Predicting the effect of CRISPR-Cas9-based epigenome editing

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

[1] Computer Science Division, University of California, Berkeley
[2] Department of Bioengineering, Rice University
[3] Department of Genetics, Stanford University
[4] Department of BioSciences, Rice University
[5] Department of Statistics, University of California, Berkeley

Abstract

Epigenetic regulation orchestrates mammalian transcription, but functional links between them remain elusive. To tackle this problem, we here 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 samples. In addition to recapitulating known associations between histone PTMs and expression patterns, our models predict 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 engineered vs. natural deposition of H3K27ac might differentially affect expression, we apply the synthetic dCas9-p300 histone acetyltransferase system to 8 genes in the HEK293T 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.

[*]These authors contributed equally to this work.
[†]To whom correspondence should be addressed: `isaac.hilton@rice.edu`, `yss@berkeley.edu`

1 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 core 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). 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), and DNA methylation (Zhong et al. 2019). However, as statistical associations can be driven by non-causal mechanisms, it is unclear whether such computational models learn

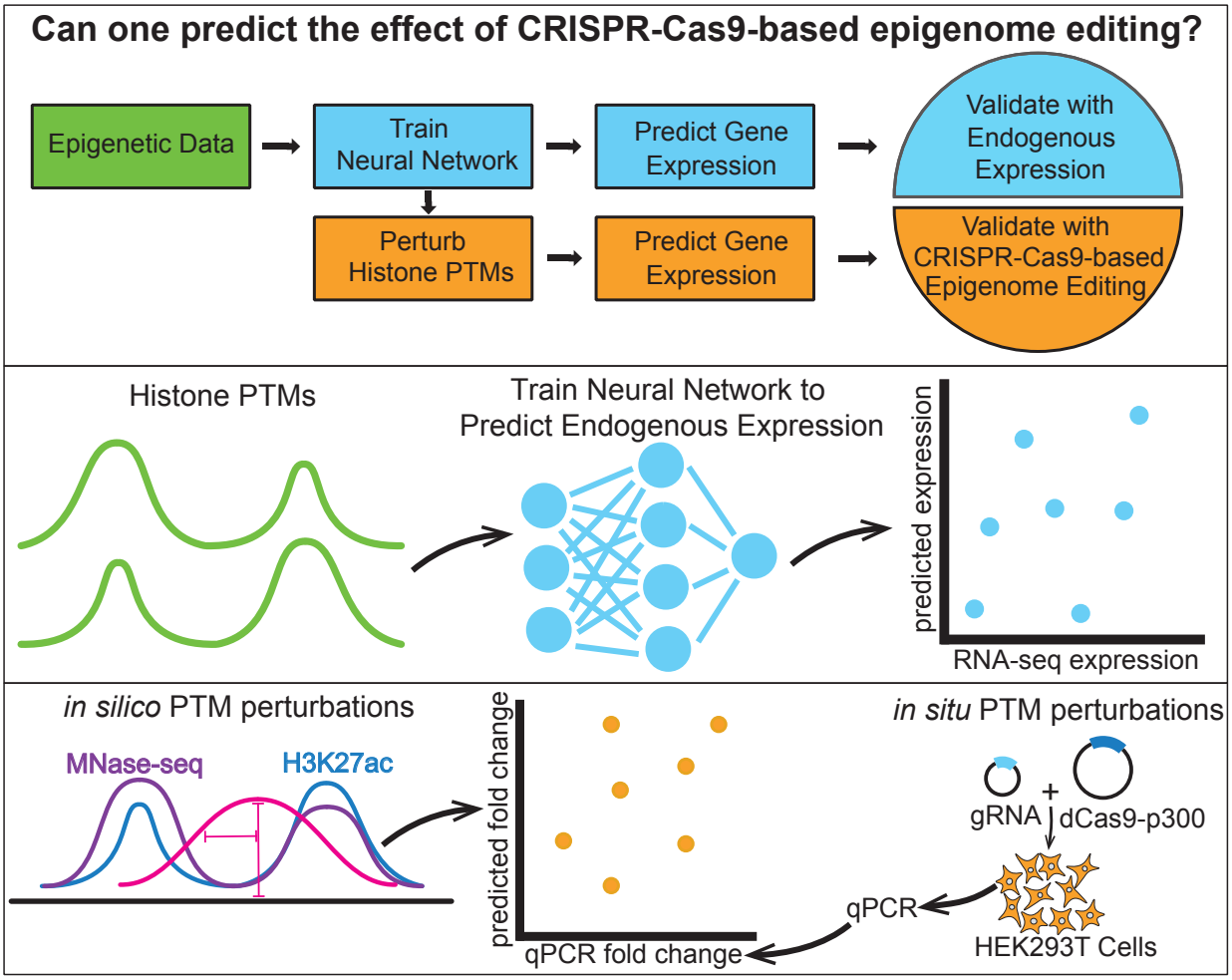

Figure 1: **Schematic of the epigenome editing prediction pipeline.** The pipeline uses epigenetic data to train models to predict endogenous gene expression. These models were used to predict fold change in gene expression based on perturbed histone PTM input data, and their predictions were validated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

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 (**Fig. 1**). We found that our models were effective in ranking relative fold changes among genes in response to the dCas9-p300 system, achieving a Spearman's rank correlation of $\sim$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.

