Nucleic Acids Research, 2025, **53**, gkae1212
https://doi.org/10.1093/nar/gkae1212
Advance access publication date: 11 December 2024
Gene regulation, Chromatin and Epigenetics

Predicting gene expression from histone marks using chromatin deep learning models depends on histone mark function, regulatory distance and cellular states

Alan E. Murphy [1,2,*], Aydan Askarova [1,2], Boris Lenhard [3], Nathan G. Skene [1,2] and Sarah J. Marzi [2,4,5,*]

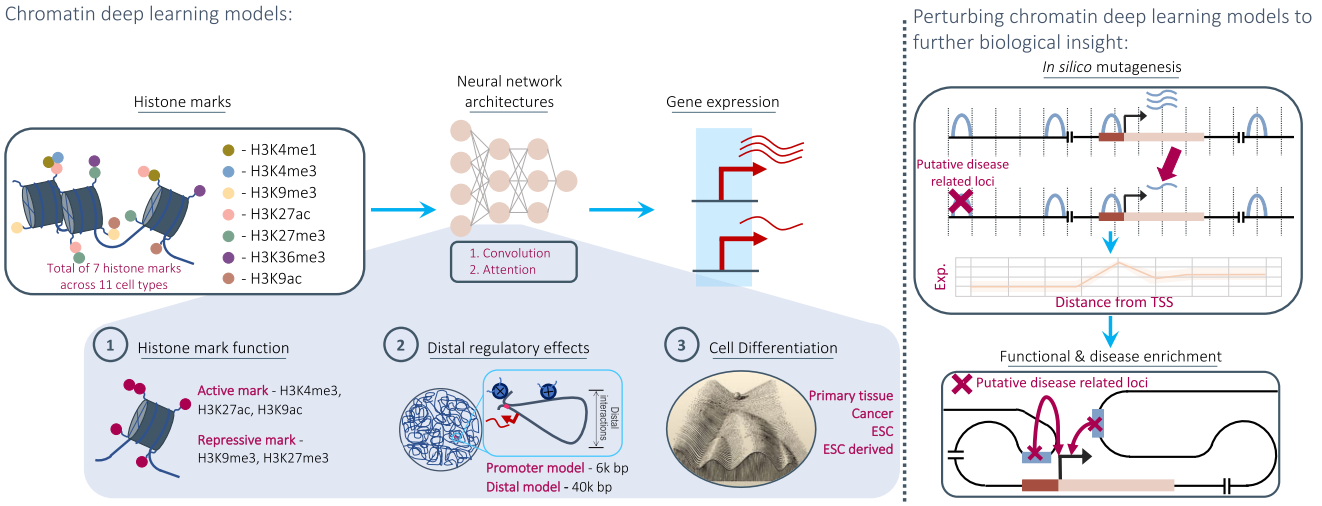
[1]UK Dementia Research Institute at Imperial College London, 86 Wood Lane, London W12 0BZ, UK
[2]Department of Brain Sciences, Imperial College London, 86 Wood Lane, London W12 0BZ, UK
[3]MRC London Institute of Medical Sciences, Imperial College London, Du Cane Road, London W12 0HS, UK
[4]UK Dementia Research Institute at King's College London, 338 Euston Road, London SE5 9RT, UK
[5]Department of Basic and Clinical Neuroscience, Institute of Psychiatry Psychology & Neuroscience, King's College London, 16 De Crespigny Park, London SE5 9RT, UK

*To whom correspondence should be addressed. Tel: +44 20 7594 5863; Email: a.murphy@imperial.ac.uk
Correspondence may also be addressed to Sarah J. Marzi. Tel: Email: sarah.marzi@kcl.ac.uk

Abstract

To understand the complex relationship between histone mark activity and gene expression, recent advances have used *in silico* predictions based on large-scale machine learning models. However, these approaches have omitted key contributing factors like cell state, histone mark function or distal effects, which impact the relationship, limiting their findings. Moreover, downstream use of these models for new biological insight is lacking. Here, we present the most comprehensive study of this relationship to date – investigating seven histone marks in eleven cell types across a diverse range of cell states. We used convolutional and attention-based models to predict transcription from histone mark activity at promoters and distal regulatory elements. Our work shows that histone mark function, genomic distance and cellular states collectively influence a histone mark's relationship with transcription. We found that no individual histone mark is consistently the strongest predictor of gene expression across all genomic and cellular contexts. This highlights the need to consider all three factors when determining the effect of histone mark activity on transcriptional state. Furthermore, we conducted *in silico* histone mark perturbation assays, uncovering functional and disease related loci and highlighting frameworks for the use of chromatin deep learning models to uncover new biological insight.

Graphical abstract



Introduction

Post-translational modifications on the N-terminal tails of histone proteins, known as histone marks, form a key epigenetic mechanism by which eukaryotic cells regulate transcriptional activity via altering chromatin structure and interacting with other transcriptional regulators (1,2). These epigenetic modifications enable cell plasticity without changes to the underlying DNA sequence. Histone mark dynamics in a given cell

Received: March 29, 2024. **Revised:** October 12, 2024. **Editorial Decision:** November 15, 2024. **Accepted:** December 9, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

are mediated by both internal and extracellular cues (3,4). Alterations in histone modifications have been found to strongly associate with cellular differentiation, cell cycle stages and the development of different diseases (5–8). For example, as cells mature and differentiate, chromatin accessibility and histone acetylation become progressively restricted throughout their lineage (9).

Individual regulatory effects of histone marks on transcription have been widely studied: while H3K9ac is associated with active promoter regions (10), H3K4me1 is found at distal enhancers (11). However, less emphasis has been placed on the extent to which these histone marks directly regulate gene expression levels. To investigate if transcriptional levels in different cellular contexts can be determined solely from histone modification states, one could correlate levels of histone modifications in regulatory elements with messenger RNA (mRNA) expression individually. However, this ignores two additional levels of complexity: firstly, it is well-known that histone modifications interact with other epigenetic factors such as pioneer transcription factors, which in turn have been linked to enhancer activation (9); secondly, histone marks themselves act in concert and the interaction between different regulatory elements is not necessarily additive. To circumvent these challenges it is possible to conduct *in silico* experiments, predicting transcriptional levels from histone mark signals where the observed chromatin state is assumed to capture the other contributing epigenetic regulators without directly experimentally observing them.

This method has been applied in previous work, to investigate which histone mark is most predictive of gene expression. For example, addressing this over 10 years ago, Karlić *et al.* (12) used a linear regression model but only considered histone mark levels at promoter regions, and only tested the effect in a single cell type, CD4 + T-cells. González-Ramírez and colleagues (13) identified predictive histone marks at promoters and other regulatory regions, leveraging chromatin interaction data from a Hi-C assay to link enhancers to target genes. However, this study also only considered one cell type, mouse embryonic stem cells (ESCs). Moreover, there is some circularity in the selection of training regions based on derived (from their histone mark data) promoter and enhancer locations and the model's input measuring the same histone mark levels. Finally, Wang *et al.* (14), investigated the relationship between histone marks and transcription but inverted the problem, predicting histone mark levels from transcription. For transcription, they used Pro-seq and GRO-seq, which labels RNA as it is being transcribed, avoiding issues with RNA degradation (14,15). The authors used a Support Vector Regression model but again only investigated relationships in the K562 cell line.

Here, we expand on previous research by considering multiple cell types, histone marks and regulatory distances. We investigated the effects of seven histone marks on gene expression (Table 1) in eleven human cell or tissue types from the Roadmap Epigenomics Consortium (16). We will refer to these as cell types hereafter, but note that they also contain tissue samples and cell lines. We used two neural network architectures to predict gene expression: a simple convolutional neural network considering genes promoter regions, and a recently published transformer-based, DNA interaction-aware deep learning architecture called Chromoformer (17) (see associated publication, figure 1 for architecture). Chromoformer was originally trained to predict expression using seven histone marks. Here we adapt and retrain the model to predict based on single histone marks, and pairwise combinations of histone marks, to investigate their effect on transcription in isolation. Our work highlights how histone mark function, cellular differentiation and genomic distance to regulatory elements all collectively influence the relationship between histone modification levels and gene expression. We find that there is no universal histone mark which is consistently the most predictive of expression. We recommend that researchers consider all three of these influencing factors when determining the effect of histone mark levels on the transcriptional state of a cell in their work.

In the related field of genomic deep learning, models predict expression or epigenetic marks from DNA sequence, such as Enformer (18), Borzoi (19) and Sei (20). There has been a recent shift away from arbitrarily benchmarking performance, to prioritizing the use of these models to make new biological discoveries (21–23). This is still lacking for models linking histone mark levels to expression. We aim to address this by outlining a framework to use these models to identify the cell type-specific functional and disease related genomic loci, leading to new biological insights and expanding the utility of these models beyond simple investigation into which histone marks are the most predictive of gene expression.

Materials and methods

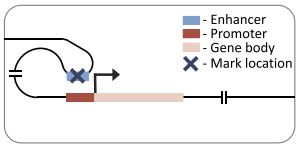
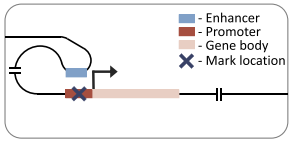
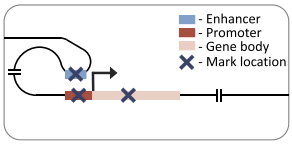
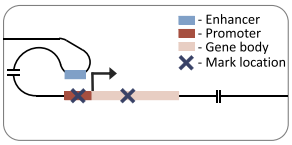
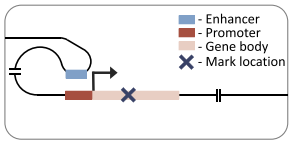
Data collection and processing

The data for our analysis was derived from the Roadmap Epigenomics Consortium (16) and follows the same preprocessing pipeline used by Chromoformer (17). We used a subset of eleven cell types from Roadmap, for which gene expression, histone mark and 3D chromatin interactions profiles were available (Supplementary Table S1). Specifically we included data from H1 ESCs (E003), H1 BMP4-derived mesendoderm (E004), H1 BMP4-derived trophoblast (E005), H1-derived mesenchymal stem cells (E006), H1-derived neuronal progenitor cultured cells (E007), HUES64 ESCs (E016), Liver (E066), Pancreatic islets (E087), A549 EtOH 0.02pct lung carcinoma (E114), GM12878 lymphoblastoid (E116) and HepG2 hepatocellular carcinoma (E118). TagAlign-formatted, ChIP-seq read alignments for seven histone marks – H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K27ac and H3K9ac were used. For consistency, the data was subsampled to 30 million reads and reads themselves were truncated to 36 base-pairs, reducing possible read length biases. The alignments were sorted and indexed using Sambamba (40) v0.6 and read depths for each base-pair position were derived along the hg19 reference genome using Bedtools (41) v2.23. Note that our goal was to predict the transcriptional signal in the same experimental conditions; same experiment, same histone mark(s), same cell type while holding out subsets of chromosomes. Thus, the experimental conditions were the same for the training and test sets and we did not compare across experiments to avoid any issues with experimental variance in signal/noise ratios or sequencing depths.

