Identification of active transcriptional regulatory elements from GRO-seq data

Charles G Danko1–3, Stephanie L Hyland4, Leighton J Core5,11, Andre L Martins6, Colin T Waters5,11, Hyung Won Lee5, Vivian G Cheung7,8, W Lee Kraus9,10, John T Lis5 & Adam Siepel3,11

Modifications to the global run-on and sequencing (GRO-seq) protocol that enrich for 5′-capped RNAs can be used to reveal active transcriptional regulatory elements (TREs) with high accuracy. Here, we introduce discriminative regulatory-element detection from GRO-seq (dREG), a sensitive machine learning method that uses support vector regression to identify active TREs from GRO-seq data without requiring cap-based enrichment (https://github.com/Danko-Lab/dREG/). This approach allows TREs to be assayed together with gene expression levels and other transcriptional features in a single experiment. Predicted TREs are more enriched for several marks of transcriptional activation—including expression quantitative trait loci, disease-associated polymorphisms, acetylated histone 3 lysine 27 (H3K27ac) and transcription factor binding—than those identified by alternative functional assays. Using dREG, we surveyed TREs in eight human cell types and provide new insights into global patterns of TRE function.

TREs such as promoters, enhancers and insulators are critical components of the genetic regulatory programs of all organisms1. These elements regulate gene expression by facilitating or inhibiting chromatin decompaction, transcription initiation and the release of RNA polymerase II (Pol II) into productive elongation, as well as by maintaining the three-dimensional architecture of the nucleus.

The comprehensive identification of TREs has emerged as a primary challenge in genomic research. At present, these elements are most effectively identified using high-throughput genomic assays that provide indirect evidence of regulatory function; these include chromatin immunoprecipitation and sequencing (ChIP-seq) of bound transcription factors (TFs) or histone modifications as well as DNase I hypersensitivity and sequencing (DNase-seq)2–4. However, these methods have important limitations. For example, ChIP-seq requires a high-affinity antibody for the targeted TF or histone modification and must be performed separately for each target. Likewise, assays that measure chromatin accessibility or histone modifications provide only circumstantial evidence that the identified sequences are actively participating in transcriptional regulation5. Even STARR-seq (self-transcribing active regulatory region sequencing), a clever high-throughput reporter-gene assay, identifies regions that are inactive in situ because the assay is independent of native local chromatin structure and genomic context6.

It has become clear that a defining characteristic of active TREs is their association with local transcription. Enhancer-templated noncoding RNAs, or eRNAs, have recently been associated with thousands of stimulus-dependent enhancers7. Like active promoters, these enhancers exhibit transcription initiation in opposing directions on each strand, a phenomenon called divergent transcription8–10. This characteristic pattern can be a powerful tool for the identification of active TREs in a cell type–specific manner7,11–15. Methods that measure the production of nascent RNAs, such as GRO-seq8 and its successor, precision run-on and sequencing (PRO-seq)16, are particularly sensitive for detecting these transient enhancer-associated RNAs because they measure primary transcription before unstable RNAs are degraded by the exosome12,17. Recently, we showed that an extension of GRO-seq that enriches the nuclear run-on RNA pool for 5′-7meGTP-capped RNAs, called GRO-cap, further improves sensitivity for eRNAs and can be used to identify tens of thousands of transcribed enhancers and promoters across the genome18,19.

Here, we introduce dREG, a computational tool for accurately identifying transcribed TREs directly from standard GRO-seq or PRO-seq data (dREG is freely available at https://github.com/Danko-Lab/dREG and as Supplementary Software). dREG uses a multiscale summary of GRO-seq or PRO-seq read counts and
then employs support vector regression\textsuperscript{20} (SVR) to detect the characteristic patterns of transcription at TREs. The software allows for high-quality predictions of TREs for any cell type with existing GRO-seq or PRO-seq data. We applied the method to four cell types for which data were previously available and four for which we provide new data. Combining these predictions with data from the Encyclopedia of DNA Elements (ENCODE) project, we found that the predicted TREs fall into four major classes. The class distinguished by a strong dREG signal was also enriched for H3K27ac, TF binding, expression quantitative trait loci (eQTL) and genome-wide association study (GWAS)-associated single-nucleotide polymorphisms (SNPs), suggesting that TREs identified using dREG actively control cell type–specific transcription.

