






Understanding relationships between epigenetic marks and their application to robust assignment of chromatin states

Leandro Murgas ^{1,2}, Gianluca Pollastri ³, Erick Riquelme ⁴, Mauricio Sáez ^{5,*}, Alberto J.M. Martin ^{2,6,*}

¹Programa de Doctorado en Genómica Integrativa, Vicerrectoría de investigación, Universidad Mayor, Camino La Pirámide 5750, 8580745 Huechuraba, Chile

²Laboratorio de Redes Biológicas, Centro Científico y Tecnológico de Excelencia Ciencia & Vida, Fundación Ciencia & Vida, Avda. del Valle 725, 8580702 Huechuraba, Chile

³School of Computer Science, University College Dublin, Belfield, Dublin 4, Dublin D04 C1P1, Ireland

⁴Department of Respiratory Diseases, Facultad de Medicina, Pontificia Universidad Católica, Avda. Libertador Bernardo O'Higgins 340, 8331150 Santiago, Chile

⁵Laboratorio de Investigación en Salud de Precisión, Departamento de Procesos Diagnósticos y Evaluación, Facultad de Ciencias de la Salud, Universidad Católica de Temuco, Manuel Montt 56, 4813302 Temuco, Chile

⁶Escuela de Ingeniería, Facultad de Ingeniería, Arquitectura y Diseño, Universidad San Sebastián, Bellavista 7, 8420524 Santiago, Chile

*Corresponding authors: Alberto J. M. Martin, E-mail: alberto.martin@uss.cl; Mauricio Sáez, E-mail: mauricio.saez@uct.cl

Abstract

Structural changes of chromatin modulate access to DNA for the molecular machinery involved in the control of transcription. These changes are linked to variations in epigenetic marks that allow to classify chromatin in different functional states depending on the pattern of these histone marks. Importantly, alterations in chromatin states are known to be linked with various diseases, and their changes are known to explain processes such as cellular proliferation. For most of the available samples, there are not enough epigenomic data available to accurately determine chromatin states for the cells affected in each of them. This is mainly due to high costs of performing this type of experiments but also because of lack of a sufficient amount of sample or its degradation. In this work, we describe a cascade method based on a random forest algorithm to infer epigenetic marks, and by doing so, to identify relationships between different histone marks. Importantly, our approach also reduces the number of experimentally determined marks required to assign chromatin states. Moreover, in this work we have identified several relationships between patterns of different histone marks, which strengthens the evidence in favor of a redundant epigenetic code.

Keywords: Random Forest; epigenetic marks; chromatin states

Introduction

Structural changes in chromatin modulate access to DNA by all proteins involved in transcription. This process is linked to variations in histone modifications, which are also known to be related to an ATP-dependent remodeling complex that causes dynamic restructuring of nucleosomes [1]. Histone Post-Translational Modifications (PTMs) or histone marks are covalent modifications usually located in histone tails, that include methylation, phosphorylation, acetylation, ubiquitylation, and sumoylation [2]. These PTMs are known to affect chromatin structure and can be classified as active or repressive marks based on their effect on transcriptional regulation [3, 4].

The hypothesis of the 'histone code', proposed by Strahl and Allis in the early 2000s [5], states that a combination of histone modifications at a specific genomic locus determines the activity of underlying genes. The histone code was next expanded to an 'epigenetic code' to include other epigenetic marks, such as DNA methylation [6, 7]. This code is considered to be conserved in each species; that is, the effect of a pattern of marks is identical for all cells of the same organism and all individuals of the same species [2]. Moreover, the code is considered to be redundant, since different combinations of histone modifications contain the

same message regarding gene regulation [2]. Additionally, different studies have shown that the existing patterns of histone modifications are also strongly related to chromatin structure [5, 8].

Chromatin was traditionally divided into two functional states, euchromatin transcriptionally active and heterochromatin inactive [9]. More recent chromatin annotation methods have classified chromatin into different number of states, ranging from two to several tens, each state associated with a different functional condition [10]. For example, a large-scale integrative genome-wide analysis of 53 chromatin-associated proteins in *Drosophila melanogaster* recognized five chromatin states [11]. Most classifications of chromatin states rely on epigenomic data for many histone marks and DNA methylation [12, 13]. ChromHMM is one of the most widely used tools to classify chromatin states [14]. This methodology is based on multivariate Hidden Markov Models, an unsupervised algorithm, to assign states from epigenetic information. Using information from different epigenetic marks in various cell lines, the authors of ChromHMM expanded the number of known chromatin states to 18 [15].