Both the promoter and distal models used the averaged $\log_2$-transformed 100 base-pair binned signal with our distal model also averaging at 500 and 2000 base-pairs to also use as model input features. Using three different resolutions of the histone mark signal in the distal model is intended to represent prior knowledge that epigenetic regulation operates

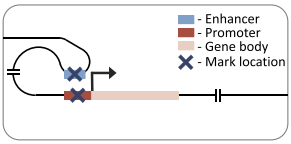
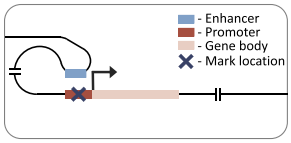
Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

Table 1. Information on the primary genomic location, transcriptional relationship and proposed function for the seven histone marks used to predict expression

Histone mark	Genomic location	Transcriptional relationship	Proposed function	References
H3K4me1		Activating	• Enriched at active and poised enhancers. • Suggested to fine-tune, rather than tightly control, enhancer activity and function by recruiting key transcription factors.	(11,24)
H3K4me3		Activating	• Found at promoter regions • Has a direct preferential association with the Plant HomeoDomain (PHD) finger of nucleosome remodelling factor complex which remodels chromatin, making the DNA accessible for gene transcription.	(25,26)
H3K9me3		Repressive	• Involved in the formation of heterochromatin, • Found at transposable elements, satellite repeats and genes, where it ensures transcriptional silencing. • These heterochromatin has also been found to relate to cell lineage-dependent, transcriptional silencing.	(27–29)
H3K27me3		Repressive	• Act as silencers in promoters and gene bodies that regulate gene expression via proximity or looping. • Function has been linked to Polycomb repressive complexes (PCR1, PCR2) which can be recruited and contribute to chromatin compaction.	(30–33)
H3K36me3		Repressive	• Enriched in gene bodies. • A binding partner for histone deacetylases (HDACs) which prevent run-away RNA polymerase II (Pol II) transcription.	(34–36)

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

Table 1. Continued

Histone mark	Genomic location	Transcriptional relationship	Proposed function	References
H3K27ac		Activating	• Enriched at active enhancer and promoter regions (differing from H3K4me1 which also indicates poised enhancers). • Recruits transcription factors to increase transcription. For example, bromodomain-containing protein 4 (BRD4) which enhances Pol II recruitment and increases transcription.	(37,38)
H3K9ac		Activating	• Enriched at promoter regions. • Mediates super elongation complex and pol II chromatin occupancy on the proximal promoter region thus aiding in the switch from transcription initiation to elongation.	(10,39)

on differing scales and has been shown to improve performance of other models (42). Promoter-capture Hi-C 3D chromatin interaction data (43) was incorporated into the distal model and mapped to the Roadmap cell types using the same approach as previously described for Chromoformer (17). Reads per kilobase of transcript, per million mapped reads (RPKM)-normalized gene expression levels from protein coding genes were downloaded from Roadmap for the matching cell types. Both the promoter and distal models predict the log2-transformed RPKM [$\log_2$(RPKM + 1)] gene expression levels and account for gene strandedness. The transcriptional start site (TSS) was identified using RefSeq annotations (release v210) (44) for each gene. Our final dataset included a total of 18 955 genes.

Promoter model

Our promoter model is a custom convolutional neural network, similar in architecture to DeepChrome (45). The model takes in a symmetrical 6000 base-pair genomic window averaged at 100 base-pair bins, centred on the TSS of the gene of interest. This 6000 base-pair window was used as a lenient cut-off to include all relevant local regions. The model architecture is composed of three standard convolutional blocks. These blocks each consist of a 1D convolutional layer, batch normalization, rectified linear unit (ReLU) activation, max-pooling and dropout. This was followed by two fully connected blocks, which have dropout (in the first block), a dense layer, ReLU activation and a final output layer with linear activation. The convolutional blocks and their sliding window converted the histone mark signal into a position-wise representation highlighting genomic loci that correlate with expression. The fully connected blocks scaled down the size of the representation, to finally produce a single score representing the gene's RPKM. The size of each layer is provided in our github repository (https://github.com/neurogenomics/chromexpress) (46).

Distal model

Our distal model architecture was based on the Chromoformer model (17). This is an attention-based model which uses cell type-specific promoter capture Hi-C data to identify interacting regions in a 40 000 base-pair genomic window centred on the TSS. This approach captures the histone mark signal both at the TSS and at putative *cis*-regulatory regions. The model has three independent modules at different resolutions (100, 500 and 2000 base-pairs), producing a multi-scale representation of the histone mark landscape. Each module goes through a transformer block before being combined and passed through a full-connected block with ReLU activation and a final output layer with linear activation. The architecture of the model is discussed in more detail in the original publication (17).

Model training

The same model training approach was used for both the promoter and distal model. Model training and evaluation was based on a 4-fold cross-validation regime to give a stronger estimate of model performance. The total 18 955 genes were split into four sets, 5045, 4751, 4605 and 4554 respectively, with genes from the same chromosome in the same split to avoid data leakage (47). For every fold, one set was used as the blind test set, while the other three sets were used for model training and validation. Performance on the test set for each fold was measured with Pearson's correlation coefficient. A separate model was trained for each histone mark, cell type and cross-validation fold combination.

The models were trained using the ADAM (48) optimizer with default parameters with a batch size of 64 over a max-

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

imum of 100 epochs. An early stopping regularization was implemented based on the model's validation loss with a patience of twelve epochs. The initial learning rate was set at 0.001 and decayed by a factor of 0.2 when the loss did not improve after a patience of three epochs. Mean squared error was used as the loss function.

Histone mark levels

Histone mark occupancy was measured separately for our promoter and distal models and for each cell type, histone mark and cross-validation fold. It was measured as the average $\log_2$-transformed, 100 base-pair binned read depth of the histone mark signal. For the promoter model, this signal was taken from the 6000 base-pairs around the TSS of each gene and for the distal model, from the full 40 000 base-pairs.

Gene expression state

Histone mark occupancy was measured for both active and inactive genes. A gene was defined as active or inactive based on whether its expression level was above or below the median for that cell type. This calculation was performed for each cell type independently. This approach was first implemented in DeepChrome (45) and has been frequently used in the literature (17,45,49,50).

Correlation analysis

We measured agreement in model performance for both our promoter and distal models by matching the cross-validation fold and cell type for each model trained on a (or pair of) different histone mark(s). The Pearson correlation coefficient was used to quantify agreement. The performance across different cell types and folds for each mark (or pairs of marks) was aggregated to get the reported mean Pearson R and standard deviation. When comparing to the distal model based on all histone marks, the mean performance for each cell type – after aggregating based on the different folds – was used to report mean Pearson R and standard deviation.

In silico perturbation of histone mark activity levels

In silico histone mark perturbation was performed on the distal model trained on a single mark. Although we have trained Chromoformer with multiple histone marks as input, we chose to use it trained on a single mark for the perturbation analysis since perturbing one mark will likely affect the histone mark occupancy of other marks in the same region which would not be possible to accurately account for in the model.

We selected one active and one repressive mark which are found at both promoter and distal regulatory elements – H3K27ac and H3K27me3. Perturbation experiments were carried out on active genes for the active mark model and inactive genes for the inactive mark model (see the 'Gene expression state' section in the 'Materials and methods' section), to measure the effect on expression of reducing the levels of the histone mark. The predictions from the different k-fold versions of the model were averaged, similar to the approach commonly used in sequence to expression models (19,51–53). For the promoter histone signal, the full 6000 base pairs around the TSS were perturbed, whereas for the distal histone signal, bins of 2000 base pairs across the 40 000 base pair receptive field were perturbed iteratively (similar to the approach for DNA sequence perturbation used by the genomic deep learning model CRÈME (54). The implemented perturbation levels were between 0 and 1 inclusively in 0.1 steps. The code to perform the *in silico* histone mark perturbation is available on our github repository (https://github.com/neurogenomics/chromexpress) (46).

As well as averaging the predictions from the different k-fold versions, we also tested the correlation between the different folds to ensure the model is learning consistent regulatory code. Moreover, we benchmarked this concordance against Borzoi (19), a genomic deep learning model with the current largest receptive field of 524 000 base pairs. Here, for a fair comparison, we only tested Borzoi's concordance in the centre 40 000 base-pairs to match Chromoformers receptive field, for the RNA predictions in the same cell types as those used in our analysis and added up to 500 random genetic variants to the sequences of 1000 genes to match our perturbation test.

In silico perturbation enrichment in quantitative trait loci studies

The averaged *in silico* perturbation experiments on the active (H3K27ac) model for each cell type were filtered to those >6000 base-pairs upstream and 4000 base-pairs downstream of the TSS to concentrate on distal, cell type-specific regulatory regions as opposed to promoter regions. To control for the greater effect of changes in gene bodies (Figure 6E), which have a median length of ∼25 000 base pairs (55) – longer than the downstream receptive field of the model – these predictions were sorted based on their predicted change in expression and split into deciles separately for upstream and downstream regions.

The fine-mapped expression quantitative trait loci (eQTL) data based on the UK Biobank population was sourced from Wang *et al.*, 2021 (56). Causal single nucleotide polymorphisms (SNPs) were identified from those in linkage disequilibrium (LD) using FINEMAP v1.3.1 (57) and SuSiE v0.8.1 (58). The resulting fine-mapped SNPs were filtered to those with a SuSiE causal probability [posterior inclusion probability (PIP)] > 0.9 in the tissue of interest and with a PIP < 0.1 in other tissues to get just the high confidence, tissue-specific fine-mapped SNPs. The ROADMAP cell types were matched to five of the tissues used in eQTL study where the tissue assayed were identical across the two (available on our github repository: https://github.com/neurogenomics/chromexpress) (46).

To test for enrichment of the fine-mapped, tissue-specific SNPs, a bootstrap sampling experiment was implemented where the proportion of SNPs found in each decile were compared against 10 000 randomly sampled regions from all deciles. *P*-values were derived and adjusted using false discovery rate (FDR) correction for multiple testing.

Since the distal model uses Hi-C chromatin interaction data as input, we also ran this eQTL enrichment test on the matched cell type and gene, significant promoter capture Hi-C interactions (2000 base-pair resolution) to compare against the model's enrichment performance. To match the model's tested regions, the chromatin interaction data was filtered to just those upstream of the gene. Furthermore, we benchmarked against the regions of maximum histone mark activity for each gene up to 20 000 base-pairs upstream and downstream of the TSS, averaged at 2000 base-pairs to match the Hi-C and model approach, and also against a proximal loci

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

of the 6000 base-pairs just upstream or downstream of the TSS. Scripts detailing the approach are available on our github repository (https://github.com/neurogenomics/chromexpress) (46).

In silico perturbation disease enrichment

To test for disease enrichment, the top decile of the same averaged *in silico* perturbation experiments on the active model, filtered to just those upstream or downstream of the TSS, were considered. For this analysis, predictions in the liver and neuronal progenitor cells (NPCs) were used due to their potential respective relationships with liver and neuronal diseases. A third group of regions comprising the top decile across all cell types was included to look for cell type-consistent disease enrichment.

Summary statistics for genome-wide association studies (GWAS) for liver diseases – non-alcoholic fatty liver disease (NAFLD) (59) and hepatitis (60), glial diseases – Parkinson's (61) and Alzheimer's (62) and neuronal diseases – amyotrophic lateral sclerosis (63), schizophrenia (64), autism spectrum disorder (65) and bipolar disorder (66) were downloaded from the IEU GWAS portal (67) and the BioStudies database (68) and were uniformly processed with MungeSumstats v1.11.3 (69) (default settings, converting build to hg19 where necessary and saving in 'LDSC' format).

We applied stratified LD score regression (s-LDSC) (70) v1.0.1 (https://github.com/bulik/ldsc) to test for disease enrichment. Specifically, annotation files for each of the three groups of genomic loci were first created with Phase 3 of the 1000 genomes reference. Followed by the generation of LD scores with a window size of 1 centiMorgan (cM) i.e. ∼1 million base pairs, filtering to HapMap3 SNPs to match the baseline model. Finally, the enrichment analysis was run for the GWAS summary statistics across the three groups as well as those in the baseline model whilst excluding the major histocompatibility complex (due to the known difficulties predicting LD in this region) (70).

Results

Active histone marks prove most informative at promoter regions

We first focused on the performance of histone mark levels from the promoter regions of the gene of interest using our promoter model (Figure 1). As a naïve, baseline performance measure we compared performance against absolute correlation of histone mark and gene expression. Each histone mark was averaged at different local distances around the transcriptional start site (1500, 3000 and 6000 base-pairs) to capture the best possible promoter correlation. Our promoter model outperformed these baselines for all histone marks, which was largely expected due to neural network's ability to capture non-linear relationships, with notably better performance in regressive marks. Moreover, the lenient inclusion of 3000 base-pairs up and down stream of the TSS in the promoter model led to improved performance over a smaller region more focused on the putative gene promoter functional regions, i.e. 2500 base-pairs upstream and 500 base-pairs downstream of the TSS (Supplementary Figure S1).