RESULTS

Regulatory element identification in eight cell types

The key to dREG is a feature vector that summarizes the patterns of aligned GRO-seq reads near each candidate element at multiple scales (Fig. 1a and Supplementary Fig. 1). This feature vector consists of read counts for windows ranging from 10 bp to 5 kb, standardized using the logistic function (Supplementary Fig. 2a). The feature vector is passed to an SVR, which scores sites with high PRO-seq signal for similarity to a training set of TREs. To train our classifier, we used TREs identified from GRO-cap data\textsuperscript{19} as positive examples and regions of matched PRO-seq signal intensity lacking additional marks associated with TREs as negative examples. After training and optimization of several tuning parameters (Supplementary Tables 1 and 2), the program displayed excellent performance when applied to PRO-seq data for K562 cells (area under the receiver operating characteristic (ROC) curve (AUC) = 0.99; Supplementary Fig. 2b).

We ran dREG to predict the location of TREs genome wide in K562 cells, adopting a prediction threshold that limits the genome-wide false discovery rate to 10%. At this threshold, we recovered 94% of 21,082 GRO-cap ‘paired’ sites (i.e., sites for which divergent Pol II initiation was detected in both directions) and 94% of 9,940 active transcription start sites (TSSs) detected by cap analysis of gene expression (CAGE) (Fig. 1b). Furthermore, we observed high sensitivity within the subsets of GRO-cap peaks overlapping gene bodies (84%) as well as promoters (95%) and enhancers (80%) defined by the ChromHMM software. We applied dREG to an independent cell type, GM12878 lymphoblastoid cells, without retraining the classifier. On the basis of GRO-cap data available for GM12878, dREG achieved similar performance in this cell type for all classes of regulatory elements tested (Fig. 1b), indicating that the method generalized well across cell types. Finally, we examined the sensitivity of dREG to sequencing depth and data quality and found that sensitivity was satisfactory with as few as 40 million mapped reads (Supplementary Fig. 3). Together, these findings demonstrated that dREG accurately identified active TREs across a broad spectrum of GRO-seq and PRO-seq data sets.

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

Four major classes of transcriptional regulatory elements

We compared dREG predictions with two complementary sets of TREs: (i) ChromHMMPredictions of promoters, enhancers and insulators\textsuperscript{24} and (ii) DNase I–hypersensitive sites (DHSs)\textsuperscript{25,26}. ChromHMMPredictions are based on genome-wide ChIP-seq assays targeting histone modifications and CTCF binding, whereas the DHSs identify regions of open chromatin where the DNA is accessible to DNase I cleavage. For the DHSs, we used high-confidence DNase I–accessible sites, defined as the

Figure 1 | dREG schematic and validation. (a) High PRO-seq signal intensity marks TREs and gene bodies. dREG is a shape detector trained to recognize the characteristic pattern of divergent transcription on the positive (red) and negative (blue) strand near TREs in PRO-seq data (1). After training, dREG can be used to identify TREs using a new PRO-seq data set (red peaks) (2). For comparison, a screenshot of ChromHMM–predicted promoters (red), enhancers (yellow) and insulators (dark blue) is shown (3). (b) Bar charts represent the genome-wide sensitivity of dREG for various classes of TRE at a 5% (line) or 10% (bar) false discovery rate in K562 and GM12878 cells. Classes of regulatory elements represent GRO-cap transcribed DHS (Transcribed DHS), transcription start sites identified by CAGE (CAGE TSS), histone acetylation associated with DHS (Acetyl DHS), GRO-cap transcribed ChromHMM promoters (Promoters), GRO-cap transcribed ChromHMM enhancers (Enhancers), GRO-cap TSSs inside annotated gene bodies (Gene body) and GRO-cap pairs. Pie charts represent the fraction of sites aligning within RefSeq TSSs, introns or intergenic regions in each validation set.
Functional properties of distinct TRE classes

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

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

Transcription factors activate and suppress eRNA synthesis

Fundamental differences among the four TRE classes are likely to be mirrored by patterns of TF binding. Therefore, we examined binding of 91 TFs for which high-resolution ChIP-seq data are available

![Image](https://example.com/image.png)
in K562 cells. Almost 70% of TREs in the MCO class do not bind any TFs (Fig. 2d), and most −dREG TREs bind small numbers (i.e., 1–10). dREG TREs, by contrast, display a striking enrichment for binding many TFs (18 on average, and 39% bind >20 TFs).

