RESEARCH ARTICLE

REVISED Predicting gene expression changes upon epigenomic drug treatment

[version 3; peer review: 4 approved, 2 not approved]

Piyush Agrawal[1,2], Vishaka Gopalan[2], Monjura Afrin Rumi[3], Sridhar Hannenhalli[2]

[1]Division of Medical Research, SRM Medical College Hospital & Research Centre, SRMIST, Kattankulathur, Chennai, Tamil Nadu, India
[2]Cancer Data Science Laboratory, National Cancer Institute, Bethesda, Maryland, 20814, USA
[3]Faculty research assistant, Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, 20742, USA

v3 **First published:** 01 Sep 2023, **12**:1089
https://doi.org/10.12688/f1000research.140273.1

Second version: 19 Dec 2023, **12**:1089
https://doi.org/10.12688/f1000research.140273.2

Latest published: 02 May 2025, **12**:1089
https://doi.org/10.12688/f1000research.140273.3

Abstract

Background

Tumors are characterized by global changes in epigenetic modifications such as DNA methylation and histone modifications that are functionally linked to tumor progression. Accordingly, several drugs targeting the epigenome have been proposed for cancer therapy, notably, histone deacetylase inhibitors (HDACi) such as vorinostat and DNA methyltransferase inhibitors (DNMTi) such as zebularine. However, a fundamental challenge with such approaches is the lack of genomic specificity, *i.e.*, the transcriptional changes at different genomic loci can be highly variable, thus making it difficult to predict the consequences on the global transcriptome and drug response. For instance, treatment with DNMTi may upregulate the expression of not only a tumor suppressor but also an oncogene, leading to unintended adverse effect.

Methods

Given the pre-treatment transcriptome and epigenomic profile of a sample, we assessed the extent of predictability of locus-specific changes in gene expression upon treatment with HDACi using machine learning.

Open Peer Review

Approval Status ✗ ✗ ✓ ✓ ✓ ✓

	1	2	3	4	5	6
version 3 (revision) 02 May 2025			✓ view		✓ view	✓ view
version 2 (revision) 19 Dec 2023		✗ view	? view	✓ view		
version 1 01 Sep 2023	✗ view					

1. **Angelika Merkel**, Josep Carreras Leukemia Research Institute (IJC), Barcelona, Spain
2. **Daiqing Liao**, University of Florida, Gainesville, USA
   Zhiguang Huo, University of Florida, Gainesville, USA
3. **Sukhen Das Mandal**, Ghani Khan Choudhary Institute of Engineering and Technology, Malda, West Bengal, India
4. **Indrakant K. Singh**, Deshbandhu College, University of Delhi, Delhi, India

Results

We found that in two cell lines (HCT116 treated with Largazole at eight doses and RH4 treated with Entinostat at 1µM) where the appropriate data (pre-treatment transcriptome and epigenome as well as post-treatment transcriptome) is available, our model distinguished the post-treatment up *versus* downregulated genes with high accuracy (up to ROC of 0.89). Furthermore, a model trained on one cell line is applicable to another cell line suggesting generalizability of the model.

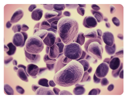
Conclusions

Here we present a first assessment of the predictability of genome-wide transcriptomic changes upon treatment with HDACi. Lack of appropriate omics data from clinical trials of epigenetic drugs currently hampers the assessment of applicability of our approach in clinical setting.

Keywords
Epigenetics, HDACi, DNMTi, Cancer therapy, Machine Learning, Transcriptomics

This article is included in the Bioinformatics gateway.

This article is included in the Genomics and Genetics gateway.



This article is included in the Bioinformatics in Cancer Research collection.

5. **Vivek Dhar Dwivedi**, Quanta Calculus, Greater Noida, India
6. **Ashwini Kumar**, Sharda University, Greater Noida, India

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Piyush Agrawal (apiyush74@gmail.com), Sridhar Hannenhalli (sridhar.hannenhalli@nih.gov)

Author roles: Agrawal P: Conceptualization, Data Curation, Formal Analysis, Methodology, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Gopalan V**: Conceptualization, Data Curation, Formal Analysis, Methodology, Writing – Original Draft Preparation; **Rumi MA**: Data Curation, Formal Analysis, Methodology; **Hannenhalli S**: Conceptualization, Formal Analysis, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the National Cancer Institute (Intramural Research funding).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Agrawal P, Gopalan V, Rumi MA and Hannenhalli S. **Predicting gene expression changes upon epigenomic drug treatment [version 3; peer review: 4 approved, 2 not approved]** F1000Research 2025, **12**:1089 https://doi.org/10.12688/f1000research.140273.3

First published: 01 Sep 2023, **12**:1089 https://doi.org/10.12688/f1000research.140273.1

REVISED **Amendments from Version 2**

In the revised manuscript, we have updated the gene ontology section providing an explanation about why genes are associated with different types of GO terms for different cell lines. Also, we have rewritten the results associated with the Figure 6.

Any further responses from the reviewers can be found at the end of the article

Introduction

Phenotypic state of a cell or tissue, either in normal homeostasis or in disease such as cancer, is intimately linked to its transcriptional state, which in turn is profoundly determined by its global epigenome.[1] Cancer genomes display a substantially altered epigenome relative to their non-malignant counterparts. For instance, global DNA hypomethylation and focal hypermethylation, notably, at tumor suppressor gene promoters, have been noted as a general feature of many cancers.[2] Accordingly, drugs that alter the epigenome have emerged as potential candidates for cancer therapy.[3] Two of the most common classes of epigenome-altering drugs are DNA methyltransferase inhibitors (DNMTi), such as zebularine, and histone deacetylase inhibitors (HADCi), such as vorinostat. While DNMTi's are standard of care in some hematological malignancies, across most cancers, the efficacy of epigenomic drugs has been mixed.

One of the many reasons adversely affecting the success of epigenomic drugs is its lack of locus specificity. Note that the intent of DNMTi, for instance, is partly to reactivate aberrantly silenced genes by demethylating their (aberrantly methylated) promoters.[3,4] However, the drug is locus-agnostic and, *a priori*, can activate many other loci in the genome, some of which may have toxic side effects[5] and in the worst case, have pro-tumor effects; indeed, DNMTi's are known to activate cancer testis antigens, which are known to be pro-tumor. Currently, we lack the sufficient knowledge to predict locus-specific effects of an epigenomic drug on the gene expression, to be able to develop more rational therapies. One of the facts complicating this understanding is incompletely understood interactions between different epigenomic marks. For instance, there is broad antagonism between two key mechanisms of transcriptional suppression, namely, histone modification H3K27me3 and DNA methylation[6] which may result in redistribution of one upon perturbation of the other. Because of the complex interactions between epigenomic marks as well as feed-forward and feed-back loops between the epigenome and the transcriptome, the ultimate effect of epigenomic perturbation on the global transcriptome may not be easily predictable, especially based only on the local genomic context.

In previous studies, researchers tried to predict gene expression from the epigenome in a specific context.[7–9] However, we are interested in predicting the effect on the expression upon drug treatment, given the epigenomic profile in the pre-treatment sample. Our question necessitates availability of pre- and post-treatment gene expression and pre-treatment epigenomic profile, while these previous approaches are not concerned with the changes upon drug treatment. Hence, we believe, our work could substantially help assess the clinical efficacy of an epigenomic drug.

Methods

Processing of RH4 and HCT116 sequencing data

FASTQ files were downloaded from Sequence Read Archive (SRA) (HCT116 dataset accession: SRP113250, RH4 accession: SRP151465). HCT116 is human colorectal carcinoma cell line initiated from an adult male whereas RH4 cell line is for studying alveolar rhabdomyosarcoma and is belongs to soft tissue lineage. We uniformly re-processed both RNA-seq and H3K27ac ChIP-seq data to minimize biases. We ran the fastqc toolkit (v0.11.9) to ensure quality. Trimgalore (v0.6.7) was run with default options to trim any adaptor sequence contamination in reads. For ChIP-seq data, bwa-mem2 (v2.2.1)[10] was used to align trimmed reads while salmon (v1.7.0)[11] was run to align trimmed reads with the –validateMappings option enabled. For ChIP-seq data, the read counts in each genomic bin (defined below) were normalized to TPM (transcripts per million) scale with genomic bin counts quantified using the Rsubread package.[12] Since the RH4 ChIP-seq data contained spike-in reads from the *Drosophila melanogaster* genome, bwa-mem2 was used to align reads using a joint BWA index of the hg38 and dm6 genome. Thus, for the RH4 data, in addition to the library size normalization that is applied to each sample, we additionally divided the TPM values by the total number of reads that aligned to the dm6 genome assembly following the recommendation in the source publication describing the *Drosophila melanogaster* spike-in protocol for ChIP-seq data.[13] All pseudo-aligned RNA-seq data from salmon was normalized to a TPM scale using the tximport function (v1.28.0).

