arXiv:1210.3313v1 [q-bio.QM] 11 Oct 2012

Identifying and Mapping Cell-type Specific Chromatin Programming of Gene Expression

Troels T. Marstrand[1] and John D. Storey[1,2*]

[1]Lewis-Sigler Institute for Integrative Genomics
[2]Department of Molecular Biology
Princeton University, Princeton, NJ 08544, USA.
*Corresponding author: jstorey@princeton.edu

Abstract

A problem of substantial interest is to systematically map variation in chromatin structure to gene expression regulation across conditions, environments, or differentiated cell types. We developed and applied a quantitative framework for determining the existence, strength, and type of relationship between high-resolution chromatin structure in terms of DNaseI hypersensitivity (DHS) and genome-wide gene expression levels in 20 diverse human cell lines. We show that ∼25% of genes show cell-type specific expression explained by alterations in chromatin structure. We find that distal regions of chromatin structure (e.g., ±200kb) capture more genes with this relationship than local regions (e.g., ±2.5kb), yet the local regions show a more pronounced effect. By exploiting variation across cell-types, we were capable of pinpointing the most likely hypersensitive sites related to cell-type specific expression, which we show have a range of contextual usages. This quantitative framework is likely applicable to other settings aimed at relating continuous genomic measurements to gene expression variation.

Abbreviations: ARS, Angle Ratio Statistic; DHS, DNaseI hypersensitivity; SI, Supplementary Information

Note: The Supplementary Information may be found among the source files in arxiv_SI.pdf.

1 Introduction

Humans, like all other multicellular organisms, possess a large number of distinct cell-types, each of which is specialized for a particular function within the body. Cells from a variety of tissue types exhibit different gene expression profiles relating to their function, where typically only a fraction of the genome is expressed. As all somatic cells share the same genome, specialization is in part achieved by physically sequestering regions containing non-essential genes into heterochromatin structures. Genes which are needed for the particular task of the cell-type display an accessible chromatin structure allowing for the binding of transcription factors and other related DNA machinery and subsequent gene expression.

To date, most studies have been limited to considering the chromatin accessibility surrounding the promoter region of genes, which is typically proximal ($<$10kb) to the transcription region in just one or very few cell-types or experimental conditions [53, 4, 43] . However, it is also of interest to understand how larger regions ($\gg$10kb) of chromatin structure relate to a gene's expression variation across multiple cell types, disease states, or environmental conditions. Recently, several large-scale international collaborations have started to generate data that can be used for this purpose [38], although doing so requires new developments in computational methods [21, 22, 11].

To this end, we undertook a genome-wide investigation to characterize the relationship between variations in chromatin structure and gene expression levels across 20 diverse human cell lines (SI, Table S1). We utilized data on chromatin structure as ascertained through DNaseI hypersensitivity (DHS) measured by next-generation deep sequencing technology, and gene expression data measured by Affymetrix exon arrays. Replicated data on 10 cell lines were also utilized to assess the robustness of our method.

Relating DHS to gene expression levels across multiple cell-types is challenging because the DHS represents a continuous variable along the genome not bound to any specific region, and the relationship between DHS and gene expression is largely uncharacterized. In order to exploit variation across cell-types and test for cell-type specific relationships between DHS and gene expression, the measurement units must be placed on a common scale, the continuous DHS measure associated to each gene in a well-defined manner, and all measurements considered simultaneously. Moreover, the chromatin and gene expression relationship may only manifest in a single cell-type, making standard measures of correlation between the two uninformative because their relationship is not linear over a continuous range, as shown in Fig. 1 (further details in SI and Figs. S2-S6).

The computational approach developed here provides a powerful, tractable, and intuitive way of representing these data and capturing biologically informative relationships. We were able to characterize the level to which variation of chromatin accessibility is associated with gene expression variation in a cell-type specific manner. Within genomic segments of significant chromatin gene expression concordance, our methodology is further capable of pinpointing the most likely local sites related to the detected association. We show that such sites are context specific and can be shared across genes within a single cell-type or across several cell-types. Our quantitative framework has some generality in that it may be readily applied to associate any quantitative measure along the genome to gene expression variation.

