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, NY 14853.
[2] Graduate field of Computational Biology, Cornell University, Ithaca, NY 14853.
[3] Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.

[*] **Address correspondence to:**
Charles G. Danko, Ph.D.
Baker Institute for Animal Health
Cornell University
Hungerford Hill Rd.
Ithaca, NY 14853
Phone: (607) 256-5620
E-mail: dankoc@gmail.com

Abstract:

Our genomes encode a wealth of transcription initiation regions (TIRs) that can be identified by their distinctive patterns of transcription initiation. 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% improvements in accuracy, 3-4-fold 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-I-hypersensitivity. Novel TIRs discovered by dREG were often associated with RNA polymerase III initiation or bound by transcription factors that recognize DNA concurrently with a nucleosome. We provide a web interface to dREG that can be used by the scientific community (http://dREG.DNASequence.org).

Introduction

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 are indirect measures of genome function, and mark a variety of nuclease accessible regions in our genomes, such as binding sites for the insulator protein CTCF or inactive regulatory elements (Danko et al. 2015; Xi et al. 2007). 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.

Active transcription initiation by Pol II 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; Kwak et al. 2013; Churchman and Weissman 2011; 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. 2017). 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. Combined with new innovations to identify regions enriched for transcription initiation (called dREG "transcription initiation regions") and an extremely accurate new dREG model, we show that this strategy is useful approach for detecting regulatory elements genome-wide in a number of applications. Importantly, we show that our new strategy is more sensitive in certain types of regulatory regions, for instance Pol III promoters or transcription start sites in heterochromatin domains, than DNase-I hypersensitivity or other genomic tools. 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 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 datasets which took 30-40 hours 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 and 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 dataset where bona-fide positive TREs 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 datasets 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 and Supplemental Fig. S3**). We estimate the probability that dREG scores were drawn from the negative class of sites (i.e., non-TREs) by modelling 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 non-overlapping 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 (non-regulatory 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 two datasets, PRO-seq in K562 and GRO-seq in GM12878, that were held out during model training. Holdouts were selected because they cover a range of library sequencing depths and a new cell type that together allowed us to determine whether the dREG model generalized to additional datasets. dREG predicted 34,677 and 71,131 TIRs in K562 and GM12878, 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 PRO-cap pairs) and 58.4% or 71.8% of DHSs marked by the acetylation of histone 3 lysine 27 (H3K27ac) (**Fig. 1B**). Sensitivity for both PRO-cap and H3K27ac-DHSs was >2-fold higher for dREG than for the elegant model-based Tfit program (Azofeifa and Dowell 2016) when run on the same data. Transcription initiation regions display a range in the efficiency of initiation on the two strands (Scruggs et al. 2015; Duttke et al. 2015), and dREG was able to identify the location of both uni- and bi-directional transcription initiation sites (**Supplemental Fig. S5**).