Many complex diseases have been linked to changes in chromatin states associated with misregulation of gene expression [16]. Among these diseases, it is possible to highlight most types

Received: March 31, 2024. Revised: September 9, 2024. Accepted: December 2, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

For commercial re-use, please contact journals.permissions@oup.com

of cancer, diabetes, and several neurological and cardiovascular complex diseases [17–19]. For example, several studies have associated colorectal cancer with alterations in chromatin states, identifying abnormalities in DNA methylation patterns, and different histone modifications [20, 21]. Melanoma progression has also been associated with variations in normal patterns of epigenetic marks [22]. Importantly, for most complex diseases, there are not enough epigenomic data available to neither determine chromatin states nor their alterations.

Machine Learning (ML) has proven to be indispensable for interpreting large genomic datasets, integrating several types of 'omic' data and for the study and diagnosis of diseases [23]. Thus, ML methods can be used to learn how to recognize the location of Transcription Start Sites in a genome sequence [24], to infer gene expression [25], or to annotate new Transcription Factor Binding Sites (TFBSs) [26]. ML algorithms can use input data generated by different genomic assays, such as expression data, chromatin accessibility, histone modifications, or TFBSs [23]. There are several ML-based methodologies that help solve the scarcity of experimental information. For example, CHROMIMPUTE [27] and AVOCADO [28] use ML strategies to predict PTM enrichment signals from epigenetic information. CHROMIMPUTE relies on regression tree algorithms to impute PTMs from other marks from the same sample. Similarly, AVOCADO uses multi-scale deep tensor factorization to represent epigenetic information in a neural network model that allows predictions of PTMs from other histone marks. Both methodologies use large amounts of information to make optimal predictions, and in the case of AVOCADO, it could be difficult to employ by non-expert users. In addition, the information provided by these tools only indicates the presence or absence of the target PTM, without explaining how this prediction was made or what relationships can be found between the different PTMs used.

Here we describe a new Random Forest (RF)-based method to predict histone modifications from other marks. Notably, owing to the inherent properties of this algorithm, our approach also identifies different relationships between PTMs that are employed to predict one PTM based on the others. In addition, our tool also allows the reduction of the number of experiments required for the robust assignment of chromatin states, reducing the cost associated to chromatin state assignment.

Methods and materials

Data

Enhancer annotations were obtained from GeneCards [29], whereas promoters and gene coordinates are from ENSEMBL [30], both for the human genome version GRCh38. We used 33 PTMs ChIP-seq experiments from Fizev et al. [22], each in two biological conditions annotated as tumorigenic (Tum) and non-tumorigenic (noTum) melanocytes (GSE58953) in two biological replicates, Hmel and Pmel. Fastq files of the ChIP-seq experiments were reanalyzed and aligned against genome GRCh38 with Bowtie2 [31]. PCR duplicates were removed using SAMTools [32], and BEDgraph files were obtained with BamCoverage [33] with default parameters, different bin sizes, and normalizing the reads to RPKM. For colorectal cancer data, ChIP-seq experiments of 11 PTMs of the HCT116 cell line available in ENCODE [34] were analyzed using the same pipeline.

Processed ChIP-seq experiments were used to create training and testing datasets for the ML algorithms (Fig 1A). The genome was segmented into fragments of different sizes. Each fragment is then described by a vector where each element represents

the RPKM value per fragment of reads for each histone mark. Genomic features were represented by a number in the [0,1] range indicating the proportion of bases covered by that feature in the fragment. Finally, each vector is associated with a label representing the characteristic to predict for its central fragment, for example, a specific PTM. Fragments up y downstream of this central fragment within a fixed distance limit are also added to the vector.

Parameter determination

RF regressors from Scikit-Learn [35] were trained and tested using the 32 other PTMs ChIP-seq data and genomic features as input to predict PTMs H3K4me1, H3K4me3, and H3K27ac from the Pmel-noTum cell had of the PTM, one at the time using only chromosome 1 data. Different training parameters were tested to define the fragment size, distance, and RF depth, using 75% of the dataset to train each combination of parameters and the remaining 25 % as the test set. The performance was evaluated using Pearson's correlation coefficient between the label of the data used for testing and the predicted values.