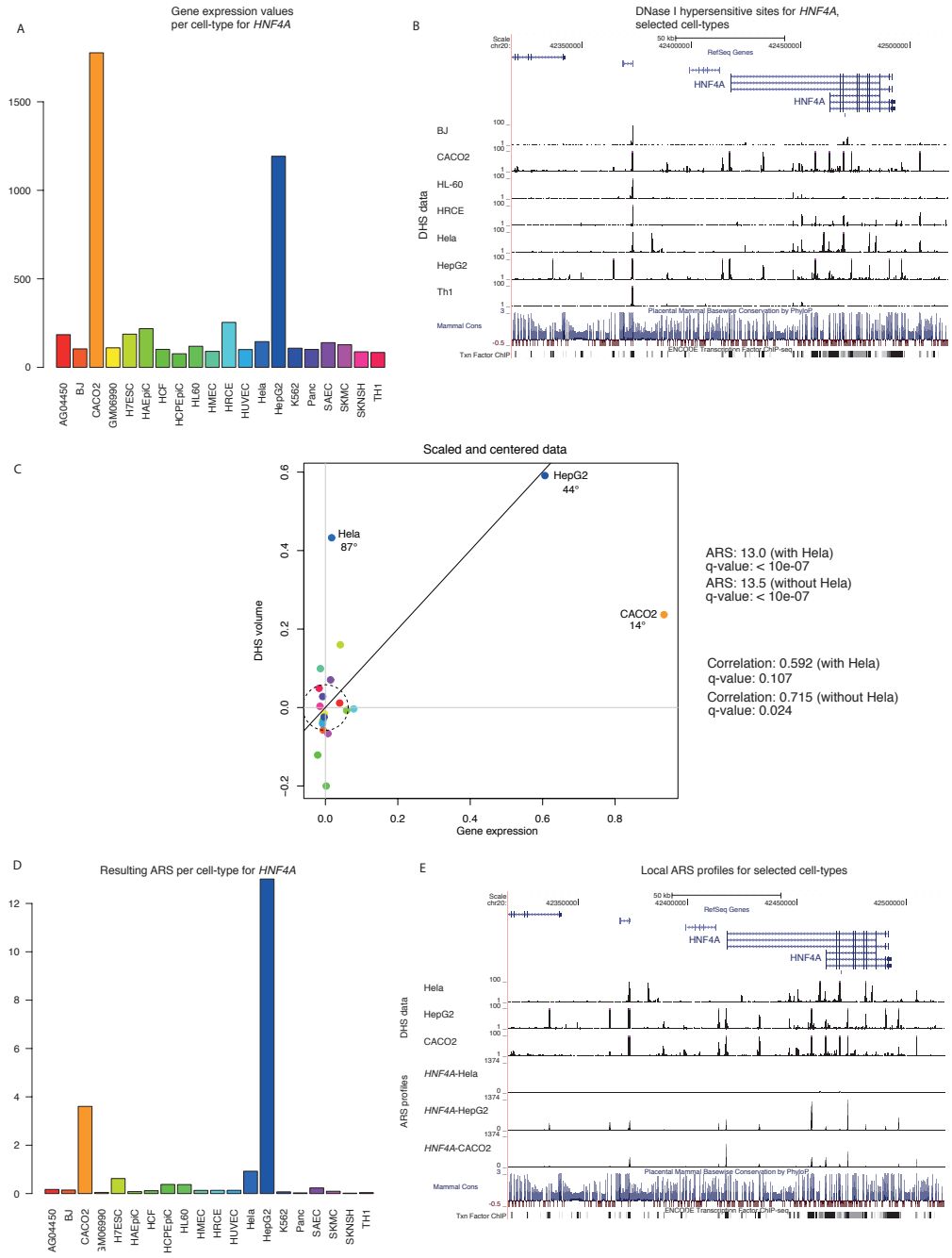


Figure 1: **Overview of Data and Proposed Approach.** (A) Gene expression measurements for twenty cell lines on an example gene, *HNF4A*. (B) DNase-I Hypersensitivity (DHS) fragment sequencing counts in a region about the gene. (C) The DHS signal is captured by summing the overall number of fragments over a given segment size (e.g., $\pm$ 100kb) about the gene's transcriptional start sight (TSS) to obtain a "DHS volume". After global normalization, the gene expression data and DHS volume measures are scaled to lie on the unit interval [0,1] and the data are centered about the origin according to the two-dimensional medoid. For the *HNF4A* example, three outliers are clearly visible; for example, HepG2 displays both chromatin accessibility and active gene expression, whereas Hela displays only chromatin accessibilty. The goal is to quantitatively capture the isolated relationship seen in HepG2 and assess whether this relationship is statistically significant. Traditional measures of linear correlation are not suitable for identifying this type of signal, as shown by the substantial change seen after removal of a single cell line, Hela, even though the data for Hela are expected to exist for many genes and cell lines. The proposed ARS is robust to Hela since the measure is based on angular placement and the median distance to the medoid of the data (dashed circle). (D) The ARS statistic is calculating by first quantifying the relative distance to the origin for each cell line in a robust manner. An angular penalty for each cell line is then calculated to quantify cell-types concordant in both expression and DHS measured. This quantity is measured in terms of angular distance from the 45° degree line, and it is then multiplied times its respective relative distance to give and overall score for each cell line. The maximum score is taken as the statistic for the given gene, allowing a comparison across all genes. (E) A local version of the ARS statistic we introduce can pinpoint DHS "peaks" contributing the most to the detected association. See main text for details on the proposed methods.

2 Results

2.1 Genome-wide profiling of chromatin accessibility and gene expression

We utilized data on genome-wide, high-resolution chromatin accessibility measurements for 20 distinct human primary and culture cell lines that were obtained by an established sequencing-based method [34]. In principle, accessible "open" chromatin is cleaved by the non-specific endonuclease DNaseI, and the cleaved fragments are sequenced to provide a high-resolution, genome-wide map of DNaseI hypersensitivity (DHS) for every cell-type (SI, Table S2). The interpretation of these data is that increased fragment counts within a region are indicative of greater chromatin accessibility. To investigate the impact of regional chromatin accessibility on gene expression variation, we likewise utilized genome-wide gene expression measurements in each cell line from Affymetrix exon arrays (SI, Table S4). A total of 19,215 genes were analyzed after preprocessing (Methods).

With these quantifications, we sought to characterize the relationship between chromatin accessibility and gene expression in a cell-type specific manner, summarized in Fig. 1. To this end, the cell-type specific chromatin profiles were quantified by integrating the DHS fragment counts over increasingly larger genomic segments relative to the gene of interest (SI, Fig. S7) to obtain a cell-type specific regional DHS volume. We selected a range of segments that were likely to encompass all proximal (TSS $\pm$2.5kb) and most distal regulatory elements (TSS $\pm$50kb, $\pm$100kb, $\pm$150kb, $\pm$200kb, $\pm$100kb minus proximal 2.5kb, and $\pm$200kb minus proximal 2.5kb). Additionally, to account for copy number variation and chromosome arm related effects, the obtained DHS volumes were scaled on either side of the centromere to arrive at equilibrium across samples (SI, Fig. S8). Alternative representations of DHS signal [11, 6] could be utilized at this step, although we did not identify any advantages in doing so. Gene expression values were summarized as the mean intensity across all probe-sets linked to a given RefSeq-gene.

2.2 Detecting cell-type specific chromatin accessibility and gene expression concordance

Due to the "on-off" nature of DHS and subsequent transcription, there will not necessarily be a linear relationship between DHS and gene expression measures. Using correlation or correlation-like statistics to associate the two measurements across all cell-types proved to be unreliable and uninformative (further details in SI and Figs. S2-S6). One of the key types of relationships we sought to detect is of the type shown in Fig. 1, where one or very few cell types are outliers from the others. The standard Pearson correlation statistic is not well-suited for this scenario. First, it requires the data to be jointly Normal in order to obtain parametric p-values, but the Normal assumption does not hold for these data (SI, Fig. S3). Second, this correlation statistic is unstable when there are outliers, even when using permutation based p-values, demonstrated directly on these data (SI, Figs. S2 and S4). The rank-based Spearman correlation statistic is a potential alternative, but it shows very poor power relative to the method proposed here as shown in Fig. 2. (See also SI, Figs. S5 and S6). For example, at a false discovery rate (FDR) ≤ 0.05, the proposed method identifies 2538 genes with a cell-type specific DHS and gene expression relationship whereas the Spearman statistic identifies only 286 (Figs. 2, S5, and S6).

The new statistic proposed here is designed to be appropriate for scenarios when both measurements are restricted to a narrow relative range with one or very few cell-types appearing as distinct outliers. To evaluate the relationship between the DHS volume of a genomic segment and

gene expression, we took into account the compactness of the measurements versus any distinct outliers in both dimensions and whether the outliers were concordant in both measurements (i.e., a simultaneous increase or decrease) to form an overall composite measure called an Angle Ratio Statistic (ARS) (detailed in Fig. 1, Fig. S1, Methods, and SI). To summarize, we first scale and median center the DHS volume and expression data, respectively, for a given gene. We then calculate the relative distance of each cell type to the overall center of the data, which serves as a way to measure the degree to which each cell type is an outlier. In order to measure concordance of DHS volume and gene expression, we calculate the angular distance between each point and the 45° line of identity, penalizing points further away from the line of identity according to a data-derived exponential function. These two quantities are then multiplied to form an ARS_i value for each cell type ($i = 1, 2, \ldots, 20$), and the maximal value $\mathrm{ARS}_{\max}$ is the overall statistic that quantifies cell-type specific DHS volume and gene expression concordance for a given gene.

To identify statistically significant genes from $\mathrm{ARS}_{\max}$, we constructed a null distribution based on randomization of the observed experimental data (see Methods, SI, and Fig. S9). $\mathrm{ARS}_{\max}$ values obtained from the randomized data were used as a basis for determining a p-value of the observed $\mathrm{ARS}_{\max}$ for each gene. False-discovery rate (FDR) based statistical significance and the proportion of genes with a true chromatin accessibility and gene expression relationship were estimated from the p-values [45] (Figure 2b).

We estimate that ∼25% of genes show concordance between chromatin accessibility and gene expression variation in a cell-type specific manner. While our strategy is capable of detecting outliers showing negative concordance (decreased chromatin accessibility and decreased gene expression), none were found to be significant at FDR ≤ 0.05. The number of significant genes increased by inclusion of distal DHS volume (Fig. 2b, column 2), indicating that distal chromatin programming effects are more widespread in a genome-wide sense. On the other hand, using the proximal DHS volume we observe a greater empirical effect size compared to the distal DHS volumes (Fig. 2b, columns 3-5).

