Identification of regulatory elements from nascent transcription using dREG

Zhong Wang,[1] Tinyi Chu,[1,2] Lauren A. Choate,[1] and Charles G. Danko[1,3]

[1] *Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, New York 14853, USA;* [2] *Graduate Field of Computational Biology, Cornell University, Ithaca, New York 14853, USA;* [3] *Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, New York 14853, USA*

Our genomes encode a wealth of transcription initiation regions (TIRs) that can be identified by their distinctive patterns of actively elongating RNA polymerase. We previously introduced dREG to identify TIRs using PRO-seq data. Here, we introduce an efficient new implementation of dREG that uses PRO-seq data to identify both uni- and bidirectionally transcribed TIRs with 70% improvement in accuracy, three- to fourfold higher resolution, and >100-fold increases in computational efficiency. Using a novel strategy to identify TIRs based on their statistical confidence reveals extensive overlap with orthogonal assays, yet also reveals thousands of additional weakly transcribed TIRs that were not identified by H3K27ac ChIP-seq or DNase-seq. Novel TIRs discovered by dREG were often associated with RNA polymerase III initiation, bound by pioneer transcription factors, or located in broad domains marked by repressive chromatin modifications. Our results suggest that transcription initiation can be a powerful tool for expanding the catalog of functional elements.

[Supplemental material is available for this article.]

Our genomes encode a wealth of distal and proximal control regions that are collectively known as transcriptional regulatory elements. These regulatory DNA sequence elements regulate gene expression by affecting the rates of a variety of necessary steps during the RNA polymerase II (Pol II) transcription cycle (Fuda et al. 2009), including chromatin accessibility, transcription initiation, and the release of Pol II from a paused state into productive elongation.

Identifying regulatory elements at a genome scale has recently become a subject of intense interest. Regulatory elements are generally identified using genome-wide molecular assays that provide indirect evidence that a particular locus is associated with regulatory activity. For example, nucleosomes tagged with post-translational modifications can be identified by chromatin immunoprecipitation and sequencing (ChIP-seq) (Barski et al. 2007; Heintzman et al. 2007). Likewise, nucleosome-free DNA can be enriched using DNase I or Tn5 transposase (Boyle et al. 2008; Hesselberth et al. 2009; Buenrostro et al. 2013). However, each of these strategies has important limitations. Histone modification ChIP-seq has a poor resolution compared with the ~110-bp nucleosome-free region that serves as the regulatory element core (Core et al. 2014; Scruggs et al. 2015; Chen et al. 2016). Likewise, nuclease accessibility assays mark a variety of nuclease-accessible regions in our genomes, such as binding sites for the insulator protein CTCF or inactive regulatory elements, without the capacity to distinguish between these types of functional elements (Xi et al. 2007; Danko et al. 2015). Each of these tools is also limited by a high background, which prevents the detection of weakly active regulatory elements which may nevertheless have important functional roles.

Transcription initiation has recently emerged as an alternative mark for the location of active regulatory elements (Andersson et al. 2014a; Core et al. 2014; Danko et al. 2015). Both proximal and distal regulatory elements are associated with RNA polymerase initiation (Kim et al. 2010; Core et al. 2014; Andersson et al. 2015; Henriques et al. 2018; Mikhaylichenko et al. 2018). RNAs produced at these elements are often degraded rapidly by the nuclear exosome complex (Andersson et al. 2014b; Core et al. 2014), and as a result, these patterns are most reliably detected by nascent RNA sequencing techniques that map the genome-wide location of RNA polymerase itself (Core et al. 2008; Churchman and Weissman 2011; Kwak et al. 2013; Scruggs et al. 2015). Transcription leaves a characteristic signature at these sites that can be extracted from nascent RNA sequencing data using appropriate computational tools (Melgar et al. 2011; Hah et al. 2013; Danko et al. 2015; Azofeifa and Dowell 2016).

We recently introduced dREG (Danko et al. 2015), a sensitive machine learning tool for the detection of regulatory elements using maps of RNA polymerase derived from run-on and sequencing assays, including GRO-seq (Core et al. 2008), PRO-seq (Kwak et al. 2013), and ChRO-seq (Chu et al. 2018). dREG was trained to recognize characteristic signatures of nascent RNAs to accurately discover the coordinates of regulatory elements genome-wide. However, our preliminary version of dREG was limited by a slow and cumbersome implementation that made it challenging to use in practice. Here, we present an efficient new implementation of dREG that leverages a general purpose graphical processing unit to accelerate computation. Our new version of dREG is available to the community by a public web server at https://dreg.dnasequence.org/.

Results

A new machine learning tool for the discovery of TIRs

We recently introduced a machine learning tool for the detection of regulatory elements using GRO-seq and other run-on and

Corresponding author: dankoc@gmail.com
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.238279.118.

sequencing assays (dREG) (Danko et al. 2015). Here, we introduce a new implementation of dREG which makes several important optimizations to identify regulatory elements with improved sensitivity and specificity using the multiscale feature vector introduced in dREG. We implemented dREG on a general purpose graphical processing unit (GPU) using Rgtsvm (Wang et al. 2017). Our GPU implementation decreased run-times by >100-fold, allowing analysis of data sets which took 30–40 h using 32 threads in the CPU-based version of dREG to be run in under an hour.

We used the speed of our GPU-based implementation to train a new support vector regression (SVR) model that improved dREG accuracy. We trained dREG using 3.3 million sites obtained from five independent PRO-seq or GRO-seq experiments in K562 cells (Supplemental Fig. S1; Supplemental Table S1). To improve the accuracy of dREG predictions in the unbalanced setting typical for genomic data, where negative examples greatly outnumber positive examples, dREG was trained on a data set where bona fide positive regulatory elements represent just 3% of the training data. Together, these improvements in the composition and size of the training set increased the area under the precision-recall curve by 70% compared with the original dREG model when evaluated on two data sets that were held out during training (Supplemental Fig. S2).

We developed a novel strategy to identify regions enriched for dREG signal, which we call transcription initiation regions (TIRs), and filter these based on statistical confidence (see Methods; Fig. 1A; Supplemental Fig. S3). We estimate the probability that dREG scores were drawn from the negative class of sites (i.e., non-TIRs) by modeling dREG scores using the Laplace distribution. The Laplace distribution was used to model SVR scores previously (Lin and Weng 2004) and fits dREG scores in negative sites reasonably well (Supplemental Fig. S4). To improve our statistical power to identify bona fide regulatory elements, we merge nearby candidate sites into nonoverlapping genomic intervals, or candidate TIRs, each of which contains approximately one divergently oriented pair of paused RNA polymerases (Core et al. 2014; Scruggs et al. 2015). We compute the joint probability that five positions within each TIR are all drawn from the negative (nonregulatory element) training set using the covariance between adjacently positioned dREG scores (see Methods). This novel peak calling strategy provides a principled way to filter the location of TIRs based on SVR scores estimated using dREG.

Comparison to orthogonal genomic data

To evaluate the performance of dREG in real-world examples, we analyzed three data sets in K562 (PRO-seq), GM12878 (GRO-seq), and HCT116 (GRO-seq) that were held out during model training. Holdouts were selected because they cover a range of library sequencing depths and new cell types that together allowed us to determine whether the dREG model generalized to additional data sets. dREG predicted 34,631, 71,097, and 62,934 TIRs in K562, GM12878, and HCT116, respectively. dREG recovered the location of the majority of regulatory elements defined using orthogonal strategies at an estimated 5% false discovery rate: 81.3% or 96.1% of DNase I hypersensitive sites (DHSs) marked by transcription (using GRO-cap pairs) and 58.4%, 71.8%, or 84.9% of DHSs marked by the acetylation of histone 3 lysine 27 (H3K27ac) (Fig. 1B). Sensitivity for both GRO-cap and H3K27ac-DHSs was greater than twofold higher for dREG than for the elegant model-based Tfit program (Azofeifa et al. 2018) when run on the same data. Transcription initiation regions display a range in the efficiency of initiation on the two strands (Duttke et al. 2015; Scruggs et al. 2015), and dREG was able to identify the location of both uni- and bidirectional transcription initiation sites (Supplemental Fig. S5).