Initial tests had correlation values between 0.88 and 0.95 for H3K4me1, 0.82 and 0.93 for H3K4me3, and in the range 0.94 and 0.98 for H3K27ac. These analyses also showed that the fragment size is the hyperparameter with the greater impact on the performance compared to distance or tree depth (Table S1). Tests using fragment sizes of 2K, 3K, 4K, and 5K bp showed small differences, so we opted for shortest fragments to increase the resolution of our model. According to these tests, we decided to generate all models using fragment size of 2K bases, distances of 10K bases, and tree depth 25.

Identification of relationships between HPTMs

To determine the relationships between different PTMs, we trained regression RFs to predict each of the marks by employing different combinations of other marks selected according to the relevance of these marks to predict each other. This is expected to determine four types of relationships (Fig. 2): dependency, if a mark is necessary to predict another mark; interdependence—if two marks or combinations of them can better predict another mark, there is an interdependence between the marks used to predict and the target; redundancy— if both one mark and another can predict a third, there is redundancy between them; finally, non-informative relationships, if one mark, or a combination of brands, cannot predict another or worsens its prediction significantly.

Assignment of chromatin states

Predicted PTMs were next employed as input to train a multi-label RF classifier to assign chromatin states (Fig 1B). In the first instance, ChromHMM [15] and the experimental data (BAM files) of the 33 PTMs of melanocytes in the two biological conditions (Hmel-noTum and Hmel-Tum) were used to make an assignment of 8 chromatin states based on the work presented by Jieang et al. [36]. We also using the same size of 2000 bp chromatin fragments as for the other predictors. This outcome was used as target to train the predictor, where '1' was assigned for state presence and '0' for its absence. In the case of the features, these correspond to predicted PTMs in the fragment. Precision (P), which corresponds to the total percentage of elements correctly classified, and Recall (R), the number of elements correctly identified as positives out of the total number of true positives, were used to evaluate the predictor performance (See formula below). We also analyzed the enrichment of the PTMs in each state and the presence of