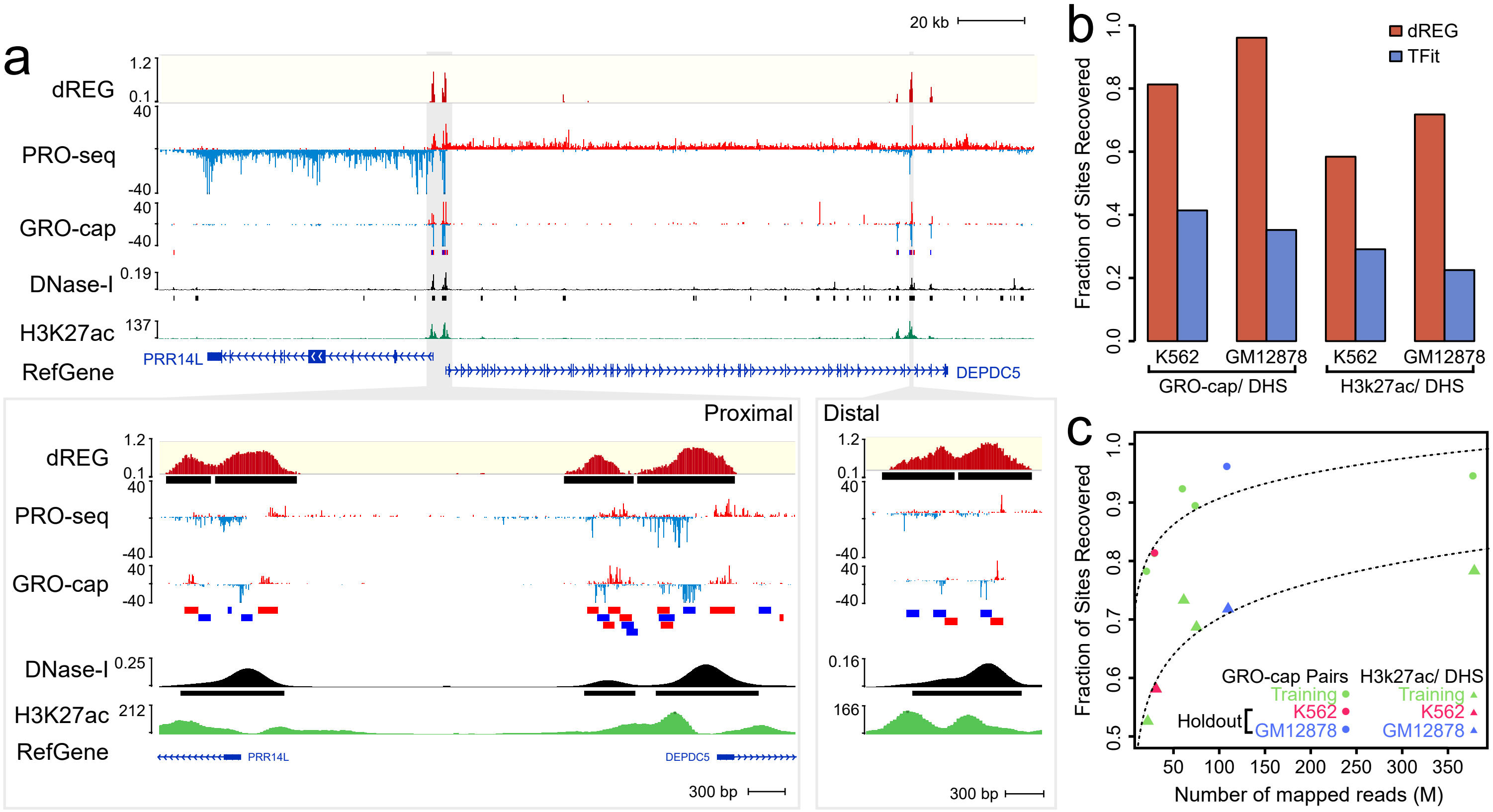

Figure 1. dREG identifies regions of transcription initiation. (A) WashU Epigenome Browser visualization of dREG signal, PRO-seq data, GRO-cap, DNase-I hypersensitivity, and H3K27ac ChIP-seq near the PRR14L and DEPDC5 genes. Inserts show an expanded view of gene-proximal promoter elements (left) and a distal enhancer (right), each encoding multiple transcription initiation sites. **(B)** Barplots show the fraction of transcribed DHSs (left) and H3K27ac+ DHSs (right) in two holdout datasets that were discovered by dREG (red) and Tfit (blue). **(C)** Scatterplot shows the fraction of sites recovered (Y-axis) as a function of sequencing depth (X-axis) for seven datasets shown in Supplementary Table 1. The best fit lines are shown. The color represents whether the dataset was used for training (green) or is a holdout dataset (K562, red) or cell type (GM12878, blue).

The sensitivity of dREG varied systematically by the library sequencing depth (**Fig. 1C**). After accounting for sequencing depth, we did not observe any systematic difference between datasets that were held out or used during training, suggesting that dREG was not noticeably overfitting to the training data. dREG achieved a reasonable sensitivity on a K562 holdout dataset with 27M uniquely mapped reads (81.3% of DHSs overlapping GRO-cap pairs were recovered), and saturated the discovery of enhancers supported by ENCODE data at between 60-100M uniquely mapped reads. 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 (2,924 bp on average), closer in size to high-resolution DNase-I-seq data (322 bp on average). Histone modification ChIP-seq or DNase-I-seq data aligned to the center of human dREG sites revealed good agreement with the center of the nucleosome free region (**Fig. 2A**). 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 (400-8,000 TIRs, depending on the dataset), and intriguingly did not saturate even in datasets sequenced to a depth of 350M uniquely mapped reads (**Fig. 2B**). As expected, TIRs had lower dREG scores and lower polymerase abundance when they were found uniquely by dREG (**Fig. 2A**), 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-I hypersensitivity 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 dataset) overlapped ChIP-seq peak calls for sequence specific transcription factors (**Fig. 3A, Supplemental Fig. S6**). 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 non-coding RNAs (**Fig. 2C**). Often the promoter of these transcription units lacked sufficiently robust enrichment of histone modifications or DNase-I hypersensitivity to make confident peak calls, and many lacked sufficient paused Pol II to be represented in GRO-cap data. Nevertheless, examination of these TIRs genome-wide revealed a local increase in the abundance of reads in the average profiles of DNase-I-seq and active histone modification ChIP-seq data (**Fig. 2A and Supplemental Fig. S7**), 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. S8**). 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.

