Predicting the effect of CRISPR-Cas9-based epigenome editing

Sanjit Singh Batra[1,2], **Alan Cabrera**[1,3], **Jeffrey P. Spence**[1,4], **Jacob Goell**[3], **Selvalakshmi S. Anand**[5], **Isaac B. Hilton**[3,5,*], **Yun S. Song**[2,6,*]

***For correspondence:**
isaac.hilton@rice.edu;
yss@berkeley.edu

[1]Equally contributing authors; [2]Computer Science Division, University of California, Berkeley, CA 94720; [3]Department of Bioengineering, Rice University, TX 77005; [4]Department of Genetics, Stanford University, CA 94305; [5]Systems, Synthetic, and Physical Biology Graduate Program, Rice University, TX 77005; [6]Department of Statistics, University of California, Berkeley, CA 94720

Abstract Epigenetic regulation orchestrates mammalian transcription, but functional links between them remain elusive. To tackle this problem, we use epigenomic and transcriptomic data from 13 ENCODE cell types to train machine learning models to predict gene expression from histone post-translational modifications (PTMs), achieving transcriptome-wide correlations of $\sim 0.70 - 0.79$ for most cell types. Our models recapitulate known associations between histone PTMs and expression patterns, including predicting that acetylation of histone subunit H3 lysine residue 27 (H3K27ac) near the transcription start site (TSS) significantly increases expression levels. To validate this prediction experimentally and investigate how natural vs. engineered deposition of H3K27ac might differentially affect expression, we apply the synthetic dCas9-p300 histone acetyltransferase system to 8 genes in the HEK293T cell line and to 5 genes in the K562 cell line. Further, to facilitate model building, we perform MNase-seq to map genome-wide nucleosome occupancy levels in HEK293T. We observe that our models perform well in accurately ranking relative fold-changes among genes in response to the dCas9-p300 system; however, their ability to rank fold-changes within individual genes is noticeably diminished compared to predicting expression across cell types from their native epigenetic signatures. Our findings highlight the need for more comprehensive genome-scale epigenome editing datasets, better understanding of the actual modifications made by epigenome editing tools, and improved causal models that transfer better from endogenous cellular measurements to perturbation experiments. Together these improvements would facilitate the ability to understand and predictably control the dynamic human epigenome with consequences for human health.

Introduction

All cells within a multicellular organism have the same genetic sequence up to a minuscule number of somatic mutations. Yet, many cell types exist with diverse morphological and functional traits. Epigenetics is an important regulator and driver of this diversity by allowing differences in cellular state and gene expression despite having the same genotype (***Taherian Fard and Ragan, 2019***). Indeed, cells traversing the trajectory from pluripotency through terminal differentiation have essentially the same genotype.

Epigenetic modifications such as post-translational modifications (PTMs) to histone proteins are involved in many vital regulatory processes influencing genomic accessibility, nuclear compartmentalization and transcription factor binding and recognition (***Reik et al., 2001***; ***Kouzarides, 2007***; ***Gibney and Nolan, 2010***; ***Klemm et al., 2019***; ***Hafner and Boettiger, 2022***; ***Zhang and Reinberg, 2001***). The Histone Code Hypothesis suggests that combinations of different histone PTMs specify distinct chromatin states thereby regulating gene expression (***Strahl and Allis, 2000***; ***Jenuwein and Allis, 2001***).

The field of epigenome editing has produced new tools for understanding the outcomes of epigenetic perturbations that promise to be useful for therapeutics by enabling fine-tuned control of gene expression (***Matharu and Ahituv, 2020***; ***Thakore et al., 2016***; ***Goell and Hilton, 2021***; ***Stricker et al., 2017***). Currently small molecule drugs are used to potently interfere with epigenetic regulation of gene expression. For example Vorinostat inhibits histone deacetylases thereby impacting the epigenetic landscape (***Estey, 2013***; ***Yoon and Eom, 2016***). However, small molecules globally disrupt the epigenome and transcriptome, and therefore are not suitable for targeting individual dysregulated genes nor clarifying epigenetic regulatory mechanisms (***Swaminathan et al., 2007***). Meanwhile, numerous tools have been designed to harness catalytically dead Cas9 (dCas9) to target epigenetic modifiers to DNA sequences encoded in guide RNAs (gRNAs) (***Jinek et al., 2012***; ***Mali et al., 2013***; ***Hilton et al., 2015***; ***Stepper et al., 2017***; ***Kwon et al., 2017***; ***Li et al., 2021***). CRISPR-Cas9-based epigenome editing strategies facilitate unprecedented, precise control of the epigenome and gene activation providing a path to epigenetic-based therapeutics (***Cheng et al., 2019***).

A major challenge for epigenome editing is designing gRNAs that can achieve a desired level of transcriptional or epigenetic modulation. Finding effective gRNAs currently typically requires expensive and low throughput experimental strategies (***Mohr et al., 2016***; ***Liu et al., 2020***; ***Mahata et al., 2023***). An alternative approach would be to computationally model how epigenome editing impacts histone PTMs as well as how perturbing these PTMs would consequently impact gene expression.

To understand how histone PTMs relate to gene expression, large epigenetic and transcriptomic datasets are required. Advancements in high-throughput sequencing have allowed quantification of gene expression and profiling of histone PTMs. Large consortia have performed an extensive number of assays across a wide variety of cell types (***ENCODE Project Consortium, 2012***; ***Roadmap Epigenomics Consortium et al., 2015***; ***Barrett et al., 2012***).

These include measurements of histone PTMs, transcription factor binding, gene expression, and chromatin accessibility. These data have enhanced our understanding of how histone PTMs and other chromatin dynamics impact transcriptional regulation (***Keung et al., 2015***; ***Rao et al., 2014***; ***Holoch and Moazed, 2015***).

Studying the function of these histone PTMs, however, has been largely limited to statistical associations with gene expression, which may not capture causal relationships (***Karlić et al., 2010***; ***Stillman, 2018***; ***Singh et al., 2016***). For example deep learning has been successful in predicting gene expression from epigenetic modifications, such as transcription factor binding (***Schmidt et al., 2017***), chromatin accessibility (***Schmidt et al., 2020***), histone PTMs (***Singh et al., 2016***; ***Sekhon et al., 2018***; ***Frasca et al., 2022***; ***Singh et al., 2017***; ***Hamdy et al., 2022***; ***Chen et al., 2022***), and DNA methylation (***Zhong et al., 2019***). However, these studies predict gene expression as binary levels instead of a continuous quantity. Finally, as statistical associations can be driven by non-causal mech-

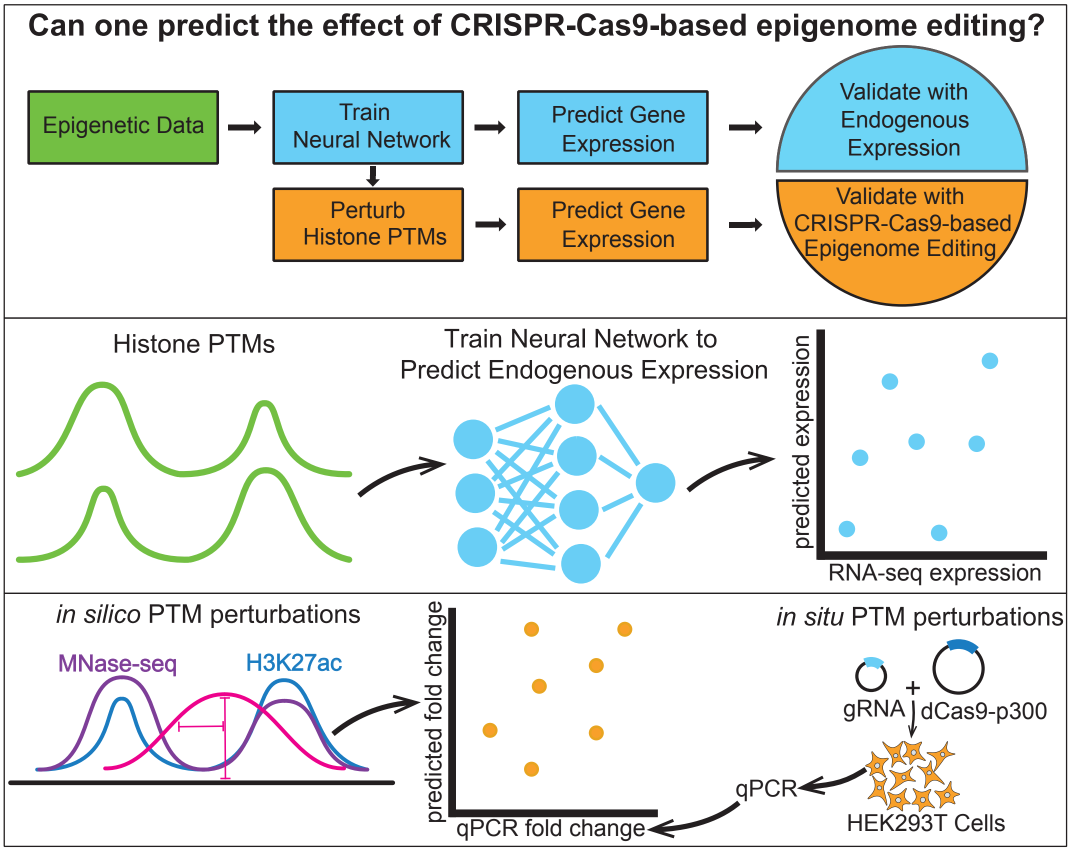
Can one predict the effect of CRISPR-Cas9-based epigenome editing?
Epigenetic Data
Train Neural Network
Predict Gene Expression
Validate with Endogenous Expression
Perturb Histone PTMs
Predict Gene Expression
Validate with CRISPR-Cas9-based Epigenome Editing
Histone PTMs
Train Neural Network to Predict Endogenous Expression
predicted expression
RNA-seq expression
in silico PTM perturbations
MNase-seq
H3K27ac
predicted fold change
qPCR fold change
in situ PTM perturbations
gRNA
+
dCas9-p300
qPCR
HEK293T Cells

Figure 1. Schematic of the epigenome editing prediction pipeline. The pipeline uses epigenetic data to train models to predict endogenous gene expression. These models were used to predict fold-change in gene expression based on perturbed histone PTM input data, and their predictions were validated using CRISPR-Cas9-based epigenome editing data.

anisms, it is unclear whether such computational models learn mechanistic, causal relationships between various epigenetic modifications and gene expression. Beyond modeling the relationship between histone PTMs and gene expression, to fully describe how a particular gRNA would affect gene expression, a model of how epigenome editing affects histone PTMs is also required. To our knowledge, there currently are no computational models that can accurately model, *in silico*, the impact of epigenome editing on histone PTMs.