Overall, we found H3K4me3, a mark located in active promoter regions (25), to be the best performing from our promoter model (Figure 1A). Moreover, active promoter marks H3K4me3, H3K27ac (37) and H3K9ac (10) made up the three top performing marks, all with a Pearson's correlation above 0.73. The fourth best performing histone mark, H3K4me1, is found at active enhancers (11). It likely performed worse than the other active marks due to the limited range of the model, which only took into consideration a gene's promoter region.

Repressive marks proved less informative with H3K9me3 (27), H3K27me3 (30) and H3K36me3 (71) making up the three worst performing marks. Importantly, there was high variability in performance across the histone marks with a correlation difference of 0.25 between the best active and worst repressive mark (range of Pearson correlation coefficients: 0.52–0.76).

The predictive performance of active and repressive marks differ based on cell state at promoter regions

Splitting the models' performance across the different cell types highlighted histone mark groups with similar variations in their scores across cell types (Figure 1B). This is most notable for active promoter marks (H3K4me3, H3K9ac and H3K27ac). To formally evaluate this trend, we calculated the correlation between the models' predictions across the different genes, cell types, histone marks and cross-validation folds (Figure 2A). One distinct group of highly correlated histone marks were apparent (highlighted in blue in Figure 2A), corresponding to the active histone marks previously observed. Interestingly, H3K9me3 showed the lowest correlation with the other histone marks, including with H3K36me3 and H3K27me3, the other repressive marks. The samples collected for Roadmap (16) can be classified into those taken from ESCs, cells differentiated from ESCs, adult bulk tissues and cancer cell lines (Supplementary Table S1). Active histone mark activity levels were significantly more predictive of expression in ESC than primary tissues whereas the opposite was noted for repressive histone mark activity levels which was more predictive in primary tissues than ESC (Figure 2B). The multi-modal performance, visible in Figure 2B, is the result of a combination of the histone mark and cell type being tested (Supplementary Figure S2). These results indicated that the most accurate method by which to predict gene expression from the promoter region depends on the extent to which the cell type of interest has differentiated – active marks like H3K4me3, H3K9ac and H3K27ac were most predictive for cells at earlier stages of their differentiation process, including ESCs. In contrast, repressive marks, like H3K9me3, fared better in fully differentiated tissues or cells.

Higher histone mark levels result in better predictive performance at promoter regions

We first queried whether the promoter model's performance was just a result of the quantity of histone mark signal observed, rather than learnt histone mark and cell type-specific regulatory relationships. H3K4me3 did follow this relationship with the highest levels of activity (Supplementary Figure S3A), best performance and positive correlation between them (Supplementary Figure S3B and C). However, on other marks, including H3K27ac, the model showed consistent performance irrespective of levels of training or test activity (Supplementary Figure S3B and C). Moreover, H3K27me3 displayed a negative relationship, where more activity led to

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

Figure 1. The performance of the promoter model shows the best gene expression predictions based on active histone marks H3K4me3, H3K27ac and H3K9ac. (**A**) Promoter model performance (pale blue) measured by the Pearson correlation coefficient on the blind test sets across each histone mark. Also shown is the absolute correlation of each histone mark with expression at different regions around the transcriptional start site (1500, 3000 and 6000 base-pairs) to act as a baseline performance measure. The whiskers represent the standard deviation across the different cell types and the 4-fold cross-validation. (**B**) Promoter model performance split by different cell types from Roadmap.

worse performance. If simply more histone mark activity led to better performance overall, a strong correlation would have been observed for all histone marks.

To investigate what was driving the difference in performance and how this related to active and repressive histone marks in the different cell states, we measured the average histone mark levels in the promoter regions for highly and lowly expressed genes (Figure 3, see the 'Materials and methods' section for details). We observed a strong correlation between model performance and histone mark levels for active histone marks in highly expressed genes, and conversely, for repressive marks in lowly expressed genes, indicating that the model learns the functional significance of the different histone marks (Figure 3A). For highly expressed genes (Figure 3B), repressive marks, H3K9me3, H3K36me3 and H3K27me3, had higher histone mark levels in ESCs than primary tissues (although for H3K9me3 this was not significant after multiple test correction). Conversely, active marks, H3K4me3, H3K9ac and H3K27ac, showed varying activity across primary tissues and ESCs. For lowly expressed genes (Figure 3C), histone mark levels tended to be higher in ESCs than in primary tissues across histone marks. Overall, our analysis highlighted that higher histone mark levels in a gene where the expression status matched the function of the histone mark (active versus repressive), led to better performance of the model. RNA-seq typically has a bias for 3′ ends, which might lead to underestimation of abundance of some long transcripts (72,73). To account for this, we also tested whether there was a bias in our model's predictions based on transcript lengths (Supplementary Figure S4). There appeared to be no relationship, suggesting our model's performance is not biased against genes with long transcripts.

Active marks are most predictive in the distal model

To consider histone mark levels outside of the genes' promoter regions, we next tested a model architecture with a much larger receptive field (up to 40 000 base-pairs around the TSS). We used Chromoformer (17), a transformer-based architecture, which accounts for distal histone mark levels,

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

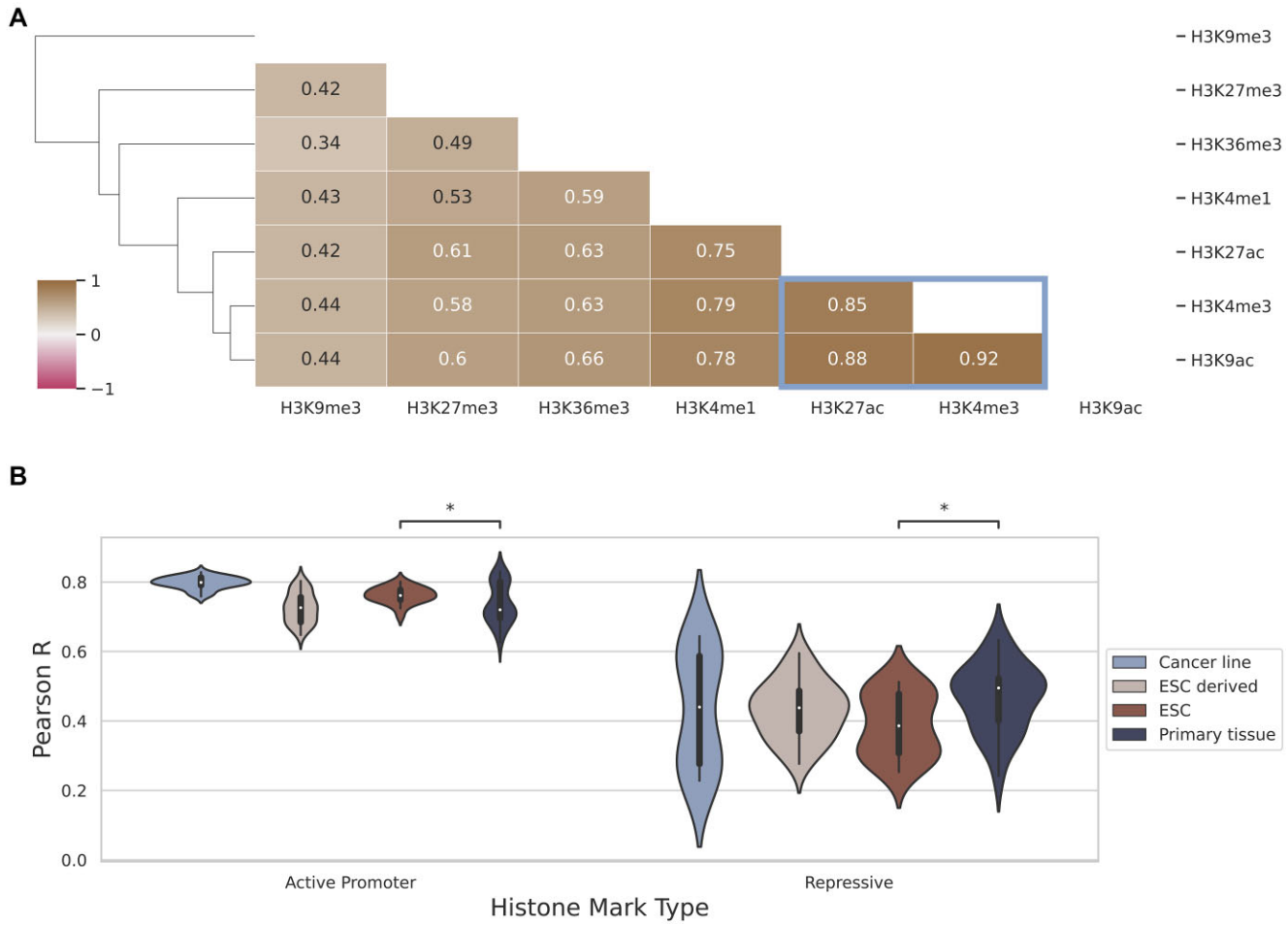


Figure 2. The predictive performance of the promoter model across samples clusters by histone mark function. (**A**) Correlation matrix of the promoter model's predictions by the different histone marks across each gene, cell type and cross-validation fold. H3K4me3, H3K9ac and H3K27ac were characterized by high positive pairwise correlations (highlighted in blue). Bars along the y-axis show the hierarchical clustering dendrogram. (**B**) Violin plot of the model's performance on active and repressive histone marks measured by their Pearson correlation coefficient on the blind test sets. The cell type performance is grouped by the cell state — ESC, ESC-derived cell, adult primary tissue or cancer cell line. A Wilcoxon rank-sum test was used to compare ESC and primary tissue performance for active and repressive marks. Significance was based on a FDR correction for multiple testing, with *P*-value indicator: * <0.05.

weighting important genomic regions using cell type-specific DNA interaction data. We trained this distal model on each single histone mark and benchmarked the performance across histone marks, and also against the performance when trained on all seven histone marks combined (Figure 4A). Again, we found H3K4me3, H3K9ac and H3K27ac to be the top three performing marks with very little difference in overall performance between them (mean Pearson R of 0.762, 0.757, and 0.749, respectively). All three histone modifications are active marks while only H3K27ac (37) is found at both promoters and distal regulatory regions (enhancers).

Receptive field expansion and multi-histone mark predictions yield diminishing returns

The performance of incorporating distal histone mark levels in a model consistently but marginally increased the Pearson correlation coefficients. The range of improvement in correlation varied from 0.01 to 0.15, despite the substantial increase in receptive field and model architecture complexity (Figure 4C). Compared to the promoter model, although the same histone mark performance ranking was observed, marks which are known to affect genomic locations outside of promoter regions showed the highest relative improvement, specifically H3K36me3, H3K27me3 and H3K27ac (Figure 4C and Table 1).

The relationship between histone mark type and cell state found for the promoter model, where active histone marks were more predictive in ESC and repressive in primary tissues, was similarly observed for the distal model (Figure 4B and Supplementary Figure S5A and B). Moreover, we investigated the histone mark levels across the full 40 000 base-pair receptive field of the distal model and found the same trend as for the promoter model where the model picks up on known biology of histone mark prevalence: We observed strong correlations between model performance and histone mark levels for active histone marks around highly expressed genes, and conversely, for repressive marks around lowly expressed genes (Supplementary Figure S5C).

