specific DHSs across two cell lines (pink boxes). Note that we called a DHS cell-type–specific if it did not overlap another DHS by more than half in any of the 18 other cell lines. (*B*) Bar graph showing the proportions of cell-type–specific DHSs across different genomic locations averaged across all cell lines. (*C*) TSSs were divided by the number of cell lines that they overlapped in a region of open chromatin. For each set of TSSs, normalized CG content in the promoter regions (−900,100) of the TSSs are shown.

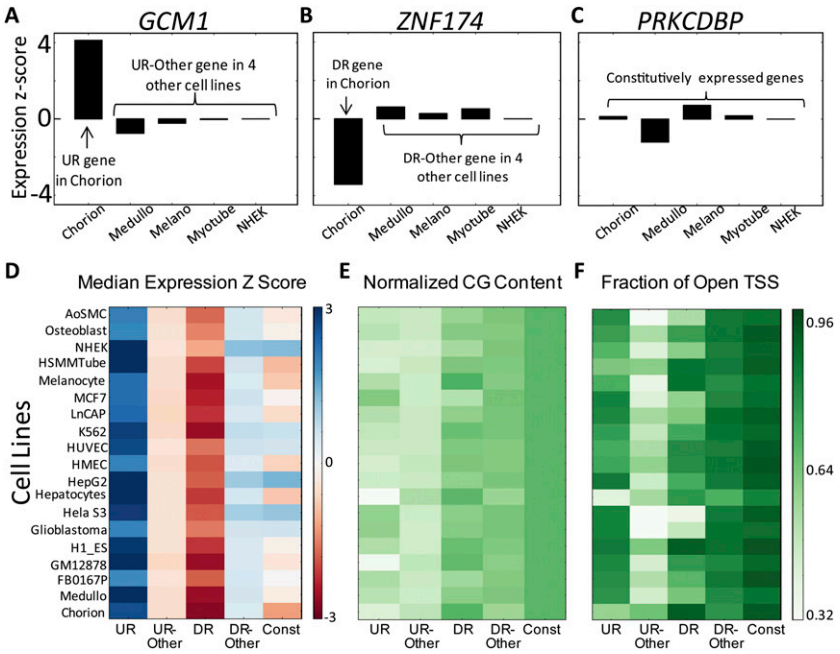


Figure 3. Cell-type–specific gene expression and definition of gene classes. (*A–C*) Representative examples of different patterns of gene expression. Note that *Z*-score values are calculated from expression across all 19 cell lines. (*A*) A gene where the expression is specifically up-regulated in the first cell line (UR gene). (*B*) A gene that is specifically down-regulated in the first cell line (DR gene). (*C*) A gene that has low variability in expression (constitutively expressed gene). (*D*) Median expression *Z*-scores for the genes in each set in each cell line. (*E*) Normalized CG content from the promoter regions of genes. (*F*) The fraction of TSS in each gene set that were in a region of open chromatin. *E* and *F* share the same color map.

as DR-Other had to be classified as down-regulated in another cell line and had an expression *Z*-score > 0 in the cell type of consideration (Fig. 3B). In this way, we defined different classes of transcriptionally active (UR, DR-Other, and Constitutive) and transcriptionally inactive genes (UR-Other, DR) from the point of view of each cell line in comparison to other cell lines.

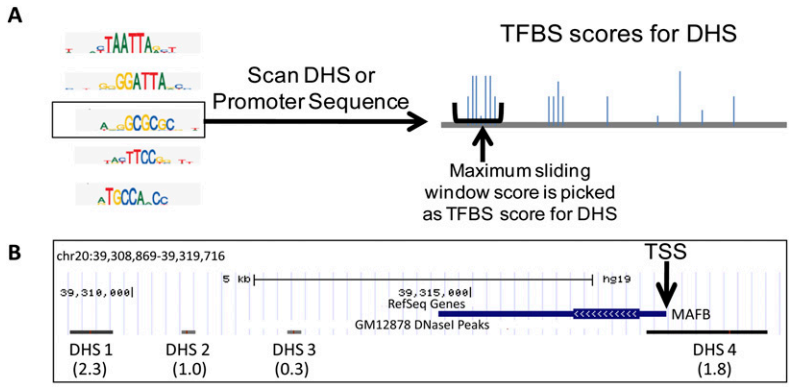
By definition, UR and DR genes displayed the highest and lowest *Z*-score gene expression, respectively (Fig. 3D). UR genes were consequently enriched in functions related to the tissue type of origin (Supplemental Table 4). DR-Other and Constitutive genes showed similar expression values and had higher expression values than genes in the UR-Other class. To understand whether these different classes had different properties in sequence composition and chromatin state, we first inspected normalized CG content in the proximal promoter region (Fig. 3E). UR and UR-Other genes had the lowest average CG content compared with the other classes of genes. Constitutively expressed genes displayed a particularly high CG content in their proximal promoter regions, as previously reported (Yamashita et al. 2005; Carninci et al. 2006; Zhu et al. 2008); however, this observation clearly extended to the DR and DR-Other gene classes, which had a CG content slightly lower than constitutively expressed genes.

Constitutively expressed and DR-Other genes had the highest proportion of their TSSs in regions of accessible chromatin (Fig. 3F). DR genes displayed slightly lower chromatin accessibility compared with DR-Other, indicating that repression of DR genes largely occurs while maintaining chromatin accessibility at the TSS. UR genes also showed a high proportion of genes containing an accessible TSS, at similar levels to DR and DR-Other. In stark contrast, UR-Other genes had the lowest fraction of TSSs that overlapped a DHS. These results indicate that even though UR-Other and DR genes are both transcriptionally inactive in a cell type of consideration compared with other cell types, they are likely to be regulated via different chromatin-remodeling mechanisms. Specifically, genes that are up-regulated in a small number of cell types likely maintain a closed chromatin conformation until cellular processes require up-regulation. In contrast, down-regulated genes may be viewed as constitutively expressed genes that are repressed in a single cell type. UR genes had intergenic and gene body DHSs associated with them (Supplemental Fig. 1A,B), in agreement with previous results indicating that cell-type–specific expression is mediated by distal *cis*-regulatory regions (Song et al. 2011). Overall, these results indicate that different classes of transcriptionally active and inactive genes have different CG content and chromatin accessibility at their TSS.

