Identifying and mapping cell-type-specific chromatin programming of gene expression

Troels T. Marstrand and John D. Storey

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved January 2, 2014 (received for review July 2, 2013)

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 and genome-wide gene-expression levels in 20 diverse human cell types. We show that ~25% of genes show cell-type-specific expression explained by alterations in chromatin structure. We find that distant regions of chromatin structure (e.g., ±200 kb) capture more genes with this relationship than local regions (e.g., ±2.5 kb), 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 uses. This quantitative framework is likely applicable to other settings aimed at relating continuous genomic measurements to gene-expression variation.

Human cells, 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 nonessential genes into heterochromatin structures. Genes that 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 (<10 kb) to the transcription region in just one or very few cell types or experimental conditions (1–3). However, it is also of interest to understand how larger regions (>10 kb) 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 (4, 5), although doing so requires new developments in computational methods (6–8). A collection of landmark papers from the Encyclopedia of DNA Elements (ENCODEx project) were recently published that summarize their most recent efforts to comprehensively understand functional elements in the human genome (e.g., refs. 5, 9, 10). Using ENCODE data, we undertook a well-targeted genome-wide investigation to characterize the relationship between variations in chromatin structure and gene-expression levels across 20 diverse human cell lines (SI Appendix, Table S1). We used data on chromatin structure as ascertainment 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 used to assess the robustness of our method.

Revisiting 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. 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 Appendix and Figs. S1–S5).

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.

Results

Genome-Wide Profiling of Chromatin Accessibility and Gene Expression

We used 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 (11). In principle, accessible “open” chromatin is cleaved

Significance

In order for genes to be expressed in humans, the DNA corresponding to a gene and its regulatory elements must be accessible. It is hypothesized that this accessibility and its effect on gene expression plays a major role in defining the different cell types that make up a human. We have only recently been able to make the measurements necessary to model DNA accessibility and gene-expression variation in multiple human cell types at the genome-wide level. We develop and apply a new quantitative framework for identifying locations in the human genome whose DNA accessibility drives cell-type-specific gene expression.

Author contributions: J.D.S. designed research; T.T.M. and J.D.S. performed research; T.T.M. and J.D.S. contributed new reagent/analytic tools; T.T.M. analyzed data; and T.T.M. and J.D.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. Free access online through the PNAS open access option.

To whom correspondence should be addressed. E-mail: jstorey@princeton.edu.

This article contains supporting information online at www.pnas.org/cgi/suppl/doi:10.1073/pnas.1312523111/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1312523111

PNAS | Published online January 27, 2014 | E645–E654
RefSeq Genes

The isolated relationship seen in HepG2 and assess whether this relationship is statistically significant. Traditional measures of linear correlation are not suitable for HepG2 displays both chromatin accessibility and active gene expression, whereas HeLa displays only chromatin accessibility. The goal is to quantitatively capture about the gene

interval [0,1] and the data are centered about the origin according to the 2D medoid. For the about the gene

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° line, and it is then multiplied times its respective relative distance to give and overall score for each cell line. The maximum score is

exist for many genes and cell lines. The proposed ARS is robust to HeLa because the measure is based on angular placement and the median distance to the medoid of the data (dashed circle). (Fig. 1.

Overview of data and proposed approach. (A) Gene-expression measurements for 20 cell lines on an example gene, HNF4A. (B) 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., ±100 kb) about the gene’s 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 2D 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 accessibility. 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 because the measure is based on angular placement and the median distance to the medoid of the data (dashed circle). (D) The ARS 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° 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 we introduce can pinpoint DHS “peaks” contributing the most to the detected association. See main text for details on the proposed methods.

by the nonspecific endonuclease DNaseI, and the cleaved fragments are sequenced to provide a high-resolution, genome-wide map of DHS for every cell type (SI Appendix, 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 used genome-wide gene-expression measurements in each cell line from Affymetrix exon arrays (SI Appendix, Table S4). A total of 19,215 genes were analyzed after preprocessing (Materials and Methods).

With these quantifications, we sought to characterize the relationship between chromatin accessibility and gene expression in a cell-type-specific manner, summarized in Figs. 1 and 2. What we mean by “cell-type specific” is that, when pairing gene-expression and chromatin-accessibility measures according to cell type, we observe a relationship between the two measures, typically for one
or very few cell types. 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 Appendix], Fig. S6) to obtain a cell-type-specific regional DHS volume. We selected a range of segments that were likely to encompass all proximal [transcriptional start site (TSS) ± 2.5 kb] and most distal regulatory elements (TSS ± 50, ± 100, ± 150, ± 200, and ± 100 kb minus proximal 2.5 kb, and ± 200 kb minus proximal 2.5 kb). In addition, 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 Appendix], Fig. S7). Alternative representations of DHS signal (8, 12) could be used 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.

**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 Appendix and Figs. S1–S5). 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 to obtain parametric P values, but the normal assumption does not hold for these data (SI Appendix, Fig. S2). Second, this correlation statistic is unstable when there are outliers, even when using permutation-based P values, demonstrated directly on these data (SI Appendix, Figs. S1 and S3). 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. 3 (see also SI Appendix, Figs. S4 and S5).

For example, at a false discovery rate (FDR) ≤ 0.05, the proposed method identifies 2,538 genes with a cell-type-specific DHS and gene-expression relationship, whereas the Spearman statistic identifies only 286 (Fig. 3 and SI Appendix, Figs. S4 and S5).

The 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 Figs. 1 and 2, Materials and Methods, and SI Appendix). 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. 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 farther away from the line of identity according to a data-derived exponential function. These two quantities are then multiplied to form an ARS value for each cell type (i = 1, 2, ..., 20), and the maximal value $ARS_{max}$ 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 $ARS_{max}$, we constructed a null distribution based on randomization of the observed experimental data (Materials and Methods, SI Appendix and Fig. S8). $ARS_{max}$ values obtained from the randomized data were used as a basis for determining a P value of the observed $ARS_{max}$ for each gene. FDR-based statistical significance and the proportion of genes with a true chromatin accessibility and gene-expression relationship were estimated from the P values (ref. 13; Fig. 3B).

We estimate that ~25% of genes show concordance between chromatin accessibility and gene-expression variation in a cell-type-specific manner. Although 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. 3B, column 2), indicating that distal chromatin programming effects are more widespread in a genome-wide sense (14). On the other hand, using the proximal DHS volume we observe a greater empirical effect size compared with the distal DHS volumes (Fig. 3B, columns 3–5).

This observation is explained by the aggregation of genes significant for the same cell type along the genome (15). Testing whether one or more significant genes within a ±100-kb 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 $P < 2.2e^{-16}$; SI Appendix and Fig. S14). 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. 3A).

**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, we calculated the number of statistically significant genes at various FDR cutoffs. Although the 2.5-kb window shows more significant genes at the stringent FDR cutoffs, indicating a larger effect size, the overall percentage of genes showing a relationship is notably lower than the more distal DHS volumes. Compared with Spearman correlation, ARS is more powerful at detecting these associations (see SI Appendix for further details). (C) The relative ARS values across all cell types for significant genes in the ±100 kb region versus the analogous components for Spearman correlation (the cross-product terms that sum to form the overall correlation) are given in the color-coded cross-products for each cell type, with 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 (SI Appendix, Fig. S13).
gene expression, respectively, we calculated the fraction of predictions preserved in all four-way comparisons (SI Appendix). We found that between 86 and 91% of significant genes (FDR < 0.05) were identical (SI Appendix, Fig. S15).

**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 (16). The method computes enrichment within the process and function components of GO categories and assigns a numerical significance 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 (17, 18) were likewise enriched in a cell-type-specific manner. For example, HepG2 (a hepatocellular carcinoma cell line) showed significant enrichment for genes within the bile-acid synthesis and drug metabolism, and HL60 (a human promyelocytic leukemia cell line) showed significant enrichment within the hematopoietic cell lineage.

Furthermore, all genes detected within each cell type at FDR < 0.05 (±100 kb DHS volume) were analyzed through the use of Ingenuity Pathways Analysis (Ingenuity Systems, www.ingenuity.com). For all but 3 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 Appendix, Fig. S13). 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 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 in a few tissues, which prompted us to investigate these instances further.

Among genes with a statistically significant ARSmax statistic, additional inspection of the remaining ARS were explored for detection of possible substructures. We calculated relative ARS values within each gene dividing all ARS by ARSmax. 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 (SI Appendix, Fig. S16).

**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 ARSmax 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 nonoverlapping consecutive regions at a high-resolution 60-base-pair windows. The ARS was calculated for each 60-bp window, which can then be plotted over the entire region used in identifying the gene as statistically significant. (The original sequence alignments for the DHS data were mapped into 20-bp windows spanning the entire genome, so we chose windows of size 60 bp to smooth the local DHS measure across neighboring sites.) For example, for a gene significant with respect to a ±200-kb DHS volume, we calculated ~6,700 local ARS values 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 logarithm of odds (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 Appendix).

We first selected the subset of local ARS profile “peaks” by thresholding the local ARS profiles in a principled manner (Materials and 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.

**Positional Bias of Putative Regulatory DHS.** Because the overall statistical significance increases when calculating DHS volume over more distant regions up to 200 kb (Fig. 3), we investigated the positional bias of local ARS peaks in a cell-type-specific manner. Fig. 4A 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 Appendix, Fig. S17). This points to the existence of cell-type-specific differences in the base-pair distance of regulatory DHS to TSS.

**Sequence Analysis of Peaks in Local ARS Profiles.** We next sought to characterize the functional significance of sequences corresponding to local ARS peaks. Because a general indicator of functionality is conservation, we extracted the conservation track values (percentCons4wayPrimate, hg18) (19) corresponding to the local ARS peaks and to the negative control set (Materials and 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 < 2.2e-16; SI Appendix, Fig. S18), 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 whether known cell-type-specific transcription-factors binding sites (TFBSs) are overrepresented in the local ARS peaks relative to the negative control set (Materials and Methods). Because 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 200 kb from the TSS.