Extending dREG analysis to 14 data sets in six cell types, we found that the sensitivity of dREG varied systematically by the library sequencing depth (Fig. 1C). dREG achieved a reasonable sensitivity on a K562 holdout data set with 27 M uniquely mapped reads (81.3% of DHSs overlapping GRO-cap pairs were recovered)

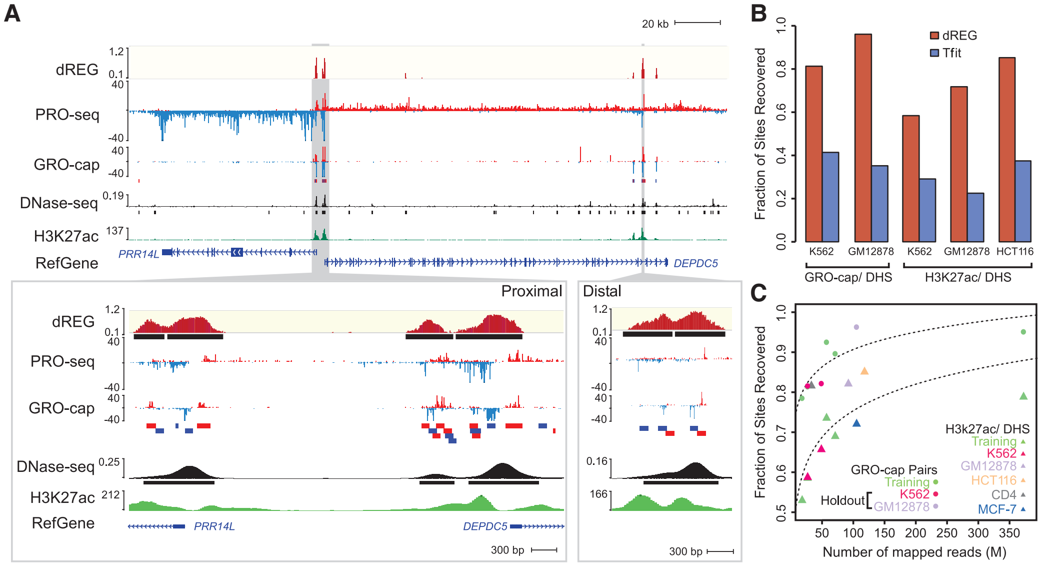

Figure 1. dREG identifies regions of transcription initiation. (*A*) WashU Epigenome Browser visualization of dREG signal, PRO-seq data, GRO-cap, DNase-seq, and H3K27ac ChIP-seq near the *PRR14L* and *DEPDC5* genes. *Inserts* (cf. gray shaded pointers) show an expanded view of gene-proximal promoter elements (*left*) and a distal enhancer (*right*), each encoding multiple transcription initiation sites. (*B*) Bar plots show the fraction of transcribed DHSs (*left*) and H3K27ac+ DHSs (*right*) that were discovered by dREG (red) and Tfit (blue) in holdout data sets. (*C*) Scatterplot shows the fraction of sites recovered (*y*-axis) as a function of sequencing depth (*x*-axis) for 12 data sets shown in Supplemental Table S1. The best fit lines are shown. The color represents whether the data set was used for training (green) or is a holdout data set (K562, red) or cell type (GM12878, lavender; HCT116, orange; CD4+ T-cells, gray; MCF-7, blue).

and saturated the discovery of enhancers supported by ENCODE data at between 60 and 100 M uniquely mapped reads. After accounting for sequencing depth, we did not observe any systematic difference between data sets that were held out or used during training, suggesting that dREG was not noticeably overfitting to the training data. We did not notice any systematic bias in sensitivity for either PRO-seq or GRO-seq data, for any specific cell type or based on the lab or origin (Fig. 1B,C; Supplemental Fig. S6A,B). Finally, we also obtained reasonably good performance using dREG to analyze publicly available mNET-seq data in HeLa cells (Supplemental Fig. S6A,C; Mayer et al. 2015; Nojima et al. 2015). These results suggest that our new dREG model is highly extensible to nascent transcription data from a variety of different sources.

Despite a high degree of overlap with histone modification ChIP-seq assays, dREG had a higher resolution for the regulatory element core region, consisting of divergently opposing RNA polymerase initiation sites (Core et al. 2014). Regions identified by dREG were on average 6.4-fold shorter (460 bp for dREG sites) than H3K27ac ChIP-seq peaks (2924 bp on average), closer in size to high-resolution DNase-seq data (322 bp on average) (Fig. 2A). dREG frequently separated out individual TIRs in clusters of initiation sites that could not be distinguished based on histone modification ChIP-seq peak calls, for instance, in the *MYC* enhancer locus (Fig. 2B; Fulco et al. 2016). Histone modification ChIP-seq or DNase-seq data aligned to the center of human dREG sites revealed good agreement with the center of the nucleosome-free region (Fig. 2C). Thus, our new dREG implementation substantially improved both resolution and accuracy compared with alternative genomic tools.

Discovery of novel regulatory elements using dREG

Despite a high degree of overlap, up to 10% of TIRs did not overlap other marks expected at active enhancers. The number of TIRs found uniquely by dREG depended on sequencing depth (400–8000 TIRs, depending on the data set) and did not saturate even in data sets sequenced to a depth of 350 M uniquely mapped reads (Fig. 3A). As expected, TIRs had lower dREG scores and lower polymerase abundance when they were found uniquely by dREG (Supplemental Fig. S7), suggesting that these sites were often either weaker regulatory elements that were more difficult for all assays to distinguish from background or false positives.

We asked whether TIRs that were not supported by DNase-seq or H3K27ac ChIP-seq peak calls reflect bona fide novel regulatory elements or false positives by dREG. TIRs detected uniquely by dREG frequently (>50% depending on the data set) overlapped ChIP-seq peak calls for sequence-specific transcription factors