Classifying tissue-specific expression from sequence features in open chromatin

To predict gene expression patterns from sequence, approaches have frequently used features contained within fixed-size proximal promoter sequences. We used DHS data from a large number of cell types to determine whether using both proximal and distal regulatory regions with open chromatin would improve predictive models for cell-type–specific expression patterns. Position weight matrices (PWMs) for TFs in vertebrates were compiled from TRANSFAC, JASPAR, and UniProbe databases (Matys et al. 2006; Bryne et al. 2008; Newburger and Bulyk 2009). For each DHS, 789 PWMs were used to calculate TFBS scores that accounted for local dinucleotide composition. The maximum sliding window score for each PWM was used as the TFBS score for that DHS (Fig. 4A; Methods). To associate DHSs with specific genes that they are likely to regulate, we applied a simple approach of associating each DHS with the closest TSS (closest gene DHS). For each TF, we then chose the maximum TFBS score across all DHSs associated with a gene (Fig. 4B). As an alternative approach, we split DHSs into distal sites (a set including both Gene-Body and Intergenic DHSs) and TSS DHS sites and used the maximum TFBS in each set as individual features (split DHSs). This doubled the number of features and allowed us to identify different characteristics of TSS-overlapping versus distal DHSs. To compare our models with previous approaches, we also used TFBS features calculated in proximal promoters, defined here as −900 to +100 nt surrounding the TSS (Landolin et al. 2010).

We used the TFBS scores as features for sparse logistic regression classifiers to discriminate between different gene classes. These classifiers balance the use of many available features against model complexity, effectively selecting a small subset of informative features that are used in the classification. We trained cell-type–specific classifiers on the task to discern whether a gene belonged to a specific expression pattern (e.g., UR vs. UR-Other, UR vs. DR, UR vs. Constitutive, etc.). The area under the receiver operating



Closest Gene DHS		Split DHS	
DHS associated with MAFB	TFBS score	DHS associated with MAFB	TFBS score
DHS 1, 2, 3, 4	2.3	DHS 1, 2, 3	2.3
		DHS 4	1.8

Figure 4. Transcription factor binding site features. (*A*) DHS and promoter sequences are scanned with PWMs. TFBS scores are log-likelihood ratios of PWM over the background model. A sliding window is used to identify the score for each DHS or promoter. (*B*) Example to show association of DHSs with genes. Numbers in the brackets are example TFBS scores for the DHS for a specific DHS. Two methods of association were used. In closest gene DHS, DHSs 1–4 from the GM12878 cell line are associated with the gene *MAFB*. For the TF in consideration, the maximum of all TFBS scores is 2.3. In Split DHS, we separated DHSs overlapping the TSS and other DHSs. This resulted in two features for each gene for each TF.

characteristic curve (AuROC) metric was used to evaluate the performance of a model, where a value of 0.5 indicates random assignments and 1.0 indicates perfect classification (see Methods). To not bias results due to different amounts of training data, the positive sets of up- and down-regulated genes were all of the same size.

The performance of the classifier using only proximal promoter information is close to that of a random classifier, across all tasks. All of the classifiers using DHS sequences display strong improvements in performance over this baseline in discriminating genes that are up-regulated in different cell types (UR vs. UR-Other) (Fig. 5A), with a greater improvement in performance coming from the Split DHS approach with separate features for the TSS and Distal DHSs (median AuROC ~0.73). Similar results were obtained when training classifiers to distinguish between specifically up- and down-regulated genes from the same cell types (UR vs. DR) (Fig. 5B), and to distinguish up-regulated from constitutively expressed genes (UR vs. Const.) (Fig. 5C). Discriminating down-regulated genes from different cell types (DR vs. DR-Other) and down-regulated from constitutively expressed genes (DR vs. Const.) resulted in lower accuracies but still showed the trend of better performance with DHSs compared with proximal promoter sequence (Supplemental Fig. 2A,B). All results clearly indicate that strong performance improvement is achieved by scanning for TFBS matches in open chromatin regions.

Evaluating the influence of CG dinucleotide content

CG dinucleotide content in the proximal promoter sequence of genes is a common sequence feature that is directly or implicitly used to distinguish various classes of genes. Adding CG dinucleotide content as an additional feature led to a variable impact on classifier performance depending on the classification being considered (Fig. 5D–F). Specifically, when using open chromatin information, adding CG content did not substantially improve the performance of classifying UR genes from UR-Other genes (Fig. 5D). In the case of the Split DHS, only six of the cell lines had a significant coefficient for the CG dinucleotide coefficient (mean across all cell lines = −0.66, SD = 1.40; the coefficient was set to 0 when not significant). Due to the means being close to zero and the standard deviations being large, the effect of using CG content to discriminate between UR and UR-Other genes was largely negligible. Only for a few cell types, such as hepatocytes, did we observe a negative regression coefficient of significant magnitude (−5.11 for Split DHS and −3.31 for Promoters). This is in agreement with previous results showing that liver-specific genes have promoters with lower CpG content (Smith et al. 2005).

As anticipated by the trends observed in Figure 3E, UR vs. DR classification tasks benefited more by the addition of CpG content. Here, this feature was deemed to be significant in 17 of the cell lines for the Split DHSs (Fig. 5E). Furthermore, the regression coefficients were largely negative, which indicated the higher CG content among the DR genes (mean = −2.88, SD = 1.55). As has been shown before (Yamashita et al. 2005; Carninci et al. 2006; Zhu et al. 2008), we observed that high CpG content is predictive of constitutively expressed genes when compared with UR genes (Fig. 5F). The regression coefficient for the feature was significant in all cases (mean = −3.25, SD = 1.44). The CpG content feature had almost no impact in classifying DR from DR-Other genes (Supplemental Fig. 3). Finally, CpG content had a significant coefficient in classifying DR from constitutively expressed genes in only one cell line (mean = −0.09, SD = 0.39).

Adding CG content to the baseline proximal promoter models reconciled the apparent discrepancies between previous studies and the results reported in Figure 5A–C, because all classification tasks were improved upon for the proximal promoter. However, the DHS models with CG content outperformed baseline proximal promoter models with the inclusion of CG content (paired t-test < 0.05). In fact, in all cases except UR vs. Constitutive genes, DHS models even without CG content perform significantly better than both proximal promoter models (paired t-test < 0.05). Note that while adding CG content provided enormous performance gains for certain classification tasks (UR genes vs. Constitutive genes), this could be considered misleading. If TFBS scores are not explicitly normalized for local nucleotide composition, as we have done here, decent performance results can be achieved based solely on the different CG content observed for down-regulated and constitutively expressed genes compared with up-regulated genes. CG content is predictive in the case of classifying constitutive and DR genes from UR genes, but is not very useful in differentiating between genes that are up-regulated or down-regulated

