Accurate imputation of histone modifications using transcription

Zhong Wang[1], Alexandra G. Chivu[1], Lauren A. Choate[1], Edward J. Rice[1], Donald C. Miller[1], Tinyi Chu[1], Shao-Pei Chou[1], Nicole B. Kingsley[3], Jessica L. Petersen[4], Carrie J. Finno[5], Rebecca R. Bellone[3], Douglas F. Antczak[1], and Charles G. Danko[1,2,*]

[1] Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.
[2] Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.
[3] Veterinary Genetics Laboratory, School of Veterinary Medicine, University of California, Davis, CA 95616.
[4] Department of Animal Science, University of Nebraska-Lincoln, NE 68583.
[5] Department of Population Health and Reproduction, University of California, Davis, CA 95616.

[*] **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:

We trained a sensitive machine learning tool to infer the distribution of histone marks using maps of nascent transcription. Transcription captured the variation in active histone marks and complex chromatin states, like bivalent promoters, down to single-nucleosome resolution and at an accuracy that rivaled the correspondence between independent ChIP-seq experiments. The relationship between active histone marks and transcription was conserved in all cell types examined, allowing individual labs to annotate active functional elements in mammals with similar richness as major consortia. Using imputation as an interpretative tool uncovered cell-type specific differences in how the PRC2-dependent repressive mark, H3K27me3, corresponds to transcription, and revealed that transcription initiation requires both chromatin accessibility and an active chromatin environment demonstrating that initiation is less promiscuous than previously thought.

Introduction

The discovery that core histones are post-transcriptionally modified fueled nearly six decades of speculation about the role that histone modifications play in transcriptional regulation by RNA polymerase II (Pol II) (*1*). Many of the best-studied histone modifications are deeply conserved within eukaryotes, indicating important functional roles (*2–4*). Indeed numerous examples illustrate how the disruption of histone modifications, or their associated writer and eraser enzymes, lead to defects in transcription and cellular phenotypes (*5–8*). Histone modifications are found in highly stereotyped patterns across functional elements, including promoters, enhancers, and over the body of transcribed genes and non-coding RNAs. Promoters and enhancers are associated with a pattern of chromatin organization consisting of a nucleosome depleted core flanked by +1 and -1 nucleosomes marked with specific histone modifications, including histone 3 lysine 4 trimethylation (H3K4me3), lysine 27 acetylation (H3K27ac), and lysine 4 monomethylation (H3K4me1) (*9–15*). Actively transcribed gene bodies are marked by histone 3 lysine 36 trimethylation (H3K36me3), lysine 79 trimethylation (H3K79me3), and histone 4 lysine 20 monomethylation (H4K20me1) (*9*, *16*). Finally, two modifications are enriched in transcriptionally depleted regions, including histone 3 lysine 27 trimethylation (H3K27me3) and lysine 9 trimethylation (H3K9me3) (*16*).

The stereotyped pattern of histone modifications makes them useful in the annotation of functional elements in eukaryotic genomes. A collection of 11 histone modifications, first used to broadly analyze different cell types by the ENCODE project, has been applied to identify functional elements in metazoans (*2*, *16–18*). Numerous studies have used histone modifications to reveal the location of enhancers, lincRNAs, and other types of functional elements (*10*, *19–22*). Histone modifications aid in interpreting phenotype-associated genetic variation (*23*, *24*) and discovering molecular changes in disease (*25–30*). Likewise, histone modifications have been proposed for applications in selecting individualized therapeutic strategies (*31*). Applications such as these have led to genome annotation efforts in a myriad of mammals (*32*), plants (*33–35*), and other eukaryotic organisms (*36*). New annotation efforts will be launched alongside moonshot goals to sequence and annotate genomes across the tree of life (*37*). However, despite extensive efforts to decrease cost and improve the throughput of experimental methods (*38–42*), and to “impute” (or guess) the abundance of marks that were not directly observed (*43*, *44*), genome annotation still requires concerted efforts of large, well-funded, interdisciplinary consortia.

Despite the widespread use of histone modifications in genome annotation, the precise nature of the relationship between histone modifications and transcription remains unknown. As a topic of singular importance, the extent to which specific histone modifications have a direct role in transcriptional regulation or an indirect role as “cogs” in the transcription machinery, remains debated (*45*). Certain combinations of histone modifications, most notably the bivalent chromatin signature consisting of H3K4me3 and H3K27me3, are speculated to mark specific genes for transcriptional activation in later developmental stages (*46*). In another example, the balance between H3K4me1 and H3K4me3, which has long been known to correlate with enhancer and promoter activity (*10*), has been proposed to establish these two regulatory roles (*47*). Another question which remains heavily debated is the extent to which distinct histone modifications mark DNA sequence elements that otherwise have similar functional activities. H3K27ac, H3K64ac, and H3K122ac are all reported to denote distinct sets of enhancers (*48*). The nature of the quantitative relationship between transcription and histone modification lies at the crux of both of

these open questions. Large amounts of histone modification that are not explained by current transcription events leaves open the possibility that marks serve as a placeholder which might contribute to transcriptional regulation in a distinct cellular state. Alternatively, if histone modifications serve as “cogs” in the transcriptional machinery, we might expect that they are nearly completely correlated with on-going transcription.

Here we trained sensitive machine learning models that decompose maps of primary transcription into ChIP-seq profiles representing 10 distinct histone modifications. We show that transcription measured using precision run-on and sequencing (PRO-seq) (*49–51*) can recover the pattern of active histone modifications at nucleosome resolution and with an accuracy that rivals the correlation between independent ChIP-seq experiments in holdout cell types. Surprisingly, transcription also recovered the distribution of the repressive chromatin mark H3K27me3. However, unlike active marks, H3K27me3 was found in two distinct patterns in different cell types: one pattern in fully differentiated cells covered broad regions with low levels of transcription, and a second pattern in stem cells was associated with transcription start sites of weakly transcribed genes. Transcription accurately identified bivalent promoters in embryonic stem cells. Although transcription accurately predicted nearly all histone modifications, we found a subset of DNase-I hypersensitive sites that were refractory to prediction. Collectively, our results (1) support models in which histone modifications are “cogs” with a supportive role, rather than a direct regulatory role, in transcription, (2) preclude models in which transcription initiates pervasively as a consequence of open chromatin, and (3) provide a new strategy for genome annotation using a single functional assay that is tractable for a single lab to perform.

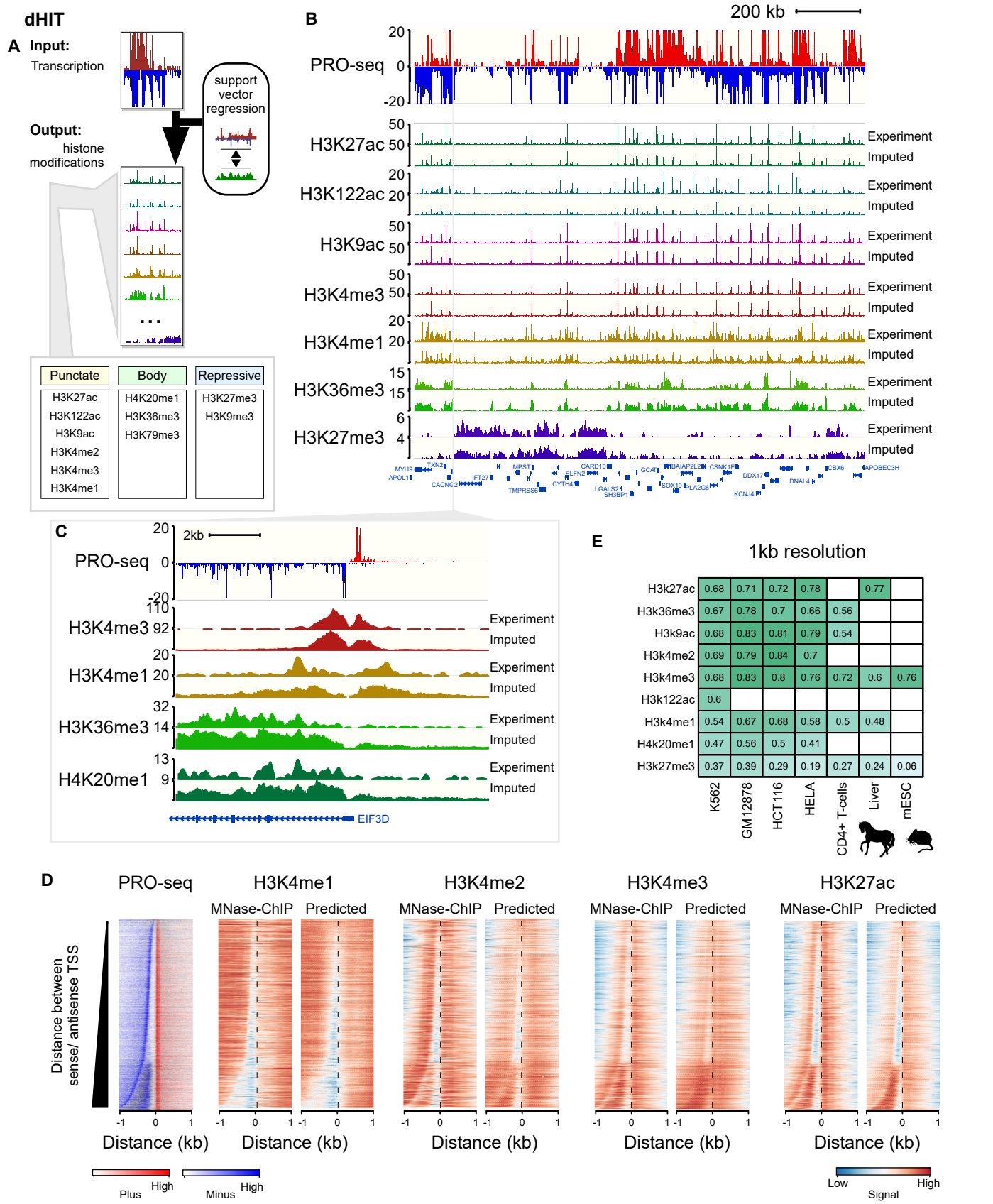

Figure 1. dHIT imputes histone modifications using nascent transcription.
(A) Schematic of the dHIT algorithm. PRO-seq and ChIP-seq data in K562 cells was used to train a support vector regression (SVR) classifier to impute 10 different histone modifications.
(B) Genome-browser compares experimental and predicted histone modifications on a holdout chromosome (chr22). PRO-seq data used to generate each imputation is shown on top.
(C) Genome-browser compares experimental and predicted histone marks near the promoter of EIF3D. PRO-seq data used to generate each imputation is shown on top.
(D) Heatmaps show the distribution of transcription (left) and histone modifications (right) measured using MNase ChIP-seq or predicted using transcription. Rows represent transcription initiation domains in K562 cells. Heatmaps were ordered by the distance between the most frequently used TSS in each transcription initiation domain on the plus and minus strand.
(E) Pearson's correlation between predicted and expected values for nine histone modifications. Values are computed on the holdout chromosome (chr22) in humans, chr1 in horse, and chr1 in mice. Empty cells indicate that no experimental data is available for comparison in the cell type shown.