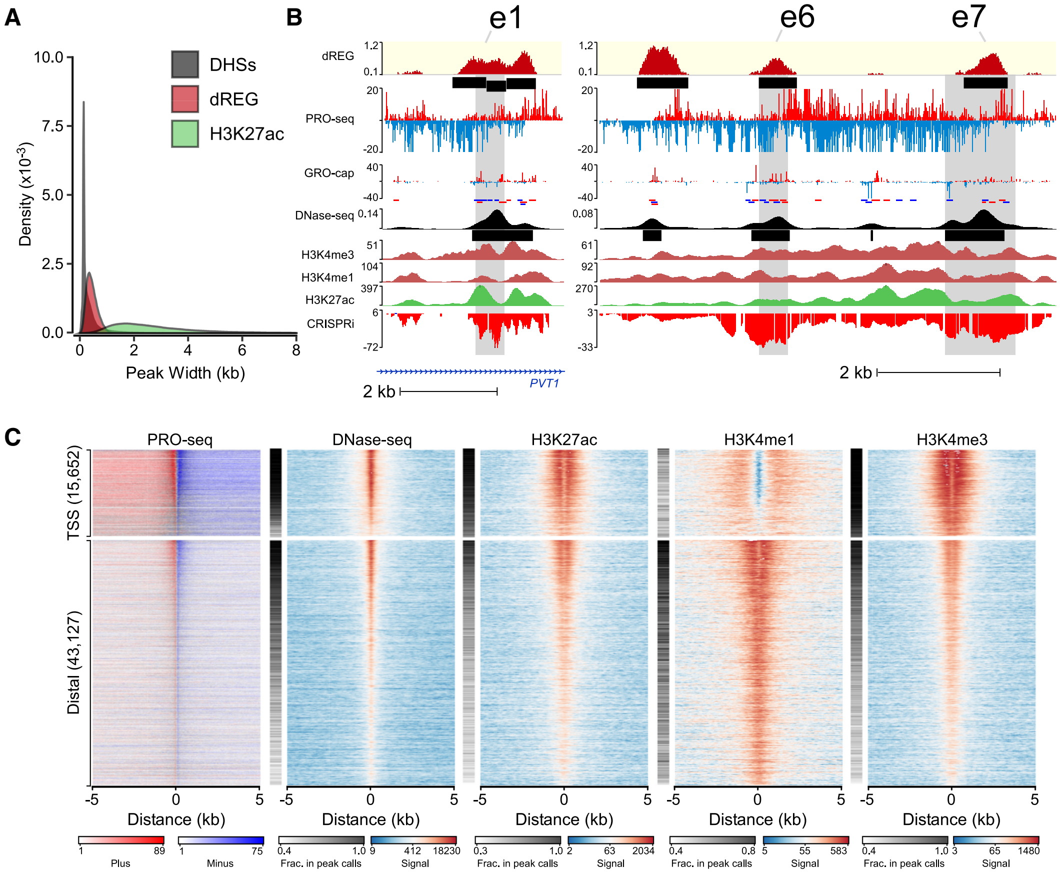

Figure 2. dREG calls are often concordant with other molecular assays. (*A*) Histogram shows the size distribution of dREG TIRs, H3K27ac ChIP-seq peaks, or DHSs. (*B*) WashU Epigenome Browser visualization of dREG signal, PRO-seq data, GRO-cap, DNase-seq, H3K4me3, H3K4me1, H3K27ac ChIP-seq, and CRISPR interference score (CRISPRi) at three enhancers (e1, e6, and e7) that affect transcription of *MYC* in K562 cells based on CRISPR interference (CRISPRi). (*C*) Heat maps show the log-signal intensity of PRO-seq, DNase-seq, or ChIP-seq for H3K27ac, H3K4me1, and H3K4me3. The fraction of sites intersecting ENCODE peak calls is shown in the white-black color map *beside* each plot. Color scales for signal and the fraction in peak calls are shown *below* the plot. Each row represents TIRs found overlapping an annotated transcription start site ($n = 15{,}652$) or >5 kb to a start site ($n = 43{,}127$).

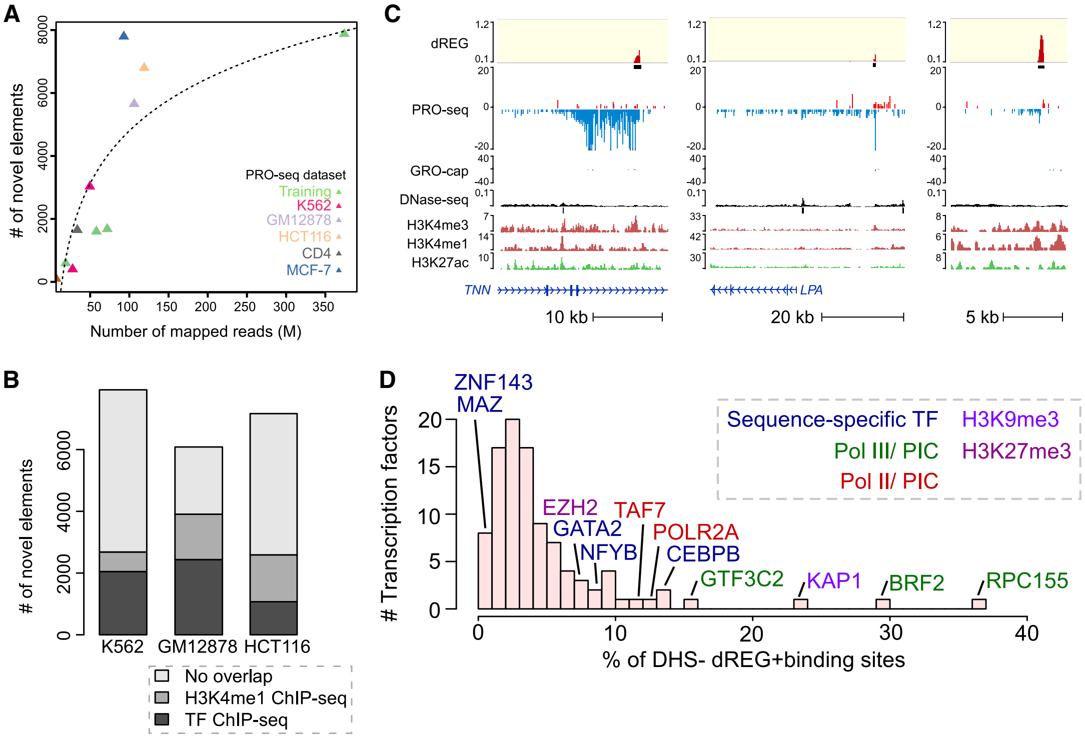

Figure 3. dREG identifies new regions that were not found using other molecular assays. (*A*) Scatterplot shows the number of new TIRs that were not discovered in DNase-seq or H3K27ac ChIP-seq data (*y*-axis) as a function of sequencing depth (*x*-axis) for seven data sets shown in Supplemental Table S1. The best fit line is shown. The color represents whether the data set was used for training (green) or is a holdout data set (K562, red) or cell type (GM12878, lavender; HCT116, orange; CD4^{+} T-cells, gray; MCF-7, blue). (*B*) Stacked bar charts show the number of elements discovered using dREG, but not found in DNase-seq or H3K27ac ChIP-seq (*y*-axis) for PRO-seq or GRO-seq data sets in K562, GM12878, and HCT116 cells. The color denotes other functional marks intersecting sites discovered only using dREG. (*C*) Three separate genome browser regions that denote TIRs discovered using dREG, but were not found in DNase-seq or H3K27ac ChIP-seq data. Tracks show dREG signal, PRO-seq data, GRO-cap, DNase-seq, H3K27ac ChIP-seq, and annotated genes. (*D*) Histogram representing the fraction of binding sites for 100 transcription factors supported by a dREG TIR that was not also discovered in DNase-seq data. Several of the outliers are shown. The color denotes whether the factor is a member of the RNA polymerase III (Pol III) preinitiation complex (green), Pol II preinitiation complex (red), associated with H3K9me3 (light purple), or H3K27me3 heterochromatin (purple), or is a sequence-specific transcription factor (blue).

(Fig. 3B; Supplemental Fig. S8). A small number of TIRs were enriched for H3K4me1, a mark associated with both active and inactive enhancers. Examining examples on the WashU Epigenome Browser (Zhou et al. 2011) revealed clearly defined transcription units that initiate long intergenic noncoding RNAs (Fig. 3C). Often the promoter of these transcription units lacked sufficiently robust enrichment of histone modifications or DNase-seq signal to make confident peak calls, and many lacked sufficient paused Pol II to be represented in GRO-cap data (Fig. 3C; Supplemental Fig. S9). Nevertheless, examination of these TIRs genome-wide revealed a local increase in the abundance of reads in the average profiles of active histone modification ChIP-seq data (Fig. 2C; Supplemental Fig. S10), suggesting that at least some were false negatives by other assays. Finally, sites detected only by dREG in K562 cells were often DHSs in a related cell type (Supplemental Fig. S11). Taken together, these findings suggest that TIRs uniquely identified by dREG were frequently novel regulatory elements but were enriched below the level of detection of other molecular assays in K562 cells.

Transcription factor binding predicts DHS status

An alternative, but not mutually exclusive, explanation for TIRs identified uniquely by dREG is that some regulatory elements tolerate differences in the core marks reported to correlate with regulatory function. We hypothesized that certain transcription factors are more tolerant of deviations from the core regulatory architecture than others. We focused on DNase-seq as a general marker for the nucleosome-depleted region in the center of regulatory elements. As a control for differences between K562 clones, growth conditions, or cell handling, we performed ATAC-seq to confirm low levels of chromatin accessibility in our own K562 cell stocks, closely related to those used to generate PRO-seq data (Supplemental Fig. S12).

To determine whether specific transcription factors may be more permissive to binding in sites having low levels of chromatin accessibility, we trained a logistic regression model to predict whether TIRs discovered using dREG intersect a DHS. Transcription factor binding site ChIP-seq data alone predicted the presence of DHSs better than using the dREG score in a matched set of holdout sites (ROC = 0.88 [TF binding], ROC = 0.75 [dREG score]) (Supplemental Fig. S13). Thus, ChIP-seq data for specific transcription factors was predictive of which TIRs lacked nuclease hypersensitivity.

To identify transcription factors that contribute to this signal, we computed the ratio of ChIP-seq peak calls that were found using dREG but not DNase-seq to those that were found using both assays (referred to as dREG+DHS−/dREG+DHS+). As expected, only a small fraction of most transcription factors were bound without creating a DHS (Fig. 3D). However, different transcription factors exhibited a broad range of binding in dREG+DHS− sites. The highest scoring outliers were frequently members of the core Pol II and Pol III transcription machinery (e.g., RPC155, BRF2, CHD1, POLR2A, and TAF7), consistent with PRO-seq detecting transcription more directly than DNase-seq and potentially suggesting that some bona fide transcription initiation sites were not sensitive to DNase I.