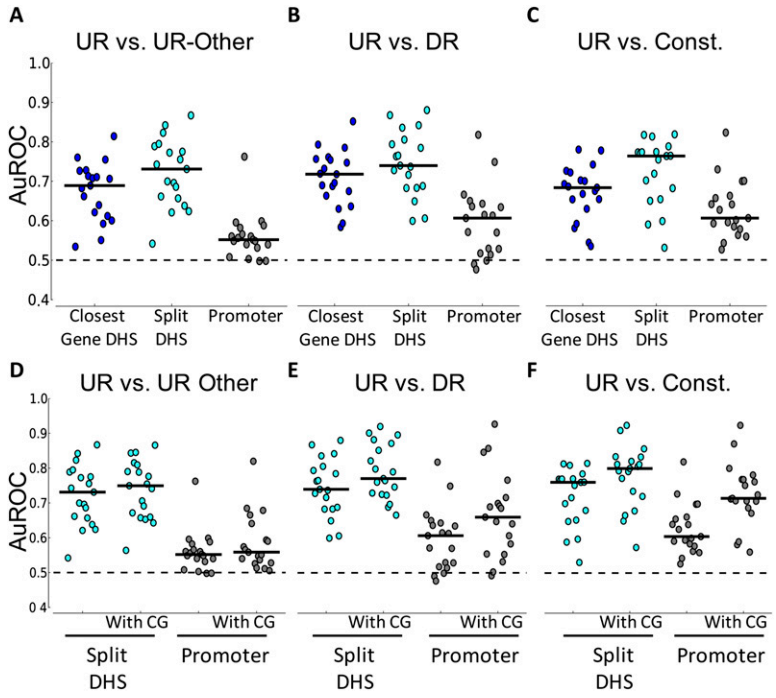


Figure 5. Classifier performance for various classification tasks. (*A–C*) Performance of the classifier using all PWMs. Each figure compares the performance of two methods of associating DHSs to genes (Closest Gene DHS and Split DHS) with the proximal promoter. The solid black lines across the dots indicate the median. Across all figures, the promoter sequence classifier does not perform as well as the performance achieved by using Closest Gene DHS and Split DHS and is significant at the 0.05 level (paired *t*-test). (*D–F*) Impact of normalized CG dinucleotide content on classifier performance. Results using the Split DHS and promoter sequence are shown. Without CG, columns are the same as in *A–C*. All figures show average results from five iterations of fourfold cross-validation. The dotted line indicates an AuROC of 0.5, which is the performance of a random classifier.

in different cell types. It is notable that the categories that are less aided by CG content are exactly those where our classifiers displayed the most predictive value.

Identifying candidate regulators

In addition to classifying genes belonging to different groups, we inspected the classifiers to identify motifs that were most informative in the classification task, i.e., those PWMs that had large regression coefficients (Supplemental Table 4). This identified several TFs with known impact on transcriptional output in the cell line of interest. For example, *YY1*, *SPI1*, and *IRF8* are crucial in the specification of B-cells (GM12878 cell line) (Lu et al. 2003; Liu et al. 2007; Sokalski et al. 2011). We also identified the *REST* motif as a positive regulator of UR genes in the medulloblastoma cell line that is of neural origin (Supplemental Table 6). *REST* specifically down-regulates neuron-specific genes in many non-neuronal cell lines, and its expression is suppressed in neurons (Schoenherr and Anderson 1995). As a result, the model identified the *cis*-elements that are present in the DHSs associated with neuron-specific genes as the factor that separates these genes from the genes up-regulated elsewhere. This example illustrates that the inactivation of a repressor can also explain up-regulation of genes. Other well-characterized factors included *ETS1* in HUVEC cells and *HNF4A* for HepG2 cells (Cereghini 1996; Oda et al. 1999; Yordy et al. 2005).

The feature set described thus far was comprehensive in that it used available PWM information from multiple sources, independent of the expression levels of transcription factors or the potential redundancy of features. To assess how much cell-type–specific regulation can be explained by the cell-type–specific expression of transcription factors themselves, we selected the top 10 TFs with highest absolute *Z*-scores from each cell line and had PWMs that were not similar to each other (Supplemental Table 7).

Using sparse logistic regression classifiers trained on these small sets of variables, we observed similar predictive trends, which indicated that a subset of cell-type–specific TFs were predictive of tissue-specific expression (Supplemental Fig. 4). Using only promoter CG content or the status of the chromatin at the TSS as features for a baseline comparison shows that motifs in DHS regions significantly contribute to the performance improvement across all comparisons. In addition, we used genomic regions identified as conserved in the 46-way placental mammal phastCons track from the UCSC Genome Browser. We note that using conserved sequences and particularly conserved noncoding regions improved performance compared with the promoter. However, the AuROC was still highest when DHS sequences were used, indicating that the presence of motifs in weakly conserved DHS regions contributes to the performance improvement. Finally, to assess the potential influence of insulators, we excluded DHSs that overlapped *CTCF* binding sites for classifiers trained specifically for the nine cell types for which genome-wide *CTCF* ChIP data were available (Supplemental Fig. 5). While this did not impact classification of UR genes, it reduced the accuracy of identifying DR genes, demonstrating that regions containing insulator sites are likely to contain regulatory information for the repression of genes.

Knowing both the regression coefficient in our model and the expression level of a potential regulator provided clues as to whether the TF in question is an activator or a repressor in the cell line, as highlighted for *REST* in medulloblastoma cells (Table 1; Supplemental Table 7). As another example, *NR2F2* was identified as a positive predictor of up-regulated genes for embryonic stem cells. However, *NR2F2* is a known negative regulator of *POU5F1*, a critical gene involved in pluripotency (Rosa and Brivanlou 2011). As expected, *NR2F2* is down-regulated in ES cells (Supplemental Table 7). We also identified other known positive regulators, such as *GATA1* in K562 cells (Huang et al. 2005) and *MYF6* in myotubes (Fan et al. 2011). Note that genes that have both positive and negative coefficients have different effects when in TSS DHSs and Distal DHSs.

For *HNF4A* in HepG2 and *GATA1* in K562 cells, ChIP data are available from the ENCODE Project. To validate the predictions made by our model, we looked for overlap of these ChIP sites with DHS sites associated with different sets of genes. In HepG2 cells, 19% of all genes with an associated DHS overlapped an *HNF4A* binding site. Strikingly, 64.5% of the UR genes had a DHS overlapping an *HNF4A* ChIP peak (P-value $< 1 \times 10^{-12}$, binomial test). Conversely, only 10.5% of DR genes had a DHS that overlapped an *HNF4A* site (P-value $< 1 \times 10^{-3}$). In K562 cells, we found that 6% of

Table 1. Candidate TFs identified by the classifier for each cell line using Split DHS from the top 10 highest absolute *Z*-score of expression and nonredundant TFs