2 Results

2.1 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) (**Supplementary Table S1**). 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 (**Supplementary Fig. S1**).

We corrected these batch effects by adapting S3norm (Xiang et al. 2020) (Section 4, **Supplementary Fig. S2**). 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 (Section 4). Importantly, we observed that H3K27ac and H3K4me3 histone PTM signal strength positively covaried with gene expression quantile (representative cell types shown in **Fig. 2**; all cell types shown in **Supplementary Fig. S3**). Conversely, repressive histone PTMs such as H3K27me3 and H3K9me3 seemed strongly anti-correlated with gene expression quantile. 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).

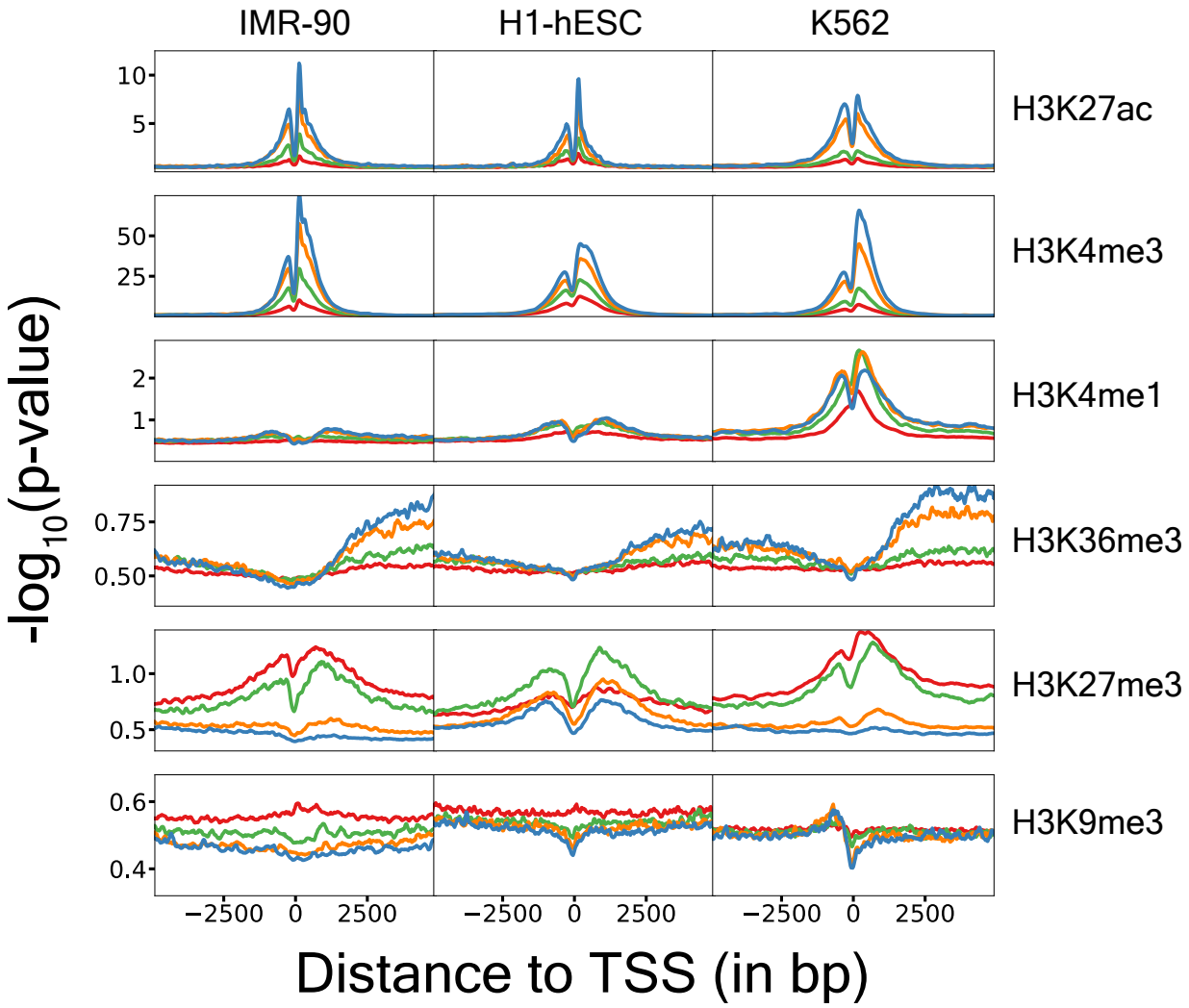

Figure 2: **Metagene plots show histone PTMs are consistent across cell types and recapitulate established relationships between histone PTMs and gene expression.** Colors represent genes binned into quantiles based on gene expression. Blue 75-100%, Orange 50-75%, Green 25-50%, Red 0-25% of gene expression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

