Supplementary Information for:

Prediction of histone post-translational modification patterns based on nascent transcription data

Zhong Wang[1,2,*], Alexandra G. Chivu[1,3,*], Lauren A. Choate[1], Edward J. Rice[1], Donald C. Miller[1], Tinyi Chu[1], Shao-Pei Chou[1], Nicole B. Kingsley[5], Jessica L. Petersen[6], Carrie J. Finno[7], Rebecca R. Bellone[5], Douglas F. Antczak[1], John T. Lis[3], and Charles G. Danko[1,4,*]

[*] Denotes equal contribution and interchangeable ordering.

[1] Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.
[2] School of Software Technology, Dalian University of Technology, Dalian 116023, China
[3] Department of Molecular Biology & Genetics, Cornell University, Ithaca, NY 14853.
[4] Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.
[5] Veterinary Genetics Laboratory, School of Veterinary Medicine, University of California, Davis, CA 95616.
[6] Department of Animal Science, University of Nebraska-Lincoln, NE 68583.
[7] 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

Supplementary Notes

Supplementary Note 1: Types of errors common to dHIT.

We examined systematic errors observed between experimental and dHIT imputed ChIP-seq data. Several types of recurring error we noticed are described in this Supplementary Note.

- **Poor prediction in background regions:** Several quality control metrics show evidence of a weak correlation between dHIT and ChIP-seq signal in background windows. Errors in background reflect small differences in the predicted and experimental read counts at individual genomic positions in background regions, which add up when summed over large window sizes. Errors in background signal can be observed clearly in the lower-left quadrant in scatter plots comparing experimental and imputed data (see **Supplementary Fig. 1**). Also, whereas predictions made using other ChIP-seq experimental data results in mean squared error that is significantly lower genome-wide than in peak regions (Durham et al., 2018; Schreiber et al., 2020) (due to the relative ease of predicting background signals using other ChIP-seq datasets), dHIT mean squared error is about the same genome wide or in peak regions (see **Supplementary Fig. 5**).

  We think the most likely explanation is that these errors reflect differences in the way ChIP-seq and PRO-seq assays capture background. In ChIP-seq, there is substantial background pulldown of DNA due to non-specific binding of DNA to beads, tubes, tips or other sources of contamination. This background signal varies across the genome due to a variety of technical and biological factors (e.g., mappability; copy number alterations in some cell lines; etc.). In PRO-seq, the background is generally much lower and distributed in a very different way than ChIP-seq. Differences in the background distribution reflect fundamental differences in the assays: PRO-seq signal is derived from RNA (ChIP-seq is DNA), and PRO-seq has less background signal because the assay has three affinity purification steps (ChIP-seq has one). Differences in the background distribution between assays make it more difficult for dHIT to predict the number of ChIP-seq reads in regions that do not have much signal, especially in larger window sizes.

- **Mismapping in blacklist regions:** ENCODE blacklist regions had lower quality control metrics. Previous work has shown that ChIP-seq signal is not reliable in blacklist regions (Amemiya et al., 2019). Blacklist regions were excluded from all analyses and therefore did not affect quality control metrics.

- **Regions of focal amplification:** dHIT frequently predicted lower signal than observed in experimental data in regions with high copy number. This error was particularly noticeable in genome browser tracks (but also affected other quality control metrics) in cell types with abnormal karyotypes (e.g., K562). Our intuition is that this error occurred because increased DNA content is more difficult to detect using PRO-seq than ChIP-seq signal, due to fundamental differences in the way both assays capture background and signal.

- **Differences in the distribution of ChIP-seq signal within peaks:** Although dHIT captured most of the variation within experimental ChIP-seq peaks, the imputed signal was often

spread over slightly larger regions than experimental signal (see especially **Fig. 1C**, **Supplementary Fig. 4, Supplementary Fig. 7**). This may indicate either systematic biological variation in the distance between Pol II and marked nucleosomes or it may reflect uncertainty in the model due to noise in either PRO-seq or ChIP-seq assays.

- **Clear disagreement between imputed and experimental data:** We identified a handful of cases where there were clearly defined peaks in the imputed histone modification data, but not the experimental data (or vice versa). Many of these examples are located in intergenic regions, and cannot be explained by signal in gene bodies or other adjacent regulatory elements. An outstanding example of this type of error at the *CERK* promoter is shown in **Supplementary Fig. 12A**. We found that these differences between dHIT imputed and ENCODE data were not reproducible in experimental data collected by our own lab in stocks of K562 cells that closely match those used for PRO-seq (**Supplementary Fig. 12B**). This implies that these systematic differences most likely reflect biological differences in cell stocks, handling, or other environmental conditions between cell lines.