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

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

**Predicting transcription factor binding using dREG**

Having shown that TF binding is predictive of transcription initiation at TREs, we next addressed an inverse question: is transcription at TREs predictive of whether or not a TF is bound to DNA sequences matching its cognate motif? Most TFs bind only a small fraction of DNA sequences matching their motif30, making TF binding site prediction a challenging computational problem. We asked whether dREG could be useful as a surrogate for, or complement to, DNase-seq data, which are widely used to help identify TF binding31–33. As a proof of concept, we chose four transcriptional activators (NRF1, ELF1, SP1 and MAX) with a range of motif information contents34 and positive regression coefficients in the analysis described above, but otherwise selected at random. For all four TFs, we found that dREG scores alone predict the occupancy of motif matches with accuracy similar to that of the protein interaction quantitation (PIQ) program32, which makes use of DNase-seq data in predicting TF binding. For example, for ELF1 (Fig. 3), dREG produces a ROC score 3.8% lower than that of PIQ (AUC = 0.88 (dREG) and 0.92 (PIQ)), and both assays identify TF binding sites more accurately than motif matches alone (AUC = 0.67). Jointly modeling DNase-seq, dREG and the motif match score improves classification accuracy 2.6–6.6% (AUC = 0.94), exceeding the PIQ score in this task. Thus, dREG appears to be a useful complement to DNase-seq–based models of TF-DNA interaction for sequence-specific activators.

**Enrichment for eQTL and GWAS hits in dREG predictions**

We asked whether +dREG TREs contain the subset of open-chromatin sites that are actively regulating gene expression. To explore this possibility, we compared the density of eQTL identified in lymphoblastoid cell lines (LCLs)35 among +dREG, −dREG and MCO TREs. We found that +dREG TREs in GM12878 LCLs contain 6.4- to 26.3-fold higher eQTL densities in LCLs than in other classes of TRE (Fig. 4a) and account for 571 out of 755 of the eQTL that intersect with the functional marks considered here (~76%). This observation is partially explained by systematic variation in the composition of the LCL population (Supplementary Table 7).

**Figure 4 | eQTL and GWAS SNP enrichments in the four classes of functional element. (a) Density of eQTL (n = 755) per site found in +dREG (further divided into promoters and enhancers using ChromHMM), −dREG and MCO TREs. We found that +dREG TREs in GM12878 LCLs contain 6.4- to 26.3-fold higher eQTL densities in LCLs than in other classes of TRE (Fig. 4a) and account for 571 out of 755 of the eQTL that intersect with the functional marks considered here (~76%). This observation is partially explained by systematic variation in the composition of the LCL population (Supplementary Table 7). (b) Density of GWAS SNPs that correlate with cell type–specific phenotypes (autoimmune disorders) in the GM12878 B-cell line and primary CD4+ T cells.**
biases in eQTL density for gene promoters, yet if we focus on TREs associated with enhancers only, we still observe a 2.4- to 9.8-fold enrichment in eQTL densities in +dREG TREs relative to the –dREG, MCO and Insulator classes ($P < 2 \times 10^{-5}$; Fisher’s exact test). This residual enrichment cannot be explained by differences in the distributions of the distance of these site classes relative to TSS annotations (Supplementary Fig. 12) and suggests that dREG TREs are more likely to be actively regulating gene expression than are other TREs.

GWAS generally implicate long haplotype blocks of SNPs, making it challenging to identify SNPs that are causally associated with disease processes. Because dREG identifies a relatively small subset of active TREs, we speculated that it might be a useful tool for prioritizing GWAS SNPs for functional validation. To illustrate the utility of dREG for this application, we obtained a set of putatively functional GWAS SNPs35. We found that dREG sites detected in relevant primary cell types are substantially enriched in GWAS-associated SNPs. For example, SNPs associated with autoimmune disorders are enriched in dREG sites in CD4+ T cells and GM12878 lymphoblastoid cells (B cells), including SNPs for celiac disease (7.4- and 9.7-fold, respectively), rheumatoid arthritis (6.9- and 11.2-fold) and type 1 diabetes (4.5- and 5.3-fold). As we observed for eQTL, cell type–specific GWAS SNPs are found at higher densities in +dREG TREs compared with other functional classes (Fig. 4b). These observations demonstrate that dREG can be useful for prioritizing GWAS validation experiments.