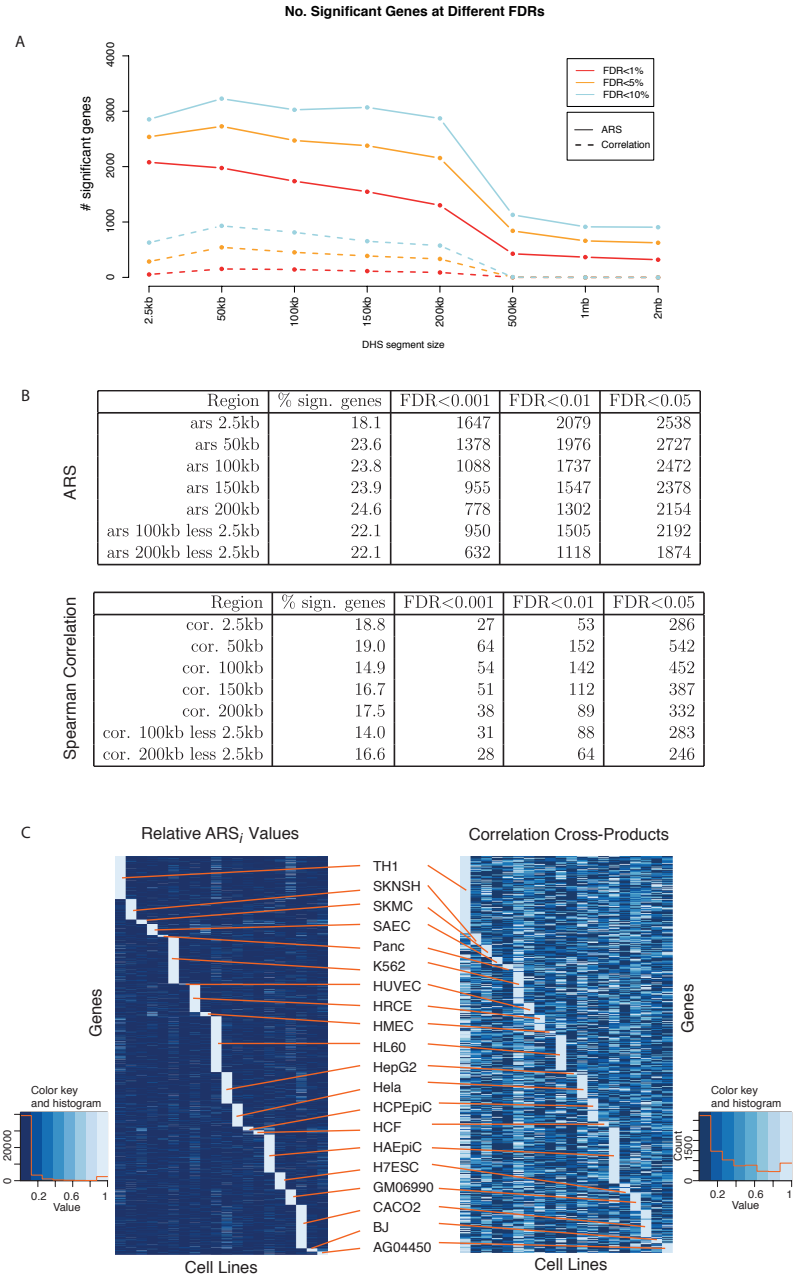
This observation is explained by the aggregation of genes significant for the same cell-type along the genome [44]. Testing whether one or more significant genes within a ±100kb region were associated with the same cell-type we found that 481 out of 668 significant genes within the specified boundary stem from the same cell-type (Fishers exact test p-value $<$ 2.2e-16; SI and Fig. S15). It is however important to note that inclusion of increasingly distal regions also increases the noise in the DHS volume, wherefore the effect size and ultimately the number of true associations starts to decline (Fig. 2a).

2.3 Experimental replication

To assess reproducibility, we tested the concordance of significant results among replicated data for 10 cell-types. Based on two independent measurements of DHS and gene expression, respectively, we calculated the fraction of predictions preserved in all four-way comparisons (SI). We found that between 86% to 91% of significant genes (FDR ≤ 0.05) were identical (SI, Fig. S16).

2.4 Gene ontology and pathway analysis

To determine the biological coherence of the set of genes found to be significant for each cell-type, we performed a gene ontology (GO) enrichment analysis [10]. The method computes enrichment within the process and function components of GO categories and assigns a numerical significance



B

ARS

Region	% sign. genes	FDR<0.001	FDR<0.01	FDR<0.05
ars 2.5kb	18.1	1647	2079	2538
ars 50kb	23.6	1378	1976	2727
ars 100kb	23.8	1088	1737	2472
ars 150kb	23.9	955	1547	2378
ars 200kb	24.6	778	1302	2154
ars 100kb less 2.5kb	22.1	950	1505	2192
ars 200kb less 2.5kb	22.1	632	1118	1874

Spearman Correlation

Region	% sign. genes	FDR<0.001	FDR<0.01	FDR<0.05
cor. 2.5kb	18.8	27	53	286
cor. 50kb	19.0	64	152	542
cor. 100kb	14.9	54	142	452
cor. 150kb	16.7	51	112	387
cor. 200kb	17.5	38	89	332
cor. 100kb less 2.5kb	14.0	31	88	283
cor. 200kb less 2.5kb	16.6	28	64	246

Figure 2: **Statistical significance for ARS and correlation across genomic segments** (A) Depicts the number of significant genes found at increasingly larger genomic segments for ARS and Spearman correlation, respectively (solid line is ARS and dashed line is Spearman correlation). (B) Statistical significance according to DHS volume segment size. Column 2 shows the percentage of genes estimated to have concordant DHS volume and gene expression variation as captured by $\text{ARS}_{\max}$ ($1-\hat{\pi}_0$, as estimated in [45]). Columns 3-5 show the number of statistically significant genes at various FDR cut-offs. While the 2.5Kb window shows more significant genes at the stringent FDR cut-offs, indicating a larger effect size, the overall percentage of genes showing a relationship is notably lower than the more distal DHS volumes. Compared to Spearman correlation, ARS is more powerful at detecting these associations (see SI for further details). (C) The relative ARS_i values across all cell-types for significant genes in the $\pm$100kb region versus the analogous components for Spearman correlation (the cross product terms that sum to form the overall correlation). The ARS_i values distinguish cell lines that have a strong DHS and expression concordance substantially more clearly than the Spearman correlation, showing that the traditional correlation is more likely to generate spurious results from small changes to the data. Enrichment of biological functions for the significant genes found by either method corroborates this finding (see SI, Fig. S14)

to the findings. In nearly all cases the results were in agreement with the actual biology; see `http://encode.princeton.edu/` for results on all DHS segment sizes. For example, human T-cells showed a strong enrichment of T-cell receptor related genes, whereas hepatic cells showed enrichment of lipid metabolism related genes. KEGG pathways [24, 7] were likewise enriched in a cell-type specific manner. For example, HepG2 showed significant enrichment for genes within the bile acid synthesis and drug metabolism, while HL60 showed significant enrichment within the hematopoietic cell lineage (data not shown).

Furthermore, all genes detected within each cell type at FDR < 0.05 ($\pm$ 100kb DHS volume) were analyzed through the use of Ingenuity Pathways Analysis (Ingenuity Systems®, `www.ingenuity.com`). For all but three cases out of 20 (two cell-types likely had too few significant genes detected to get reliable annotations), the category "Physiological System Development and Function" was in clear correspondence with that expected given the cell type, (SI, Fig. S14). For instance, TH1 was enriched for "cell-mediated immune response", K562 for "hematological system development and function", and H7ESC for "embryonic development". For each gene, there tended to be low relative ARS_i across the remaining cell types, indicating that we detected truly cell-type specific genes as clear outliers on a genome-wide scale. However, some cases showed large relative ARS_i in a few tissues, which prompted us to investigate these instances further.

Among genes with a statistically significant $\mathrm{ARS}_{\max}$ statistic, additional inspection of the remaining ARS_i were explored for detection of possible sub-structures. We calculated relative ARS values within each gene dividing all ARS_i by $\mathrm{ARS}_{\max}$. In addition to many instances of singular outliers, we detected a gradient behavior among significant genes, where a few cell-types were evident as outliers (Fig. S17).