Motivated by these observations, we explored models for how epigenome editing impacts histone PTMs as well as how histone PTMs impact gene expression. We used data available through ENCODE (***Schreiber et al., 2020a***; ***ENCODE Project Consortium, 2012***) to train a model of how histone PTMs impact gene expression. Our model is highly predictive of endogenous expression and learns an understanding of chromatin biology which is consistent with known patterns of various histone PTMs (***Kimura, 2013***). To test this model in the context of epigenome editing, we generated perturbation data using the dCas9-p300 histone acetyltransferase system (***Hilton et al., 2015***). The dCas9-p300 system is thought to act primarily through local acetylation of histone lysine residues, particularly histone subunit H3 lysine residue 27 (H3K27ac). Therefore, we modeled the impact of dCas9-p300 on the epigenome as a local increase in the H3K27ac profile near the target site; since the precise effect of these perturbations is unknown, we tried a variety of potential modification patterns. We then applied our trained model to predict the impact of these putative H3K27ac modifications on gene expression (Figure 1). We found that our models, which are designed to predict gene expression values, were effective in ranking relative fold-changes among genes in response to the dCas9-p300 system, achieving a Spearman's rank correlation of ~0.8. However, their performance in ranking fold-changes within individual genes was less successful when compared to the prediction of gene expression across cell types from their native epigenetic signatures. We offer possible explanations in the discussion section.

Results

Histone PTM data are highly predictive of gene expression

Genome-scale datasets are required to train models to predict gene expression using histone PTMs. Therefore, we obtained histone PTM ChIP-seq and RNA-seq data for 13 different human cell types from ENCODE (***Schreiber et al., 2020a***; ***ENCODE Project Consortium, 2012***) (Appendix Table 1). We inspected metagene plots (histone PTMs averaged across genes within gene expression quantiles) describing 6 histone PTMs in each of these 13 different cell types. Based on different overall signal levels across cell types, we concluded that batch effects, likely due to inconsistent sequencing depths, would need to be corrected prior to training models (Figure 2–figure supplement 1).

We corrected these batch effects by adapting S3norm (***Xiang et al., 2020***) (Materials and Methods, Figure 2–figure supplement 2). These corrected histone PTM tracks were then used for the remainder of our analyses along with RNA-seq data for each of the 13 cell types (Figure 2–figure supplement 3).

Importantly, we observed that H3K27ac and H3K4me3 histone PTM signal strengths positively covaried with gene expression quantile (representative cell types shown in Figure 2; all cell types shown in Figure 2–figure supplement 3). Conversely, repressive histone PTMs such as H3K27me3 and H3K9me3 were strongly inversely correlated with gene expression quantiles. Spatial patterns in the metagene plots for H3K36me3 suggested that this mark covaried more strongly with gene expression in the gene body than near the TSS. Taken together, these observations recapitulated the current understanding of these well-studied histone PTMs with respect to their associations to gene expression (***Kimura, 2013***; ***Millán-Zambrano et al., 2022***; ***Zhao et al., 2021***).

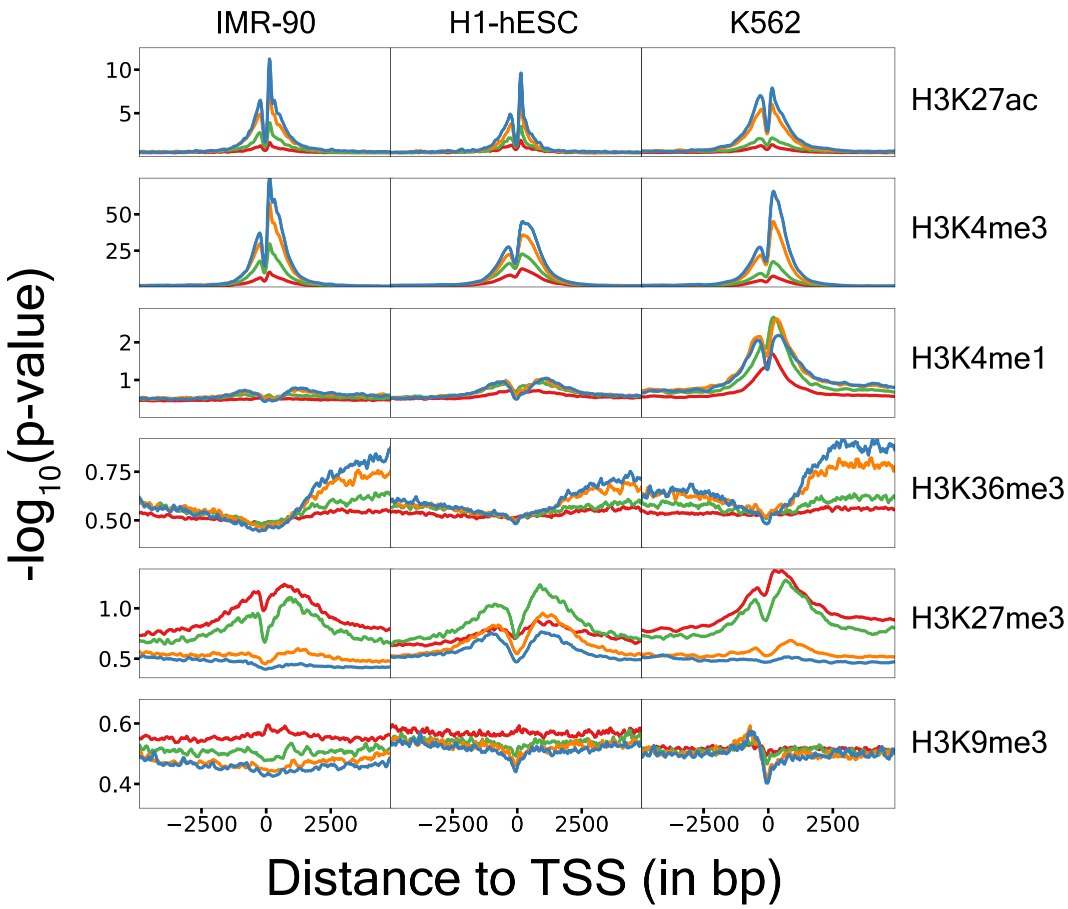

Figure 2. Metagene plots show histone PTMs are consistent across cell types and recapitulate established relationships between histone PTMs and gene expression. Colors represent genes binned into quantiles based on gene expression. Blue 75-100%, Orange 50-75%, Green 25-50%, Red 0-25% of gene expression within a cell type. The y-axis represents $-\log_{10}$(p-value) obtained from ChIP-seq data.

Figure 2–figure supplement 1. Metagene plots for different cell types for uncorrected ChIP-seq data across gene expression quantiles.

Figure 2–figure supplement 2. S3norm-based approach for correcting ChIP-seq $-\log_{10}$(p-value).

Figure 2–figure supplement 3. Metagene plots for different cell types for batch effect corrected ChIP-seq data across gene expression quantiles.

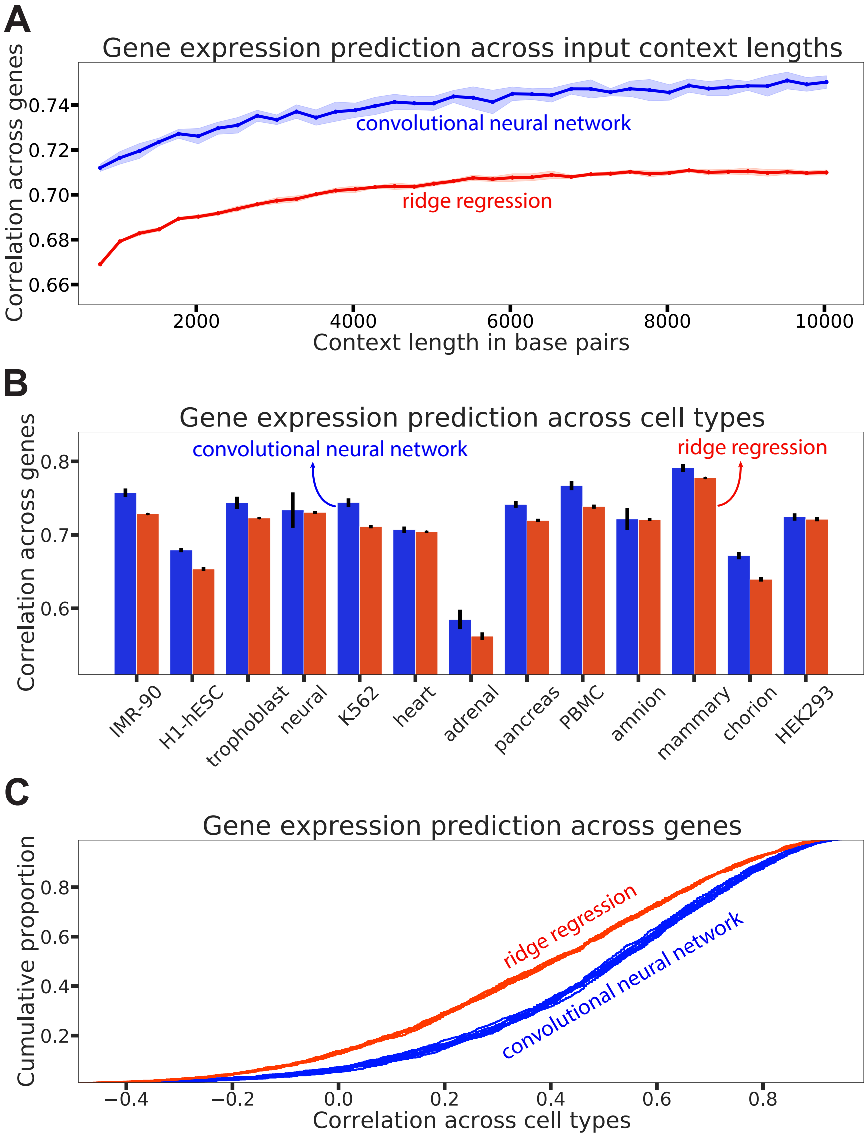

Figure 3. Histone PTMs accurately predict endogenous gene expression. (A) Spearman correlation on genes from held out chromosomes for different input context lengths, with all cell types pooled together. Blue curve is the mean across 10 computational replicates of CNNs and the red is the mean across 10 computational replicates of ridge regression. Shaded area represents standard deviation in the Spearman correlation across the 10 computational replicates. **(B)** Spearman correlation on genes of cell types held out during training. The bar plots represent the mean across 10 computational replicates and the error bars represent the corresponding standard deviations. **(C)** Distribution of Spearman correlations across genes, computed for each gene in test chromosomes by comparing predictions across the 13 cell types. The different curves represent 10 computational replicates for each model type.

Figure 3–figure supplement 1. Spearman correlation distribution across all cell types, for each cell type.

Histone PTMs accurately predict endogenous gene expression

To predict how epigenome editing affects gene expression, we first trained models to predict gene expression from endogenous histone PTMs. We trained several convolutional neural networks (CNNs) and ridge regression models to predict the gene expression of each gene in each of the 13 cell types, using only histone PTM data proximal to the TSS as features (Materials and Methods, Figure 2). We observed that Spearman's rank correlation between the true gene expression and the models' predicted gene expression on held-out chromosomes improves as the input context size increases; and for all input context sizes, the CNNs outperform ridge regression models (Figure 3**A**). Therefore, for the remainder of the analyses, we use a context size of 10,000 base pairs.