To further interrogate the contributions of histone marks to the prediction of expression in our distal model, we benchmarked performance across pairs of histone marks with the top three performing marks, H3K4me3, H3K9ac and H3K27ac (Figure 5A). All combinations with the top three

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

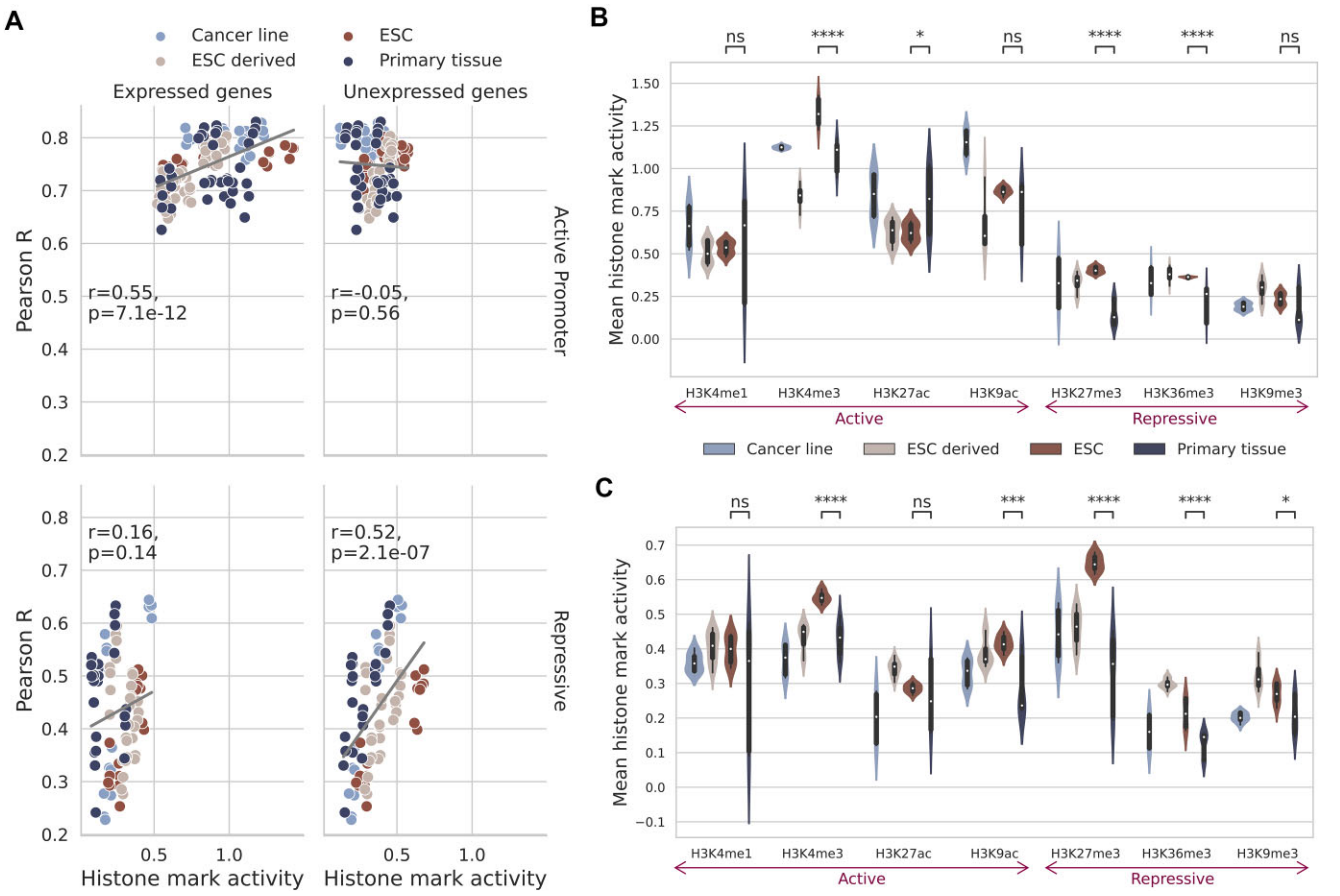


Figure 3. Higher histone mark levels are associated with better predictive performance in specific gene expression states. (**A**) Correlation between the histone mark levels and model performance for each cell type, k-fold combination. This is split by repressive and active marks, as well as highly and lowly expressed genes. The histone mark activity is firstly averaged across the receptive field of the model then it is averaged separately for high/low expression genes for every cell for each of the 4-fold versions of the model. (**B**) Violin plot of histone mark levels measured by the average $\log_2$-transformed, read depth in the promoter region (6000 base-pairs around the transcription start site) for highly expressed genes. The cell types are grouped by the cell state – ESC, ESC-derived cell, adult primary tissue or cancer cell line and averaged at the level of cell type and k-fold. (**C**) Average histone mark levels for lowly expressed genes. Significance was based on the FDR multiple test correction, with *P*-value indicators: **** <1e-4, *** <1e-3, ** <1e-2, * <0.05, ns ≥ 0.05.

histone marks were better performing than any of the top three marks by themselves. However, this improvement was relatively small (<0.04 mean increase in Pearson R for the best histone mark combination) and was only significant for a handful of combinations.

This result highlights that incorporating additional histone marks in the prediction led to a consistent, albeit small, boost in performance, regardless of the histone mark type (active or repressive). This became most evident when we combined pairs of the top three performing marks: all three are active marks and two are confined to promoter regions, but their combination still resulted in improved performance over the individual marks (Figure 5B). Notably though, combinations including H3K27ac, an enhancer mark, showed greater improvement over combinations including only promoter marks. Importantly, combining the top three marks with another mark showed a mean improvement of 0.043, compared to a mean improvement of 0.062 when using all marks (red dashed line in Figure 5A). This means that adding an additional five histone marks to the distal model would increase performance (Pearson R) by another 0.02, highlighting the diminishing returns of including additional histone mark information and giving insight for researchers looking to best capture gene regulation in their cell type of interest with minimal experimental work in the future.

Our analysis further highlighted the combinatorial predictive capabilities of H3K36me3 (Figure 5A and C). H3K36me3 is a repressive mark (74) with strong distal effects on gene expression. This mark showed the largest improvement from the promoter to the distal model (Figure 4C) and when combined with the top three scoring histone marks, was its best performing pair, even when compared to combinations of the top three performing marks (Figure 5A) and far improved performance compared to the addition of the other marks (Figure 5C). Conversely, when paired with the repressive mark H3K27me3, H3K36me3 did not result in the best performing pair (Supplementary Figure S6) but did still improve performance on the individual mark. This highlights the complementary information H3K36me3 provides in addition to the top three performing active promoter and enhancer marks.

We also investigated whether the performance increase for the top three marks when coupled with H3K36me3 was driven by bivalent genes. Bivalent genes are characterized by the presence of both repressive and active histone mark signals at their promoter and are known to silence develop-

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

Figure 4. Active histone modifications performed best at predicting gene expression in the distal model. (**A**) Distal model performance was measured by the Pearson correlation coefficient on the blind test sets for each histone mark. The whiskers represent the standard deviation across the different cell types and the 4-fold cross-validation. The red dashed line and shaded box shows the model's mean performance and standard deviation when trained on all seven histone marks together. (**B**) Predictive performance is shown split by the different cell types in Roadmap. (**C**) The improvement in performance for each histone mark by expanding the receptive field outside of the promoter region with the distal model. The largest increase in performance is observed for H3K36me3. The whiskers indicate the standard deviation across cell types and folds.

mental genes in ESCs while keeping them poised for activation (75). However, the model performance in bivalent genes for ESCs did not notably improve over non-bivalent genes (Supplementary Figure S7).

In silico histone mark perturbation prioritises functionally relevant genomic loci and disease relevant cell types

Up to this point, our work has shown the performance of chromatin to expression deep learning models and how they encode known biological relationships between histone mark levels and expression. However, none of this highlights new information about gene regulation or disease. *In silico* perturbation enables the investigation of the effect on gene expression of varying histone mark levels in a cell type and gene-specific manner that would be impractical to undertake experimentally *in vivo*. This analysis was inspired by the recent development of systematic histone mark editing toolkits to measure the relationship with expression (76) and follows a comparable approach of epigenomic editing but *in silico*. Our *in silico* perturbation experiments enable the systematic, high throughput quantification and comparison of the effect of histone mark activity at different genomic loci on gene expression.

We first investigated the effect of *in silico* perturbation experiments at an aggregate level, varying levels of histone mark activity, as well as distances from the TSS using the distal model trained on one active (H3K27ac) and one repressive (H3K27me3) mark (Figure 6). We used models trained on single marks to avoid issues where perturbing one type of histone mark will affect another mark's activity in the region. Here, we permuted either the entire TSS or the distal regions in 2000 base pair bins (see the '*In silico* perturbation of histone mark activity levels' section in the 'Materials and methods' section). Furthermore, we averaged predictions across the four k-fold model versions, a standard approach *in silico* mutagenesis experiments for genomic deep learning models

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

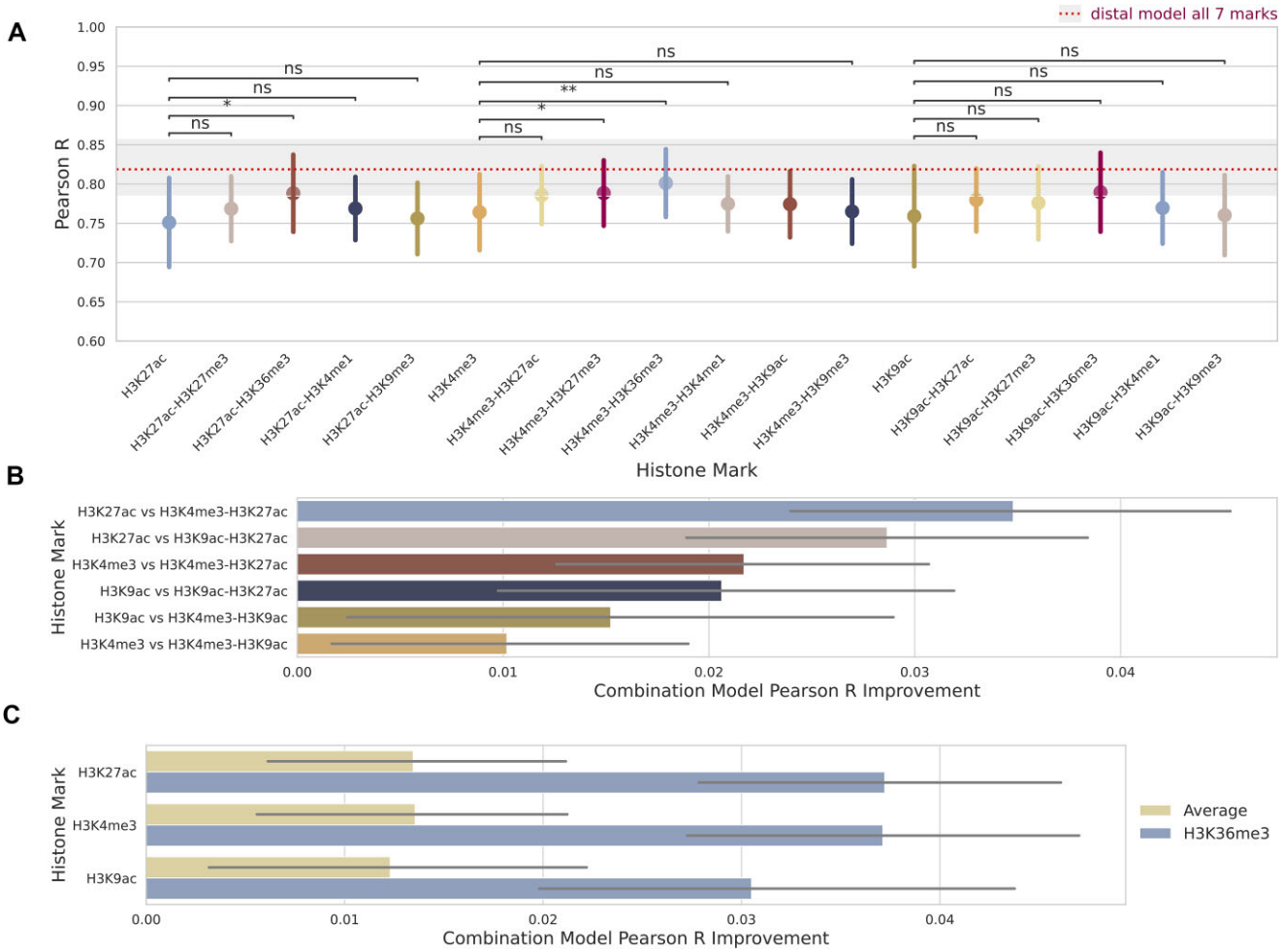


Figure 5. Pairwise combinatorial performance of distal model in predicting gene expression based on two histone marks. (**A**) Paired distal model performance was measured by the Pearson correlation coefficient on the blind test sets across combinations of histone marks. The whiskers represent the standard deviation across the different cell types and the 4-fold cross-validation. Data is averaged at the level of cell type and k-fold. The red dashed line and shaded box shows the model's mean performance and standard deviation when trained on all seven histone marks together. Significance based on FDR multiple test correction, with *P*-value indicators: ** <1e-2, * <0.05, ns ≥ 0.05. (**B**) Performance improvement for combinations of the top three performing marks from the single histone mark distal model. (**C**) Performance improvement over the single histone mark distal model for combinations using H3K36me3 versus the average of the other marks.