2.5 Local ARS Profiles

The DHS data itself provides a rich source of information about regulatory elements in the genome. However, when used in conjunction with gene expression data across differing cell types, there is an opportunity to discover which locations of chromatin accessibility drive gene expression in a cell-type specific manner. This goal prompted us to develop a technique to model the relationship for fine-scale segments of DHS volume across the larger segments. As the above strategy focused on examination of chromatin gene expression interactions over genomic segments, investigation of fine-scale patterns allowed us to: (i) validate that distal regulatory regions were indeed present as peaks in chromatin accessibility concordant with gene expression in a cell-type specific manner, (ii) perform sequence analyses of these chromatin accessibility peaks, (iii) compare localized associations across cell-types or within a single gene, and (iv) provide a framework for quantifying regions of interest on a continuous scale for investigation of regulatory elements.

We therefore extended our approach to allow one to identify and map DHS sites to genes on which they show strong evidence for playing a regulatory role in a cell-type specific manner. This was carried out by providing a fine-scale version of the ARS quantification, called a "local ARS profile" for genes with a statistically significant $\mathrm{ARS}_{\max}$ statistic over a larger segment. The peaks of the local ARS profiles pinpoint which DHS are most influential in explaining the cell-type specific gene expression variation, thereby indicating that they have the most regulatory potential. We retained the gene expression values for a given significant gene, and now considered the DHS volume within non-overlapping consecutive regions at a high resolution 60 base pair windows. The ARS statistic was calculated for each 60bp window, which can then be plotted over the entire region used in identifying the gene as statistical significant. For example, for a gene significant

with respect to a ±200kb DHS volume, we calculated ∼6700 local ARS statistics for each cell type. These can then be plotted in such a way that the signal emanating from that location is visible, loosely analogous to a LOD score profile in linkage analysis. Additional steps were taken, involving scaling across the 60 bp windows to preserve a valid interpretation of their relative magnitudes (SI).

We first selected the subset of local ARS profile "peaks" by thresholding the local ARS profiles in a principled manner (Methods), and we analyzed both positional biases and sequence compositions as they relate to function. We then analyzed the entire trajectories of local ARS profiles at specific loci, showing that they identify both known and putative regulatory DHS for given genes.

2.6 Positional bias of putative regulatory DHS

Because the overall statistical significance increases when calculating DHS volume over more distal regions up to 200kb (Fig. 2), we investigated the positional bias of local ARS peaks in a cell-type specific manner. Figure 3a shows smoothed densities of positional local ARS peak counts by cell type, which exhibit high cell-type specific differences, specifically the density around the TSS. Random densities were generated by randomly assigning positional counts to tissues in equal proportions to the observed counts, where it can be seen that the cell-type differences are no longer present (SI, Fig. S18). This points to the existence of cell-type specific biases in the base-pair distance of regulatory DHS to TSS.

2.7 Sequence analysis of peaks in local ARS profiles

We next sought to characterize the functional significance of sequences corresponding to local ARS peaks. Since a general indicator of functionality is conservation, we extracted the conservation track values (phastCons44wayPrimate, hg18) [41] corresponding to the local ARS peaks and to the negative control set (Methods). Values range from 0 to 1, with 1 indicating the most conserved. The regions with local ARS peaks were significantly more conserved than regions from the negative control set (Kolmogorov-Smirnov p-value $< 2.2e\text{-}16$, SI, Fig. S19), indicating substantial conservation of sequences corresponding to local ARS peaks.

DNase-I hypersensitive sites are well established markers of regulatory and other DNA binding proteins. We therefore sought to establish if known cell-type specific transcription factors binding sites (TFBSs) are over-represented in the local ARS peaks relative to the negative control set (Methods). Since regions distal to the TSS are rarely studied in this context, we eliminated all local ARS peaks and negative controls that fell within ±10 kb of the TSS. This step was taken to demonstrate that the proposed approach is capable of detecting distal TFBS, up to 200kb from the TSS.

We utilized the JASPAR database [31] to identify TFBS that are differentially represented in the local ARS peaks relative to the negative control set (Methods). The over- and under-represented TFBS show distinct cell-type specific patterns and provide a rich insight into cell-type specific gene regulation (Fig. 3b), several of which are listed here:

- Among the hepatocyte nuclear factors we found *HNF1B* (TCF-2) and *HNF4A* to have significant chromatin gene expression concordance in HRCE and HepG2, respectively (SI, Fig. S20 and Fig. S21). Furthermore we found the local ARS profiles in the respective tissues to display a marked over-representation of the factor in question, *HNF1B* in HRCE and *HNF4A*