	UR–UR Other genes			UR–DR genes		
Cell type	AuROC	TFs Positive coefficient	TFs Negative coefficient	AuROC	TFs Positive coefficient	TFs Negative coefficient
Chorion	0.55		***OSR2***, *ELK1*	0.83	*ZFP161*	*E2F3*, *ELK1*
Medulloblastoma	0.77	***CRX***, *REST*		0.8	***CRX***, *REST*, *NR2F2*, ***SOX11***	
FB0167P	0.72	*BACH2*, ***ZIC1***, *AIRE*		0.72	***STAT1***, ***ZBTB12***, *BACH2*, ***HOXC11***, ***ZIC1***, *AIRE*	
GM12878	0.75	***EGR2***, ***SPIB***		0.64	***SPIB***	***ARID5A***
H1_ES	0.67	***ZIC3***, ***OTX2***, *NR2F2*		0.69	*NR2F2*, ***ZIC3***, ***OTX2***	*MEIS1*, *NR2F2*
Glioblastoma	0.81	***ZIC1***, ***IRF3***, ***BAPX1***		0.74	***ZIC1***, ***HOXD10***	
HeLa S3	0.84	***PAX6***, ***E2F2***, ***FOXF2***, ***ELK1***, ***ARNT***, ***ESRRA***		0.84	***MEOX1***, ***ARNT***, ***FOXF2***, ***PAX6***, ***ELK1***, ***ESRRA***	
Hepatocytes	0.70	*FOXJ3*, ***HNF4A***, ***RXRA***, ***STAT3***	***RXRA***, *RFX7*, *HOXA6*, *FOXJ3*, *E2F3*, ***STAT3***	0.78		*E2F3*
HepG2	0.77	***GFI1***, ***HNF4A***		0.67	***SOX9***, ***FOXA2***, ***HNF4A***	
HMEC	0.6	***STAT4***		0.68	***STAT4***, ***IRF6***	
HUVEC	0.67	***SOX17***		0.71	***HIC1***, ***SOX17***	
K562	0.66	***GATA1***		0.64	***GATA1***	
LnCAP	0.64	***NKX3-1***		0.6	***ZBTB7B***	
MCF7	0.74	***GATA3***		0.68	***GATA1***, ***ESR1***	
Melanocyte	0.76	***MAF***, ***LEF1***, ***IRF4***, *CUX1*, *GABPA*		0.83	***IRF4***, *GABPA*	***TBX5***, *GABPA*
HSMMtube	0.61	*HLXB9*, ***MYF6***, *ZBTB3*	***SOX11***, *SIX1*, *ZBTB12*, *STRA13*, *ZBTB3*	0.67	***MYF6***	***MYF6***, *STRA13*, *ZBTB12*, ***SOX11***, *GATA6*, *ZBTB6*
NHEK	0.72	***MTF1***, ***MAF***		0.67	***MTF1***	
Osteoblast	0.63	***BACH1***, ***STAT4***, ***GLIS2***		0.56	***STAT4***, ***BACH1***	
AoSMC	0.83	***MEIS1***, ***OSR1***, *HOXC11*, *SOX8*, *PITX3*, *PAX4*		0.82	*PITX3*, ***MEIS1***, ***OSR1***, *HOXC11*	

UR vs. UR-Other and UR vs. DR classification tasks are shown. TFs with positive and negative coefficients are shown for both tasks. Genes in bold are up-regulated and other genes are down-regulated in the cell line. Several of the same factors help in classifying UR vs. UR-Other genes and UR vs. DR genes.

all genes had an associated DHS with a *GATA1* ChIP peak. However, 31.5% (P-value < 1×10^{-12}) of UR genes and only 3.5% (P-value < 0.1) of DR genes had a DHS with a *GATA1* ChIP peak. The ChIP binding data provided strong and independent evidence that our models identify relevant factors that regulate the transcriptional program in these cells.

In addition, we investigated the accuracy of our predictions of TFBS locations in DHSs. In HepG2, 5215 of the 6597 *HNF4A* ChIP peaks overlapped the predicted TFBS in DHSs. Furthermore, using TFBS scores led to a high accuracy on discriminating between positive and negative sets defined by ChIP peaks (AuROC of 0.79). In K562 cells, only 315 of the 1704 *GATA1* ChIP peaks overlapped the predicted TFBS in the DHS; yet, the AuROC still remained high at a value of 0.88. This indicated that high-scoring TFBSs accurately predicted binding of *GATA1* to these sites. We note that the low percent overlap may arise from nonspecific or indirect binding of *GATA1*.

To assess the presence of additional sequence motifs not accounted for by the sets of known PWMs, we used the discriminative version of MEME to perform motif finding (Bailey et al. 2010), identifying motifs differentially enriched between UR and UR-Other, respectively, DR genes (Supplemental Table 8). While some of the identified motifs corroborated the importance of features from the set of top 10 TFs (*FOXA2* [formerly *HNF3B*] in HepG2), others corresponded to TFs that were not in this list. These are candidate TFs that are not among the most differentially expressed, but still might be involved in the transcriptional program, potentially through other steps of activation. We note that we largely did not recover the motifs recently identified in a subset of seven of the 19 cell lines (Song et al. 2011). In contrast to this study, which used the sequences from cell-type–specific DHSs as foreground and the subsets of cell-type–specific DHSs in other cell types as background, we analyzed the sequences from all DHSs associated with a gene and defined the background according to the classification tasks.

Footprints in DNase-seq data show evidence of direct TF binding

While we have shown that the presence of sequence motifs in DHS regions is predictive of cell-type–specific gene expression patterns, we were only able to validate the direct binding of two factors due to the lack of ChIP data. This issue is likely to arise in several studies in which ChIP data are not available to provide evidence of direct binding of a TF to its cognate binding site. However, if a region is bound by a TF, the profile of aggregate reads around the TF binding site will show that region to be protected from digestion by DNase I, resulting in a DNase "footprint." Based on this pattern, DNase-seq data have been recently used to identify the precise binding locations of several TFs at base-pair accuracy (Hesselberth et al. 2009; Boyle et al. 2011; Pique-Regi et al. 2011).

To assess whether the factors that had high regression coefficients in the classification tasks showed such distinctive footprints, we compiled the DNase-seq reads in a 100-bp window centered on the top motif matches across the genome. We expected to see a distinct pattern in the cell line in which a motif was predictive of gene expression. As a control, DNase profiles were compiled for cell lines in which the model did not have a high regression coefficient for the TF. The motif matches used here were chosen to reflect the genome-wide binding of the factor,

as opposed to the specific binding sites used to model gene expression.

For several factors, we observed indicative footprints in the region of the motif (Fig. 6). For example, *CRX* was predictive of UR genes in the medulloblastoma cell line, and it exhibited a protected region at the motif (Fig. 6A). Importantly, in other cell lines such as GM12878, LnCAP, and MCF7, the *CRX* motif did not display a similar level of protection. While *CRX* has been shown to be expressed in certain types of medulloblastoma subtypes (Kool et al. 2008), other factors such as *OTX2* have nearly identical PWMs and are known to be important for transcriptional regulation in medulloblastomas (Bunt et al. 2011). This highlights a caveat in predicting expression from motifs; while we can identify biologically relevant motifs, this type of analysis only suggests a subset of factors that likely bind to a specific motif.

As mentioned earlier, we identified *REST* as a regulator in the medulloblastoma cell line. Since it is not expressed in this cell line, we observed the absence of a footprint in that cell line, and a visible footprint in other cell lines (Fig. 6B). Additional footprinting evidence is detected for *EGR2* and *SPIB* in the GM12878 cell line (Fig. 6C,D); however, the *SPIB* motif also exhibits a smaller footprint in another cell line. This could be due to expression of other factors that bind to a similar motif in this cell line. Further work is needed to quantify these encouraging observations rigorously.