(19,51,53). This step may not have been required given the notably high correlation between the different models' predictions (Supplementary Figure S8A), which was on par with, if not slightly better, than that of the genomic deep learning model Borzoi (19) (Supplementary Figure S8B). This indicates that models trained on histone mark levels show a similar ability to learn consistent regulatory code across differing training sets compared to that of their genomic counterparts.

For the active model, we noted a clear relationship between reducing histone mark levels and large predicted decreases in expression at the promoter (Figure 6A) while we found only minor decreases at distal regulatory regions (Figure 6B). This matches the known functional relationship between promoter and enhancer activity with expression and also the *in silico* perturbations of DNA sequence in genomic deep learning models (51). However, reducing repressive histone mark levels showed little relationship with increased expression (Figure 6C abd D) which may be due to the repressive mark's relatively worse performance overall (Figure 4A). Moreover, it highlights how the lack of a repressive mark is not sufficient to confer expression to a gene, but rather it is additionally associated with the presence of an active mark which we seen through the performance of the repressive mark H3K36me3 in combination active marks (Figure 5C). The relationship between perturbation and distance and their effect on expression is more apparent when we view the predicted quantile change in expression by distance from the TSS while removing histone mark activity completely (Figure 6E and F). Here, we saw the highest predicted change near the TSS, reducing as distance to the TSS increases for both the active and repressive model. Interestingly, this reduction was not symmetrical upstream and downstream of the TSS, with downstream loci having a greater effect on expression on average. We believe this was a result of the length of the gene body [median length ~25 000 base pairs (55)], which would incorporate the entire downstream receptive field of the model for most genes and thus lend to greater importance for RNA-seq predictions rather than assays of promoters like CAGE-Seq (77) or transcription initiation like PRO-seq (78). Our analysis highlighted that on average, perturbations to histone mark signals in the gene body had a greater predicted effect than distal regulatory regions upstream. These results held consistently

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

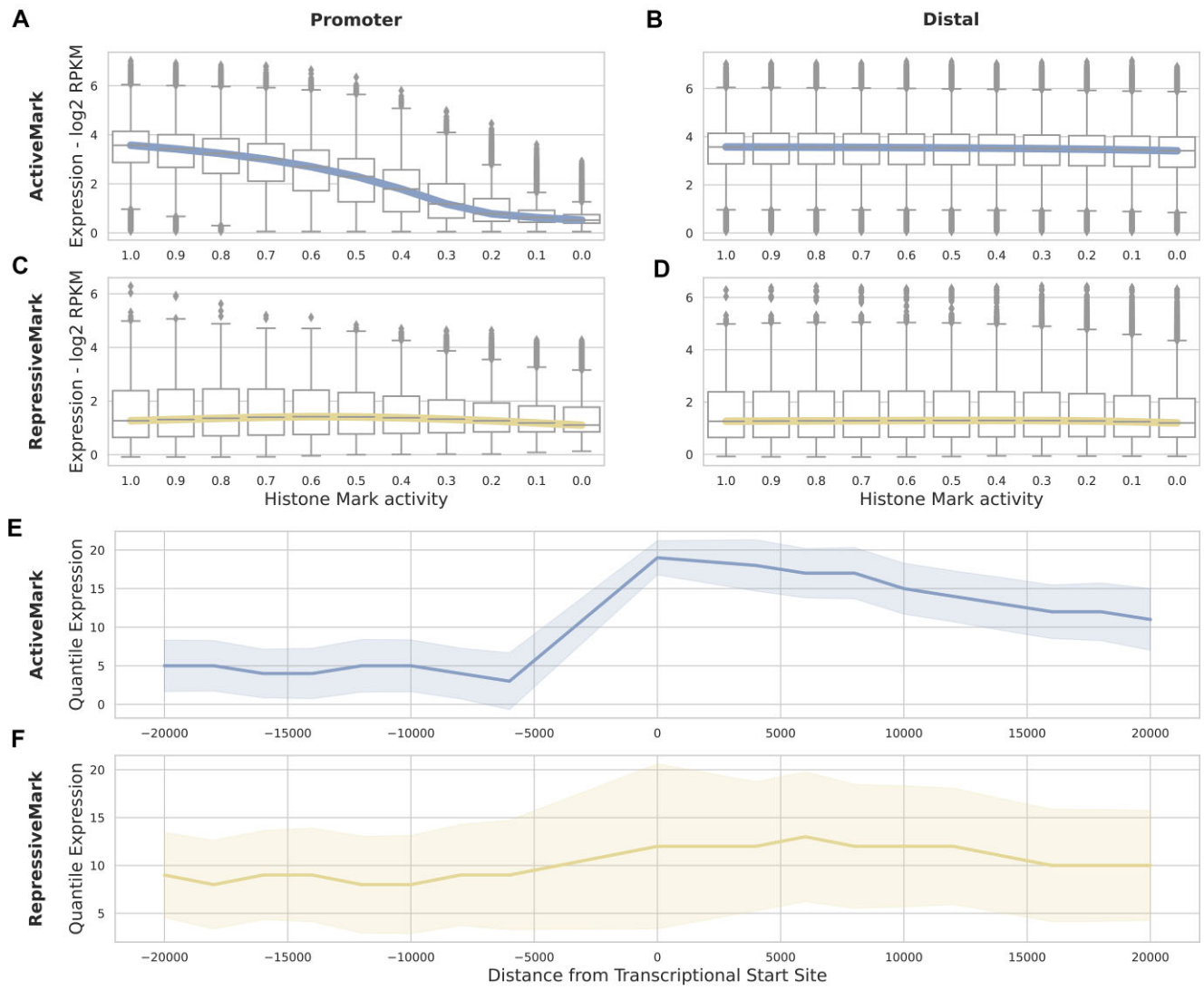


Figure 6. Effect of *in silico* histone mark perturbation on expression. The effect on predicted expression (y-axis) of changing the proportional levels of measured histone mark activity (x-axis) for all cell types and genes, averaged across the four k-fold models. The distal model trained on a single histone mark was used to measure the effects of a perturbed active mark – H3K27ac (**A,B**) – or a repressive mark – H3K27me3 (**C,D**) – in 2000 base pair bins for distal or at the promoter. (**E,F**) The effect of distance on expression change is shown when the histone mark activity is completely removed at a specific location for the active (**E**) or repressive (**F**) mark. The distribution of all gene expression changes in all cell types split into 20 quantile bins.

across all cell types and cell states for the active marks. However, this was not the case for repressive marks, which differed vastly by cell state (Supplementary Figure S9). We intentionally chose an active (H3K27ac) and repressive (H3K27me3) mark for analysis that are known to regulate expression outside of the TSS. We confirmed that a mark that is primarily associated with the TSS (H3K4me3) has higher predicted change in expression in the TSS and lower change downstream and in the gene body (Supplementary Figure S10).

We next considered whether these *in silico* perturbation experiments could be used to gain insight into the cell type-specific regulatory function in gene expression and disease, using genetic variants to test for functional and disease enrichment (Figure 7). Given that the active model captured known biological relationships in the *in silico* perturbation, we focused on this model's perturbation experiments. Moreover, we only considered upstream predictions to capture distal regulatory regions which vary across cell types as opposed to promoter and gene body signals.

To test the model's ability to capture functional loci we used a large scale, tissue-specific, fine-mapped eQTL set based on the UK Biobank population (79). First, we split the loci into deciles sorted based on the model's predicted change in expression (decile 10 having the largest predicted effect). We implemented a bootstrap sampling test to compare each decile to randomly sampled upstream and downstream loci for enrichment of the fine-mapped eQTLs (see the '*In silico* perturbation enrichment in quantitative trait loci studies' section in the 'Materials and methods' section). We found significant enrichment for the top decile for all but one cell types tested (Figure 7A), indicating that the model correctly predicted the loci which contributed most to the cell type-specific gene expression. Furthermore, to provide benchmarks against alternative loci prioritization methods, we tested how much the distal model improves upon the (i) regions of highest histone mark activity, (ii) Hi-C chromatin interacting loci, and (iii) proximal loci [≤6000 base-pairs up or down from the TSS, inspired by Wang *et al.*'s distance baselines

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

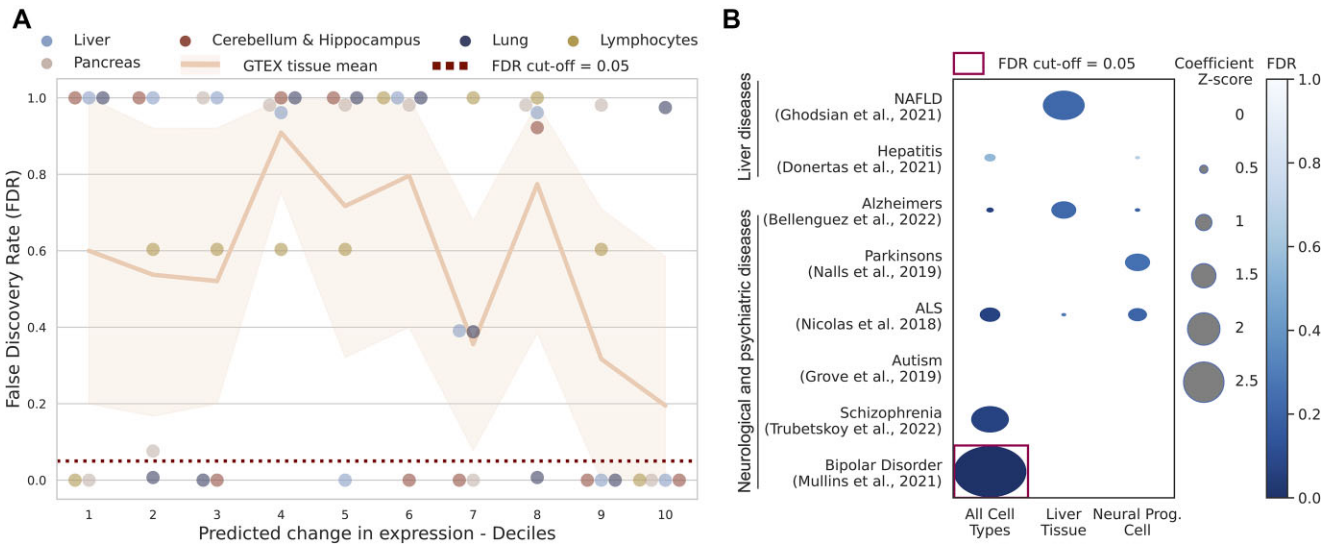


Figure 7. *In silico* histone mark perturbation experiment highlights functional and disease enrichment. (**A**) Upstream and downstream *in silico* histone mark perturbation experiments from the active model were sorted into deciles based on their predicted change in expression (x-axis). Each decile was tested for enrichment of fine-mapped eQTL interactions in matched cell types and compared against bootstrap sampling random upstream loci 10 000 times to generate *P*-values of enrichment (y-axis). (**B**) FDR *P*-value enrichment scores (colour) and coefficient Z-score (size) for the top decile (based on their predicted change in expression) derived from all cell types (non-cell type-specific), liver tissue or NPCs (x-axis). Enrichment tests were conducted with s-LDSC and genetic variants relating to different diseases (y-axis). FDR adjustment was applied for all GWAS tested.

(80)] for fine-mapped eQTL enrichment in these same loci (Supplementary Figure S11). With the exception of the lung tissue sample, the model's loci prioritization approach outperformed the region of maximum histone mark activity and Hi-C in all tissues (Supplementary Figure S11A and B) and matched the performance of the ≤6000 base-pairs approach (Supplementary Figure S11C).