Results

Accurate imputation of histone modifications at nucleosome resolution using nascent transcription

The primary goal of discriminative histone imputation using transcription (dHIT) is to use the shape and abundance of RNA polymerase, measured using PRO-seq, GRO-seq, or ChRO-seq data (henceforth referred to simply as PRO-seq), to impute the level of histone modifications genome-wide. dHIT passes transformed PRO-seq data to a support vector regression (SVR) (see Methods). During a training phase, the SVR optimized a function which mapped PRO-seq signal to the quantity of ChIP-seq signal at each position of the genome (**Fig. 1a**; see Methods). Once a dHIT model was trained using existing ChIP-seq data, it can impute steady state histone modifications in any cell type, provided that the relationship between histone modification and transcription is preserved. We trained dHIT to impute the levels of 10 different histone modifications that are widely deployed to analyze chromatin state (**Fig. 1a**) (*16*, *17*, *52*, *53*). To avoid overfitting to batch-specific features in a single run-on and sequencing dataset (*53*), training was performed using seven datasets in K562 cells that exemplify the range of variation commonly observed between data in library quality, sequencing depth, run-on strategy (PRO-seq or GRO-seq), and pausing index (**Supplementary Table 2**).

We evaluated the accuracy of each dHIT imputation model on a holdout chromosome in one of the training datasets (chr22; **Fig. 1b; Supplementary Fig. 1**). Histone modification signal intensity imputed using dHIT was highly correlated with experimental data for a variety of marks with different genomic distributions, including marks with focused signal on promoter or enhancer regions (e.g., H3K4me1/2/3, H3K27ac, H3K9ac), marks spread across active gene bodies (H3K36me3, H4K20me1), and over large domains of PRC2-dependent repressive heterochromatin (H3K27me3). In addition to well-studied and commonly used histone marks, we also obtained a high degree of correspondence for less widely studied histone modifications. For instance, acetylation of lysine 122 (H3K122ac), a residue on the lateral surface of the H3 globular domain (*54*), was reported to mark a distinct set of enhancers compared with H3K27ac (*48*). Nevertheless, dHIT models trained to impute H3K122ac had a high correlation on the holdout chromosome (**Fig. 1b**). Of the marks for which we attempted to train models, only the repressive mark H3K9me3 did not perform well.

In many cases imputation captured the fine-scale distribution of histone mark signal near the transcription start site (TSS) of annotated genes or enhancers (**Fig. 1c; Supplementary Fig. 2**). To explore the limit of the resolution for histone mark imputation using transcription, we obtained new ChIP-seq data for four active marks whose distribution correlates with enhancers and promoters (H3K4me1, H3K4me2, H3K4me3, and H3K27ac) at nucleosome resolution by using MNase to fragment DNA. We also analyzed two gene body marks, H3K36me3 and H3K79me3, that are depleted near the TSS (*9*). We trained new SVR models in K562 cells that take advantage of the higher-resolution MNase ChIP-seq data, excluding chromosome 22 as a holdout to confirm a high correlation (**Supplementary Fig. 3**). Examination of genome-browser traces near the TSS of genes on the holdout chromosome confirmed that dHIT could impute active marks with a high resolution (**Supplementary Fig. 4**).

Genome-wide, several aspects of chromatin organization were correlated with the precise location of TSSs and Pol II pause sites. These features are readily apparent when sorting by the

distance between the strongest TSS on the plus and minus strand (*13–15*) (**Fig. 1d**). First, when the distance between the maximal sense and divergent TSS was larger than ~600 bp, we observed a nucleosome between start sites that was marked predominantly with H3K4me3 and H3K27ac, but depleted for H3K4me1. Second, H3K4me3 and H3K27ac signal were highest on the +1 nucleosome, as well as the nucleosome found inside of the initiation domain. Third, H3K4me2 was highest on the -1 nucleosome. Fourth, gene body marks H3K36me3 and H3K79me2 were depleted at the promoter, and enriched in the body of transcribed genes (**Supplementary Fig. 5**). Each of these correlations between TSSs and chromatin were also observed to varying degrees in genome-wide imputation in K562 cells (**Fig. 1d**), and imputation data in a complete holdout cell type, GM12878 (**Supplementary Fig. 6**). Thus, dHIT receovered the placement of nucleosomes constrained to ordered arrays whose position correlated with transcription initiation.

Histone imputation can often generalize to new cell types, with important exceptions

We asked whether the relationship between transcription and histone modifications is a general feature that is shared across mammalian cell types. We computed the correlation between imputed and experimental histone marks in five holdout datasets without retraining the model. Holdout datasets were selected to represent a range of cultured cells, primary cells, and tissues from multiple mammalian species (**Supplementary Table 1**). Holdout data also explored a range of technical variation in both run-on assays and ChIP-seq validation experiments, including data collected by different labs, different fragmentation methods, and, for the run-on experiment, using different variants of a run-on assay (**Supplementary Table 1-2**).

Despite a variety of technical differences between ChIP-seq in holdout cell types and the ENCODE training dataset (**Supplementary Table 1**), active marks were recovered with a similar fidelity in holdout cell types as observed for K562 (**Fig 1e, Supplementary Fig. 7a-c**). At 1 kb resolution, dHIT recovered active marks indicative of promoters, enhancers, and gene-bodies at a median Pearson correlation of 0.73 (Pearson's R = 0.38-0.84), substantially higher than copying values from the training dataset (**Supplementary Fig. 7d**). Lower correlations were generally observed when the experimental ChIP-seq data (certain CD4+ T-cell datasets) or the PRO-seq data (HeLa) had fewer sequenced reads. For marks that were distributed across broad genomic regions (H3K36me3 and H3K27me3), dHIT imputation identified broad regions of high signal with reasonably high accuracy, but smoothed over fine-scale variation (**Fig. 1e; Supplementary Fig. 1a-b**). Thus, dHIT accurately recovered the distribution of active histone marks in a way that generalized to all new cell types examined here.

To place correlations observed between imputed and experimental data into context, we compared correlations between imputed and ChIP-seq data to those observed between different ChIP-seq datasets in K562 and GM12878, two cell lines for which multiple experimental datasets exist for each mark. For active marks, and for H3K27me3, correlations between dHIT imputation and experimental data were often within the range observed between experimental datasets (**Supplementary Fig. 8**). These data suggest that imputation performs similar to ChIP-seq experimental replication.

We identified one important exception on the extent to which histone imputation generalized between cell types. The repressive mark, H3K27me3, had a higher variation among

holdout samples than observed for active marks, with a reasonably high correlation in K562, GM12878, and horse liver (median Pearson's R = 0.31), but a much weaker correlation in mouse embryonic stem cells (mESCs, R = 0.06). Examination of signal tracks showed that the distribution of H3K27me3 differed dramatically from the K562 cell dataset. Whereas in K562, H3K27me3 was broadly distributed across large regions that span tens to hundreds of kilobases, in mESCs H3K27me3 appears over punctate regions near weakly transcribed promoters (**Fig. 2A**). Analysis of H3K27me3 in 375 high-quality samples showed that stem, germ, and certain progenitor cells usually had a punctate pattern, whereas most somatic cell types had the broadly distributed pattern (**Supplemental Fig. 9**). Although we cannot completely discount the possibility that technical factors contribute to this difference in H3K27me3 distribution (*55–57*), we noted the discovery of both punctate and broad H3K27me3 distributions even when libraries were prepared by the same lab (**Supplemental Fig. 9**; and ref (*58*)). These observations suggest that H3K27me3 can occur in at least two distinct profiles, both of which appear linked to features of active transcription.

Imputation of bivalent promoters and other chromatin "states"

We next asked whether dHIT could impute complex chromatin states consisting of multiple histone marks. The bivalent chromatin state consists of nucleosomes near gene promoters marked with H3K4me3 and H3K27me3, which are associated with gene activation and repression, respectively (*46*). The bivalent chromatin state is best described in ESCs and germ cells, and tends to mark the promoter of genes with developmental importance (*46*, *59–61*). We used dHIT models trained on ENCODE ChIP-seq data in K562 cells to impute H3K4me3 and H3K27me3 based on a GRO-seq dataset in mESCs (*62*). Despite the difference in H3K27me3 distribution between K562 and mESCs (noted above), we observed a strong tendency for bivalent promoters in mESCs to retain weak activity while occurring inside regions of high H3K27me3. For example, *Prox1* is located inside of a broad H3K27me3 domain, which the imputation recognized based on low transcription levels from *Prox1* and surrounding regions (**Fig. 2A**). Nevertheless, the *Prox1* promoter is weakly transcribed, and the imputation correctly places a peak of H3K4me3. The general pattern where bivalent genes were transcribed within H3K27me3-high domains was consistent enough that nearly 80% of bivalent gene promoters could be separated from promoters associated with either mark alone, or neither mark, with a precision of 80%, using a random forest on holdout data (**Fig. 2B**). Notably, promoters that carry the H3K27me3 mark in mESCs were distinguished accurately from those carrying no mark, indicating that promoters carrying the H3K27me3 mark are generally not transcriptionally silent. Taken together, these results demonstrate that bivalent genes can be identified based on the distribution of active transcription alone.