Pol III transcription initiation without chromatin accessibility

Transcription factors with the largest fraction of ChIP-seq peaks in dREG+DHS− sites were RPC155 and BRF2 (ratio of dREG+DHS −/dREG+DHS+ = 0.37 and 0.29, respectively), which encode the catalytic core of RNA polymerase III and a Pol III initiation factor. If a fraction of dREG+DHS− TIRs were explained by Pol III initiation, we expected to find a structured combination of DNA sequence motifs at these TIRs that were reported in canonical Pol III promoters (James Faresse et al. 2012). Indeed, the TATA and PSE DNA sequence elements were enriched with the correct spacing in dREG+DHS− TIRs compared to TIRs that intersect POLR2A (Pol II) ChIP-seq data ($P < 1 \times 10^{-5}$) (Supplemental Fig. S14). dREG +DHS− TIRs were enriched for Pol III promoter motifs to a similar magnitude as TIRs bound by RPC155, the core subunit of Pol III, based on ChIP-seq (Supplemental Fig. S14). These observations suggest that some Pol III promoters were not sufficiently exposed to the DNase I enzyme to be detected in DNase-seq data.

Heterochromatin domains frequently harbor TIRs

We found that dREG+DHS− TIRs were often associated with ChIP-seq for heterochromatin markers (e.g., KAP1 and EZH2) (Fig. 3D). We found 6375 dREG TIRs that overlapped heterochromatin-associated ChromHMM states in K562 cells (Polycomb-repressed and heterochromatin; low signal [Ernst et al. 2011]). In total, 55% of TIRs overlapping heterochromatin regions were not found by DNase-seq, a significant enrichment compared with all TIRs ($P < 2.2 \times 10^{-16}$, Fisher's exact test).

Next, we examined TIRs in H3K27me3 domains. Broad H3K27me3 domains frequently harbored several TIRs (Fig. 4A). Often these TIRs were supported by GRO-cap signal, suggesting that they were not false positives. H3K27me3 domains contained a median of ∼1 TIR per 50 kb of contiguous H3K27me3 (Fig. 4B). TIRs in H3K27me3 domains generally had lower levels of transcription (Fig. 4C), consistent with a causal role for H3K27me3 in reducing transcriptional activity (Hosogane et al. 2016; Coleman and Struhl 2017). Despite the lower levels of transcription, nearly 25% of TIRs in H3K27me3 domains were also supported by ChIP-seq for POLR2A (Pol II), RPC155 (Pol III), or other transcription factors. While overlap with POLR2A ChIP-seq was depleted in H3K27me3 domains as expected, RPC155 ChIP-seq was enriched by more than 40% ($P = 1 \times 10^{-8}$, Fisher's exact test), potentially suggesting that Pol III initiation may be less affected by H3K27me3 than Pol II. TIRs in H3K27me3 domains also frequently overlapped active histone marks, especially H3K4me3 and H3K4me1. Taken together, our results are consistent with recent reports that transcription start sites within heterochromatin can escape repression (Leemans et al. 2018).

Transcription factors have distinct enrichments of chromatin marks in DHS− TIRs

Several sequence-specific transcription factors were also observed to have a high fraction of sites that were dREG+DHS−. For

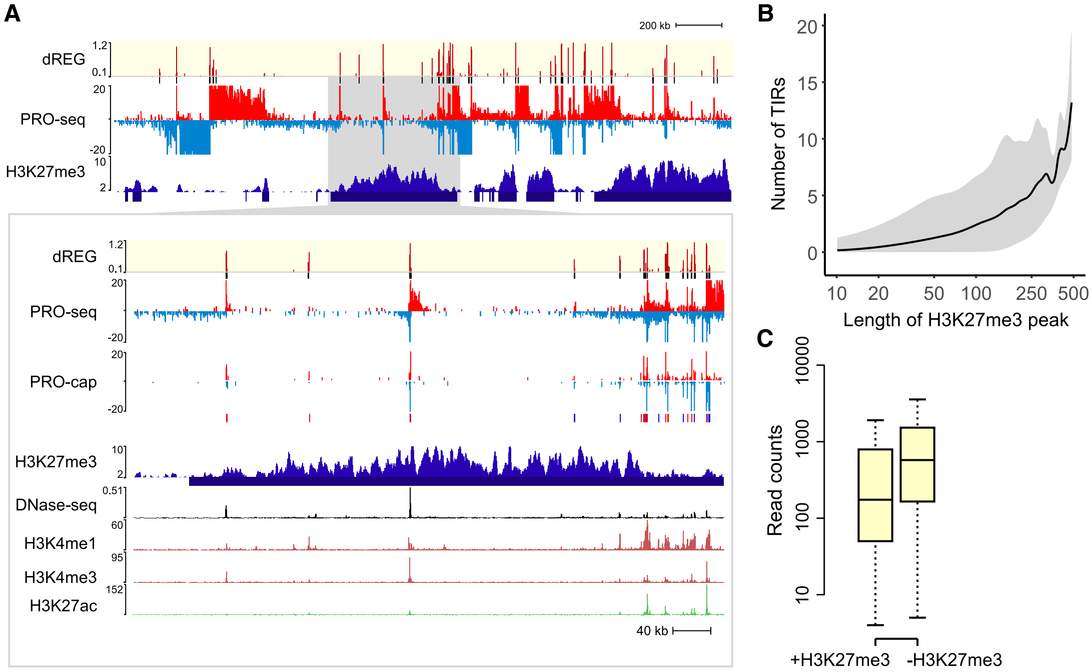

Figure 4. dREG TIRs located in H3K27me3 domains. (*A*) WashU Epigenome Browser visualization of dREG signal, PRO-seq data, GRO-cap, H3K27me3 ChIP-seq, DNase-seq, and H3K4me1, H3K4me3, and H3K27ac ChIP-seq. The *insert* (cf. gray shaded pointer) shows an expanded view of the H3K27me3 domain encoding multiple transcription initiation sites that were also supported in GRO-cap data. (*B*) The number of TIRs discovered in each H3K27me3 broad peak as a function of H3K27me3 peak size. The line represents the median, and gray shading denotes the fifth and 95th percentile. The *x*-axis is a log scale. (*C*) The box plot shows the difference in PRO-seq read counts between TIRs in an H3K27me3 peak call (+H3K27me3, *left*) and outside of an H3K27me3 peak call (−H3K27me3, *right*). The *y*-axis represents the number of reads found within 250 bp of each TIR.

example, CEBPB, NFYB, GATA2, and SPI1 had a relatively high fraction of binding sites outside of DHSs. The subset of DHS+ and DHS− binding sites for these four transcription factors had distinct profiles in the flanking chromatin. All four transcription factors were enriched for increased MNase-seq read density centered on the binding site and spanning a region ~300 bp in DHS− sites (Fig. 5), suggesting systematic differences in the chromatin environment in these regions. In contrast, binding sites for MAZ and ZNF143, which exhibited a low fraction of binding sites outside of DHSs, did not show as prominent an increase in MNase-seq signal in DHS− binding sites (Fig. 5).

Transcription factors also showed differences in their enrichment of histone post-translational modifications. NFYB exhibited no enrichment of active histone modifications in DHS− binding sites but was flanked on both sides by high levels of H3K27me3 (Fig. 5). GATA2, SPI1, and CEBPB binding sites were enriched for marks of both active and repressive chromatin, with a narrow enrichment of H3K27me3 signal localized at the putative binding site (Fig. 5). Likewise, histone modification ChIP-seq in DHS− regions lacked the dip in the center of TIRs characteristic of a nucleosome-depleted region. Thus, in some cases regulatory elements discovered by dREG, but not by DNase-seq, appear to reflect binding of strong transcriptional activators that do not meet the current description of a regulatory element.

Taken together, these results suggest that dREG identified thousands of TIRs that were not discovered using DNase-seq data but which were reproducibly associated with specific transcription factors. These observations may reflect transcription factor binding events that tolerate deviations from the core TIR architecture, preventing their discovery using more widely applied molecular tools. Collectively, these observations suggest that no molecular assay has fully saturated the repertoire of active regulatory elements, even in well-studied cell types like K562.