Discussion

In this study, we proposed a new method to predict gene expression by using DNase-seq data from 19 human cell lines. Unlike other strategies that require multiple ChIP-seq data sets for highly informative regulatory factors, a single DNase-seq experiment identifies most regions of the genome that are accessible to TF binding. We show that motifs located in these DHS sites are predictive of cell-type–specific expression.

Some of the predictive motifs we identified were found to be enriched within cell-type–specific DHSs in a previous study using a subset of the cell types used here (Song et al. 2011). Patterns of co-occurrence and conservation of TFBS have also been used to identify regulatory modules de novo (Aerts et al. 2003; Sharan et al. 2003; Fu et al. 2004; Gotea and Ovcharenko 2008). However, our approach differs from such motif finding approaches, because it is not based on the sole presence of motifs, but their predictive value for gene expression patterns. As a result, the regression coefficients in our classifier and the expression profile of the TF can be viewed as testable predictions of the activating or repressing nature of the regulatory interactions between TFs and the different patterns. We also do not restrict our analysis here to cell-type–specific DHSs, to allow for the possibility that a motif could be present in a region of ubiquitously open chromatin, but only be predictive of gene expression in a specific cell type, for instance, due to the cell-type–specific expression of the factor binding to it.

CpG islands are hallmarks of unmethylated regions in vertebrate genomes and are known to overlap promoter regions, in particular in constitutively expressed genes (Yamashita et al. 2005; Carninci et al. 2006). Our results here are in agreement with previous findings that normalized CG dinucleotide content is negatively correlated with the specificity of gene expression. Consequently, constitutively expressed genes show higher CG content than up-regulated genes. While this feature is therefore useful in differentiating constitutively expressed genes, it is a confounding feature of proximal promoters when defining tissue-specific regulatory codes. Our models are based on normalized binding site

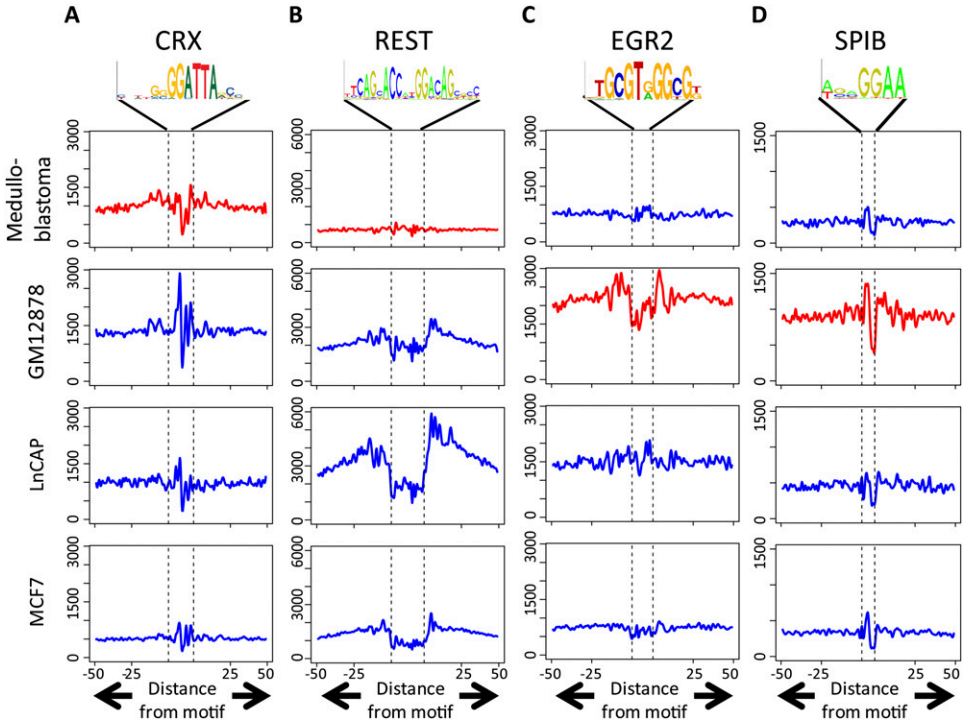


Figure 6. Aggregate plots of DNase-reads around motifs for factors with high regression coefficients. (Red lines) The cell line in which the TF is identified as a regulator. (*A*) *CRX* shows a footprint in medulloblastoma but not in the other cell lines shown. (*B*) *REST* shows a footprint in other cell lines but not in medulloblastoma, where it is not expressed. (*C*,*D*) *EGR2* and *SPIB* show footprints in the GM12878 cell line.

scores in a compendium of proximal and distal regulatory regions, and thus show consistent performance across different expression patterns. The classification performance we achieved when using the presence of motifs from open chromatin regions is significantly better than using the proximal promoter region. This is the case even when CG content is included as a feature for the classifier. We note that while using conserved regions of the genome improved the performance of the classifier over that achieved with the proximal promoter region, scanning for motifs in open chromatin regions still provided the best performance. Interestingly, in a previous study, only 43% of DHSs were found to overlap an evolutionarily conserved region (Boyle et al. 2008a), and it is known that functional enhancers are sometimes weakly conserved (Blow et al. 2010).

A related recent study monitored expression using transient transfection assays for several promoters (Landolin et al. 2010). The investigators then used sequence features in the transfected plasmids to predict expression with high accuracy. There are two main differences between this work and our study reported here. First, as pointed out by Landolin et al. (2010), the promoters are not in their endogenous context in the plasmid. Therefore, this effort reflects the role that sequence plays in determining expression outside of the chromatin context. In contrast, our work attempts to identify cell-type–specific expression from the endogenous accessibility of putative *cis*-regulatory regions. Second, the investigators defined classification tasks different from the ones we examined here. In particular, they discriminated cell-type–specifically expressed genes from ubiquitously expressed and unexpressed genes. While the first classification task is similar to the UR vs. Constitutive classifiers here, we do not attempt to define ubiquitously unexpressed genes, because genes could always be expressed in another condition not assayed or be affected by artifacts such as ineffective probes or misannotated genes. On the other hand, we build classifiers for the harder problem of predicting up- or down-regulation in one cell type versus another.

As expected, we observed an increased performance when using the comprehensive set of all PWM scores rather than just those for the most specifically expressed TFs. However, these models are harder to interpret: Many PWMs used to compute the comprehensive feature vectors are highly similar or identical. TFs with the same protein binding domains also have similar binding preferences, and a large proportion of the TFs in the current release of the UniProbe database are homeodomain TFs. This can lead to collinearity among the features that are used to classify the genes into different sets. As a result, over multiple iterations of the cross-validation, the weights assigned to each PWM are distributed to similar PWMs, and comparatively few PWMs had significant regression coefficients. To counter this, we used the subset of specifically expressed TFs, where our modeling approach allowed us to identify several known TFs that regulate gene expression but also additional candidates to study for their potential role in gene regulation in the given cell type. Future efforts will make use of recent sparse regression models that explicitly account for feature redundancy or use projection methods such as factor analysis to explain the high-dimensional feature vectors by smaller numbers of covariates.