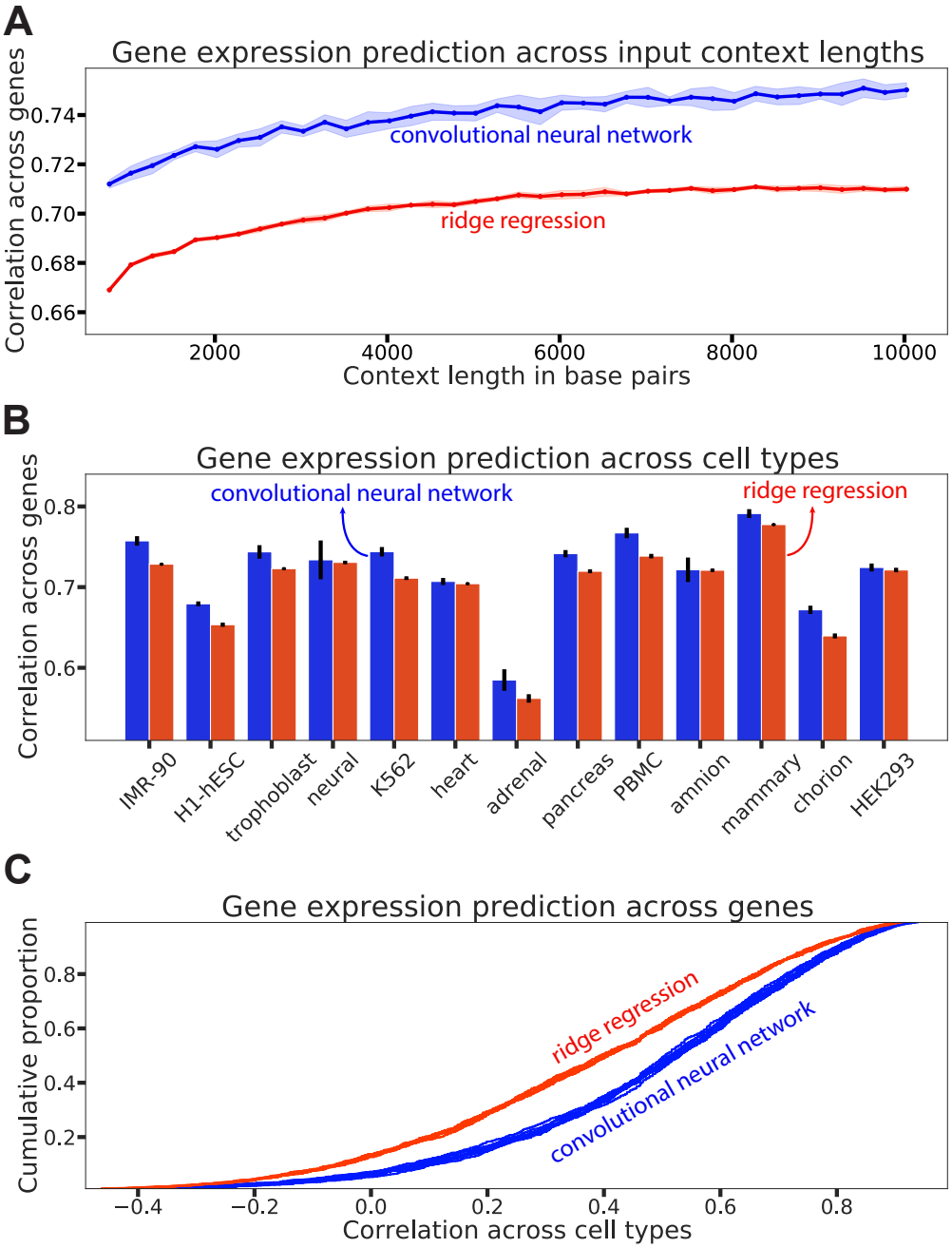

Figure 3: **Histone PTMs accurately predict endogenous gene expression.** **(A)** Spearman correlation on genes from held out chromosomes for different input context lengths, with all cell types pooled together. Blue curve is the mean across 10 computational replicates of CNNs and the red is the mean 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.

2.2 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 (Section 4, **Fig. 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 (**Fig. 3A**). 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. Concretely, 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 (**Fig. 3B**). The reduced performance on the adrenal cell type may be driven by a lower correlation of its epigenetic data with other cell types, particularly for H3K36me3 (**Supplementary Fig. S5**).

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 (**Fig. 3C**). In particular, the median cross-cell type correlation is $\sim$0.53 for CNNs compared to $\sim$0.39 for ridge regression.

2.3 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 (**Fig. 4**, Section 4).

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. 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 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.