To generalize our observations on bivalent genes to other chromatin states, we asked whether chromatin marks imputed using transcription can infer chromatin states defined by chromHMM (*63*). We used a previously reported chromHMM model that defined 18 distinct chromatin states using ChIP-seq data from six marks for which we trained imputation models (H3K4me3, H3K27ac, H3K4me1, H3K36me3, H3K9me3, and H3K27me3) (*17*, *64*). Examination on the WashU epigenome browser revealed that chromatin states were highly similar, regardless of whether they were defined using experimental data from ENCODE or dHIT imputation (**Fig. 3A**). Each chromHMM state was enriched near the TSS of annotated genes to a similar degree

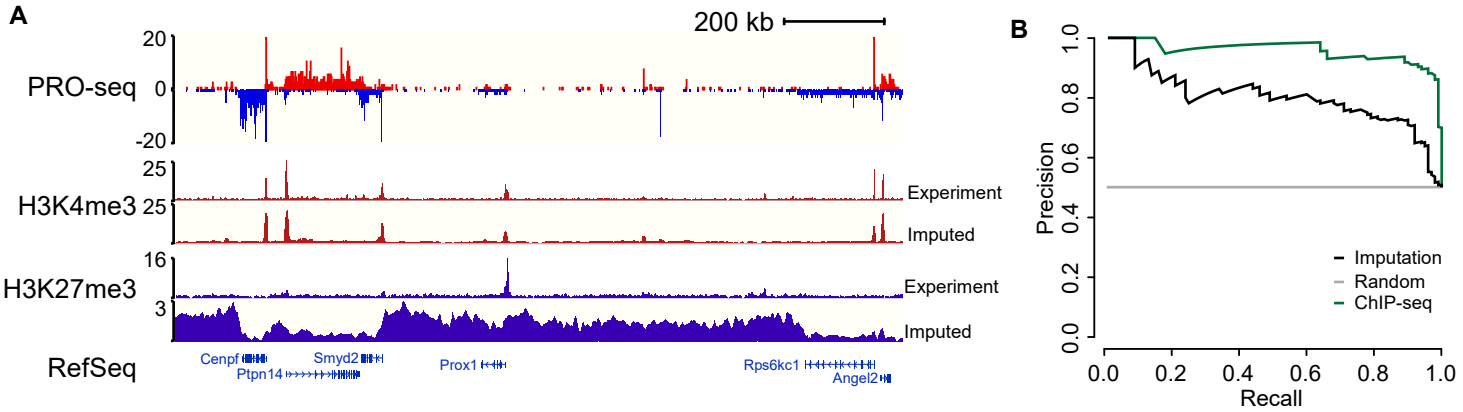

Figure 2. dHIT identifies bivalent H3K4me3, H3K27me3 marked genes.
(A) Genome-browser shows PRO-seq data and histone modification data measured by ChIP-seq or predicted using PRO-seq in the Prox1 locus. Prox1 is marked by bivalent H3K4me3 and H3K27me3 histone modifications in mESCs.
(B) Precision recall curve illustrates the accuracy of bivalent gene classification by a random forest classifier using ChIP-seq data (green) or dHIT imputation (black). The gray line denotes random classification. Classification was performed on a matched set of TSSs (50% bivalent, 50% not bivalent) that was held out during random forest training.

when using experimental or dHIT imputed data as input (**Fig 3B, Supplementary Fig. 10**). To determine the concordance expected between chromatin states defined using independent collections of experimental data, we applied chromHMM to a distinct collection of ChIP-seq data in the same cell type (**Supplementary Table 1**). Jaccard distances between imputed and experimental data were highly correlated with those observed between other ChIP-seq datasets (**Fig. 3C, Supplementary Fig. 11**). Taken together, these results suggest that transcription alone is sufficient to infer complex chromatin states, especially active chromatin states.

Genome annotation using a single functional assay

To illustrate the utility of chromHMM using dHIT imputed data, we analyzed chromatin states in 20 primary glioblastomas recently analyzed using ChRO-seq (*51*). ChromHMM analysis revealed both broad similarities and putative differences in chromatin states between different GBMs. For instance, a subset of samples were characterized by active transcription in *ADM2* and *MIOX*, and a subset carried marks of the polycomb repressed states (**Fig. 3D**). Analysis of this same dataset would have taken 120 separate ChIP-seq experiments (**Fig. 3E**). Moreover, sample quantity and quality would have limited independent ChIP-seq experiments to fewer than 9 modifications, especially if using conventional ChIP-seq instead of low-input alternatives (*40*).

Another critical application is to efficiently annotate functional elements in diverse tissues from understudied species. To illustrate this use of dHIT, we obtained ChRO-seq data from the liver of two horses that serve as the focus of the Functional Annotation of Animal Genomes (FAANG) project (*32*, *65*, *66*). Using dHIT and models trained in K562 cells to impute histone modifications, we obtained patterns of H3K27ac, H3K4me3, H3K4me1, and H3K27me3 that were highly correlated with experimental data from the same tissues (**Fig. 1E; Supplementary Fig. 12A**). In addition to those histone marks measured by FAANG, dHIT also imputed patterns for five additional histone marks, providing new information about chromatin state that was not obtained by the FAANG consortium. Next, we prepared ChRO-seq libraries in nine tissues taken from mice (**Supplementary Fig. 12B; Supplementary Fig. 13**). After accounting for biological replication in this experiment (7 replicates x 9 tissues x 9 histone marks), it would have taken 567 ChIP-seq assays to prepare this same dataset.

Thus, using dHIT to interpret ChRO-seq data provides individual labs access to consortium scale annotation of functional elements in mammalian genomes, and this information has potential applications in precision diagnostic medicine and genome annotation.

Transcription predicts active marks better than combinations of multiple ChIP-seq datasets

Many applications would benefit from data collected using the assay that provides the most information about chromatin state using the fewest experiments. To identify the best assay for this task, we trained SVR imputation models that use either PRO-seq or ChIP-seq data for each of the 10 different histone marks to predict each of the other experimental ChIP-seq datasets. We evaluated performance using the L1 norm, defined as the average of the median centered distance between imputed and experimental marks in 10bp windows on a holdout chromosome (**see Methods**). PRO-seq achieved a lower median L1 norm than any other individual assay by a fairly wide margin (**Fig. 4, black**). Examining imputation tracks led us to attribute the relative success of PRO-seq to two features. First, PRO-seq captured the boundaries

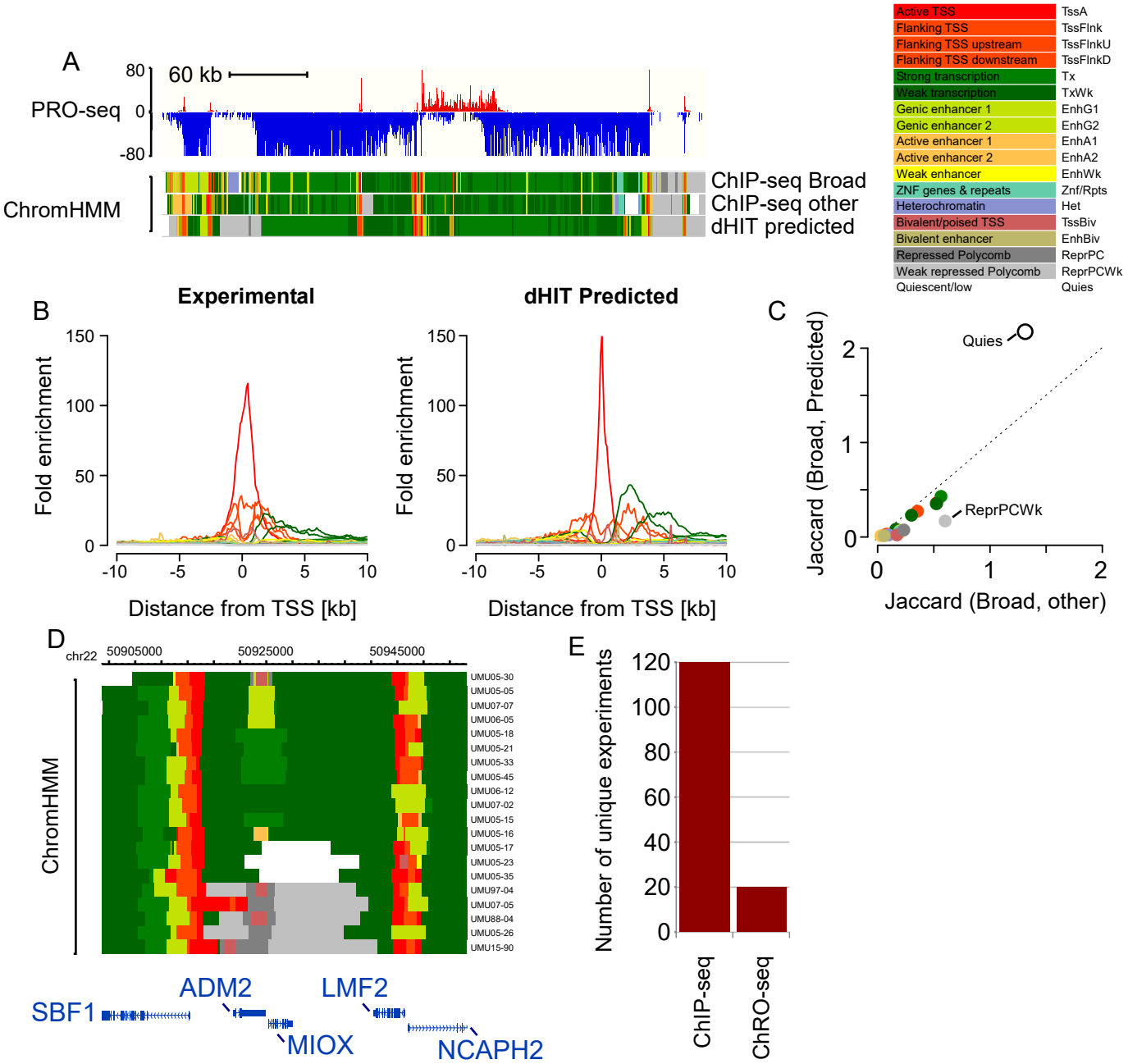

Figure 3. Inference of chromatin states defined by chromHMM using transcription.
(A) Genome-browser in K562 cells shows 18 state chromHMM model using either ChIP-seq data used to train the model (Broad), alternative ChIP-seq data in K562 (other), or based on imputation (dHIT predicted). PRO-seq data used during dHIT imputation is shown on top.
(B) Enrichment in each of 18 chromatin states as a function of distance from RefSeq annotated transcription start sites.
(C) Jaccard distance between chromHMM states inferred using ChIP-seq from Broad and predicted data (y-axis) and states inferred using ChIP-seq from Broad and an alternative compilation of high-quality ChIP-seq data (x-axis).
(D) ChromHMM states inferred using ChRO-seq data from 20 primary glioblastomas.
(E) The number of unique ChRO-seq or ChIP-seq libraries required to analyze chromatin states in 20 primary glioblastomas.