Distribution of histone marks in the genic region

We identified 1000 most upregulated and downregulated genes post-drug treatment for HCT116 and RH4 cell line. The genes were selected based on Log2 Fold change *i.e.,* Log2 Treated – Log2 Untreated. For every gene, we created 21 genomic bins to analyze the pattern of histone marks. The genomic bins include promoter region, transcription Start

Site (TSS), and Gene Body (GB) region. TSS coordinates were obtained from the ENSEMBL Genes v101 database.[14] We defined the promoter as the 2kb region upstream to the TSS which was further divided into 10 equal-sized bins where the TSS was the single nucleotide position. Finally, the gene body was defined as the entire transcribed region and was also divided into 10 equal-sized bins. Overall, this resulted in a total of 21 bins for every gene. H3K27Ac read density was calculated in each of these 21 bins and was used to compare the up and down genes and as features for the prediction of up and down-regulated genes.

Machine learning model to predict post-treatment transcriptional effect

We used the histone mark distribution in 21 genic bins as features to develop machine learning models to distinguish up versus downregulated genes after HDACi-treatment separately for both HCT116 and RH4 cell lines. Using the conventional five-fold cross-validation we computed the area under curve (AUC) as performance measure. We used a python-based library known as Scikit-learn[15] and implemented three different machine learning techniques which include Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting. Models were developed in four different categories (i) using 10 Promoter features; (ii) using single TSS feature; (iii) using 10 GB features, and lastly (iv) using all 21 features. We further performed cross cell line prediction where a model trained on one cell line data was used to predict other cell line data.

Gene Ontology analysis

We used clusterProfiler 4.0[16] to identify biological processes associated with the identified up and downregulated genes. We used the following command to get the enriched significant processes:

```
“ego <- enrichGO (de$Entrezid, OrgDb = “org. Hs.eg.db”, ont=“BP”, readable = TRUE, minGSSize = 10,
maxGSSize = 500, keyType=“SYMBOL”)”
```

As there are many redundant processes, we further obtained the parent processes using the following command:

```
“ego2 <- simplify (ego, cutoff=0.8, by=“p.adjust”, select_fun=min, measure = “Wang”)”
```

Dotplot of the above obtained processes were created using ggplot2 library in R.[17]

Results

Distinct patterns of epigenomic marks in gene locus between up- and down-regulated genes upon HDACi treatment

For each of the two cell lines (HCT116 and RH4), for the respective dosage of HDACi drugs (eight doses of Largazole for HCT116, one dose of 1 μM Entinostat for RH4), we first identified the top 1000 up-regulated and 1000 down-regulated genes (Methods). Genes classified as up, down, and unchanged post-treatment for various doses in HCT116 and RH4 cell lines are provided in Tables S1 and S2 respectively of *Extended data.*[18] TPM value of each gene, untreated as well as treated for various concentrations in HCT116 cell line and single concentration for RH4 is also provided in Tables S3 and S4 of *Extended data*[18] respectively.

We first identified enriched GO terms in each set of up and down-regulated genes (three pairs of gene sets for three representative doses in HCT116 [4.68 nM, 75 nM and 300 nM] and one pair for RH4). In general, the upregulated genes in both the cell lines were broadly enriched for the developmental and signaling processes (Figure 1). The developmental process is in the direction of epithelial to mesenchymal transition (EMT). In the case of HCT116 cell line, we additionally observed response to hypoxia. Likewise, processes associated with downregulated genes are broadly associated with the cell cycle and cell division, whereas for RH4 cell line, additional processes such as histone modification and RNA splicing were also seen (Figure 2).

Gene Ontology (GO) enrichment analysis revealed that genes upregulated by HDACi treatment were broadly involved in developmental processes and cell signaling pathways, suggesting a potential shift toward differentiation or activation of lineage-specific programs. In contrast, downregulated genes were strongly enriched for cell cycle-related processes, consistent with the known anti-proliferative effects of HDAC inhibitors. Notably, while this overall pattern was observed across all cell lines, the specific GO terms and pathways varied between them. This indicates that cell line-specific factors—such as differences in chromatin accessibility, baseline gene expression, or mutational background—may influence the transcriptional response to HDACi. Some variability may also stem from the use of different HDAC inhibitors, which may target different HDAC isoforms and thus modulate distinct regulatory networks. Disentangling these two possibilities however will necessitate a broader set of experiments involving multiple cell lines and multiple HDACi with proper experimental design.

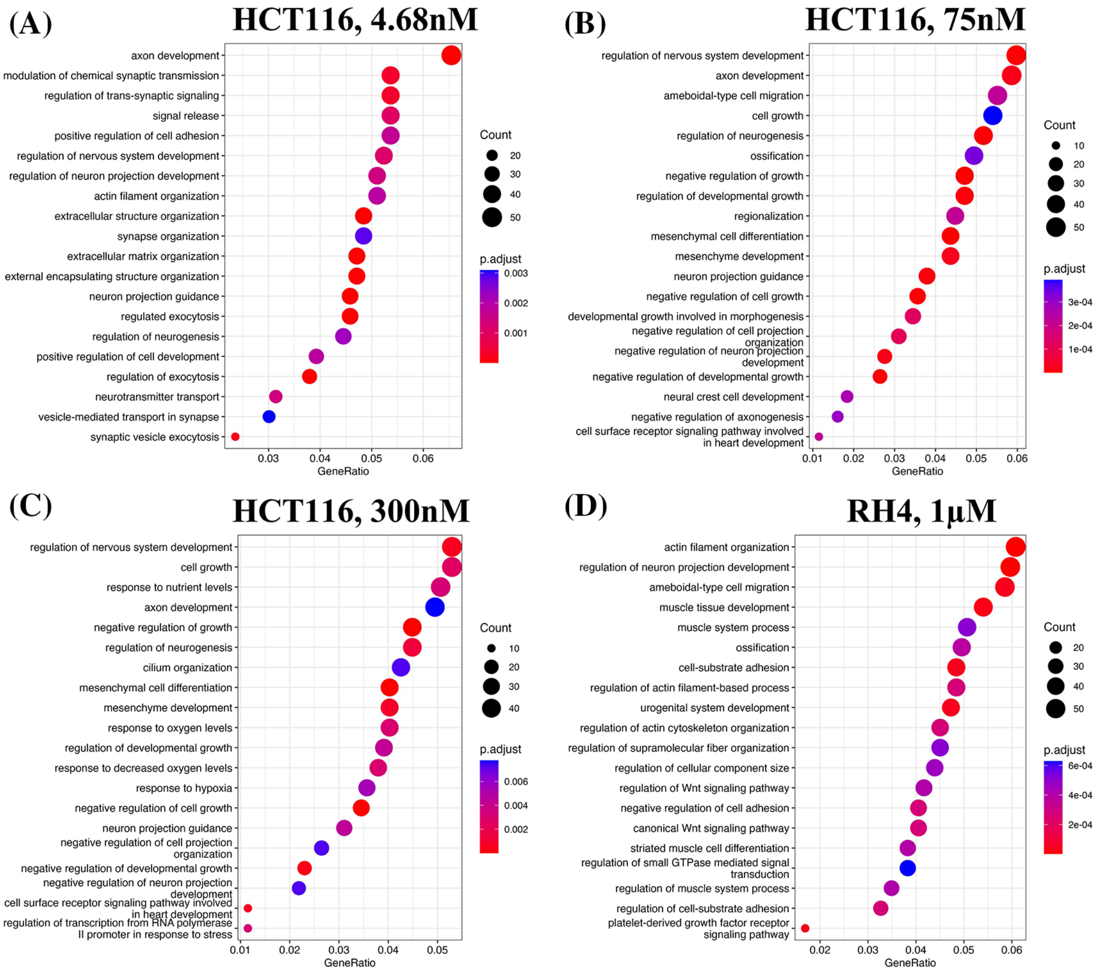


Figure 1. Upregulated genes are broadly enriched for the developmental and signaling processes. Top 20 enriched biological processes associated with upregulated genes in HCT116 cell line after treating with epigenetic drug largazole at 4.68 nM (A); 75 nM (B); 300 nM (C); and Top 20 enriched biological processes associated with upregulated genes in RH4 cell line after treating with epigenetic drug entinostat at 1 µM (D).