Pol II and Pol III transcription initiation without chromatin accessibility

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 tolerate sites that lack specific aspects of the core regulatory architecture better than others. We focused on DNase-I hypersensitivity as a general marker for the nucleosome depleted region in the center of regulatory elements. As a control we performed ATAC-seq to confirm low levels of chromatin accessibility in K562 clones that were closely related to those used to generate PRO-seq data (**Supplemental Fig. S9**). To determine whether specific transcription factors may be more permissive to binding in sites having low levels of

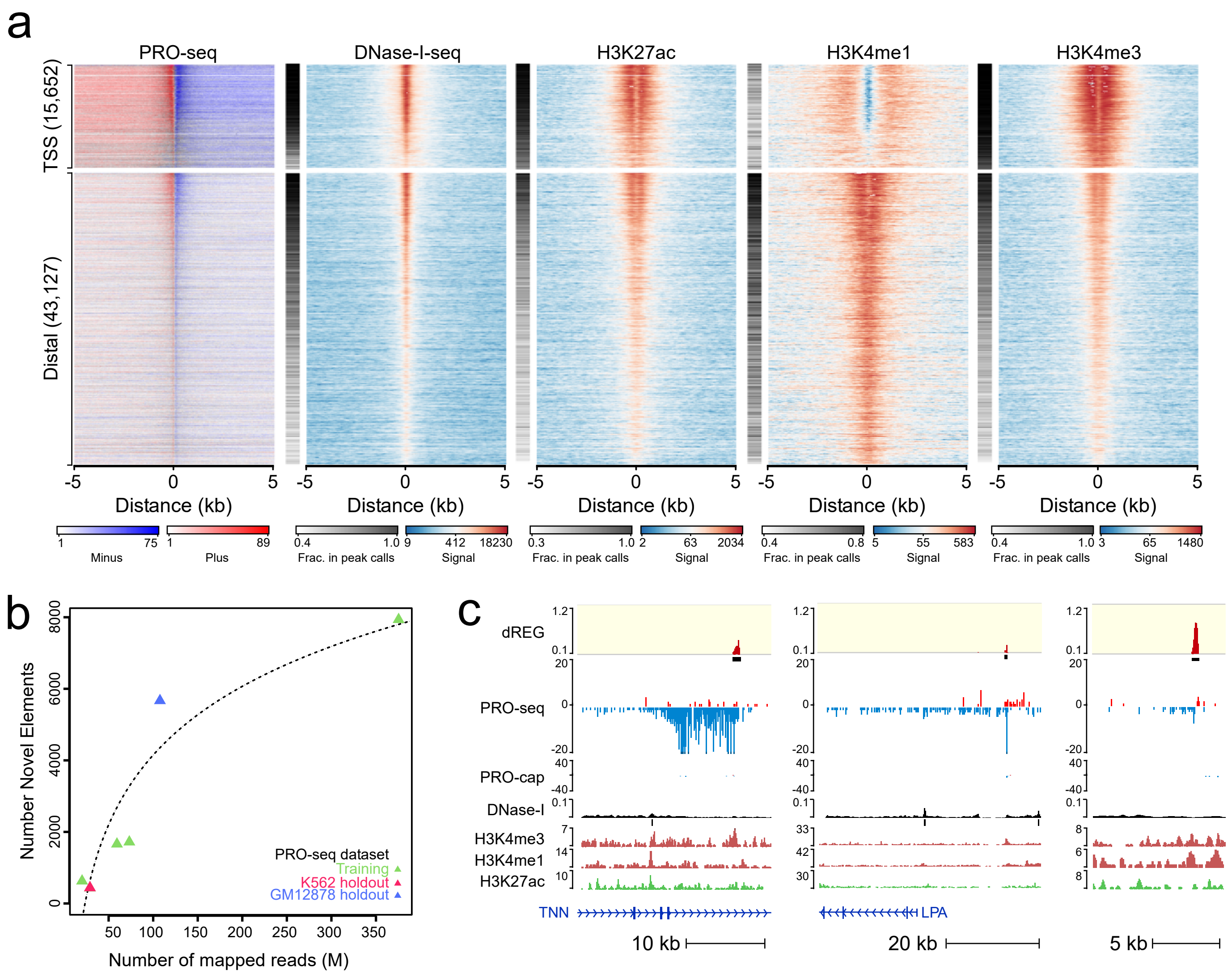

Figure 2. dREG identifies new regions that were not found using other molecular assays. (A) Heatmaps show the log-signal intensity of PRO-seq, DNase-I-seq, or ChIP-seq for H3K27ac, H3K4me1, and H3K4me3. The fraction of sites intersecting ENCODE peak calls is shown in the white-black color map beside 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 >5kb to a start site (n= 43,127) **(B)** Scatterplot shows the number of new TIRs that were not discovered in DNase-I-seq or H3K27ac ChIP-seq data (Y-axis) as a function of sequencing depth (X-axis) for seven datasets shown in Supplementary Table 1. The best fit line is shown. The color represents whether the dataset was used for training (green) or is a holdout dataset (K562, red) or cell type (GM12878, blue). **(C)** Three separate genome-browser regions that denote TIRs discovered using dREG, but were not found in DNase-I-seq or H3K27ac ChIP-seq data. Tracks show dREG signal, PRO-seq data, GRO-cap, DNase-I hypersensitivity, H3K27ac ChIP-seq, and annotated genes.

chromatin accessibility, we trained a logistic regression model to predict whether TIRs discovered using dREG intersect a DHS. Transcription factor binding 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. S10**). Thus specific transcription factors were predictive of which TIRs lacked nuclease hypersensitivity.