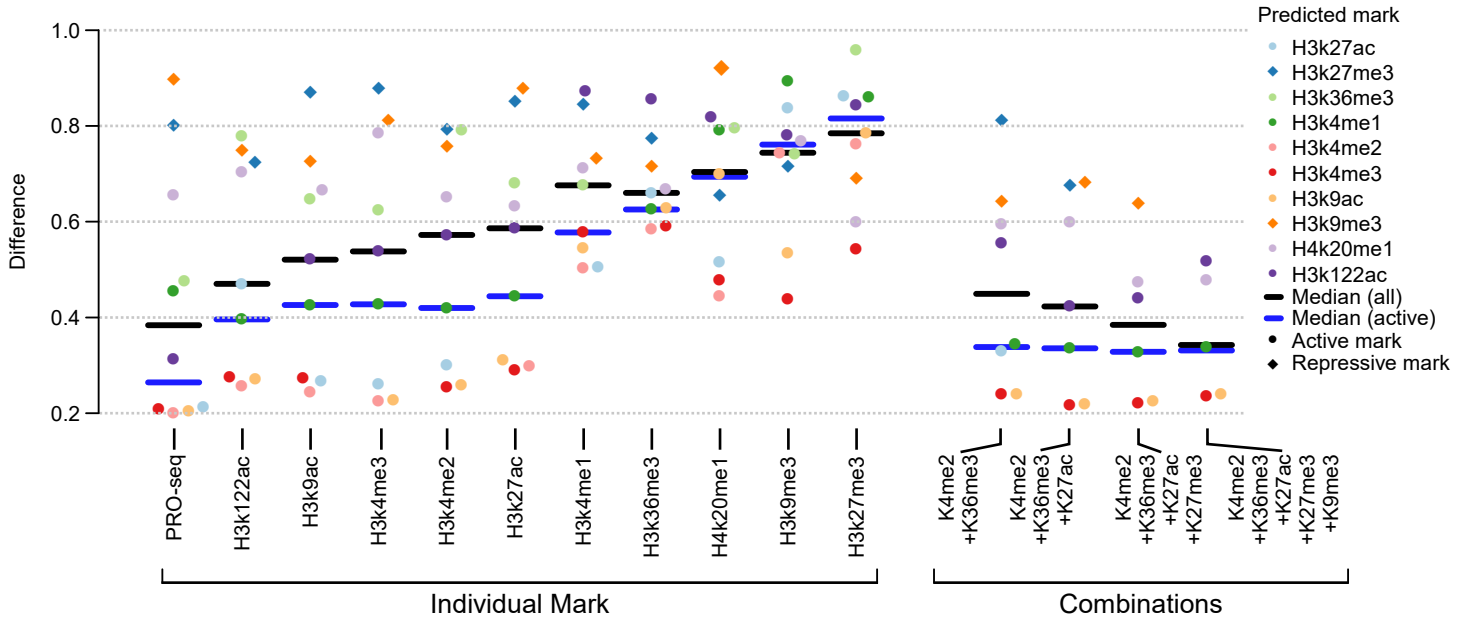

Figure 4. Transcription infers unobserved chromatin marks more accurately than other ChIP-seq data.
The mean difference between predicted and experimental ChIP-seq data on a holdout chromosome (chr22) (Y-axis). SVR models were trained using the indicated experimental mark (left) or the indicated combination of histone marks (right).

and direction of gene bodies in a manner that could not be achieved by other marks (**Supplementary Fig. 14A**). Second, PRO-seq was the most accurate at recovering the relative distribution of signal intensities in focal marks near the TSS (**Supplementary Fig. 14B**). Thus, we conclude that PRO-seq improved the accuracy of histone mark imputation by encoding signals from multiple functional regions and by improving spatial resolution compared with ChIP-seq data.

We next trained SVRs using combinations of multiple histone marks to determine whether training on multiple experimental datasets improved imputation performance. Because the space of potential histone mark combinations was extremely large and training was time consuming, we manually selected combinations of histone marks that provide orthogonal information to each other. We first selected H3K4me2 and H3K36me3, which combines a mark denoting promoter/ enhancer regions with one denoting gene bodies (*9*, *67*). The pair of experimental datasets together slightly improved the imputation of most ChIP-seq marks relative to the best performing individual mark, for instance H3K4me1 and H3K9me3 (**Fig. 4, right**). However, the median L1 norm was still worse than PRO-seq. We tested combinations where larger numbers of marks were observed by adding H3K27ac, H3K27me3, and H3K9me3, and evaluating the accuracy with which imputation could recover experimental marks. In most cases using additional marks made only a minor difference in performance (**Fig. 4**). Although we observed a decrease in the median accuracy using multiple marks (**Fig. 4, black**), this was explained largely by replacing the worst performing marks with experimental data. Our results therefore suggest that capturing information about the relative position of TSSs and gene bodies was enough to saturate performance using our current framework. Thus, PRO-seq data predicted ChIP-seq signals of unobserved active histone marks at least as well as ChIP-seq data for five different histone marks.

Chromatin accessibility is not sufficient for transcription initiation

In classical models, gene regulation in eukaryotes primarily involves removing nucleosomes from the promoter of active genes, at which point Pol II initiates in an indiscriminate manner (*68*). More recent studies support such accessibility models by observations that Pol II initiates at nearly all types of open chromatin regions (*69*, *70*). We used imputation to ask whether transcription initiates stochastically at any accessible DNA sequence, as previously proposed (*69*, *70*), or specifically at well-defined enhancer and promoter regions. We trained an SVR to impute smoothed DNase-I-seq data using PRO-seq in the same manner as we used for histone modifications. The best model predicted a holdout chromosome (chr22) with an accuracy of 0.61 or 0.77 (R^2) at resolutions of 100 and 1,000 bp (**Fig. 5A-C**), consistent with a strong correlation between chromatin accessibility and transcription initiation (*13*, *69*).

Nevertheless a substantial number of DNase-I hypersensitive sites had predicted values near zero, indicating a subset of sites that were refractory to prediction (**Fig. 5a, red arrow**). Intersecting experimental and imputed DNase-I-seq intensities (100 bp windows) with ChIP-seq data revealed that poorly performing windows were enriched for binding of CTCF (**Fig. 5c**), or to a lesser extent for transcriptional repressors and co-repressors such as REST, RFX5, or HDAC2 (**Supplementary Fig. 15**). In contrast H3K27ac peaks were depleted for 100 bp windows with poor matches between experimental and imputed DNase-I-seq data (**Fig. 5b**).

To confirm the absence of transcription, and further investigate the chromatin environment at each of these sites, we divided 100 bp windows into those in which DNase-I-seq was predicted well by PRO-seq, and those for which it was predicted poorly (**Fig. 5b-c, red boxes**). Windows in

which DNase-I-seq was predicted well by dHIT for both CTCF and H3K27ac had a high signal for transcription initiation in GRO-cap data and active histone modifications (H3K27ac, H3K4me3, and H3K4me1) (**Fig. 5d-k**). Windows in which DNase-I-seq was predicted poorly had a high CTCF signal, but virtually no evidence of transcription initiation based on GRO-cap, and weak signal for active histone modifications (**Fig. 5l-o**). Yet despite substantial differences in histone marks, the quantity of DNase-I-seq signal was similar in these regions (**Fig. 5g,k,o**). Thus, a substantial portion of DNase-I accessible regions show no robust evidence of transcription initiation. Our analysis supports a model in which both chromatin accessibility and chromatin environment are important factors to facilitate transcription initiation by Pol II.

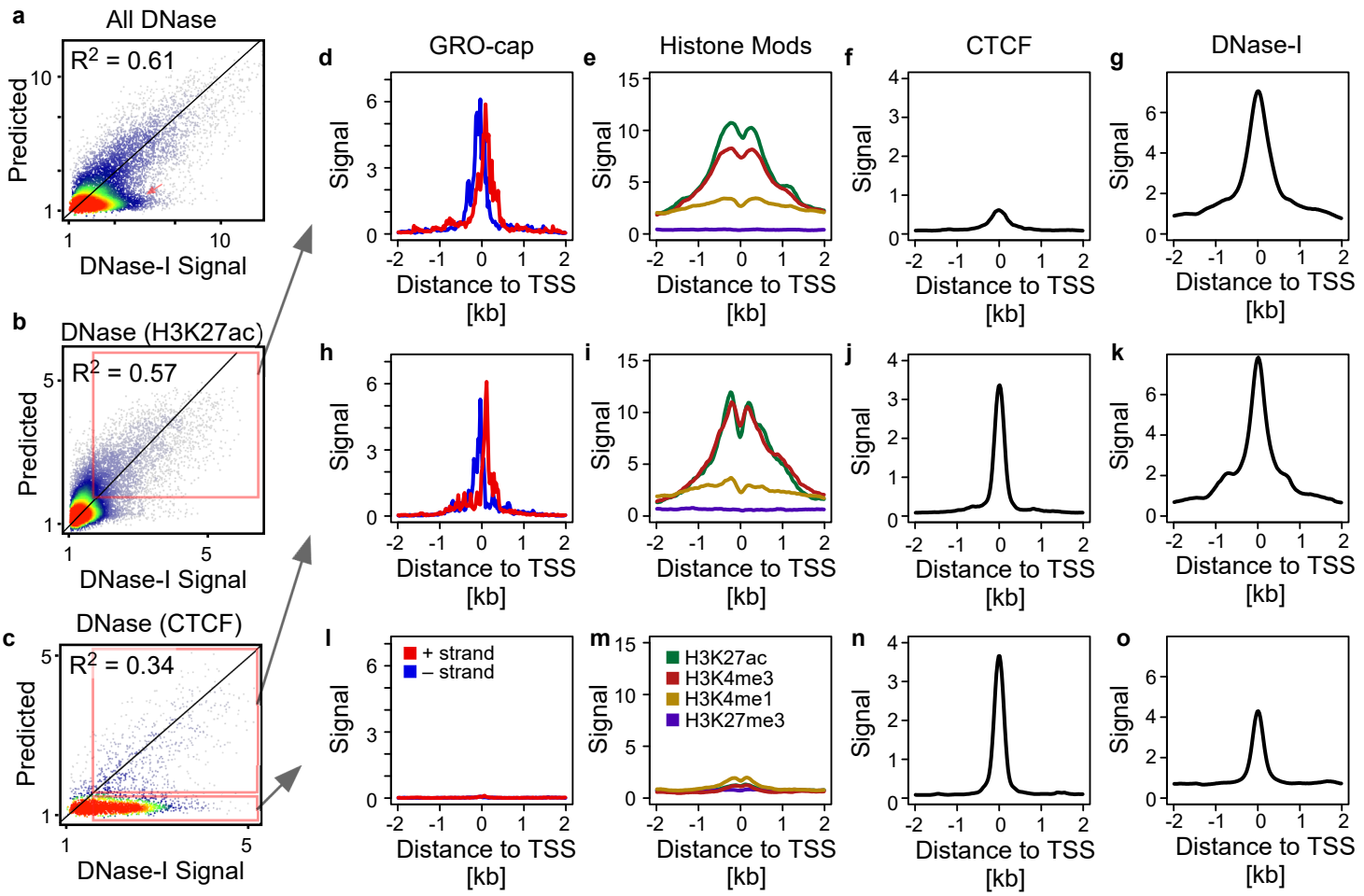

Figure 5. Chromatin accessibility is not sufficient for transcription initiation.
(A-C) Scatterplots show experimental DNase-I hypersensitivity (x-axis) as a function of predicted DNase-I hypersensitivity (y-axis) in 100 bp windows intersected with DNase-I hypersensitive sites (A), H3K27ac peaks (B), or CTCF peaks (C) on a holdout chromosome (chr22).
(D-G) Meta plots show GRO-cap, histone modifications, CTCF binding, and DNase-I hypersensitivity signal near H3K27ac peaks in which DNase-I hypersensitivity signal was accurately predicted by transcription.
(H-K) Meta plots show GRO-cap, histone modifications, CTCF binding, and DNase-I hypersensitivity signal near CTCF peaks in which DNase-I hypersensitivity signal was accurately predicted by transcription.
(L-O) Meta plots show GRO-cap, histone modifications, CTCF binding, and DNase-I hypersensitivity signal near CTCF peaks in which DNase-I hypersensitivity signal was not accurately predicted by transcription.