Complete lists of processes associated with the upregulated and downregulated genes in HCT116 [4.68 nM, 75 nM and 300 nM] and RH4 cell lines are provided in Tables S5-S7 and S8 respectively of *Extended data.*[18]

Next, we compared the pre-treatment epigenomic profiles of the up- and downregulated genes by plotting the H3K27Ac mark intensity (normalized read counts) in the pre-treatment sample along 21 genic bins (Methods). Distributions for three representative doses for HCT116 (4.68 nM (lowest), 75 nM, and 300 nM (highest)) and 1 µM dose for RH4 are included in Figure 3; all other distributions for the HCT116 cell line are provided in Figures S1-S5 of *Extended data.*[18] Overall, the following general trends emerged: (1) There was substantial variability across the bins around the genic locus in the upregulated *versus* downregulated H3K27Ac mark density, (2) in the upstream regions downregulated genes had a higher H3K27Ac pre-treatment; (3) this trend was also true in gene body but only at mid and higher dosage, while (4) at low dosage the trend was opposite in gene body where the downregulated genes had lower H3K27Ac; (5) RH4 trends at 1 µm dose of Entinostat most resembled the patterns at 75 nM dose of largazole in HCT116. Overall, while there was a variable pre-treatment epigenomic pattern within the gene body across cell lines, drug, and dosages, there were nevertheless sufficient differences between up- and downregulated genes, motivating us to develop machine learning models to predict transcription effects given the H3K27Ac pattern at a gene locus.

Predicting HDACi treatment impact on gene expression from the epigenome

Here, we assess whether the pre-treatment epigenetic profile at a gene locus can predict whether the gene will be upregulated or downregulated upon treatment with HDACi. The top 1000 upregulated and 1000 downregulated genes were compiled. For every gene, pre-treatment H3K27Ac read count in 21 regions relative to the gene (Methods) were

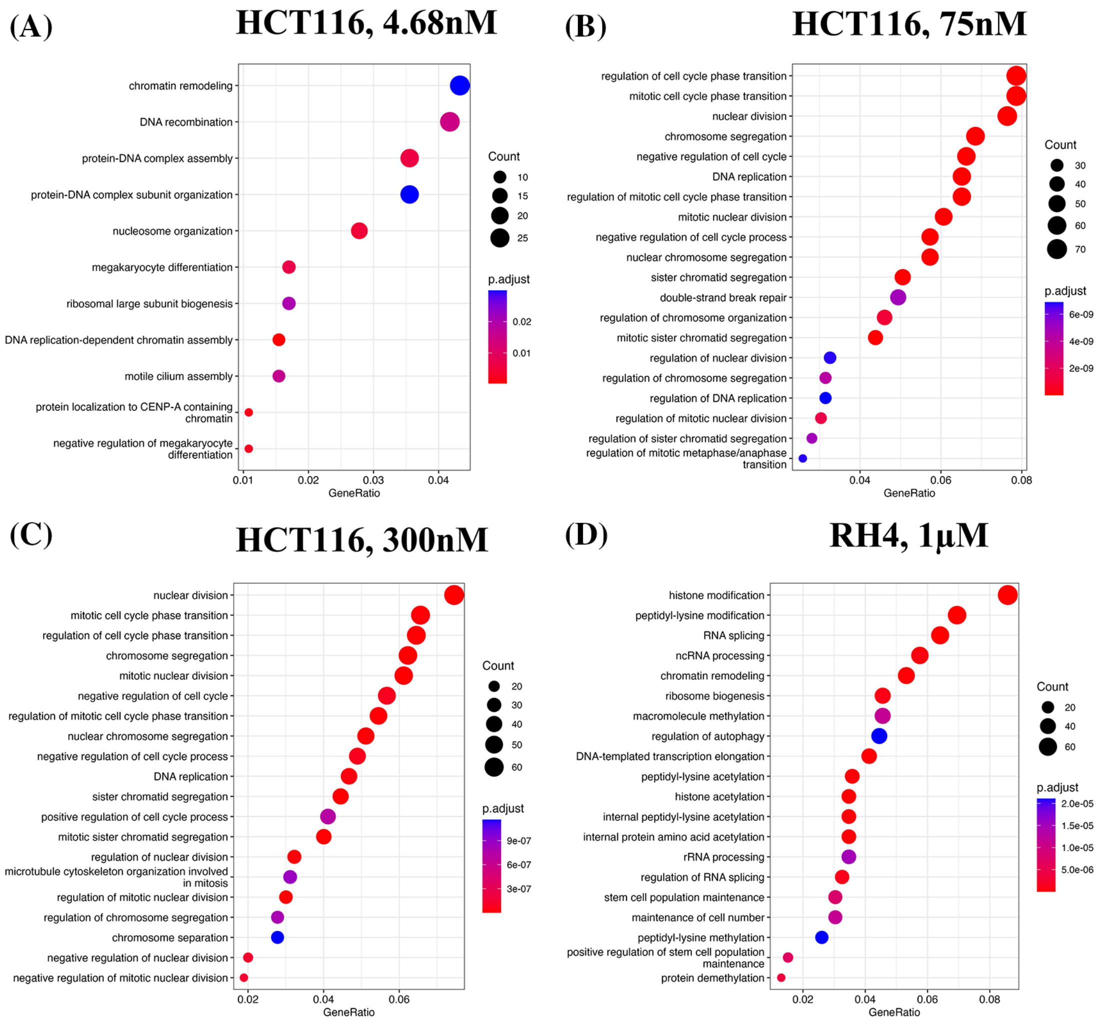


Figure 2. Downregulated genes are broadly enriched for the cell cycle processes. Top 20 enriched biological processes associated with downregulated genes in HCT116 cell line after treating with epigenetic drug largazole at 4.68 nM (A); 75 nM (B); 300 nM (C); and Top 20 enriched biological processes associated with upregulated genes in RH4 cell line after treating with epigenetic drug entinostat at 1 μM (D).

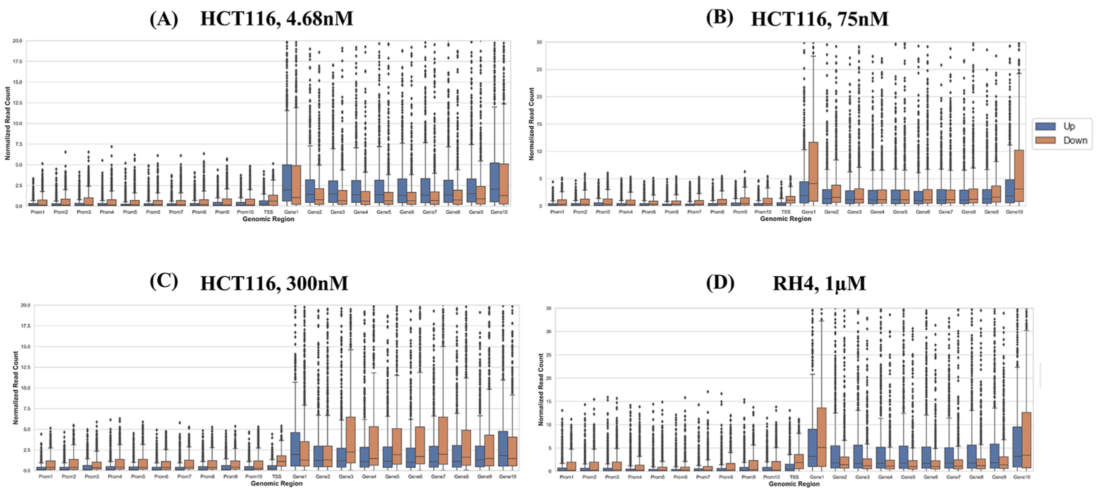


Figure 3. H3K27Ac mark distribution across genomic bins. Boxplot distribution of H3K27Ac marks across 21 genomic bins (10 equal sized bins of Promoter, Gene body and 1 bin of TSS) associated with upregulated (blue bars) and downregulated (brown bars) genes when HCT116 cell line is treated with largazole at concentrations 4.68 nM (A); 75 nM (B); 300 nM (C); and when RH4 cell line is treated with entinostat at 1 μM (D).

used as features and three machine learning models – Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (GB), were benchmarked based on five-fold cross-validation and accuracy was quantified as area under the ROC curve (AUC). A separate model was benchmarked for each of the eight drug dosages in HCT116 data. As shown in Table 1, overall, various machine learning approaches performed comparably and using all features was preferable; specifically, the best performance was achieved by SVM for 75 nM dosage with AUC of 0.89 (Figure 4). Analogous benchmarking for RH4 cell line data available at the single dosage using all features yielded comparable AUC

Table 1. Performance of various machine learning models on HCT116 and RH4 cell line testing dataset. Here the concentration of drugs is in nM and μM. P stands for Promoter; TSS stands for Transcription Start Site; GB stands for Gene Body; and All is the combination of all three features.