To identify transcription factors that contribute to this signal, we computed the ratio of binding sites that were found using dREG but not DNase-I-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. 3B**). However, different transcription factors exhibited a broad range of binding in dREG+DHS- sites. The highest scoring transcription factors were RPC1155 and BRF2 (ratio of dREG+DHS-/dREG+DHS+ = 0.37 and 0.29, respectively), which encode the catalytic core of RNA Polymerase (Pol) III and a Pol III initiation factor. Thus, Pol III promoters were often not sufficiently exposed to the DNase-I enzyme to be detected in DNase-I-seq data. Members of the core Pol II transcriptional machinery were also represented as outliers (CHD1, POLR2A, and TAF7) consistent with some Pol II promoters being inaccessible as well. Many of the other transcription factors enriched in the dREG only class were associated with heterochromatin (KAP1, ZNF274, and EZH2), consistent with the lower amount of transcription that distinguishes sites without DHSs (**Fig. 2A**).

Several sequence specific transcription factors were also observed to have a high fraction of sites that were dREG+DHS-. For example, CEBPB, NFYB, GATA2, and SPI1 had a relatively high fraction of binding sites outside of DHSs. Intriguingly, 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 approximately 300 bp in DHS- sites (**Fig. 3C**), suggesting systematic differences in the chromatin environment in these regions. By contrast, binding sites for MAZ and ZNF143, which exhibited a low fraction of binding sites outside of DHSs, did not show as promanant of an increase in MNase-seq signal in DHS- binding sites (**Fig 3C**). 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. 3C**). 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. 3C**). Likewise, histone modification ChIP-seq in DHS- regions notably 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-I-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-I-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 TRE 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 provides 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 hours 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 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 and Lawrence; 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/.