DISCUSSION

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

By comparing dREG sites to other functional genomic assays, we demonstrated the existence of at least four major classes of TREs in human cells. Several lines of evidence, including enrichments for eQTL, transcriptional activators and histone acetylation, suggest that dREG identifies genomic sites that play a direct and active role in gene regulation. We also discovered three independent classes of TREs that are untranscribed (–dREG, MCO and Insulator) and were found to have relatively lower regulatory activities by several metrics. Insulators appeared to be a distinct functional class, as they were found to be depleted for the functional marks examined here, yet their relatively high evolutionary conservation (Supplementary Fig. 13), as well as prior work37, strongly suggest that insulators function in various aspects of cellular biology. The other two inactive classes of TREs, MCO and –dREG, appeared to have distinct mechanisms of inactivation, including the presence of a central nucleosome that occludes activator binding (MCO), and either DNA sequence-dependent binding by transcriptional repressors, or a lack of binding by transcriptional activators, at open chromatin (–dREG). In some cases, we observed changes between these TRE classes in different cell types (Supplementary Fig. 14 and Supplementary Table 4), suggesting that they might reflect intermediates in the assembly of active regulatory elements. Future studies are needed to identify the functional mechanisms that are responsible for the assembly and activation of TREs and to further elucidate the relationships and mechanistic transitions among these classes of regulatory elements.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. NCBI Gene Expression Omnibus: GSE66031 (CD4+ and Jurkat T cells, as well as dREG scores). Data are also available for visualization on the UCSC Genome Browser track hubs at http://www.charlesdanko.org/hub/dreg/.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

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

AUTHOR CONTRIBUTIONS

C.G.D. designed the dREG tool. C.G.D., A.S. and S.L.H. designed and implemented the software. C.G.D., S.L.H., A.L.M., L.J.C., J.T.L. and A.S. analyzed the data and interpreted the results. L.J.C., C.T.W., C.G.D., H.W.L., J.T.L., W.L.K. and V.G.C. contributed data and helped to troubleshoot experiments. C.G.D., A.S., J.T.L., L.J.C., S.L.H. and A.L.M. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? Mol. Cell 49, 825–837 (2013).
2. ENCODE Project Consortium. et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
3. Giresi, P.G. & Lieb, J.D. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). Methods 48, 233–239 (2009).
4. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013).
5. Cusanovich, D.A., Pavlovic, B., Pritchard, J.K. & Gilad, Y. The functional consequences of variation in transcription factor binding. PLoS Genet. 10, e1004226 (2014).
6. Arnold, C.D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339, 1074–1077 (2013).
7. Kim, T.-K. et al. Widespread transcription at neuronal activity-regulated enhancers. Nature 465, 182–187 (2010).
8. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322, 1845–1848 (2008).
9. Seita, A.C. et al. Divergent transcription from active promoters. Science 322, 1849–1851 (2008).
10. Kapranov, P. et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316, 1486–1488 (2007).
11. Hah, N. et al. A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. Cell 145, 622–634 (2011).
12. Hah, N., Murakami, S., Nagari, A., Danko, C. & Kraus, W.L. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res.* 23, 1210–1223 (2013).

13. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461 (2014).

14. Melgar, M.F., Collins, F.S. & Sethupathy, P. Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biol.* 12, R113 (2011).

15. Wu, H. et al. Tissue-specific RNA expression marks distant-acting developmental enhancers. *PLoS Genet.* 10, e1004610 (2014).

16. Kwak, H., Fuda, N.J., Core, L.J. & Lis, J.T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* 339, 950–953 (2013).

17. Preker, P. et al. RNA exosome depletion reveals transcription upstream of active human promoters. *Science* 322, 1851–1854 (2008).