Web server access to dREG

We developed a web interface for users to run dREG on their own PRO-seq, GRO-seq, or ChRO-seq data. Users upload PRO-seq data as two bigWig files representing raw counts mapped to the plus and minus strand. A typical run takes ~4–12 h depending on server usage. Users are required to register for an account, which keeps track of previous jobs. Once dREG completes successfully, users can download dREG peak calls and raw dREG signal. Additionally, the dREG web interface provides a link to visualize input PRO-seq data, dREG signal, and dREG peak calls as a private track hub on the WashU Epigenome Browser. dREG is available on the Extreme Science and Engineering Discovery Environment (XSEDE) as a science gateway (Gesing et al. 2017; Knepper et al. 2017) and is implemented using the Airavata middleware (Marru et al. 2011; Pierce et al. 2015). The dREG science gateway is available at https://dreg.dnasequence.org/.

Discussion

In this article, we have introduced an optimized version of our dREG software package, a sensitive machine learning method that identifies the location of regulatory elements using data from run-on and sequencing assays, including PRO-seq, GRO-seq, and ChRO-seq (Core et al. 2008; Kwak et al. 2013; Chu et al. 2018). Our optimization efforts have achieved substantial improvements in computational efficiency, sensitivity, specificity, and site resolution. We developed a new approach to identify dREG peaks, called transcription initiation regions, based on a hypothesis testing framework that controls false discovery rates. Finally, we provide dREG as a web service where users can easily upload their own run-on and sequencing data.

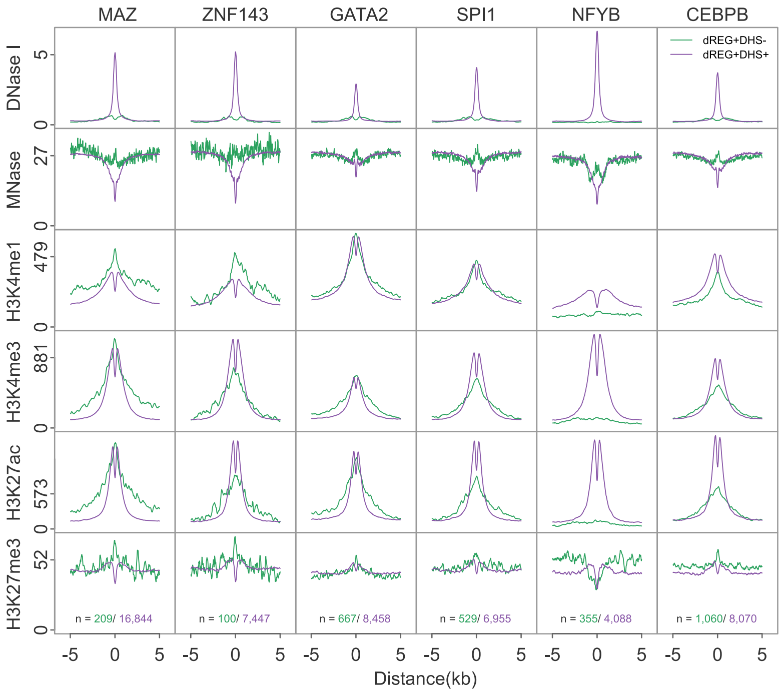

Figure 5. dREG TIRs with specific transcription factor binding show distinct chromatin marks. Metaplots show the raw signal of DNase-seq, MNase-seq, and ChIP-seq for H3K4me1, H3K4me3, H3K27ac, and H3K27me3 near binding sites for six transcription factors, including MAZ, ZNF143, GATA2, SPI1, NFYB, and CEBPB. Signals are shown for dREG+DHS− (green) and dREG+DHS+ (purple) sites. The number of sites contributing to each signal is shown (*bottom*).

Taken together, our dREG implementation has a number of advantages compared with alternative approaches. dREG offers substantial improvements in resolution for transcription factor binding sites, which tend to be located between divergently initiating RNA polymerase (Core et al. 2014). Likewise, dREG provides information about local patterns of transcription initiation, improved signal to noise ratio, and a higher sensitivity for certain types of active regulatory elements. Compared with GRO-cap (Core et al. 2014), dREG is less dependent on paused Pol II and can also be used to detect the levels of gene transcription in the same molecular assay. Most importantly, dREG/PRO-seq allows users to measure multiple aspects of gene regulation, including the precise position of regulatory elements, gene expression, and pausing levels using a single genomic experiment. When paired with ChRO-seq (Chu et al. 2018), which applies run-on assays in solid tissues, dREG allows the discovery of regulatory elements in primary tumors and other clinical isolates, in which the application of genomics technologies are limited by sample quantity and the cost of applying multiple assays across large cohorts.

By comparing TIRs to other functional genomic assays, we identified >8000 regulatory elements that were not detected using DNase-seq or H3K27ac ChIP-seq. Differences between assays may in part reflect false negatives in DNase-seq and ChIP-seq, where signals drop below the background level, or false positives by dREG. Several lines of evidence outlined in our results suggest that most TIRs are unlikely to reflect false positives. For instance, we observed a residual enrichment in the average profiles of other functional marks near TIRs that lack peak calls, which suggests that at least some fraction of TIRs reflect weak enrichment in other molecular assays that were not detected as peaks. Our results may contribute additional support to experiments assigning regulatory function to rare sites which lack canonical promoter and enhancer marks (Rajagopal et al. 2016; Diao et al. 2017). Nascent transcription may be an effective tool to expand the catalog of functional elements.

TIRs may also reflect weakly bound transcriptional activators that are relatively tolerant of binding to sites lacking DNase-seq. Indeed, several of the transcription factors with a relatively large fraction of dREG+DHS− binding sites were identified as having pioneer factor activity, including GATA2, SPI1, NFYB, and CEBPB (Heinz et al. 2010; Grøntved et al. 2013; Barozzi et al. 2014; Sherwood et al. 2014). It is possible that some of these elements may denote distinct architectures of functional elements that are better identified using nascent transcription. Consistent with this, we found an enrichment of MNase protection at sites lacking DNase-seq signal. At least one of the transcription factors that we discovered having this property (GATA2) was from a family reported to bind concurrently with a nucleosome in vitro (Cirillo and Zaret 1999; Takaku et al. 2016). Thus, one interpretation is that many of these sites reflect weak binding events in which the transcription factor and nucleosome are both present on the DNA.

A major open question following our study is whether weaker regulatory elements that lack DNase-seq signal or chromatin modifications have a distinct biological function. NFYB is an interesting example, as we observed enrichment of H3K27me3 in flanking sites, a unique pattern of MNase-seq signal, and binding inside of H3K27me3 chromatin domains. Transcription may be required within H3K27me3 domains either to maintain silencing or to establish new profiles during cellular differentiation or in response to environmental signals. We anticipate that future studies will use transcription to categorize these distinct groups of functional elements in additional detail and will determine their biological relevance in a myriad of cell types and biological conditions.

Methods

Overview of the dREG method

We devised a method to detect the location of transcriptional regulatory elements from GRO/PRO/ChRO-seq data (dREG). The basic idea behind dREG is to differentiate between two types of regions that show high levels of RNA polymerase: (1) positions where new RNA polymerase initiates; and (2) positions where RNA polymerase transcribes through after initiating at an upstream site. Our strategy for dREG prediction and scoring closely follows our prior work (Danko et al. 2015), except with modifications that leverage our new and considerably faster implementation to achieve higher classification accuracy. In addition, we have also added a novel strategy to improve the resolution for the region between divergently initiating transcription start sites.

We used support vector regression to score 50-bp intervals along the genome. Loci that were low in PRO-seq reads were prefiltered and excluded from both training and prediction tasks (see below for details). We summarized PRO-seq read counts near each position by integrating reads in nonoverlapping windows centering around the informative positions, followed by transformations that are the same as in our prior work (Danko et al. 2015). Nonoverlapping windows were taken at multiple scales, spanning both plus and minus strands and both upstream and downstream directions. dREG scores can be interpreted as the degree to which each genomic position resembles a position that falls inside of a region in which transcription initiates. We use dREG scores to identify nonoverlapping regions enriched for transcription initiation. We call these dREG "peaks" because they are analogous in most respects to ChIP-seq peaks.