Discussion

Currently genome annotation requires conducting assays for multiple independent histone modifications to identify functional elements. Since functional elements are known to be highly tissue specific and their activity is dependent on environmental conditions, assays must be performed in numerous tissues and conditions to exhaustively identify functional elements. As a result, genome annotation efforts still benefit greatly from the coordinated efforts of large consortia. However, consortium efforts are not tractable to apply in all species and tissues, especially as major efforts to sequence eukaryotic organisms begin to produce large numbers of high-quality reference genomes (*37*). This creates a great need for individual communities to annotate functional elements using the most efficient molecular and computational tools. We show here that nascent transcription measured using PRO-seq provides at least as much information about chromatin state as the combination of multiple ChIP-seq datasets. In addition to chromatin state, nascent transcription is also known to provide direct information about gene expression (*71*), transcription factor binding (*72*), the location of transcription start sites, and the grammar of transcription initiation domains (*11*, *15*, *53*, *73*). Moreover, the introduction of new biochemical tools that allow the application of PRO-seq techniques with greater ease in solid tissue samples and other samples that have proven challenging to measure using conventional genomic techniques has the potential to further democratize these technologies (*51*, *74*). Thus, using dHIT to decompose PRO-seq data into separate information about active chromatin modifications is a supremely efficient strategy to gain information about functional elements using a single molecular assay for each tissue and condition.

The high correlation between histone modifications and transcription has important consequences for the way we understand the role of histone modifications. Despite decades of speculation about the regulatory role of many histone modifications, recent studies depleting histone modifications widely believed to be critical for transcription result in surprisingly limited effects on gene expression (*75*–*77*). Our incomplete knowledge about the role that histone modifications play in transcription results in part from a lack of information about precisely how strong the correspondence between histone modifications and transcription actually is. For instance, do histone modifications serve in part to “bookmark” critical functional elements for later activation by developmental or environmental processes? We show here that the correlation between histone modifications and transcription is nearly as strong as the correlation between biological replicates of experimental histone modification ChIP-seq data. Our findings likely underestimate the actual correlation between transcription and histone modifications, due to technical factors including imperfections in the model fit and biological differences between cells cultured in different labs.

The strong correspondence between histone modification and transcription is not compatible with models where histone modifications routinely encode future transcription events that are not already reflected in the current cellular environment. Although our work does not directly rule out histone modifications encoding transcription programs which are currently in use, a regulatory role implies that histone modifications must be stable and inherited through successive cell divisions. Numerous reports indicate active histone modifications are unstable due to a continual dynamic competition between writers and erasers (*78*, *79*). Therefore our data, in conjunction with published work, is most compatible with models in which histone modifications serve as cogs in the transcription machinery.

Our results also have implications for the debate about whether transcription initiates pervasively at open chromatin regions. Classical models of gene regulation during the 1970s held that histones were general suppressor proteins that passively silenced gene expression (*68*). Although the importance of chromatin has been recognized since the discovery of histone modifications in 1964 (*1*), recent variations on this basic theme posit that all nucleosome depleted regions initiate transcription with some frequency, regardless of whether they show histone modifications or regulatory activity (*69*). We show that, unlike histone modifications, there were a substantial number of DNase-I accessible open chromatin regions that were not identified by dHIT imputation. These DNase-I accessible regions had no evidence of either transcription or histone modifications, but were enriched for other chromatin binding proteins like CTCF. Our findings indicate that transcription requires not just chromatin accessibility, but it also requires the correct chromatin environment for transcription initiation.

Methods

Experimental methods

Cell culture: K562 cells (ATCC, CCL-243) were cultured at 37ºC, 5% CO_2 at a density between 0.3-1 x 10^6cells/mL in RPMI medium (VWR 45000-396) topped up with 10% Fetal Bovine Serum (Genesee Scientific, cat: #25-514). Cells were split at a consistent interval of 3 days, when the cells reach 10^6 cells/mL.

Cells culture for Triptolide time course: 24h prior to Triptolide treatment, K562 cells were resuspended in fresh (RPMI) medium at a density of 0.6 x 10^6 cells/mL. On the day of the experiment, cells were recounted, aliquoted in equal cell numbers to 6 T-100 ThermoFisher Tissue Culture Flasks (each flask corresponding to one time point) and treated with Triptolide (Sigma-Aldrich, T3652-1MG) to a final concentration of 500 nM Triptolide. The Triptolide treatment was performed for 0 min, 15 min, 30 min, 1h, 2h, and respectively 4h.

Cells cross-linking for ChIP: After Triptolide treatment, K562 cells were cross-linked in 1% CH_2O freshly prepared in 1x PBS on the day of the experiment to reach the final concentration of 0.1% CH_2O in the media. Following a 5 min incubation at room temperature on a rocking platform, the cross-linker was quenched with 1M Glycine to reach a final concentration of 0.135 M Glycine. Lastly, cells were washed twice in 1x PBS, then harvested and snap frozen on dry ice.

MNase ChIP-seq - chromatin extraction: We prepared MNase ChIP-seq data for seven histone marks in K562 cells, including H3K4me1 (ab8895, lot: GR3206285-1), H3K4me2 (ab7766, lot: GR102810-4), H3K4me3 (ab8580, lot: GR3197347-1), H3K27ac (ab4729, lot: GR3231937-1), H3K36me3 (ab9050, lot: GR3257952-2), H3K79me3 (ab2621, lot: GR3173217), and H3K27me3 (ab6002, lot: GR3228496-2). All buffers and solutions used were provided by Cell Signaling Technology (91820S Simple ChIP kit)

Cross-linked K562 cells were thawed on ice and resuspended in 1 mL cold Buffer A, mixed well, and centrifuged at 2000x g for 5 min at 4ºC. The pellet was then mixed in 0.5 mL cold Buffer B, centrifuged at 2000x g for 5 min at 4ºC andresspended again in Buffer B. While still in Buffer B, chromatin was digested with 0.5 uL MNase for 13 min at 37ºC. Tubes were inverted every 2 min during the incubation time. Finally, the reaction was stopped by the addition of 40 uL 0.5 M EDTA, and the tubes were moved to 4ºC. The cell suspension got topped up with 1.5 mL cold ChIP Buffer, transferred to a 7 mL glass dounce homogenizer, and dounced ~30 times with a tight pestle to release the chromatin. The chromatin was further diluted with 1 mL cold ChIP Buffer and aliquoted to 1.5 mL Eppendorf tubes to be centrifuged at 12000x g for 10 min at 4ºC. The supernatant was collected and total chromatin quantified before each immunoprecipitation.

MNase ChIP-seq - Immunoprecipitation: Total digested chromatin was diluted to a total volume of 1 mL in cold ChIP Buffer. ChIP samples were incubated with 3ug anti-histone antibody at 4ºC overnight rotating, then incubated for an extra 2h at 4ºC with 20 ug magnetic beads (50% protein A, 50% protein G). After incubation, samples were placed on a magnetic rack and washed three times with 1 mL Low Salt Wash Buffer for 5 min at 4ºC, and three times with High Salt Wash Buffer for 5 min at 4ºC. Lastly, the beads were resuspended in 150 uL Elution Buffer and incubated on a shaking Thermomixer for 1.5 h at 65ºC. The eluted fractions were saved, treated with 2 uL 5M NaCl and 10 uL Proteinase K, and incubated overnight at 65ºC to reverse the cross-linker. Samples were cleaned up, the DNA quantified with Qubit, and library prep was performed using the NEBNext Ultra II DNA Library Prep Kit for Illumina (E7645S). The barcodes used were purchased from NEB:NEBNext Multiplex Oligos for Illumina (E6440S). Before Bioanalyzer and Illumina sequencing, all libraries were size selected by being run on a 6% Native PAGE. The fragments corresponding to 200-700 bp were cut out of the gel and the DNA extracted from the polyacrylamide using 3 volumes of a DNA extraction buffer (10mM Trip pH=8, 300mM NaAc, 20mM $MgCl_2$, 1mM EDTA, 0.1% SDS) per gram gel slice. The tubes were closed, covered with parafilm, and incubated overnight at 50ºC shaking, on a Thermomixer. The following day, Spin-X columns (CLS8160, Millipore Sigma) were used to remove gel bits from the eluate which got Phenol/Chloroform precipitated. The precipitated DNA was resuspended in a 15 uL nuclease-free H_2O and the library quantified using Qubit.

PRO-seq/ ChRO-seq library prep: New PRO-seq or ChRO-seq libraries were prepared from cultured K562 cells, and from equine liver tissue samples. We prepared PRO-seq libraries in K562 cells, matched to the MNase ChIP-seq.

Training dHIT SVRs to predict histone marks using PRO-seq, GRO-seq or ChRO-seq data

Overview: The primary goal of dHIT is to map the signal intensity and “shape” in a run-on and sequencing dataset (PRO-seq, GRO-seq or ChROseq; henceforth referred to simply as PRO-seq) to the specific quantity of a histone modification at each position in the reference genome. The dHIT algorithm passes standardized read count data to a support vector regression (SVR) classifier. During a training phase, the SVR model optimized an objective function which mapped PRO-seq signal to the quantity of ChIP-seq signal at each position of the genome. Once a dHIT model is trained using existing ChIP-seq data, it can impute steady state histone modifications in

any cell type, provided that the relationship between histone modification and transcription is preserved.

Training dataset: We trained each model using five different run-on and sequencing datasets that were generated by different laboratories, thereby reducing the potential for overfitting to batch-specific features of a single dataset (see **Supplementary Table 2**) (*53*). Training data was distributed between PRO-seq and GRO-seq data. Sequencing depth of the training data ranged from 18 to 374 million uniquely mapped reads, and all five training datasets were highly correlated when comparing RPKM normalized read counts in gene bodies (*53*).

We trained SVR models for ten different histone modifications in K562 cells, primarily using data from the ENCODE project (*16*). Data for H3K122ac ChIP-seq in K562 cells was obtained from a recent paper (*48*). Lastly, we trained models to recognize high-resolution ChIP-seq data using an MNase ChIP-seq protocol for H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K36me3, and H3K79me3. For validation in holdout cell types, we obtained ChIP-seq data from six additional cell types from a variety of sources. All ChIP-seq data used in training or for validation is listed in **Supplementary Tables 1 and 2**.

SVR feature vector: We passed dHIT PRO-seq data from non-overlapping windows of multiple sizes that were centered on the position for which ChIP-seq signal intensity was being imputed. We passed data from windows at multiple size scales, including 10, 50, 500, and 5,000 bp windows (n = 10, 20, 20, and 20 windows, respectively), representing read data as far as 100 KB from the genomic region in question. PRO-seq data was standardized across each length scale in a similar fashion as we use for dREG (*73*), using a logistic function, $F(t)$, to transform raw read counts using two free parameters, α and β:

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

Where t denotes the read counts in each window. Tuning parameters α and β were defined in terms of two parameters, x and y. Intuitively, y gives the value of the logistic function at a read count of 0, and x represents the fraction of the maximal read count at which the logistic function approaches 1. Values of x and y are related to the parameters α and β by the following equations:

$$\beta = x \max(t)$$

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

We have previously found that x = 0.05 and y = 0.01 optimized the discovery of transcription initiation regions (TIRs) (*73*), and these values were used throughout this study.