To assess the models' ability to generalize to unseen cell types, we trained a set of 10 models for each cell type. In particular, we held out the histone PTMs for a given cell type during training and then tested the models on that held-out cell type.

We observed that the CNNs outperformed ridge regression models on this cross-cell type generalization task across essentially all cell types (Figure 3**B**). The reduced performance on the adrenal cell type may be driven by a cell-type-specific biological mechanism that leads a lower correlation of its epigenetic data with other cell types, particularly for H3K36me3 (Figure 3–figure supplement 1).

Although our models accurately predicted endogenous gene expression, this does not guarantee their ability to accurately predict the relationship between local histone PTM variations and with gene expression for a particular gene across different cell types. Therefore, we determined Spearman's rank correlations between the observed expression and the predicted expression for each held-out gene across the different cell types. The distribution of these correlations suggests that overall the CNNs can better rank cell types by gene expression than ridge regression (Figure 3**C**). In particular, the median cross-cell type correlation is ∼0.53 for CNNs compared to ∼0.39 for ridge regression.

Models recover established relationships between histone PTMs and gene expression

We investigated what features of the data the models used to predict gene expression. For a given gene, we modified the input histone PTMs one-by-one at nucleosome-scale and measured the predicted fold-change in gene expression (Figure 4, Figure 4–figure supplement 1, Materials and Methods).

We observed considerable changes to the predicted fold-change upon modifying different histone PTMs. In particular, our CNN models predict that repressive marks such as H3K27me3 and H3K9me3 proximal to the TSS result in a slight decrease in expression. In contrast, activating histone PTMs such as H3K27ac and H3K4me3 result in an almost two-fold increase in predicted gene expression near the TSS. Activating both of these markers exhibits a periodic pattern, likely reflecting nucleosome occupancy. However, activation of H3K4me3 results in a sharp increase in gene expression downstream of the TSS. Additionally, we observed that H3K36me3 is predicted to increase expression, but only if it is deposited in the gene body, and the degree of activation gradually increases as it is deposited further inside of the gene body. The consistency of these observations with established mechanisms, observed previously in the literature, via which these histone PTMs modulate gene expression (***Kimura, 2013***) lend credence to our gene expression models and show that these models learn the spatial patterns of histone PTMs.

dCas9-p300 differentially activates genes depending on gRNA-targeted site

To test if our gene expression models could accurately predict the outcome of *in situ* epigenome editing experiments, we first generated dCas9-p300 data in the HEK293T cell line for 8 genes (Figure 5). We assayed at least 5 gRNAs per gene with at least 3 replicates for each gRNA. We used the HEK293T cell line because it is a widely-used testbed for epigenome editing strategies (***Hilton et al., 2015***; ***Nuñez et al., 2021***; ***O'Geen et al., 2017***; ***Mahata et al., 2023***; ***Escobar et al., 2022***; ***Wang et al.,***

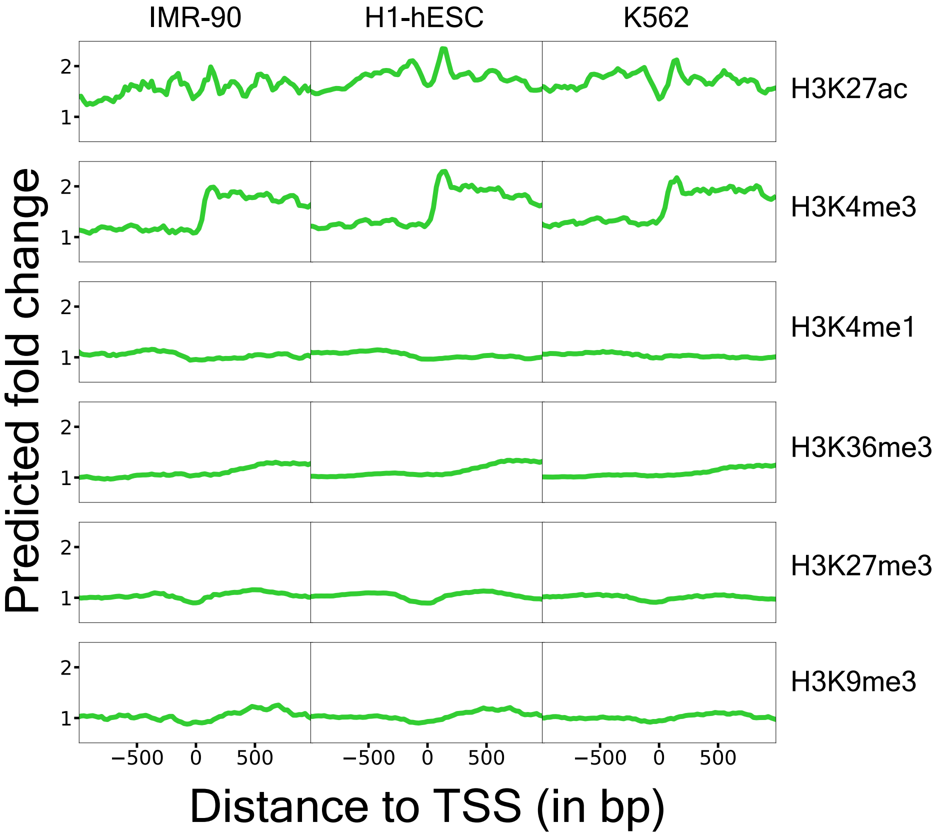

Figure 4. Features learned by gene expression models. Each point on the x-axis corresponds to *in silico* perturbation of that assay at that position and the y-axis measures the predicted fold-change in gene expression, averaged across a set of 100 trained models. The fold-changes were averaged across 500 randomly chosen genes.

Figure 4–figure supplement 1. Features learned by gene expression models for H3K9me3 in K562.

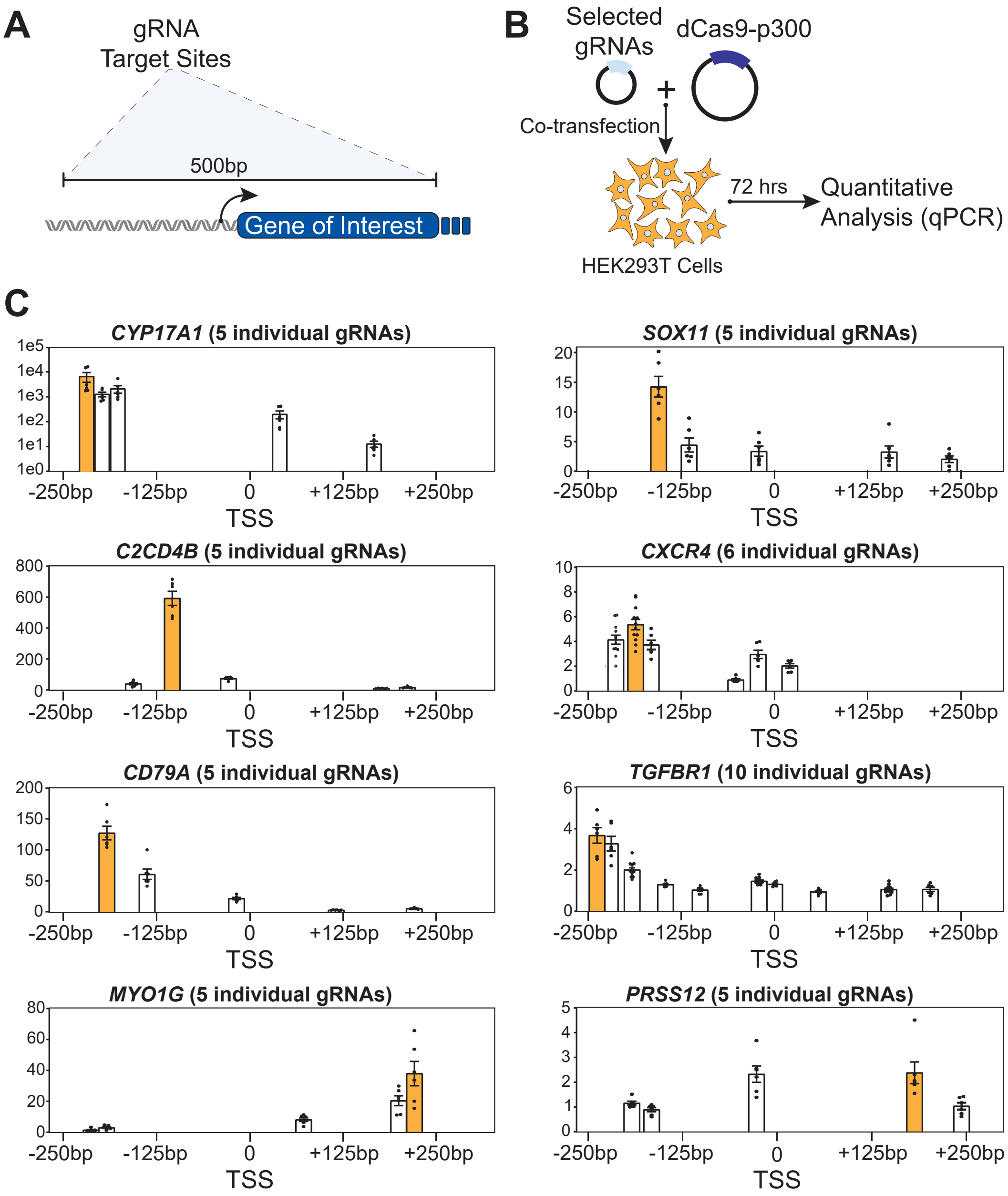

Figure 5. dCas9-p300 epigenome editing at eight endogenous genes identifies gene specific responses. The genes tested are *CYP17A1*, *SOX11*, *C2CD4B*, *CXCR4*, *CD79A*, *TGFBR1*, *MYO1G*, and *PRSS12*.
(A) gRNA targeting +/− 250 bp of each gene were selected. **(B)** These Selected gRNA were individually co-transfected with dCas9-p300 with relative mRNA determined with qPCR. **(C)** Relative mRNA associated with selected guide position are displayed with the highest activating guide position marked in orange. The Y-axis corresponds to qPCR fold-change.
Figure 5–figure supplement 1 Transfection efficiency is shared across experiments.

2022). Based on Figure 2 and Figure 4, the largest changes in H3K27ac across gene expression quantiles occur within 500 base pairs of the TSS, so we constrained gRNA targeting to this critical window. We filtered gRNAs for predicted specificity (***Concordet and Haeussler, 2018***) and on-target activity scores (***Sanson et al., 2018***). Each gRNA was tested individually, and relative mRNA abundance was measured using quantitative PCR (qPCR).