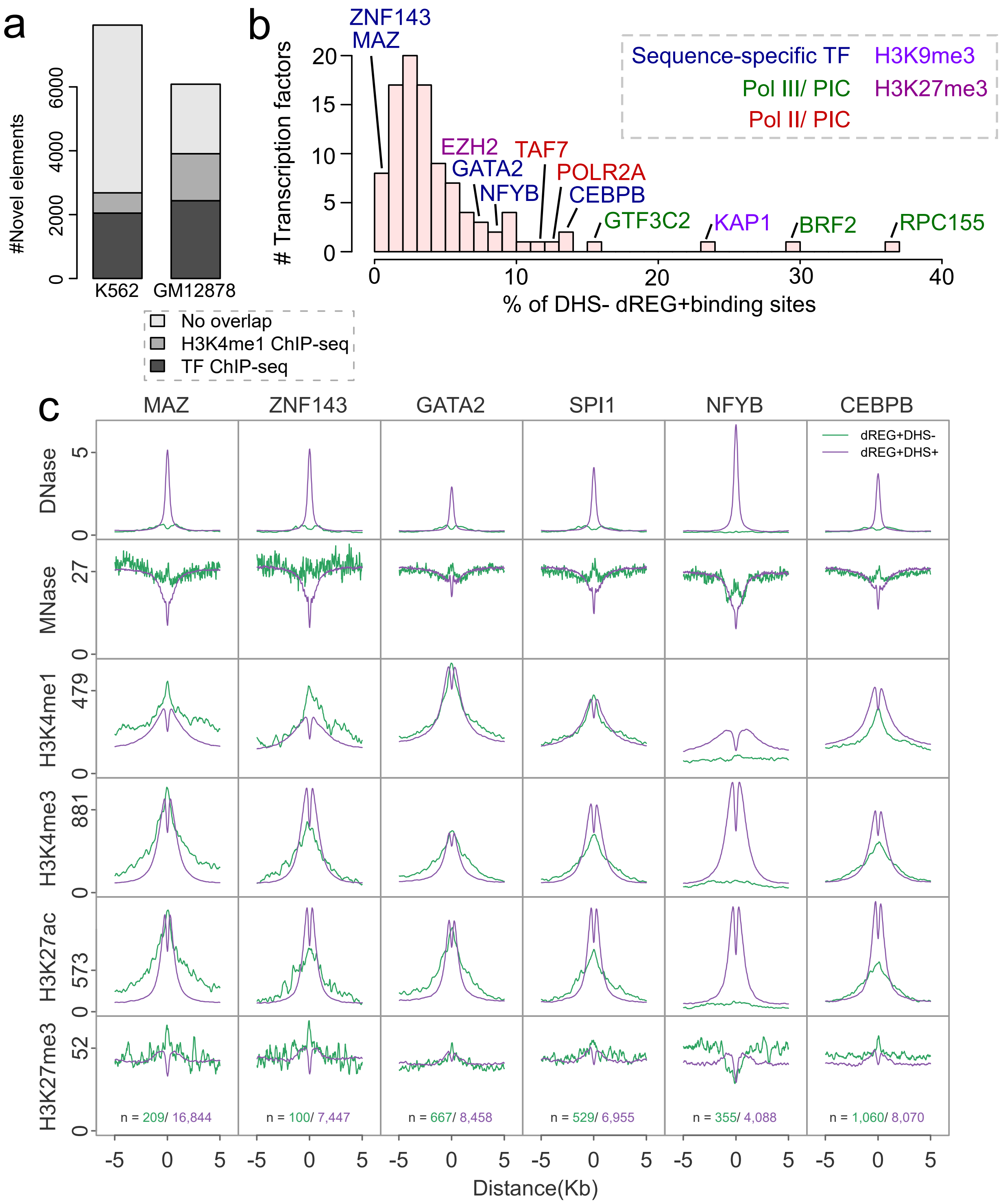

Figure 3. dREG TIRs were reproducibly associated with specific transcription factor binding. (A) Plot shows the number of elements discovered using dREG, but not found in DNase-I hypersensitivity or H3K27ac ChIP-seq (Y-axis) for PRO-seq datasets in K562 and GM12878 cells. The color denotes other functional marks intersecting sites discovered only using dREG. **(B)** Histogram representing the fraction of binding sites for 100 transcription factors supported by a dREG TIR that was not also discovered in DNase-I hypersensitivity data (i.e., dREG+ DHS- / dREG+ DHS+). Several of the outliers are shown. The color denotes whether the factor is a member of the Pol III pre initiation complex (green), Pol II preinitiation complex (red), associated with H3K9me3 or H3K27me3 heterochromatin (purple), or is a sequence specific transcription factor (blue). **(C)** Metaplots show the raw signal of DNase-I hypersensitivity, MNase-seq, and ChIP-seq for H3K4me1, H3K4me3, H3K27ac, and H3K27me3 for six transcription factors, including MAZ, ZNF143, GATA2, SPI1, NFYB, and CEBPB. Signals are shown for dREG+DHS- (green) and dREG+DHS+ (purple) sites.

Discussion

In this article we introduced a fully optimized version 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 (Chu et al. 2017; Core et al. 2008; Kwak et al. 2013). 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. Taken together, our dREG implementation has a number of advantages compared with alternative approaches. dREG offers substantial improvements in regulatory element resolution (compared with ChIP-seq), information about activity (DNase-I-seq and ATAC-seq), improved signal to noise ratio (ChIP-seq, DNase-I-seq, and ATAC-seq), and a higher sensitivity for certain types of active regulatory elements (ChIP-seq, DNase-I-seq, and ATAC-seq). 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 expression in the same molecular assay. Most importantly, dREG/ PRO-seq allows users to measure multiple aspects of gene regulation, including the position of regulatory elements, gene expression, and pausing levels using a single genomic experiment. When paired with ChRO-seq (Chu et al. 2017), 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 >8,000 regulatory elements that were not detected using DNase-I-hypersensitivity or H3K27ac ChIP-seq. Differences between assays may in part reflect false negatives in DNase-I-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 (Diao et al. 2017; Rajagopal et al. 2016). 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-I hypersensitivity. It is possible that some of these elements may denote distinct architectures of functional element that are better identified using nascent transcription. Consistent with this, we found an enrichment of MNase protection at sites lacking DNase-I-seq signal. At least one of the transcription factors that that we discovered having this property (GATA) was from a family reported to bind concurrently with a nucleosome *in vitro* (Takaku et al. 2016; Cirillo and Zaret 1999). 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 TREs that lack DNase-I hypersensitivity or other chromatin modifications have a distinct biological function. NFYB is an interesting example, as its enrichment of H3K27me3 in flanking sites, as well as a unique pattern of MNase-seq signal, might suggest 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.

Acknowledgements

We thank XSEDE allocation numbers TG-BIO160048 and TG-MCB160061 for providing computational resources required in this gateway, including Web server, Apache service, Airavata service, Web storage and GPU nodes. 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, Anni Viherva and Don Miller. We thank James Lewis for critical comments on the manuscript, as well as other members of the Danko lab for valuable discussions and suggestions. Work in this publication was supported by an NHGRI (National Human Genome Research Institute) grant R01-HG009309 to CGD. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

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: (i) positions where new RNA polymerase initiates, and (ii) 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 more precisely identify regions in which transcription start sites occur, representing the location of TREs.