Selecting training positions: We trained models using 3 million training examples divided evenly among five K562 training datasets (n = 600 thousand positions in each dataset). In all cases, human chromosome 22 was excluded from training to use as a holdout.

We found it convenient to use heuristics that identify regions with a high PRO-seq signal intensity when choosing training samples. We broke the genome into non-overlapping 10bp

windows. 10bp windows were defined as “informative positions” when the window had more than 3 reads within 100 bp on the single strand or at least one read within 1000 bp on both the positive and negative strands. Within the five training datasets, informative positions accounted for 27.3% (855.9M), 6.7% (209.4M), 14.7% (460.0M), 13.8% (433.9M), and 9.4% (294.0M) of 10 bp windows, respectively.

Training examples were selected at random, according to the following criteria: In order to increase the frequency of windows with a strong signal intensity in the training dataset, we selected 5% of the training data from positions in the informative positions pool (defined above) that also intersected a transcription start site (TSS), defined using GRO-cap (*11*), and a DNase-I hypersensitive site (*16*), 93% from the non-TSS informative sites, and the remaining 2% from the non-informative position pool. This was done to enrich the frequency of GRO-cap TSSs (these were 0.78% of hg19), and to increase the frequency of regions with substantial PRO-seq signal intensity, in the training dataset.

Training computations were conducted using Rgtsvm, a fast, GPU-based SVR implementation (*80*). We trained 3M samples with 360 features for each sample from 5 data sets with an average training time of 27.9 hours (18.0~37.8 hours) on an NVIDIA Tesla TITAN XP GPU. Training achieved an average Pearson correlation of 0.48 (0.109~0.725) on holdout positions that matched the training dataset at 10bp resolution.

SVR imputation: We imputed histone modifications every 10 bp using the run-on and sequencing datasets outlined in **Supplementary Table 2**. We tested the accuracy of imputation on human chr22 (which was withheld during training) in four holdout cell lines HCT116, HeLa, and CD4+ T-cells (*81–83*). Imputation was conducted using ChRO-seq data from 20 primary glioblastoma cases (*51*). We also imputed data from two additional mammals: mouse embryonic stem cells (mESCs) (*62*) and horse liver (new data). Computing imputed values on human chr22 (5.1M loci) took 3-5 hours on a Tesla TITAN XP GPU.

Training models that impute histone marks using other histone marks

We selected 1M samples from chromosome 1 to train SVR models in which histone marks were used to predict other histone marks. In order to make a fair comparison with models trained to predict histone marks using PRO-seq data, we also trained new models from PRO-seq (using the dataset G1) using 1M samples. To select training positions when training models using histone marks, we calculated the maximum read count in every 50 bp windows on chr1 (4.99M regions), and selected 1/3 of the samples from regions that contain more read counts than median value in either the training or the experimental data (for instance, if using H3K4me1 to predict H3K4me3, we selected 33% of training positions that had higher read counts than the median H3K4me1 or H3K4me3 signal). We selected another 1/3 from regions which contained read counts that were less than 20% of the median value in either the training or the experimental data. We selected the last 1/3 of the training regions from remaining regions at random. To obtain training datasets when multiple histone marks were used to jointly predict a histone mark, we merged multiple experimental histone mark data together and sampled windows as described above. The feature vector and standardization for histone marks were identical to those used for PRO-seq data (see above). When generating the feature vectors for multiple histone marks, we concatenated the feature vectors extracted from multiple experimental histone marks together.

We compared the difference between imputation and original experimental data using the L1 norm, by median centering and scaling each dataset, as follows:

$$L1_norm === abs(((x_i - median(x)) / sd(x)) - ((y_i - median(y)) / sd(y)))$$

Where x_i is the imputed signal, and y_i is the experimental signal for a particular comparison, and i represents the set of all genomic positions on chr22. We use sd() to denote the standard deviation of the mark.

Computing performance metrics using dHIT SVRs

Imputed profiles for 10 histone modifications in seven cell lines were compared to a variety of publicly available and newly generated ChIP-seq data available from ENCODE, Epigenome Roadmap, and a variety of other sources, as outlined in **Supplementary Table 1**. When measuring correlations, we subtracted the background (median) value from all positions, and applied a series of filters that were designed to remove artifacts of mappability or repeat content. Filters used to compute correlations include: 1) We masked all positions in which 30bp, the size of many of the older ENCODE ChIP-seq datasets, can not map uniquely to the reference genome; 2) We removed ENCODE blacklist regions in hg19 and mm9 genomes, in which they were defined (*84*); 3) We identified and masked "spikes" in the data, caused by putative experimental or mapping artifacts, that were not filtered by the above two criteria. Our filter identified blocks with a high signal intensity (top 2%) for which the sum of the absolute value of the two maximal derivatives was higher than the number of read counts in the region (i.e., $[abs(d_1) + abs(d_2)] > h$, where d_1 and d_2 are the maximal and second highest change in ChIP-seq signal intensity, and h is the total read density between the positions at which d_1 and d_2 occur). When comparing performance metrics between two experimental datasets, this filter was applied to both ChIP-seq datasets.

After masking the types of regions indicated above, we divided the whole genome or the entire chromosome into four granularities, 10 bp windows, 100 bp windows, 1,000 bp windows, and 10,000 bp windows. After collecting the sum of the read counts from experimental data and imputed data in each window, we compared the relationship between two datasets using four statistics: Pearson correlation, Spearman correlation, MAD, and JSD. Windows with 0 counts were removed from estimates of Pearson and Spearman correlation when using 10kb windows, as large regions without any ChIP-seq signal were likely driven by mappability issues.

ChromHMM analysis

Chromatin state annotations were generated using ChromHMM (*63*). We used the 18 state core model (model_18_core_K27ac) trained using ENCODE data (*52*), because we had already imputed all of the histone modifications used in this model. To convert imputed histone modifications into data that met the requirements of ChromHMM, we fit the sum of imputed signal in 200 bp windows to a Poisson distribution, and identified windows with values higher than the 0.999th quantile. Chromatin segmentation was performed using the *MakeSegmentation* command, following the instructions from the authors (*64*). We also made chromatin segmentations using an alternative source of experimental data for six histone marks, including H3K27ac, H3K27me3, H3K36me3, H3K4m1, H3K4me3, and H3K9me3 from ENCODE and other

sources, as outlined in **Supplementary Table 1**. Chromatin segmentations were compared between experimental datasets, and between imputed and experimental data, using the Jaccard distance between each pair of states. All computations were performed with bedtools (*85*). When comparing enrichments of each state to those expected at random, we randomized the position of each state using bedtools random.

Predicting bivalent TSSs

Bivalent genes in mESCs were identified using data from ref (*59*) and converted into mm9 coordinates using liftOver. Bivalent transcription start sites were predicted using a random forest. We used features representing H3K4me3 within 1,000 bp in 250 bp bins and H3K27me3 within 60,000 bp in 15,000 bp bins surrounding each promoter. All imputed histone modification data was based on models trained in K562 cells. We trained on a matched set of 100 bivalent and 100 non-bivalent promoters. The model was tested on a random set of 100 bivalent and 100 non-bivalent promoters that excluded promoters held out during training.

Acknowledgements

We thank XSEDE allocation number TG-MCB160061 as well as an NVIDIA GPU Grant for providing computational resources required in this study. We thank James Lewis, Haiyuan Yu, Anniina Vihervaara, Mike DeBerardine, John Lis and all members of the Danko lab for valuable discussions and suggestions. Work in this publication was supported by R01-HG009309 (NHGRI) to CGD and a grant from the Zweig Memorial Fund for Equine Research to DFA and CGD. DFA is an Investigator of the Dorothy Russell Havemeyer Foundations, Inc. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

References

1. V. G. Allfrey, R. Faulkner, A. E. Mirsky, ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS. *Proc. Natl. Acad. Sci. U. S. A.* **51**, 786–794 (1964).

2. J. W. K. Ho, Y. L. Jung, T. Liu, B. H. Alver, S. Lee, K. Ikegami, K.-A. Sohn, A. Minoda, M. Y. Tolstorukov, A. Appert, S. C. J. Parker, T. Gu, A. Kundaje, N. C. Riddle, E. Bishop, T. A. Egelhofer, S. S. Hu, A. A. Alekseyenko, A. Rechtsteiner, D. Asker, J. A. Belsky, S. K. Bowman, Q. B. Chen, R. A.-J. Chen, D. S. Day, Y. Dong, A. C. Dose, X. Duan, C. B. Epstein, S. Ercan, E. A. Feingold, F. Ferrari, J. M. Garrigues, N. Gehlenborg, P. J. Good, P. Haseley, D. He, M. Herrmann, M. M. Hoffman, T. E. Jeffers, P. V. Kharchenko, P. Kolasinska-Zwierz, C. V. Kotwaliwale, N. Kumar, S. A. Langley, E. N. Larschan, I. Latorre, M. W. Libbrecht, X. Lin, R. Park, M. J. Pazin, H. N. Pham, A. Plachetka, B. Qin, Y. B. Schwartz, N. Shoresh, P. Stempor, A. Vielle, C. Wang, C. M. Whittle, H. Xue, R. E. Kingston, J. H. Kim, B. E. Bernstein, A. F. Dernburg, V. Pirrotta, M. I. Kuroda, W. S. Noble, T. D. Tullius, M. Kellis, D. M. MacAlpine, S. Strome, S. C. R. Elgin, X. S. Liu, J. D. Lieb, J.

Ahringer, G. H. Karpen, P. J. Park, Comparative analysis of metazoan chromatin organization. *Nature*. **512**, 449–452 (2014).

3. A. Weiner, T.-H. S. Hsieh, A. Appleboim, H. V. Chen, A. Rahat, I. Amit, O. J. Rando, N. Friedman, High-resolution chromatin dynamics during a yeast stress response. *Mol. Cell*. **58**, 371–386 (2015).

4. A. Sebé-Pedrós, C. Ballaré, H. Parra-Acero, C. Chiva, J. J. Tena, E. Sabidó, J. L. Gómez-Skarmeta, L. Di Croce, I. Ruiz-Trillo, The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal Multicellularity. *Cell* (2016), doi:10.1016/j.cell.2016.03.034.