18. Kruesi, W.S., Core, L.J., Waters, C.T., Lis, J.T. & Meyer, B.J. Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. *eLife* 2, e00808 (2013).

19. Core, L.J. et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* 46, 1311–1320 (2014).

20. Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A. & Vapnik, V. in *Proc Adv. Neural Inf. Process. Syst.* (NIPS, 1996).

21. Danko, C.G. et al. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol. Cell* 50, 212–222 (2013).

22. Luo, X., Chae, M., Krishnakumar, R., Danko, C.G. & Kraus, W.L. Dynamic reorganization of the AC16 cardiomyocyte transcriptome in response to TNFα signaling revealed by integrated genomic analyses. *BMC Genomics* 15, 155 (2014).

23. Wang, I.X. et al. RNA-DNA differences are generated in human cells within seconds after RNA exits polymerase II. *Cell Rep.* 6, 906–915 (2014).

24. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49 (2011).

25. John, S. et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* 43, 264–268 (2011).

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

27. Creighton, M.P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA* 107, 21931–21936 (2010).

28. Guertin, M.J., Martins, A.L., Siepel, A. & Lis, J.T. Accurate prediction of inducible transcription factor binding intensities in vivo. *PLoS Genet.* 8, e1002610 (2012).

29. Heintzman, N.D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39, 311–318 (2007).

30. Guertin, M.J. & Lis, J.T. Chromatin landscape dictates HSF binding to target DNA elements. *PLoS Genet.* 6, e1001114 (2010).

31. Pique-Regí, R. et al. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* 21, 447–455 (2011).

32. Sherwood, R.I. et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* 32, 171–178 (2014).

33. Jolma, A. et al. DNA-binding specificities of human transcription factors. *Cell* 152, 327–339 (2013).

34. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* 22, 1748–1759 (2012).

35. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511 (2013).

36. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* 22, 1748–1759 (2012).

37. Hadjur, S. et al. Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature* 460, 410–413 (2009).
ONLINE METHODS

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

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

GRO-seq feature vector. GRO-seq read counts were summarized in our model as a multiscale feature vector, as illustrated by the bar chart (Supplementary Fig. 2a). We counted GRO-seq reads that mapped in nonoverlapping windows on either side of a central base that met the signal intensity requirements (as described above). Our approach represented the genome at multiple scales (window sizes). For each scale, we counted reads in the specified number of nonoverlapping windows both upstream and downstream of the central base. Each scale could represent redundant information in the GRO-seq read counts. The final feature vector was constructed by concatenating the vectors representing read counts at each scale and strand. The specific parameters of the scales and number of windows at each scale were optimized using cross-validation (as described below and depicted in Supplementary Table 2).

Data standardization. GRO-seq data were standardized using the logistic function, \( F(t) \), with parameters \( \alpha \) and \( \beta \), as follows:

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

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

\[
\beta = x \max(t)
\]

\[
\alpha = \frac{1}{\beta} \log \left( \frac{1}{y} - 1 \right)
\]

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

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

Training the dREG support vector regression model. We fit an epsilon-SVR model using the e1071 R package\(^{38,39} \), which is based on the libsvm SVM implementation\(^{39} \). When training dREG, we assigned a label of 1 to sites intersecting both GRO-cap transcription start sites\(^\text{19} \) and high-confidence DHSs and excluded from the training set any sites intersecting a functional mark indicative of a regulatory element but not a GRO-cap peak (including ChromHMM enhancers or promoters). All other positions in the genome meeting the GRO-seq signal requirements (described above) were assigned a score of 0. The final SVR was trained on a matched set of 100,000 loci (composed of 50,000 positive and 50,000 negative examples) using PRO-seq data in K562 cells. Sites in the positive set (i.e., GRO-cap peaks) were chosen at random. When selecting the set of negative (i.e., non-transcription start site) examples, we chose 25% of sites to enrich for positions that were commonly associated with false positives during preliminary testing. These include 15% of the negative set that were selected to be within 1–5 kbp of the positive regions (to improve the resolution of dREG) and 10% in regions where the 3’ ends of annotated genes on opposite strands converged (to eliminate a common source of false positives). The remaining 75% of negative sites were selected at random from the set of positions across the genome meeting the GRO-seq signal requirements (described above).