We successfully increased gene expression of all 8 genes with fold-change activation using the most effective respective gRNA for each gene ranging from 3-fold to ~6,500-fold relative to a non-targeting control gRNA (Figure 5**C**). Some of this variation may be explained by differences in endogenous gene expression levels, with the targeting of lowly expressed genes resulting in higher fold-change measurements (Appendix Table 2), as observed previously (***Wang et al., 2022***). Nevertheless, substantial variability was observed in gRNA efficacy for all targeted genes. In particular, two (*MYO1G* and *PRSS12*) out of eight genes had the most efficacious gRNA downstream of the TSS. This contrasts with other reports where targeting CRISPR/Cas based activators upstream of the TSS leads to the highest activation (***Mohr et al., 2016***; ***Gilbert et al., 2014***).

These data indicate that the rules governing the outcomes for successful dCas9-p300-based epigenome editing – and subsequent increased transcriptional activation – are complex, and highlights the fact that locus-specific nuances can be important factors in epigenome editing experiments. For example, two gRNAs targeting within ~50 base pairs of each other on *C2CD4B* have a 100-fold difference in measured mRNA (Figure 5**C**). Further, gRNAs targeting the same position in different genes can have vastly different effects. For instance, several gRNAs targeting ~250 base pairs upstream of the *CYP17A1* TSS result in a high fold-change while two gRNAs targeting roughly the same position in *MYO1G* failed to produce substantial activation (Figure 5**C**).

Computationally predicting the outcome of dCas9-p300 epigenome editing experiments

To test the hypothesis that dCas9-p300 acts through the local deposition of H3K27ac, we modeled this process *in silico* and used these perturbations as inputs to our models trained on endogenous gene expression.

We modeled the effect of dCas9-p300 on histone PTMs based on evidence from the literature as well as additional experiments we performed. The key assumptions of this model are: 1) there exists steric hindrance of dCas9 by nucleosomes (***Makasheva et al., 2021***; ***Horlbeck et al., 2016***; ***Isaac et al., 2016***; ***Radzisheuskaya et al., 2016***); 2) dCas9-p300 acts locally, altering H3K27ac levels near the gRNA target locus (***Gemberling et al., 2021***; ***Dominguez et al., 2022***) (we adopted this simplifying assumption since off-target effects are unpredictable and underexplored (***Dominguez et al., 2022***; ***Gemberling et al., 2021***; ***Weinert et al., 2018***)); 3) dCas9-p300 can deposit H3K27ac at nucleosomes, as defined by MNase activity (see Materials and Methods) (***Segelle et al., 2022***; ***Zhou et al., 2016***). Our resulting *in silico* perturbation model had a number of free parameters that we briefly describe below. Wherever possible, we used values for these parameters obtained from the literature or tested a range of plausible values. For a more complete description of the model, see Materials and Methods.

The first component of our perturbation model is steric hindrance of dCas9-p300 by nucleosomes (Figure 6**A**). Intuitively, if DNA is tightly wound around a nucleosome, the gRNA would be less likely to bind successfully. Mathematically, we modeled this as an inverse relationship between the amount of H3K27ac deposited and the MNase activity at the gRNA target locus.

It is widely assumed that dCas9-p300 activates genes through the local deposition of H3K27ac (***Klann et al., 2017***; ***Dominguez et al., 2022***). To model this, we increased local levels of H3K27ac relative to endogenous levels according to a Gaussian kernel centered at the gRNA target locus (Figure 6**B**). This adds acetylation primarily within a distance controlled by the standard deviation (σ) of the kernel. We performed CUT&RUN experiments (see Appendix) that suggest that this distance is at least 1,000 base pairs (Figure 6–figure supplement 1). Since we also do not know the degree to which dCas9-p300 alters H3K27ac levels, we modeled this as another free parameter, λ, which we varied over a range of plausible values (Materials and Methods).

Finally, we assumed that dCas9-p300 does not affect the positioning of nucleosomes and hence can only add H3K27ac at positions currently occupied by histones (***Zhou et al., 2016***). As such, we expect H3K27ac levels to only increase at loci where there is MNase activity. In particular, we modulated the Gaussian kernel described above, by performing point-wise multiplication with MNase activity (Figure 6**C**).

Since nucleosome positioning plays a crucial role in our perturbation model, we generated, to our knowledge, the first MNase-seq data for the HEK293T cell line (see Appendix).

To get a baseline of how well our perturbation model might be able to predict the effect of dCas9-p300 on gene expression, we considered the 13 distinct cell types as being analogous to natural perturbations of local histone PTMs. Across the 8 genes discussed above, which were excluded from the training set, we observed a Spearman's rank correlation of ∼0.8 between the endogenous expression and that predicted by our expression model (Figure 6**D**). This correlation was in line with the correlation observed across the endogenous transcriptome (Figure 3**A,B**). We further observed that our expression models were able to accurately rank gene expression across cell types within individual genes (Figure 6–figure supplement 2).

We then computed fold-changes between the expression predicted using endogenous histone PTMs and the expression predicted using *in silico* perturbations of these histone PTMs. We observed that our models were effective in ranking relative fold-changes across genes in response to dCas9-p300, achieving a Spearman's rank correlation of ∼0.8 between these predicted fold-changes and the experimentally determined mRNA fold-changes induced by dCas9-p300 (Figure 6**E**). However, the performance in ranking fold-changes within individual genes was less accurate (Figure 6–figure supplement 3) when compared to the prediction of cell-type-specific gene expression from native epigenetic signatures (Figure 6–figure supplement 2).

We extended this analysis to a Perturb-seq dataset, consisting of gRNAs targeting proximal to the TSS of 5 genes in the K562 cell line to further assess the model's ability to estimate gene expression changes. Consistent with the performance observed in Figure 6E, the model demonstrated robustness in predicting gene expression fold-changes across these 27 gRNAs targeting these 5 genes. Notably, these predictions achieved a Spearman's rank correlation of ∼ 0.47 with the experimentally determined mRNA fold-changes measured by Perturb-seq, as shown in Figure 6–figure supplement 4 (see Appendix). These results reinforce the model's effectiveness in capturing the nuanced effects of epigenome editing across different genes and cell types.

Discussion

Here, we sought to investigate whether we could predict how targeted epigenome editing affects endogenous gene expression. First, we collected data from ENCODE which reflects how post-translational modifications (PTMs) to histones covary with gene expression across cell types. We trained models to predict endogenous gene expression from these histone PTMs and found that these models were highly predictive (Figure 3). We further showed that such models learned known relationships between histone PTMs and gene expression (Figure 4). To test whether these expression models could predict the outcomes of epigenome editing experiments, we generated dCas9-p300 epigenome editing data in the HEK293T cell line for eight genes along with genome-wide MNase-seq data for this testbed cell line. We anticipate that the genome-wide nucleosome occupancy information for the HEK293T cell line provided by our MNase-seq experiment will be a useful resource for the genomics community. We also generated dCas9-p300 epigenome editing data via a perturb-seq experiment in K562 cells with gRNAs targeting the promoter regions of five genes. In this study, we focused on the histone changes induced by dCas9-p300 epigenome editing, but future studies may use the framework described in our manuscript and apply it to other transcriptional editors as well.

We modeled dCas9-p300's impact on local H3K27ac using a variety of parameter choices and found that these models accurately predicted fold-changes across genes. However, they were less

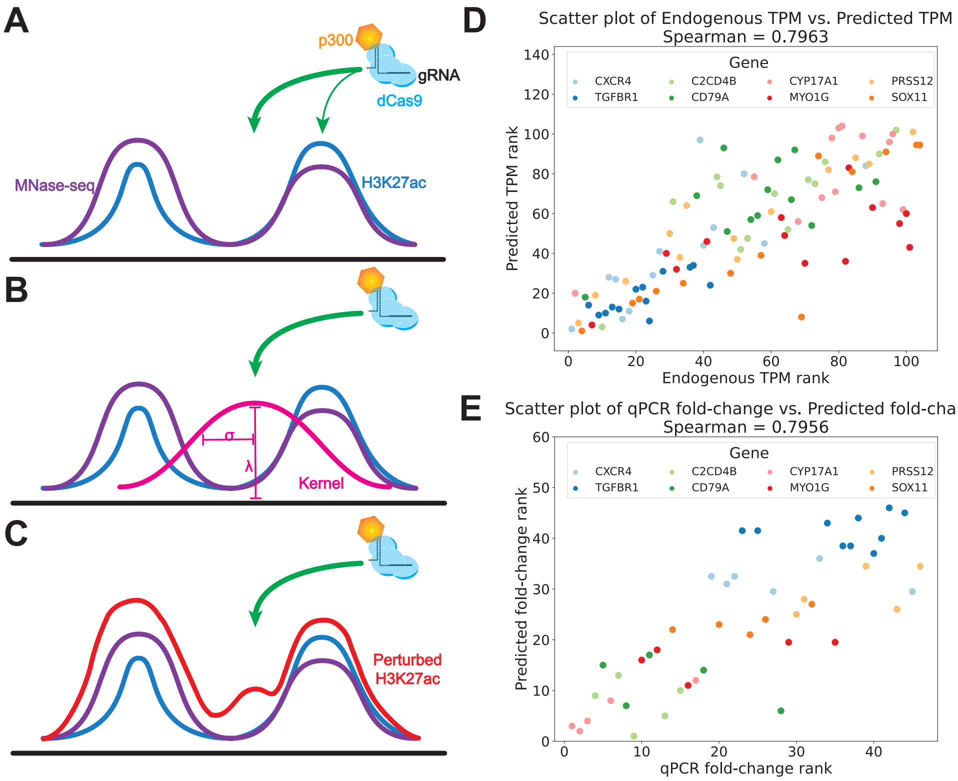

Figure 6. *In silico* model for dCas9-p300-based epigenome editing (A) dCas9-p300 is more likely to bind to a position not occupied by the nucleosome. Thicker green arrow represents higher probability of binding for a gRNA targeting that site. **(B)** The *in silico* perturbation is modeled as a Gaussian kernel parameterized by a standard deviation, σ, and the amount of H3K27ac deposited, λ. **(C)** The final perturbed H3K27ac is obtained by point-wise multiplication of the Gaussian kernel with nucleosome occupancy quantified by MNase activity since dCas9-p300 can only acetylate histones within nucleosomes. **(D)** Ranks for predicted and endogenous expression across 8 genes and 13 cell types. Rank 1 corresponds to the highest numerical value. **(E)** Ranks for predicted and empirically measured expression fold-changes following perturbation by dCas9-p300 for 8 genes in HEK293T cells. Rank 1 corresponds to the highest numerical value.

Figure 6–source data 1. Raw qPCR data.

Figure 6–source data 2. Raw CUT&RUN qPCR data.

Figure 6–source data 3. Primer sequences, sources, assay use, and corresponding direction.

Figure 6–figure supplement 1. H3K27ac levels elevation is similar across quantified regions following gRNA dCas9-p300 targeting.

Figure 6–figure supplement 2. Gene-wise predicted vs experimental gene expression TPM ranks.

Figure 6–figure supplement 3. Gene-wise predicted vs experimental fold-change ranks.