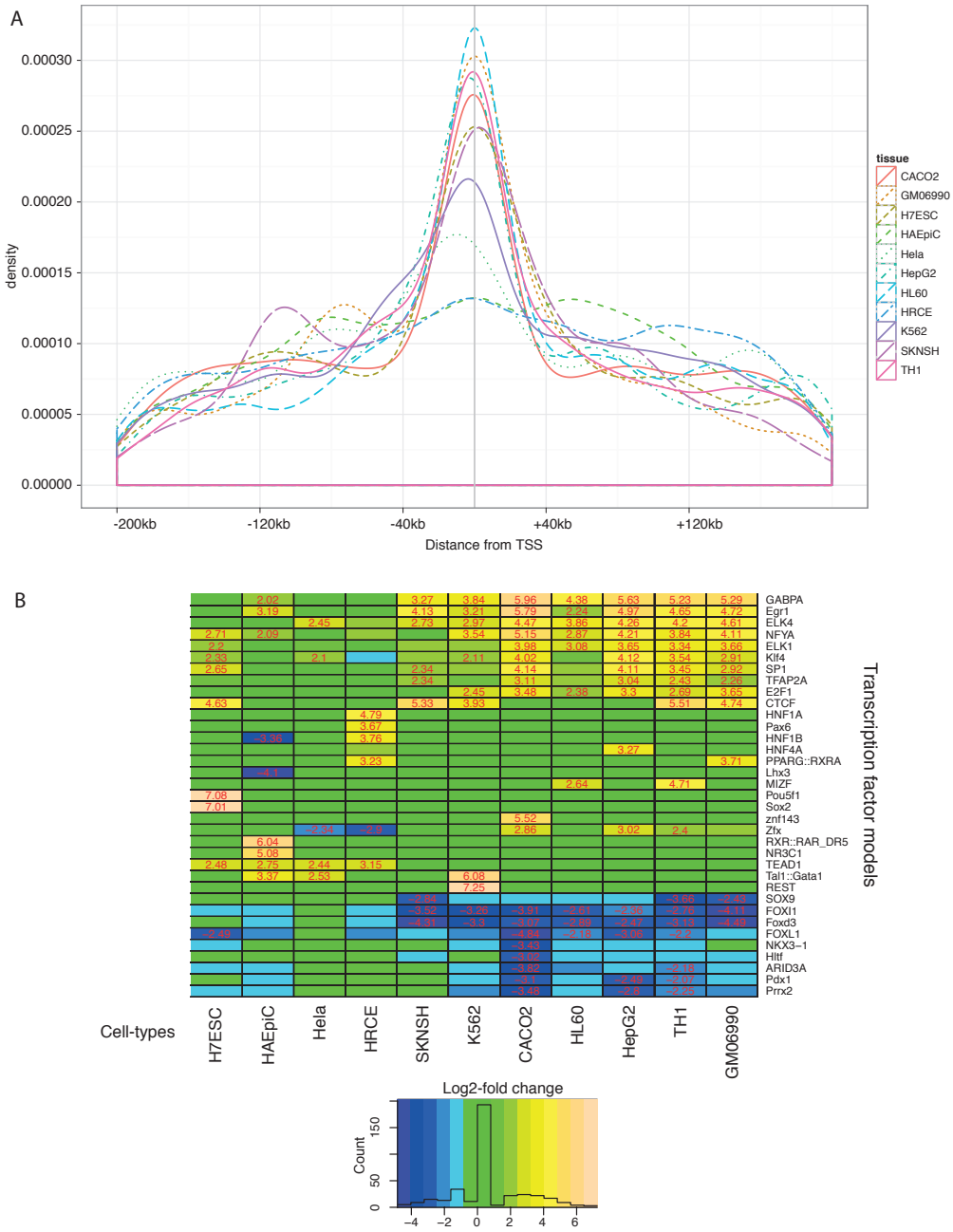


Figure 3: **Analysis of local ARS profiles.** (A) *Distribution of local ARS peaks relative to the TSS according to cell type.* The positional bias of cell-type specific local ARS peaks as measured by the density of local ARS peaks within cell lines with respect to position from to the TSS. Clear differences in the amount of distal regulation are seen across the cell-types and the density around the TSS differ markedly among cell types. For example, HL60 shows a more proximal signal relative to that of HAEpiC. (B) *Transcription factor binding site analysis among local ARS peaks occurring 10kb to 200kb from the TSS.* Sequences corresponding to local ARS peaks within significant cell-type specific genes were searched with known transcription factor binding site models, and the relative over- and under-representation was assessed based on a negative control set. Instances of absolute $\log_2$ fold-change ≥ 2 are displayed within the relevant cell types. Over-representation is indicative of a preferential transcription factor binding site, and is therefore a likely regulatory candidate for the observed gene expression. Under-represented sites indicate factors which should be avoided to maintain proper cell-type specific expression profiles. For instance, Sox2 and Pou5f1 (Oct4) were observed solely over-represented in the embryonic cells, H7ESC.

in HepG2. Mutations in *HNF1B* have been associated with a broad range of renal diseases [50], and *HNF4A* is essential for hepatocyte differentiation and morphology [29].

- H7ESC was found to show over-representation of *SOX2* and *POU5F1* (Oct-4) both essential for self-renewal in undifferentiated stem cells.

- *NFYA* (a CCAAT-binding protein) was found over-represented in almost all tissues. This factor is essential for enhancer function by requiting distal transcription factors to the proximal promoter region [52]. The ubiquitous CCAAT-binding factor family is linked to cellular differentiation in a variety of tissues [26].

- RXR-RAR was found in HAEpiC (human amniotic epithelial cells). The co-expression of the retinoic acid receptors (RARs) and the retinoid X receptors (RXRs) [37] are essential for proper placental development, and retinoid X receptor (RXR) null mouse mutants are lethal after 10 days due to placental defects [36].

- Forkhead binding sites were found to be primarily under-represented, specifically *FOXD3* was under-represented in, among others, the leukemic cell-types. Silencing of *FOXD3* by aberrant chromatin modification has been implicated in leukemogenesis [37]. Over-expression of *FOXD3* prevents neural crest formation [30]. Interestingly, binding sites for the factor were under-represented in SKNSH, a neuroblastoma derived from neural crest cells.

- NF-κB was found over-represented in TH1, where it promotes the expression of, among others, interleukin 12 (IL-12) essential for TH1 development [28].

The differentially represented TFBSs were distributed largely distal. For all cell-types, from 68% to 79% were located more than $\pm$50kb away from the TSS. We repeated the analysis with only the proximal regions ($\pm$10kb from the TSS), and we found that important known cell-type specific motifs were no longer detected (SI, Fig. S22).

2.8 Mapping putative regulatory DHS to genes

We also investigated the utility of considering the entire trajectory of local ARS profiles at a locus to characterize the regulatory architecture of cell-type specific expression. We investigated in detail two well-characterized examples of regulatory interactions at the β-globin (*HBB*) locus control region and at the stem cell leukemia (*SCL*) gene, also known as *TAL1*, with several more appearing in SI (Figs. S23-S35). It can be seen from these analyses that the local ARS profiles provide a means to map DHS sites to genes in a cell-type specific manner.

The *HBB* (β-globin) locus control region (LCR) comprises an array of functional elements that *in vivo* gives rise to five major DNase I hypersensitive sites (HS1-HS5 [49, 14, 19], Fig. 4) upstream of *HBE1* (ϵ-globin) on the short arm of chromosome 11. All five sites were present in cell line K562 according to our DHS data (see Figs. S23- S27 for complete data across all 20 cell-types, and Fig. S28 for local ARS profiles across all genes at this locus control region). Although the DHS volume at these sites contributed to both *HBE1* and *HBB* yielding statistically significant ARS values, the relative importance of HS1-5 differs significantly between these two genes, clearly detected by the local ARS profiles (Fig. 4a).

In the case of *HBE1*, we observed local ARS peaks for HS1 at -6.1kb and to a lesser extent HS3 and HS4 (-14.7 and -18kb respectively). For *HBB* we observed similar local ARS profiles for