We next tested whether these loci also harbour known disease related genetic variants. We used s-LDSC (70) with GWAS for eight different liver and brain diseases (59–66). Test regions were based on the top cell type-specific decile with the largest predicted effect for liver and NPCs, as well as the averaged top decile across all cell types of upstream and downstream loci (Figure 7B). S-LDSC measures enrichment of disease genetic variants accounting for the obscuring nature of LD (70). Here, we used the neuronal cells and liver tissue for their predicted relationship with brain and liver diseases, respectively. We used the top decile for all cell types to look for non-cell type-specific disease enrichment. While none of the expected cell type-specific significant enrichments were detected, the NAFLD GWAS was tending towards significance in liver tissue. Moreover, one significant association for bipolar disorder was detected after multiple testing correction for the non-cell type-specific disease enrichment, highlighting the functional importance to disease of regions where changes in the histone mark activity levels was predicted to have a large effect on expression across cell types.

Results of the *in silico* perturbation analysis can be visualized and investigated locally by approximating the effect histone mark activity at each genomic locus has on the predicted gene expression by calculating the partial derivatives of the model with respect to the input, i.e. the gradient on the input (81). Using this approach, we uncovered a regulatory region upstream of *DSTN* in the lung carcinoma cell line (Figure 8). This locus was in the top decile of predicted effects on expression for this cell type from the *in silico* perturbation analysis. The region is shown to have a small effect across the three input resolutions our distal model uses prior to perturbing the H3K27ac signal, but once removed, has a large effect. This locus contains the fine-mapped SNP rs611572 from the UK Biobank population eQTL analysis in a matched cell type (79) and *DSTN* has been noted to promote malignancy in lung adenocarcinoma. It has potential as a prognostic marker and therapeutic target (82), confirming its predicted functional importance based on our findings. Of note, we see a similar relationship when visualizing the embedding attention matrix, where the weights for this region go to zero after perturbation, indicating the effect of the lack of histone mark signal (Supplementary Figure S12).

Discussion

We present the most comprehensive deep learning study of the relationship between histone mark levels and transcription to date. We considered multiple cell types and histone marks at differing receptive fields and demonstrated how the prediction of gene expression is dependent on three key contributing factors – histone mark function, regulatory distance and cellular states.

For our analysis, we used the Roadmap (16) data repository, benefiting from the standardized experimental approach. We investigated the genome-wide activity of seven histone marks across eleven cell types. For the histone mark ChIP-Seq read alignments, we subsampled to 30 million reads to enable a fair comparison across marks. While for gene expression, we utilized RPKM values which measures the mRNA abundance of transcripts normalized by gene length, avoiding any potential within sample bias. Since model predictions were made in the same cell type as training, there was no need to standardize gene expression levels across cells (83). The use of mRNA abundance here is one limitation of the study as it may not be reflective of true cellular regulation.

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

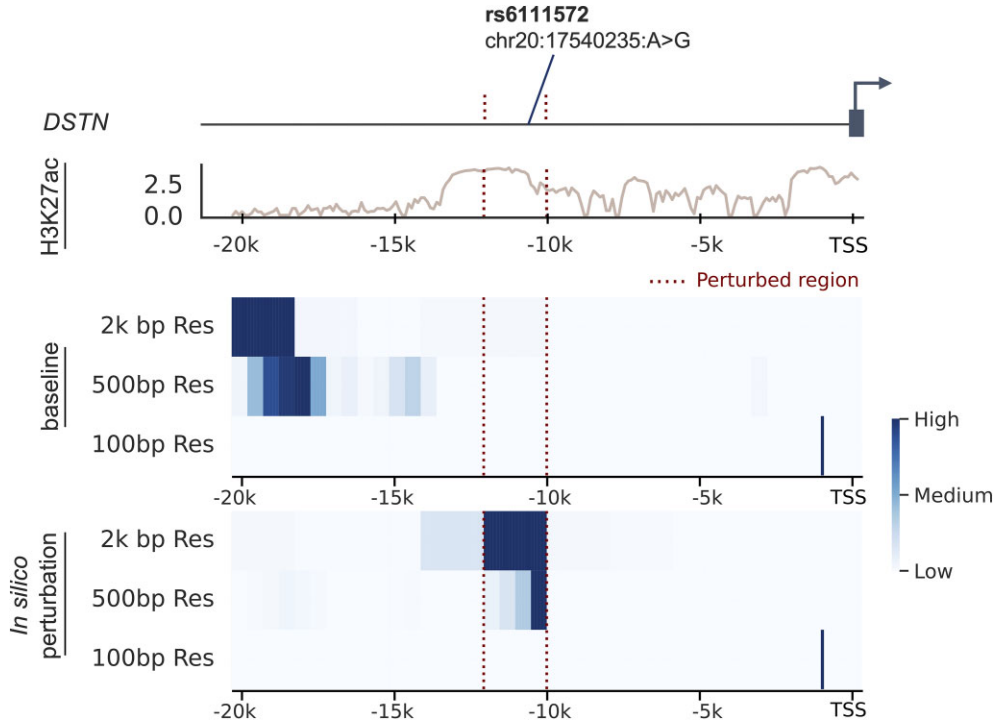


Figure 8. *In silico* histone mark perturbation visualized for DSTN in lung carcinoma cell line. Upstream *in silico* histone mark perturbation of *DSTN* in the lung carcinoma cell line using the distal model trained on the H3K27ac signal. The experimental H3K27ac signal is shown along with an approximation of the effect each region of histone mark activity has on the prediction, calculated using the partial derivatives of the model with respect to the input, before and after ablation of the H3K27ac signal in the noted region. The fine-mapped SNP rs611572 from the UK Biobank population eQTL analysis in a matched cell type is also shown (79). This region was chosen as the *in silico* perturbation analysis identified it in the top decile of predicted effects on expression for this cell type.

RNA-seq read counts are impacted by mRNA turnover, stability and nuclear export, not solely deriving from transcriptional changes (72,73). Thus, it is possible that the noted limited correlation between transcription and repressive histone marks is a result of this (Figures 1A and 4A). For instance, particularly stable mRNAs might obscure the effects of transcriptional repression, potentially leading to underestimates of the actual repressive contribution of histone marks at these specific loci. Future work may benefit from investigating measures of transcription which avoid these issues, for example, by profiling transcription initiation by assays such as precision nuclear run-on sequencing (PRO-seq) (78).

To ensure robust benchmarking results, we repeated training of both our promoter and distal models across a 4-fold cross-validation, ensuring the test set genes were grouped by chromosomes to avoid any data leakage (47) – where data in the training set is related to data in the test set, artificially inflating model performance. To run each cell type, histone mark or combination of histone marks, for each cross-validation fold using both the promoter and distal models, was a computationally intensive task. This resulted in 1276 training and prediction iterations which were all run using an A100 80 GB GPU. To avoid overfitting over such a substantial number of iterations, we automated hyperparameter tuning for both models using a learning rate decay and early stopping regime, holding out an independent validation set of genes for monitoring.

Both our promoter and distal models were developed as quantitative, regression models, predicting a gene's $\log_2$ RPKMs, which has been shown to yield better generalization and interpretability than binary classification models (84). For the promoter model, following the same approach as past benchmarking work (84), we implemented an intentionally simple convolutional neural network architecture based a relatively small receptive field around the TSS to compare histone mark performance. On the other hand, our distal model, Chromoformer (17), was a transformer-based architecture accounting for distal histone mark levels in a weighted manner based on DNA interaction data. One limitation is the receptive field of our distal model which extends 40 000 base-pairs around the TSS. Although this is a large window and computationally intensive to include in such a model, it is a fraction of the known distance at which DNA interactions can occur. For example, Hi-C experiments capture *cis*-interactions up to 1 million base-pairs away (85).

The results of our promoter region analysis showed that the active marks H3K4me3 and H3K9ac were the most predictive of gene expression (Figure 1). However, their optimal performance was dependent on the cell state, performing better in ESCs whereas repressive marks like H3K9me3 performed relatively better in adult primary tissues (Figure 2B). We concluded that the stage of cell differentiation was the driving factor of performance for active and repressive marks: active marks performed better at early stages of lineage commitment, i.e. ESCs, whereas repressive marks were more predictive in fully differentiated cells, i.e differentiated tissue. This relationship is in line with the distinctive ESC chromatin landscape, which features more accessible DNA and less heterochromatin and correspondingly, less repressive and more active histone mark activity when compared with differentiated cells (86).

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

The effect of repressive mark performance across different cell states is further substantiated by our *in silico* permutation analysis (Supplementary Figure S9). We observed a consistent relationship for active marks, where perturbations to histone mark signals in the gene body had a greater predicted effect than distal regulatory regions. This pattern was only replicated for inactive marks in differentiated cell types. We hypothesize that this relates to the lack of repressive signal in the other cell states (86).

Furthermore, we noted that for active genes, the relationship between a histone mark's levels and performance replicated known biology: Observing a strong correlation for active histone marks in highly expressed genes, and conversely for repressive marks in lowly expressed genes (Figure 3A). This highlighted that higher histone mark levels were beneficial for the model, leading to greater predictive performance in the correct context. In relation to histone mark levels at the TSS, we also noted that for inactive genes, histone mark activity is reduced with cell differentiation (Figure 3B). The concept that cell lineage commitment leads to globally lower histone mark levels has been previously noted (87).

For our distal model, active marks H3K4me3, H3K9ac and H3K27ac were the best performing (Figure 4A). Interestingly, these three marks, two of which are linked to the TSS of genes they regulate, outperformed H3K4me1 which is associated with distal enhancers (11). One possible explanation for this was investigated in a recent study (51) which found that, when predicting expression from DNA sequence, a similar attention model prioritized sequences at the promoter region over distal regulatory regions. The reason being that given the multiple choice of enhancer and other regulatory regions and their relatively small influence on gene expression, a model will prioritize the information at the promoter region. We believe the same effect could have contributed to our results whereby the active promoter mark information contributed to expression to a greater degree than distal regulatory regions. This same relationship was clearly notable in our *in silico* perturbation experiments (Figure 6E and F). Moreover, given that H3K4me1 is indicative of poised rather than active enhancers (11), it would presumably be less predictive of gene activity.

A key point of our findings is the marginal return in performances by: (i) Extending from an intentionally simple local promoter model to an attention based, computationally complex model which accounts for distal histone mark levels (Figure 4C), and (ii) Increasing the number of histone marks included in the model (Figure 5B). We noted that understanding the cell state (undifferentiated or fully differentiated) and the gene state (highly or lowly expressed) of interest and choosing the most appropriate mark for these had a greater impact on performance than the number or receptive field of histone mark levels considered.

Comparing performance across the promoter and distal models, H3K27ac showed the biggest gain of the top three performing marks (Figure 4). This was expected given its relationship with active promoters and enhancer regions (37). However, its performance based solely in promoter regions was still relatively strong, which was reassuring given the mark's prevalence in complex disease research.

We also trained our distal model on pairs of histone marks, showing that the added performance of incorporating additional histone marks diminishes markedly after this point. This result could benefit researchers wishing to capture transcription in a cell type of interest from limited histone mark information. The paired analysis also highlighted the strong combinatorial performance of H3K36me3. This repressive mark was the best performing choice as a pair with any of our three top marks (Figure 5A) and was the fourth best performing mark of the single histone mark analysis (Figure 4A). H3K36me3 is a canonical mark of transcription, serving as a binding partner for HDACs which prevent run-away transcription of RNA polymerase II (35). H3K36me3 is generally enriched in gene bodies of mRNAs (34), outside of the TSS, which may explain its relatively poor performance in the promoter model and why the improvement with the distal model for this mark was the highest of any mark tested (Figure 4C).

Finally, we performed an array of *in silico* histone mark level perturbation experiments, showing the relationship between distance from the TSS and a regulatory region's effect on gene expression (Figure 6). Our analysis highlighted the very high correlation for the *in silico* perturbations between the different cross-validation fold models (Supplementary Figure S8). A possible advantage of histone mark deep learning models over genomic deep learning models trained on DNA sequence is that they offer an alternative to making single base pair level changes such as genetic variants which are notoriously difficult to interpret, particularly as the genomic window considered by the model increases (19,53,54,87). For example, Sasse *et al.* found that Enformer, a long-range genomic deep learning model, did a poor job of predicting the effect of inserting multiple single nucleotide variants based on personalized genomes, even predicting the wrong direction of effect up to 40% of the time (53). This work highlighted that the current training paradigm for genomic deep learning models of training across the genome for variability is insufficient to accurately capture the effect of genetic variants. Moreover, these variants are subject to LD and as of yet, it has not been proven that genomic deep learning models can accurately differentiate between causal genetic variants and those in LD. By identifying genomic loci of interest based on perturbing histone mark levels, our model captures significant enrichment of eQTLs in the most predictive regions, offering an alternative to genomic deep learning models trained on DNA. Additionally, these results based on epigenetic data further underlines the functional importance of these eQTLs (Figure 7A).