2.4 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 (**Fig. 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). Based on **Fig. 2** and **Fig. 4**, the largest

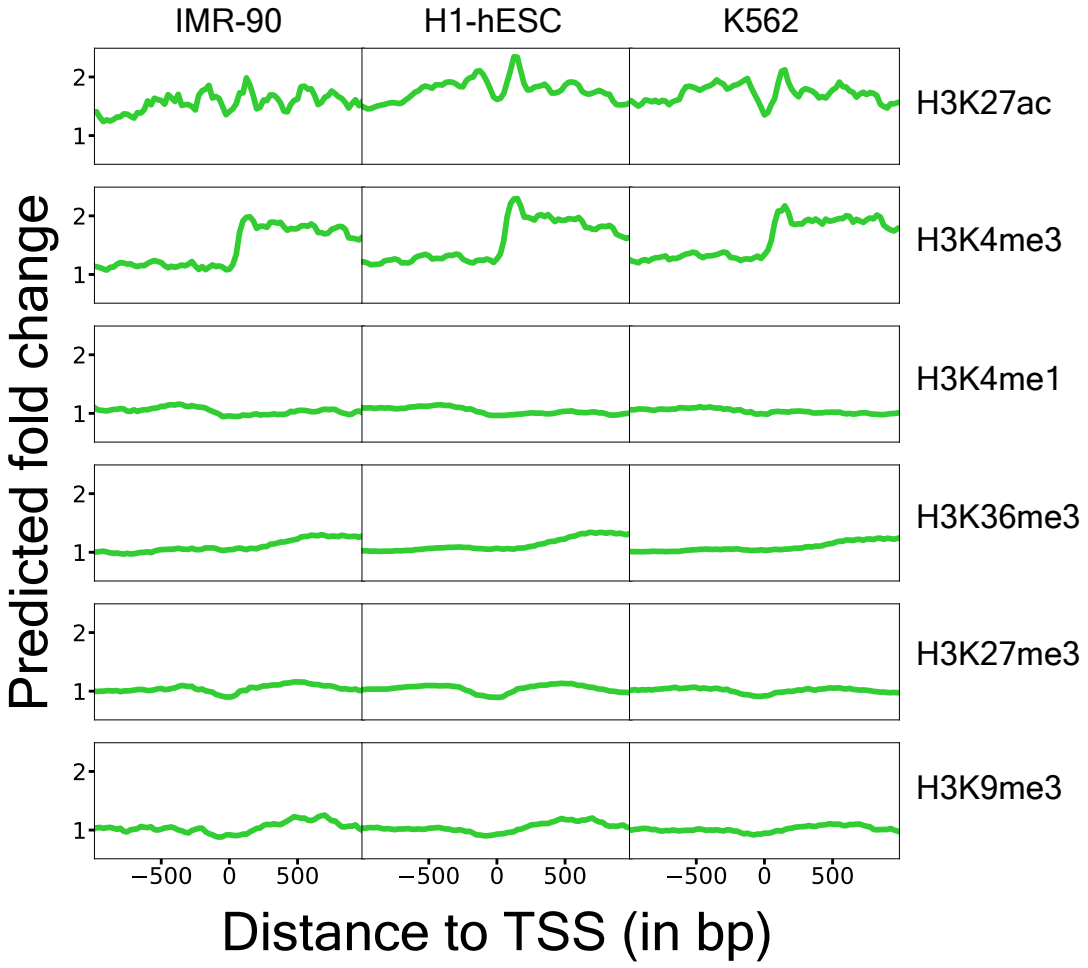

Figure 4: **Features learned by gene expression models.** Each point on the x-axis corresponds to *in silico* perturbation of that assay at that position and the y-axis measures the predicted fold change in gene expression, averaged across a set of 100 trained models. The fold changes were averaged across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