5. J. Schwartzentruber, A. Korshunov, X.-Y. Liu, D. T. W. Jones, E. Pfaff, K. Jacob, D. Sturm, A. M. Fontebasso, D.-A. K. Quang, M. Tönjes, V. Hovestadt, S. Albrecht, M. Kool, A. Nantel, C. Konermann, A. Lindroth, N. Jäger, T. Rausch, M. Ryzhova, J. O. Korbel, T. Hielscher, P. Hauser, M. Garami, A. Klekner, L. Bognar, M. Ebinger, M. U. Schuhmann, W. Scheurlen, A. Pekrun, M. C. Frühwald, W. Roggendorf, C. Kramm, M. Dürken, J. Atkinson, P. Lepage, A. Montpetit, M. Zakrzewska, K. Zakrzewski, P. P. Liberski, Z. Dong, P. Siegel, A. E. Kulozik, M. Zapatka, A. Guha, D. Malkin, J. Felsberg, G. Reifenberger, A. von Deimling, K. Ichimura, V. P. Collins, H. Witt, T. Milde, O. Witt, C. Zhang, P. Castelo-Branco, P. Lichter, D. Faury, U. Tabori, C. Plass, J. Majewski, S. M. Pfister, N. Jabado, Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature*. **482**, 226–231 (2012).

6. W. Béguelin, R. Popovic, M. Teater, Y. Jiang, K. L. Bunting, M. Rosen, H. Shen, S. N. Yang, L. Wang, T. Ezponda, E. Martinez-Garcia, H. Zhang, Y. Zheng, S. K. Verma, M. T. McCabe, H. M. Ott, G. S. Van Aller, R. G. Kruger, Y. Liu, C. F. McHugh, D. W. Scott, Y. R. Chung, N. Kelleher, R. Shaknovich, C. L. Creasy, R. D. Gascoyne, K.-K. Wong, L. Cerchietti, R. L. Levine, O. Abdel-Wahab, J. D. Licht, O. Elemento, A. M. Melnick, EZH2 is required for germinal center formation and somatic EZH2 mutations promote lymphoid transformation. *Cancer Cell*. **23**, 677–692 (2013).

7. Y. Gu, T. Nakamura, H. Alder, R. Prasad, O. Canaani, G. Cimino, C. M. Croce, E. Canaani, The t(4;11) chromosome translocation of human acute leukemias fuses the ALL-1 gene, related to Drosophila trithorax, to the AF-4 gene. *Cell*. **71** (1992), pp. 701–708.

8. T. A. Milne, Y. Dou, M. E. Martin, H. W. Brock, R. G. Roeder, J. L. Hess, MLL associates specifically with a subset of transcriptionally active target genes. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 14765–14770 (2005).

9. A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, K. Zhao, High-resolution profiling of histone methylations in the human genome. *Cell*. **129**, 823–837 (2007).

10. N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford, B. Ren, Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).

11. L. J. Core, A. L. Martins, C. G. Danko, C. T. Waters, A. Siepel, J. T. Lis, Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).

12. R. Andersson, A. Sandelin, C. G. Danko, A unified architecture of transcriptional regulatory elements. *Trends Genet.* **31**, 426–433 (2015).

13. B. S. Scruggs, D. A. Gilchrist, S. Nechaev, G. W. Muse, A. Burkholder, D. C. Fargo, K. Adelman, Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol. Cell*. **58**, 1101–1112 (2015).

14. Y. Chen, A. A. Pai, J. Herudek, M. Lubas, N. Meola, A. I. Järvelin, R. Andersson, V. Pelechano, L. M. Steinmetz, T. H. Jensen, A. Sandelin, Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. *Nat. Genet.* **48**, 984–994 (2016).

15. J. M. Tome, N. D. Tippens, J. T. Lis, Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nat. Genet.* (2018), doi:10.1038/s41588-018-0234-5.

16. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature*. **489**, 57–74 (2012).

17. Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, Integrative analysis of 111 reference human epigenomes. *Nature*. **518**, 317–330 (2015).

18. F. Yue, Y. Cheng, A. Breschi, J. Vierstra, W. Wu, T. Ryba, R. Sandstrom, Z. Ma, C. Davis, B. D. Pope, Y. Shen, D. D. Pervouchine, S. Djebali, R. E. Thurman, R. Kaul, E. Rynes, A. Kirilusha, G. K. Marinov, B. A. Williams, D. Trout, H. Amrhein, K. Fisher-Aylor, I. Antoshechkin, G. DeSalvo, L.-H. See, M. Fastuca, J. Drenkow, C. Zaleski, A. Dobin, P. Prieto, J. Lagarde, G. Bussotti, A. Tanzer, O. Denas, K. Li, M. A. Bender, M. Zhang, R. Byron, M. T. Groudine, D. McCleary, L. Pham, Z. Ye, S. Kuan, L. Edsall, Y.-C. Wu, M. D. Rasmussen, M. S. Bansal, M. Kellis, C. A. Keller, C. S. Morrissey, T. Mishra, D. Jain, N. Dogan, R. S. Harris, P. Cayting, T. Kawli, A. P. Boyle, G. Euskirchen, A. Kundaje, S. Lin, Y. Lin, C. Jansen, V. S. Malladi, M. S. Cline, D. T. Erickson, V. M. Kirkup, K. Learned, C. A. Sloan, K. R. Rosenbloom, B. Lacerda de Sousa, K. Beal, M. Pignatelli, P. Flicek, J. Lian, T. Kahveci, D. Lee, W. J. Kent, M. Ramalho Santos, J. Herrero, C. Notredame, A. Johnson, S. Vong, K. Lee, D. Bates, F. Neri, M. Diegel, T. Canfield, P. J. Sabo, M. S. Wilken, T. A. Reh, E. Giste, A. Shafer, T. Kutyavin, E. Haugen, D. Dunn, A. P. Reynolds, S. Neph, R. Humbert, R. S. Hansen, M. De Bruijn, L. Selleri, A. Rudensky, S. Josefowicz, R. Samstein, E. E. Eichler, S. H. Orkin, D. Levasseur, T. Papayannopoulou, K.-H. Chang, A. Skoultchi, S. Gosh, C. Disteche, P. Treuting, Y. Wang, M. J. Weiss, G. A. Blobel, X. Cao, S. Zhong, T. Wang, P. J. Good, R. F. Lowdon, L. B. Adams, X.-Q. Zhou, M. J. Pazin, E. A. Feingold, B.

Wold, J. Taylor, A. Mortazavi, S. M. Weissman, J. A. Stamatoyannopoulos, M. P. Snyder, R. Guigo, T. R. Gingeras, D. M. Gilbert, R. C. Hardison, M. A. Beer, B. Ren, Mouse ENCODE Consortium, A comparative encyclopedia of DNA elements in the mouse genome. *Nature*. **515**, 355–364 (2014).

19. A. Visel, M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, B. Ren, E. M. Rubin, L. A. Pennacchio, ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*. **457**, 854–858 (2009).

20. A. Rada-Iglesias, R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, J. Wysocka, A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*. **470**, 279–283 (2011).

21. S. Bonn, R. P. Zinzen, C. Girardot, E. H. Gustafson, A. Perez-Gonzalez, N. Delhomme, Y. Ghavi-Helm, B. Wilczyński, A. Riddell, E. E. M. Furlong, Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.* **44**, 148–156 (2012).

22. M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, E. S. Lander, Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*. **458**, 223–227 (2009).

23. M. Claussnitzer, S. N. Dankel, K.-H. Kim, G. Quon, W. Meuleman, C. Haugen, V. Glunk, I. S. Sousa, J. L. Beaudry, V. Puviindran, N. A. Abdennur, J. Liu, P.-A. Svensson, Y.-H. Hsu, D. J. Drucker, G. Mellgren, C.-C. Hui, H. Hauner, M. Kellis, FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373**, 895–907 (2015).

24. J. K. Pickrell, Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).

25. B. Akhtar-Zaidi, R. Cowper-Sal-lari, O. Corradin, A. Saiakhova, C. F. Bartels, D. Balasubramanian, L. Myeroff, J. Lutterbaugh, A. Jarrar, M. F. Kalady, J. Willis, J. H. Moore, P. J. Tesar, T. Laframboise, S. Markowitz, M. Lupien, P. C. Scacheri, Epigenomic enhancer profiling defines a signature of colon cancer. *Science*. **336**, 736–739 (2012).

26. M. F. Barber, E. Michishita-Kioi, Y. Xi, L. Tasselli, M. Kioi, Z. Moqtaderi, R. I. Tennen, S. Paredes, N. L. Young, K. Chen, K. Struhl, B. A. Garcia, O. Gozani, W. Li, K. F. Chua, SIRT7 links H3K18 deacetylation to maintenance of oncogenic transformation. *Nature*. **487**, 114–118 (2012).

27. S. K. Kurdistani, Histone modifications as markers of cancer prognosis: a cellular view. *Br. J. Cancer*. **97**, 1–5 (2007).

28. Y. Chervona, M. Costa, Histone modifications and cancer: biomarkers of prognosis? *Am. J. Cancer Res.* **2**, 589–597 (2012).

29. T. S. Furey, P. Sethupathy, S. Z. Sheikh, Redefining the IBDs using genome-scale molecular phenotyping. *Nat. Rev. Gastroenterol. Hepatol.* (2019), doi:10.1038/s41575-019-0118-x.

30. C. Abi Khalil, The emerging role of epigenetics in cardiovascular disease. *Ther. Adv. Chronic Dis.* **5**, 178–187 (2014).

31. M. M. Kronfol, M. G. Dozmorov, R. Huang, P. W. Slattum, J. L. McClay, The role of epigenomics in personalized medicine. *Expert Rev Precis Med Drug Dev*. **2**, 33–45 (2017).

32. E. Giuffra, C. K. Tuggle, FAANG Consortium, Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. *Annu Rev Anim Biosci*. **7**, 65–88 (2019).

33. C. Wang, C. Liu, D. Roqueiro, D. Grimm, R. Schwab, C. Becker, C. Lanz, D. Weigel, Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. *Genome Res.* **25**, 246–256 (2015).

34. X. Li, X. Wang, K. He, Y. Ma, N. Su, H. He, V. Stolc, W. Tongprasit, W. Jin, J. Jiang, W. Terzaghi, S. Li, X. W. Deng, High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *Plant Cell*. **20**, 259–276 (2008).

35. K. Baker, T. Dhillon, I. Colas, N. Cook, I. Milne, L. Milne, M. Bayer, A. J. Flavell, Chromatin state analysis of the barley epigenome reveals a higher-order structure defined by H3K27me1 and H3K27me3 abundance. *Plant J.* **84**, 111–124 (2015).

36. P. V. Kharchenko, A. A. Alekseyenko, Y. B. Schwartz, A. Minoda, N. C. Riddle, J. Ernst, P. J. Sabo, E. Larschan, A. A. Gorchakov, T. Gu, D. Linder-Basso, A. Plachetka, G. Shanower, M. Y. Tolstorukov, L. J. Luquette, R. Xi, Y. L. Jung, R. W. Park, E. P. Bishop, T. K. Canfield, R. Sandstrom, R. E. Thurman, D. M. MacAlpine, J. A. Stamatoyannopoulos, M. Kellis, S. C. R. Elgin, M. I. Kuroda, V. Pirrotta, G. H. Karpen, P. J. Park, Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*. **471**, 480–485 (2011).