We used support vector regression (SVR) to score 50 bp intervals along the genome. Loci that were low in PRO-seq reads were pre-filtered and excluded from both training and prediction tasks. We selected loci for analysis that meet either of the following heuristics: 1) contain more than 3 reads in a 100 bp interval on either strand, or 2) more than 1 reads in 1 kbp interval on both strands. We refer to positions meeting these criteria as “informative positions”. We summarized PRO-seq read counts near each position by integrating reads in non-overlapping windows centering around the informative positions, followed by transformations that are the same as in our prior work (Danko et al. 2015). Non-overlapping windows were taken at multiple scales, spanning both plus and minus strand, 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 non-overlapping regions enriched for transcription initiation. We call these dREG “peaks” because they are analogous in most respects to ChIP-seq peaks.

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 approximately two years. 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 dataset (G7) and a dataset representing an independent cell type (GM12878) were held out during model training to evaluate whether the final was able to generalize to additional datasets. **Table S1** lists all data sources.

PRO-seq and GRO-seq data were downloaded from Gene Expression Omnibus (see accession numbers in **Table S1**). We verified that all libraries were highly correlated with one another (**Figure 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-I-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 datasets we trained dREG on an unbalanced training set. In practice the number of informative genomic positions within and outside of bona-fide TSS 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 20K positive examples and 640K negative examples from each of the five datasets, which amounted to 3.3M training examples. Since the size of the dataset was beyond the capacity of conventional CPU-based SVM implementations, we developed a GPU-based SVM/SVR package *Rgtsvm* to handle this dataset, accomplishing the training within ~28.5hrs 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 dataset 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 dataset, 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 (**Figure 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):

$$\sigma = \frac{\sum_{⍰=1}^{⍰} \quad |⍰_{⍰}|}{⍰} \tag{1}$$

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 TREs often form clusters of distinct divergently oriented initiation sites within a local genomic region (Chen et al. 2016; Scruggs et al. 2015). 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).

$$\overline{r_i} = \frac{1}{16}r_{i-3} + \frac{2}{16}r_{i-2} + \frac{3}{16}r_{i-1} + \frac{4}{16}r_i + \frac{3}{16}r_{i+1} + \frac{2}{16}r_{i+2} + \frac{1}{16}r_{i+3} \tag{2}$$

We identified points representing local maxima within each peak in which the numerical 1st 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 each peak and the valley. The random forest was trained on a manually curated dataset on chromosome 22 of the G1 PRO-seq dataset. dREG regions that contained three or more local maxima were split iteratively until no two adjacent ignored local maxima regions existed. The boundaries of 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 long side and short side within 2:1. The result of this procedure was a set of non-overlapping 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-TRE) distribution. We consider five dREG scores around the peak center (i.e., peak center - 40bp, peak center - 20bp, peak center, peak center + 20bp, peak center + 40bp). 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 homogenous 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 dataset. We compute the variance, σ^2, between sites every 20 bp using all of the dREG scores in the dataset.

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

We calculated the p value based on the conditional cumulative distribution of a multivariate Laplace (i.e., p($S_i \geq ps_i$ | X_i=0, for $i \in [1,...,5]$), 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 (XSEDE) high performance computing resource (Gesing and Lawrence; 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 is 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 using 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 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 hours (mean = 6.7 hours) after the job starts running on the GPU server.

Using TFit

The Tfit software (most recent on April 28th 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), and treating input data as both full Illumina reads and specifying the 3' ends. We present the parameters that achieved the highest sensitivity for transcribed DHSs (without the -tss parameter, and with fully extended reads).

Comparison to DNase-I-seq and ChIP-seq data

DNase-I hypersensitive sites for the ENCODE reference cell types were processed using a uniform pipeline that we recently described (Chu et al. 2017). Sites detected using dREG were classified into DHS+ (defined as TIRs having peak calls in both Duke and UW DNase-I hypersensitivity 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 (*bedtools intersect*), closest distances (*bedtools closest*) and jaccard scores (*bedtools jaccard*). Downstream processing, data analyses, heatmaps, and other and visualizations were performed in R using the bigWig package (https://github.com/andrelmartins/bigWig). Our scripts are posted on GitHub (https://github.com/Danko-Lab/dREG/tree/master/dREG_submit_2018).