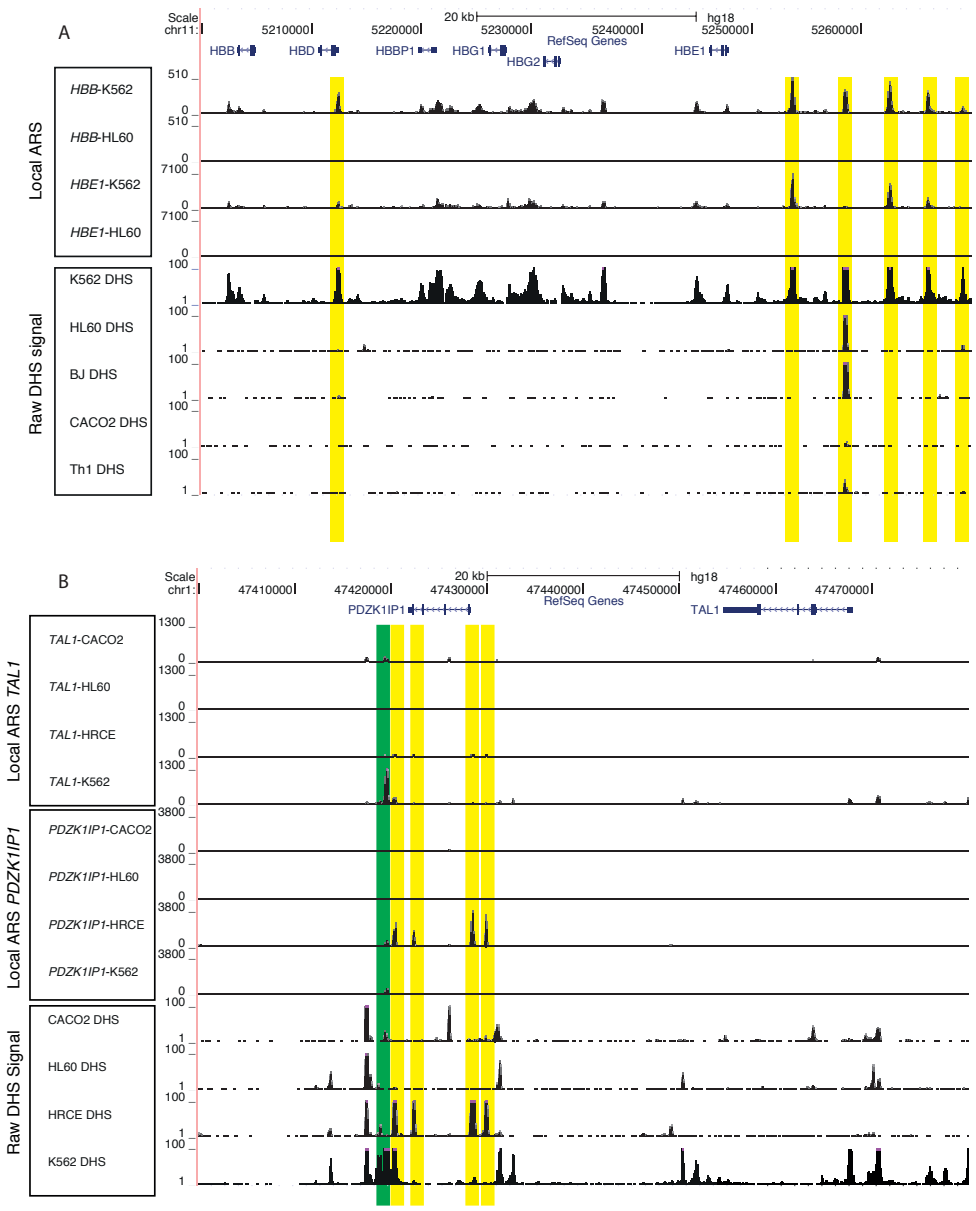


Figure 4: **Mapping putative regulatory DHS with local ARS profiles at two loci.** (A) *β-globin locus control region.* The DHS data for five cell lines (out of 20) are shown, as well as the local ARS profiles for *HBB* and *HBE1* in the K562 and HL60 cell lines. The transparent yellow boxes indicate regulatory regions, specifically hypersensitive regions 1-5 (HS1-5), together with a less characterized site upstream of *HBD*. It can be seen that *HBB* and *HBE1* show different local ARS profiles, indicative of differences in usage of regulatory elements. The local ARS profile shows no peak in HL60 despite the existence of a hyper-sensitive site when considering DNaseI profile alone. The full data and local ARS profiles for all 20 cell lines and both genes are displayed in Figs. S23-S27. (B) *TAL1 locus.* We identified *TAL1* as statistically significant with its maximal ARS in the K562 cell line across all tested genomic segments. Local ARS profiles show a dominant effect from the +40 enhancer region (green box), spanning *PDZK1IP1*. DHS signals across multiple cell-types were correctly not detected to be associated with the expression of *TAL1*. Furthermore, note that even though the DHS data for *TAL1* and *PDZK1IP1* are largely overlapping, they nevertheless have distinct local ARS profiles due to their different patterns of gene expression. This demonstrates that ARS is capable of separating interwoven signals across cell-types for neighboring genes, and that there is information to be gained by combining DHS and gene expression profiling. The full data for all 20 cell lines and local ARS profiles are displayed in Figs. S29-S31.

HS1, HS3 and HS4, and smaller local ARS values for HS2 (-10.9kb). It has previously been shown that HS1 is a stable chromatin structure [14] through out development and essential for *HBE1* expression [40] due to a GATA-1 binding site, while HS2, 3 and 4 show synergistic enhancement of expression of *HBB* by formation of the LCR holocomplex [15, 27, 48]. Finally, the element upstream of *HBD* has also been reported to specifically enhance transcription of *HBB* [1]. While HS5 is present in the DHS data for K562, similar open chromatin structures were detected in other tissues. HS5 (-21kb) is not in concordance with tissue-specific gene expression of either *HBE1* or *HBB*, an observation in line with this site's function as an insulator and CTCF binding site [47].

TAL1 encodes a basic helix-loop-helix protein which is essential for the formation of all hematopoietic lineages (SI, Figs. S29- S31 for all data across the 20 cell-types). Previous studies using chromatin structure, comparative sequence analysis, transfection assays [13, 18], and transgenic mice [5, 17, 16, 35, 42] have identified a total of five enhancers modulating the expression of *TAL1*. We detect *TAL1* as significant with maximal cell line K562 across all tested genomic segments (from $\pm$2.5KB to $\pm$200KB) with the most significant ARS_{max} occurring for $\pm$50kb. Further investigation by the local ARS profile (Fig. 4b) showed that while proximal regulatory sites were correctly identified, the most dominant signal is by far confined to the +40 enhancer region and is an order of magnitude greater than other signals. While the *TAL1* +40 region is downstream of *PDZK1IP1*, it was not linked to the expression of this gene which was detected as significant in HRCE. The +40 enhancer region has been shown to direct expression in transgenic mice to primitive, but not definitive erythoblasts, such as the phenotype displayed by K562. This example demonstrates that our methodology is capable of identifying regions of regulatory potential, which otherwise requires laborious effort to annotate.

Local ARS profiles showed both differences and similarities across genes as well as cell-types. A few examples included:

- *CCR2* and *CCR5* were significant for two different cell-types, HL60 and TH1, respectively (SI, Fig. S32).
- Part of the HOX-cluster crucial for kidney development in mammals (*HOXD8*, *HOXD4*, and *HOXD3*) showed identical local ARS profiles (SI, Fig. S33), and all were significant genes in HRCE [9].
- Another example of shared profiles, but across several cell-types instead of across several genes, was seen with *LOXL2*, a gene essential for biogenesis of connective tissue, which is detected as an outlier in SKMC and has high relative ARS values in HAEpiC and BJ (SI, Fig. S34). Further fine-scale investigation showed a solid overlap in the local ARS profiles (Fig. S35).

These observations point to a potentially widespread sharing of regulatory mechanisms both across genes and cell-types.

3 Discussion

As the epigenome in multicellular organisms is a dynamic entity whose variation leads to reprogramming of gene expression [51], it is a likely candidate in the etiology of disease complementary

to that of mutations in DNA [39, 20]. It is therefore of considerable interest to identify and characterize the regulatory regions contributing to gene expression variation with respect to a given disease.

We have presented a framework for quantifying relationships between chromatin structure and gene expression across multiple conditions (here, cell-types), facilitating a new avenue for understanding cellular responses by localizing and characterizing regions of regulatory potential. The local ARS profiles we introduced allow specific hypersensitive regions to be associated with condition-specific gene expression, thereby conferring contextual regulatory information not obtainable using DHS data alone. This effectively pinpoints a shortlist of primary candidates for further functional studies. We found the peaks from the local ARS profiles in statistically significant segments to be both highly conserved and enriched for known transcription factor binding sites as far as 200kb from transcription start sites. While beyond the scope of the current work, we believe our approach could be used in conjunction with quantitative trait analyses (QTL) to increase the power for detecting true cis- and trans-acting SNP by interfering with transcription factor binding sites which in turn leads to altered DHS signals in a similar manner as Degner et al. [8].

As measurements from high throughput sequencing platforms become commonplace in molecular biology, there will be an increasing demand for the development of new statistical approaches for these data. A major challenge is that sequencing measurements are rarely in units directly relatable to one another; e.g., DHS measures chromatin accessibility, ChIP-seq measures binding affinity, RNA-seq measures RNA molecule abundance, etc. Our framework provides the initial development of a statistic which captures relationships among these measurements and enables statistical testing of associations among them. Moreover, by exploiting variation across multiple conditions, the sensitivity of our approach should only increase with additional data and sources of variation. Hence, the presented framework can likely be applied to test for associations between appropriate continuous quantitative genomic measurements and gene expression, thereby facilitating a comparable basis for meta-analyses on the interplay of epigenetic features.

4 Materials and Methods

4.1 DHS and gene expression data

The data used in this study were generated through the ENCODE consortium and are publicly available. Established cell lines and primary cells used in this study were procured from commercial or other sources as listed in Table S1. The cells were cultured as per the vendor recommendations, and individual cell growth protocols are available in the UCSC human genome browser. The DHS data are available at the UCSC genome browser by downloading the track IDs listed in Table S2 and the web address shown therein. Normalized probe-level expression data were obtained from the Gene Expression Omnibus (GEO); the accession numbers for all arrays are shown in Table S4. Probes were mapped to genes according to HG18 using bowtie [25] allowing for 2 mismatches and up to 10 maps to the genome, including the best match. Only probe sets for which all probes had a unique best match and fully corresponded to exon boundaries found in RefSeq annotations (HG18) were retained for further analysis. If a RefSeq gene had multiple splice variants, these were aggregated to a meta-gene structure. In the rare event that a gene mapped to currently ambiguous regions (e.g., chr6_random) such regions were not included. To arrive at a gene specific expression value, the mean expression across all probe sets within the exon boundaries of the gene model was

calculated. This yielded expression measures for 19,215 genes on 20 cell lines.