Figure 6–figure supplement 4. Predicted vs experimental fold-change ranks.

accurate at predicting the outcome of these experiments within a given gene, as compared to predicting gene expression from the endogenous epigenetic signatures (Figure 6–figure supplement 2, Figure 6–figure supplement 3). Since the endogenous epigenetic signatures could be different across genes, these *global* factors might drive the models' accurate inter-gene fold-change prediction accuracy. However, since ranking fold-changes within a gene requires a detailed understanding of the epigenetic profiles before and after dCas9-p300 epigenome editing, the reduction in performance from predicting endogenous expression to predicting the outcome of epigenome editing experiments is likely explained by one or more of the following hypotheses: 1) dCas9-p300 activates gene expression by mechanisms other than the *local* acetylation of H3K27 or dCas9-p300 functions differently from native p300; 2) differences in gRNA efficacy are not accurately explained by existing computational scores; or 3) our models, trained on endogenous gene expression across various cell types, failed to generalize even if dCas9-p300 perturbations are correctly modeled. We discuss these possible explanations more in depth below.

We considered numerous models of how dCas9-p300 affects local histone PTMs. These models span current hypotheses of how dCas9-p300 alters local histone PTMs such as H3K27ac. The poor generalization of our models in predicting intra-gene epigenome editing fold-changes could be explained by dCas9-p300 acting via mechanisms beyond local acetylation of histone proteins and H3K27 (***Zhao et al., 2021***). For example, p300 is a promiscuous lysine acetyltransferase and dCas9-p300 could be broadly acetylating across the proteome impacting *trans* factors (***Weinert et al., 2018***). Alternatively, local acetylation could be contingent on unmodeled factors such as *trans*-acting proteins or other histone PTMs present at the locus (***Zhao et al., 2021***; ***Zheng et al., 2021***). Furthermore, the genome-wide specificity of dCas9-p300-mediated histone acetylation – although likely better than small molecule-based perturbations – remains imperfect (***Gemberling et al., 2021***; ***Dominguez et al., 2022***). Our inability to accurately predict the relative fold-change of different gRNAs targeting the same gene suggests that these unmodeled factors would have to differentially affect neighboring loci within the same gene. This highlights that the current understanding of the mechanism via which dCas9-p300 drives gene expression is potentially incomplete. To better understand this mechanism, it would be immensely helpful to generate a compendium of histone PTM profiles before and after performing epigenome editing, which would enable us to train better machine learning models to predict the impact of dCas9-p300 on gene expression.

Another possible explanation for the drop in accuracy is varying gRNA efficacies. For example, gRNAs might have different levels of on-target and off-target effects. Although we ensured that all of the gRNAs used in generating the dCas9-p300 epigenome editing data were predicted to have high on-target and low off-target scores, we observed examples of gRNAs that targeted roughly the same genomic position but had vastly different impacts on gene expression. This suggests that these differences could be driven by inconsistencies in gRNA efficacy instead of local acetylation dynamics. Generating a large number of pairs of gRNAs, such as through CRISPR screens (***Schmidt et al., 2022***), targeting nearby positions could help to elucidate the factors that drive differential gRNA efficacy for epigenome editing.

The ambiguity in how to accurately model the impact of epigenome editing stands in contrast to the simpler case of DNA sequence changes, where perturbations are relatively trivial to model. Indeed, dCas9-p300 changes histone PTMs in complex ways rendering the modeling of such perturbations much more challenging. In contrast, models like Enformer (***Avsec et al., 2021***) that predict gene expression directly from DNA sequence may be able to generalize to DNA sequence perturbations better due to their relative simplicity.

Another source of generalization error could be extrapolating beyond the range of the training data. Massively increasing the amount of H3K27ac at a locus may make a gene look different than any other endogenous gene observed during training. Regression approaches including neural networks are known to have limitations in extrapolation (***Xu et al., 2020***).

Our research indicates that we can predict endogenous gene expression accurately based on histone PTMs. By creating a comprehensive dataset of epigenome editing, which assays histone

PTMs before and after *in situ* perturbations, we can enhance machine learning models. This will improve our understanding of the effects of dCas9-p300 on gene expression and assist in the design of gRNAs for achieving fine-tuned control over gene expression levels. These advancements are vital for devising experiments that deepen our mechanistic insight and offer effective strategies for human epigenome editing.

Materials and Methods

Data preparation

We obtained $-\log_{10}(p\text{-value})$ ChIP-seq tracks created by running the MACS2 peak-caller (***Feng et al., 2012***) on read count data, from the ENCODE Imputation Challenge (***Schreiber et al., 2020a***). For three tracks where data were not available, we downloaded Avocado (***Schreiber et al., 2020b***) imputations from the ENCODE data portal (***ENCODE Project Consortium, 2012***). We binned each epigenetic track at 25 base pair resolution and pre-processed them with an additional $\log$ operation before inputting them into the models for training.

We downloaded polyA-plus RNA-seq gene expression Transcripts Per Million (TPM) values for each of the 13 cell types in Appendix Table 1, from the ENCODE data portal (***ENCODE Project Consortium, 2012***) and preprocessed them with a $\log$ operation.

Normalizing *p*-values by adapting S3norm

We assigned IMR-90 to be a reference cell type, for each of the 6 histone PTMs and kept its p-values unchanged. We then performed a transformation for each of the remaining cell types adapted from the core technique developed by S3norm (***Xiang et al., 2020***), in order to normalize each histone PTM track in each of these remaining cell types, with respect to the corresponding histone PTM track in IMR-90.

First, we computed *peaks* in both, the reference as well as the target cell type. *Peaks* were defined as the 25 base pair bins corresponding to FDR-adjusted p-values less than 0.05 (***Benjamini and Hochberg, 1995***). For histone PTM tracks that were obtained from Avocado imputations (due to lack of availability of experimental data), *peaks* were defined to be the 1000 bins containing the smallest Avocado imputed p-values (based on suggestions from the authors of Avocado (***Schreiber et al., 2020b***)). All the remaining bins were defined to be *background*, for both, the reference as well as the target cell types.

We then computed the list of *peaks* that were common to both the reference and the target cell types. These were termed, *common peaks*. Similarly, we defined *common background* as the list of bins that were assigned to be *background* in both, the reference as well as the target cell types.

The S3norm method was designed to work with count data, which is always ≥ 1. However, the histone PTM tracks, which are represented as $-\log_{10}(p\text{-values})$, are not guaranteed to always be ≥ 1, hence, we transformed all the histone PTM tracks by adding 1 to the $-\log_{10}(p\text{-values})$, in both the reference as well as the target cell types.

Additionally, since the histone PTM tracks obtained from imputations performed by Avocado were not guaranteed to be distributed similar to experimental $-\log_{10}(p\text{-values})$, we scaled all the histone PTM tracks (both experimental as well as Avocado imputations) by dividing them by the minimum observed value in *common peaks* and *common background*, in order to bring experimental data and Avocado imputations onto a similar footing. In particular, before applying the S3norm normalization, we transformed $-\log_{10}(p\text{-values})$ in *common peaks* and *common background* for both the reference as well as the target cell type as following:

$$\text{TransformedCommonPeaks}_{i,\text{reference}} = \frac{1 + \text{CommonPeaks}_{i,\text{reference}}}{\min_i(\text{CommonPeaks}_{i,\text{reference}})} \tag{1}$$

$$\text{TransformedCommonPeaks}_{i,\text{target}} = \frac{1 + \text{CommonPeaks}_{i,\text{target}}}{\min_i(\text{CommonPeaks}_{i,\text{target}})} \tag{2}$$

$$\text{TransformedCommonBackground}_{i,\text{reference}} = \max\left(\frac{1 + \text{CommonBackground}_{i,\text{reference}}}{\min_i(\text{CommonBackground}_{i,\text{reference}})}, 0\right) \tag{3}$$

$$\text{TransformedCommonBackground}_{i,\text{target}} = \max\left(\frac{1 + \text{CommonBackground}_{i,\text{target}}}{\min_i(\text{CommonBackground}_{i,\text{target}})}, 0\right) \tag{4}$$

The normalization procedure of S3norm then wishes to find two positive parameters, α and β that are to be learned from the data such that both the following equations are satisfied:

$$\text{mean}(\text{TransformedCommonPeaks}_{\text{reference}}) = \text{mean}(\alpha \times \text{TransformedCommonPeaks}^{\beta}{}_{\text{target}}) \tag{5}$$

$$\text{mean}(\text{TransformedCommonBackground}_{\text{reference}}) = \text{mean}(\alpha \times \text{TransformedCommonBackground}^{\beta}_{\text{target}}) \tag{6}$$

Specifically, α is a scale factor that shifts the transformed $-\log_{10}(p\text{-values})$ of the target data set in log scale, and β is a power transformation parameter that rotates the transformed $-\log_{10}(p\text{-values})$ of the target data set in log scale (Figure 2–figure supplement 2). There is one and only one set of values for α and β that can simultaneously satisfy both the above equations for *common peaks* and the *common background* (***Xiang et al., 2020***).

The values of α and β were estimated by the Powell minimization method implemented in scipy (***Fletcher and Powell, 1963***; ***Virtanen et al., 2020***). The resulting normalized $-\log_{10}(p\text{-values})$ were used for all downstream analyses (Figure 2–figure supplement 3).

Training endogenous gene expression models

We trained convolutional neural network and ridge regression models, each, to predict gene expression using histone PTM tracks. Input features for each gene were centered at its TSS. We used an input context size of $10,000$ base pairs for all analyses subsequent to Figure 3. For all analyses we obtained predictions from our models by averaging predictions ensembled across 100 computational replicates.

To train convolutional neural network models, the normalized histone PTM tracks for each gene were processed with successive convolutional blocks. Each convolutional block consisted of a batch-normalization layer, rectified linear units (ReLU), a convolutional layer consisting of 32 convolutional kernels, each of width 5, followed by a dropout with 0.1 probability. Finally a pooling layer was applied to gradually reduce the dimension of the features. After being processed with 5 such convolutional blocks, the output was flattened and passed through a fully connected layer consisting of 16 neurons and a ReLU activation. This was ultimately processed with a fully connected layer with a single output and a linear activation (since this was a regression task). The models were trained with a mean squared error loss using the Adam optimizer with a learning rate of 0.001 for the first 50 epochs and 0.0005 for the remaining 50 epochs. Training convolutional neural network models took about 1.5 hours on 1 NVIDIA A100 Tensor Core GPU.

Interrogating the features learned by CNNs

To see how different features affected predicted levels of expression, we systematically perturbed each input feature and determined how much the perturbation affected predicted expression levels. To be concrete, we denoted the epigenetic feature at position i of gene g in cell type CT as $E_i^{CT,g}$. We then defined a perturbation function that added a scalar value of $\lambda_0 = 2500$ to the epigenetic features within 3 bins of a focal position, say, j:

$$F_j\left(E_1^{CT,g}, \dots, E_W^{CT,g}\right) := \left(E_1^{CT,g}, \dots, E_{j-3}^{CT,g} + \lambda_0, \dots, E_{j+3}^{CT,g} + \lambda_0, \dots, E_W^{CT,g}\right),$$

recalling that W is the number of bins of 25 base pairs considered by our models, which is set to 401, corresponding to a 10,000 base pair input context length, for all analyses subsequent to Figure 3. These perturbations corresponded to ~ 150 base pairs which is roughly the length of DNA wrapped around a nucleosome.