	Support vector machine				Random forest				Gradient boosting			
Concentration	**P**	**TSS**	**GB**	**All**	**P**	**TSS**	**GB**	**All**	**P**	**TSS**	**GB**	**All**
HCT116 (4.68 nM)	0.57	0.61	0.68	0.71	0.61	0.68	0.69	0.74	0.62	0.69	0.65	0.73
HCT116 (9.37 nM)	0.74	0.73	0.77	0.78	0.74	0.76	0.82	0.80	0.72	0.77	0.75	0.79
HCT116 (18.75 nM)	0.60	0.74	0.74	0.83	0.68	0.74	0.74	0.85	0.66	0.76	0.74	0.80
HCT116 (30 nM)	0.74	0.77	0.82	0.86	0.75	0.76	0.82	0.84	0.74	0.78	0.73	0.80
HCT116 (37.5 nM)	0.82	0.80	0.85	0.87	0.83	0.74	0.85	0.85	0.82	0.79	0.78	0.80
HCT116 (75 nM)	0.74	0.78	0.84	0.89	0.72	0.75	0.83	0.87	0.72	0.77	0.78	0.81
HCT116 (150 nM)	0.73	0.79	0.86	0.87	0.72	0.76	0.85	0.87	0.73	0.79	0.77	0.82
HCT116 (300 nM)	0.75	0.75	0.86	0.88	0.76	0.74	0.86	0.86	0.76	0.75	0.76	0.80
RH4 (1μM)	0.74	0.67	0.71	0.75	0.71	0.63	0.73	0.76	0.70	0.66	0.70	0.74

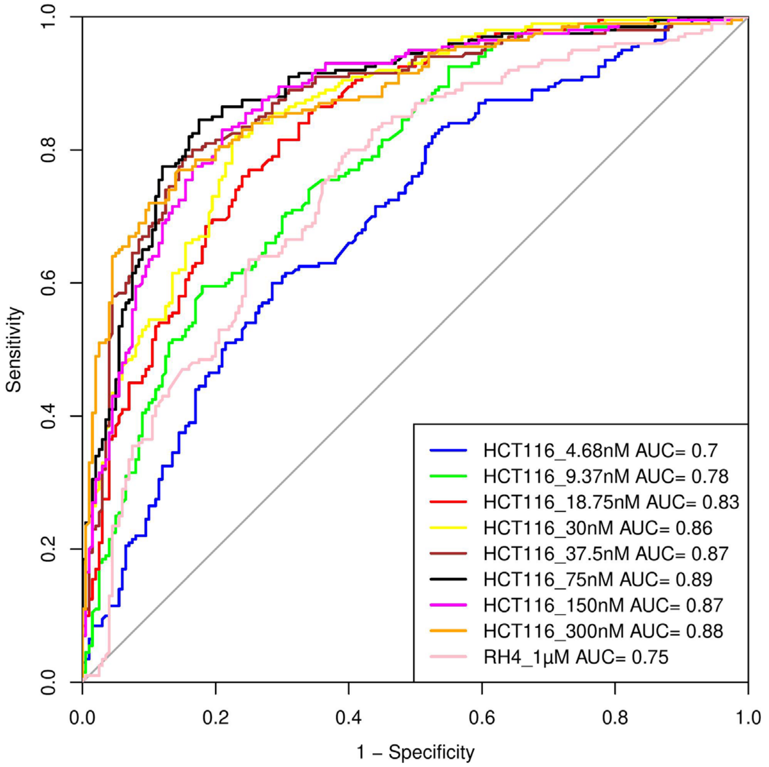


Figure 4. Performance of Support Vector Machine (SVM) model. Performance of various SVM based models in terms of Area Under Curve (AUC) at different concentrations when HCT116 cell line was treated with 8 different largazole concentration and RH4 cell line was treated with entinostat.

ranging from 0.74-0.76 for the three machine learning methods. Overall, H3K27Ac signal near the gene is informative of the gene expression changes upon treatment with HDACi.

Prediction model is generalizable across cell lines

Next, we assessed whether a model trained on one cell line to predict the transcriptional effect of a certain epigenomic drug can predict the effect in a different cell line treated with a different drug, albeit also HDACi. Toward this, first, a SVM (75 nM) model trained on HCT116 cell line data was able to achieve an AUC value of 0.71 when applied to RH4 cell line data (Figure 5A). Likewise, the model trained on RH4 cell line data when applied to HCT116 data achieved an AUC of 0.81 (Figure 5B), supporting the cross-context generalizability of the model, consistent with similarity of epigenomic profile trends between the two cell lines as shown above (Figure 3B and 3D).

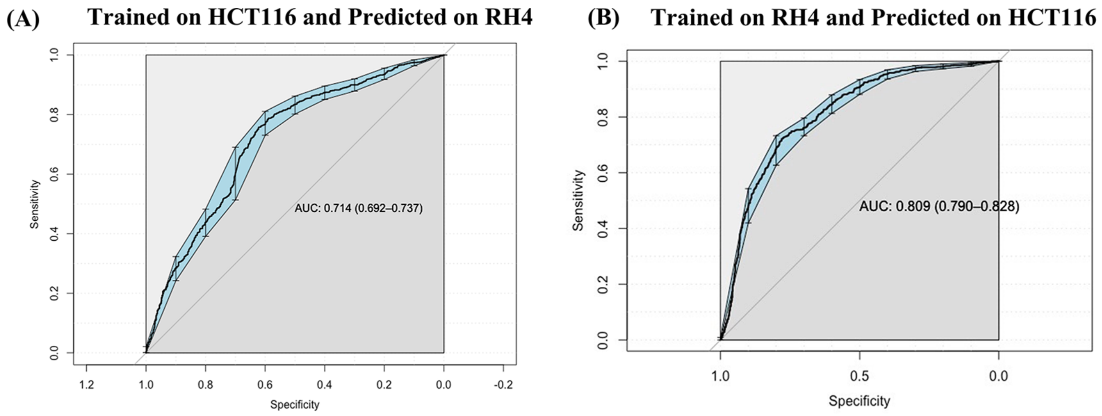


Figure 5. Cross cell line prediction: (A) Performance of HCT116 data trained Support Vector Machine (SVM) model on RH4 cell line used as testing dataset. (B) Performance of RH4 data trained SVM model on HCT116 cell line used as testing dataset.