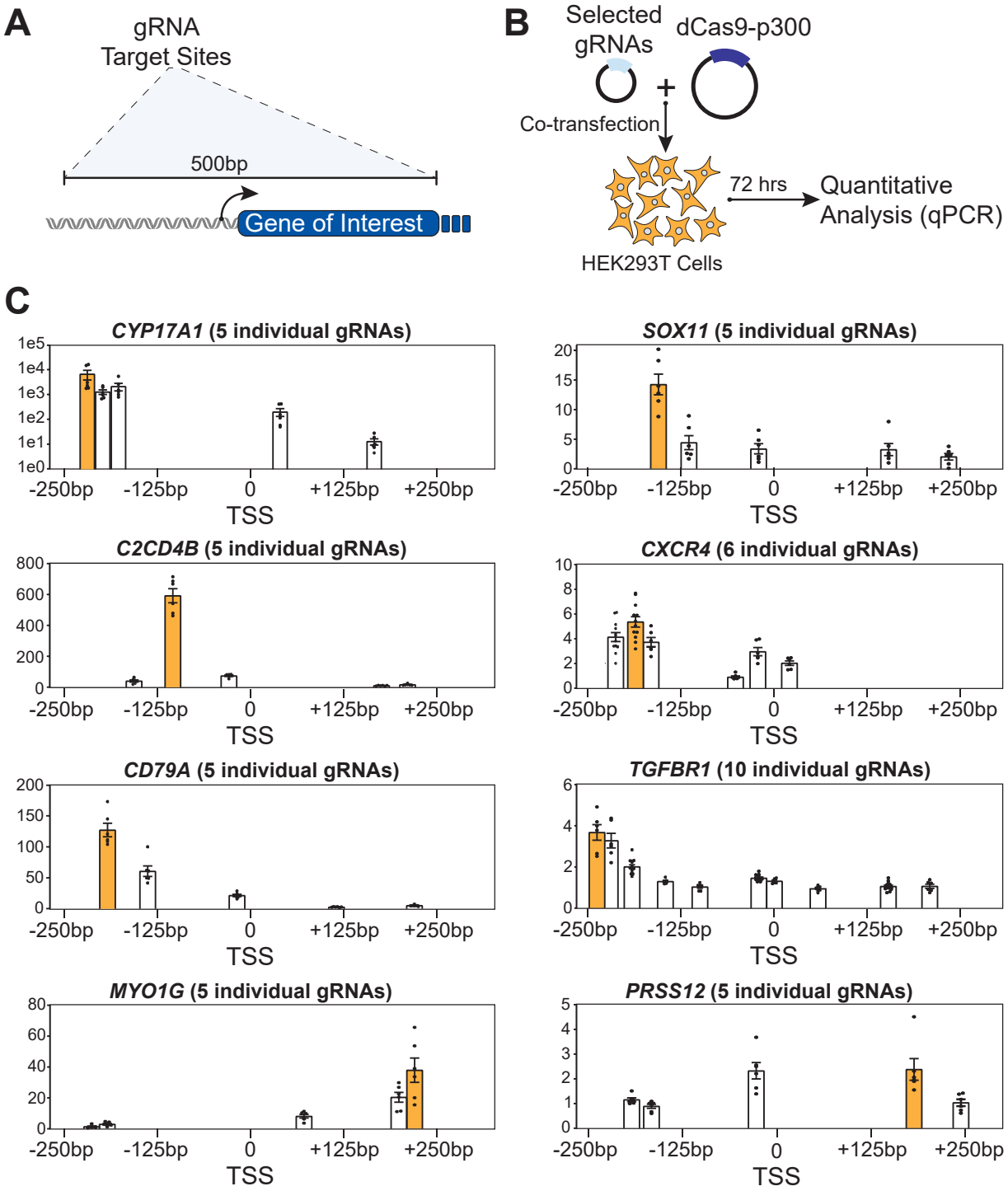

Figure 5: **dCas9-p300 epigenome editing at eight endogenous genes identifies gene specific responses.** The genes tested are CYP17A1, SOX11, C2CD4B, CXCR4, CD79A, TGFBR1, MYO1G, and PRSS12. (A) gRNA targeting +/− 250 bp of each gene were selected. (B) These Selected gRNA were individually co-transfected 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.

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 (**Fig. 5C**). 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 (**Supplementary Table S2**), as observed previously (Wang et al. 2022a). 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. For example, two gRNAs targeting within ∼50 base pairs of each other on C2CD4B have a 100-fold difference in measured mRNA (**Fig. 5C**). 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 (**Fig. 5C**).

2.5 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 unexplored); 3) dCas9-p300 can deposit H3K27ac at nucleosomes, as defined by MNase activity (see Section 4) (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 Section 4.

The first component of our perturbation model is steric hindrance of dCas9-p300 by nucleosomes (**Fig. 6A**). 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 (**Fig. 6B**). This adds acetylation primarily within a distance controlled by the standard deviation (σ) of the kernel. We performed CUT&RUN experiments (see Supplementary Material) that suggest that this distance is at least 1,000 base pairs (**Supplementary Fig. S6**). 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 (Section 4).

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. Concretely, we modulated the Gaussian kernel described above, by performing point-wise multiplication with MNase activity (**Fig. 6C**). 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 Supplementary Material).

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 $\sim$0.8 between the endogenous expression and that predicted by our expression model (**Fig. 6D**). This correlation was in line with the correlation observed across the endogenous transcriptome (**Fig. 3A,B**). We further observed that our expression models were able to accurately rank gene expression across cell types within individual genes (**Supplementary Fig. S7**).

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 $\sim$0.8 between these predicted fold changes and the experimentally determined mRNA fold changes induced by dCas9-p300 (**Fig. 6E**). However, the performance in ranking fold changes within individual genes was less accurate (**Supplementary Fig. S8**) when compared to the prediction of cell-type-specific gene expression from native epigenetic signatures (**Supplementary Fig. S7**).