4.2 Statistical methods

The ARS algorithm and statistical analyses were written in the R programming language [33]. The main ARS algorithm, results, GO-analyses, and preprocessed data are available at `http://encode.princeton.edu/`. Complete details of the ARS algorithm, including the null randomization strategy and estimation of the angular penalty, are provided in SI.

A schematic of the method is shown in Fig. 1. We represented the measurements of a single gene by two paired vectors $\mathbf{x} = (x_1, \ldots, x_m)$ for gene expression and $\mathbf{y} = (y_1, \ldots, y_m)$ for DHS volume, where m is the number of cell-types under consideration (here, $m = 20$). To place the two variables on a common scale, each vector was scaled by its maximum observation $\mathbf{x}^s = \frac{\mathbf{x}}{\max\{x_1,\ldots,x_m\}}$ and $\mathbf{y}^s = \frac{\mathbf{y}}{\max\{y_1,\ldots,y_m\}}$ so that all values are now in $[0, 1]$. Each vector was then centered by its median $\text{med}(\mathbf{x}^s)$ and $\text{med}(\mathbf{y}^s)$ to form $\mathbf{x}^* = \mathbf{x}^s - \text{med}(\mathbf{x}^s)$ and $\mathbf{y}^* = \mathbf{y}^s - \text{med}(\mathbf{y}^s)$. Hence the data for a given gene and segment are now centered around the two-dimensional medoid where the center of mass of the data lies at the origin. If there is little variation across the multiple cell-types, all points would cluster around the medoid, while singular cell-types displaying greater variation would be present as distinct outliers (SI, Fig. S9 and Fig. S10). To gauge potential outliers the Euclidian distance $d_i = \sqrt{x_i^{*2} + y_i^{*2}}$ were calculated for every cell type $i = 1, \ldots, m$ to produce the distance vector $\mathbf{d} = (d_1, \ldots, d_m)$. We formed a ratio statistic according to $r_i = \frac{d_i}{\text{med}(\mathbf{d})}$, thereby quantifying the relative distance of each point to the medoid.

While the ratios r_i describe the dispersion of the data, it does not account for any concordance between the measurements. A perfectly concordant relationship between the two measurements would result in points lying along the 45° diagonal identity line. We therefore calculated the angle θ_i for each data point (x_i^*, y_i^*) relative to the unit vector $(1, 0)$ for $i = 1, \ldots, m$, where $0 \leq \theta_i \leq 360$. The angular penalty involves first calculating the smaller of the two angular distances between θ_i and the identity line, denoted as Δ_i. For example, $\Delta_i = |45 - \theta_i|$ for $0 \leq \theta_i < 135$. The angular penalty is calculated as $a_i = \exp(c \times \Delta_i)$, where c is determined empirically to satisfy a correct null distribution (SI). Therefore, the value a_i measures the penalized angular distance of (x_i^*, y_i^*) from the identity line in a symmetric fashion (SI, Fig. S11). The statistic applied to each (x_i^*, y_i^*) pair is then $\text{ARS}_i = a_i \times r_i$, with the gene's overall statistic being the maximum, $\text{ARS}_{\max} = \max(\text{ARS}_1, \text{ARS}_2, \ldots, \text{ARS}_m)$. In addition to calculating these quantities for each gene, we also recorded the ordering of the cell types as determined by their relative ARS_i values.

Inclusion of the angular penalty had a twofold purpose. Firstly, it correctly eliminated points that were outliers in only one dimension, gene expression or DHS alone, and therefore not of interest here since there is no direct relationship between the two measurements. Secondly, penalizing such points acted as a tuning parameter adjusting for the degree of off-diagonal noise in data, and thereby ensured a correct null distribution and p-values (Fig. S12). The specific value of c was determined such that observed null p-values over ($p \geq 0.5$) had a Uniform(0,1) distribution according to a Kolmogorov-Smirnov test (SI). Reassuringly, this lead to nearly identical values for c across all genomic segment sizes of DHS volume considered (SI, Fig. S13).

The scaled data $\mathbf{x}^s$ and $\mathbf{y}^s$ for all genes were aggregated into a single distribution in the unit square $[0, 1] \times [0, 1]$. From this, randomized data sets were created by sampling 20 points that preserves the fact that either one point must lie on $(1, 1)$ or two points lie on $(x, 1)$ and $(1, y)$, respectively. The 20 sampled points are then median centered and the $\text{ARS}_{\max}$ statistic is calculated.

We performed this 100 times to generate 100 sets of null ARS_{max} statistics for every gene (for a total of $100 \times 19,215$ null statistics). A p-value was then formed for each gene by calculating the frequency by which null statistics exceed the observed statistic. The p-values were then utilized to calculate FDR q-values for the genes, as previously described [45]. See SI for full details on this randomization method.

4.3 Selecting local ARS profile peaks for further analysis

We first identified genes called significant at FDR < 0.10 for the ARS analysis performed on the segment size of $\pm$ 200kb about the TSS. We recorded the maximal cell-type for each of these genes (i.e., the cell type yielding the ARS_{max} value), producing a list of significant gene/cell-type pairs. We limited our selection of gene/cell-type pairs to those cell types that were maximal at this threshold for at least 100 genes. For each of these selected gene/cell-type pairs, we scaled its local ARS profile by the maximal value in the $\pm$ 200kb segment about the TSS. All DNA sequences $\pm$ 50bp with scaled local ARS profile value > 0.5 were then selected as "local ARS peaks." Likewise, all DNA sequences $\pm$ 50bp with scaled local ARS profile value < 0.2 were selected as the "negative control set." The local ARS peak set consisted of a total of 38,819 100bp regions, and the negative control set consisted of 156,060 100bp regions.

4.4 TFBS analysis

We took the above local ARS peaks and negative control set, and we eliminated all segments within $\pm$ 10kb of the TSS, reducing the number of local ARS peak segments from 38,819 to 32,063 and negative control segments from 156,060 to 148,423. These were searched with all non-redundant vertebrate positions count matrices in the JASPAR database [31]. The position count matrices were converted to position weight matrices using a uniform background, and a matrix specific thresholding of 0.8 of the maximal log-odds score was used. Significant over- or under-representation was determined by exact binomial tests where the probability was based on the frequency of hits per base pair in the negative control sequences. Effect-size was calculated as $\log_2$ fold-change between number of hits per base pair in the local ARS peaks versus the negative control set.

Web Resource

To provide an interface for the community to utilize the results from this work, the local ARS tracks across any given gene in any of the 20 cell-types can be calculated via our web-service at `http://encode.princeton.edu/`, where all results encompassing the larger DHS regions are also searchable.

Acknowledgments

We thank the Stamatoyannopoulos lab for useful discussions and suggestions, Shane Neph for help with collating gene ontology analyses, Richard Sandstrom for information on experimental details, Michael Hudock for assistance with computations, and Lance Parsons for building the web site. The publicly available ENCODE data utilized in this work were generated by the Stamatoyannopoulos lab. This research was supported in part by NIH grants U54 HG004592 and R01 HG002913.

References

[1] S. Acuto, et. al. (1996) An element upstream from the human delta-globin-encoding gene specifically enhances beta-globin reporter gene expression in murine erythroleukemia cells *em Gene*, 168(2):237–241

[2] E. Ammirati, et al. (2008) Expansion of T-cell receptor zeta dim effector T cells in acute coronary syndromes *Arterioscler Thromb Vasc Biol*, 28(12):2305–2311

[3] Z. Bian, et al. (2009) Cellular repressor of E1A-stimulated genes attenuates cardiac hypertrophy and fibrosis *J Cell Mol Med*, 13(7):1302–1313

[4] A. P. Boyle, et al. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322

[5] M. A. Chapman, at al.(2003) Comparative and functional analyses of LYL1 loci establish marsupial sequences as a model for phylogenetic footprinting. *Genomics*, 81(3):249–259

[6] N. Day, A. Hemmaplardh, R. E. Thurman, J. A. Stamatoyannopoulos, and W. S. Noble. (2007) Unsupervised segmentation of continuous genomic data. *Bioinformatics*, 23(11):1424–1426.