We used the JASPAR database (20) to identify TFBS that are differentially represented in the local ARS peaks relative to the negative control set (Materials and Methods). The over- and underrepresented TFBS show distinct cell-type-specific patterns (21) and provide a rich insight into cell-type-specific gene regulation (Fig. 4B), several of which are listed here:
Among the hepatocyte nuclear factors we found HNF1B (transcription factor 2, TCF-2) and HNF4A to have significant chromatin gene expression concordance in HRCE (a human renal cortical epithelial cell line) and HepG2, respectively (SI Appendix, Figs. S19 and S20). Furthermore we found the local ARS profiles in the respective tissues to display a marked overrepresentation of the factor in question, HNF1B in HRCE and HNF4A in HepG2. Mutations in HNF1B have been associated with a broad range of renal diseases (22), and HNF4A is essential for hepatocyte differentiation and morphology (23).

H7ESC (a human embryonic stem cell line) was found to show overrepresentation of SOX2 and Pou5f1 (Oct4) to the proximal promoter region (24). The ubiquitous CCAAT-binding-factor family is linked to cellular differentiation in a variety of tissues (25).

Retinoic X receptors (RXR)—retinoic acid receptors (RAR) were found in human amniotic epithelial cells (HAEpiC). The coexpression of RAR and RXR (26) is essential for proper placental development, and RXR null mouse mutants are lethal after 10 d due to placental defects (27).

Forhead binding sites were found to be primarily underrepresented, specifically Foxd3 was under-represented in, among others, the leukemic cell types. Silencing of Foxd3 by aberrant chromatin modification has been implicated in leukemogenesis (26). Overexpression of Foxd3 prevents neural crest formation (28). It is interesting to note that binding sites for the factor were underrepresented in SKNSH, a neuroblastoma derived from neural crest cells.
NF-xB was found over-represented in TH1, where it promotes the expression of, among others, interleukin 12 (IL-12) essential for TH1 development (29).

The differentially represented TFBSs were distributed largely distal. For all cell types, from 68 to 79% were located more than ±50 kb away from the TSS. We repeated the analysis with only the proximal regions (±10 kb from the TSS), and we found that several important known cell-type-specific motifs were no longer detected (SI Appendix, Figs. S21).

To validate our in silico TFBS predictions with an independent data source, we downloaded the UCSC genome tracks for POU5F1 (Oct-4) in HESC and NFYA in K562 (http://genome.ucsc.edu/ENCODE/). (Note that HESC is not completely identical to H7ESC, but both are embryonic stem cells.) We next calculated the overlap between the POU5F1 binding sites determined from the ChIP-seq data and the positive local ARS peaks that we identified in H7ESC. Of the 1,614 peaks used for TFBS analysis in H7ESC, 200 overlapped with the POU5F1 ChIP-seq calls; of the remaining 57,205 positive local ARS peaks in other cell types, where POU5F1 was not predicted to be enriched, only 119 overlapped. Similarly for K562, of the 4,861 positive local ARS peaks 439 overlapped ChIP-seq calls for NFYA. For cell types where NFYA was not detected [HRCE, SNKSH (a neuroblastoma cell line), and HeLa], of 8,176 positive local ARS peaks, only 72 had overlapping ChIP-seq calls. Hence, we obtain an enrichment of 39-fold for POU5F1 ($P < 10^{-10}$) and 10-fold for NFYA ($P < 10^{-10}$), respectively. Overall, this supports the method’s ability to map relevant regulatory regions on a fine scale.

Single-Nucleotide Polymorphisms. The local ARS peaks were also investigated with respect to SNPs found to be significant in genome-wide association studies (GWAS). The database compiled by Johnson and O’Donnell (30) containing a total of 52,622 unique SNPs associated with disease phenotypes (56,411 total associations) was mapped against the genomic coordinates of the local ARS peaks. We found that only 42 SNPs fell within local ARS peaks, and the expected value under completely random placement of SNPs is 68. A statistical test indicates that there is significant underrepresentation of GWAS SNPs (two-sided exact test $P < 0.0001$). This is likely linked to the conserved state of the sequences corresponding to local ARS peaks (shown above) and that these regions harbor cell-type-specific regulatory elements, as indicated by the above TFBS analysis.

This does not preclude GWAS SNPs from appearing in local ARS peaks. Indeed, such an occurrence may be particularly noteworthy. An interesting example is that of rs1890131 (31), which lies in a local ARS peak driving the ARS-based statistical significance of three different genes (SI Appendix, Fig. S22): CREG1 in HL60, RCSD1 in TH1, and CD247 in TH1 (SI Appendix, Fig. S23). The SNP rs1890131 has been associated with coronary spasms via CREG1, where cellular repressor of E1A-stimulated genes (CREG) attenuates cardiac hypertrophy (32) by blocking mitogen-activated protein kinases/extracellular signal-regulated kinase (MEK-ERK)1/2-dependent signaling. RCSD1 (also termed capZ-interacting protein) is suggested to regulate the ability of CapZ to remodel actin filament assembly in vivo (33), and thereby affects cardiac isometric tension generation (34). Last, CD247 (TCR) is a master regulator of adaptive immune responses, where loss of TCR- chain expression may contribute to the inflammatory process that triggers coronary instability by accumulating in coronary plaques (35). The SNP rs1890131 is centered in a hub of transcription-factor binding sites as detected by the ENCODE transcription-factor track (SI Appendix, Fig. S24). The local ARS peak containing this SNP contributed significantly to the association detected for all three genes, which are distal and in different orientation from the site of the SNP (SI Appendix, Fig. S22).

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 Appendix, Figs. S25–S37. 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; refs. 36–38; Fig. 5) upstream of HBE1 (ε-globin) on the short arm of chromosome 119, Fig. S23). The SNP rs1890131 has been associated with κ- chain expression may contribute to 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 Appendix, Figs. S25–S37. 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; refs. 36–38; Fig. 5) upstream of HBE1 (ε-globin) on the short arm of chromosome.

Fig. 5. Mapping putative regulatory DHS with local ARS profiles at two loci. (A) The 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 elements. The local ARS profile shows no peak in HL60 despite the existence of a hypersensitive site when considering DNaseI profile alone. The full data and local ARS profiles for all 20 cell lines and both genes are displayed in SI Appendix, Figs. S25–S29. (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 no peak in HL60 despite the existence of a hypersensitive site when considering DNaseI profile alone. The full data and local ARS profiles for all 20 cell lines and both genes are displayed in SI Appendix, Figs. S31–S33.
All five sites were present in cell line K562 according to our DHS data (see SI Appendix, Figs. S25–S29 for complete data across all 20 cell types, and SI Appendix, Fig. S30 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. 5A).

In the case of HBE1, we observed local ARS peaks for HS1 at −6.1 kb and to a lesser extent HS3 and HS4 (−14.7 and −18 kb, respectively). For HBB we observed similar local ARS profiles for HS1, HS3, and HS4, and smaller local ARS values for HS2 (−10.9 kb). It has previously been shown that HS1 is a stable chromatin structure (37) throughout development and essential for HBE1 expression (39) due to a GATA-1 binding site, and HS2, 3, and 4 show synergistic enhancement of expression of HBB by formation of the LCR holocomplex (40–42). Finally, the element upstream of HBD has also been reported to specifically enhance transcription of HBB (43). Although HS5 is present in the DHS data for K562, similar open chromatin structures were detected in other tissues. HS5 (−21 kb) 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 (44).

T-cell acute lymphocytic leukemia protein 1 (TAL1) encodes a basic helix-loop-helix protein, which is essential for the formation of all hematopoietic lineages (SI Appendix, Figs. S31–S33 for all data across the 20 cell types). Previous studies using chromatin structure, comparative sequence analysis, transcription assays (45, 46), and transgenic mice (47–51) 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 ±2.5 to ±200 kb) with the most significant ARS max occurring for ±50 kb. Further investigation by the local ARS profile (Fig. 5B) showed that although proximal regulatory sites were correctly identified, the most dominant signal is by far confined to the +40 enhancer region and is of order of magnitude greater than other signals. Although the TAL1 +40 region is downstream of PDZKIPI, 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 erythroblasts, 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 Appendix, Fig. S34).
- Part of the HOX-cluster crucial for kidney development in mammals (HOXD8, HOXD4, and HOXD3) showed identical local ARS profiles (SI Appendix, Fig. S35), and all were significant genes in HRCE (52).
- 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 human skeletal muscle cells (SKMC) and has high relative ARS values in HAEpiC and BJ (skin fibroblast) (SI Appendix, Fig. S36). Further fine-scale investigation showed a solid overlap in the local ARS profiles (SI Appendix, Fig. S37).

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

Web Resource
To provide an interface for the community to use the results from this work, the local ARS profiles 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.

Discussion
As the epigenome in multicellular organisms is a dynamic entity whose variation leads to reprogramming of gene expression (53), it is a likely candidate in the etiology of disease complementary to that of mutations in DNA (54, 55). 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 avenues 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 short list 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 200 kb from transcription start sites. Although beyond the scope of the current work, we believe our approach could be used in conjunction with quantitative trait analyses to increase the power for detecting true cis- and trans-acting SNPs by interfering with transcription factor binding sites, which in turn leads to altered DHS signals in a similar manner as Degner et al. (56).

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 that 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 metaanalyses on the interplay of epigenetic features.

Materials and Methods
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 SI Appendix, Table S1. The cells were cultured as per the vendor recommendations, and individual cell growth protocols are available in the University of California, Santa Cruz (UCSC) human genome browser. The DHS data are available at the UCSC genome browser by downloading the track IDs listed in SI Appendix, Table S2 and the web address shown therein. Normalized probe-level expression data were obtained from the Gene Expression Omnibus; the accession numbers for all arrays are shown in SI Appendix, Table S4. Probes were mapped to genes according to HG18 using bowtie (57) allowing for two mismatches and up to 10 maps to the genome, including the best match. Only probe sets for which all probes had the best match and fully corresponded to exon boundaries were retained. In RefSeq annotations (HG18) were retained for further analysis. If a RefSeq gene had multiple splice variants, these were aggregated to a metagene 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.

Statistical Methods. The ARS algorithm and statistical analyses were written in the R programming language (58). The main ARS algorithm, results, GO analyses, and preprocessing 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 Appendix.

A schematic of the method is shown in Fig. 1. We represented the measurements of a single gene by two paired vectors \( x = (x_1, x_2, \ldots, x_m) \) for gene expression and \( y = (y_1, y_2, \ldots, y_m) \) 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 \( x^* = \frac{x}{x_{\text{max}}} \) and \( y^* = \frac{y}{y_{\text{max}}} \), so that all values are now in \([0, 1]\). Each vector was then centered by its median \( \text{med}(x) \) and \( \text{med}(y) \) to form \( x^* = x - \text{med}(x) \) and \( y^* = y - \text{med}(y) \). Hence the data for a given gene and segment are now centered around the 2D 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, and singular cell types displaying greater variation would be present as distinct outliers (SI Appendix, Figs. S8 and S9). To gauge potential outliers the Euclidean distances \( d_i = \sqrt{x_i^2 + y_i^2} \) were calculated for every cell type \( i = 1, \ldots, m \) to produce the distance vector \( d = (d_1, \ldots, d_m) \). We formed a ratio statistic according to \( r = \frac{d}{d_{\text{med}}} \), thereby quantifying the relative distance of each point to the medoid.

Although the ratios \( r \) 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 \( \theta_i \) 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 \( \theta_i \) and the identity line, denoted as \( \Delta_i \). For example, \( \Delta_i = 45° - \theta_i \) for \( 0 \leq \theta_i \leq 135 \). The angular penalty is calculated as \( a = \exp(c \times \Delta_i) \), where \( c \leq 0 \) and is determined empirically to satisfy a correct null distribution (SI Appendix). Therefore, the value \( a \) measures the penalized angular distance of \((x_i, y_i)\) from the identity line in a symmetric fashion (SI Appendix, Fig. S10). The statistic applied to each \((x_i, y_i)\) pair is then \( \bar{ARS} = a \times r_i \), with the gene’s overall statistic being the maximum, \( \bar{ARS}_{\text{max}} = \max(\bar{ARS}_1, \bar{ARS}_2, \ldots, \bar{ARS}_n) \). In addition to calculating these quantities for each gene, we also recorded the ordering of the cell types as determined by their relative \( \bar{ARS} \) values.

Inclusion of the angular penalty had a twofold purpose. First, it correctly eliminated points representing outliers in only one dimension, gene expression or DHS alone, and therefore not of interest here because there is no direct relationship between the two measurements. Second, penalizing such points acted as a tuning parameter adjusting for the degree of off-diagonal noise in the data, and thereby ensured a correct null distribution and \( P \) values (SI Appendix, Fig. S11). The specific value of \( c \) was determined such that ob-

erved null \( P \) values over \((p > 0.5)\) had a Uniform(0,1) distribution according to a Kolmogorov–Smirnov test (SI Appendix). Reassuringly, this lead to nearly identical values for across all genomic segment sizes of DHS volume considered (SI Appendix, Fig. S12).

The scaled data \( x^* \) and \( y^* \) 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 \((1,0)\) and \((0,1)\), respectively. The 20 sampled points are then median centered and the \( \bar{ARS}_{\text{max}} \) statistic is calculated. We performed this 100 times to generate 100 sets of null \( \bar{ARS}_{\text{max}} \) 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 used to calculate FDR \( q \) values for the genes, as previously described (13). See SI Appendix for full details on this randomization method.

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 \( \pm 200 \) kb about the TSS. We recorded the maximal cell type for each of these genes (i.e., the cell type yielding the \( \bar{ARS}_{\text{max}} \) 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 \( \pm 200 \) kb segment about the TSS. All DNA sequences \( \pm 50 \) bp with scaled local ARS profile value > 0.5 were then selected as local ARS peaks. Likewise, all DNA sequences \( \pm 50 \) bp 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 38,819 100-bp regions, and the negative control set consisted of 156,060 100-bp regions.

TFBS Analysis. We took the above local ARS peaks and negative control set, and we eliminated all segments within \( \pm 10 \) kb 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 nonredundant vertebrate positions count matrices in the JASPAR database (20). 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 logfold change between number of hits per base pair in the local ARS peaks versus the negative control set.

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 Hudson for assistance with computations, and Lance Parsons for help with the website. This work was supported in part by National Institutes of Health Grants US4 HG004592 and RO1 HG002913.

Marstrand and Storey


SUPPLEMENTARY INFORMATION:
Identifying and mapping cell-type-specific chromatin programming of gene expression

Troels T. Marstrand\textsuperscript{a} and John D. Storey\textsuperscript{a,b,1}

\textsuperscript{a}Lewis-Sigler Institute for Integrative Genomics
\textsuperscript{b}Department of Molecular Biology
Princeton University, Princeton, NJ 08544, USA.
\textsuperscript{1}Corresponding author: jstorey@princeton.edu

Contents

Supplementary Text \hspace{1cm} 7

Supplementary Tables \hspace{1cm} 12

Supplementary Figures \hspace{1cm} 17

List of Figures

S1 Four examples of correlation versus angle ratio statistics (ARS).

Data points are represented by black dots and the outlier (i.e., most distant point to the origin) is the red point. The dashed line is the 45 degree identity line and the solid line is a standard regression line. The ARS and Pearson p-values are based on permuted data. To calculate ARS, the data, DHS and gene expression, were scaled by their maximum observations respectively and median centered. The distance and angle of the “outlier” relative to all other points determines the final test-statistic. ARS show robust detection of outliers while standard Pearson correlation may give spurious signals as data is not bivariate Normal. The genes under consideration are (A) \textit{A2M}, (B) \textit{AADAT}, (C) \textit{ACSL3}, (D) \textit{ABCB6}.

\hspace{1cm} 17
Kolomogorov-Smirnov p-values testing whether gene expression and DHS data are Normal. All expression data, and proximal (± 2.5kb) DHS were centered and scaled on a per gene basis, and tested against Normal using a Kolmogorov-Smirnov test. The histograms show the distribution of the resulting p-values in both cases. If the data were Normal, a Uniform distribution would be present. In testing the resulting p-values against Uniform, using a Kolomogorov-Smirnov test we obtain p-values < 2.2e-16 for both cases demonstrating that data cannot be assumed to be Normal.

Comparison of p-values obtained from randomized null data with ARS and Pearson correlation. Using the randomization algorithm on the data 100 times, we calculated permutation based p-values for both AR-Statistics and Pearson correlations. ARS generates well-behaved p-values (e.g., p-values > 0.5 are Uniform), while Pearson correlation results in a bimodal histogram since the Pearson correlation statistic is not designed for the signal in these data.

Comparison of statistical significance from ARS and rank-based Spearman correlation. Using the proximal ± 2.5kb DHS data, ARS shows superior performance relative to the rank based Spearman correlation. At given FDR thresholds, the ARS method identifies many more significant genes.

Comparison of statistical significance from ARS and rank-based Spearman correlation. Using the proximal ± 200kb DHS data, ARS shows superior performance relative to the rank based Spearman correlation. At given FDR thresholds, the ARS method identifies many more significant genes.

Genomic segments for DHS volume. Integrated DHS measurements across increasingly larger genomic segments relative to the TSS of the gene of interest were used to quantify gene-specific DHS volume. All data within the segment was used irrespectively of adjacent genes etc.

Chromosome arm scaling factors. To account for copy number variation difference in read density, etc., each DHS volume was scaled such that the total DHS volume per chromosomal arm was constant across cell-types. The above show the scaling parameter for each cell-type by chromosome.

Distribution of scaled data. By scaling the DHS volume and gene expression by their per gene maximum observations and median centered, it is evident that most data points inhabit a fan-like pattern which forks away from the diagonal. Points away from the diagonal were penalized by an angular decay rate. Colors indicate log₂ of the bin counts, and data were counted on a 20 × 20 grid.
Distribution of scaled and median centered data. Scaled and median centered data for significant genes, non-significant genes, and their difference, where the color indicates the spatial distribution of points. Here, the definition of significance is determined by FDR < 0.05. All x-axes display scaled and median centered expression, all y-axes display scaled and median centered DHS volume. Significant genes tend to have maximum values in both dimensions, and inhabit the upper right corner at (1,1).

Angular penalty. A plot of the angular penalty applied to the scaled and center data. The x-axis and y-axis correspond to expression and DHS volume, and the z-axis to the angular penalty applied to any data point at the corresponding \((x, y)\) location. It can be seen that points falling along the 45° identity line are not penalized, and that the farther a point is from the line the more it is penalized.

The influence of \(c\) on the p-value histograms. In cases where the angular distribution is not correctly captured by \(c\), the estimated p-values would either be conservative or anti-conservative in the right tail as seen above. In the above example, a value of \(c = -0.05\) gave Uniform null p-values in the interval [0.5,1] according to a Kolomogorov-Smirnov test. For the presented analyses a range of values of \(c\) was tested (see Figure S12), if multiple values of \(c\) gave rise to Uniform null p-values the value of \(c\) with the largest p-value was chosen. Due to differences in the distribution of the DHS values, values of \(c\) were slightly different depending on the genomic segment investigated.

Kolomogorov-Smirnov tests to determine \(c\) in the angular penalty. For each value of \(c\), a Kolomogorov-Smirnov test was performed to test for Uniform null p-values defined as p-values in the interval [0.5,1]. A non-significant p-value indicates a valid choice for \(c\). It can be seen that all DHS segment sizes yielded similar valid \(c\) values. If multiple values of \(c\) gave non-significant p-values, the value of \(c\) yielding the largest p-value was chosen.

Ingenuity enrichments for biological functions. Significant genes within the ±100kb region found with ARS and Spearman correlation were tested for enrichment of genes with particular biological function (similar to GO-term analysis) using Ingenuity Pathways Analysis (Ingenuity Systems®, www.ingenuity.com).

In general, genes called significant with ARS show a good correspondence between cell-type and the functions found enriched. Indicating that truly cell-type specific genes were indeed detected. This pattern is not as clearly seen using Spearman correlation, likely due to the fact that using correlation on this type of data leads to many false predictions.
Aggregation of cell-type specific genes. Four different tests were done to test for aggregation of significant genes according to genomic location and maximal cell type. Here, the distance between significant genes are plotted against the distance between non-significant genes. The number of null genes were chosen to match the number of significant genes from that cell-type, such that the effect is not primarily driven by cell-type. A clear and significant trend show that cell-type significant genes aggregate along the genome.

Assessment of reproducibility. The p-values corresponding to genes significant (FDR < 0.05) from one DHS and gene expression replicate are plotted for the complementary analysis. It can be seen that small p-values from one data set yield small p-values from the replicated data set, indicating a high level of reproducibility. The fraction recaptured at FDR<0.05 is listed below each histogram.

Relative ARS scores. For each gene we can scale the ARS values by ARS_max in order to distinguish between significant genes with an “on-off” effect and significant genes which show a more gradual trend (i.e., more than a single cell-type may act as an outlier compared to the remaining cell-types). Significance is here determined as FDR < 0.05. In the color image cell-types are sorted, column-wise, by decreasing scaled ARS_i, the image is further divided into three blocks, representing the three major distributions of ARS_i observed. (A) Significant genes with a singular outlier, i.e. “on-off”, (B) significant genes with a gradual trend, and finally (C) non-significant genes with no distinct outliers as can be seen from the high relative ARS_i values extending across multiple cell-types. Right of each sub-panel is a 2D plot of the scaled and centered gene expression by DHS volume showing a representative gene from the given distribution: IFNG, LOXL2, and ABCG4 respectively.

Positional densities of local ARS peaks by cell-type. The positional bias of cell-type specific local ARS peaks can be seen in the upper panel. For example, HL60 shows a more proximal signal relative to that of HAEpiC. A clear difference in the amount of distal regulation is seen across the different cell-types. The lower panel shows the same analysis applied to the data where the cell types are randomly shuffled among the cases considered in the upper panel.

Conservation of sequences corresponding to local ARS peaks. Phastcons scores within local ARS peaks and a negative control set were recorded and displayed in the quantile-quantile plot. The local ARS peaks show a marked and significantly (Kolmogorov-Smirnov p-value < 2.2e-16) higher conservation relative to the negative control set, indicating that local ARS peaks are strong candidates for functional regions.

HNF1B as chromatin / gene expression concordant in HRCE. While genes significant in HRCE have a clear over-representation of the HNF1B motif, the gene itself is also in chromatin / gene expression concordance.
**S20** HNF4A as chromatin / gene expression concordant in HepG2. While genes significant in HepG2 have a clear over-representation of the HNF4A motif, the gene itself is also in chromatin / gene expression concordance.

**S21** Over- and under-representation of TFBS within cell-type specific proximal (±10kb) promoters. Using only the proximal (±10kb) regions for the local ARS peak TFBS analysis, known cell-specific interactions disappear, e.g., SOX2 and POU5F1 (Oct4) for H7ESC.

**S22** Location of genes relative to the SNP rs1890131 and corresponding local ARS profiles. All three local ARS profiles for the genes CREG1 in HL60, RCSI1 in TH1, and CD247 in TH1, show specific and high association to the site of rs1890131, indicated by the salmon colored bar.

**S23** Plots of gene expression and DHS volumes for the three genes related to rs1890131. Concordance between the entire ±100kb DHS segment and gene expression for the three genes for which rs1890131 falls within a local ARS peak. For RCSI1, HL60 is a clear secondary outlier.

**S24** The SNP rs1890131 in a transcription factor hub. The SNP is centered in a hub of transcription factor binding sites as detected by the ENCODE transcription factor ChIP-seq track. This site is strongly associated in the local ARS profiles with three adjacent genes: CREG1 in HL60, RCSI1 in TH1, and CD247 in TH1.

**S25** Raw DHS fragments for HBB and HBE1. The DHS values within a region of interest for HBB and HBE1 across all cell-types are highlighted by salmon colored bars. These regions correspond to known hypersensitive regions. Interestingly, one region is present in almost all cell-types, whereas, others are unique to K562.

**S26** Scaled and centered gene expression and DHS volume for HBB. The gene expression and DHS volume [0-1] scaled and medoid centered for all 20 cell-types, across the entire ±100kb DHS segment. This show HBB as a clear and singular outlier for K562.

**S27** ARS profiles for HBB. Local ARS profiles across all 20 cell-types, where the salmon colored bars indicate ARS peaks. The 6th hypersensitive site (from the left) is not seen as an ARS peak, due to its presence in multiple cell-types. The hypersensitive site present in the 3rd bar (from the left) appears as an ARS peak, despite being present in all cell-types due to its high local value.

**S28** Scaled and centered gene expression and DHS volume for HBE1. The gene expression and DHS volume [0-1] scaled and medoid centered for all 20 cell-types, across the entire ±100kb DHS segment. HBE1 is a clear and singular outlier for K562.

**S29** ARS profiles for HBE1. Local ARS profiles across all 20 cell-types, where the salmon colored bars show that the local ARS profile is markedly different from HBB, indicating different regulatory importance of the hypersensitive regions for HBB relative to HBE1.
ARS profiles for entire β-globin locus. Local ARS profiles for all genes for which we had expression data in the β-globin control locus region. The hypersensitive regions are utilized differently by the genes, indicating a difference in regulatory regime. ARS peaks are highlighted by salmon colored bars. The third ARS peak from the left shows significant variation across the genes, despite being chromatin / gene expression concordant in the same cell-type.

Raw data DHS fragment counts for TAL1. The DHS values within a region of interest for TAL1 across all cell-types, where the salmon colored bar indicates the area detected as a local ARS peak for TAL1.

Scaled and centered gene expression and DHS volume for TAL1. The gene expression and DHS volume [0-1] scaled and medoid centered for all 20 cell-types, across the entire ± 100kb DHS segment. TAL1 is an outlier for K562, and a secondary outlier for CACO2.

ARS profiles for TAL1. Local ARS profiles across all 20 cell-types, where the salmon colored bars indicate the ARS peaks. Interestingly, a clear hypersensitive site from Figure S31 does not yield a local ARS peak as it is present in all tissues, and therefore not in cell-type specific concordance with the gene expression values.

Distinct local ARS profiles for adjacent genes. Adjacent significant genes show largely distinct local ARS profiles related to the specific cell-type, indicative of separate regulatory programs. Salmon colored bars show ARS peaks for CCR2 in HL60 and teal colored bars indicate ARS peaks for CCR5 in TH1. This example demonstrates the ability of the ARS approach to detect differences in chromatin / gene expression concordance between adjacent genes.

Shared ARS profile within a single cell-type. Neighboring HOXD genes (HOXD8, HOXD4, HOXD3) all significant in HRCE show identical usage of local hypersensitive sites indicative of shared regulatory mechanisms. Salmon colored bars indicate local ARS peaks, which show similar ranking in magnitude within each gene, but differences in total magnitude due to differences in gene expression of the individual genes.

Sub-maximal outliers for LOXL2. Visualization of the sub-maximal outliers BJ and HAEpiC with similar local ARS profiles as SKMC shown below in Figure S37. While statistical significance is based on ARS max., secondary outliers can be seen for other cell-types, indicating shared chromatin / gene expression concordance across cell-types.

LOXL2 share local ARS profiles across cell-types. While ARS max. identified SKMC as the significant outlier, HAEpiC and BJ show large relative local ARS values. Closer inspection shows overlapping utilization of local hypersensitive sites across all three cell-types indicative of possible shared regulatory mechanisms (salmon colored bars). However, individual sites are also detected (teal bars) only shared among SKMC and HAEpiC.
Supplementary Text

ARS versus standard correlation measures. While the standard Pearson correlation statistic is capable of capturing some instances of cell-type specific outliers that are also detected by our proposed ARS method, the operating characteristics of the correlation statistic are unreliable when the data are not Normal and there are outliers. In comparing a test of non-zero correlation to our method, we predicted a total of 1566 significant genes using the parametric correlation test (\texttt{cor.test} in R) compared to 2538 with our proposed method at FDR < 0.05. However, upon further investigation it became clear that the correlation test identifies cases that are not of biological interest and it has unreliable behavior due to outliers (which are of biological interest). Furthermore, the data are not Normal distributed, which means that the p-values produced by the function \texttt{cor.test} are not trustworthy.

To illustrate the point that the Pearson statistic identifies cases that are not of biological interest, Figure S1 shows instances where the Pearson correlation statistic and the ARS either agree or differ. The top left panel shows an example where both ARS\textsubscript{max} and correlation produce significant p-values. The bottom left shows an example where correlation produces a biological false positive. The top and bottom right are more or less indistinguishable, but the correlation test gives widely different results while our ARS remains stable and rightfully calls both genes as having a distinct outlier.

Furthermore the parametric p-values generated from the standard Pearson correlation test are not valid when data for a given gene are not bivariate Normal, as is the case with the data considered here. We investigated the Normal assumption by testing the data, DHS and gene expression for each gene, against the Normal by using a Kolomogorov-Smirnov test. The \( \sim 21,000 \) resulting p-values were then further tested against Uniform(0,1), which identifies if there are systematic departures from Normality. We would expect the p-values to be Uniform had the tests against Normal been non-significant across all genes. The p-values from this nested Kolomogorov-Smirnov procedure proved data to be not Normal, with p-value < 2.2e-16 in all cases (Figure S2 shows the results for DHS segment size \( \pm 2.5 \text{kb} \), and DHS segment size \( \pm 200 \text{kb} \)).

Instead of using parametric p-values from the Pearson correlations, we investigated the performance of the Pearson correlation using our randomization strategy. We assigned p-values to the obtained correlations from the observed data, based on correlations obtained using randomized data exactly as it was generated for the ARS method (100 total iterations, same as the ARS method). This resulted in the estimated proportion of true null genes to be \( \hat{\pi}_0 = 1 \), implying that there is no biological signal since no test was significant when correcting for multiple testing (Figure S3). This is in contrast to the statistic reported in the main text that the ARS method estimates that about 25\% of the genes show the signal of interest. Finally, log-transforming the data in order to better approximate the Normal does not alleviate the problem and Pearson correlations still resulted in \( \hat{\pi}_0 = 1 \).

We further tested the performance of the rank-based Spearman correlation test, as this would indicate a simple monotonic trend between the variables, gene expression and DHS. However, the power of the Spearman correlation over an order of magnitude less than the ARS, as can be seen in Figure S4, and S5.
**Aggregation test.** To test for an aggregation of significant genes from the same cell-type in terms of genomic location, we counted the number of cell-type matching significant genes within a boundary of 100kb. We tested against two different null models: (i) pairs of random significant genes (no boundary) or (ii) the fraction of neighboring genes with an outlier in matching cell-types, irrespective of significance. Both null models confirmed the presence of aggregation and possible regulatory islands at extreme significance (Fishers exact test p-values < 2.2e-16). We also tested aggregation by considering the number of genes as well as the distance between significant gene pairs from the same cell-type versus that of non-significant genes pairs. Using both measurements we find that significant ARS genes tend to aggregate (Kolmogorov-Smirnov test p-values < 2.2e-16 for both, Figure S14).

**Experimental replication.** Assessment of reproducibility was done by pairing DHS volume and gene expression of two replicates from 10 cell-types, such that we got a total of four pairings A1, B1, A2, and B2, with A, B representing DHS replicates and 1, 2 representing gene expression replicates. Next we calculated the fraction of significant genes (FDR < 0.05) from one pairing retained in another, and did this for all four-way comparisons (Figure S15).

**Determining the angular penalty rate** $c$. Recall from the main text that for a given segment size and gene, the DHS volume data and gene expression data are scaled and medoid centered, represented by $m$-vectors $x^*$ and $y^*$, respectively (here, $m = 20$ cell lines). We calculate $d_i = \sqrt{x_i^2 + y_i^2}$ for $i = 1, 2, \ldots, m$ to obtain the “ratio” component of the angle-ratio statistic (ARS) calculated as $r_i = \frac{d_i}{\text{med}(d)}$. We then calculate the angle $\theta_i$ for each data point $(x_i^*, y_i^*)$ relative to the unit vector (1,0), where $0 \leq \theta_i \leq 360$. The angular penalty involves first calculating the smaller of the two angular distances between $\theta_i$ and the identity line, calculated by:

$$\Delta_i = \begin{cases} 
|45 - \theta_i| & \text{for } 0 \leq \theta_i < 135 \\
|225 - \theta_i| & \text{for } 135 \leq \theta_i < 315 \\
|45 - (\theta_i - 360)| & \text{for } 315 \leq \theta_i < 360 
\end{cases}$$

The angular penalty is calculated as $a_i = \exp(c \times \Delta_i)$ where $c \leq 0$. ARS is then based on the product $ARS_i = a_i \times r_i$ as described in the text. The correct value of $c$ is that which captures the distribution of angles among true null genes. The most direct way to determine this value is to identify the value of $c$ such that the genes ranked least significant according to ARS produce valid null p-values. This can be accomplished by performing a statistical test of whether the large p-values follow the Uniform distribution. As can be seen from the p-value histograms in Figure S11, incorrect values of $c$ lead to either conservative or anti-conservative p-value estimates in the right tail. Within a narrow range of $c$ values, the null p-values were Uniform according to a Kolmogorov-Smirnov test, and this range was similar across all segments of DHS volume tested (Figure S11). Note that one should check carefully the normalized data (Figure S9) as well as the Kolmogorov-Smirnov test results (Figure S11) to make sure that this approach is valid. For example, if one of the two data types is nonlinearly transformed in a poor manner, then the approximate symmetry we see in the angular density in Figure S9 may not hold.
Null data randomization strategy. In resampling the data to obtain a correct null ARS distribution, the dependence between expression and DHS volume is retained by treating each 2-tuple as the unit to resample \((x_i, y_i)\) as well as the angle \(\theta_i\) associated with 2-tuple. In order to make the data exchangeable across genes, we actually resample the scaled data \((x^s_i, y^s_i)\), as defined in the main text. Finally, as “outliers” are not uniformly distributed across the cell-types under consideration, our permutation strategy should be balanced across cell-types.

The data for a given gene can take one of two possible configurations:

\[
\begin{align*}
\text{Case 1:} & \quad \begin{bmatrix} x^s_1 < 1 & y^s_1 < 1 \\ \vdots & \vdots \\ x^s_i = 1 & y^s_i = 1 \\ \vdots & \vdots \\ x^s_m < 1 & y^s_m < 1 \end{bmatrix}, \\
\text{Case 2:} & \quad \begin{bmatrix} x^s_1 < 1 & y^s_1 < 1 \\ \vdots & \vdots \\ x^s_i = 1 & y^s_i < 1 \\ \vdots & \vdots \\ x^s_j < 1 & y^s_j = 1 \\ \vdots & \vdots \\ x^s_m < 1 & y^s_m < 1 \end{bmatrix}
\end{align*}
\]

This can be understood in terms of Figure S8, which is a plot of the scaled data. Case 1 is a gene such that one of the cell line’s 2-tuple lies on the upper right corner, point \((1, 1)\), and all other \(m - 1\) points lie in the interior of the unit square. Case 2 is a gene such that one cell line’s 2-tuple lies on the top boundary \((x, 1)\), another cell line’s 2-tuple lies on the right boundary \((1, y)\), and the other \(m - 2\) points lie in the interior of the unit square. This phenomenon is simply due to the way in which the data are scaled. We take it into account, however, when generating randomized data.

The strategy we take to generate a “null” gene’s data is to randomly sample \(m\) points from Figure S8 but do so in a way that the cell lines are balanced and the two cases are represented proportionally. Moreover, when we randomly sample a point, we take with it the observed angle and utilize that angle in the ARS calculation for the null gene. The following provides the details of the algorithm.

We extend the notation by writing 2-tuples as \((x^s_{ij}, y^s_{ij})\) where \(i\) denotes the cell line as before and \(j\) denotes the gene \((i = 1, \ldots, m; j = 1, \ldots, N)\). Here, \(m = 20\) and \(N = 19,215\). This allows us to classify all such 2-tuples into \(m + 3\) different sets:

\[
\begin{align*}
\mathcal{B}_{\text{corner}} & = \{ (x^s_{ij}, y^s_{ij}) : (x^s_{ij}, y^s_{ij}) = (1, 1) \} \\
\mathcal{B}_{\text{top}} & = \{ (x^s_{ij}, y^s_{ij}) : x^s_{ij} < 1, y^s_{ij} = 1 \} \\
\mathcal{B}_{\text{right}} & = \{ (x^s_{ij}, y^s_{ij}) : x^s_{ij} = 1, y^s_{ij} < 1 \} \\
\mathcal{C}_k & = \{ (x^s_{ij}, y^s_{ij}) : i = k, x^s_{kj} < 1, y^s_{kj} < 1 \},
\end{align*}
\]

where \(\mathcal{C}_k\) is defined for \(k = 1, 2, \ldots, m\). It can be seen that \(\mathcal{B}_{\text{corner}}\) contains top right corner points, \(\mathcal{B}_{\text{top}}\) contains the top boundary points, \(\mathcal{B}_{\text{right}}\) contains the right boundary points, and the \(\mathcal{C}_k\) contain the interior points for each respective cell line.

Let \(r\) be the proportion of genes that fall into Case 1 and \(1 - r\) the proportion that fall into Case 2. The algorithm to generate a randomized data set works as follows:
1. Assign a fraction $r$ of genes to Case 1 and $1 - r$ to Case 2.

2. For each gene of Case 1, randomly assign a 2-tuple from $B_{\text{corner}}$, and record its cell line, $k_{\text{corner}}$. Randomly assign one 2-tuple from each $C_k$ for all $k \neq k_{\text{corner}}$ to the gene. This yields a gene of exactly $m$ 2-tuples of type Case 1, with one 2-tuple from each cell line.

3. For each gene of Case 2, randomly assign a 2-tuple from $B_{\text{top}}$, and record its cell line, $k_{\text{top}}$. Also randomly assign a 2-tuple from $B_{\text{right}}$, and record its cell line, $k_{\text{right}}$. We do so such that $k_{\text{top}} \neq k_{\text{right}}$. Then randomly assign to the gene one 2-tuple from each $C_k$ for all $\{k : k \neq k_{\text{top}} \text{ and } k \neq k_{\text{right}}\}$. This yields a gene of exactly $m$ 2-tuples of type Case 2, with one 2-tuple from each cell line.

Once the randomized genes are formed, the data for each gene are median centered, and then the ratio statistic is calculated. Recall that the angles are tied their original 2-tuples. Therefore, the angular penalty applied to each 2-tuple is exactly that from the original data set. From these $\text{ARS}_i$ statistics, the $\text{ARS}_{\text{max}}$ statistic is then calculated for each randomized gene.

This entire process is repeated for $B$ total permutation data sets to produce $B$ sets of $N$ randomized $\text{ARS}_{\text{max}}$ statistics. The proportion of these exceeding the observed $\text{ARS}_{\text{max}}$ statistics then forms the basis of the p-values. We performed this procedure for $B=3$, 10 and 100. We then compared the distribution of the null-statistics to verify that the procedure quickly converges for a low number of permutations. If the distributions of randomized ARS values is identical among the values of $B$, then the p-values generated from any number of permutations would likewise be stable. In all cases the distribution of null statistics were identical according to a Kolmogorov-Smirnov test, 3 versus 10: p-value 0.75, 3 versus 100: p-value 0.54, 10 versus 100: p-value 0.67. The difference in estimated $\pi_0$ between 3 and 100 permutations was $\sim 0.01$ (0.8221462 versus 0.8315243), with the lower $\pi_0$ estimate achieved at 100 permutations resulting in 9 more genes being detected at FDR<0.01, and 30 more genes at FDR<0.05. To ensure stable, and high resolution p-values all reported p-values in the paper are based on 100 permutations.

**Local ARS profiles.** For a given gene, the gene expression data $x = (x_1, \ldots, x_m)^T$ are fixed. The larger DHS segment is broken into $L$ consecutive non-overlapping smaller regions, forming a matrix

$$A = \begin{bmatrix} y_{1,1} & \cdots & y_{1,l} & \cdots & y_{1,L} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ y_{i,1} & \cdots & y_{i,l} & \cdots & y_{i,L} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ y_{m,1} & \cdots & y_{m,l} & \cdots & y_{m,L} \end{bmatrix}$$

Now the standard ARS is calculated for each column vector of $A$, $y_1 = (y_1, \ldots, y_m)^T$ against the gene expression vector $x$. From these data, we can obtain a matrix of $\text{ARS}_{i,l}$ for $i = 1, \ldots, m$ and $l = 1, \ldots, L$. 

10
However, since each $y_l$ is scaled individually (to produce $y^*_l$ in forming the $ARS_{i,l}$), the resolution is lost across regions. We therefore scale the resulting $ARS_{i,l}$ for a fixed $l$ by its maximal DHS value relative to the average maximal DHS value across all regions. Specifically, let $\max_l = \max(y_{1,l}, \ldots, y_{m,l})$. We form

$$ARS^*_{i,l} = \frac{\max_l}{\sum_{k=1}^{L} \max_k / L} ARS_{i,l}.$$ 

thereby obtaining a matrix of local ARS scores:

$$ARS_{local} = \begin{bmatrix}
ARS^*_{1,1} & \ldots & ARS^*_{1,l} & \ldots & ARS^*_{1,L} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
ARS^*_{i,1} & \ldots & ARS^*_{i,l} & \ldots & ARS^*_{i,L} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
ARS^*_{m,1} & \ldots & ARS^*_{m,l} & \ldots & ARS^*_{m,L}
\end{bmatrix}.$$ 

By selecting the proper row vector $(ARS^*_{1,1}, \ldots, ARS^*_{i,L})$ for cell type $i$ where significance was established on the larger DHS segment, we now have a local measure of which hypersensitive sites contributed the most to the detected association.
## Supplementary Tables

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Vendor</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG04450</td>
<td>fetal lung fibroblast</td>
<td>Coriell AG04450</td>
</tr>
<tr>
<td>BJ</td>
<td>skin fibroblast</td>
<td>ATCC CRL-2522</td>
</tr>
<tr>
<td>CACO2</td>
<td>colorectal adenocarcinoma</td>
<td>ATCC HTB-37</td>
</tr>
<tr>
<td>GM06990</td>
<td>lymphoblastoid (CEPH)</td>
<td>Coriell GM06990</td>
</tr>
<tr>
<td>H7-hESC (H7ESC)</td>
<td>Undifferentiated human embryonic stem cells</td>
<td>WiCell WA07 (H7)</td>
</tr>
<tr>
<td>HAEpiC</td>
<td>Human Amniotic Epithelial Cells</td>
<td>ScienCell 7110</td>
</tr>
<tr>
<td>HCF</td>
<td>Human cardiac fibroblasts</td>
<td>ScienCell 6300</td>
</tr>
<tr>
<td>HCPEpiC</td>
<td>Human choroid plexus epithelial cells</td>
<td>ScienCell 1310</td>
</tr>
<tr>
<td>HL-60</td>
<td>Human promyelocytic leukemia cells</td>
<td>ATCC CCL-240</td>
</tr>
<tr>
<td>HeLa</td>
<td>cervical carcinoma</td>
<td>ATCC CCL-2.2</td>
</tr>
<tr>
<td>HepG2</td>
<td>liver carcinoma</td>
<td>ATCC HB-8065</td>
</tr>
<tr>
<td>HMEC</td>
<td>Human mammary epithelial cells</td>
<td>Lonza CC-3150</td>
</tr>
<tr>
<td>HRCEpiC (HRCE)</td>
<td>Human renal cortical epithelial cells</td>
<td>ScienCell #4110</td>
</tr>
<tr>
<td>HUVEC</td>
<td>Human Umbilical Vein Endothelial Cell</td>
<td>Lonza CC-2517</td>
</tr>
<tr>
<td>K562</td>
<td>leukemia</td>
<td>ATCC CCL-243</td>
</tr>
<tr>
<td>PANC-1</td>
<td>pancreatic carcinoma</td>
<td>ATCC CRL-1469</td>
</tr>
<tr>
<td>SAEC</td>
<td>Small airway epithelial cells</td>
<td>Lonza CC-2547</td>
</tr>
<tr>
<td>SK-N-SH_RA (SKNSH)</td>
<td>neuroblastoma cell line diff. w. retinoic acid</td>
<td>ATCC HTB-11</td>
</tr>
<tr>
<td>SKMC</td>
<td>Human skeletal muscle cells</td>
<td>Lonza CC-2561</td>
</tr>
<tr>
<td>TH1</td>
<td>primary human Th1 T cells</td>
<td>None</td>
</tr>
</tbody>
</table>

Table S1: **Cell-lines and vendors.** The 20 cell lines and primary cells used in the study were procured from commercial or other sources as listed above. Further information on growth conditions and protocols are available via:

http://genome.ucsc.edu/cgi-bin/hgEncodeVocab?type=cellType
Table S2: DNase I hypersensitive data. The track IDs for each cell-type are specified and can be accessed via:

[http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeUwDnaseSeq/](http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeUwDnaseSeq/)

<table>
<thead>
<tr>
<th>Internal id</th>
<th>Cell-type</th>
<th>Alt. name</th>
<th>subId</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS12255</td>
<td>AG04450</td>
<td>AG04450</td>
<td>subId=907</td>
</tr>
<tr>
<td>DS10018</td>
<td>BJ</td>
<td>BJ</td>
<td>subId=393</td>
</tr>
<tr>
<td>DS8235</td>
<td>CACO2</td>
<td>Caco-2</td>
<td>subId=288</td>
</tr>
<tr>
<td>DS7748</td>
<td>GM06990</td>
<td>GM06990</td>
<td>subId=46</td>
</tr>
<tr>
<td>DS11909</td>
<td>hESCT0</td>
<td>H7-hESC</td>
<td>subId=917</td>
</tr>
<tr>
<td>DS12663</td>
<td>HAEpC</td>
<td>HAEpC</td>
<td>subId=918</td>
</tr>
<tr>
<td>DS12491</td>
<td>HCF</td>
<td>HCF</td>
<td>subId=921</td>
</tr>
<tr>
<td>DS12447</td>
<td>HCPEpC</td>
<td>HCPEpC</td>
<td>subId=922</td>
</tr>
<tr>
<td>DS11733</td>
<td>HL60</td>
<td>HL-60</td>
<td>subId=507</td>
</tr>
<tr>
<td>DS8680</td>
<td>HMEC</td>
<td>HMEC</td>
<td>subId=627</td>
</tr>
<tr>
<td>DS10662</td>
<td>HRCE</td>
<td>HRCEpC</td>
<td>subId=399</td>
</tr>
<tr>
<td>DS10060</td>
<td>HUVEC</td>
<td>HUVEC</td>
<td>subId=297</td>
</tr>
<tr>
<td>DS10011</td>
<td>Hela</td>
<td>HeLa-S3</td>
<td>subId=396</td>
</tr>
<tr>
<td>DS7764</td>
<td>HEPG2</td>
<td>HepG2</td>
<td>subId=48</td>
</tr>
<tr>
<td>DS9764</td>
<td>K562</td>
<td>K562</td>
<td>subId=106</td>
</tr>
<tr>
<td>DS9873</td>
<td>PANC1</td>
<td>PANC-1</td>
<td>subId=511</td>
</tr>
<tr>
<td>DS10514</td>
<td>SAEC</td>
<td>SAEC</td>
<td>subId=513</td>
</tr>
<tr>
<td>DS11939</td>
<td>SkMC</td>
<td>SKMC</td>
<td>subId=754</td>
</tr>
<tr>
<td>DS8476</td>
<td>SKNSH</td>
<td>SK-N-SH_RA</td>
<td>subId=737</td>
</tr>
<tr>
<td>DS7840</td>
<td>TH1</td>
<td>Th1</td>
<td>subId=49</td>
</tr>
<tr>
<td>Cell-type</td>
<td>subId</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BJ</td>
<td>295</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BJ</td>
<td>393</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caco-2</td>
<td>288</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caco-2</td>
<td>394</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM06990</td>
<td>46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM06990</td>
<td>110</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HRCEpiC</td>
<td>398</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HRCEpiC</td>
<td>399</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HRE</td>
<td>400</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HRE</td>
<td>401</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HeLa-S3</td>
<td>396</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HeLa-S3</td>
<td>397</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HepG2</td>
<td>48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HepG2</td>
<td>111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K562</td>
<td>106</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K562</td>
<td>106</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAEC</td>
<td>514</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAEC</td>
<td>513</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SK-N-SH_RA</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SK-N-SH_RA</td>
<td>737</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table S3: DNase I hypersensitive data - replicated. The track IDs for each cell-type of the replicated data is specified and can be accessed via: [http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeUwDnaseSeq/](http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeUwDnaseSeq/) Please note that the same subId for K562 appears twice as in the original file list from UCSC.
<table>
<thead>
<tr>
<th>GEO-accession</th>
<th>Cell-type</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM510534</td>
<td>AG04450-DS12255</td>
</tr>
<tr>
<td>GSM472924</td>
<td>BJ-DS11522</td>
</tr>
<tr>
<td>GSM472933</td>
<td>Caco-2-DS11539</td>
</tr>
<tr>
<td>GSM472903</td>
<td>GM06990-DS7748</td>
</tr>
<tr>
<td>GSM510581</td>
<td>H7-hESC-DS11909</td>
</tr>
<tr>
<td>GSM510516</td>
<td>HAEPiC-DS12663</td>
</tr>
<tr>
<td>GSM510545</td>
<td>HCF-DS12491</td>
</tr>
<tr>
<td>GSM510547</td>
<td>HCPEpiC-DS12447</td>
</tr>
<tr>
<td>GSM472942</td>
<td>HL-60-DS11733</td>
</tr>
<tr>
<td>GSM472934</td>
<td>HMEC-DS8680</td>
</tr>
<tr>
<td>GSM472920</td>
<td>HRCE-DS10662</td>
</tr>
<tr>
<td>GSM472935</td>
<td>HUVEC-DS10060</td>
</tr>
<tr>
<td>GSM472905</td>
<td>HeLa-S3-DS7930</td>
</tr>
<tr>
<td>GSM472906</td>
<td>HepG2-DS7764</td>
</tr>
<tr>
<td>GSM472926</td>
<td>K562-DS11524</td>
</tr>
<tr>
<td>GSM472938</td>
<td>PANC-1-DS9873</td>
</tr>
<tr>
<td>GSM472923</td>
<td>SAEC-DS10518</td>
</tr>
<tr>
<td>GSM472943</td>
<td>SKMC-DS11939</td>
</tr>
<tr>
<td>GSM472912</td>
<td>SK-N-SH_RA-DS8476</td>
</tr>
<tr>
<td>GSM472914</td>
<td>Th1-DS7840</td>
</tr>
</tbody>
</table>

Table S4: **GEO-accession number.** GEO-accession numbers for the Affymetrix exon-array data.
<table>
<thead>
<tr>
<th>GEO-accession</th>
<th>Cell-type</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM472924</td>
<td>BJ-DS11522</td>
</tr>
<tr>
<td>GSM472925</td>
<td>BJ-DS11523</td>
</tr>
<tr>
<td>GSM472900</td>
<td>Caco-2-DS8235</td>
</tr>
<tr>
<td>GSM472933</td>
<td>Caco-2-DS11539</td>
</tr>
<tr>
<td>GSM472903</td>
<td>GM06990-DS7748</td>
</tr>
<tr>
<td>GSM472904</td>
<td>GM06990-DS7784</td>
</tr>
<tr>
<td>GSM472920</td>
<td>HRCE-DS10662</td>
</tr>
<tr>
<td>GSM472921</td>
<td>HRCE-DS10666</td>
</tr>
<tr>
<td>GSM472918</td>
<td>HRE-DS10631</td>
</tr>
<tr>
<td>GSM472919</td>
<td>HRE-DS10641</td>
</tr>
<tr>
<td>GSM472905</td>
<td>HeLa-S3-DS7930</td>
</tr>
<tr>
<td>GSM472916</td>
<td>HeLa-S3-DS8200</td>
</tr>
<tr>
<td>GSM472906</td>
<td>HepG2-DS7764</td>
</tr>
<tr>
<td>GSM472907</td>
<td>HepG2-DS7768</td>
</tr>
<tr>
<td>GSM472926</td>
<td>K562-DS11524</td>
</tr>
<tr>
<td>GSM472927</td>
<td>K562-DS11525</td>
</tr>
<tr>
<td>GSM472922</td>
<td>SAEC-DS10514</td>
</tr>
<tr>
<td>GSM472923</td>
<td>SAEC-DS10518</td>
</tr>
<tr>
<td>GSM472912</td>
<td>SK-N-SH_RA-DS8476</td>
</tr>
<tr>
<td>GSM472913</td>
<td>SK-N-SH_RA-DS8482</td>
</tr>
</tbody>
</table>

Table S5: **GEO-accession number.** GEO-accession numbers for the replicated Affymetrix exon-array data.
Supplementary Figures

Figure S1: **Four examples of correlation versus angle ratio statistics (ARS).** Data points are represented by black dots and the outlier (i.e., most distant point to the origin) is the red point. The dashed line is the 45 degree identity line and the solid line is a standard regression line. The ARS and Pearson p-values are based on permuted data. To calculate ARS, the data, DHS and gene expression, were scaled by their maximum observations respectively and median centered. The distance and angle of the “outlier” relative to all other points determines the final test-statistic. ARS show robust detection of outliers while standard Pearson correlation may give spurious signals as data is not bivariate Normal. The genes under consideration are (A) A2M, (B) AADAT, (C) ACSL3, (D) ABCB6.
Figure S2: Kolomogorov-Smirnov p-values testing whether gene expression and DHS data are Normal. All expression data, and proximal (± 2.5kb) DHS were centered and scaled on a per gene basis, and tested against Normal using a Kolmogorov-Smirnov test. The histograms show the distribution of the resulting p-values in both cases. If the data were Normal, a Uniform distribution would be present. In testing the resulting p-values against Uniform, using a Kolomogorov-Smirnov test we obtain p-values < 2.2e-16 for both cases demonstrating that data cannot be assumed to be Normal.
**Figure S3:** Comparison of p-values obtained from randomized null data with ARS and Pearson correlation. Using the randomization algorithm on the data 100 times, we calculated permutation based p-values for both AR-Statistics and Pearson correlations. ARS generates well-behaved p-values (e.g., p-values > 0.5 are Uniform), while Pearson correlation results in a bimodal histogram since the Pearson correlation statistic is not designed for the signal in these data.
Figure S4: **Comparison of statistical significance from ARS and rank-based Spearman correlation.** Using the proximal ± 2.5kb DHS data, ARS shows superior performance relative to the rank based Spearman correlation. At given FDR thresholds, the ARS method identifies many more significant genes.
Figure S5: **Comparison of statistical significance from ARS and rank-based Spearman correlation.** Using the proximal ± 200kb DHS data, ARS shows superior performance relative to the rank based Spearman correlation. At given FDR thresholds, the ARS method identifies many more significant genes.
Figure S6: Genomic segments for DHS volume. Integrated DHS measurements across increasingly larger genomic segments relative to the TSS of the gene of interest were used to quantify gene-specific DHS volume. All data within the segment was used irrespectively of adjacent genes etc.
Figure S7: **Chromosome arm scaling factors.** To account for copy number variation difference in read density, etc., each DHS volume was scaled such that the total DHS volume per chromosomal arm was constant across cell-types. The above show the scaling parameter for each cell-type by chromosome.
Figure S8: **Distribution of scaled data.** By scaling the DHS volume and gene expression by their per gene maximum observations and median centered, it is evident that most data points inhabit a fan-like pattern which forks away from the diagonal. Points away from the diagonal were penalized by an angular decay rate. Colors indicate log₂ of the bin counts, and data were counted on a 20 × 20 grid.
Figure S9: **Distribution of scaled and median centered data.** Scaled and median centered data for significant genes, non-significant genes, and their difference, where the color indicates the spatial distribution of points. Here, the definition of significance is determined by FDR<0.05. All x-axes display scaled and median centered expression, all y-axes display scaled and median centered DHS volume. Significant genes tend to have maximum values in both dimensions, and inhabit the upper right corner at (1,1).
Figure S10: **Angular penalty.** A plot of the angular penalty applied to the scaled and center data. The x-axis and y-axis correspond to expression and DHS volume, and the z-axis to the angular penalty applied to any data point at the corresponding \((x, y)\) location. It can be seen that points falling along the 45° identity line are not penalized, and that the farther a point is from the line the more it is penalized.
Figure S11: The influence of $c$ on the p-value histograms. In cases where the angular distribution is not correctly captured by $c$, the estimated p-values would either be conservative or anti-conservative in the right tail as seen above. In the above example, a value of $c = -0.05$ gave Uniform null p-values in the interval $[0.5,1]$ according to a Kolomogorov-Smirnov test. For the presented analyses a range of values of $c$ was tested (see Figure S12), if multiple values of $c$ gave rise to Uniform null p-values the value of $c$ with the largest p-value was chosen. Due to differences in the distribution of the DHS values, values of $c$ were slightly different depending on the genomic segment investigated.
Figure S12: **Kolomogorov-Smirnov tests to determine c in the angular penalty.** For each value of c, a Kolomogorov-Smirnov test was performed to test for Uniform null p-values defined as p-values in the interval [0.5,1]. A non-significant p-value indicates a valid choice for c. It can be seen that all DHS segment sizes yielded similar valid c values. If multiple values of c gave non-significant p-values, the value of c yielding the largest p-value was chosen.

Figure S13 (following page): **Ingenuity enrichments for biological functions.** Significant genes within the ± 100kb region found with ARS and Spearman correlation were tested for enrichment of genes with particular biological function (similar to GO-term analysis) using Ingenuity Pathways Analysis (Ingenuity Systems®, www.ingenuity.com). In general, genes called significant with ARS show a good correspondence between cell-type and the functions found enriched. Indicating that truly cell-type specific genes were indeed detected. This pattern is not as clearly seen using Spearman correlation, likely due to the fact that using correlation on this type of data leads to many false predictions.
AG04450
BJ
CACO2
GM06990
HTESC
HAEpiC
HCF
HCPEpiC
Hela
HepG2
HL60
HMEC
HRCE
HUVEC
K562
Panc
SAEC
SKMC
SKNSH
TH1

Cell-types

Genes by relative ARS

Ingenuity enrichment

Reproductive system development & function
Digestive system development & function
Hematological system development & function
Embryonic development
Organismal development
Cardiovascular system development & function
Connective tissue development & function
Endocrine system development & function
Hematological system development & function
Hair and skin development & function
Renal and urological system development & function
Hematological system development & function
Endocrine system development & function
Hair and skin development & function
Skeletal and muscular system development & function
Nervous system development & function
Cell-mediated immune response

Cell-types

Genes by cross-product

Ingenuity enrichment

Cardiovascular System Development & Function
Skeletal and Muscular System Development & Function
Hematological System Development & Function
Hematological System Development & Function
Cell-mediated Immune Response
Organ Development
Cardiovascular System Development & Function
Hematological System Development & Function
Connective Tissue Development & Function
Hematological System Development & Function
Connective Tissue Development & Function
Embryonic Development
Tissue Development
Renal and Urological System Development & Function
Embryonic Development
Tissue Development
Hematological System Development & Function
Tissue Development
Figure S14: **Aggregation of cell-type specific genes.** Four different tests were done to test for aggregation of significant genes according to genomic location and maximal cell type. Here, the distance between significant genes are plotted against the distance between non-significant genes. The number of null genes were chosen to match the number of significant genes from that cell-type, such that the effect is not primarily driven by cell-type. A clear and significant trend show that cell-type significant genes aggregate along the genome.
Figure S15: **Assessment of reproducibility.** The p-values corresponding to genes significant (FDR $< 0.05$) from one DHS and gene expression replicate are plotted for the complementary analysis. It can be seen that small p-values from one data set yield small p-values from the replicated data set, indicating a high level of reproducibility. The fraction recaptured at FDR$<0.05$ is listed below each histogram.
Figure S16: Relative ARS scores. For each gene we can scale the $ARS_i$ values by $ARS_{\text{max}}$ in order to distinguish between significant genes with an “on-off” effect and significant genes which show a more gradual trend (i.e., more than a single cell-type may act as an outlier compared to the remaining cell-types). Significance is here determined as $\text{FDR} < 0.05$. In the color image cell-types are sorted, column-wise, by decreasing scaled $ARS_i$, the image is further divided into three blocks, representing the three major distributions of $ARS_i$ observed. (A) Significant genes with a singular outlier, i.e. “on-off”, (B) significant genes with a gradual trend, and finally (C) non-significant genes with no distinct outliers as can be seen from the high relative $ARS_i$ values extending across multiple cell-types. Right of each sub-panel is a 2D plot of the scaled and centered gene expression by DHS volume showing a representative gene from the given distribution: $IFNG$, $LOXL2$, and $ABCG4$ respectively.
Figure S17: **Positional densities of local ARS peaks by cell-type** The positional bias of cell-type specific local ARS peaks can be seen in the upper panel. For example, HL60 shows a more proximal signal relative to that of HAEpiC. A clear difference in the amount of distal regulation is seen across the different cell-types. The lower panel shows the same analysis applied to the data where the cell types are randomly shuffled among the cases considered in the upper panel.
Figure S18: Conservation of sequences corresponding to local ARS peaks. Phastcons scores within local ARS peaks and a negative control set were recorded and displayed in the quantile-quantile plot. The local ARS peaks show a marked and significantly (Kolmogorov-Smirnov p-value < 2.2e-16) higher conservation relative to the negative control set, indicating that local ARS peaks are strong candidates for functional regions.
Figure S19: **HNF1B** as chromatin / gene expression concordant in HRCE. While genes significant in HRCE have a clear over-representation of the *HNF1B* motif, the gene itself is also in chromatin / gene expression concordance.
Figure S20: **HNF4A** as chromatin / gene expression concordant in HepG2. While genes significant in HepG2 have a clear over-representation of the *HNF4A* motif, the gene itself is also in chromatin / gene expression concordance.
Figure S21: Over- and under-representation of TFBS within cell-type specific proximal (±10kb) promoters. Using only the proximal (±10kb) regions for the local ARS peak TFBS analysis, known cell-specific interactions disappear, e.g., SOX2 and POU5F1 (Oct4) for H7ESC.
Figure S22: Location of genes relative to the SNP rs1890131 and corresponding local ARS profiles. All three local ARS profiles for the genes CREG1 in HL60, RCSD1 in TH1, and CD247 in TH1, show specific and high association to the site of rs1890131, indicated by the salmon colored bar.
Figure S23: **Plots of gene expression and DHS volumes for the three genes related to rs1890131.** Concordance between the entire ± 100kb DHS segment and gene expression for the three genes for which rs1890131 falls within a local ARS peak. For *RCSD1*, HL60 is a clear secondary outlier.
Figure S24: The SNP rs1890131 in a transcription factor hub. The SNP is centered in a hub of transcription factor binding sites as detected by the ENCODE transcription factor ChIP-seq track. This site is strongly associated in the local ARS profiles with three adjacent genes: *CREG1* in HL60, *RCSD1* in TH1, and *CD247* in TH1.
Figure S25: **Raw DHS fragments for HBB and HBE1.** The DHS values within a region of interest for HBB and HBE1 across all cell-types are highlighted by salmon colored bars. These regions correspond to known hypersensitive regions. Interestingly, one region is present in almost all cell-types, whereas, others are unique to K562.
Figure S26: **Scaled and centered gene expression and DHS volume for HBB.** The gene expression and DHS volume [0-1] scaled and medoid centered for all 20 cell-types, across the entire ± 100kb DHS segment. This show HBB as a clear and singular outlier for K562.
Figure S27: ARS profiles for HBB. Local ARS profiles across all 20 cell-types, where the salmon colored bars indicate ARS peaks. The 6th hypersensitive site (from the left) is not seen as an ARS peak, due to its presence in multiple cell-types. The hypersensitive site present in the 3rd bar (from the left) appears as an ARS peak, despite being present in all cell-types due to its high local value.
Figure S28: Scaled and centered gene expression and DHS volume for HBE1. The gene expression and DHS volume [0-1] scaled and medoid centered for all 20 cell-types, across the entire ± 100kb DHS segment. HBE1 is a clear and singular outlier for K562.
Figure S29: **ARS profiles for HBE1.** Local ARS profiles across all 20 cell-types, where the salmon colored bars show that the local ARS profile is markedly different from HBB, indicating different regulatory importance of the hypersensitive regions for HBB relative to HBE1.
Figure S30: **ARS profiles for entire β-globin locus.** Local ARS profiles for all genes for which we had expression data in the β-globin control locus region. The hypersensitive regions are utilized differently by the genes, indicating a difference in regulatory regime. ARS peaks are highlighted by salmon colored bars. The third ARS peak from the left shows significant variation across the genes, despite being chromatin / gene expression concordant in the same cell-type.
Figure S31: Raw data DHS fragment counts for TAL1. The DHS values within a region of interest for TAL1 across all cell-types, where the salmon colored bar indicates the area detected as a local ARS peak for TAL1.
Figure S32: *Scaled and centered gene expression and DHS volume for TAL1*. The gene expression and DHS volume [0-1] scaled and medoid centered for all 20 cell-types, across the entire ± 100kb DHS segment. *TAL1* is an outlier for K562, and a secondary outlier for CACO2.
ARS across all cell-types for TAL1

Figure S33: **ARS profiles for TAL1.** Local ARS profiles across all 20 cell-types, where the salmon colored bars indicate the ARS peaks. Interestingly, a clear hypersensitive site from Figure S31 does not yield a local ARS peak as it is present in all tissues, and therefore not in cell-type specific concordance with the gene expression values.
Figure S34: **Distinct local ARS profiles for adjacent genes.** Adjacent significant genes show largely distinct local ARS profiles related to the specific cell-type, indicative of separate regulatory programs. Salmon colored bars show ARS peaks for *CCR2* in HL60 and teal colored bars indicate ARS peaks for *CCR5* in TH1. This example demonstrates the ability of the ARS approach to detect differences in chromatin / gene expression concordance between adjacent genes.
Figure S35: **Shared ARS profile within a single cell-type.** Neighboring *HOXD* genes (*HOXD8, HOXD4, HOXD3*) all significant in HRCE show identical usage of local hypersensitive sites indicative of shared regulatory mechanisms. Salmon colored bars indicate local ARS peaks, which show similar ranking in magnitude within each gene, but differences in total magnitude due to differences in gene expression of the individual genes.
Figure S36: **Sub-maximal outliers for LOXL2.** Visualization of the sub-maximal outliers BJ and HAEpiC with similar local ARS profiles as SKMC shown below in Figure S37. While statistical significance is based on ARS$_{\text{max}}$, secondary outliers can be seen for other cell-types, indicating shared chromatin / gene expression concordance across cell-types.
Figure S37: *LOXL2* share local ARS profiles across cell-types. While ARS$_{\text{max}}$ identified SKMC as the significant outlier, HAEpiC and BJ show large relative local ARS values. Closer inspection shows overlapping utilization of local hypersensitive sites across all three cell-types indicative of possible shared regulatory mechanisms (salmon colored bars). However, individual sites are also detected (teal bars) only shared among SKMC and HAEpiC.