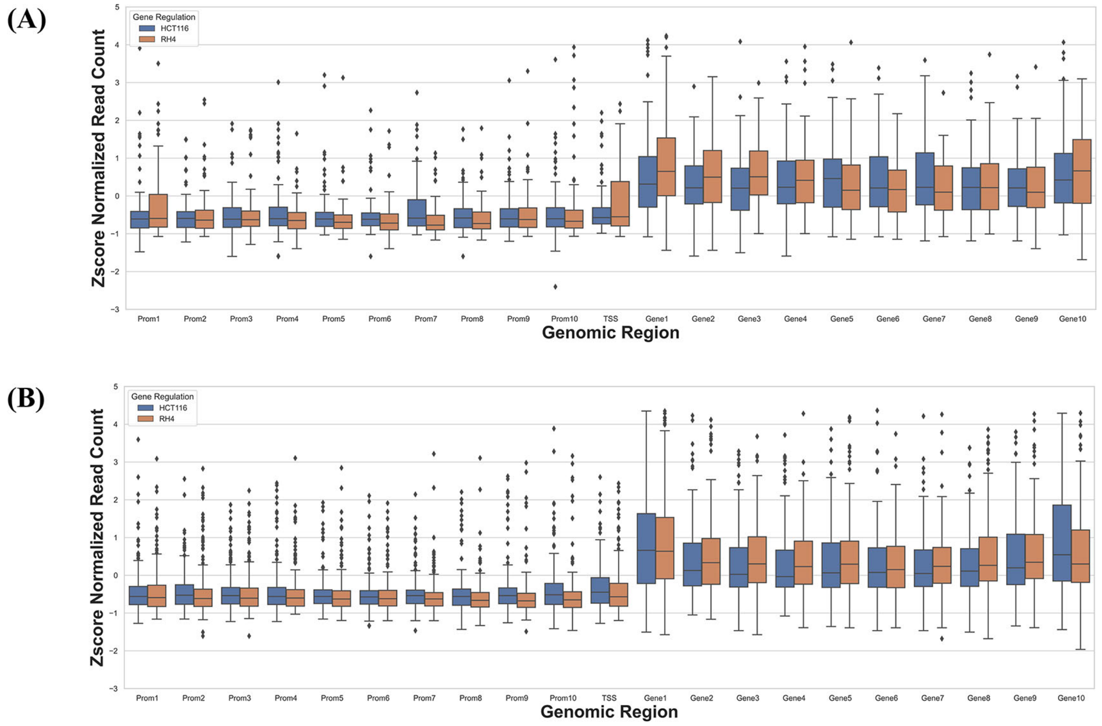


Figure 6. H3K27Ac mark distribution across genomic bins in two cell lines. Boxplot distribution of H3K27Ac marks across 21 genomic bins (10 equal sized bins of Promoter, Gene body and 1 bin of TSS) associated with genes with positive log fold change in HCT116 but negative log fold change in RH4 cell line (A) and Genes with positive log fold change in RH4 but negative log fold change in HCT116 cell line (B).

Cell line-specific expression changes are reflected in their epigenome

Next, we specifically assessed whether the context-specific differences across the two cell lines in their HDACi-induced gene expression were reflected in their context-specific pre-treatment epigenomic profile in the gene locus. Toward this we compared the data for HCT116 treated with 75 nM largazole with RH4 treated with 1 μM of entinostat. For each cell line we applied stringent criteria to identify genes which were upregulated in one cell line and downregulated in another cell line. We selected those genes whose fold change >3 in one cell line and <1/3 in another. This resulted in two gene sets: (1) 73 genes upregulated in HCT116 and downregulated in RH4, and (2) 184 genes upregulated in RH4 and downregulated in HCT116. To normalize for cell line-specific differences in H3K27Ac, we z-scored the cross-bin H3K27Ac signal for each gene. Then for these two gene sets, we plotted the normalized H3K27Ac intensities along the 21 genic bins, comparing two cell lines features. As shown below, we do not observe any clear pattern for the H3K27Ac read density among the two cell lines, except at TSS and gene body 1 region, where we observed lower H3K27Ac read distribution for the upregulated genes (Figure 6A&B).

Discussion

Epigenetic dysregulation is a key characteristic of cancers. A number of mutations have been observed in the genes encoding epigenetic modifiers such as DNA methylation and histone modification enzymes.[19] Accordingly, efforts have been made in targeting epigenetic regulators.[20] At present, seven epigenetics-targeting drugs have been approved by the FDA.[21] However, there are certain challenges associated with this class of drugs, limiting their success. Some of the key challenges include (i) Different epigenetic mutations are associated with different cancer types; (ii) The same gene may have opposite function in tumorigenesis of different cancers. For example, *EZH2* deficiency causes myeloid malignancies[22] whereas gain-of-function causes B cell lymphomas[23]; (iii) Another major issue is the selectivity of these drugs. For example, 30 enzymes of the KDM family with similar JMJC domain belong to five subfamilies. These enzymes demethylate different histone residues. Hence, drugs targeting these are broad-spectrum, affecting multiple KDM subfamilies and histone marks with potentially unintended consequences[24]; (iv) Yet another issue with epigenetics-targeting drugs, focused on in this work, is the selectivity of genomic loci. For instance, a HDACi can both increase as well as decrease histone acetylation in different genomic loci and can thus upregulate certain genes while downregulating others, again with unintended consequences.

Here, we tried to address the selectivity issue by developing a machine learning model based on pretreatment histone mark. In two cell lines, we established that the locus-specific effect of HDACi treatment on gene expression can be predicted to a reasonable accuracy from the pre-treatment histone acetylation pattern at a gene locus, and the model appears to be generalizable across cell lines. While the current study is promising and may potentially be applied to personalized therapy by predicting the transcriptomic consequence of HDACi treatment, there are a few limitations which need to be addressed. Our predictive model is based only on the H3K27ac mark. Several other marks such as H3K9ac, H3K4me3, H3K27me3, among others, should be incorporated in such modeling approaches in the future as and when such data become available. Our model was assessed only in cell lines and its efficacy in bulk tumor data representing the tumor microenvironment remains to be assessed. Last but not the least, pre- and post-treatment tumor epigenetic and transcriptomic data in clinical and pre-clinical models are still lacking, required for assessing the clinical applicability of our approach.

Author contribution

VG download and processed the data. PA, and SH perform the analysis. PA and SH perform the statistical analysis. PA, VG and SH wrote the manuscript. PA and SH supervised the study. All authors read the article and approved the submitted version.

Declarations

We declare that no third-party material was used in this study and also, we have not used AI tools at any point in the preparation of the manuscript.

Data availability

Underlying data

Sequence Read Archive: Genome-wide Dose-dependent Inhibition of Histone Deacetylases Reveals Their Roles in Enhancer Remodeling and Suppression of Oncogenic Super-enhancers, https://identifiers.org/insdc.sra:SRP113250.[25]

Sequence Read Archive: Genome-wide Dose-dependent Inhibition of Histone Deacetylases Reveals Their Roles in Enhancer Remodeling and Suppression of Oncogenic Super-enhancers, https://identifiers.org/insdc.sra:SRP151465.[26]

Extended data

Figshare: Supplementary_Figures, https://doi.org/10.6084/m9.figshare.23736882.v1.[18]

This project contains the following extended data:

- Supplementary.xlsx
- Supplementary_Figures.docx

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Analysis code

Analysis code available from: https://github.com/hannenhalli-lab/Epigenetic_Project/

Archived analysis code at time of publication: https://zenodo.org/record/8212782.[27]

License: MIT

Acknowledgements

This work utilized the computational resources of the NIH HPC Biowulf cluster. Authors are also thankful to the other lab members for their valuable suggestions. This work was supported by the Intramural Research Program of the National Cancer Institute, Center for Cancer Research.