To produce Figure 4, we applied the above perturbation functions, $F_1, \dots, F_W$ to a histone PTM track of interest, and the measured the fold-change in predicted expression. To account for differences in the endogenous histone PTM tracks between genes, we averaged these fold-changes across 500 randomly chosen genes.

In silico modeling of dCas9-p300-based epigenome editing

Our model of how dCas9-p300 perturbs local histone PTMs has three separate components. We describe each of these components in turn, and then present the full model below. Throughout, we write j for the position that the gRNA targets.

First, we modeled steric hindrance of dCas9 due to nucleosomes. We used MNase-seq signal strength as a proxy for nucleosome occupancy. Letting m_j be the MNase-seq read coverage at the gRNA binding site, we modeled steric hindrance by scaling the acetylation activity of dCas9-p300 by a factor of $\exp\left(-5 \times m_j\right)$.

Second, we assumed that dCas9-p300 primarily alters the levels of H3K27ac only locally. As such, we modeled the acetylation activity of dCas9-p300 at a particular locus as a Gaussian kernel centered at the gRNA. Concretely, the acetylation activity at position i is multiplied by a factor of $\exp\left(-(i-j)^2/2\sigma^2\right)$, where σ^2 is a parameter of the model.

Finally, we assumed that dCas9-p300 can only acetylate histones where they currently are – it cannot move histones or increase H3K27ac levels outside of histones. To model this mathematically, we multiplied the acetylation activity at site i by the MNase read coverage, m_i. Therefore, if the MNase read coverage is 0 (i.e., there is no evidence of histones at that locus) then the amount of H3K27ac added to that position is also 0.

Putting this all together, for a guide targeting at position j, the effect on H3K27ac levels at position i is proportional to

$$\exp\left(-5\, m_j\right) \times \exp\left[\frac{-(i-j)^2}{2\sigma^2}\right] \times m_i$$

The constant of proportionality (i.e., how strong we expect dCas9-p300 to be overall) is treated as another free parameter, which we denote by λ.

ENCODE has epigenetic data for the HEK293 cell line, but we performed our dCas9-p300 perturbations in the HEK293T cell line. As such, we used the HEK293 histone PTM as well as RNA-seq data as a stand in for the HEK293T histone PTM and RNA-seq levels. This substitution is justified as gene expression levels for HEK293 and HEK293T are highly concordant (Figure 3–figure supplement 2). Indeed the Spearman's rank correlation between expression levels for HEK293 and two independent measurements of expression levels in HEK293T are 0.86 and 0.88, which are comparable to the correlation between the two independent experiments in HEK293T ($\rho = 0.92$). That is, the correlation across experiments within HEK293T cells is only slightly higher than the correlation between HEK293 and and HEK293T, suggesting that cross-cell type differences between HEK293 and HEK293T are on the same order as the inherent experimental and biological noise within a single cell type.

Experimental procedure

The details of dCas9-p300 epigenome editing, qPCR, CUT&RUN, and MNase-seq experiments are provided in the Supplementary Material.

Data and Code Availability

The MNase-seq data for the HEK293T cell line is available at BioProject ID PRJNA892960 on SRA. Code and data for training the gene expression models along with code for generating the figures in the manuscript are available at https://github.com/songlab-cal/epigenome_editing_2023. The data from the dCas9-p300 K562 Perturb-seq experiments is available at GSE255610 on SRA.

Declaration of interests

The authors have no competing interests.

Author Contributions

S.S.B., A.C., J.P.S, I.B.H., and Y.S.S. conceived the project and assisted with experimental design. A.C., J.G., S.S.A. and I.B.H. led laboratory experimentation and analyses, and S.S.B., J.P.S. and Y.S.S. led computational experimentation and analyses. All authors contributed to writing the manuscript.

Acknowledgments

This research is supported in part by NIH grants R35-GM134922 and R35-GM143532.

References

Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021 Oct; 18(10):1196–1203.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research. 2012; 41(D1):D991–D995.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995; 57(1):289–300.

Chen Y, Xie M, Wen J. Predicting gene expression from histone modifications with self-attention based neural networks and transfer learning. Frontiers in Genetics. 2022; 13:1081842.

Cheng Y, He C, Wang M, Ma X, Mo F, Yang S, Han J, Wei X. Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. Signal Transduction and Targeted Therapy. 2019; 4(1):1–39.

Concordet JP, Haeussler M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Research. 2018; 46(W1):W242–W245.

Cui K, Zhao K. Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq. In: *Chromatin Remodeling* Springer; 2012.p. 413–419.

Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nature Biotechnology. 2016; 34(2):184–191.

Dominguez AA, Chavez MG, Urke A, Gao Y, Wang L, Qi LS. CRISPR-mediated Synergistic Epigenetic and Transcriptional Control. The CRISPR Journal. 2022; 5(2):264–275.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep; 489(7414):57–74.

Escobar M, Li J, Patel A, Liu S, Xu Q, Hilton IB. Quantification of genome editing and transcriptional control capabilities reveals hierarchies among diverse CRISPR/Cas systems in human cells. ACS Synthetic Biology. 2022; 11(10):3239–3250.

Estey E. Epigenetics in clinical practice: the examples of azacitidine and decitabine in myelodysplasia and acute myeloid leukemia. Leukemia. 2013; 27(9):1803–1812.

Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nature Protocols. 2012; 7(9):1728–1740.

Fletcher R, Powell MJ. A rapidly convergent descent method for minimization. The Computer Journal. 1963; 6(2):163–168.

Frasca F, Matteucci M, Leone M, Morelli MJ, Masseroli M. Accurate and highly interpretable prediction of gene expression from histone modifications. BMC Bioinformatics. 2022 Apr; 23(1):151.

Gemberling MP, Siklenka K, Rodriguez E, Tonn-Eisinger KR, Barrera A, Liu F, Kantor A, Li L, Cigliola V, Hazlett MF, et al. Transgenic mice for in vivo epigenome editing with CRISPR-based systems. Nature methods. 2021; 18(8):965–974.

Gibney E, Nolan C. Epigenetics and gene expression. Heredity. 2010; 105(1):4–13.

Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, et al. Genome-scale CRISPR-mediated control of gene repression and activation. Cell. 2014; 159(3):647–661.

Goell JH, Hilton IB. CRISPR/Cas-based epigenome editing: advances, applications, and clinical utility. Trends in Biotechnology. 2021; 39(7):678–691.

Goell JH, Li J, Mahata B, Ma AJ, Kim S, Shah S, Shah S, Contreras M, Misra S, Reed D, Bedford GC, Escobar M, Hilton IB. Tailoring a CRISPR/Cas-based Epigenome Editor for Programmable Chromatin Acylation and Decreased Cytotoxicity. bioRxiv. 2024; https://www.biorxiv.org/content/early/2024/09/22/2024.09.22.611000, doi: 10.1101/2024.09.22.611000.

Hafner A, Boettiger A. The spatial organization of transcriptional control. Nature Reviews Genetics. 2022; p. 1–16.

Hamdy R, Maghraby FA, Omar YM. Convchrome: Predicting gene expression based on histone modifications using deep learning techniques. Current Bioinformatics. 2022; 17(3):273–283.

Hilton IB, D'ippolito AM, Vockley CM, Thakore PI, Crawford GE, Reddy TE, Gersbach CA. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nature Biotechnology. 2015; 33(5):510–517.

Holoch D, Moazed D. RNA-mediated epigenetic regulation of gene expression. Nat Rev Genet. 2015 Feb; 16(2):71–84.

Horlbeck MA, Witkowsky LB, Guglielmi B, Replogle JM, Gilbert LA, Villalta JE, Torigoe SE, Tjian R, Weissman JS. Nucleosomes impede Cas9 access to DNA in vivo and in vitro. eLife. 2016; 5:e12677.

Isaac RS, Jiang F, Doudna JA, Lim WA, Narlikar GJ, Almeida R. Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. eLife. 2016; 5:e13450.

Jenuwein T, Allis CD. Translating the histone code. Science. 2001 Aug; 293(5532):1074–1080.

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science. 2012; 337(6096):816–821.

Karlić R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A. 2010 Feb; 107(7):2926–2931.

Keung AJ, Joung JK, Khalil AS, Collins JJ. Chromatin regulation at the frontier of synthetic biology. Nat Rev Genet. 2015 Mar; 16(3):159–171.

Kimura H. Histone modifications for human epigenome analysis. J Hum Genet. 2013 Jul; 58(7):439–445.

Klann TS, Black JB, Chellappan M, Safi A, Song L, Hilton IB, Crawford GE, Reddy TE, Gersbach CA. CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. Nature Biotechnology. 2017; 35(6):561–568.

Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nature Reviews Genetics. 2019; 20(4):207–220.

Kouzarides T. Chromatin modifications and their function. Cell. 2007; 128(4):693–705.

Kwon DY, Zhao YT, Lamonica JM, Zhou Z. Locus-specific histone deacetylation using a synthetic CRISPR-Cas9-based HDAC. Nature Communications. 2017; 8(1):1–8.

Li J, Mahata B, Escobar M, Goell J, Wang K, Khemka P, Hilton IB. Programmable human histone phosphorylation and gene activation using a CRISPR/Cas9-based chromatin kinase. Nature Communications. 2021; 12(1):1–10.

Liu G, Zhang Y, Zhang T. Computational approaches for effective CRISPR guide RNA design and evaluation. Computational and Structural Biotechnology Journal. 2020; 18:35–44.

Mahata B, Cabrera A, Brenner DA, Guerra-Resendez RS, Li J, Goell J, Wang K, Guo Y, Escobar M, Parthasarathy AK, et al. Compact engineered human mechanosensitive transactivation modules enable potent and versatile synthetic transcriptional control. Nature Methods. 2023; 20(11):1716–1728.

Makasheva K, Bryan LC, Anders C, Panikulam S, Jinek M, Fierz B. Multiplexed single-molecule experiments reveal nucleosome invasion dynamics of the Cas9 genome Editor. Journal of the American Chemical Society. 2021; 143(40):16313–16319.

Mali P, Esvelt KM, Church GM. Cas9 as a versatile tool for engineering biology. Nature Methods. 2013; 10(10):957–963.

Matharu N, Ahituv N. Modulating gene regulation to treat genetic disorders. Nat Rev Drug Discov. 2020 Nov; 19(11):757–775.

McKnight LE, Crandall JG, Bailey TB, Banks OG, Orlandi KN, Truong VN, Donovan DA, Waddell GL, Wiles ET, Hansen SD, et al. Rapid and inexpensive preparation of genome-wide nucleosome footprints from model and non-model organisms. STAR protocols. 2021; 2(2):100486.

Millán-Zambrano G, Burton A, Bannister AJ, Schneider R. Histone post-translational modifications—cause and consequence of genome function. Nature Reviews Genetics. 2022; p. 1–18.