Supplementary Note 2: Comparison between the information in transcription and other histone modifications.

We asked whether PRO-seq more accurately predicted unobserved histone modifications than SVR models trained using a small number of observed histone modifications. 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 (**see Methods**). PRO-seq achieved a lower median L1 norm than any other individual assay by a fairly wide margin (**Fig. 3C, black**). Examining imputation tracks led us to attribute the relative success of PRO-seq to two features. First, PRO-seq captured the boundaries and direction of gene bodies in a manner that could not be achieved by other marks (**Supplementary Fig. 19A**). Second, PRO-seq was the most accurate at recovering the relative distribution of signal intensities in focal marks near the TSS (**Supplementary Fig. 19B**). 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 (see Supplementary Methods). Pairs 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. 3**). However, the median L1 norm was still worse than PRO-seq. We then tested combinations of marks which, in most cases, made only a minor difference in performance (**Fig. 3**). Although we observed a decrease in the median accuracy using more than two marks (**Fig. 3 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.

Supplementary Note 3: Effects of Trp on transcription.

We rapidly blocked transcription initiation using the small molecule Trp and observed the immediate effects on both transcription (using PRO-seq). After spike-in normalization, PRO-seq revealed the expected pattern of changes in Pol II throughout the time-course(Jonkers et al., 2014): a large loss near active TSSs by 1h of treatment, followed by an almost complete loss in signal across the entire genome by 4h (**Fig. 4C, right; Supplementary Fig. 8**). As Trp does not affect engaged RNA polymerase, we observed a clearing wave of Pol II ~100kb from the TSSs on long genes at 1h (**Fig. 4C, left**), consistent with reported elongation rates of ~1-3kb per minute(Danko et al., 2013; Jonkers et al., 2014)

Supplementary Note 4: Transcription independent deposition of H3K4me3.

We examined a small number of sites (<5%) at which H3K4me3 or H3K27ac did not change dramatically following Trp treatment. Retention of signals on this subset of sites could be explained in part by Pol II independent histone mark deposition acting at loci with low levels of Pol II prior to Trp treatment. Indeed, all of the sites with little or no loss in signal had low levels of Pol II in untreated cells (**Fig. 4K; Supplementary Fig. 10B**). We explored one such possible mechanism: the CxxC zinc finger protein 1 (CFP1) binds unmethylated CpG dinucleotides and recruits SET1, the main H3K4me3 methyltransferase(Clouaire et al., 2012). Sites which retained H3K4me3 had significantly higher density of CpG dinucleotides (**Fig. 4L**). Moreover, this enrichment was not found for sites that retained H3K27ac (**Fig. 4L**), illustrating that retention near CpG dinucleotides was specific to H3K4me3. Thus, at least for H3K4me3, a reasonable model is that the bulk of histone modification is deposited in a manner that is dependent on Pol II. Small amounts of H3K4me3 can be deposited in a manner that depends on other factors, but Pol II is critical to achieve high levels at most loci. These findings highlight a critical role for Pol II in maintaining H3K4me3 and H3K27ac on chromatin.

Supplementary Note 5: Transcription required for H3K27me3 near PRC2 binding sites.

We also analyzed the PRC2 dependent repressive mark, H3K27me3, after Trp treatment. H3K27me3 is deposited in large domains by the Polycomb repressive complexes 1 and 2 (PRC1/2)(Kuzmichev et al., 2002; Müller et al., 2002). H3K27me3 intensity was decreased following Trp treatment near focal binding sites of Enhancer of zeste homolog 2 (EZH2), a component of the PRC2 complex (**Supplementary Fig. 9C**), consistent with a requirement for RNA in recruiting PRC2 and depositing H3K27me3(Long et al., 2020). A small number of EZH2 binding sites contained divergent transcription, detectable even at low PRO-seq sequencing depth, that was lost coincidently with H3K27me3 (**Supplementary Fig. 9D**). However, 1-4h of Trp did not cause large changes in H3K27me3 accumulation over broad regions away from PRC2 (**Supplementary Fig. 9C**), and we found no evidence for global changes in H3K27me3 by Western blotting (**Fig. 4J**). Likewise, although the deletion of the *Ephx1* TSSs led to an accumulation of H3K27me3 over gene bodies over long time-scales(Hosogane et al., 2016), acute and genome-wide inhibition of transcription by Trp did not increase H3K27me3 signal on gene bodies in general (**Supplementary Fig.9E**). Thus, while transcription may prevent

H3K27me3 spread over extended timescales, inhibiting transcription acutely (1-4h) did not change the level of H3K27me3 that is broadly distributed away from PRC2 regions. These findings suggest that H3K27me3 is consistently renewed near PRC2 binding sites in a transcription-dependent manner, but that H3K27me3 levels across heterochromatin are reasonably stable.