References

1. Henikoff S, Greally JM: **Epigenetics, cellular memory and gene regulation.** *Curr. Biol.* 2016 Jul 25 [cited 2023 Jul 17]; **26**(14): R644–R648.
Publisher Full Text | Reference Source
2. Dawson MA: **The cancer epigenome: Concepts, challenges, and therapeutic opportunities.** *Science.* 2017 Mar 17 [cited 2023 Jul 17]; **355**(6330): 1147–1152.
PubMed Abstract | Publisher Full Text
3. Jones PA, Issa JPJ, Baylin S: **Targeting the cancer epigenome for therapy.** *Nat. Rev. Genet.* 2016 Oct 1 [cited 2023 Jul 17]; **17**(10): 630–641.
Publisher Full Text | Reference Source
4. Amatori S, Bagaloni I, Donati B, *et al.*: **DNA demethylating antineoplastic strategies: a comparative point of view.** *Genes Cancer.* 2010 [cited 2023 Jul 17]; **1**(3): 197–209.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Feehley T, O'Donnell CW, Mendlein J, *et al.*: **Drugging the epigenome in the age of precision medicine.** *Clin. Epigenetics.* 2023 Dec 1 [cited 2023 Jul 17]; **15**(1): 6.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Reddington JP, Perricone SM, Nestor CE, *et al.*: **Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes.** *Genome Biol.* 2013 Mar 25 [cited 2023 Jul 17]; **14**(3): R25.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Chen Y, Xie M, Wen J: **Predicting gene expression from histone modifications with self-attention based neural networks and transfer learning.** *Front. Genet [Internet].* 2022 Dec 14 [cited 2023 Dec 7]; **13**: 1081842.
PubMed Abstract | Publisher Full Text
8. Singh R, Lanchantin J, Robins G, *et al.*: **DeepChrome: deep-learning for predicting gene expression from histone modifications.** *Bioinformatics [Internet].* 2016 Sep 1 [cited 2023 Dec 7]; **32**(17): i639–i648.
PubMed Abstract | Publisher Full Text
9. Karlić R, Chung HR, Lasserre J, *et al.*: **Histone modification levels are predictive for gene expression.** *Proc. Natl. Acad. Sci. U S A. [Internet].* 2010 Feb 16 [cited 2023 Dec 7]; **107**(7): 2926–2931.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Md V, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** *Proceedings - 2019 IEEE 33rd International Parallel and Distributed Processing Symposium, IPDPS 2019.* 2019 May 1; pp. 314–324.
11. Patro R, Duggal G, Love MI, *et al.*: **Salmon provides fast and bias-aware quantification of transcript expression.** *Nat. Methods.* 2017 [cited 2023 Jul 17]; **14**(4): 417–419.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Liao Y, Smyth GK, Shi W: **The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads.** *Nucleic Acids Res.* 2019 May 1 [cited 2023 Jul 17]; **47**(8): e47.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Orlando DA, Chen MW, Brown VE, *et al.*: **Quantitative ChIP-Seq normalization reveals global modulation of the epigenome.** *Cell Rep.* 2014 [cited 2023 Jul 17]; **9**(3): 1163–1170.
PubMed Abstract | Publisher Full Text
14. Cunningham F, Allen JE, Allen J, *et al.*: **Ensembl 2022.** *Nucleic Acids Res.* 2022 Jan 7 [cited 2023 Jul 19]; **50**(D1): D988–D995.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Pedregosa FABIANPEDREGOSAF, Michel V, Grisel OLIVIERGRISELO, *et al.*: **Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot.** *J. Mach. Learn. Res.* 2011 [cited 2023 Jul 17]; **12**: 2825–2830.
Reference Source
16. Wu T, Hu E, Xu S, *et al.*: **clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.** *Innovation (Cambridge (Mass)).* 2021 Aug 28 [cited 2023 Jul 17]; **2**(3): 100141.
Publisher Full Text | Reference Source
17. Wickham H: **ggpolt2 Elegant Graphics for Data Analysis.** *Use R! Series.* 2016; 211.
Publisher Full Text
18. Agrawal P: Supplementary_Figures.docx. [Data set]. *figshare.* 2023.
Publisher Full Text
19. Morel D, Jeffery D, Aspeslagh S, *et al.*: **Combining epigenetic drugs with other therapies for solid tumours - past lessons and future promise.** *Nat. Rev. Clin. Oncol.* 2020 Feb 1 [cited 2023 Jul 17]; **17**(2): 91–107.
PubMed Abstract | Publisher Full Text
20. Wang N, Ma T, Yu B: **Targeting epigenetic regulators to overcome drug resistance in cancers.** *Signal Transduct. Target. Ther.* 2023 Dec 1 [cited 2023 Jul 17]; **8**(1).
Publisher Full Text | Reference Source

21. Nepali K, Liou JP: **Recent developments in epigenetic cancer therapeutics: clinical advancement and emerging trends.** *J. Biomed. Sci.* 2021 Dec 1 [cited 2023 Jul 17]; **28**(1): 27.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

22. Rinke J, Chase A, Cross NCP, *et al.*: **EZH2 in Myeloid Malignancies.** *Cells.* 2020 Jul 8 [cited 2023 Jul 17]; **9**(7).
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

23. Chu L, Tan D, Zhu M, *et al.*: **EZH2 W113C is a gain-of-function mutation in B-cell lymphoma enabling both PRC2 methyltransferase activation and tazemetostat resistance.** *J. Biol. Chem.* 2023 Apr 1 [cited 2023 Jul 17]; **299**(4): 103073.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

24. Pfister SX, Ashworth A: **Marked for death: targeting epigenetic changes in cancer.** *Nat. Rev. Drug Discov.* 2017 Apr 1 [cited 2023 Jul 17]; **16**(4): 241–263.
PubMed Abstract | **Publisher Full Text**

25. Chemistry and Biochemistry, University of Colorado Boulder: Genome-wide Dose-dependent Inhibition of Histone Deacetylases Reveals Their Roles in Enhancer Remodeling and Suppression of Oncogenic Super-enhancers, Sequence Read Archive. [Dataset]. 2017.
Reference Source

26. Khan J, Genetics Branch, NCI, NIH: Genome-wide Dose-dependent Inhibition of Histone Deacetylases Reveals Their Roles in Enhancer Remodeling and Suppression of Oncogenic Super-enhancers, Sequence Read Archive. [Dataset]. 2019.
Reference Source

27. Agrawal P, *et al.*: **Github. Predicting gene expression changes upon epigenomic drug treatment.** 2023 [cited 2023 Aug 2].
Reference Source

Open Peer Review

Current Peer Review Status: ✗ ✗ ✓ ✓ ✓ ✓

Version 3

Reviewer Report 24 May 2025

https://doi.org/10.5256/f1000research.181146.r383150

✓ **Ashwini Kumar**
Sharda University, Greater Noida, Uttar Pradesh, India

The article titled "Predicting gene expression changes upon epigenomic drug treatment" is an interesting approach to show the effect of drug treatment on the genomic and epigenomic changes in cancer cells and prepare a model that can predict the post-treatment changes across the cancer types. Though drugs that target the epigenomic alterations in cancer have been clinically used, however they lack locus-specific action that may result in unpredictable adverse effects. Thus. the approach presented in this manuscript tackles pre- and post-treatment gene expression and pre-treatment epigenomic profile, and paves a way to predict the locus-specific effect(s) of the drugs that may lead us to personalize the therapy, and also lead us to develop more rational drugs.

Although such models can be made on in vitro studies involving various cancer cells, however this manuscript presents a potential new approach to tackle the issue of epigenomic changes that have substantial effect pre- and post-treatment. These models can help the researchers develop better drug development strategies while clinicians to have better therapeutic strategies minimizing the adverse effects.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Drug delivery and drug discovery

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 24 May 2025

https://doi.org/10.5256/f1000research.181146.r383152

✓

Vivek Dhar Dwivedi

Quanta Calculus, Greater Noida, India

The current study proposes a novel and interesting strategy to predict gene expression changes using histone modification patterns, with a focus on H3K27 acetylation, in response to drug exposure. The findings may play a key role in enhancing personalized treatment strategies for cancer patients undergoing histone deacetylase inhibitor therapy.

Still, a few minor but important concerns need to be addressed prior to its acceptance.

1. Authors should explain why they use only the mentioned 2 cell lines in the current study.
2. Please provide an explanation if this kind of approach will also work for other histone marks?
3. Reason behind selecting the SVM method in the case of Figure 5.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 24 May 2025

https://doi.org/10.5256/f1000research.181146.r382129

✓ **Sukhen Das Mandal**
Ghani Khan Choudhary Institute of Engineering and Technology, Malda, West Bengal, India

All my concerns have been addressed satisfactorily. Therefore, this article may be approved.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: RNA biology, Machine learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 07 May 2025

https://doi.org/10.5256/f1000research.159982.r267047

✓

Indrakant K. Singh
Deshbandhu College, University of Delhi, Delhi, India

The study presents a novel machine learning approach to predict gene expression changes from histone modification data, specifically H3K27 acetylation, after drug treatment. This could have significant implications for personalized cancer treatment with histone deacetylase inhibitors.

However, there are some concerns that need to be addressed before accepting.

1. The rationale behind choosing different parameters (e.g., z-scores of normalized read counts in Figure 6 versus normalized read counts in Figure 3) is not adequately explained. Consistency in data presentation or a clear justification for differences is needed.

2. The manuscript needs more detailed descriptions of certain methods. For example, the process of peak calling in ChIP-seq analysis and the significance of the log10 fold change in differential expression analysis are not clearly described. Detailed methodological transparency is essential for replication and understanding

3. Some results and their descriptions appear contradictory or unclear. For instance, the authors state that HCT116 genes showed higher concentrations of H3K27Ac marks, but this is not consistent with what is shown in Figure 6. Such inconsistencies need to be addressed and clarified.

I recommend minor revision.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics and Cancer Biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 01 August 2024

https://doi.org/10.5256/f1000research.159982.r243027

?

Sukhen Das Mandal
Ghani Khan Choudhary Institute of Engineering and Technology, Malda, West Bengal, India

In the paper authors developed machine learning models to predict gene expression changes for HDACi treatment. Though it will enrich current understanding about how epigenomic drugs can change expression of genes, there are multiple/some issues those need to be addressed.
Suggestions:
1. GO analysis of upregulated and down regulated genes for HDACi treatment showed that genes are associated with different types of GO terms for different cell lines. Whether it is because different cell lines respond differently to HDACi or different types of HDACi act through different mechanisms, you should discuss it more elaborately in the result section.
2. H3K27Ac profile for treated vs untreated should be checked for up and down regulated genes. This can verify whether change of H3K27Ac causes transcriptional change which leads to expression change or post transcriptional regulation like RNA stability causes the change of the expression.
3. "(3) this trend was also true in gene body but only at mid and higher dosage" or " (5) RH4 trends at 1 µm dose of Entinostat most resembled the patterns at 75 nM dose of largazole in HCT116" - These statements do not match with the figures. Kindly explain these more clearly.
4. The length of GB is different for every gene. The bin size of the genes having larger GB size is higher compared to genes having shorter GB. As a longer bin having more numbers of nucleotides gives the possibility of higher numbers of H3K27Ac count. So, authors should consider the length of GB into their analysis and should check whether the pattern persists after normalisation by the length of the GB of a gene.