3 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 (**Fig. 3**). We further showed that such models learned known relationships between histone PTMs and gene expression (**Fig. 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 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 accurate at predicting the outcome of these experiments within a given gene, as compared to predicting gene expression from the endogenous epigenetic signatures (**Supplementary Fig. S8**, **Supplementary Fig. S7**). 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

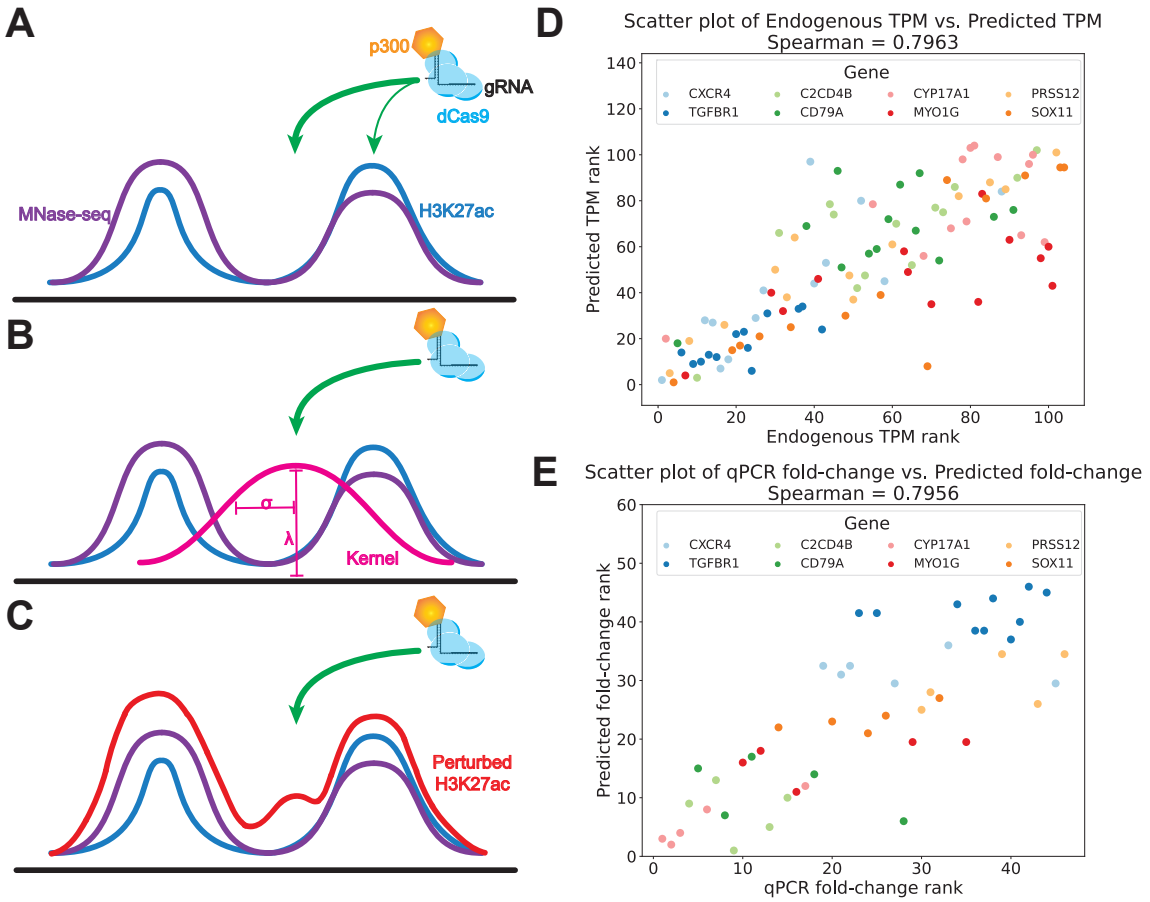

Figure 6: ***In silico* model for dCas9-p300-based epigenome editing** (A) dCas9-p300 is more likely to bind to a position not occupied by the nucleosome. Thicker green arrow represents higher probability of binding for a gRNA targeting that site. (B) The *in silico* perturbation is modeled as a Gaussian 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. Higher ranks correspond to higher values. (E) Ranks for predicted and empirically measured expression fold changes following perturbation by dCas9-p300 for 8 genes in HEK293T cells. Higher ranks correspond to higher values.

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.

Lastly, it is possible that our models of how histone PTMs affect gene expression may not generalize to the types of histone PTM profiles generated by epigenome editing experiments. One source of such a generalization error is failing to learn a *causal* model. Since we trained our models on endogenous gene expression across multiple cell types, our models may be inferring non-causal relationships. For example, epigenetic regulators within a cell could deposit histone PTMs in response to high gene expression(Wang et al. 2022b; Williams and Gehring 2020; Mazzone et al. 2019). In this case, our models would learn that this histone PTM should lead to a higher gene expression but in reality the causality is reversed. 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.

4 Materials and Methods

4.1 Data preparation

We obtained $-\log_{10}$(p-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 **Table S1**, from the ENCODE data portal (ENCODE Project Consortium 2012) and preprocessed them with a log operation.

4.2 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 (**Supplementary Fig. S2**). 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 (**Supplementary Fig. S3**).

4.3 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 **Fig. 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.

4.4 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:

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

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 **Fig. 3**. These perturbations corresponded to ~ 150 base pairs which is roughly the length of DNA wrapped around a nucleosome.