Furthermore, using these identified genomic loci, we developed a framework by which such models can be applied to test for both functional and disease enrichment in a cell type-specific manner (Figure 7). The results for the disease enrichment did not recapitulate projected relationships for the NPCs or liver, which could be a result of the imperfect cell type matching, the limited overlap between the disease related genetic variants and the relatively small window of upstream genomic loci considered, or the observed differences in the measured genetic effects on gene expression versus complex traits (88). This highlights that further work on such models is needed, hopefully increasing the receptive field and using known affected cell types, to capture an association with complex diseases. Importantly, a substantial overlap and large genomic coverage of the loci considered are key recommendations for s-LDSC analyses (70). This issue when capturing complex phenotypic enrichment is not unique to these models and is also a challenge with genomic deep learning models as highlighted recently (52). Non-cell type specific genomic loci predictive of gene expression were enriched for GWAS sig-

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

nal for bipolar disorder, indicating the importance of cell type consistent regulatory regions in complex disease. Importantly, this analysis highlights the significance of these predicted genomic loci not only in functional genomics studies (Figure 7A) but also in disease (Figure 7B). Past approaches such as ICE-BERG, a pipeline that uses CUT&RUN replicates to create a combined profile of binding events for H3K4me3, have been previously used to uncover functionally relevant regulatory events (89). However, our approach shows, for the first time, how chromatin deep learning models can be perturbed to uncover genome-wide and cell type-specific functionally and disease relevant regulatory regions.

Overall, our study shows that multiple factors influence the performance of histone marks when predicting gene expression, something which had not explicitly been considered by previous work (12–14). Our findings suggest that if one wishes to investigate the TSS of genes, promoter-specific active marks H3K4me3 and H3K9ac are the best options. Beyond the promoter region, active marks H3K4me3, H3K9ac and H3K27ac are most predictive, especially in combination with the transcriptionally associated mark H3K36me3. However, it is worth considering the cell state (differentiated or early stages of lineage commitment) and the state of the genes or interest (highly or lowly expressed) to have a better understanding of the optimally predictive histone marks. Importantly, more effort should be placed on using these models to uncover new biological insights, particularly for phenotypic and disease-based studies. Similar to genomic deep learning models, chromatin deep learning models are capable of capturing functionally relevant genomic loci.

Data availability

The Histone mark ChIP-seq read alignments, RPKM gene expression profiles were downloaded from the Roadmap Epigenomics Web Portal (https://egg2.wustl.edu/roadmap/web_portal/index.html). The promoter capture Hi-C experiments were obtained from the 3DIV database (available at http://3div.kr/), specifically the tissue mnemonics H1, ME, TB, MSC, NPC, LI11, PA, LG and GM. The UK Biobank fine-mapped eQTL data were downloaded from the supplementary material of Wang *et al*.'s study (56). The summary statistics were downloaded from the IEU GWAS portal (67) (IDs: ieu-b-7, ebi-a-GCST90027158, ebi-a-GCST90027158, ebi-a-GCST005647, ebi-a-GCST90091033, ebi-a-GCST90091033, ebi-a-GCST90038627, ieu-b-5099, ieu-a-1185, ieu-b-5110) and for hepatitis, from the BioStudies database (68) (ID: S-BSST407). All reference datasets used to run s-LDSC (70) were downloaded following the links from the source material: https://github.com/bulik/ldsc. The model architectures and all training and analysis scripts, along with scripts to download and complete all pre-processing steps on the training data [sourced from Roadmap (16) and largely replicated from Chromoformer's scripts (17)] are available at https://github.com/neurogenomics/chromexpress and https://doi.org/10.5281/zenodo.10940542 (46). The model results and weights are available for download on figshare (https://figshare.com/account/home#/projects/201105).

Supplementary data

Supplementary Data are available at NAR Online.

Funding

UK Dementia Research Institute [UKDRI-5008 and UKDRI-6009]; Medical Research Council. UK; Edmond and Lily Safra Early Career Fellowship Program (to S.J.M.); UKRI Future Leaders Fellowship [MR/T04327X/1 to N.G.S.]. Funding for open access charge: UK Dementia Research Institute [UKDRI-5008 and UKDRI-6009].

Conflict of interest statement

None declared.

References

1. Miller,J.L. and Grant,P.A. (2013) The role of DNA methylation and histone modifications in transcriptional regulation in humans. *Subcell. Biochem.*, **61**, 289–317.
2. Bannister,A.J. and Kouzarides,T. (2011) Regulation of chromatin by histone modifications. *Cell Res.*, **21**, 381–395.
3. McBrian,M.A., Behbahan,I.S., Ferrari,R., Su,T., Huang,T.-W., Li,K., Hong,C.S., Christofk,H.R., Vogelauer,M., Seligson,D.B., *et al.* (2013) Histone acetylation regulates intracellular pH. *Mol. Cell*, **49**, 310–321.
4. Niu,Y., DesMarais,T.L., Tong,Z., Yao,Y. and Costa,M. (2015) Oxidative stress alters global histone modification and DNA methylation. *Free Radical Biol. Med.*, **82**, 22–28.
5. Jenuwein,T. and Allis,C.D. (2001) Translating the histone code. *Science*, **293**, 1074–1080.
6. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R., *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
7. Marzi,S.J., Leung,S.K., Ribarska,T., Hannon,E., Smith,A.R., Pishva,E., Poschmann,J., Moore,K., Troakes,C., Al-Sarraj,S., *et al.* (2018) A histone acetylome-wide association study of Alzheimer's disease identifies disease-associated H3K27ac differences in the entorhinal cortex. *Nat. Neurosci.*, **21**, 1618–1627.
8. Zhao,Z. and Shilatifard,A. (2019) Epigenetic modifications of histones in cancer. *Genome Biol.*, **20**, 245.
9. Atlasi,Y. and Stunnenberg,H.G. (2017) The interplay of epigenetic marks during stem cell differentiation and development. *Nat. Rev. Genet.*, **18**, 643–658.
10. Karmodiya,K., Krebs,A.R., Oulad-Abdelghani,M., Kimura,H. and Tora,L. (2012) H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics*, **13**, 424.
11. Bae,S. and Lesch,B.J. (2020) H3K4me1 distribution predicts transcription state and poising at promoters. *Front. Cell Dev. Biol.*, **8**, 289.
12. Karlić,R., Chung,H.-R., Lasserre,J., Vlahoviček,K. and Vingron,M. (2010) Histone modification levels are predictive for gene expression. *Proc. Natl Acad. Sci. U.S.A.*, **107**, 2926–2931.
13. González-Ramírez,M., Ballaré,C., Mugianesi,F., Beringer,M., Santanach,A., Blanco,E. and Croce,L.D. (2021) Differential contribution to gene expression prediction of histone modifications at enhancers or promoters. *PLoS Comput. Biol.*, **17**, e1009368.
14. 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.* (2022) Prediction of histone post-translational modification patterns based on nascent transcription data. *Nat. Genet.*, **54**, 295–305.
15. Core,L.J., Waterfall,J.J. and Lis,J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.
16. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J.,

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

Ziller,M.J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
17. Lee,D., Yang,J. and Kim,S. (2022) Learning the histone codes with large genomic windows and three-dimensional chromatin interactions using transformer. *Nat. Commun.*, **13**, 6678.
18. 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. (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, **18**, 1196–1203.
19. Linder,J., Srivastava,D., Yuan,H., Agarwal,V. and Kelley,D.R. (2023) Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. bioRxiv doi: https://doi.org/10.1101/2023.08.30.555582, 01 September 2023, preprint: not peer reviewed.
20. Chen,K.M., Wong,A.K., Troyanskaya,O.G. and Zhou,J. (2022) A sequence-based global map of regulatory activity for deciphering human genetics. *Nat. Genet.*, **54**, 940–949.
21. Avsec,Ž., Weilert,M., Shrikumar,A., Krueger,S., Alexandari,A., Dalal,K., Fropf,R., McAnany,C., Gagneur,J., Kundaje,A., *et al.* (2021) Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.*, **53**, 354–366.
22. Nair,S., Ameen,M., Sundaram,L., Pampari,A., Schreiber,J., Balsubramani,A., Wang,Y.X., Burns,D., Blau,H.M., Karakikes,I., *et al.* (2023) Transcription factor stoichiometry, motif affinity and syntax regulate single-cell chromatin dynamics during fibroblast reprogramming to pluripotency. bioRxiv doi: https://doi.org/10.1101/2023.10.04.560808, 21 October 2023, preprint: not peer reviewed.
23. Taskiran,I.I., Spanier,K.I., Dickmänken,H., Kempynck,N., Pančíková,A., Ekşi,E.C., Hulselmans,G., Ismail,J.N., Theunis,K., Vandepoel,R., *et al.* (2024) Cell-type-directed design of synthetic enhancers. *Nature*, **626**, 212–220.
24. Rada-Iglesias,A. (2018) Is H3K4me1 at enhancers correlative or causative? *Nat. Genet.*, **50**, 4–5.
25. Schneider,R., Bannister,A.J., Myers,F.A., Thorne,A.W., Crane-Robinson,C. and Kouzarides,T. (2004) Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nat. Cell Biol.*, **6**, 73–77.
26. Wysocka,J., Swigut,T., Xiao,H., Milne,T.A., Kwon,S.Y., Landry,J., Kauer,M., Tackett,A.J., Chait,B.T., Badenhorst,P., *et al.* (2006) A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, **442**, 86–90.
27. Foret,M.R., Sandstrom,R.S., Rhodes,C.T., Wang,Y., Berger,M.S. and Lin,C.-H.A. (2014) Molecular targets of chromatin repressive mark H3K9me3 in primate progenitor cells within adult neurogenic niches. *Front. Genet.*, **5**, 252.
28. Padeken,J., Methot,S.P. and Gasser,S.M. (2022) Establishment of H3K9-methylated heterochromatin and its functions in tissue differentiation and maintenance. *Nat. Rev. Mol. Cell Biol.*, **23**, 623–640.
29. Wang,C., Liu,X., Gao,Y., Yang,L., Li,C., Liu,W., Chen,C., Kou,X., Zhao,Y., Chen,J., *et al.* (2018) Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development. *Nat. Cell Biol.*, **20**, 620–631.
30. Cao,R., Wang,L., Wang,H., Xia,L., Erdjument-Bromage,H., Tempst,P., Jones,R.S. and Zhang,Y. (2002) Role of histone H3 lysine 27 methylation in polycomb-group silencing. *Science*, **298**, 1039–1043.
31. Cai,Y., Zhang,Y., Loh,Y.P., Tng,J.Q., Lim,M.C., Cao,Z., Raju,A., Lieberman Aiden,E., Li,S., Manikandan,L., *et al.* (2021) H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. *Nat. Commun.*, **12**, 719.
32. Ku,M., Koche,R.P., Rheinbay,E., Mendenhall,E.M., Endoh,M., Mikkelsen,T.S., Presser,A., Nusbaum,C., Xie,X., Chi,A.S., *et al.* (2008) Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.*, **4**, e1000242.
33. Sanz,L.A., Chamberlain,S., Sabourin,J.-C., Henckel,A., Magnuson,T., Hugnot,J.-P., Feil,R. and Arnaud,P. (2008) A mono-allelic bivalent chromatin domain controls tissue-specific imprinting at Grb10. *EMBO J.*, **27**, 2523–2532.
34. Li,W., Notani,D. and Rosenfeld,M.G. (2016) Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet.*, **17**, 207–223.
35. Carrozza,M.J., Li,B., Florens,L., Suganuma,T., Swanson,S.K., Lee,K.K., Shia,W.-J., Anderson,S., Yates,J., Washburn,M.P., *et al.* (2005) Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell*, **123**, 581–592.
36. Joshi,A.A. and Struhl,K. (2005) Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation. *Mol. Cell*, **20**, 971–978.
37. Herrera-Uribe,J., Liu,H., Byrne,K.A., Bond,Z.F., Loving,C.L. and Tuggle,C.K. (2020) Changes in H3K27ac at gene regulatory regions in porcine alveolar macrophages following LPS or PolyIC exposure. *Front. Genet.*, **11**, 817.
38. Ibragimov,A.N., Bylino,O.V. and Shidlovskii,Y.V. (2020) Molecular basis of the function of transcriptional enhancers. *Cells*, **9**, 1620.
39. Gates,L.A., Shi,J., Rohira,A.D., Feng,Q., Zhu,B., Bedford,M.T., Sagum,C.A., Jung,S.Y., Qin,J., Tsai,M.-J., *et al.* (2017) Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation. *J. Biol. Chem.*, **292**, 14456–14472.
40. Tarasov,A., Vilella,A.J., Cuppen,E., Nijman,I.J. and Prins,P. (2015) Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, **31**, 2032–2034.
41. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
42. Schreiber,J., Durham,T., Bilmes,J. and Noble,W.S. (2020) Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome Biol.*, **21**, 81.
43. Jung,I., Schmitt,A., Diao,Y., Lee,A.J., Liu,T., Yang,D., Tan,C., Eom,J., Chan,M., Chee,S., *et al.* (2019) A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.*, **51**, 1442–1449.
44. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D., *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic. Acids. Res.*, **44**, D733–D745.
45. Singh,R., Lanchantin,J., Robins,G. and Qi,Y. (2016) DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, **32**, i639–i648.
46. Murphy,A. (2024) neurogenomics/chromexpress: initial release. https://doi.org/10.5281/zenodo.10940543.
47. Whalen,S., Schreiber,J., Noble,W.S. and Pollard,K.S. (2022) Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet.*, **23**, 169–181.
48. Kingma,D.P. and Ba,J. (2017) Adam: a method for stochastic optimization. arXiv doi: https://doi.org/10.48550/arXiv.1412.6980, 22 December 2024, preprint: not peer reviewed.
49. Kang,M., Lee,S., Lee,D. and Kim,S. (2020) Learning cell-type-specific gene regulation mechanisms by multi-attention based deep Learning with regulatory latent space. *Front. Genet.*, **11**, 869.
50. Sekhon,A., Singh,R. and Qi,Y. (2018) DeepDiff: dEEP-learning for predicting DIFFerential gene expression from histone modifications. *Bioinformatics*, **34**, i891–i900.
51. Karollus,A., Mauermeier,T. and Gagneur,J. (2023) Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol.*, **24**, 56.