[7] G. Dennis, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, 4(5)

[8] J. F. Degner et al (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 7385(482):390–394.

[9] N. Di-Poï, J. Zákány, and D. Duboule. (2007) Distinct roles and regulations for HoxD genes in metanephric kidney development. *PLoS Genet*, 3(12)

[10] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10:48–48

[11] J. Ernst and M. Kellis. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*, 28(8):817–825

[12] C. E. Eyers, et al. (2005) The phosphorylation of CapZ-interacting protein (CapZIP) by stress-activated protein kinases triggers its dissociation from CapZ. *Biochem J*, 389(Pt 1):127–135.

[13] J. L. Fordham, B. Göttgens, F. McLaughlin, and A. R. Green. (1999) Chromatin structure and transcriptional regulation of the stem cell leukaemia (SCL) gene in mast cells. *Leukemia*, 13(5):750–759

[14] W. C. Forrester, C. Thompson, J. T. Elder, and M. Groudine. (1986) A developmentally stable chromatin structure in the human beta-globin gene cluster. *Proc Natl Acad Sci U S A*, 83(5):1359–1363

[15] P. Fraser, J. Hurst, P. Collis, and F. Grosveld. (1990) DNaseI hypersensitive sites 1, 2 and 3 of the human beta-globin dominant control region direct position-independent expression. *Nucleic Acids Res*, 18(12):3503–3508.

[16] B. Göttgens, et al. (2002) Transcriptional regulation of the stem cell leukemia gene (SCL)–comparative analysis of five vertebrate SCL loci. *Genome Res*, 12(5):749–759.

[17] B. Göttgens, et al. (2001) Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res*, 11(1):87–97.

[18] B. Göttgens et al. (2002) Establishing the transcriptional programme for blood: the SCL stem cell enhancer is regulated by a multiprotein complex containing Ets and GATA factors. *EMBO J*, 21(12):3039–3050.

[19] F. Grosveld, et al. (1990) The dominant control region of the human beta-globin domain. *Ann N Y Acad Sci*, 612:152–159.

[20] E. Hatchwell and J. M. Greally. (2007) The potential role of epigenomic dysregulation in complex human disease. *Trends Genet*, 23(11):588–595

[21] R. D. Hawkins, G. C. Hon, and B. Ren. (2010) Next-generation genomics: an integrative approach. *Nat Rev Genet*, 11(7):476–486.

[22] N. D. Heintzman, et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112.

[23] A. D. Johnson and C. J. O'Donnell. (2009) An open access database of genome-wide association results. *BMC Med Genet*, 10:6–6.

[24] M. Kanehisa.(2002) The KEGG database. *Novartis Found Symp*, 247:91–101.

[25] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3).

[26] J. Lekstrom-Himes and K. G. Xanthopoulos. (1998) Biological role of the CCAAT/enhancer-binding protein family of transcription factors. *J Biol Chem*, 273(44):28545–28548.

[27] J. M. Molete, et al. (2001) Sequences flanking hypersensitive sites of the beta-globin locus control region are required for synergistic enhancement. *Mol Cell Biol*, 21(9):2969–2980.

[28] T. L. Murphy, M. G. Cleveland, P. Kulesza, J. Magram, and K. M. Murphy. (1995) Regulation of interleukin 12 p40 expression through an NF-kappa B half-site. *Mol Cell Biol*, 15(10):5258–5267.

[29] F. Parviz, et al. (2003) Hepatocyte nuclear factor 4alpha controls the development of a hepatic epithelium and liver morphogenesis. *Nat Genet*, 34(3):292–296.

[30] B. S. Pohl and W. Knöchel. (2001) Overexpression of the transcriptional repressor FoxD3 prevents neural crest formation in Xenopus embryos. *Mech Dev*, 103(1-2):93–106.

[31] E. Portales-Casamar, et al. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 38(Database issue):105–110.

[32] W. G. Pyle, G. La Rotta, P. P. de Tombe, M. P. Sumandea, and R. J. Solaro. (2006) Control of cardiac myofilament activation and PKC-betaII signaling through the actin capping protein, CapZ. *J Mol Cell Cardiol*, 41(3):537–543.

[33] R Development Core Team. (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[34] P. J. Sabo, et al. (2004) Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc Natl Acad Sci U S A*, 101(48):16837–16842.

[35] M. Sánchez, et al. (1999) An SCL 3' enhancer targets developing endothelium together with embryonic and adult haematopoietic progenitors. *Development*, 126(17):3891–3904.

[36] V. Sapin, P. Dollé, C. Hindelang, P. Kastner, and P. Chambon. (1997) Defects of the chorioallantoic placenta in mouse rxralpha null fetuses. *Dev Biol*, 191(1):29–41.

[37] V. Sapin, S. J. Ward, S. Bronner, P. Chambon, and P. Dollé. (1997) Differential expression of transcripts encoding retinoid binding proteins and retinoic acid receptors during placentation of the mouse. *Dev Dyn*, 208(2):199–210.

[38] J. S. Satterlee, D. Schübeler, and H. H. Ng. (2010) Tackling the epigenome: challenges and opportunities for collaboration. *Nat Biotechnol*, 28(10):1039–1044.

[39] E. Schneider, et al. (2010) Spatial, temporal and interindividual epigenetic variation of functionally important DNA methylation patterns. *Nucleic Acids Res*, 38(12):3880–3890.

[40] M. Shimotsuma, E. Okamura, H. Matsuzaki, A. Fukamizu, and K. Tanimoto. (2010) DNase I hypersensitivity and epsilon-globin transcriptional enhancement are separable in locus control region (LCR) HS1 mutant human beta-globin YAC transgenic mice. *J Biol Chem*, 285(19):14495–14503.

[41] A. Siepel, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–1050.

[42] A. M. Sinclair, et al. (1999) Distinct 5' SCL enhancers direct transcription to developing brain, spinal cord, and endothelium: neural expression is mediated by GATA factor binding sites. *Dev Biol*, 209(1):128–142.

[43] F. Song, et al. (2005) Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proc Natl Acad Sci U S A*, 102(9):3336–3341.

[44] D. Sproul, N. Gilbert, and W. A. Bickmore. (2005) The role of chromatin structure in regulating the expression of clustered genes. *Nat Rev Genet*, 6(10):775–781.

[45] J. D. Storey and R. Tibshirani.(2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445.

[46] S. Suzuki, et al. (2007) A novel genetic marker for coronary spasm in women from a genome-wide single nucleotide polymorphism analysis. *Pharmacogenet Genomics*, 17(11):919–930.

[47] K. Tanimoto, et al. (2003) Human beta-globin locus control region HS5 contains CTCF- and developmental stage-dependent enhancer-blocking activity in erythroid cells. *Mol Cell Biol*, 23(24):8946–8952.

[48] B. Tolhuis, R. J. Palstra, E. Splinter, F. Grosveld, and W. de Laat. (2002) Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell*, 10(6):1453–1465.

[49] D. Tuan, W. Solomon, Q. Li, and I. M. London. (1985) The "beta-like-globin" gene domain in human erythroid cells. *Proc Natl Acad Sci U S A*, 82(19):6384–6388.

[50] T. Ulinski, et al. (2006) Renal phenotypes related to hepatocyte nuclear factor-1beta (TCF2) mutations in a pediatric cohort. *J Am Soc Nephrol*, 17(2):497–503.

[51] A. H. Wong, I. I. Gottesman, and A. Petronis. (2005) Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Hum Mol Genet*, 14 Spec No 1:11–18.

[52] K. L. Wright, et al (1994) CCAAT box binding protein NF-Y facilitates in vivo recruitment of upstream DNA binding transcription factors. *EMBO J*, 13(17):4042–4053.

[53] H. Xi, et al. (2007) Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet*, 3(8).