Selecting positions to score

We used the SVR to score loci that meet either of the following heuristics: (1) contain more than three reads in a 100-bp interval on either strand; or (2) more than one reads in 1-kbp interval on both strands (called "informative positions"). These heuristics were designed to reduce the number of sites that we scored with each data set, while at the same time scoring at least one site near each bona fide TIR. To select these heuristics, we defined the upper bound of sensitivity for TIRs as the fraction of all GRO-cap peaks (extended by 500 bp and merged) that we recovered. We computed the fraction of TIRs that were missed (this quantity is the lower bound of our false negative rate [FNR]), and the number of positions meeting these criteria which we would have to score over different values of each of these heuristics (Supplemental Fig. S15), including the number of reads on either strand in a 100-bp window; the number of reads on the plus and minus strand within a window of 1 kb; combining separate thresholds for reads on either strand; and for reads on both the plus and minus strand. We found that the FDR was minimized to a reasonable value of 7.8% using both thresholds, as described above, without expanding the number of sites beyond what is reasonable for computation in a data set sequenced to a depth that is typical for PRO-seq data (~40 M reads). In a more deeply sequenced data set (400 M reads), we found this heuristic resulted in a theoretical lower-bound FDR of <1%.

dREG training

The new dREG model was trained using PRO/GRO-seq signal in K562 cells obtained from five independent experiments conducted by different hands in different labs over a period of ~2 yr. This diversity of training data was designed to accommodate variation in experimental conditions, batch-specific effects caused by a variety of technical factors, and detection factors such as sequencing depth. A sixth K562 data set (G7) and a data set representing an independent cell type (GM12878) were held out during model training to evaluate whether the final model was able to generalize to additional data sets. Supplemental Table S1 lists all data sources.

PRO-seq and GRO-seq data were downloaded from Gene Expression Omnibus (see accession numbers in Supplemental Table S1). We verified that all libraries were highly correlated with one another (Supplemental Fig. S1). Using this data, dREG was trained on a positive set of transcribed DHSs, defined as the intersection between DHSs identified by Duke and UW DNase-seq assays (Thurman et al. 2012) and GRO-cap HMM calls (Core et al. 2014). We defined a negative set as informative positions that do not intersect with Duke DHSs, UW DHSs, or GRO-cap HMM calls in K562 cells. We labeled each informative position as 1 or 0 according to whether it was found within a positive or negative region. To improve performance in unbalanced data sets, we trained dREG on an unbalanced training set. In practice, the number of informative genomic positions within and outside of bona fide TSSs differ greatly. To reduce the generalization error on genome-wide predictions, we optimized the ratio between positive and negative sets to best mimic this scenario. We selected 20 K positive examples and 640 K negative examples from each of the five data sets, which amounted to 3.3 M training examples. Since the size of the data set was beyond the capacity of conventional CPU-based SVM implementations, we developed a GPU-based SVM/SVR package *Rgtsvm* to handle this data set, accomplishing the training within ~28.5 h in a NVIDIA K80 GPU (Wang et al. 2017). The final models can be obtained from: ftp://cbsuftp.tc.cornell.edu/danko/hub/dreg.models/asvm.gdm.6.6M.20170828.rdata.

Discovering peaks enriched for dREG signal

We devised a statistical framework to identify genomic regions that are enriched for evidence of transcription initiation. We break the discovery of sites into three separate stages: First, we identify regions enriched for high dREG scores. Second, we stitch these regions into candidate peaks. Third, we estimate the probability that these peaks are drawn from the negative set of sites. Final predictions for genomic regions that contain transcription start sites are corrected using the false discovery rate correction for multiple testing and reported to the user.

During the first stage, our goal is to obtain an initial and inclusive set of sites and to stitch these into candidate peaks. We developed a statistical framework that determines a threshold dynamically for each data set beyond which sites are likely to be located near a transcription start site. We estimate the distribution of dREG scores in negative sites using the Laplace distribution, following previous work using this distribution for the same task (Lin and Weng 2004). The Laplace distribution is parameterized by a mean and a scale (σ in Equation 1). We assume that negative sites have a mean value of 0. The distribution of dREG scores represents a mixture distribution comprised of both negative and positive regions, and therefore fitting the scale parameter to all of the data tends to systematically overestimate the scale. To estimate the scale for a given data set, we take advantage of the fact that the Laplace distribution is symmetric about its mean. Negative dREG scores are depleted for transcription start sites and provide an estimate of the scale parameter which is empirically close to that obtained from the entire set of negative training examples when labels are available (Supplemental Fig. S4). Therefore, we estimate the scale parameter using negative dREG scores. Under these assumptions, the maximum likelihood estimate of the scale parameter is given as shown in Equation 1:

$$\sigma = \frac{\sum_{i=1}^{l} |\xi_i|}{l}, \tag{1}$$

where ξ represents the dREG scores in training examples and l is the number of training examples. Genomic loci with dREG scores higher than 99.95% under the background model were selected and stitched together into intervals by extending genomic loci that pass the threshold by ±100 bp and merging these extended loci that were in 500-bp proximity. These broad regions are similar to those introduced in our first dREG publication (Danko et al. 2015).

We next designed heuristics to refine the resolution of preliminary broad regions into narrow dREG peaks. Our approach was motivated by reports that TIRs often form clusters of distinct divergently oriented initiation sites within a local genomic region (Scruggs et al. 2015; Chen et al. 2016). Conceptually, our strategy increases the density of sites that are scored by dREG within the region and defines heuristics to identify local maxima. We first increased the local density of SVR predictions within the boundaries of preliminary dREG peaks, from 50 bp (in the initial prediction of broad dREG regions) to 10 bp. The dREG scores were smoothed by computing a weighted average of the seven dREG scores, representing ±60 bp of DNA (Equation 2).

$$\bar{r}_i = \frac{1}{16}r_{i-3} + \frac{2}{16}r_{i-2} + \frac{3}{16}r_{i-1} + \frac{4}{16}r_i + \frac{3}{16}r_{i+1} + \frac{2}{16}r_{i+2} + \frac{1}{16}r_{i+3}. \tag{2}$$

We identified points representing local maxima within each peak in which the numerical first order derivatives changed from positive to negative. This resulted in one or more local maxima for each preliminary dREG region, each pair of which had a local minima between them. We trained a random forest to decide whether to break neighboring local maxima into separate transcription initiation regions at the local minima between them. The random forest employed dREG scores, ratio of scores between the peak and valley, and the distance between each peak and the valley. The random forest was trained on a manually curated data set on Chromosome 22 of the G1 PRO-seq data set. dREG regions that contained three or more local maxima were split iteratively until no two adjacent ignored local maxima regions existed. The boundaries of a final dREG peak were defined by two valleys between the split local maxima region. For the unsymmetric broad final peaks (≥900 bp), we trimmed the longer trail to limit the width ratio between the long side and short side within 2:1. The result of this procedure was a set of nonoverlapping transcription initiation regions which were often found in clusters.

To estimate the statistical confidence of each candidate dREG peak, we devised a hypothesis testing framework in which we test the null hypothesis that points within each peak are drawn from the null (i.e., non-TIR) distribution. We consider five dREG scores around the peak center (i.e., peak center − 40 bp, peak center − 20 bp, peak center, peak center + 20 bp, peak center + 40 bp). Small peaks (<50 bp) were removed. We model dREG scores using a multivariate Laplace distribution parameterized by a mean vector and a covariance. We set the mean vector to 0, which corresponds to our null hypothesis that all five of these points are in negative regions. Nearby dREG scores have a complex correlation structure, requiring us to account for the covariance between sites. The covariance structure was specified by the Toeplitz matrix with homogeneous variances and heterogeneous correlations (Equation 3), because this formulation provides the most flexibility to fit complex data, and plenty of data is available for training in each data set. We compute the variance, σ², between sites every 20 bp using all of the dREG scores in the data set.

$$\sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_4 & \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix} \tag{3}$$

We calculated the *P*-value based on the conditional cumulative distribution of a multivariate Laplace (i.e., $p(S_i \geq ps_i \mid X_i = 0$, for $i \in [1,\ldots,5])$, where ps_i denotes the predicted score for locus i). Each dREG peak is associated with an estimated *P*-value. *P*-values are corrected for multiple testing using the Benjamini-Hochberg false discovery rate (Benjamini and Hochberg 1995). By default, dREG reports peaks with an FDR corrected *P*-value ≤ 0.05.