Supplementary Note 6: Increased residence time of the PIC explains decreased H3 following Trp.

Two mechanistic models could explain the loss of nucleosomes near transcription initiation regions upon Trp inhibition of transcription. First, Trp inhibition of TFIIH activity could increase the time that Pol II spends in the PIC or initiation mode, thereby keeping nucleosomes at bay. Second, pioneer factors, which bind and cooperate with SWI/SNF chromatin remodelers to open chromatin(Judd et al., 2020; Krietenstein et al., 2016), could be locked in a histone evicting mode by the failure of Pol II to properly pass through initiation. To distinguish between these models, we performed CUT&RUN for TATA-binding protein (TBP), which showed increased TBP occupancy following 30 min of Trp (**Supplementary Fig. 26G-H**). This result supports the idea that the residence time of the pre-initiation complex plays some role in establishing chromatin accessibility following Trp treatment. Notably, our results mirror observations of chromatin accessibility during mitosis(Teves et al., 2018), a cellular context during which Pol II is depleted from chromatin, but accessibility to Tn5 and the signal for TBP are increased. Thus, although paused Pol II may help to establish the position of +1 and -1 nucleosomes near open chromatin regions(Gilchrist et al., 2008), chromatin accessibility can be established by multiple factors and does not necessarily require Pol II initiation or pausing.

Supplementary Note 7: ChIP-seq normalization strategy for MNase-seq triptolide time course.

In our experiments, both the human and spike-in samples were mixed and treated with MNase together, before the antibody incubation. To correct IP signals for biases in MNase cutting efficiency, handling, and other errors, we used the spike adjusted procedure (SAP) method (Bonhoure et al., 2014). We assume that ChIP-seq data reflects a linear combination of three factors: signal from the mark of interest, background which may be partially correlated with the mark, and random noise. SAP assumes that the background signals should be the same in treated and untreated samples and enforces this assumption by subtracting the expected background read count observed in the input. Because the data is noisy and we cannot assume input samples are sequenced deeply enough to estimate the background directly, SAP subtracts the expected background estimated using a linear regression fit in background regions. To define background regions, we selected a set of coordinates in the human genome with the following properties:

- They are untranscribed. The number of reads aligning within these sites should not change during the Trp time course. We first masked human coordinates for all annotated gencode transcripts, then removed regions near PRO-seq reads aligned to hg19;

- They are found outside MACS2-called ChIP-seq peaks (Zhang et al., 2008). The number of reads aligning within these sites should not be affected by differences in IP efficiencies;
- They are located in accessible chromatin and have a broad range of MNase sensitivities that cover the range observed in loci of interest.

To satisfy all these requirements we found a set of ENCODE CTCF peaks in K562 that were located at least 40kb away from any annotated transcription initiation region (TIRs). All TIRs were identified using dREG (Chu et al., 2018; Danko et al., 2015). In the case of punctate histone modifications (H3K27ac, H3K4me1, H3K4me3), we counted the input and IP reads in each 500bp bin over a 5.5kb region centered on the CTCF peak. To capture the variation in MNase accessibility for marks deposited within broader domains (H3K36me3 and H3K27me3) we enlarged the bin size to 1kb spanning 30kb adjacent to the CTCF peaks. To account for differences in IP efficiency and MNase accessibility due to biases in chromatin accessibility, we then ranked CTCF peaks by their DNase-I hypersensitivity (DHS) and summed up the counts in each bin into 10 deciles.

Box 1: Notations:

I = window ID (where I ∈ [1,11] for H3K27ac, H3K4me1, and H3K4me3
OR
I ∈ [1,30] for H3K36me3, H3K27me3)

i = sample name (where *i* ∈{ H3K27ac, H3K4me1, H3K4me3, H3K36me3, H3K27me3})

j = time point (where *j* ∈{ 0h, 1h, 4h })

(1) $\varphi(reads)I,i,j$ = sum of reads in background

(2) $\delta(reads)I,i,j$ = The expected number of IP reads based on the background ($\varphi(reads)I,i,j$) using the calculated linear regression equation

(3) $\omega(reads)I,i,j$ = reads counted from the true IP samples

(4) $ResI,i,j$ = residuals computed between (2) - (3)

(5) $Norm$ = positive residuals (from 4) divided by the total number of spike-in reads in a given (I,*i*,*j*) samples

---

We denote the raw, observed signal in each window as:

Let $\varphi(reads)I,i,j$ represent the raw signal per background window (I) in each IP sample or its corresponding input control (*i*) and time point (*j*).