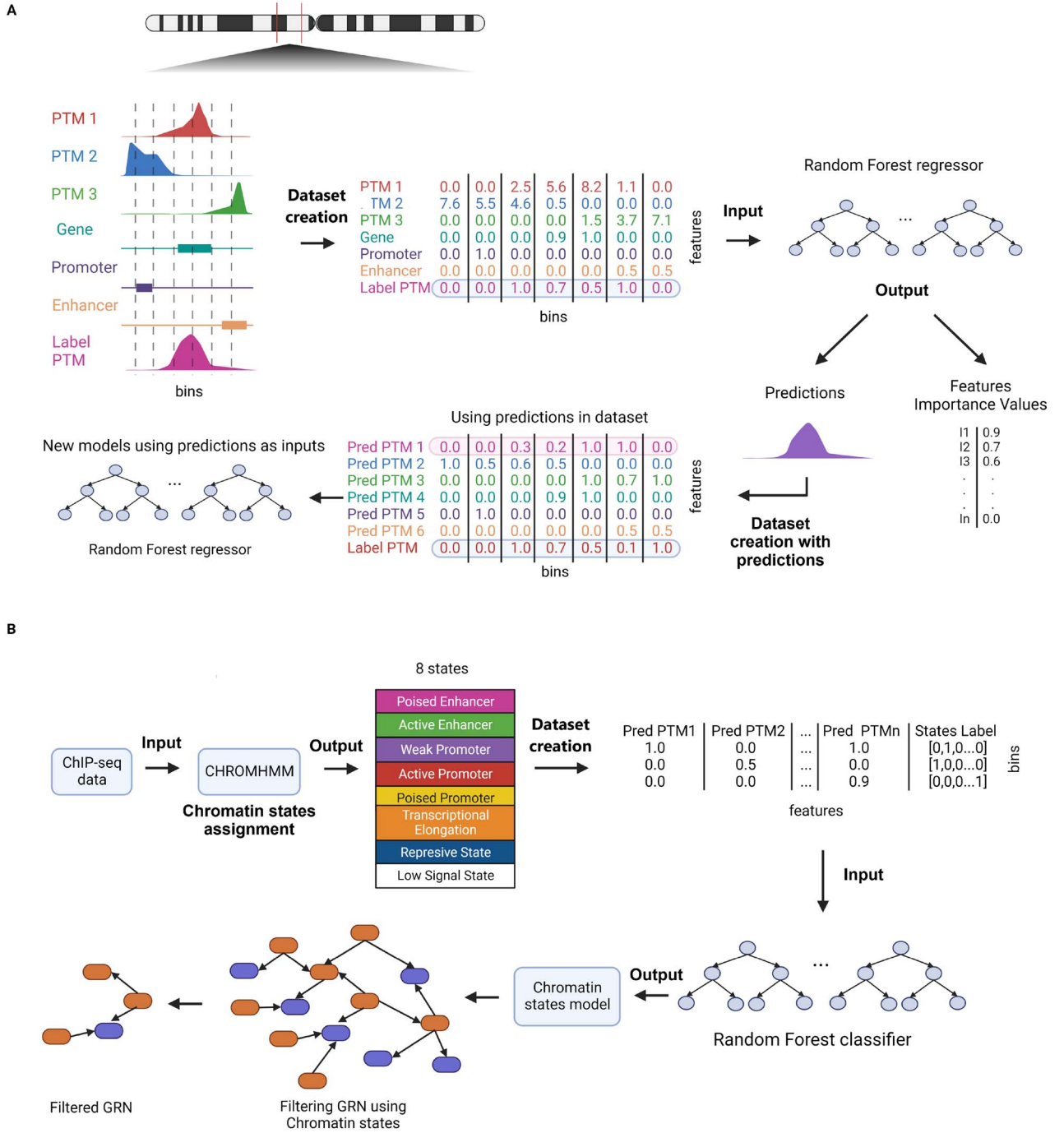


Figure 1. Methodology: (A) we first created dataset files using ChIP-seq data and enhancers, gene, and promoter genomic annotations. These files were employed to train and test RF models and to create them the genome was divided into bins. The presence or absence of the different features was represented using RPKMs for ChIP-seq peaks in each bin, and the percentage of coverage for each of the genome annotation features. RFs were trained to predict PTMs from other marks, to determine the most important features for each prediction and to determine relationships between features. (B) In a second stage, predicted PTMs were employed to assign chromatin states. First, ChromHMM was used to annotate 8 chromatin states from all 33 available experimental marks. Then, this assignment was coded for the training and testing of another RF to classify chromatin states from predicted marks. Finally, we generated a Gene Regulatory Network (GRN) based on the regulatory nature associated to each chromatin state predicted for each chromatin region.

genome annotations such as coordinates of promoters, genes, and enhancers. The models were trained using a five-fold cross validation with chromosome 1 data.

$$P = \frac{TP}{TP+FP}; R = \frac{TP}{TP+FN}$$

Where a TP (true positive) is a chromatin fragment which state is correctly predicted, an FP (false positive) is a fragment predicted to be in a state different to the one it actually is in, and an FN

(false negative) is a fragment that should have been predicted to be in a certain state but it was not.

GRN analysis

To further validate chromatin states assigned from predicted epigenetics marks we constructed GRNs. In brief, a reference network is filtered using epigenetic data to remove unlikely regulatory interactions. Here, instead we removed interactions that arise

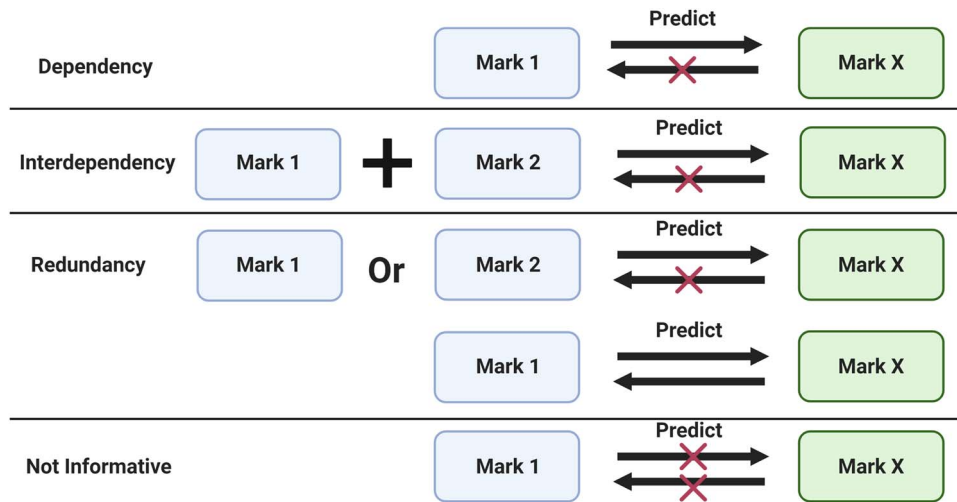


Figure 2. Possible relationships between histone marks. Dependency, if a mark is necessary to predict another mark; interdependency, if two marks or a combination of marks are able to predict better another mark; Redundancy, if both mark and another can predict a third there is redundancy between them; not-informative relationships, if a mark, or a combination of marks, cannot predict another.