Web-based implementation using Apache Arvitata

The public web-based version of dREG is hosted as a Science gateway in the Extreme Science and Engineering Discovery Environment high-performance computing resource (Gesing et al. 2017; Knepper et al. 2017). The dREG gateway is hosted on the JetStream server as a web service which can submit compute jobs and download the results of dREG peaks. From the view of software architecture, it can be divided into two parts: the secured web service and the high-performance computing (HPC) resource. The secured web service built with PGA (PHP gateway with Airavata) on an Apache web server performs user authentication, data upload, sequence data transfer, and jobs submission to GPU servers via Apache Airavata middleware (Marru et al. 2011; Pierce et al. 2015). The HPC resources are GPU servers hosted by XSEDE. The dREG gateway uses a job scheduler to call the *dREG* package to complete the peak calling on GPU nodes. Once the calculation is completed, Apache Airavata copies the results from the HPC storage into the user's web storage. Since this gateway uses GPUs to speed up dREG prediction with the aid of the *Rgtsvm* package (Wang et al. 2017), a typical run takes ∼4–12 h (mean = 6.7 h) after the job starts running on the GPU server.

Using Tfit

The Tfit software (most recent on April 28, 2017) was obtained from https://github.com/azofeifa/Tfit. The Tfit software package was run using the default parameters, following instructions from the package authors. We tried using a variety of different settings (both with and without optimizing the template density function by promoter or TSS associated regions; -tss parameter). We also explored treating input data as both the full Illumina mapped read or representing the position of RNA polymerase using the single base corresponding to the 3′ end of each read. We present the parameters that achieved the highest sensitivity for transcribed DHSs (without the -tss parameter, and using the complete Illumina read in the input bigWig).

Comparison to DNase-seq and ChIP-seq data

Public PRO-seq and GRO-seq data were collected from published resources (Hah et al. 2011; Danko et al. 2013, 2018; Allen et al. 2014; Core et al. 2014; Niskanen et al. 2015; Dukler et al. 2017; Vihervaara et al. 2017). DNase I hypersensitive sites for the ENCODE reference cell types were processed using a uniform pipeline that we recently described (Chu et al. 2018). Sites detected using dREG were classified into DHS+ (defined as TIRs having peak calls in both Duke and UW DNase-seq data), and DHS− (defined as having peak calls in neither Duke nor UW data). All computations on BED files were performed using BEDTools (Quinlan and Hall 2010). BEDTools was used to calculate overlap regions (using the command *bedtools intersect*), closest distances (*bedtools closest*), and Jaccard scores (*bedtools jaccard*). ENCODE hg19 blacklist regions were excluded from all analyses. Downstream processing, data analyses, heat maps, and other visualizations were performed in R (R Core Team 2019) using the bigWig package (https://github.com/andrelmartins/bigWig). Our scripts are posted on GitHub (https://github.com/Danko-Lab/dREG/tree/master/GR_submit_2018).

Cell culture

Cell lines were obtained from the American Type Culture Collection (ATCC) and cultured using standard cell culture procedures and sterile technique. Human K562 suspension cells were cultured in RPMI-1640 media supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin. Human HeLa adherent cells were cultured with Dulbecco's Modified Eagle Media supplemented with 10% FBS and 1% penicillin/streptomycin. Media and antibiotics were from Corning and FBS was from Atlanta Biologicals.

ATAC-seq data preparation and processing

ATAC-seq was performed on K562 and HeLa as described in Buenrostro et al. (2013). Briefly, nuclei were isolated from 50,000 K562 and HeLa cells in duplicate, tagmented using the Nextera DNA Sample Preparation kit, and amplified for seven PCR cycles using the NEBNext DNA Library Prep kit. All libraries were pooled and sequenced using an Illumina NextSeq 500. Raw sequencing data was aligned to hg19 using BWA-MEM (Li 2013). The hg19 genome build was used to maintain compatibility with existing PRO-seq alignments. Aligning reads to GRCh38 is unlikely to substantially affect the conclusions, as the two reference assemblies are highly similar in the uniquely mappable euchromatin regions analyzed here.

Data access

All new sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE121993.

Acknowledgments

Computational resources required in this study were provided by XSEDE allocation numbers TG-BIO160048 and TG-MCB160061 as well as an NVIDIA GPU Grant. We thank the PHP gateway Airavata team for technical assistance building the web service, especially Marcus A. Christie, Eroma Abeysinghe, Marlon Pierce, and Suresh Marru. We also thank beta testers, Anniina Vihervaara and Donald Miller. We thank James Lewis and Anniina Vihervaara for critical comments on the manuscript, as well as other members of the Danko lab for valuable discussions and suggestions. The work in this publication was supported by an NHGRI (National Human Genome Research Institute) grant R01-HG009309 to C.G.D. The content is solely the responsibility of the authors and does not necessarily represent the official views of the U.S. National Institutes of Health.

References

Allen MA, Andrysik Z, Dengler VL, Mellert HS, Guarnieri A, Freeman JA, Sullivan KD, Galbraith MD, Luo X, Kraus WL, et al. 2014. Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. *eLife* **3:** e02200. doi:10.7554/eLife.02200

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014a. An atlas of active enhancers across human cell types and tissues. *Nature* **507:** 455–461. doi:10.1038/nature12787

Andersson R, Refsing Andersen P, Valen E, Core LJ, Bornholdt J, Boyd M, Heick Jensen T, Sandelin A. 2014b. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat Commun* **5:** 5336. doi:10.1038/ncomms6336

Andersson R, Sandelin A, Danko CG. 2015. A unified architecture of transcriptional regulatory elements. *Trends Genet* **31:** 426–433. doi:10.1016/j.tig.2015.05.007

Azofeifa JG, Dowell RD. 2016. A generative model for the behavior of RNA polymerase. *Bioinformatics* **33:** 227–234. doi:10.1093/bioinformatics/btw599

Azofeifa JG, Allen MA, Hendrix JR, Read T, Rubin JD, Dowell RD. 2018. Enhancer RNA profiling predicts transcription factor activity. *Genome Res* **28:** 334–344. doi:10.1101/gr.225755.117

Barozzi I, Simonatto M, Bonifacio S, Yang L, Rohs R, Ghisletti S, Natoli G. 2014. Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Mol Cell* **54:** 844–857. doi:10.1016/j.molcel.2014.04.006

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837. doi:10.1016/j.cell.2007.05.009

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* **57:** 289–300. doi:10.2307/2346101

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132:** 311–322. doi:10.1016/j.cell.2007.12.014

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10:** 1213–1218. doi:10.1038/nmeth.2688

Chen Y, Pai AA, Herudek J, Lubas M, Meola N, Järvelin AI, Andersson R, Pelechano V, Steinmetz LM, Jensen TH, et al. 2016. Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. *Nat Genet* **48:** 984–994. doi:10.1038/ng.3616

Chu T, Rice EJ, Booth GT, Salamanca HH, Wang Z, Core LJ, Longo SL, Corona RJ, Chin LS, Lis JT, et al. 2018. Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. *Nat Genet* **50:** 1553–1564. doi:10.1038/s41588-018-0244-3

Churchman LS, Weissman JS. 2011. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469:** 368–373. doi:10.1038/nature09652

Cirillo LA, Zaret KS. 1999. An early developmental transcription factor complex that is more stable on nucleosome core particles than on free DNA. *Mol Cell* **4:** 961–969. doi:10.1016/S1097-2765(00)80225-7

Coleman RT, Struhl G. 2017. Causal role for inheritance of H3K27me3 in maintaining the OFF state of a *Drosophila* HOX gene. *Science* **356:** eaai8236. doi:10.1126/science.aai8236

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322:** 1845–1848. doi:10.1126/science.1162228

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46:** 1311–1320. doi:10.1038/ng.3142

Danko CG, Hah N, Luo X, Martins AL, Core L, Lis JT, Siepel A, Kraus WL. 2013. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol Cell* **50:** 212–222. doi:10.1016/j.molcel.2013.02.015

Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, Cheung VG, Kraus WL, Lis JT, Siepel A. 2015. Identification of active transcriptional regulatory elements from GRO-seq data. *Nat Methods* **12:** 433–438. doi:10.1038/nmeth.3329

Danko CG, Choate LA, Marks BA, Rice EJ, Wang Z, Chu T, Martins AL, Dukler N, Coonrod SA, Tait Wojno ED, et al. 2018. Dynamic evolution of regulatory element ensembles in primate CD4$^+$ T cells. *Nat Ecol Evol* **2:** 537–548. doi:10.1038/s41559-017-0447-5

Diao Y, Fang R, Li B, Meng Z, Yu J, Qiu Y, Lin KC, Huang H, Liu T, Marina RJ, et al. 2017. A tiling-deletion-based genetic screen for *cis*-regulatory element identification in mammalian cells. *Nat Methods* **14:** 629–635. doi:10.1038/nmeth.4264