SAP assumes that reads in each IP sample should be proportional to the input sample in background regions. To estimate the expected signal between input and IP, we fit a linear regression between the 10 deciles in the background regions of input and IP, resulting in

estimates of two regression coefficients, α and β. Next, we fixed α and β and used those parameters to estimate the expected IP signal within given sets of MACS2-called peaks (δ $(reads)I, i, j$) as a function of the input:

$$\delta(reads)I, i, j \ = \ \alpha \ * \ \varphi(reads)I, i = Input, j_j + \ \beta$$

To estimate the contribution of histone mark to the signal in each IP sample in each window of interest (e.g., a MACS2 peak), we then computed the residual ($ResI, i, j$) between the number of reads expected based on the linear regression equation (δ$(I)i, j$) and the number of IP reads observed experimentally within the same regions ($\omega(reads)I, i, j$).

$$ResI, i, j \ = \ \ \omega(reads)I, i, j \text{ - } \delta(reads)I, i, j$$

As residuals can be negative, and it does not make sense to have a negative signal for a mark, we set all negative residuals to zero, as in SAP. We note that the vast majority of loci we expected to have signals (e.g., because they were in ENCODE peaks) had residuals that were larger than 0.

$$Res'I, i, j \ = \ max(0, \ ResI, i, j)$$

Last, to account for global changes, we divide each normalized region by the total number of spike-in reads (Si, j) in a given sample (*i*) of time point (*j*).

$$Normalized\ ChIP\ signal \ = \ Res'\ i, t \ * \ \frac{1}{Si,j}$$

Supplementary References

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

Bonhoure, N., Bounova, G., Bernasconi, D., Praz, V., Lammers, F., Canella, D., Willis, I.M., Herr, W., Hernandez, N., Delorenzi, M., et al. (2014). Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. Genome Res. *24*, 1157–1168.

Chu, T., Wang, Z., Chou, S.-P., and Danko, C.G. (2018). Discovering Transcriptional Regulatory Elements From Run-On and Sequencing Data Using the Web-Based dREG Gateway. Curr. Protoc. Bioinformatics e70.

Clouaire, T., Webb, S., Skene, P., Illingworth, R., Kerr, A., Andrews, R., Lee, J.-H., Skalnik, D., and Bird, A. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. Genes Dev. *26*, 1714–1728.

Danko, C.G., Hah, N., Luo, X., Martins, A.L., Core, L., Lis, J.T., Siepel, A., and Kraus, W.L. (2013). Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. Mol. Cell *50*, 212–222.

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

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

Gilchrist, D.A., Nechaev, S., Lee, C., Ghosh, S.K.B., Collins, J.B., Li, L., Gilmour, D.S., and Adelman, K. (2008). NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. Genes Dev. *22*, 1921–1933.

Hosogane, M., Funayama, R., Shirota, M., and Nakayama, K. (2016). Lack of Transcription Triggers H3K27me3 Accumulation in the Gene Body. Cell Rep. *16*, 696–706.

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

Judd, J., Duarte, F.M., and Lis, J.T. (2020). Pioneer factor GAF cooperates with PBAP and NURF to regulate transcription.

Krietenstein, N., Wal, M., Watanabe, S., Park, B., Peterson, C.L., Pugh, B.F., and Korber, P. (2016). Genomic Nucleosome Organization Reconstituted with Pure Proteins. Cell *167*, 709–721.e12.

Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P., and Reinberg, D. (2002). Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. Genes Dev. *16*, 2893–2905.

Long, Y., Hwang, T., Gooding, A.R., Goodrich, K.J., Rinn, J.L., and Cech, T.R. (2020). RNA is essential for PRC2 chromatin occupancy and function in human pluripotent stem cells. Nat. Genet. 1–8.

Müller, J., Hart, C.M., Francis, N.J., Vargas, M.L., Sengupta, A., Wild, B., Miller, E.L., O'Connor, M.B., Kingston, R.E., and Simon, J.A. (2002). Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. Cell *111*, 197–208.

Schreiber, J., Durham, T., Bilmes, J., and Noble, W.S. (2020). Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. Genome Biology *21*.

Teves, S.S., An, L., Bhargava-Shah, A., Xie, L., Darzacq, X., and Tjian, R. (2018). A stable mode of bookmarking by TBP recruits RNA polymerase II to mitotic chromosomes. Elife *7*.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.