DNase data also showed footprints of cell-type–specific binding of some factors at a high resolution. This analysis therefore corroborates recent analyses that demonstrated that the DNase-seq assay improves the signal-to-noise when attempting to identify functional locations of TF binding (Boyle et al. 2011; Pique-Regi et al. 2011). Future work in predicting gene expression will attempt to understand the utility of these high-resolution data in predicting gene expression.

We note that the approach of predicting cell-type–specific expression from *cis*-regulatory sequence as presented here is impeded by several limitations, even when the location of distal enhancers is known. First, only a small fraction of all TF motifs are known, making it likely that more comprehensive knowledge of motifs will improve the performance of the classifier. Second, accounting for long-range regulatory interactions by methods like 3C, 4C, 5C, Hi-C, and ChIA-PET will allow for more accurate connections of DHSs to the correct target gene (van Steensel and Dekker 2010). Third, quantitative nucleotide-level accessibility scores may perform better than simple binary DHS peak calling. Fourth, an important extension to our work lies in the identification of combinatorial TF codes that improve classification accuracy. Finally, transcript abundance is affected by several factors including post-transcriptional regulation, for instance, by microRNAs. A complete model that takes all of these factors into account will likely be necessary to provide even better predictive models of gene expression.

Methods

DNase-seq

DNase-seq was performed on 19 human cell lines representing a wide variety of tissue types, and aligned reads were used to define DNase hypersensitive sites (DHSs). Data from seven cell lines were previously published (Song et al. 2011), and remaining libraries were processed as described in that study. The reads generated were aligned to the hg19 genome using BWA and were then smoothed using a kernel density estimator, F-seq (Boyle et al. 2008b; Li and Durbin 2009). Following this, DHS peaks were identified as having a $-\log_{10}$ (*P*-value) ≥ 1.3. We refer to these regions as DHSs or regions of open chromatin. Note that the AoSMC cell line was cultured in serum-free media.

Classifying DHSs based on genomic location

The RefSeq hg19 database was downloaded from the UCSC Genome Browser and used to classify DHSs based on their genomic location. If a DHS overlapped the TSS of any transcript variant of a gene, it was classified as being a TSS DHS for that gene. Other DHSs were similarly classified as Gene Body DHSs if they overlapped any region of the gene excluding the TSS. All other DHSs were classified as Intergenic DHSs.

Microarrays

We used Affymetrix Human Exon 1.0ST microarrays to measure gene expression following ENCODE protocols. We normalized 110 microarrays (measuring 40 cell lines) together, then extracted the subset (19 cell lines) used in the present study. Probesets flagged as cross-hybridizing were first removed from the analysis (Salomonis et al. 2010). Although these arrays provide exon-level probesets, we sought gene-level expression estimates, so we grouped probesets by gene for normalization (Bemmo et al. 2008). To normalize, we used Affymetrix Power Tools (APT) with the chipstream command "rma-bg, med-norm, pm-gcbg, med-polish." This chipstream calls for an RMA normalization with gc-background correction using antigenomic background probes. After normalizing, we noticed an effect due to a switch in microarray reagents. Some of our arrays were processed differently, because our earlier array reagents become unavailable partway through the experiment.

Using hierarchical clustering and multidimensional scaling, we found the arrays to group on the basis of reagent used, rather than by biological relatedness. To make the arrays comparable, we used an R script (ComBat) to correct for this batch effect (Johnson et al. 2007). After correction with ComBat, the arrays grouped according to expected biological similarity.

Cell-type–specific expression

We identified the genomic location of genes based on matching gene symbols to the RefSeq hg19 database. If a gene from the array did not have a matching gene symbol, it was dropped from the analysis. If a gene did not have expression above background (>4.8) in at least one cell type, it was dropped from the analysis. For the remaining genes, expression values across the 19 cell lines were *Z*-transformed. The *Z*-scores for expression in each cell line were sorted. The top 200 genes were classified as UR genes and the bottom 200 genes as DR genes in that cell type. For each cell type, the UR-Other genes were compiled as follows: We first made a set comprising all UR genes from all cell types. We then removed the current cell-type UR genes from this global UR gene set. We further removed genes from this set that had expression *Z*-score ≥ 0 in the current cell type to exclude genes that had higher than mean expression in that cell type. This ensured that this set of genes was purely composed of genes that were up-regulated in other cell types and had lower than mean expression in the current cell type. A similar procedure identified DR-Other genes. To identify constitutively expressed genes, we selected genes in neither UR or DR sets across all cell lines that were expressed above background in all cell lines and had a maximum $|Z\text{-score}| < 1.7$, which resulted in 168 genes. This size compared well with the other positive sets of UR and DR genes. This gave us a list of genes that did not have a significant variance in their expression.

GO analysis

DAVID was used to identify functional categories of genes up-regulated in each cell type (Huang et al. 2009a,b).We used the GO categories and also the UP_Tissue category to identify the tissue showing the closest gene expression profile to the cell type in question. A *P*-value < 0.05 was used as the significance threshold.

PWM scans of DHS and promoter sequences

We collected PWMs for vertebrate TFs from the TRANSFAC, JASPAR, and UniProbe databases. This gave us a collection of 789 PWMs, some of which refer to the same TF. Note that we allowed multiple PWMs for each TF in our data set since it is generally not known which one reflects binding affinities more accurately. Furthermore, multiple PWMs may also reflect different binding modalities for the same factor. This could be due to, for example, the presence of specific cooperative binding partners in one cell type but not in another.

We scanned the sequence from each DHS site and (proximal) promoter sequence (−900 to +100 relative to each TSS of a gene) using these PWMs. A score was assigned to each location in the sequence based on the log-likelihood ratio of the PWM score (probability of the PWM generating the specific sequence) versus the probability that the sequence was generated by a background model. The background model used was a first-order Markov Model trained over a 500-bp window centered at the base pair being scored. This effectively corrects for the underlying dinucleotide composition and allows us to separate signal from noise (Megraw et al. 2009). Scores with a log-likelihood ratio less than 0 were not included in further analysis.

After these scores were generated for each base pair, we summed scores across a sliding window of size 60 to account for local clusters of multiple, potentially overlapping binding sites. Clusters of binding sites have been shown to be more likely to be bound by TFs as opposed to single binding sites (Gotea et al. 2010). For each DHS or promoter region, we assigned the maximum sliding window score as the TFBS score for that TF for that region.