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

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

dREG sensitivity to sequencing depth and library quality. To evaluate the sensitivity of dREG to sequencing depth, we subsampled the K562 data by removing reads at random from the BED files representing mapped reads. We ran the dREG algorithm as described, either with or without retraining the model on the reduced read depth (both are plotted in Supplementary Fig. 2). Artificial low-quality data sets were created by choosing genomic coordinates with mapped reads at random and redistributing their reads to neighboring sites in a 50-kbp (nonoverlapping) window. In each window, locations were retained with probability proportional to the original read density at that site. This procedure was designed to recreate the profile observed in low-quality data, in which large numbers of reads tend to align on a small number of positions, creating the appearance of ‘spikes’ when viewed on the genome browser. The “asymptotic unique reads” metric used to evaluate data quality was defined as the number of unique genomic coordinates in a GRO-seq library in the limit as the number of mapped reads approaches infinity. This value was estimated by subsampling the read depth and measuring the slope of the number of unique locations covered as a function of the library sequencing depth. We interpolated the number of uniquely covered genomic coordinates to 1% of its final value, assuming that the slope of the read depth did not change.

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

GRO-seq and PRO-seq library preparation. Extraction of primary CD4+ T cells from blood samples. Blood samples (80–100 mL) from three human individuals were collected at Gannett Health Services at Cornell University in compliance with Cornell IRB guidelines. Informed consent was obtained from all donors. Mononuclear cells were isolated using density gradient centrifugation, and CD4+ cells were extracted using CD4 from all donors. Mononuclear cells were isolated using density

Comparison to ChromHMM and DNase I data. We compared dREG TREs to ENCODE DNase I and ChromHMM data. For ChromHMM data, we selected the set of sites annotated as promoter, enhancer or insulator using data from GM12878, K562 (ref. 24 and GSE60456), HeLa40 or CD4+ T cells41. We collected ENCODE DNase I peak calls from the UW or Duke DNase I–seq protocol2 and selected peaks identified using both experimental assays. To compare different experimental assays, we merged sites identified by ChromHMM, DNase I-seq and dREG, and labeled each merged site according to experimental assays that identified it. TREs were subsequently divided into four independent, nonoverlapping classes on the basis of the set of experimental peak calls that they intersected. Site classes were defined as those sites that intersected: (i) dREG, DNase I–seq and ChromHMM (+dREG), (ii) ChromHMM insulators but not dREG (Insulator), (iii) DNase I–seq and ChromHMM but not dREG (−dREG), and (iv) ChromHMM but not DNase I–seq or dREG (modified chromatin only, or MCO). All operations in these analyses were performed using the Bedops42, BEDTools43 or bigWig software packages.

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

We modeled the presence or absence of dREG at a particular DNase I peak as the response in a logistic regression. Covariants consist of the presence or absence of each TF asayed in the cell type of interest. To determine the presence or absence of each TF, we collected uniform peak calls for all ChIP-seq data from the ENCODE project. For MCF-7 cells, ENCODE data were supplemented with a set of 37 TFs for which ChIP-chip data were available44. TFs having multiple ChIP biological replicates were associated with each peak if any of the replicates was enriched at that peak. The significance of the direction of effect for each TF on the presence of a dREG signal was determined using a
1,000-sample bootstrap, in which we chose one TF at random to omit from the regression analysis during each iteration. Supplementary Figure 10b plots the set of all TFs for each cell type whose direction of effect is consistent across each of the bootstrap iterations.

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

38. Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. & Weingessel, A. e1071: Misc Functions of the Department of Statistics (e1071) (TU Wien, 2010).
39. Chang, C.-C. & Lin, C.-J. in ACM Trans. Intell. Syst. Technol. 2, 27 (ACM, 2011).
40. Hoffman, M.M. et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 41, 827–841 (2013).
41. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. 28, 817–825 (2010).
42. Neph, S. et al. BEDOPS: high-performance genomic feature operations. Bioinformatics 28, 1919–1920 (2012).
43. Quinlan, A.R. & Hall, I.M. BEDtools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
44. Kittler, R. et al. A comprehensive nuclear receptor network for breast cancer cells. Cell Rep. 3, 538–551 (2013).