Mohr SE, Hu Y, Ewen-Campen B, Housden BE, Viswanatha R, Perrimon N. CRISPR guide RNA design for research applications. The FEBS Journal. 2016; 283(17):3232–3238.

Nuñez JK, Chen J, Pommier GC, Cogan JZ, Replogle JM, Adriaens C, Ramadoss GN, Shi Q, Hung KL, Samelson AJ, et al. Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. Cell. 2021; 184(9):2503–2519.

O'Geen H, Ren C, Nicolet CM, Perez AA, Halmai J, Le VM, Mackay JP, Farnham PJ, Segal DJ. dCas9-based epigenome editing suggests acquisition of histone methylation is not sufficient for target gene repression. Nucleic Acids Research. 2017; 45(17):9901–9916.

Radzisheuskaya A, Shlyueva D, Müller I, Helin K. Optimizing sgRNA position markedly improves the efficiency of CRISPR/dCas9-mediated transcriptional repression. Nucleic Acids Research. 2016; 44(18):e141–e141.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014 Dec; 159(7):1665–1680.

Reik W, Dean W, Walter J. Epigenetic reprogramming in mammalian development. Science. 2001; 293(5532):1089–1093.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb; 518(7539):317–330.

Sanson KR, Hanna RE, Hegde M, Donovan KF, Strand C, Sullender ME, Vaimberg EW, Goodale A, Root DE, Piccioni F, et al. Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nature Communications. 2018; 9(1):1–15.

Schmidt F, Gasparoni N, Gasparoni G, Gianmoena K, Cadenas C, Polansky JK, Ebert P, Nordström K, Barann M, Sinha A, Fröhler S, Xiong J, Dehghani Amirabad A, Behjati Ardakani F, Hutter B, Zipprich G, Felder B, Eils J, Brors B, Chen W, et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. Nucleic Acids Res. 2017 Jan; 45(1):54–66.

Schmidt F, Kern F, Schulz MH. Integrative prediction of gene expression with chromatin accessibility and conformation data. Epigenetics Chromatin. 2020 Feb; 13(1):4.

Schmidt R, Steinhart Z, Layeghi M, Freimer JW, Bueno R, Nguyen VQ, Blaeschke F, Ye CJ, Marson A. CRISPR activation and interference screens decode stimulation responses in primary human T cells. Science. 2022; 375(6580):eabj4008.

Schreiber J, Bilmes J, Noble WS. Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples. Genome Biol. 2020 Mar; 21(1):82.

Schreiber J, Durham T, Bilmes J, Noble WS. Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. Genome Biology. 2020; 21(1):1–18.

Segelle A, Núñez-Álvarez Y, Oldfield AJ, Webb KM, Voigt P, Luco RF. Histone marks regulate the epithelial-to-mesenchymal transition via alternative splicing. Cell Reports. 2022; 38(7):110357.

Sekhon A, Singh R, Qi Y. DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. Bioinformatics. 2018 Sep; 34(17):i891–i900.

Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, Heckl D, Ebert BL, Root DE, Doench JG, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014; 343(6166):84–87.

Singh R, Lanchantin J, Robins G, Qi Y. DeepChrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics. 2016 Sep; 32(17):i639–i648.

Singh R, Lanchantin J, Sekhon A, Qi Y. Attend and predict: Understanding gene regulation by selective attention on chromatin. Advances in neural information processing systems. 2017; 30.

Stepper P, Kungulovski G, Jurkowska RZ, Chandra T, Krueger F, Reinhardt R, Reik W, Jeltsch A, Jurkowski TP. Efficient targeted DNA methylation with chimeric dCas9–Dnmt3a–Dnmt3L methyltransferase. Nucleic Acids Research. 2017; 45(4):1703–1713.

Stillman B. Histone Modifications: Insights into Their Influence on Gene Expression. Cell. 2018 Sep; 175(1):6–9.

Strahl BD, Allis CD. The language of covalent histone modifications. Nature. 2000; 403(6765):41–45.

Stricker SH, Köferle A, Beck S. From profiles to function in epigenomics. Nature Reviews Genetics. 2017; 18(1):51–66.

Swaminathan V, Reddy BAA, Ruthrotha Selvi B, Sukanya MS, Kundu TK. Small molecule modulators in epigenetics: implications in gene expression and therapeutics. Subcell Biochem. 2007; 41:397–428.

Taherian Fard A, Ragan MA. Quantitative Modelling of the Waddington Epigenetic Landscape. Methods Mol Biol. 2019; 1975:157–171.

Thakore PI, Black JB, Hilton IB, Gersbach CA. Editing the epigenome: technologies for programmable transcription and epigenetic modulation. Nature Methods. 2016; 13(2):127–137.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020; 17(3):261–272.

Wang K, Escobar M, Li J, Mahata B, Goell J, Shah S, Cluck M, Hilton IB. Systematic comparison of CRISPR-based transcriptional activators uncovers gene-regulatory features of enhancer–promoter interactions. Nucleic Acids Research. 2022; 50(14):7842–7855.

Weinert BT, Narita T, Satpathy S, Srinivasan B, Hansen BK, Schölz C, Hamilton WB, Zucconi BE, Wang WW, Liu WR, et al. Time-resolved analysis reveals rapid dynamics and broad scope of the CBP/p300 acetylome. Cell. 2018; 174(1):231–244.

Xiang G, Keller CA, Giardine B, An L, Li Q, Zhang Y, Hardison RC. S3norm: simultaneous normalization of sequencing depth and signal-to-noise ratio in epigenomic data. Nucleic Acids Res. 2020 May; 48(8):e43.

Xu K, Zhang M, Li J, Du SS, Kawarabayashi Ki, Jegelka S. How neural networks extrapolate: From feedforward to graph neural networks. arXiv preprint arXiv:200911848. 2020; .

Yoon S, Eom GH. HDAC and HDAC inhibitor: from cancer to cardiovascular diseases. Chonnam Medical Journal. 2016; 52(1):1–11.

Zhang Y, Reinberg D. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. Genes Dev. 2001 Sep; 15(18):2343–2360.

Zhao W, Xu Y, Wang Y, Gao D, King J, Xu Y, Liang FS. Investigating crosstalk between H3K27 acetylation and H3K4 trimethylation in CRISPR/dCas-based epigenome editing and gene activation. Scientific Reports. 2021; 11(1):1–11.

Zheng X, Cui J, Wang Y, Zhang J, Wang C. CRSIPR-AI: a webtool for the efficacy prediction of CRISPR activation and interference. bioRxiv. 2021; .

Zhong H, Kim S, Zhi D, Cui X. Predicting gene expression using DNA methylation in three human populations. PeerJ. 2019 May; 7:e6757.

Zhou X, Blocker AW, Airoldi EM, O'Shea EK. A computational approach to map nucleosome positions and alternative chromatin states with base pair resolution. Elife. 2016; 5:e16970.

Appendix

Experimental Methods

Identification and Selection of gRNA

All gRNAs were designed following the same in silico identification algorithm. Genomic sequences were first identified using the UCSC Genome Browser and manipulated in Benchling. CRISPOR was used to separately identify all gRNA within a 500 bp window of each gene's TSS. TSS coordinates were determined using Phantom prediction, gene annotation, and DNase Hypersensitivity data as visualized in the Genome Browser with the hg38 genome assembly. gRNA were selected from this list to optimize for coverage, predicted specificity (CFD score >=80), and predicted on target activity (Doench 16' score) (***Doench et al., 2016***). When no gRNA were available in a region, predicted score constraints were minimally relaxed to identify a gRNA.

Plasmid & Guide Cloning, Transfection, mRNA extraction, and qPCR

dCas9-p300 was cloned into a lentiviral plasmid backbone (Addgene# 83889) was a gift from Gersbach lab. gRNA were cloned using the molecular cloning pipeline described by the Zhang group. These gRNA were cloned into an isogenic minimal guide expression backbone utilizing the Gecko guide cloning strategy (***Shalem et al., 2014***). This minimal guide cloning plasmid was a gift from Gersbach lab (Addgene#47108). Following sequence verification, gRNA tiling experiments were completed.

Cell Culture

HEK293T (ATCC, CRL-11268) were purchased from ATCC and cultured using supplemented DMEM (10%FBS (Millipore), 1% penicillin/streptomycin(Gibco)). These cells were initially expanded and cryopreserved in 0.5% (vol/vol) DMSO containing supplemented DMEM at a concentration of 2E6/ml per vial. HEK293T cells were consistently passaged at 80% confluence using Trypsin/EDTA (Gibco) dissociation and passaged at a 1:10 ratio. Cells were disregarded after their 10th passage.

gRNA Tiling qPCR Experiments

On day 0 healthy HEK293T cells (<passage 10) were lifted with Trypsin EDTA, centrifuged, resuspended, and counted with a manual hemocytometer using Trypan blue to assess health. Cells with >95% viability were seeded into 24 well plates with a consistent cell number per well (1.5×10^5). 24 hours post plating, cells with confluence of between 70% and 90% with healthy phenotype were cotransfected with individual gRNA and dCas9-p300 plasmid DNA (mass = 500 ng, 125ng, gRNA:375ng dCas9-p300, using 1.5ul Lipofectamine 3000) according to the manufacturer's protocol. All qPCR experiments were conducted using 24 samples (2 biological samples per condition) measured in the 96-well format. Additionally, these experiments individually test 11 uniquely targeting gRNA and utilized a non-targeting gRNA as a negative control for downstream analysis. Total cell mRNA was extracted using the Qiagen RNeasy kit and protocol. Reverse transcription was then carried out using iScript Advanced reverse transcriptase (Biorad) 750ng of total RNA in a 10 ul reaction. From there, cDNA was diluted to 10ng/ul based on the initial total RNA input. Then qPCR reactions were assembled in technical duplicate and consisted of the following: 45ng (original mass) of reverse transcribed and diluted cDNA, Luna qPCR Mastermix (NEB), forward primer and reverse primer. The appropriate primer set was used to target (a) the gene intended for transcriptional modulation and (b) GAPDH a ubiquitously expressed gene used to normalize input cDNA mass.

MNase-seq

MNase sample processing was completed similarly to the previously methods (***Cui and Zhao, 2012***) with modifications. HEK293T cells were grown in parallel for > 3 passages. 3 biological replicates were processed together to minimize variance. Crosslinking was carried out on 20E7 HEK293T cells with 1% formaldehyde incubated for 10 minutes at 37C prior to Glycine quenching. Next, lysis and washing occurred followed by nuclei isolation via 600g centrifugation. An initial optimization was performed on 2E6 purified nuclei using MNase amounts between 0.1-64 units of enzyme. RNase treatment as well as Proteinase K treatment and removal of crosslinks were performed as previously described (***McKnight et al., 2021***). QIAquick PCR purification, sample DNA were visualized with the use of a 2% agarose gel and Tape station (Agilent). The mono nucleosomal band 150 bp was cut from the gel and purified using the QIAquick Gel purification. Heat was not used when melting gel to preserve AT rich regions. Following mono nucleosomal band purification samples were quantified and size verified using a Tape Station (Agilent). Illumina libraries were produced using the NEBNext Ultra II DNA library preparation kit with NEBNext Dual Index Multiplex Oligos for Illumina using 1 μg of purified DNA as input and SPRI bead size selection after adapter ligation prior to index addition. Color balanced unique i5 and i7 indices were used for each biological replicate to reduce confounds associated with index hopping. Prepared library concentrations and purity were determined on a tape station. Following verification, the 3 biological replicates were admixed with the same mass and sent to Azenta for sequencing on a single lane on the HISeq 3000/4000 platform with expected yield of 350 million paired end 150 bp length reads. This sequencing scheme was expected to yield a coverage of ∼10x for each biological replicate sample.