Figure Legends

Figure 1. dREG identifies regions of transcription initiation. (A) WashU Epigenome Browser visualization of dREG signal, PRO-seq data, GRO-cap, DNase-I hypersensitivity, and H3K27ac ChIP-seq near the *PRR14L* and *DEPDC5* genes. Inserts show an expanded view of gene-proximal promoter elements (left) and a distal enhancer (right), each encoding multiple transcription initiation sites. (B) Barplots show the fraction of transcribed DHSs (left) and H3K27ac+ DHSs (right) in two holdout datasets that were discovered by dREG (red) and Tfit (blue). (C) Scatterplot shows the fraction of sites recovered (Y-axis) as a function of sequencing depth (X-axis) for seven datasets shown in Supplementary Table 1. The best fit lines are shown. The color represents whether the dataset was used for training (green) or is a holdout dataset (K562, red) or cell type (GM12878, blue).

Figure 2. dREG identifies new regions that were not found using other molecular assays. (A) Heatmaps show the log-signal intensity of PRO-seq, DNase-I-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 >5kb to a start site (n= 43,127) (B) Scatterplot shows the number of new TIRs that were not discovered in DNase-I-seq or H3K27ac ChIP-seq data (Y-axis) as a function of sequencing depth (X-axis) for seven datasets shown in Supplementary Table 1. The best fit line is shown. The color represents whether the dataset was used for training (green) or is a holdout dataset (K562, red) or cell type (GM12878, blue). (C) Three separate genome-browser regions that denote TIRs discovered using dREG, but were not found in DNase-I-seq or H3K27ac ChIP-seq data. Tracks show dREG signal, PRO-seq data, GRO-cap, DNase-I hypersensitivity, H3K27ac ChIP-seq, and annotated genes.

Figure 3. dREG TIRs were reproducibly associated with specific transcription factor binding. (A) Plot shows the number of elements discovered using dREG, but not found in DNase-I hypersensitivity or H3K27ac ChIP-seq (Y-axis) for PRO-seq datasets in K562 and GM12878 cells. The color denotes other functional marks intersecting sites discovered only using dREG. (B) Histogram representing the fraction of binding sites for 100 transcription factors supported by a dREG TIR that was not also discovered in DNase-I hypersensitivity data (i.e., dREG+ DHS- / dREG+ DHS+). Several of the outliers are shown. The color denotes whether the factor is a member of the Pol III preinitiation complex (green), Pol II preinitiation complex (red), associated with H3K9me3 or H3K27me3 heterochromatin (purple), or is a sequence specific transcription factor (blue). (C) Metaplots show the raw signal of DNase-I hypersensitivity, MNase-seq, and ChIP-seq for H3K4me1, H3K4me3, H3K27ac, and H3K27me3 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).

Supplementary Figure Legends

Supplemental Figure S1. PRO-seq datasets used during dREG model training. Heatmap shows Spearman's rank correlation (upper left) and raw gene-body correlations (lower right) between five PRO-seq and GRO-seq datasets used during dREG model training.

Supplemental Figure S2. dREG accurately detects the location of regulatory elements. Precision-recall curves show the precision (Y-axis; true positives/ [true positives + false positives]) and recall (X-axis; true positives / [true positives + false negatives]) of the new and previously published dREG models on the indicated dataset. The gold standard positive set was defined as DNase-I hypersensitive sites having a GRO-cap-defined transcription start site. The negative set was defined as sites lacking both GRO-cap and DNase-I hypersensitivity. The figure reflects all informative positions having a threshold of PRO-seq signal scored by dREG in the indicated dataset. Both datasets were held out during model training.

Supplemental Figure S3. Procedure for discovering transcription initiation regions (TIRs). We devised a new method for finding peaks of dREG signal, called transcription initiation regions (TIRs). dREG selects informative positions and predicts dREG signal, increases the local density in high scoring regions, and smoothes to identify local increases in signal intensity.

Supplemental Figure S4. Laplace distribution fits dREG scores in negative regions. Histogram shows the density (Y-axis) of dREG scores (X-axis) in regions that were defined as true negatives using orthogonal sources of genomic data. The lines represent the best fits to the distribution based on all true-negative sites (blue) or based on negative dREG scores (red).

Supplemental Figure S5. dREG identifies unidirectional TREs. WashU Epigenome Browser visualization of dREG signal, PRO-seq data, GRO-cap, DNase-I hypersensitivity, and H3K27ac ChIP-seq near the *ASTN1* and *HHAT* genes. Two TIRs (bottom) are supported by reads on only one strand. TIR indicated by the gray bar (top) lacks signal in both H3K27ac and DNase-I hypersensitivity Two TIRs (bottom) are supported by reads on only one strand.

Supplemental Figure S6. Novel elements discovered using dREG frequently overlap transcription factor ChIP-seq peaks. Plot shows the number of novel elements discovered using dREG, but not found in DNase-I hypersensitivity or H3K27ac ChIP-seq (Y-axis) for seven PRO-seq datasets. Six datasets were used from K562 cells (G1-7) and one dataset was used from GM12878. The number of novel dREG sites overlapping transcription factor ChIP-seq peaks (dark gray),H3K4me1 ChIP-seq peaks (gray), or not overlapping either (light gray) are shown.