5. "As shown below, with few exceptions, for the first gene set, HCT116 genes showed higher concentration of H3K27Ac marks and for the second gene set, the opposite was true (Figure 6B)". Again these statements are not matching with the figures rather the opposite may be true or there is no clear pattern to mention. Kindly recheck your results and explain it properly.
6. The reference of the figure 5A is not there in the text.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: RNA biology, Machine learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 25 Apr 2025

Piyush Agrawal

Reviewer 3
In the paper authors developed machine learning models to predict gene expression changes for HDACi treatment. Though it will enrich current understanding about how epigenomic drugs can change expression of genes, there are multiple/some issues those need to be addressed.
Suggestions:
1. GO analysis of upregulated and downregulated genes for HDACi treatment showed that genes are associated with different types of GO terms for different cell lines. Whether it is because different cell lines respond differently to HDACi or different types of HDACi act through different mechanisms, you should discuss it more elaborately in the result section.
Response: We thank the reviewer for this insightful comment. In response, we have

expanded the discussion in the Results section to clarify the biological relevance of the distinct GO term enrichments observed across different cell lines following HDACi treatment.
Specifically, we observed that **upregulated genes were consistently enriched for developmental and cell signaling processes**, while **downregulated genes were predominantly associated with cell cycle-related functions**. This pattern suggests that HDAC inhibition may generally suppress proliferation while activating differentiation or lineage-specifying pathways. However, the specific GO terms varied somewhat across different cell lines, which may either reflect context-specific effects of HDACi, or the **differences in the specific HDAC inhibitors used**; however, since different HDACi compounds were used in different cell lines, it is not possible to distinguish between these two possibilities.
We have now added this interpretation to the Results section (page 5, lines 139–149).

The following section has been added in the revised manuscript:
“Gene Ontology (GO) enrichment analysis revealed that genes ***upregulated by HDACi treatment*** *were broadly involved in* ***developmental processes and cell signaling pathways****, suggesting a potential shift toward differentiation or activation of lineage-specific programs. In contrast,* ***downregulated genes were strongly enriched for cell cycle-related processes****, consistent with the known anti-proliferative effects of HDAC inhibitors. Notably, while this overall pattern was observed across all cell lines, the specific GO terms and pathways varied between them. This indicates that* ***cell line-specific factors—such as differences in chromatin accessibility, baseline gene expression, or mutational background—may influence the transcriptional response to HDACi****. Some variability may also stem from the* ***use of different HDAC inhibitors****, which may target different HDAC isoforms and thus modulate distinct regulatory networks. Disentangling thse two possibilities however will necessitate a broader set of experiments involving multiple cell lines and multiple HDACi with proper experimental design.”*

2. H3K27Ac profile for treated vs untreated should be checked for up and down regulated genes. This can verify whether change of H3K27Ac causes transcriptional change which leads to expression change or post transcriptional regulation like RNA stability causes the change of the expression.

Response: We completely agree and had considered this. Unfortunately, however the post-treatment H3K27Ac data is currently not available.

3. "(3) this trend was also true in gene body but only at mid and higher dosage" or " (5) RH4 trends at 1 µm dose of Entinostat most resembled the patterns at 75 nM dose of largazole in HCT116" - These statements do not match with the figures. Kindly explain these more clearly.

Response: We regret the lack of clarity. For point 3, we are talking about higher distribution of H3K27Ac epigenomic marks in gene body regions of the downregulated genes for HCT cell line with concentration of 75nM and 300nM. In the figure, in case of 75nM, gene body bin 1-3, 6 and 8-10 shows higher H3K27Ac read density concentration, where in remaining bins its nearly equal among up and downregulated genes. Likewise, in case of 300nM, majority of the gene body shows higher H3K27Ac distribution then promoter region except bin 1, and 10. In revised manuscript, we have now revised the statement for clarity.

4. The length of GB is different for every gene. The bin size of the genes having larger GB size is higher compared to genes having shorter GB. As a longer bin having more numbers of nucleotides gives the possibility of higher numbers of H3K27Ac count. So, authors should consider the length of GB into their analysis and should check whether the pattern persists after normalisation by the length of the GB of a gene.

Response: We appreciate this. However, we would like to clarify that **our analysis already incorporates normalization that accounts for these differences**.
Specifically, as mentioned in the Methods section, **we normalized the read counts within each genomic bin to TPM (Transcripts Per Million)** using the **Rsubread** package. TPM normalization adjusts for both sequencing depth and **bin length**, thereby mitigating the potential bias introduced by longer bins containing more nucleotides.
Thus, the enrichment patterns we observed are not confounded by gene length variation and reflect true biological signal distribution rather than technical artifacts. We have now made this point more explicit in the manuscript for clarity.

5. "As shown below, with few exceptions, for the first gene set, HCT116 genes showed higher concentration of H3K27Ac marks and for the second gene set, the opposite was true (Figure 6B)". Again, these statements are not matching with the figures rather the opposite may be true or there is no clear pattern to mention. Kindly recheck your results and explain it properly.

Response: We thank the reviewer for pointing out this mistake and we also regret the misinterpretation of the results. In the revised manuscript, we have clarified this for the better explanation. **Page 7; Lines 217:220.**
“As shown below, we do not observe any clear pattern for the H3K27Ac read density among the two cell lines, except at TSS and gene body 1 region, where we observed lower H3K27Ac read distribution for the upregulated genes **(Figure 6A&B)”**.

6. The reference of the figure 5A is not there in the text.
Response: Figure 5A is referenced on Page No. 7, Line No 201.

Competing Interests: No competing interests were disclosed.

Reviewer Report 22 February 2024

https://doi.org/10.5256/f1000research.159982.r239600

✗

Daiqing Liao
Department of Anatomy and Cell Biology, College of Medicine & UF Health Cancer Center, University of Florida, Gainesville, USA
Zhiguang Huo
Department of Biostatistics, University of Florida, Gainesville, Florida, USA

Cancer epigenome underpins oncogenic gene expression. Histone acetylation is a major epigenomic event that impacts the transcriptome. The dynamic action of histone acetyltransferases (HATs) and deacetylases (HDACs) shape the acetylome. Small molecule inhibitors of both HATs and HDACs are promising agents for cancer therapy. Several HDAC inhibitors (HDACi) have been approved for treating cutaneous T-cell lymphoma and peripheral T-cell lymphoma. HDACi profoundly perturbs the transcriptome[1]. Thousands of genes can be downregulated and upregulated after treatment with an HDACi [2,3]. Pretreatment epigenome could influence transcriptomic perturbation by HDACi, although this has not been thoroughly examined. The manuscript titled “Predicting gene expression changes upon epigenomic drug treatment” by Agrawal, Gopalan, and Hannenhalli assessed whether the pretreatment epigenomic profile of cells could predict gene expression alterations induced by HDACi. The authors used the levels of a single histone mark, namely H3K27ac, in divided bins (21 genomic bins in total) from 2 kb upstream of the transcription start site (TSS) to the end of gene body and data of differentially expressed genes (DEGs) as inputs to study the relationship between the pretreatment H3K27ac levels and the transcription outputs in several machine learning models. The authors found that the pretreatment H3K27ac patterns could predict the post-treatment up versus downregulated genes with reasonable accuracy. Essentially, gene expression levels due to HDACi treatment seem to be inversely correlated with the levels of H3K27ac before treatment, which is observed in two cell lines treated with two HDACi of different chemical classes (entinostat and largazole). While the results are interesting, several significant issues should be addressed.