H3K27ac CUT&RUN qPCR

H3K27ac CUT&RUN qPCR CUT&RUN was completed using the CUTANA™ ChIC/CUT&RUN Kit by Epicypher (Catalog #: 14-1048). H3K27ac was bound using the Anti-Histone H3 (acetyl K27) antibody (Catalog#: ab4729) sold by Abcam. E. Coli spike-in DNA and Rabbit IgG (components of kit# : 14-1048) were used for qPCR input normalization and negative control respectively. All experiments were performed in duplicate with 3 independent experiments. Briefly, p300 and individual gRNA were co-transfected into HEK293T cells, after 72 hours cells were detached, and CUT&RUN was completed with identical cell number were used for each sample. qPCR was completed in technical duplicate using a primer set designed for amplification near the targeted promoter (*CXCR4* or *TGFBR1*). qPCR reactions were also completed using a previously described primer set for quantification of the e.Coli gene uida. The ddT relative qPCR methods was used to analyze data, where uida Ct was used to normalized input and fold over rabbit IgG was calculated for each sample.

Computational Methods

Analysis of Perturb-seq Data

We analyzed single-cell RNA sequencing (scRNA-seq) data generated on the 10X Genomics platform as described previously (***Goell et al., 2024***), corresponding to G genes and r guide RNAs (gRNAs) across C cells using the following steps:

Quality Control

- Cells were retained if the total gene count in a cell was between 1,000 and 10,000, ensuring adequate complexity and excluding potential empty droplets or doublets.
- Cells with mitochondrial gene content exceeding 10% were excluded to avoid including dying or stressed cells.

Normalization

To normalize the expression data, the following was applied to each gene's raw count in each cell:

$$\text{Normalized Expression} = \frac{\text{Raw Count}}{\text{Total Counts per Cell} + 1}$$

Computing Perturb-seq Gene Expression

For each gene g, targeted by a set of gRNAs denoted as r_g, we determined the impact of each specific gRNA r_g^i on its gene expression by performing the following steps:

- **Expression Thresholding:** Only cells with non-zero expression of gene g were selected for further analysis.
- **gRNA-specific Selection:** From these cells, only those expressing the specific gRNA r_g^i and none other from r_g were retained.
- **Pseudobulk Quantification:** For these cells, we computed the pseudobulk mean (μ) and standard deviation (σ) of the expression levels of gene g.

Computing Perturb-seq fold-change

To establish the baseline expression of gene g, we considered cells not targeted by any gRNAs r_g^i. The pseudobulk mean ($\mu_{\text{control cells}}$) and standard deviation ($\sigma_{\text{control cells}}$) of gene g's expression in these cells were calculated. The fold-change for gene g due to gRNA r_g^i was then quantified as:

$$\text{fold-change} = \frac{\mu}{\mu_{\text{control cells}}}$$

Plotting Perturb-seq fold change against model predictions

In order to prepare Figure 6–figure supplement 4, we performed the following steps:

- genes were included if there were at least two distinct 25 bp bins within 250 base pairs of the TSS with a gRNA targeting that gene and having a distinct expression fold-change.
- gRNAs were included if the number of cells expressing r_g^i was ≥ 2.
- For each bin, the average Perturb-seq fold-change μ and the average predicted fold-change $\mu_{\text{predicted}}$ were calculated as:

$$\mu_{\text{bin}} = \frac{\sum \text{fold-change}}{\text{number of fold-change observations in the bin across gRNAs targeting the same 25 bp bin}}$$

$$\mu_{\text{predicted, bin}} = \frac{\sum \text{predicted fold-change}}{\text{number of models used to make a prediction for the fold-change within this bin}}$$

- Ranks were assigned to μ_{bin} and $\mu_{\text{predicted, bin}}$ for comparison.
- These ranks were then plotted against each other to evaluate the correlation between observed Perturb-seq fold-change and the model predicted fold-change.

Cell Type	polyA Plus RNA-seq	H3K36me3	H3K27me3	H3K27ac	H3K4me1	H3K4me3	H3K9me3
IMR-90	T	T	T	T	T	T	T
H1-hESC	T	T	T	T	T	T	T
trophoblast cell	T	T	T	T	T	T	T
neural stem progenitor cell	T	T	T	T	T	T	T
K562	T	T	T	T	T	T	T
heart left ventricle	T	T	T	T	T	T	T
adrenal gland	T	T	T	T	T	T	T
endocrine pancreas	T	T	T	T	T	T	T
peripheral blood mononuclear cell	T	T	T	T	T	T	T
amnion	T	T	T	T	T	T	T
myoepithelial cell of mammary gland	T	T	T	A	T	T	T
chorion	T	T	T	T	T	T	A
HEK293	T	T	A	T	T	T	T

Appendix Table 1. ChIP-seq $-\log_{10}$(p-values) were obtained from the ENCODE Imputation Challenge where the ground truth data were available (corresponding to entries labeled **T** in the table). Avocado imputations were downloaded from the ENCODE data portal , where ground truth data were not available (corresponding to entries labeled **A** in the table).

Gene	HEK293 (SRR3997504) TPM	HEK293T (SRR13341848) TPM	HEK293T (SRR15013784) TPM	Maximum fold-change in dCas9-p300 data	Cross-cell type Spearman
PRSS12	12.710	8.448	6.910	2.380	0.896
CXCR4	11.974	2.826	8.216	5.365	0.852
TGFBR1	0.725	3.254	8.029	3.675	0.689
C2CD4B	0.306	0.000	0.000	591.312	0.726
CD79A	0.280	0.207	0.127	127.094	0.364
SOX11	0.051	0.131	0.209	14.245	0.846
MYO1G	0.000	0.016	0.000	37.948	0.621
CYP17A1	0.000	0.000	0.000	6,549.110	0.397

Appendix Table 2. Endogenous gene expression of genes for which we generated dCas9-p300 epigenome editing data indicates that genes for which high fold-change was obtained are more likely to have low endogenous gene expression in HEK293T. Cross-cell type Spearman provides a metric to assess how accurate our CNN model predictions are, on any given gene, across the 13 cell types.

Supplementary Figure and Table Captions

Figure 2–figure supplement 1. Metagene plots for different cell types for uncorrected ChIP-seq data across gene expression quantiles. Blue is the highest and red is the lowest gene expression quantile. $*$ represents data from HEK293 and (A) represents Avocado imputed data.

Figure 2–figure supplement 2. S3norm-based approach for correcting ChIP-seq $-\log_{10}$(p-values). On the left panel, the p-values of a target cell type's ChIP-seq data, which are to be corrected are plotted on the Y-axis. While the ChIP-seq data for the reference cell type, chosen to be IMR-90, is shown on the X-axis. After correction with this devised procedure, the resulting corrected p-values are shown on the Y-axis of the right panel.

Figure 2–figure supplement 3. Metagene plots for different cell types for batch effect corrected ChIP-seq data across gene expression quantiles. Blue is the highest and red is the lowest gene expression quantile. $*$ represents data from HEK293 instead and (A) represents Avocado imputed data.

Figure 3–figure supplement 1. Spearman correlation distribution across all cell types, for each cell type. Each panel corresponds to a different assay where the epigenetic data for that assay in chromosome 17 (which is part of the test dataset) is considered.

Figure 3–figure supplement 2. Endogenous RNA-seq expression levels of HEK293 and HEK293T cell lines are highly concordant. Spearman correlations between TPM values from RNA-seq datasets of two biological replicates of the HEK293T cell line (with SRA accessions shown in parentheses) are on par with Spearman correlation with RNA-seq TPM values for the HEK293 cell line.

Figure 4–figure supplement 1. Features learned by gene expression models for H3K9me3 in K562. Each point on the X-axis corresponds to *in silico* perturbation of H3K9me3 at that position and the Y-axis measures the predicted fold-change in gene expression, averaged across a set of 100 trained models. The fold-changes were averaged across 500 randomly chosen genes. This is a zoomed-in version of the subplot in Figure Figure 4 corresponding to H3K9me3 in K562.

Figure 5–figure supplement 1. Transfection efficiency is shared across experiments. This figure shows consistent transfection efficiency across multiple gene targets. Histograms show the distribution of fluorescent signal intensity, indicating the percentage of cells (right) successfully transfected with the reporter construct containing mCherry-p300. We selected 2 gRNAs (gRNA1 and gRNA2) for 2 gene targets (*CYP17A1* and *SOX11*) and a scramble gRNA to measure the transfection efficiency. An average transfection efficiency of 17% was achieved across the different samples with no transfection in the untreated cells.

Figure 6–source data 1. Raw qPCR data. Each row has an individual measurement which includes pertinent information used to generate *in silico* and compare with model predictions. Columns include corresponding guide information regarding gRNA position and coordinates as well as gene information such as orientation and coordinates.

Figure 6–source data 2. Raw CUT&RUN qPCR data. This table includes measurements with corresponding sgRNA used and there distance with respect to the TSS. Gene information and amplicon centerpoint distance to the TSS.

Figure 6–source data 3. Primer sequences, sources, assay use, and corresponding direction. CUT&RUN primers have their corresponding genomic coordinates reported corresponding to the regions they amplify.

Figure 6–figure supplement 1. H3K27ac levels elevation is similar across quantified regions following gRNA dCas9-p300 targeting. Each colored line corresponds to a gRNA targeting proximal to CXCR4 and TGFBR1 in HEK293T cells. The X-axis represents the distance between gRNA and the CUT&RUN amplicon. The Y-axis represents H3K27ac fold enrichment estimated through CUT&RUN.

Figure 6–figure supplement 2. Gene-wise predicted vs experimental gene expression TPM ranks. Each dot corresponds to a cell type and the title of each plot shows the Spearman correlation and the corresponding p-values. Rank 1 corresponds to the highest numerical value.

Figure 6–figure supplement 3. Gene-wise predicted vs experimental fold-change ranks. Each dot corresponds to a gRNA targeting a locus near the TSS of the gene (each gRNA corresponds to atleast three replicates and hence the fold-change shown here is the experimental mean). Rank 1 corresponds to the highest numerical value.

Figure 6–figure supplement 4. Predicted vs experimental fold-change ranks. Each dot corresponds to a gRNA targeting a locus near the TSS of the gene. Rank 1 corresponds to the highest numerical value.