Associating DHS TFBS scores with genes

To associate each DHS with a gene that it was potentially regulating, we found the closest TSS to the midpoint of the DHS. The DHS was then associated with that gene. In general, there were several associated DHSs for each gene, and these were assumed to be the putative regulatory regions for that gene. To assign one score for a TFBS to each gene, we picked the maximum TFBS score across all of the associated DHSs (closest gene DHS). In Split DHS, we separated DHSs into two groups—overlapping TSSs and those in other parts of the genome. We selected the maximum for each set, therefore having two features per gene per TF.

Sparse logistic regression classifier

We used a sparse logistic regression classifier that minimizes an objective function that is a linear combination of the sum of squared residuals and the ℓ_1-norm of the weights (Koh et al. 2007). We divided our data into four parts to perform a fourfold cross-validation. Three parts of the divided data were used as the training set. This training set was further divided into six parts, one of which was used as the validation set to learn the hyperparameter for the contribution of the ℓ_1-norm in the objective function. The optimal hyperparameter was selected from 10 values (0.001, 0.011, . . ., 0.091). This value was then used to evaluate the performance of the model on the original test set, and the area under the receiver operating characteristics (AuROC) was computed as a measure of performance. We performed five iterations of each fourfold cross-validation, with the data shuffled before each iteration. Results for each classification task are therefore averages over 20 different partitions of the data, which makes our results more robust to chance arrangements of the data.

Cell-type–specific expressed nonredundant TFs

To identify TFs that were cell-type–specifically expressed, we used the absolute *Z*-score and extracted gene symbols for TFs with available PWMs. We again used the gene symbol from the Affymetrix arrays to match expression to TF names in our PWM list. By using absolute *Z*-scores, we picked out genes that were cell-type–specifically up- and down-regulated. To ensure that we were picking a nonredundant set of TFs, we used STAMP (Mahony and Benos 2007) with the default settings. Starting from the TF with the highest absolute *Z*-score, we only added a TF when it had a significantly different PWM (*E*-value > 0.25) from the TFs already chosen. We stopped adding to the set when we had 10 TFs.

Conservation analysis

Genomic regions from the 46-way placental mammal track were downloaded from the UCSC Genome Browser. Only regions of size 100 bp or more were used. Coding exon sequences were extracted from the RefSeq hg19 database. BedTools (Quinlan and Hall 2010) was used to subtract exonic coding sequences from conserved regions to find noncoding conserved regions. Regions were scanned for motifs, and TFBS scores were assigned to genes in the same way as open chromatin regions.

Discriminative motif finding using MEME

We used MEME to first calculate a position-specific prior using the psp-gen tool. For example, if we wanted to identify the motifs in the DHS sequences related to the UR genes when classifying against DR genes, the positive file was the UR genes and the negative file was the DR genes. This was then used to identify motifs from 1000 positive sequences, UR gene DHS sequences in the above example. The motif width was chosen to be between 8 and 12 bp, and motifs were allowed to have zero or one occurrence per sequence. Motifs with an *E*-value < 0.05 were then compared with either the top 10 nonredundantly expressed TFs or the full list of TFs using STAMP. A STAMP *E*-value < 0.05 was accepted as a good match of a motif to a TF.

ChIP analysis

ChIP-seq peak coordinates were obtained from the ENCODE webpage in BED (narrowPeaks) file format. To assess whether DHSs are really bound by the TF or not, we checked for overlap between the coordinates of DHS and ChIP-seq peaks for that TF. To calculate AuROC, the TFBS scores in the DHS were compared against ChIP-seq peaks. Overlaps with the 60-bp window around the TFBS site were considered true positives, and others were considered false positives for AuROC calculations.

Footprinting analysis

DNase reads were used to plot the distribution of DNase-seq reads around transcription factor binding sites. The number of reads mapping to 100 bp surrounding transcription factor binding sites was counted for each site and aggregated over the 10,000 highest-scoring binding sites. A trough centered at the binding sites in such plots is called a DNase footprint, indicative of protection of the binding site against DNase digestion by a bound TF.

Data access

DNase-seq data are publicly available on the UCSC Genome Browser, the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) (accession no. GSE32970), and the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) (accession no. SRA047031). Expression data can be downloaded from the GEO database (accession nos. GSE15805, GSE12760, GSE14863). Data sets can also be found via http://labs.genome.duke.edu/ohler/research/transcription.

Acknowledgments

We thank Terry Furey and Darin London for initial processing of DNase-seq data, Molly Megraw and Ashlee Benjamin for providing code for feature generation and classification, and Danielle Maatouk for help with conservation analysis. We also thank Alexander Hartemink for helpful discussions. This work is supported by NHGRI grant U54 HG004563 to G.E.C. and NHGRI award R01 HG004065 to U.O. N.C.S. is supported by an NSF Graduate Research Fellowship and the Norwegian Research Council.

References

Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B. 2003. Computational detection of *cis*-regulatory modules. *Bioinformatics* (Suppl 2) **19:** ii5–ii14.

Bailey TL, Boden M, Whitington T, Machanick P. 2010. The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics* **11:** 179. doi: 10.1186/1471-2105-11-179.

Bemmo A, Benovoy D, Kwan T, Gaffney DJ, Jensen RV, Majewski J. 2008. Gene expression and isoform variation analysis using Affymetrix Exon Arrays. *BMC Genomics* **9:** 529. doi: 10.1186/1471-2164-9-529.

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2010. ChIP-seq identification of weakly conserved heart enhancers. *Nat Genet* **42:** 806–810.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008a. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132:** 311–322.

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008b. F-seq: A feature density estimator for high-throughput sequence tags. *Bioinformatics* **24:** 2537–2538.

Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS. 2011. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res* **21:** 456–464.

Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: A theory. *Science* **165:** 349–357.

Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A. 2008. JASPAR, the open access database of transcription factor-binding profiles: New content and tools in the 2008 update. *Nucleic Acids Res* **36:** D102–D106.

Bunt J, Hasselt NE, Zwijnenburg DA, Hamdi M, Koster J, Versteeg R, Kool M. 2011. OTX2 directly activates cell cycle genes and inhibits differentiation in medulloblastoma cells. *Int J Cancer* **131:** E21–E32.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38:** 626–635.

Cereghini S. 1996. Liver-enriched transcription factors and hepatocyte differentiation. *FASEB J* **10:** 267–282.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, et al. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133:** 1106–1117.

Das D, Nahle Z, Zhang MQ. 2006. Adaptively inferring human transcriptional subnetworks. *Mol Syst Biol* **2:** 2006.0029. doi: 10.1038/msb4100067.

Fan H, Cinar MU, Phatsara C, Tesfaye D, Tholen E, Looft C, Schellander K. 2011. Molecular mechanism underlying the differential MYF6 expression in postnatal skeletal muscle of Duroc and Pietrain breeds. *Gene* **486:** 8–14.

Fu Y, Frith MC, Haverty PM, Weng Z. 2004. MotifViz: An analysis and visualization tool for motif discovery. *Nucleic Acids Res* **32:** W420–W423.