1. The authors should also generate the feature importance score, e.g., the importance score for the random forest model (all gene models). This would be informative for readers to understand which genomic bin is most important to predict gene expression regulation.
2. The authors should show the p-value for each comparison pair in Figure 3.
3. The authors’ description “with few exceptions, for the first gene set, HCT116 genes showed higher concentration of H3K27Ac marks, and for the second gene set, the opposite was true (Figure 6B), consistent with the patterns in Figure 3” (page 9) seems contradictory to what is shown in Figure 6. This must be clarified.
4. In Figure 6, the z-scores of normalized read counts are used, while normalized read counts are shown in Figure 3. What is the rationale for using different parameters? Also, the p-value for each pair comparison should be shown in Figure 6.
5. Although the authors acknowledged that different epigenomic marks should be included in future studies, they are encouraged to analyze at least whether a different histone mark associated with active transcription, such as H3K4me3, would exhibit a similar predictive power as H3K27ac.
6. Another major limitation of the manuscript is that data based on only two cell lines and two HDACi are analyzed. An expanded study of data from a diverse set of samples (cells and tumors) treated with additional HDACi should improve the validity and rigor of the study.

References

1. Gryder BE, Wu L, Woldemichael GM, Pomella S, et al.: Chemical genomics reveals histone deacetylases are required for core regulatory transcription.*Nat Commun*. 2019; **10** (1): 3004 PubMed Abstract | Publisher Full Text

2. Lauffer BE, Mintzer R, Fong R, Mukund S, et al.: Histone deacetylase (HDAC) inhibitor kinetic rate constants correlate with cellular histone acetylation but not transcription and cell viability.*J Biol Chem*. 2013; **288** (37): 26926-43 PubMed Abstract | Publisher Full Text

3. Xiao Y, Hale S, Awasthee N, Meng C, et al.: HDAC3 and HDAC8 PROTAC dual degrader reveals roles of histone acetylation in gene regulation.*Cell Chem Biol*. 2023; **30** (11): 1421-1435.e12 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics (Zhiguang Huo) and epigenetics (Daiqing Liao)

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Version 1

Reviewer Report 20 September 2023

https://doi.org/10.5256/f1000research.153610.r203232

✕ **Angelika Merkel**

Bioinformatics Unit, Josep Carreras Leukemia Research Institute (IJC), Barcelona, Spain

Agrawal et al have presented a comparison of machine learning approaches to predict gene expression from the distribution of histone marks, specifically H3K27 acetylation.

Albeit interesting and potentially very useful for cancer specific and personalized treatment with histone deacetylase inhibitors as the authors suggest, the latter has actually not been addressed. The authors present a proof-of-concept study for predicting gene expression from histone marks but not do not address variations in tissues, individual or disease type.

As such, this is not a novel approach as there have been numerous studies before addressing the prediction of gene expression (e.g. Karlic et 2010[1], Singh et al 2016[2], Chen et al 2022[3]). The authors have failed to cite any of the current literature on the topic.

Methods and results are presented in a somewhat rough manner and need improvement:

- Abstract and introduction contain poor grammar and sentence structure ('genomic specificity" is not a new term to be defined, better to to use target specificity)
- Why were the HCT166 and RH4 cell lines chosen and what are these exactly (need description)?
- Was peak calling performed to eliminate background noise in the Chip-seq analysis?
- Was differential expression analysis performed (normalization, statistical analysis, etc.)? What is the significance of the log10fold change?
- The GO analysis is repeated in the methods and results section
- Figure 3: The scale among plots differ; gene bins are of variable size and larger promoter bins. Why have they been plotted them together?

References

1. Karlić R, Chung HR, Lasserre J, Vlahovicek K, et al.: Histone modification levels are predictive for gene expression.*Proc Natl Acad Sci U S A*. 2010; **107** (7): 2926-31 PubMed Abstract | Publisher Full Text
2. Singh R, Lanchantin J, Robins G, Qi Y: DeepChrome: deep-learning for predicting gene expression from histone modifications.*Bioinformatics*. 2016; **32** (17): i639-i648 PubMed Abstract | Publisher Full Text
3. Chen Y, Xie M, Wen J: Predicting gene expression from histone modifications with self-attention based neural networks and transfer learning.*Front Genet*. 2022; **13**: 1081842 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: transcriptomics, epigenetics, bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 09 Dec 2023

Piyush Agrawal

Agrawal et al have presented a comparison of machine learning approaches to predict gene expression from the distribution of histone marks, specifically H3K27 acetylation.

Albeit interesting and potentially very useful for cancer specific and personalized treatment with histone deacetylase inhibitors as the authors suggest, the latter has actually not been addressed. The authors present a proof-of-concept study for predicting gene expression from histone marks but not do not address variations in tissues, individual or disease type.

Response: We thank the reviewers for finding our work and interesting and potentially useful. Although we agree with reviewer regarding the importance of addressing variations in tissues, individual or disease type, unfortunately, lack of appropriate data currently prohibits such analysis. Besides the 2 cell lines used in this study, we didn’t come across any dataset (clinical/cell line) where for a given patient, both epigenomic marks and gene expression data is provided for both pre and post drug treatment. We would be happy to do the additional analysis if the reviewer can kindly point us to such data.

As such, this is not a novel approach as there have been numerous studies before addressing the prediction of gene expression (e.g. Karlic et 2010[1], Singh et al 2016[2], Chen et al 2022[3]). The authors have failed to cite any of the current literature on the topic.

Response: We thank the reviewer for pointing out these studies, however, we want to clarify that the above-mentioned studies address a different question than ours, which is “to predict gene expression from the epigenome in a specific context”. Whereas we are interested in predicting the effect on the expression upon drug treatment, given the epigenomic profile in the pre-treatment sample. Our question necessitates availability of pre- and post-treatment gene expression and pre-treatment epigenomic profile, while these previous approaches are not concerned with the changes upon drug treatment. In the revised manuscript, we have highlighted this difference in the Introduction.

Methods and results are presented in a somewhat rough manner and need improvement:

Abstract and introduction contain poor grammar and sentence structure ('genomic specificity" is not a new term to be defined, better to to use target specificity)

Response:

Why were the HCT166 and RH4 cell lines chosen and what are these exactly (need description)?

Response: We regret lack of clarity. HCT116 and RH4 cell lines were chosen because the data we required for our study was available only for these 2 cell lines, namely, pre-treatment epigenomic profile and pre- and post-treatment gene expression data. HCT116 is human colorectal carcinoma cell line initiated from an adult male whereas RH4 cell line is for studying alveolar rhabdomyosarcoma and is belongs to soft tissue lineage. In the revised manuscript, we have added the description of these cell lines in the Methods section.

Was peak calling performed to eliminate background noise in the Chip-seq analysis?

Response: No, peak calling wasn’t performed in the current study. The goal of our analysis was to compare changes in H3K27ac read density upon HDAC treatment. Peak calling represents a 0/1 binarization of a continuous variable (in this case, normalized read counts in promoters and gene bodies). Our model uses the epigenomic signal intensity in specific windows around the gene magnitude, making peak-calling unnecessary.

Was differential expression analysis performed (normalization, statistical analysis, etc.)? What is the significance of the log10fold change?

Response: The mean Log-TPM values of each gene was computed for both treated and untreated conditions, with the Log-FC computed as the difference between treated and untreated conditions. The top 1000 most up-regulated and 1000 most down-regulated genes were considered to be differentially expressed. We mistakenly mentioned Log10, which is now corrected to Log2. We have corrected this in the revised version**.**

The GO analysis is repeated in the methods and results section

Response: We want to clarify that in Methods Section, we mentioned the software and command only used for GO analysis, however, in Result section, we discussed the enriched terms associated with the up and downregulated genes.

Figure 3: The scale among plots differ; gene bins are of variable size and larger promoter bins. Why have they been plotted them together?

Response: We agree with the reviewer that scale among plots differ and gene bins are of variable size. The only reason for plotting them together was to analyze all the 21 features together in single image i.e. change in distribution of the pre-treatment epigenomic marks in up and down genes post treatment.

Is the work clearly and accurately presented and does it cite the current literature?
No

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

References

1. Karlić R, Chung HR, Lasserre J, Vlahovicek K, et al.: Histone modification levels are predictive for gene expression.*Proc Natl Acad Sci U S A*. 2010; **107** (7): 2926-31 PubMed Abstract | Publisher Full Text
2. Singh R, Lanchantin J, Robins G, Qi Y: DeepChrome: deep-learning for predicting gene expression from histone modifications.*Bioinformatics*. 2016; **32** (17): i639-i648 PubMed Abstract | Publisher Full Text
3. Chen Y, Xie M, Wen J: Predicting gene expression from histone modifications with self-attention based neural networks and transfer learning.*Front Genet*. 2022; **13**: 1081842 PubMed Abstract | Publisher Full Text

Competing Interests: No competing interests