37. H. A. Lewin, G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington, K. A. Crandall, R. Durbin, S. V. Edwards, F. Forest, M. T. P. Gilbert, M. M. Goldstein, I. V. Grigoriev, K. J. Hackett, D. Haussler, E. D. Jarvis, W. E. Johnson, A. Patrinos, S. Richards, J. C. Castilla-Rubio, M.-A. van Sluys, P. S. Soltis, X. Xu, H. Yang, G. Zhang, Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4325–4333 (2018).

38. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, 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).

39. Q. He, J. Johnston, J. Zeitlinger, ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat. Biotechnol.* **33**, 395–401 (2015).

40. P. J. Skene, S. Henikoff, An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife Sciences*. **6**, e21856 (2017).

41. M. J. Rossi, W. K. M. Lai, B. F. Pugh, Simplified ChIP-exo assays. *Nat. Commun.* **9**, 2842 (2018).

42. H. S. Kaya-Okur, S. J. Wu, C. A. Codomo, E. S. Pledger, T. D. Bryson, J. G. Henikoff, K. Ahmad, S. Henikoff, CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* (2019), p. 568915.

43. J. Ernst, M. Kellis, Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.* **33**, 364–376 (2015).

44. T. J. Durham, M. W. Libbrecht, J. J. Howbert, J. Bilmes, W. S. Noble, PREDICTD PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition. *Nat. Commun.* **9**, 1402 (2018).

45. S. Henikoff, A. Shilatifard, Histone modification: cause or cog? *Trends Genet.* **27**, 389–396 (2011).

46. B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, E. S. Lander, A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*. **125**, 315–326 (2006).

47. N. S. Outchkourov, J. M. Muiño, K. Kaufmann, W. F. J. van Ijcken, M. J. Groot Koerkamp, D. van Leenen, P. de Graaf, F. C. P. Holstege, F. G. Grosveld, H. T. M. Timmers, Balancing of histone H3K4 methylation states by the Kdm5c/SMCX histone demethylase modulates promoter and enhancer function. *Cell Rep.* **3**, 1071–1079 (2013).

48. M. M. Pradeepa, G. R. Grimes, Y. Kumar, G. Olley, G. C. A. Taylor, R. Schneider, W. A. Bickmore, Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat. Genet.* (2016), doi:10.1038/ng.3550.

49. L. J. Core, J. J. Waterfall, J. T. Lis, Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*. **322**, 1845–1848 (2008).

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

51. T. Chu, E. J. Rice, G. T. Booth, H. H. Salamanca, Z. Wang, L. J. Core, S. L. Longo, R. J. Corona, L. S. Chin, J. T. Lis, H. Kwak, C. G. Danko, Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. *Nat. Genet.* (2018), doi:10.1038/s41588-018-0244-3.

52. J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, B. E. Bernstein, Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. **473**, 43–49 (2011).

53. Z. Wang, T. Chu, L. A. Choate, C. G. Danko, Identification of regulatory elements from nascent transcription using dREG. *Genome Res.* (2018), doi:10.1101/gr.238279.118.

54. P. Tropberger, S. Pott, C. Keller, K. Kamieniarz-Gdula, M. Caron, F. Richter, G. Li, G. Mittler, E. T. Liu, M. Bühler, R. Margueron, R. Schneider, Regulation of transcription through acetylation of H3K122 on the lateral surface of the histone octamer. *Cell*. **152**, 859–872 (2013).

55. J. S. Becker, R. L. McCarthy, S. Sidoli, G. Donahue, K. E. Kaeding, Z. He, S. Lin, B. A. Garcia, K. S. Zaret, Genomic and Proteomic Resolution of Heterochromatin and Its Restriction of Alternate Fate Genes. *Mol. Cell*. **68**, 1023–1037.e15 (2017).

56. R. K. Auerbach, G. Euskirchen, J. Rozowsky, N. Lamarre-Vincent, Z. Moqtaderi, P.

Lefrançois, K. Struhl, M. Gerstein, M. Snyder, Mapping accessible chromatin regions using Sono-Seq. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 14926–14931 (2009).

57. R. N. Shah, A. T. Grzybowski, E. M. Cornett, A. L. Johnstone, B. M. Dickson, B. A. Boone, M. A. Cheek, M. W. Cowles, D. Maryanski, M. J. Meiners, R. L. Tiedemann, R. M. Vaughan, N. Arora, Z.-W. Sun, S. B. Rothbart, M.-C. Keogh, A. J. Ruthenburg, Examining the Roles of H3K4 Methylation States with Systematically Characterized Antibodies. *Mol. Cell*. **0** (2018), doi:10.1016/j.molcel.2018.08.015.

58. R. D. Hawkins, G. C. Hon, L. K. Lee, Q. Ngo, R. Lister, M. Pelizzola, L. E. Edsall, S. Kuan, Y. Luu, S. Klugman, J. Antosiewicz-Bourget, Z. Ye, C. Espinoza, S. Agarwahl, L. Shen, V. Ruotti, W. Wang, R. Stewart, J. A. Thomson, J. R. Ecker, B. Ren, Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell*. **6**, 479–491 (2010).

59. M. Ku, R. P. Koche, E. Rheinbay, E. M. Mendenhall, M. Endoh, T. S. Mikkelsen, A. Presser, C. Nusbaum, X. Xie, A. S. Chi, M. Adli, S. Kasif, L. M. Ptaszek, C. A. Cowan, E. S. Lander, H. Koseki, B. E. Bernstein, Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.* **4**, e1000242 (2008).

60. B. J. Lesch, G. A. Dokshin, R. A. Young, J. R. McCarrey, D. C. Page, A set of genes critical to development is epigenetically poised in mouse germ cells from fetal stages through completion of meiosis. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 16061–16066 (2013).

61. B. J. Lesch, S. J. Silber, J. R. McCarrey, D. C. Page, Parallel evolution of male germline epigenetic poising and somatic development in animals. *Nat. Genet.* **48**, 888–894 (2016).

62. I. Jonkers, H. Kwak, J. T. Lis, Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife*. **3**, e02407 (2014).

63. J. Ernst, M. Kellis, Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).

64. J. Ernst, M. Kellis, Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).

65. E. N. Burns, M. H. Bordbari, M. J. Mienaltowski, V. K. Affolter, M. V. Barro, F. Gianino, G. Gianino, E. Giulotto, T. S. Kalbfleisch, S. A. Katzman, M. Lassaline, T. Leeb, M. Mack, E. J. Müller, J. N. MacLeod, B. Ming-Whitfield, C. R. Alanis, T. Raudsepp, E. Scott, S. Vig, H. Zhou, J. L. Petersen, R. R. Bellone, C. J. Finno, Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. *Anim. Genet.* **49**, 564–570 (2018).

66. N. B. Kingsley, C. Kern, C. Creppe, E. N. Hales, H. Zhou, T. S. Kalbfleisch, J. N. MacLeod, J. L. Petersen, C. J. Finno, R. R. Bellone, Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq. *Genes* . **11** (2019), doi:10.3390/genes11010003.

67. H. H. He, C. A. Meyer, H. Shin, S. T. Bailey, G. Wei, Q. Wang, Y. Zhang, K. Xu, M. Ni, M. Lupien, P. Mieczkowski, J. D. Lieb, K. Zhao, M. Brown, X. S. Liu, Nucleosome dynamics define transcriptional enhancers. *Nat. Genet.* **42**, 343–347 (2010).

68. G. Felsenfeld, A brief history of epigenetics. *Cold Spring Harb. Perspect. Biol.* **6** (2014), doi:10.1101/cshperspect.a018200.

69. R. S. Young, Y. Kumar, W. A. Bickmore, M. S. Taylor, Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers. *Genome Biol.* **18**, 242 (2017).

70. K. Struhl, Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* **14**, 103–105 (2007).

71. A. Blumberg, Y. Zhao, Y.-F. Huang, N. Dukler, E. J. Rice, K. Krumholz, C. G. Danko, A. Siepel, Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data. *bioRxiv* (2019), p. 690644.

72. J. G. Azofeifa, M. A. Allen, J. R. Hendrix, T. Read, J. D. Rubin, R. D. Dowell, Enhancer RNA profiling predicts transcription factor activity. *Genome Res.* (2018), doi:10.1101/gr.225755.117.

73. C. G. Danko, S. L. Hyland, L. J. Core, A. L. Martins, C. T. Waters, H. W. Lee, V. G. Cheung, W. L. Kraus, J. T. Lis, A. Siepel, Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods*. **12**, 433–438 (2015).

74. S. S.-Y. Kim, A. Dziubek, S. A. Lee, H. Kwak, Nascent RNA sequencing of peripheral blood leukocytes reveal gene expression diversity. *bioRxiv* (2019), p. 836841.

75. K. M. Dorighi, T. Swigut, T. Henriques, N. V. Bhanu, B. S. Scruggs, N. Nady, C. D. Still, B. A. Garcia, K. Adelman, J. Wysocka, Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol. Cell*. **0** (2017), doi:10.1016/j.molcel.2017.04.018.

76. R. Rickels, H.-M. Herz, C. C. Sze, K. Cao, M. A. Morgan, C. K. Collings, M. Gause, Y.-H. Takahashi, L. Wang, E. J. Rendleman, S. A. Marshall, A. Krueger, E. T. Bartom, A. Piunti, E. R. Smith, N. A. Abshiru, N. L. Kelleher, D. Dorsett, A. Shilatifard, Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nat. Genet.* (2017), doi:10.1038/ng.3965.

77. T. Zhang, Z. Zhang, Q. Dong, J. Xiong, B. Zhu, Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biol.* **21**, 45 (2020).

78. S. Henikoff, J. M. Greally, Epigenetics, cellular memory and gene regulation. *Curr. Biol.* **26**, R644–R648 (2016).

79. T. Lappalainen, J. M. Greally, Associating cellular epigenetic models with human phenotypes. *Nat. Rev. Genet.* (2017), doi:10.1038/nrg.2017.32.

80. Z. Wang, T. Chu, L. A. Choate, C. G. Danko, Rgtsvm: Support Vector Machines on a GPU in R. *arXiv [stat.ML]* (2017), (available at http://arxiv.org/abs/1706.05544).

81. R. Andersson, P. Refsing Andersen, E. Valen, L. J. Core, J. Bornholdt, M. Boyd, T. Heick Jensen, A. Sandelin, Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat. Commun.* **5**, 5336 (2014).

82. M. A. Allen, Z. Andrysik, V. L. Dengler, H. S. Mellert, A. Guarnieri, J. A. Freeman, K. D. Sullivan, M. D. Galbraith, X. Luo, W. Lee Kraus, R. D. Dowell, J. M. Espinosa, Global

analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. *eLife*. **3** (2014), , doi:10.7554/elife.02200.

83. C. G. Danko, L. A. Choate, B. A. Marks, E. J. Rice, Z. Wang, T. Chu, A. L. Martins, N. Dukler, S. A. Coonrod, E. D. Tait Wojno, J. T. Lis, W. L. Kraus, A. Siepel, Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. *Nature Ecology & Evolution* (2018), doi:10.1038/s41559-017-0447-5.

84. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).

85. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).