Gotea V, Ovcharenko I. 2008. DiRE: Identifying distant regulatory elements of co-expressed genes. *Nucleic Acids Res* **36:** W133–W139.

Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* **20:** 565–577.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39:** 311–318.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459:** 108–112.

Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al. 2009. Global mapping of protein–DNA interactions *in vivo* by digital genomic footprinting. *Nat Methods* **6:** 283–289.

Huang DY, Kuo YY, Chang ZF. 2005. GATA-1 mediates auto-regulation of Gfi-1B transcription in K562 cells. *Nucleic Acids Res* **33:** 5331–5342.

Huang DW, Sherman BT, Lempicki RA. 2009a. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37:** 1–13.

Huang DW, Sherman BT, Lempicki RA. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4:** 44–57.

Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8:** 118–127.

Koh K, Kim S-J, Boyd S. 2007. An interior-point method for large-scale ℓ_1-regularized logistic regression. *J Mach Learn Res* **8:** 1519–1555.

Kool M, Koster J, Bunt J, Hasselt NE, Lakeman A, van Sluis P, Troost D, Meeteren NS, Caron HN, Cloos J, et al. 2008. Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. *PLoS ONE* **3:** e3088. doi: 10.1371/journal.pone.0003088.

Landolin JM, Johnson DS, Trinklein ND, Aldred SF, Medina C, Shulha H, Weng Z, Myers RM. 2010. Sequence features that drive human promoter function and tissue specificity. *Genome Res* **20:** 890–898.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760.
Liu H, Schmidt-Supprian M, Shi Y, Hobeika E, Barteneva N, Jumaa H, Pelanda R, Reth M, Skok J, Rajewsky K. 2007. Yin Yang 1 is a critical regulator of B-cell development. *Genes Dev* **21:** 1179–1189.
Lu R, Medina KL, Lancki DW, Singh H. 2003. IRF-4,8 orchestrate the pre-B-to-B transition in lymphocyte development. *Genes Dev* **17:** 1703–1708.
Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M. 2005. Functional evolution of a *cis*-regulatory module. *PLoS Biol* **3:** e93. doi: 10.1371/journal.pbio.0030093.
Mahony S, Benos PV. 2007. STAMP: A web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* **35:** W253–W258.
Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al. 2006. TRANSFAC and its module TRANSCompel: Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34:** D108–D110.
Megraw M, Pereira F, Jensen ST, Ohler U, Hatzigeorgiou AG. 2009. A transcription factor affinity-based code for mammalian transcription initiation. *Genome Res* **19:** 644–656.
Newburger DE, Bulyk ML. 2009. UniPROBE: An online database of protein binding microarray data on protein–DNA interactions. *Nucleic Acids Res* **37:** D77–D82.
Oda N, Abe M, Sato Y. 1999. ETS-1 converts endothelial cells to the angiogenic phenotype by inducing the expression of matrix metalloproteinases and integrin β3. *J Cell Physiol* **178:** 121–132.
Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39:** 730–732.
Ouyang Z, Zhou Q, Wong WH. 2009. ChIP-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci* **106:** 21521–21526.
Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* **21:** 447–455.
Quinlan AR, Hall IM. 2010. BED Tools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842.
Ramsey SA, Klemm SL, Zak DE, Kennedy KA, Thorsson V, Li B, Gilchrist M, Gold ES, Johnson CD, Litvak V, et al. 2008. Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. *PLoS Comput Biol* **4:** e1000021. doi: 10.1371/journal.pcbi.1000021.
Rosa A, Brivanlou AH. 2011. A regulatory circuitry comprised of miR-302 and the transcription factors OCT4 and NR2F2 regulates human embryonic stem cell differentiation. *EMBO J* **30:** 237–248.
Salomonis N, Schlieve CR, Pereira L, Wahlquist C, Colas A, Zambon AC, Vranizan K, Spindler MJ, Pico AR, Cline MS, et al. 2010. Alternative splicing regulates mouse embryonic stem cell pluripotency and differentiation. *Proc Natl Acad Sci* **107:** 10514–10519.
Schoenherr CJ, Anderson DJ. 1995. The neuron-restrictive silencer factor (NRSF): A coordinate repressor of multiple neuron-specific genes. *Science* **267:** 1360–1363.
Sharan R, Ovcharenko I, Ben-Hur A, Karp RM. 2003. CREME: A framework for identifying *cis*-regulatory modules in human–mouse conserved segments. *Bioinformatics* (Suppl 1) **19:** i283–i291.
Sinha S, Adler AS, Field Y, Chang HY, Segal E. 2008. Systematic functional characterization of *cis*-regulatory motifs in human core promoters. *Genome Res* **18:** 477–488.
Smith AD, Sumazin P, Zhang MQ. 2005. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci* **102:** 1560–1565.
Sokalski KM, Li SK, Welch I, Cadieux-Pitre HA, Gruca MR, DeKoter RP. 2011. Deletion of genes encoding PU.1 and Spi-B in B cells impairs differentiation and induces pre-B cell acute lymphoblastic leukemia. *Blood* **118:** 2801–2808.
Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D, et al. 2011. Open chromatin defined by DNase I and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res* **21:** 1757–1767.
Stormo GD, Zhao Y. 2010. Determining the specificity of protein–DNA interactions. *Nat Rev Genet* **11:** 751–760.
Suzuki H, Forrest AR, van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, Lassmann T, Ravasi T, Hasegawa Y, de Hoon MJ, et al. 2009. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet* **41:** 553–562.
Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126:** 663–676.
Thorrez L, Laudadio I, Van Deun K, Quintens R, Hendrickx N, Granvik M, Lemaire K, Schraenen A, Van Lommel L, Lehnert S, et al. 2011. Tissue-specific disallowance of housekeeping genes: The other face of cell differentiation. *Genome Res* **21:** 95–105.
van Steensel B, Dekker J. 2010. Genomics tools for unraveling chromosome architecture. *Nat Biotechnol* **28:** 1089–1095.
Vavouri T, Elgar G. 2005. Prediction of *cis*-regulatory elements using binding site matrices—the successes, the failures and the reasons for both. *Curr Opin Genet Dev* **15:** 395–402.
Wu C, Gilbert W. 1981. Tissue-specific exposure of chromatin structure at the 5′ terminus of the rat preproinsulin II gene. *Proc Natl Acad Sci* **78:** 1577–1580.
Yamashita R, Suzuki Y, Sugano S, Nakai K. 2005. Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene* **350:** 129–136.
Yordy JS, Moussa O, Pei H, Chaussabel D, Li R, Watson DK. 2005. SP100 inhibits ETS1 activity in primary endothelial cells. *Oncogene* **24:** 916–931.
Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R, et al. 2007. Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318:** 1917–1920.
Zhu J, He F, Hu S, Yu J. 2008. On the nature of human housekeeping genes. *Trends Genet* **24:** 481–484.

Received November 21, 2011; accepted in revised form June 8, 2012.