from genes that are in chromatin states known to contain inactive regulatory regions. We built a human reference network using data of promoter coordinates available at GeneCards [29]. Promoters were kept only if they were no more than 5 kb apart from their target gene. These promoters were then filtered using the 'Active Promoter' chromatin state assigned with the methodology mentioned above, considering that the state completely covers the promoter. The GRN was created by assigning Transcription Factors (TFs) to target genes in the promoters where their TFBSs are transcriptionally active. The results of this validation on two cell lines are shown in Supplementary file 2.

Code availability

All code employed is available under GNUv3 license at https://github.com/networkbiolab/RF_histonemarks and <https://zenodo.org/record/7547293>.

Results

Performance of HPTMs prediction

First, we generated models trained on 35 different attributes (32 PTMs and coordinates of Enhancer, Promoter, and Gene annotations) from the Pmel-noTum cell. We blind tested these models with data from the Hmel-noTum, Pmel-Tum, and Hmel-Tum cell lines. These analyses showed it was possible to obtain robust models for the 33 PTMs (positive correlation values above 0.5). When testing with the Hmel-noTum, the lowest correlation value was 0.75 for H3K27me3 and the highest for H2BK5ac with 0.97. In the case of testing the Hmel-Tum cell line, the lowest value was 0.73 for H4K20me2 and the highest for H3K79me1 with 0.96, while for Pmel-Tum, H3K4me2 obtained the lowest with 0.69 and the highest H2BK5ac with 0.93 (see Table 1 (35 attr), for the other PTMs (Table S2).

Following, we generated models to predict PTMs with a lower number of input attributes to test whether it is possible to generate robust models with a smaller amount of information. One of the advantages of the RF algorithm is that it identifies which attributes are the most relevant for prediction. We trained models with the top eight attributes with the greatest importance as reported by the RF for each PTM (Table S3). These models showed correlation values similar to those obtained with all

information (32 histone marks, Enhancers, Promoters, and Gene coordinates). For example, in the Hmel-noTum, for H2BK5ac, the same correlation coefficient of 0.97 was maintained for all and only the top eight attributes; in the case of H3K4me2, it decreased from 0.87 to 0.86, and for H3K36me1, it increased from 0.86 to 0.87 (Tables 1 and S3). Importantly, visual inspection of our predictions compared to actual experimental data with the IGV genome browser [37] shows no significant variations (Fig. 3). Moreover, when performing whole genome tests, the performance is very similar to that of using only chromosome 1 (Table S4).

HPTMs relationships

We also identified possible relationships between the analyzed PTMs using the eight most important attributes reported by the RF models. We grouped together PTMs that shared three attributes in their respective top eight predicting attributes, obtaining 12 different groups (Table 2). This grouping of PTMs suggests that it is possible to predict any mark of those belonging to same group using the same attributes. We verified this by generating new models in which only these shared attributes were used to predict each group PTMs, showing robust performance values for all predicted PTMs (Table S5).

Additionally, we trained models to determine the relationships between PTMs by varying the number of attributes, using only the most important attribute, and increasing the number of attributes until reaching the eight most important attributes for each mark. In general, all except for H3K27me3 and H3K4me3 present positive correlation values above 0.50 in all the cell lines and experimental settings using only the most important attribute. There is an improvement in performance by adding the second most important attribute, which would indicate possible interdependence relationships. Adding the third attribute, we observed a decrease in correlation for most of the PTMs evaluated but for H2BK120ac, H3K4ac, and H4K91ac, for which the performance was the same as with two input features. The fourth attribute improved the previous models, but in most cases, correlation was not greater than those obtained with the two most important attributes. Finally, improvements were observed in some of the cell lines evaluated by adding more input features (Table S6).

Table 1. