Published in final edited form as:

Nat Methods. 2015 May ; 12(5): 433–438. doi:10.1038/nmeth.3329.

Identification of active transcriptional regulatory elements with GRO-seq

Charles G. Danko[1,2,3], **Stephanie L. Hyland**[4], **Leighton J. Core**[5,11], **Andre L. Martins**[6], **Colin T Waters**[5,11], **Hyung Won Lee**[5], **Vivian G. Cheung**[7,8], **W. Lee Kraus**[9,10], **John T. Lis**[6], and **Adam Siepel**[3,11]

[1]Baker Institute for Animal Health, Cornell University, Ithaca, NY, USA

[2]Department of Biomedical Sciences, Cornell University, Ithaca, NY, USA

[3]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY,USA

[4]Tri-Institutional Training Program in Computational Biology and Medicine, New York, New York, USA

[5]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA

[6]Graduate Field in Computational Biology, Cornell University, Ithaca, NY, USA

[7]Life Sciences Institute, University of Michigan, Ann Arbor, MI, USA

[8]Howard Hughes Medical Institute, Chevy Chase, MD, USA

[9]Laboratory of Signaling and Gene Regulation, Cecil H. and Ida Green Center for Reproductive Biology Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA

[10]Division of Basic Research, Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center, Dallas, TX, USA

Abstract

Address correspondence to: Adam Siepel, Ph.D., asiepel@cshl.edu, Charles G. Danko, Ph.D., dankoc@gmail.com, John T. Lis, Ph.D., jtl10@cornell.edu.

[11]**Present addresses:**

Department of Molecular and Cell Biology, Institute for Systems Genomics, University of Connecticut, Storrs, Connecticut, USA (L.J.C.), Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA (C.T.W.) and Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA (A.S.).

Accession codes:

Raw and processed data is available from the Gene Expression Omnibus, identifiers GSE60456 (K562, GM12878), and GSE66031 (CD4+ and Jurkat T-cells, as well as dREG scores). Data are also available for visualization on the UCSC genome-browser track hubs http://www.charlesdanko.org/hub/dreg/.

Author Contributions:

CGD designed the dREG tool. CGD, ALM, and SLH designed and implemented the software. CGD, SLH, ALM, LJC, JTL, and AS analyzed the data and interpreted the results. LJC, CTW, CGD, HWL, JTL, WLK, and VGC contributed data, and helped to troubleshoot experiments. CGD, AS, JTL, LJC, SLH, and ALM wrote the manuscript.

Competing financial interests:

The authors have no competing financial interests to declare.

Transcriptional regulatory elements (TREs), including enhancers and promoters, determine the transcription levels of associated genes. We have recently shown that global run-on and sequencing (GRO-seq) with enrichment for 5'-capped RNAs reveals active TREs with high accuracy. Here, we demonstrate that active TREs can be identified by applying sensitive machine-learning methods to standard GRO-seq data. This approach allows TREs to be assayed together with gene expression levels and other transcriptional features in a single experiment. Our prediction method, called discriminative Regulatory Element detection from GRO-seq (dREG), summarizes GRO-seq read counts at multiple scales and uses support vector regression to identify active TREs. The predicted TREs are more strongly enriched for several marks of transcriptional activation, including eQTL, GWAS-associated SNPs, H3K27ac, and transcription factor binding than those identified by alternative functional assays. Using dREG, we survey TREs in eight human cell types and provide new insights into global patterns of TRE function.

Transcriptional regulatory elements (TREs), such as promoters, enhancers, and insulators, are critical components of the genetic regulatory programs of all organisms[1]. These elements regulate gene expression by facilitating or inhibiting chromatin decompaction, transcription initiation, and the release of RNA polymerase II into productive elongation, as well as by maintaining the three-dimensional architecture of the nucleus. TREs enable complex, cell-type- and condition-dependent patterns of gene expression that contribute to nearly all biological processes.

Since the completion of high-quality gene catalogs for humans and most model organisms, the comprehensive identification of TREs has emerged as a primary challenge in genomic research. At present, these elements are most effectively identified using high-throughput genomic assays that provide indirect evidence of regulatory function, such as chromatin immunoprecipitation and sequencing (ChIP-seq) of bound transcription factors (TFs) or histone modifications, and DNase-I hypersensivity and sequencing (DNase-seq)[2–4]. However, the methods currently in wide use have important limitations. For example, ChIP-seq requires a high-affinity antibody for the targeted TF or histone modification of interest and must be performed separately for each target. Likewise, assays that measure chromatin accessibility or histone modifications provide only circumstantial evidence that the identified sequences are actively participating in transcriptional regulation[5]. Even STARR-seq, a clever high-throughput reporter-gene assay, identifies only regions that are inactive *in situ*, because the assay is independent of native local chromatin structure and genomic context[6].

Recently it has become clear that a defining characteristic of active TREs is that they are associated with local transcription. Enhancer-templated non-coding RNAs, or eRNAs, have recently been associated with thousands of stimulus-dependent enhancers[7]. Like active promoters, these enhancers exhibit transcription initiation in opposing directions on each strand, a phenomenon called divergent transcription[8–10]. This characteristic pattern can be a powerful tool for the identification of active TREs in a cell-type specific manner[7,11–15]. Methods that measure the production of nascent RNAs, such as Global Run-On and sequencing (GRO-seq)[8] and its successor, Precision Run-On and sequencing (PRO-seq)[16], are particularly sensitive for detecting these transient enhancer-associated RNAs, because

they measure primary transcription before unstable RNAs are degraded by the exosome[12,17]. Recently, we have shown that an extension of GRO-seq that enriches the nuclear run-on RNA pool for 5'-7meGTP-capped RNAs, called GRO-cap, further improves sensitivity for eRNAs, and can be used to identify tens of thousands of transcribed enhancers and promoters across the genome[18,19].

Here, we introduce a new computational method for accurately identifying transcribed TREs directly from standard GRO-seq or PRO-seq data. Our method, called discriminative Regulatory Element detection from GRO-seq (dREG), uses a novel, multiscale summary of GRO-seq or PRO-seq read counts, and then employs support vector regression[20] (SVR) to detect the characteristic patterns of transcription at TREs. dREG allows high-quality predictions of TREs for any cell type with existing GRO-seq or PRO-seq data. We applied the method to four cell types for which data was previously available and four for which we provide new data. Combining these predictions with data from the ENCODE project, we found that the predicted TREs fall into four major classes. The class distinguished by a strong dREG signal was also enriched for H3K27 acetylation (H3K27ac), TF binding, eQTL, and GWAS-associated SNPs, suggesting that TREs identified using dREG are actively controlling cell-type-specific transcription.

Results

Regulatory Element Identification in Eight Cell Types

We devised a machine-learning approach, called dREG, to identify TREs, including both promoters and enhancers, from standard GRO-seq or PRO-seq data (Fig. 1a and Supplementary Fig. 1). The key to our method is a feature vector that summarizes the patterns of aligned GRO-seq reads near each candidate element at multiple scales. This feature vector consists of read counts for windows ranging in size from 10 bp to 5 kbp, standardized using the logistic function (Supplementary Fig. 2a). The feature vector is passed to a SVR, which scores sites with high PRO-seq signal for similarity to a training set of TREs. To train our classifier, we used TREs identified from GRO-cap data[19] as positive examples and regions of matched PRO-seq signal intensity lacking additional marks associated with TREs as negative examples. After training and optimization of several tuning parameters (Supplementary Tables 1 and 2), the program displayed excellent performance when applied to PRO-seq data for K562 cells (AUC= 0.99; Supplementary Fig. 2b).

We ran dREG to predict the location of TREs genome-wide in K562 cells, adopting a prediction threshold that limits the genome-wide false discovery rate to 10%. At this threshold, we recovered 94% of 21,082 GRO-cap 'paired' sites (i.e., sites for which divergent Pol II initiation was detected in both directions), and 94% of 9,940 active transcription start sites detected by CAGE (Fig. 1b). Furthermore, we observed high sensitivity within the subsets of GRO-cap peaks overlapping gene bodies (84%), chromHMM promoters (95%) and chromHMM enhancers (80%). We applied dREG to an independent cell type, GM12878 lymphoblastoid cells, without retraining the classifier. Based on GRO-cap data available for GM12878, dREG achieved similar performance in this cell type for all classes of regulatory elements tested (Fig. 1b), indicating that the method generalized well across

cell types. Finally, we examined the sensitivity of dREG to sequencing depth and data quality and found that sensitivity is satisfactory with as few as 40M mapped reads (Supplementary Fig. 3). Together, these findings demonstrated that dREG accurately identified active TREs across a broad spectrum of GRO-seq and PRO-seq data sets.

dREG enabled us to predict TREs for additional cell types for which GRO-seq or PRO-seq data is available. We analyzed existing GRO-seq data sets for MCF-7, IMR90, GM12878, and AC16 cell lines[8,11,21–23], as well as new data that we generated in four cell types analyzed by the ENCODE and Epigenome Roadmap projects, including K562, primary CD4+ T-cells, Jurkat leukemia cells, and HeLa carcinoma cells. For each of these new cell types, GRO-seq or PRO-seq libraries were produced and sequenced to a depth of 53–375 million mappable reads (Supplementary Table 3). The dREG model trained on K562 cells was applied to each data set. The dREG predictions for each cell type include ~30,000 TREs (20,848–37,545), covering ~1.3% (0.82–1.68%) of the human genome, at a median size of ~1.1 kb. Approximately half of these elements mark active promoters and half mark a subset of distal enhancers (Supplementary Fig. 4). The union of these predictions across all eight cell types includes 103,096 TREs, covering 4.3% of the human genome.

Four Major Classes of Transcriptional Regulatory Elements

We compared dREG predictions with two complementary sets of TREs: ChromHMM predictions of promoters, enhancers, and insulators[24], and DNase-I hypersensitive sites (DHSs)[25,26]. ChromHMM predictions are based on genome-wide ChIP-seq assays targeting histone modifications and CTCF binding, whereas the DHSs identify regions of ‘open’ chromatin where the DNA is accessible to DNase-I cleavage. For the DHSs, we used high-confidence DNase-I accessible sites, defined as the intersection of Duke and UW DHS predictions (Supplementary Fig. 5). After taking the union of these three sets of putative TREs (see Methods), we labeled each TRE by the collection of methods that identified it (dREG, ChromHMM, and/or DNase-seq).

Our analysis of the labeled TREs indicated that these three methods identified nested sets of elements, with the ChromHMM predictions being most inclusive, the DHSs largely forming a subset of the ChromHMM predictions, and dREG, in turn, generally narrowing those identified by DNase-seq to a smaller subset (Fig. 2a and Supplementary Fig. 6). Interestingly, the ChromHMM predictions of insulators showed limited overlap with DNase-seq or dREG predictions. Thus, we observe four main classes of TREs based on these methods: (1) actively transcribed TREs identified by dREG, DNase-seq, and ChromHMM, comprising 14–17% (depending on the cell type) of the merged set (+dREG); (2) ‘open’ but untranscribed TREs identified by DNase-seq and ChromHMM (excluding the insulator predictions) but not dREG, accounting for 10% (−dREG); (3) elements with histone modifications indicative of enhancers but that are untranscribed and display either weak (16.8%) or no (23.2%) evidence of DNase-I accessibility, accounting for 40% (Marked Chromatin Only; or MCO); and (4) the ChromHMM insulator predictions, which commonly overlap with DHSs and comprise 25% of all TREs (Insulator). Other combinations of assays account for only 1–6% of TREs, and can likely be attributed to experimental biases, false positive and/or false negative predictions. Notably, a follow-up analysis suggested that the

absence of dREG predictions in the −dREG set cannot be explained by an inadequacy of sensitivity or sequencing depth (Supplementary Fig. 7). These observations indicated that dREG identified a smaller collection of transcribed TREs that might have functional properties that distinguish them from sites predicted using chromatin modifications and/or DNase-seq alone.

Functional Properties of Distinct TRE Classes

We investigated the distinctions among the four classes of TREs by comparing their genomic distributions with those of complementary assays. First, we characterized the enrichments of three histone marks—H3K27ac, H3K9ac, and H3K4me1—among the MCO, −dREG, +dREG, and Insulator classes. H3K27ac and H3K9ac denote ‘active’ regulatory elements[27,28], and H3K4me1 is a universal mark located at both active and so-called ‘primed’ enhancers[29]. We found that dREG TREs are strongly enriched for the ‘active’ H3K27ac and H3K9ac signals and, accordingly, that the majority of ENCODE peak calls for these marks are also identified by dREG (Fig. 2b and Supplementary Fig. 8a–c). In contrast, the −dREG and MCO classes show little or no H3K27ac or H3K9ac signal. Moreover, the minority of +dREG TREs that are not associated with H3K27ac peak calls nevertheless display elevated H3K27ac ChIP-seq signal (Supplementary Fig. 9), suggesting that many simply fall below the detection threshold used in peak calling. These observations suggest that H3K27ac and +dREG point to the same class of functional element. H3K4me1 is not only enriched at dREG TREs, but is also found at high levels in the −dREG and MCO classes. Thus, dREG identifies the same genomic regions as detected using ChIP-seq for H3K27ac and H3K9ac, and a subset of H3K4me1 peaks, suggesting that it can effectively distinguish between ‘active’ and ‘poised’ enhancer classes.

The observation that TREs in the MCO class are not accessible to DNase-I cleavage suggests that access to these DNA sequences might be restricted by nucleosomes or higher-order forms of chromatin structure. We used MNase-seq data to map the locations of nucleosomes surrounding all four classes of TREs in K562 cells. We found that TREs in the MCO class have a well-positioned nucleosome near their center (Fig. 2C), which likely occludes binding by transcriptional activators as well as cleavage by DNase-I. By contrast, −dREG enhancers typically contain an array of well-positioned nucleosomes in which the central nucleosome appears to have been displaced, whereas +dREG TREs, on average, contain a large nucleosome-free region surrounding the center and extending for ~1–2 kbp in both directions (although this pattern is most prominent at promoters; Supplementary Fig. 8d). This observation of positioned nucleosomes in the MCO class, but not in the classes additionally characterized by DNase-I hypersensitivity and/or active transcription, further supports that these represent fundamentally distinct classes of TREs.

Transcription Factors Activate and Suppress eRNA Synthesis

Fundamental differences among the four TRE classes are likely to be mirrored by patterns of TF binding. Therefore, we examined binding of 91 TFs for which high-resolution ChIP-seq data is available in K562 cells. Almost 70% of TREs in the MCO class do not bind any TFs (Fig. 2d), and most −dREG TREs bind small numbers (i.e., 1–10). dREG TREs, by contrast, display a striking enrichment for binding many TFs (18 on average, and 39% bind >20 TFs).

To identify TFs that contribute to transcriptional activation at TREs, we created a logistic regression model with the transcription status of each distal TRE as the response, and the presence or absence of ChIP-seq-assayed TF binding events within the TRE (in K562, GM12878, MCF-7, and HeLa cells) as the predictors. This model predicts the transcription status of a holdout set of DHSs with remarkably high accuracy (Supplementary Fig. 10a; AUC= 0.86–0.95), and preforms notably better than a model based only on the absolute level of DNase-seq signal intensity (AUC= 0.80 in K562). These observations suggest that binding by particular TFs, more than simply the degree of chromatin accessibility, is responsible for the differential transcriptional outcomes observed in dREG TREs. This regression analysis also provides additional information about the relative importance of individual TFs in predicting whether or not a site is transcribed (Supplementary Fig. 10b). A comparison of regression coefficients indicates that components of the preinitiation complex, the histone acetyltransferase P300, and many sequence-specific activators (e.g., AP-1, PU1, CEBPB) are highly predictive of transcription initiation at TREs. By contrast, transcriptional co-repressors (e.g., HDACs and TRIM28) are associated with an absence of transcription.

Insulator-associated proteins (e.g., CTCF, RAD21, and SMC3) are also associated with an absence of transcription. This finding is consistent with the overlap observed between dREG sites and either CTCF peak calls (Supplementary Fig. 8c) or raw signal (Supplementary Fig. 11). Notably, the 18% of CTCF peak calls which do intersect dREG TREs are overwhelmingly found in promoters (77%) rather than enhancers (23%). These findings strongly suggest that CTCF plays an indirect role in transcriptional regulation.

Predicting Transcription Factor Binding using dREG

Having shown that TF binding is predictive of transcription initiation at TREs, we next addressed an inverse question: is transcription at TREs predictive of whether or not a TF is bound to DNA sequences matching its cognate motif? Most TFs bind only a small fraction of DNA sequences matching their motif[30], making TF binding site prediction a challenging computational problem. We asked whether dREG could be useful as a surrogate for, or complement to, DNase-seq data, which is widely used as an aid in the identification of TF binding[31–33]. As a proof of concept, we chose four transcriptional activators (NRF1, ELF1, SP1, and MAX) with a range of motif information contents[34] and positive regression coefficients in the analysis described above, but otherwise selected at random. For all four TFs, we found that dREG scores alone predict the occupancy of motif matches with accuracy similar to the PIQ program[32], which makes use of DNase-seq data in predicting TF binding. For example, for ELF1 (Fig. 3), dREG produces a ROC score 3.8% lower than PIQ (AUC= 0.88 [dREG], 0.92 [PIQ]) and both assays identify TF binding sites more accurately than motif matches alone (AUC= 0.67). Jointly modeling DNase-seq, dREG, and the motif match score improves classification accuracy 2.6–6.6% (AUC= 0.94), exceeding the PIQ score in this task. Thus, dREG appears to be a useful complement to DNase-seq based models of TF-DNA interaction for sequence-specific activators.

Enrichment for eQTL and GWAS hits in dREG Predictions

We asked whether +dREG TREs contain the subset of open-chromatin sites that are actively regulating gene expression. To explore this possibility, we compared the density of expression quantitative trait loci (eQTL) identified in lymphoblastoid cell lines (LCLs)[35] among +dREG, −dREG, and MCO TREs. We found that +dREG TREs in GM12878 LCLs contain 6.4–26.3-fold higher eQTL densities in LCLs than in other classes of TRE (Fig. 4a), and account for 571 out of 755 of the eQTL that intersect with the functional marks considered here (~76%). This observation is partially explained by systematic biases in eQTL density for gene promoters, yet if we focus on TREs associated with ‘enhancers’ only, we still observe a 2.4–9.8-fold enrichment in eQTL densities in +dREG TREs relative to the −dREG, MCO, and Insulator classes ($p < 2e-5$; Fisher’s Exact Test). This residual enrichment cannot be explained by differences in the distributions of the distance of these site classes relative to TSS annotations (Supplementary Fig. 12), and suggests that dREG TREs are more likely to be actively regulating gene expression than other TREs.

Genome wide association studies (GWAS) generally implicate long haplotype blocks of single nucleotide polymorphisms (SNPs), making it challenging to identify SNPs that are causally associated with disease processes. Because dREG identifies a relatively small subset of active TREs, we speculated that it might be a useful tool for narrowing GWAS SNPs for functional validation. To illustrate the utility of dREG in this application, we obtained a set of putatively functional GWAS SNPs[36]. We found that dREG sites detected in relevant primary cell types are substantially enriched in GWAS-associated SNPs. For example, SNPs associated with autoimmune disorders are enriched in dREG sites in CD4+ T-cells and GM12878 LCLs (B-cells), including SNPs for celiac disease (7.4 and 9.7-fold, respectively), rheumatoid arthritis (6.9 and 11.2), and type-1 diabetes (4.5 and 5.3). As we observed for eQTL, cell-type specific GWAS SNPs are found at higher densities in +dREG TREs compared with other functional classes (Fig. 4b). These observations demonstrate that dREG can be useful for prioritizing GWAS validation experiments.

Discussion

We have introduced a new high-throughput prediction method, called dREG, for detecting active TREs using GRO-seq or PRO-seq data. In combination with a single PRO-seq experiment, the dREG program allows investigators to interrogate many aspects of gene expression simultaneously, including not only TREs, but also TF binding, expression levels, and pausing. This efficiency is vital in a number of applications of current interest, for example in cancer genomics and personalized medicine, in which the use of genomics technologies is currently limited by sample quantities and the high cost of collecting data in large numbers of subjects.

By comparing dREG sites to other functional genomic assays, we demonstrated the existence of at least four major classes of TREs in human cells. These classes correspond to closed chromatin marked by histone modifications such as H3K4me1 (MCO), DNase-I accessible DNA without a dREG signal (−dREG), insulator factor binding (CTCF), and transcription initiation detected by dREG (+dREG). Several lines of evidence, including

enrichments for eQTL, transcriptional activators, and histone acetylation suggested that dREG identifies genomic sites that play a direct and active role in gene regulation.

We discovered three independent classes of regulatory elements that are untranscribed (−dREG, MCO, and Insulator). These TREs showed several indications of reduced regulatory activity, including a paucity of eQTL (Fig. 4a), depletion of transcription factor binding (Fig. 2d), and the absence of histone acetylation (Fig. 2b). Insulators appeared to be a distinct functional class, as they were found to be depleted for the functional marks examined here, yet their relatively high evolutionary conservation (Supplementary Fig. 13), as well as prior work[37], strongly suggested that insulators function in various aspects of cellular biology. The other two inactive classes of TRE, MCO and −dREG, appeared to have distinct mechanisms of inactivation, including the presence of a central nucleosome that occludes activator binding (MCO), and either DNA sequence-dependent binding by transcriptional repressors, or a lack of binding by transcriptional activators, at open chromatin (−dREG). In some cases, we observed changes between these TRE classes in different cell types (Supplementary Fig. 14), suggesting that they might reflect intermediates in the assembly of active regulatory elements. Future studies will identify the functional mechanisms that are responsible for the assembly and activation of TREs, and will further elucidate the relationships and mechanistic transitions among these classes of regulatory elements.

Online Methods

Training the Support Vector Regression Model

Overview—We treated transcription start site detection using GRO-seq and PRO-seq data as a regression problem (hereafter we refer only to GRO-seq, but the same methods apply to both sources of data). Our goal was to separate regions of high GRO-seq signal intensity into a class in which RNA polymerase originates by initiation and rapidly transitions to elongation (positive set, comprised of transcription start sites), and a class through which polymerase elongates (negative set, largely comprised of gene bodies). This classification problem was addressed using a standard epsilon-support vector regression (SVR), as described in the following sections.

GRO-seq Signal Intensity Requirements—We removed from consideration genomic positions with very low signal levels, implicitly assigning these positions to the negative set. We retained sites meeting either of the following two signal intensity thresholds: one or more reads on both the plus and minus strand within a window of 1 kbp, or three or more reads within a window of 100 bp on either the plus or minus strand. At these cutoff thresholds, 93% of K562 GRO-cap peaks contained at least one informative site in a PRO-seq library depth of ~40M reads. The remaining sites were segmented into non-overlapping 50 bp intervals to improve the speed of processing on large datasets.

GRO-seq Feature Vector—GRO-seq read counts were summarized in our model as a multi-scale feature vector, as illustrated by the barchart (Fig. 2a). We counted GRO-seq reads that mapped in non-overlapping windows on either side of a central base that met the signal intensity requirements (as described above). Our approach represented the genome at

multiple scales (window sizes). For each scale, we counted reads in the specified number of non-overlapping windows both upstream and downstream of the central base. Each scale could represent redundant information in the GRO-seq read counts. The final feature vector was constructed by concatenating the vectors representing read counts at each scale and strand. The specific parameters of the scales and number of windows at each scale were optimized using cross-validation (as described below, and depicted in Supplementary Table 2).

Data Standardization—GRO-seq data was standardized using the logistic function, $F(t)$, with parameters α and β, as follows:

$$F(t)=\frac{1}{1+e^{-\alpha(t-\beta)}}$$

where t denotes the read counts in each window. We find it convenient to define the 'tuning' parameters α and β in terms of a transformed pair of parameters, x and y, such that x represents the fractional portion of the maximum read count depth at which the logistic function reaches 1 and y represents the value of the logistic function at read counts of 0. The relationship of (α, β) to (x, y) is given by the following equations:

$$\beta=x \bullet \max(t) \quad (1)$$

$$\alpha=\log(1/y-1)/\beta \quad (2)$$

where max (t) denotes the maximum read depth, as computed separately for each window size and strand in the feature vector. In practice, we found it convenient to fix the value of y at 0.01 and use x for tuning. We tried values of x between 0.01 and 1.0, and found that the optimal AUC was achieved at x = 0.05 (Supplementary Table 1). Using this function in its optimized form tends to assign each position a value near 0 or 1, and consequently most of the signal for dREG is dependent on where reads are located, rather than on the relative read depths.

We also evaluated alternative standardization approaches, including simply dividing the reads in each feature vector by their maximum value, but these approaches did not perform as well as the logistic function.

Training the dREG Support Vector Regression model—We fit an epsilon-support vector regression model using the e1071 R package[38], which is based on the libsvm SVM implementation[39]. When training dREG, we assigned a label of 1 to sites intersecting both GRO-cap transcription start sites[19] and high-confidence DHS, and excluded from the training set any sites intersecting a functional mark indicative of a regulatory element but not a GRO-cap peak (including ChromHMM enhancers or promoters). All other positions in the genome meeting the GRO-seq signal requirements (described above) were assigned a score of 0. The final SVR was trained on a matched set of 100,000 loci (comprised of 50,000 positive and 50,000 negative examples) using PRO-seq data in K562 cells. Sites in

the positive set (i.e., GRO-cap peaks) were chosen at random. When selecting the set of negative (i.e., non-transcription start site) examples, we chose 25% of sites to enrich for positions that were commonly associated with false positives during preliminary testing. These include 15% of the negative set that were selected to be within 1–5 kbp of the positive regions (to improve the resolution of dREG), and 10% in regions where the 3' ends of annotated genes on opposite strands converged (to eliminate a common source of false positives). The remaining 75% of negative sites were selected at random from the set of positions across the genome meeting the GRO-seq signal requirements (described above).

Optimizing Tuning Parameters—Tuning parameters were optimized on a balanced set of 50,000 loci (comprised of 25,000 positive and 25,000 negative examples), and performance was evaluated on a holdout set of 2,000. Parameters were chosen to maximize the area under the receiver operating characteristic curve (AUC). We first selected parameters of the data transformation that maximized the AUC using a fixed feature vector (20 windows, each 10, 50, and 500 bp in size). Subsequently we fixed the optimal data standardization tuning parameter, x (see *Data Standardization* section, above), and selected the feature vector, including the number and size of windows, which maximized the AUC. False positives were defined as sites that did not overlap GRO-cap, DHSs, or ChromHMM (promoters, enhancers, or insulators). True positives were sites that overlapped GRO-cap HMM predictions[19]. False negatives were sites that were identified by GRO-cap, but were not identified by dREG. True negatives were sites that were not identified by dREG, or any of the other assays. Various tuning parameter settings are summarized in Supplementary Tables 1 and 2.

Running dREG and Post Processing—We ran dREG on GRO-seq or PRO-seq data in eight cell types. We used the SVR model trained in K562 cells to compute the predicted score at each position meeting the GRO-seq signal intensity thresholds. To call dREG ‘peaks’ we thresholded this score at 0.77, which we found returned a ~10% false discovery rate (FDR) in two datasets for which extensive data was available (K562 and GM12878). In cell types with lower read counts, this score was likely to be conservative, resulting in both a lower FDR and lower sensitivity (see Supplementary Fig. 2). Regions meeting the dREG signal requirement within 500bp of one another were merged to prevent the independent detection of the same promoter or enhancer elements.

dREG Sensitivity to Sequencing Depth and Library Quality—To evaluate the sensitivity of dREG to sequencing depth, we subsampled the K562 data by removing reads at random from the bed files representing mapped reads. We ran the dREG algorithm as described, either with or without re-training the model on the reduced read depth (both are plotted in Supplementary Fig. 2). Artificial low-quality datasets were created by choosing genomic coordinates with mapped reads at random and redistributing their reads to neighboring sites in a 50 kbp (non-overlapping) window. In each window, locations were retained with probability proportional to the original read density at that site. This procedure was designed to re-create the profile observed in low-quality data, in which large numbers of reads tend to align on a small number of positions, creating the appearance of ‘spikes’ when viewed on the genome browser. The *asymptotic unique reads* metric used to evaluate data

quality was defined as the number of unique genomic coordinates in a GRO-seq library in the limit as the number of mapped reads approaches infinity. This value was estimated by subsampling the read depth and measuring the slope of the number of unique locations covered as a function of the library sequencing depth. We interpolated the number of uniquely covered genomic coordinates to 1% of its final value assuming that the slope of the read depth did not change.

Software availability—A software package implementing the dREG approach to TRE identification is freely available for download from https://github.com/Danko-Lab/dREG.

GRO-seq and PRO-seq Library Prep

Extraction of Primary CD4+ T-cells from Blood Samples—Blood samples (80–100mL) from three human individuals were collected at Gannett Health Services at Cornell University in compliance with Cornell IRB guidelines. Informed consent was obtained from all donors. Mononuclear cells were isolated using density gradient centrifugation, and CD4+ cells were extracted using CD4 microbeads from Miltenyi Biotech (130-045-101), following the manufacturer's instructions. Primary CD4+ T-Cells were kept in culture (RPMI-1640, supplemented with 10% FBS) for 1–3 hours to recover homeostasis.

Cell Culture Conditions and PRO-seq Library Preparation—Both primary and Jurkat CD4+ T-cells were maintained in RPMI-1640 media supplemented with 10% FBS, and treated for 30 minutes with low amounts of DMSO and ethanol (as they are controls for a separate experiment, manuscript in preparation). To isolate nuclei, cells were resuspended in 1mL lysis buffer (10mM Tris-Cl pH 8, 300mM sucrose, 10mM NaCl, 2mM MgAc2, 3mM CaCl2, and 0.1% NP-40). Nuclei were washed in 10mL of wash buffer (10mM Tris-Cl pH 8, 300mM sucrose, 10mM NaCl, and 2mM MgAc2) to dilute free NTPs. Nuclei were washed in 1mL, and subsequently resuspended in 50uL, of storage buffer (50mL Tris-Cl pH 8.3, 40% glycerol, 5mM MgCl2, and 0.1mM EDTA), snap frozen in liquid nitrogen, and kept for up to 6 months before preforming PRO-seq. HeLa cells were maintained in DMEM media supplemented with 10% FBS and 1× pen/strep (Gibco). Cells were harvested by rinsing the tissue culture plate several times in 1× PBS followed by scraping in 10ml of 1× PBS. Cells were pelleted by centrifugation and nuclei were isolated as described above. K562 cells were maintained in culture and nuclei were isolated exactly as previously described[19]. K562 and HeLa carcinoma cells were purchased from the American Type Culture Collection (ATCC; CCL-243 and CCL-2.2, respectively). GM12878 cells were obtained from the Coriell Institute for Medical Research (GM12878). Jurkat T-cells were obtained from the Mangelsdorf and Kliewer labs at UT Southwestern. All cells, except primary and Jurkat CD4+ T-cells, were tested for mycoplasma prior to starting experiments. For all cell types, PRO-seq or GRO-seq was performed as exactly described[8,16], and sequenced using an Illumina Hi-Seq 2000 at the Cornell University Biotechnology Resource Center.

Comparison to ChromHMM and DNase-I data

We compared dREG TREs to ENCODE DNase-I and ChromHMM data. For ChromHMM data, we selected the set of sites annotated as promoter, enhancer, or insulator using data

from GM12878, K562[24], HeLa[40], or CD4+ T-cells[41]. We collected ENCODE DNase-I peak calls from the UW or Duke DNase-I-seq protocol[2], and selected peaks identified using both experimental assays. To compare different experimental assays, we merged sites identified by ChromHMM, DNase-I-seq, and dREG, and labeled each merged site based on the experimental assays which identified it. TREs were subsequently divided into four independent, non-overlapping classes based on the set of experimental peak calls that they intersected. Site classes were defined as those sites that intersected: (1) dREG, DNase-I-seq, and ChromHMM (+dREG), (2) ChromHMM insulators but not dREG (Insulator), (3) DNase-I-seq and ChromHMM, but not dREG (−dREG), and (4) ChromHMM, but not DNase-I-seq or dREG (Modified Chromatin Only; or MCO). All operations in these analyses were performed using the bedops[42], bedtools[43], or bigWig software packages.

Logistic Regression Classifier of DNase-I peaks with and without dREG

We used a logistic regression classifier to evaluate the how accurately transcription factors (TFs) could be used to distinguish between DNase-I peaks with and without the presence of dREG. We collected the set of all high confidence DNase-I peaks, consisting of the intersection between the UW and Duke assays. To improve our confidence about the transcription status of each DNase-I peak, we required that dREG-positive sites contain dREG scores greater than 0.8, and dREG-negative sites have scores less than 0.3.

We modeled the presence or absence of dREG at a particular DNase-I peak as the response in a logistic regression. Co-variants consist of the presence or absence of each TF assayed in the cell type of interest. To determine the presence or absence of each TF, we collected uniform peak calls for all ChIP-seq data from the ENCODE project. For MCF-7 cells, ENCODE data was supplemented with a set of 37 TFs for which ChIP-chip data was available[44]. TFs having multiple ChIP biological replicates were associated with each peak if any of the replicates was enriched at that peak. The significance of the direction of effect for each TF on the presence of a dREG signal was determined using a 1,000-sample bootstrap, in which we chose one TF at random to omit from the regression analysis during each iteration. Supplementary Fig. 13b plots the set of all TFs for each cell type whose direction of effect is consistent across each of the bootstrap iterations.

Identification of TF Binding using dREG, DNase-I and a joint model

We identified all occurrences of motifs associated with four transcription factors (NRF1, ELF1, MAX, and SP1) in hg19 using the PIQ program with the default log-odds score threshold of 5. Each position was classified as ‘bound’ or ‘unbound’ to the TF of interest using ENCODE ChIP-seq peak calls in the appropriate cell type. ROC plots profiling the accuracy of binding detection were collected by varying the max dREG score in a 200 bp window (treating unscored sites as a score of 0), DNase-I read counts in a 200 bp window around each putative motif matching the canonical PWM, or more stringent matches to the canonical TF motif. PIQ was run using the instructions provided by the authors. To evaluate the accuracy of PIQ we varied a threshold for the predicted positive-predictive value (PPV) output by the PIQ program at each site. We also evaluated a joint model which used data from dREG, PIQ PPV, the motif, and the absolute amount of Pol II mapping to the forward and reverse strand (within 200 bp), using each data source as a covariate in a logistic

regression, and modeling the presence of a ChIP-seq peak at each motif match as the response variable.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank I. Jonkers and N. Dukler for comments and helpful discussions on an early manuscript draft, and B. Gulko for critical discussions about support vector machines. This work was made possible by generous seed grants from the Cornell University Center for Vertebrate Genomics (CVG), Center for Comparative and Population Genetics (3CPG), an NHGRI grant (5R01HG007070-02) to AS and JTL, and NIH R01 (DK058110) to WLK. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

References

1. Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? Mol. Cell. 2013; 49:825–837. [PubMed: 23473601]
2. Bernstein BE, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
3. Giresi PG, Lieb JD. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). Methods. 2009; 48:233–239. [PubMed: 19303047]
4. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods. 2013; 10:1213–1218. [PubMed: 24097267]
5. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. The Functional Consequences of Variation in Transcription Factor Binding. PLoS Genet. 2014; 10:e1004226. [PubMed: 24603674]
6. Arnold CD, et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013; 339:1074–1077. [PubMed: 23328393]
7. Kim T-K, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010; 465:182–187. [PubMed: 20393465]
8. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008; 322:1845–1848. (80-.). [PubMed: 19056941]
9. Seila AC, et al. Divergent transcription from active promoters. Science. 2008; 322:1849–1851. [PubMed: 19056940]
10. Kapranov P, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 2007; 316:1484–1488. [PubMed: 17510325]
11. Hah N, et al. A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. Cell. 2011; 145:622–634. [PubMed: 21549415]
12. Hah N, Murakami S, Nagari A, Danko C, Kraus WL. Enhancer Transcripts Mark Active Estrogen Receptor Binding Sites. Genome Res. 2013
13. Andersson R, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014; 507:455–461. [PubMed: 24670763]
14. Melgar MF, Collins FS, Sethupathy P. Discovery of active enhancers through bidirectional expression of short transcripts. Genome Biol. 2011; 12:R113. [PubMed: 22082242]
15. Wu H, et al. Tissue-Specific RNA Expression Marks Distant-Acting Developmental Enhancers. PLoS Genet. 2014; 10:e1004610. [PubMed: 25188404]
16. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science. 2013; 339:950–953. [PubMed: 23430654]

17. Preker P, et al. RNA exosome depletion reveals transcription upstream of active human promoters. Science. 2008; 322:1851–1854. [PubMed: 19056938]
18. Kruesi WS, Core LJ, Waters CT, Lis JT, Meyer BJ. Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. Elife. 2013; 2:e00808. [PubMed: 23795297]
19. Core LJ, et al. Analysis of transcription start sites from nascent RNA supports a unified architecture of mammalian promoters and enhancers. Nat. Genet. 2014 In Press.
20. Drucker H, Burges CJC, Kaufman L, Smola AJ, Vapnik VN. Support Vector Regression Machines. Adv. Neural. Inf. Process. Syst. 1996:9.
21. Danko CG, et al. Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells. Mol. Cell. 2013; 50:212–222. [PubMed: 23523369]
22. Luo X, Chae M, Krishnakumar R, Danko CG, Kraus WL. Dynamic reorganization of the AC16 cardiomyocyte transcriptome in response to TNFα signaling revealed by integrated genomic analyses. BMC Genomics. 2014; 15:155. [PubMed: 24564208]
23. Wang IX, et al. RNA-DNA differences are generated in human cells within seconds after RNA exits polymerase. II. Cell Rep. 2014; 6:906–915. [PubMed: 24561252]
24. Ernst J, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011; 473:43–49. [PubMed: 21441907]
25. John S, et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat. Genet. 2011; 43:264–268. [PubMed: 21258342]
26. Boyle AP, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008; 132:311–322. [PubMed: 18243105]
27. Creyghton MP, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. U. S. A. 2010; 107:21931–21936. [PubMed: 21106759]
28. Guertin MJ, Martins AL, Siepel A, Lis JT. Accurate prediction of inducible transcription factor binding intensities in vivo. PLoS Genet. 2012; 8:e1002610. [PubMed: 22479205]
29. Heintzman ND, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. 2007; 39:311–318. [PubMed: 17277777]
30. Guertin MJ, Lis JT. Chromatin landscape dictates HSF binding to target DNA elements. PLoS Genet. 2010; 6:15.
31. Pique-Regi R, et al. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 2011; 21:447–455. [PubMed: 21106904]
32. Sherwood RI, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat. Biotechnol. 2014 advance on.
33. Neph S, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012; 489:83–90. [PubMed: 22955618]
34. Jolma A, et al. DNA-binding specificities of human transcription factors. Cell. 2013; 152:327–339. [PubMed: 23332764]
35. Lappalainen T, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501:506–511. [PubMed: 24037378]
36. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012; 22:1748–1759. [PubMed: 22955986]
37. Hadjur S, et al. Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature. 2009; 460:410–413. [PubMed: 19458616]

References (supplemental)

38. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. 2010
39. Chang C-C, Lin C-J. LIBSVM. ACM Trans. Intell. Syst. Technol. 2011; 2:1–27.

40. Hoffman MM, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013; 41:827–841. [PubMed: 23221638]
41. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. 2010; 28:817–825. [PubMed: 20657582]
42. Neph S, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012; 28:1919–1920. [PubMed: 22576172]
43. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]
44. Kittler R, et al. A comprehensive nuclear receptor network for breast cancer cells. Cell Rep. 2013; 3:538–551. [PubMed: 23375374]

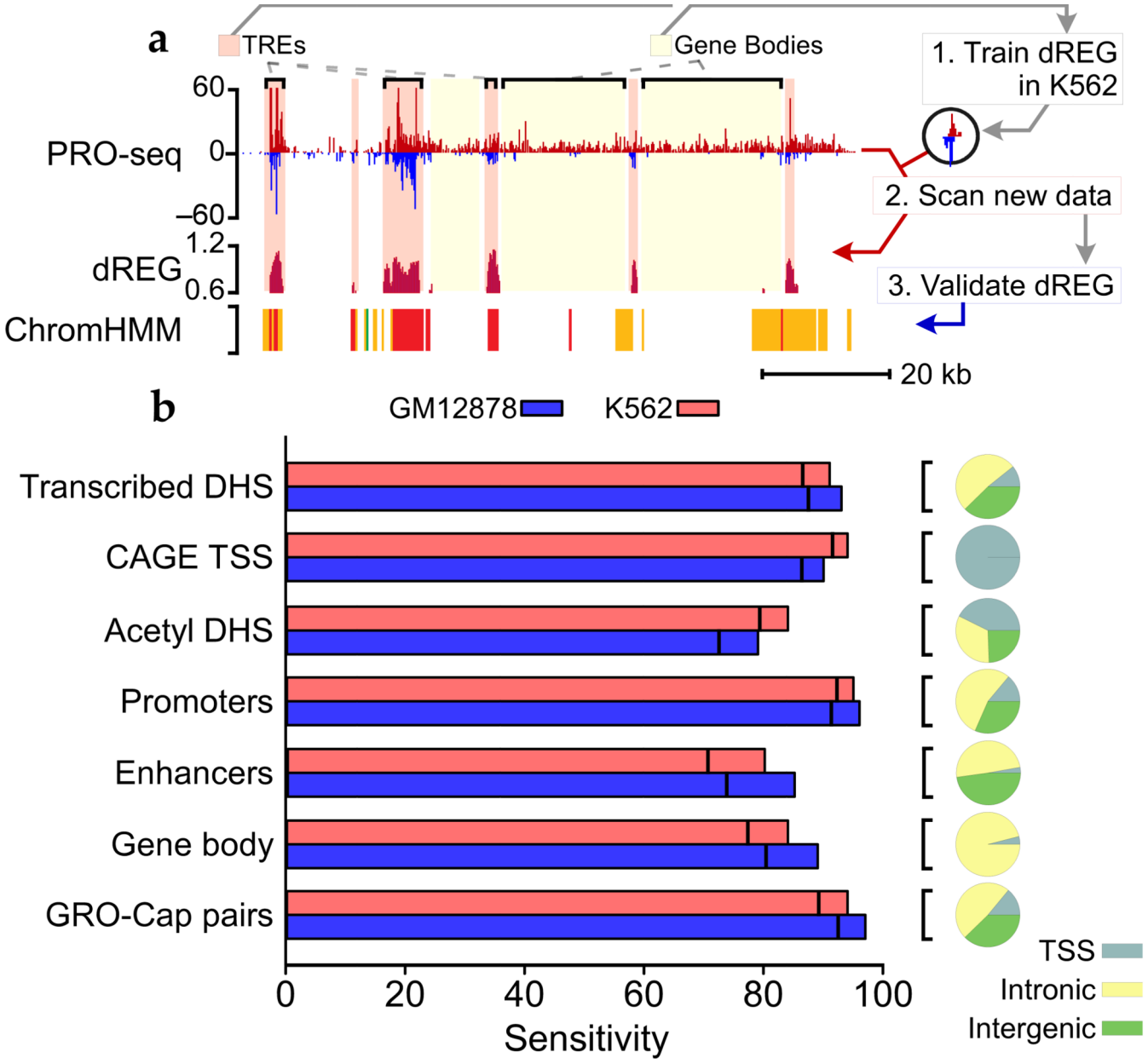

Figure 1.

dREG schematic and validation. (**a**) High PRO-seq signal intensity marks TREs (highlighted with pink background) and gene bodies (yellow background). dREG is a shape detector trained to recognize the characteristic pattern of TREs in PRO-seq data (#1). After training, dREG can be used to identify TREs using a new PRO-seq data set (red peaks) (#2). Browser shot compares dREG-predicted TREs to ChromHMM-predicted promoters (red), enhancers (yellow), and insulators (green) (#3). (**b**) Bar charts (left) represent the genome-wide sensitivity of dREG for various classes of TRE at a 5% (line) or 10% (bar) false discovery rate in K562 (pink) and GM12878 (blue) cells. Classes of regulatory elements represent GRO-cap transcribed DHS (Transcribed DHS), transcription start sites identified by CAGE (CAGE TSS), histone acetylation associated with DHS (Acetyl DHS), GRO-cap transcribed

ChromHMM promoters (Promoters), GRO-cap transcribed chromHMM enhancers (Enhancers), GRO-cap TSS inside annotated Gene Bodies (Gene Body), and GRO-cap pairs (GRO-cap Pairs). Pie charts (right) represent the fraction of sites aligning within RefSeq transcription start sites (TSS), introns, or intergenic regions in each validation set.

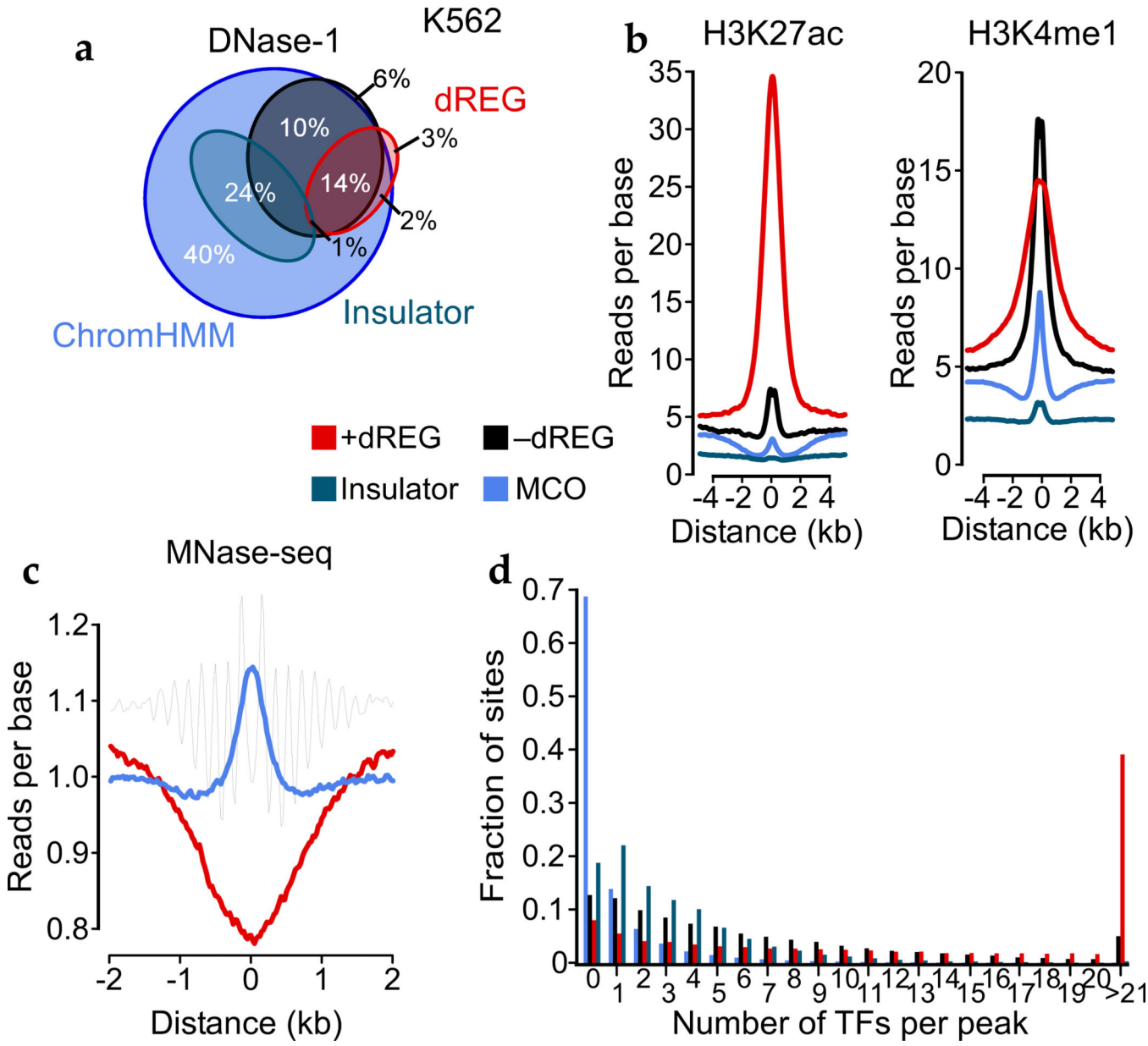

Figure 2.

Comparison of putative TREs detected using dREG, DNase-I, and ChromHMM. (**a**) Four-way Venn diagram depicting the relationships among separate genomic assays, which support the existence of four distinct classes of regulatory element. Numbers give the rounded overall fraction of TREs that fall into the specified intersection. TREs discovered using multiple assays were classified as +dREG, −dREG, Insulator, or as modified chromatin only (MCO). (**b**) Comparison of read-densities for H3K27ac (left) and H3K4me1 (right) in each class of functional element. (**c**) Distribution of MNase-seq reads in the +dREG (red), −dREG (black), and modified chromatin only classes (MCO; blue). (**d**) Histogram compares the number of transcription factors found in each of the four functional classes.

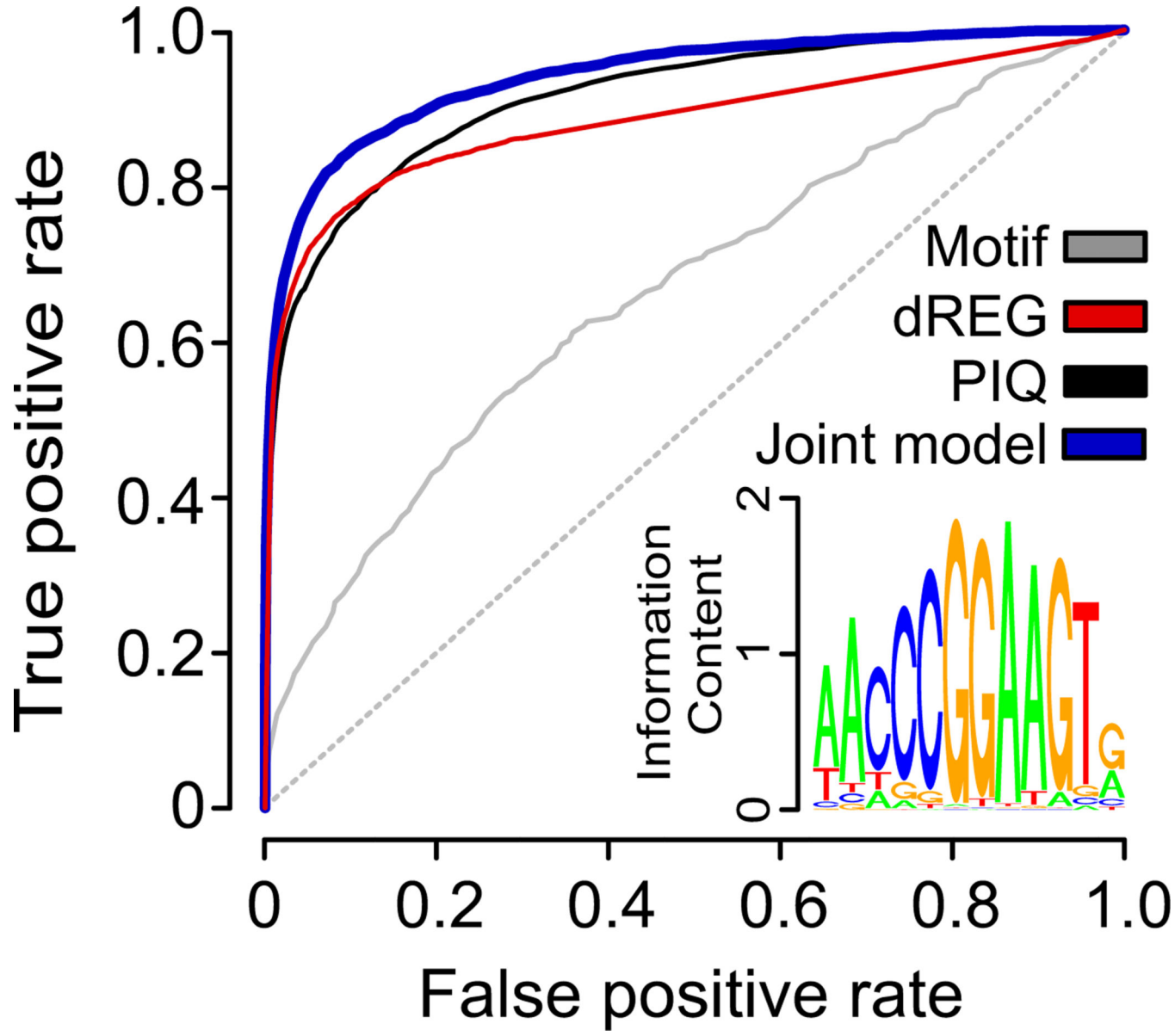

Figure 3.

Sequence-specific transcription factors identified using dREG transcribed TREs. ROC plot shows the accuracy of predicting ELF1 binding to strong matches to the ELF1 consensus binding motif (sequence logo shown) using PIQ (black; AUC= 0.92), dREG (red; AUC= 0.88), the DNA sequence motif (gray; AUC= 0.67), or a joint logistic regression model considering all three variables (blue; AUC= 0.94). Motif matches that intersect ENCODE ChIP-seq peak calls were used as the set of true binding sites.

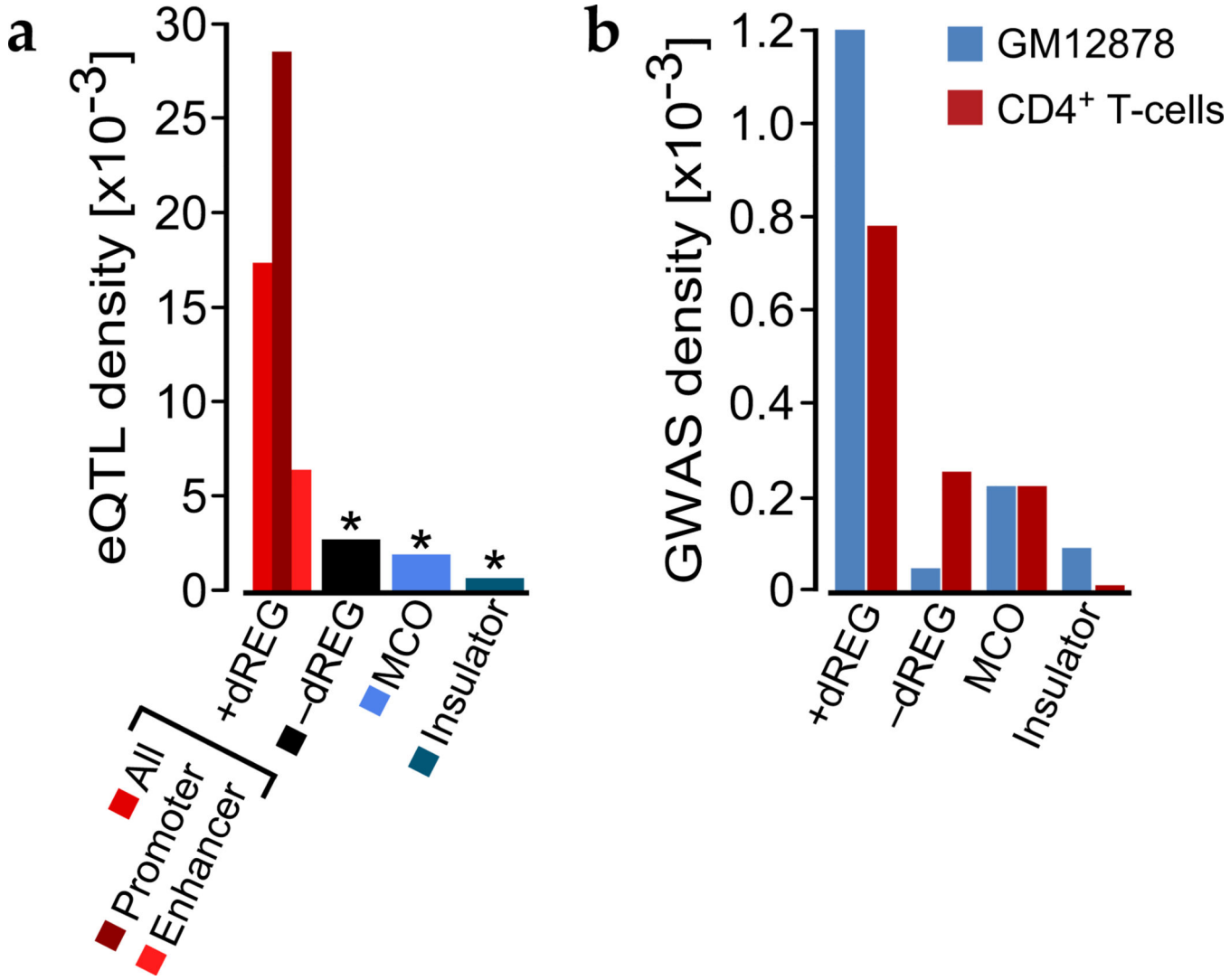

Figure 4.

eQTL and GWAS SNP enrichments in the four classes of functional element. (**a**) The density of eQTL (n= 755) per site found in +dREG (further divided into promoters and enhancers using ChromHMM), −dREG, modified chromatin only (MCO), and Insulator classes. The asterisk indicates significantly lower eQTL densities than in dREG enhancers by a Fisher's exact test ($P<2\times10^{-5}$). (**b**) The density of GWAS SNPs that correlate with cell-type specific phenotypes (autoimmune disorders) in GM12878, a B-cell line (blue), and primary CD4+ T-cells (red).