To produce **Fig. 4**, we applied the above perturbation functions, $F_1, \ldots, F_W$ 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.

4.5 *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 $\exp(-5 \times m_j)$.

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

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

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 (**Supplementary**

Table S2). 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.

4.6 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.

Declaration of interests

The authors have no competing interests.

Author Contributions

All authors conceived the project and assisted with experimental design. A.C. 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, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. October 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, **18,**(10) 1196–1203.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., et al. 2012. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, **41,**(D1) D991–D995.

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

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

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

Cui, K. and Zhao, K. Genome-wide approaches to determining nucleosome occupancy in metazoans using mnase-seq. In *Chromatin Remodeling*, pages 413–419. Springer, 2012.

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, **34,**(2) 184–191.

Dominguez, A. A., Chavez, M. G., Urke, A., Gao, Y., Wang, L., and Qi, L. S. 2022. CRISPR-mediated synergistic epigenetic and transcriptional control. *The CRISPR Journal*, **5,**(2) 264–275.

ENCODE Project Consortium. September 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489,**(7414) 57–74.

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

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X. S. 2012. Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, **7,**(9) 1728–1740.

Fletcher, R. and Powell, M. J. 1963. A rapidly convergent descent method for minimization. *The Computer Journal*, **6,**(2) 163–168.

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

Gemberling, M. P., Siklenka, K., Rodriguez, E., Tonn-Eisinger, K. R., Barrera, A., Liu, F., Kantor, A., Li, L., Cigliola, V., Hazlett, M. F., et al. 2021. Transgenic mice for in vivo epigenome editing with CRISPR-based systems. *Nature Methods*, **18,**(8) 965–974.

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

Gilbert, L. A., Horlbeck, M. A., Adamson, B., Villalta, J. E., Chen, Y., Whitehead, E. H., Guimaraes, C., Panning, B., Ploegh, H. L., Bassik, M. C., et al. 2014. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell*, **159,**(3) 647–661.

Goell, J. H. and Hilton, I. B. 2021. Crispr/cas-based epigenome editing: advances, applications, and clinical utility. *Trends in Biotechnology*, **39,**(7) 678–691.

Hafner, A. and Boettiger, A. 2022. The spatial organization of transcriptional control. *Nature Reviews Genetics*, pages 1–16.

Hilton, I. B., D'ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E., and Gersbach, C. A. 2015. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology*, **33,**(5) 510–517.

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

Horlbeck, M. A., Witkowsky, L. B., Guglielmi, B., Replogle, J. M., Gilbert, L. A., Villalta, J. E., Torigoe, S. E., Tjian, R., and Weissman, J. S. 2016. Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *eLife*, **5,** e12677.

Isaac, R. S., Jiang, F., Doudna, J. A., Lim, W. A., Narlikar, G. J., and Almeida, R. 2016. Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. *eLife*, **5,** e13450.

Jenuwein, T. and Allis, C. D. August 2001. Translating the histone code. *Science*, **293,**(5532) 1074–1080.

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

Karlić, R., Chung, H.-R., Lasserre, J., Vlahovicek, K., and Vingron, M. February 2010. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, **107,**(7) 2926–2931.

Keung, A. J., Joung, J. K., Khalil, A. S., and Collins, J. J. March 2015. Chromatin regulation at the frontier of synthetic biology. *Nat. Rev. Genet.*, **16,**(3) 159–171.

Kimura, H. July 2013. Histone modifications for human epigenome analysis. *J. Hum. Genet.*, **58,**(7) 439–445.

Klann, T. S., Black, J. B., Chellappan, M., Safi, A., Song, L., Hilton, I. B., Crawford, G. E., Reddy, T. E., and Gersbach, C. A. 2017. CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nature Biotechnology*, **35,**(6) 561–568.

Klemm, S. L., Shipony, Z., and Greenleaf, W. J. 2019. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, **20,**(4) 207–220.

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

Kwon, D. Y., Zhao, Y.-T., Lamonica, J. M., and Zhou, Z. 2017. Locus-specific histone deacetylation using a synthetic CRISPR-Cas9-based HDAC. *Nature Communications*, **8,**(1) 1–8.

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

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

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

Mali, P., Esvelt, K. M., and Church, G. M. 2013. Cas9 as a versatile tool for engineering biology. *Nature Methods*, **10,**(10) 957–963.

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

Mazzone, R., Zwergel, C., Artico, M., Taurone, S., Ralli, M., Greco, A., and Mai, A. 2019. The emerging role of epigenetics in human autoimmune disorders. *Clinical Epigenetics*, **11,**(1) 1–15.

McKnight, L. E., Crandall, J. G., Bailey, T. B., Banks, O. G., Orlandi, K. N., Truong, V. N., Donovan, D. A., Waddell, G. L., Wiles, E. T., Hansen, S. D., et al. 2021. Rapid and inexpensive preparation of genome-wide nucleosome footprints from model and non-model organisms. *STAR protocols*, **2,**(2) 100486.