Dukler N, Booth GT, Huang Y-F, Tippens N, Waters CT, Danko CG, Lis JT, Siepel A. 2017. Nascent RNA sequencing reveals a dynamic global transcriptional response at genes and enhancers to the natural medicinal compound celastrol. *Genome Res* **27:** 1816–1829. doi:10.1101/gr.222935.117

Duttke SHC, Lacadie SA, Ibrahim MM, Glass CK, Corcoran DL, Benner C, Heinz S, Kadonaga JT, Ohler U. 2015. Human promoters are intrinsically directional. *Mol Cell* **57:** 674–684. doi:10.1016/j.molcel.2014.12.029

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473:** 43–49. doi:10.1038/nature09906

Fuda NJ, Ardehali MB, Lis JT. 2009. Defining mechanisms that regulate RNA polymerase II transcription *in vivo*. *Nature* **461:** 186–192. doi:10.1038/nature08449

Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, Kane M, Cleary B, Lander ES, Engreitz JM. 2016. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354:** 769–773. doi:10.1126/science.aag2445

Gesing S, Wilkins-Diehr N, Dahan M, Lawrence K, Zentner M, Pierce M, Hayden L, Marru S. 2017. Science gateways: the long road to the birth of an institute. In *Proceedings of the 50th Hawaii International Conference on System Sciences*. HICSS, Waikoloa, HI. https://scholarspace.manoa.hawaii.edu/bitstream/10125/41919/1/paper0770.pdf.

Grøntved L, John S, Baek S, Liu Y, Buckley JR, Vinson C, Aguilera G, Hager GL. 2013. C/EBP maintains chromatin accessibility in liver and facilitates glucocorticoid receptor recruitment to steroid response elements. *EMBO J* **32:** 1568–1583. doi:10.1038/emboj.2013.106

Hah N, Danko CG, Core L, Waterfall JJ, Siepel A, Lis JT, Kraus WL. 2011. A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145:** 622–634. doi:10.1016/j.cell.2011.03.042

Hah N, Murakami S, Nagari A, Danko CG, Kraus WL. 2013. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res* **23:** 1210–1223. doi:10.1101/gr.152306.112

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39:** 311–318. doi:10.1038/ng1966

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38:** 576–589. doi:10.1016/j.molcel.2010.05.004

Henriques T, Scruggs BS, Inouye MO, Muse GW, Williams LH, Burkholder AB, Lavender CA, Fargo DC, Adelman K. 2018. Widespread transcriptional pausing and elongation control at enhancers. *Genes Dev* **32:** 26–41. doi:10.1101/gad.309351.117

Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al. 2009. Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nat Methods* **6:** 283–289. doi:10.1038/nmeth.1313

Hosogane M, Funayama R, Shirota M, Nakayama K. 2016. Lack of transcription triggers H3K27me3 accumulation in the gene body. *Cell Rep* **16:** 696–706. doi:10.1016/j.celrep.2016.06.034

James Faresse N, Canella D, Praz V, Michaud J, Romascano D, Hernandez N. 2012. Genomic study of RNA polymerase II and III SNAP$_C$-bound promoters reveals a gene transcribed by both enzymes and a broad use of common activators. *PLoS Genet* **8:** e1003028. doi:10.1371/journal.pgen.1003028

Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465:** 182–187. doi:10.1038/nature09033

Knepper R, Coulter E, Pierce M, Marru S, Pamidighantam S. 2017. Using the Jetstream research cloud to provide science gateway resources. In *2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, pp. 753–757. IEEE/ACM, Madrid.

Kwak H, Fuda NJ, Core LJ, Lis JT. 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339:** 950–953. doi:10.1126/science.1229386

Leemans C, van der Zwalm M, Brueckner L, Comoglio F, van Schaik T, Pagie L, van Arensbergen J, van Steensel B. 2018. Promoter-intrinsic and local chromatin features determine gene repression in lamina-associated domains. bioRxiv doi:10.1101/464081

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].

Lin CJ, Weng RC. 2004. *Simple probabilistic predictions for support vector regression*. National Taiwan University, Taipei, Taiwan. https://www.researchgate.net/profile/Ruby_Weng/publication/228573389_Simple_probabilistic_predictions_for_support_vector_regression/links/5555f92208ae980ca60c7ee3.pdf.

Marru S, Gunathilake L, Herath C, Tangchaisin P, Pierce M, Mattmann C, Singh R, Gunarathne T, Chinthaka E, Gardler R, et al. 2011. Apache Airavata: a framework for distributed applications and computational workflows. In *Proceedings of the 2011 ACM Workshop on Gateway Computing Environments, GCE '11*, pp. 21–28. ACM, New York.

Mayer A, di Iulio J, Maleri S, Eser U, Vierstra J, Reynolds A, Sandstrom R, Stamatoyannopoulos JA, Churchman LS. 2015. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **161:** 541–554. doi:10.1016/j.cell.2015.03.010

Melgar MF, Collins FS, Sethupathy P. 2011. Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biol* **12:** R113. doi:10.1186/gb-2011-12-11-r113

Mikhaylichenko O, Bondarenko V, Harnett D, Schor IE, Males M, Viales RR, Furlong EEM. 2018. The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes Dev* **32:** 42–57. doi:10.1101/gad.308619.117

Niskanen EA, Malinen M, Sutinen P, Toropainen S, Paakinaho V, Vihervaara A, Joutsen J, Kaikkonen MU, Sistonen L, Palvimo JJ. 2015. Global

SUMOylation on active chromatin is an acute heat stress response restricting transcription. *Genome Biol* **16:** 153. doi:10.1186/s13059-015-0717-y

Nojima T, Gomes T, Grosso ARF, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M, Proudfoot NJ. 2015. Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161:** 526–540. doi:10.1016/j.cell.2015.03.027

Pierce ME, Marru S, Gunathilake L, Wijeratne DK, Singh R, Wimalasena C, Ratnayaka S, Pamidighantam S. 2015. Apache Airavata: design and directions of a science gateway framework. *Concurr Comput* **27:** 4282–4291. doi:10.1002/cpe.3534

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

R Core Team. 2019. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Rajagopal N, Srinivasan S, Kooshesh K, Guo Y, Edwards MD, Banerjee B, Syed T, Emons BJM, Gifford DK, Sherwood RI. 2016. High-throughput mapping of regulatory DNA. *Nat Biotechnol* **34:** 167–174. doi:10.1038/nbt.3468

Scruggs BS, Gilchrist DA, Nechaev S, Muse GW, Burkholder A, Fargo DC, Adelman K. 2015. Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol Cell* **58:** 1101–1112. doi:10.1016/j.molcel.2015.04.006

Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK. 2014. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* **32:** 171–178. doi:10.1038/nbt.2798

Takaku M, Grimm SA, Shimbo T, Perera L, Menafra R, Stunnenberg HG, Archer TK, Machida S, Kurumizaka H, Wade PA. 2016. GATA3-dependent cellular reprogramming requires activation-domain dependent recruitment of a chromatin remodeler. *Genome Biol* **17:** 36. doi:10.1186/s13059-016-0897-0

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489:** 75–82. doi:10.1038/nature11232

Vihervaara A, Mahat DB, Guertin MJ, Chu T, Danko CG, Lis JT, Sistonen L. 2017. Transcriptional response to stress is pre-wired by promoter and enhancer architecture. *Nat Commun* **8:** 255. doi:10.1038/s41467-017-00151-0

Wang Z, Chu T, Choate LA, Danko CG. 2017. Rgtsvm: support vector machines on a GPU in R. arXiv:1706.05544 [stat.ML].

Xi H, Shulha HP, Lin JM, Vales TR, Fu Y, Bodine DM, McKay RDG, Chenoweth JG, Tesar PJ, Furey TS, et al. 2007. Identification and characterization of cell type–specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet* **3:** e136. doi:10.1371/journal.pgen.0030136

Zhou X, Maricque B, Xie M, Li D, Sundaram V, Martin EA, Koebbe BC, Nielsen C, Hirst M, Farnham P, et al. 2011. The Human Epigenome Browser at Washington University. *Nat Methods* **8:** 989–990. doi:10.1038/nmeth.1772

Received May 13, 2018; accepted in revised form December 18, 2018.