Supplemental Figure S7. Novel dREG TIRs overlap local increases in histone marks associated with enhancers. Meta plots show the raw signal for H3K27ac, H3K4me1, and H3K4me3 near TIRs identified using dREG, but were not found in peak calls for H3K27ac or DNase-I hypersensitivity.

Supplemental Figure S8. Histogram shows the distribution of jaccard distance between dREG sites in K562 cell with respect to DNase-I hypersensitivity sites in ENCODE reference cell types. Jaccard distance was calculated for (A) all dREG sites in K562 cells, (B) dREG sites in K562 cells that do not overlap with DNase-I hypersensitive sites, and (C) dREG sites in K562 cells that do not overlap with DNase-I hypersensitive sites nor with H3K27ac ChIP-seq peaks. Major cell types among the outliers were colored and labeled.

Supplemental Figure S9. dREG+DHS- sites do not reflect clonal differences in between K562 cells grown by ENCODE and by our lab. Heatmaps show raw signal for two replicates of ATAC-seq data from K562 cells grown in our lab and clonally related to K562 cells used to produce PRO-seq data. Data is shown near dREG+DHS+ (n= 29,828) and dREG+DHS- (n= 7,350). Sites were ordered by dREG score.

Supplemental Figure S10. Accuracy of classifying DHS status of dREG TIRs. Receiver operating characteristic (ROC) plot shows the accuracy of predicting whether TIRs identified using dREG were also DHSs recognized by DNase-I hypersensitivity using either 100 transcription factor ChIP-seq datasets in K562 cells (black, auROC= 0.88) or the dREG score alone (gray, auROC= 0.75).

Supplementary Tables

Supplemental Table S1: Sources of PRO-seq data used in dREG model training and evaluation. Columns show the data identification number, cell line, number of mapped sequence reads, informative sites in which dREG scores were computed, reference, and the role (training or holdout validation) in this particular study.

Data ID	Cell line	Mapped Reads	Informative Sites	GEO	Reference	Role
G1	K562	374,946,808	17,120,769	GSM1480327	(Core et al. 2014)	Training
G2	K562	18,129,333	4,189,045	GSM1480325	(Core et al. 2014)	Training
G3	K562	57,520,888	9,201,344	SUBMIT	Herein	Training
G5	K562	71,452,942	8,678,964	GSE89230	(Vihervaara et al. 2017)	Training
G6	K562	26,972,822	5,890,140	GSM2545324	(Dukler et al. 2017)	Training
G7	K562	27,046,373	5,839,230	GSM2545325	(Dukler et al. 2017)	Holdout
GM	GM12878	105,936,649	9,738,488	GSM1480326	(Core et al. 2014)	Holdout

References

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.

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.

Andersson R, Sandelin A, Danko CG. 2015. A unified architecture of transcriptional regulatory elements. *Trends Genet* **31**: 426–433.

Azofeifa JG, Dowell RD. 2016. A generative model for the behavior of RNA polymerase. *Bioinformatics*. http://dx.doi.org/10.1093/bioinformatics/btw599.

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.

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.

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.

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.

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.

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

Chu T, Rice EJ, Booth GT, Salamanca HH, Wang Z, Core LJ, Longo SL, Corona RJ, Chin LS, Lis JT, et al. 2017. Chromatin run-on reveals nascent RNAs that differentiate normal and malignant brain tissue. *bioRxiv* 185991. https://www.biorxiv.org/content/early/2017/09/07/185991 (Accessed October 2, 2017).

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.

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.

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

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.

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*. http://dx.doi.org/10.1038/nmeth.4264 (Accessed April 18, 2017).

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*. http://dx.doi.org/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.

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

Gesing S, Lawrence K. Proceedings of the 50th Hawaii International Conference on System Sciences | 2017. https://scholarspace.manoa.hawaii.edu/bitstream/10125/41919/1/paper0770.pdf.

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

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.

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*. http://dx.doi.org/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.

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.

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.

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.

Lin CJ, Weng RC. 2004. Simple probabilistic predictions for support vector regression. *National Taiwan*

University, Taipei. 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, NY, USA.

Melgar MF, Collins FS, Sethupathy P. 2011. Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biol* **12**: 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*. http://genesdev.cshlp.org/content/early/2018/01/29/gad.308619.117.abstract.

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.

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

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.

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.

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.

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.

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.

Wang Z, Chu T, Choate LA, Danko CG. 2017. Rgtsvm: Support Vector Machines on a GPU in R. *arXiv [statML]*. http://arxiv.org/abs/1706.05544.

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.

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.