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

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

Nuñez, J. K., Chen, J., Pommier, G. C., Cogan, J. Z., Replogle, J. M., Adriaens, C., Ramadoss, G. N., Shi, Q., Hung, K. L., Samelson, A. J., et al. 2021. Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. *Cell*, **184,**(9) 2503–2519.

O'Geen, H., Ren, C., Nicolet, C. M., Perez, A. A., Halmai, J., Le, V. M., Mackay, J. P., Farnham, P. J., and Segal, D. J. 2017. dCas9-based epigenome editing suggests acquisition of histone methylation is not sufficient for target gene repression. *Nucleic Acids Research*, **45,**(17) 9901–9916.

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

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., and Aiden, E. L. December 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159,**(7) 1665–1680.

Reik, W., Dean, W., and Walter, J. 2001. Epigenetic reprogramming in mammalian development. *Science*, **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, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. February 2015. Integrative analysis of 111 reference human epigenomes. *Nature*, **518,**(7539) 317–330.

Sanson, K. R., Hanna, R. E., Hegde, M., Donovan, K. F., Strand, C., Sullender, M. E., Vaimberg, E. W., Goodale, A., Root, D. E., Piccioni, F., et al. 2018. Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nature Communications*, **9,**(1) 1–15.

Schmidt, F., Gasparoni, N., Gasparoni, G., Gianmoena, K., Cadenas, C., Polansky, J. K., 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., Hengstler, J. G., Hamann, A., Lengauer, T., Rosenstiel, P., Walter, J., and Schulz, M. H. January 2017. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.*, **45,**(1) 54–66.

Schmidt, F., Kern, F., and Schulz, M. H. February 2020. Integrative prediction of gene expression with chromatin accessibility and conformation data. *Epigenetics Chromatin*, **13,**(1) 4.

Schmidt, R., Steinhart, Z., Layeghi, M., Freimer, J. W., Bueno, R., Nguyen, V. Q., Blaeschke, F., Ye, C. J., and Marson, A. 2022. CRISPR activation and interference screens decode stimulation responses in primary human T cells. *Science*, **375,**(6580) eabj4008.

Schreiber, J., Bilmes, J., and Noble, W. S. March 2020a. Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples. *Genome Biol.*, **21,** (1) 82.

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

Segelle, A., Núñez-Álvarez, Y., Oldfield, A. J., Webb, K. M., Voigt, P., and Luco, R. F. 2022. Histone marks regulate the epithelial-to-mesenchymal transition via alternative splicing. *Cell Reports*, **38,**(7) 110357.

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

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., et al. 2014. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, **343,**(6166) 84–87.

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

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

Stillman, B. September 2018. Histone modifications: Insights into their influence on gene expression. *Cell*, **175,**(1) 6–9.

Strahl, B. D. and Allis, C. D. 2000. The language of covalent histone modifications. *Nature*, **403,** (6765) 41–45.

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

Swaminathan, V., Reddy, B. A. A., Ruthrotha Selvi, B., Sukanya, M. S., and Kundu, T. K. 2007. Small molecule modulators in epigenetics: implications in gene expression and therapeutics. *Subcell. Biochem.*, **41,** 397–428.

Taherian Fard, A. and Ragan, M. A. 2019. Quantitative modelling of the Waddington epigenetic landscape. *Methods Mol. Biol.*, **1975,** 157–171.

Thakore, P. I., Black, J. B., Hilton, I. B., and Gersbach, C. A. 2016. Editing the epigenome: technologies for programmable transcription and epigenetic modulation. *Nature Methods*, **13,**(2) 127–137.

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

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

Wang, Z., Chivu, A. G., Choate, L. A., Rice, E. J., Miller, D. C., Chu, T., Chou, S.-P., Kingsley, N. B., Petersen, J. L., Finno, C. J., et al. 2022b. Prediction of histone post-translational modification patterns based on nascent transcription data. *Nature Genetics*, **54,**(3) 295–305.

Weinert, B. T., Narita, T., Satpathy, S., Srinivasan, B., Hansen, B. K., Schölz, C., Hamilton, W. B., Zucconi, B. E., Wang, W. W., Liu, W. R., et al. 2018. Time-resolved analysis reveals rapid dynamics and broad scope of the CBP/p300 acetylome. *Cell*, **174,**(1) 231–244.

Williams, B. P. and Gehring, M. 2020. Principles of epigenetic homeostasis shared between flowering plants and mammals. *Trends in Genetics*, **36,**(10) 751–763.

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

Xu, K., Zhang, M., Li, J., Du, S. S., Kawarabayashi, K.-i., and Jegelka, S. 2020. How neural networks extrapolate: From feedforward to graph neural networks. *arXiv preprint arXiv:2009.11848*.

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

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

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

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

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

Zhou, X., Blocker, A. W., Airoldi, E. M., and O'Shea, E. K. 2016. A computational approach to map nucleosome positions and alternative chromatin states with base pair resolution. *Elife*, **5,** e16970.