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

52. Murphy,A.E., Beardall,W., Rei,M., Phuycharoen,M. and Skene,N.G. (2024) Predicting cell type-specific epigenomic profiles accounting for distal genetic effects. *Nat Commun*, **15**, 9951.
53. Sasse,A., Ng,B., Spiro,A.E., Tasaki,S., Bennett,D.A., Gaiteri,C., De Jager,P.L., Chikina,M. and Mostafavi,S. (2023) Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nat. Genet.*, **55**, 2060–2064.
54. Toneyan,S. and Koo,P.K. (2024) Interpreting cis-regulatory interactions from large-scale deep neural networks for genomics. *Nat Genet*, **56**, 2517–2527.
55. Grishkevich,V. and Yanai,I. (2014) Gene length and expression level shape genomic novelties. *Genome Res.*, **24**, 1497.
56. Wang,Q.S., Kelley,D.R., Ulirsch,J., Kanai,M., Sadhuka,S., Cui,R., Albors,C., Cheng,N., Okada,Y., Aguet,F., *et al.* (2021) Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. *Nat. Commun.*, **12**, 3394.
57. Benner,C., Spencer,C.C.A., Havulinna,A.S., Salomaa,V., Ripatti,S. and Pirinen,M. (2016) FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, **32**, 1493–1501.
58. Wang,G., Sarkar,A., Carbonetto,P. and Stephens,M. (2020) A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **82**, 1273–1300.
59. Ghodsian,N., Abner,E., Emdin,C.A., Gobeil,É., Taba,N., Haas,M.E., Perrot,N., Manikpurage,H.D., Gagnon,É., Bourgault,J., *et al.* (2021) Electronic health record-based genome-wide meta-analysis provides insights on the genetic architecture of non-alcoholic fatty liver disease. *Cell Rep. Med.*, **2**, 100437.
60. Dönertaş,H.M., Fabian,D.K., Valenzuela,M.F., Partridge,L. and Thornton,J.M. (2021) Common genetic associations between age-related diseases. *Nat. Aging*, **1**, 400–412.
61. Nalls,M.A., Blauwendraat,C., Vallerga,C.L., Heilbron,K., Bandres-Ciga,S., Chang,D., Tan,M., Kia,D.A., Noyce,A.J., Xue,A., *et al.* (2019) Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.*, **18**, 1091–1102.
62. Bellenguez,C., Küçükali,F., Jansen,I.E., Kleineidam,L., Moreno-Grau,S., Amin,N., Naj,A.C., Campos-Martin,R., Grenier-Boley,B., Andrade,V., *et al.* (2022) New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet.*, **54**, 412–436.
63. Nicolas,A., Kenna,K.P., Renton,A.E., Ticozzi,N., Faghri,F., Chia,R., Dominov,J.A., Kenna,B.J., Nalls,M.A., Keagle,P., *et al.* (2018) Genome-wide analyses identify KIF5A as a novel ALS gene. *Neuron*, **97**, 1267–1288.
64. Trubetskoy,V., Pardiñas,A.F., Qi,T., Panagiotaropoulou,G., Awasthi,S., Bigdeli,T.B., Bryois,J., Chen,C.-Y., Dennison,C.A., Hall,L.S., *et al.* (2022) Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*, **604**, 502–508.
65. Grove,J., Ripke,S., Als,T.D., Mattheisen,M., Walters,R.K., Won,H., Pallesen,J., Agerbo,E., Andreassen,O.A., Anney,R., *et al.* (2019) Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.*, **51**, 431–444.
66. Mullins,N., Forstner,A.J., O'Connell,K.S., Coombes,B., Coleman,J.R.I., Qiao,Z., Als,T.D., Bigdeli,T.B., Børte,S., Bryois,J., *et al.* (2021) Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat. Genet.*, **53**, 817–829.
67. Elsworth,B., Lyon,M., Alexander,T., Liu,Y., Matthews,P., Hallett,J., Bates,P., Palmer,T., Haberland,V., Smith,G.D., *et al.* (2020) The MRC IEU OpenGWAS data infrastructure. bioRxiv doi: https://doi.org/10.1101/2020.08.10.244293, 10 August 2020, preprint: not peer reviewed.
68. Sarkans,U., Gostev,M., Athar,A., Behrangi,E., Melnichuk,O., Ali,A., Minguet,J., Rada,J.C., Snow,C., Tikhonov,A., *et al.* (2018) The BioStudies database—One stop shop for all data supporting a life sciences study. *Nucleic Acids Res.*, **46**, D1266–D1270.
69. Murphy,A.E., Schilder,B.M. and Skene,N.G. (2021) MungeSumstats: a bioconductor package for the standardization and quality control of many GWAS summary statistics. *Bioinformatics*, **37**, 4593–4596.
70. Finucane,H.K., Bulik-Sullivan,B., Gusev,A., Trynka,G., Reshef,Y., Loh,P.-R., Anttila,V., Xu,H., Zang,C., Farh,K., *et al.* (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, **47**, 1228–1235.
71. Soldi,M., Mari,T., Nicosia,L., Musiani,D., Sigismondo,G., Cuomo,A., Pavesi,G. and Bonaldi,T. (2017) Chromatin proteomics reveals novel combinatorial histone modification signatures that mark distinct subpopulations of macrophage enhancers. *Nucleic Acids Res.*, **45**, 12195–12213.
72. Tani,H., Mizutani,R., Salam,K.A., Tano,K., Ijiri,K., Wakamatsu,A., Isogai,T., Suzuki,Y. and Akimitsu,N. (2012) Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res.*, **22**, 947–956.
73. Schwanhäusser,B., Busse,D., Li,N., Dittmar,G., Schuchhardt,J., Wolf,J., Chen,W. and Selbach,M. (2011) Global quantification of mammalian gene expression control. *Nature*, **473**, 337–342.
74. Chantalat,S., Depaux,A., Héry,P., Barral,S., Thuret,J.-Y., Dimitrov,S. and Gérard,M. (2011) Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin. *Genome Res.*, **21**, 1426–1437.
75. Bernstein,B.E., Mikkelsen,T.S., Xie,X., Kamal,M., Huebert,D.J., Cuff,J., Fry,B., Meissner,A., Wernig,M., Plath,K., *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.
76. Policarpi,C., Munafò,M., Tsagkris,S., Carlini,V. and Hackett,J.A. (2024) Systematic epigenome editing captures the context-dependent instructive function of chromatin modifications. *Nat. Genet.*, **56**, 1168–1180.
77. Takahashi,H., Kato,S., Murata,M. and Carninci,P. (2012) CAGE-Cap analysis gene expression: a protocol for the detection of promoter and transcriptional networks. *Methods Mol. Biol.*, **786**, 181–200.
78. Mahat,D.B., Kwak,H., Booth,G.T., Jonkers,I.H., Danko,C.G., Patel,R.K., Waters,C.T., Munson,K., Core,L.J. and Lis,J.T. (2016) Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.*, **11**, 1455–1476.
79. Sudlow,C., Gallacher,J., Allen,N., Beral,V., Burton,P., Danesh,J., Downey,P., Elliott,P., Green,J., Landray,M., *et al.* (2015) UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, **12**, e1001779.
80. Wang,L., Khunsriraksakul,C., Markus,H., Chen,D., Zhang,F., Chen,F., Zhan,X., Carrel,L., Liu,D.J. and Jiang,B. (2024) Integrating single cell expression quantitative trait loci summary statistics to understand complex trait risk genes. *Nat. Commun.*, **15**, 4260.
81. Hechtlinger,Y. (2016) Interpretation of prediction models using the input gradient. arXiv doi: https://doi.org/10.48550/arXiv.1611.07634, 23 November 2016, preprint: not peer reviewed.
82. Zhang,H.-J., Chang,W.-J., Jia,C.-Y., Qiao,L., Zhou,J., Chen,Q., Zheng,X.-W., Zhang,J.-H., Li,H.-C., Yang,Z.-Y., *et al.* (2020) Destrin contributes to lung adenocarcinoma progression by activating wnt/β-catenin signaling pathway. *Mol. Cancer Res.*, **18**, 1789–1802.
83. Zhao,S., Ye,Z. and Stanton,R. (2020) Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA*, **26**, 903–909.
84. Toneyan,S., Tang,Z. and Koo,P.K. (2022) Evaluating deep learning for predicting epigenomic profiles. *Nat Mach Intell*, **4**, 1088–1100.

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

85. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O., *et al.* (2009) Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
86. Liang,G. and Zhang,Y. (2013) Embryonic stem cell and induced pluripotent stem cell: an epigenetic perspective. *Cell Res.*, **23**, 49–69.
87. Wu,K., Fan,D., Zhao,H., Liu,Z., Hou,Z., Tao,W., Yu,G., Yuan,S., Zhu,X., Kang,M., *et al.* (2023) Dynamics of histone acetylation during human early embryogenesis. *Cell Discov.*, **9**, 1–24.
88. Mostafavi,H., Spence,J.P., Naqvi,S. and Pritchard,J.K. (2023) Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat. Genet.*, **55**, 1866–1875.
89. Nordin,A., Pagella,P., Zambanini,G. and Cantù,C. (2024) Exhaustive identification of genome-wide binding events of transcriptional regulators. *Nucleic Acids Res.*, **52**, e40.

Downloaded from https://academic.oup.com/nar/article/53/4/gkae1212/7921050 by Imperial College London Library user on 10 March 2025

Received: March 29, 2024. **Revised:** October 12, 2024. **Editorial Decision:** November 15, 2024. **Accepted:** December 9, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.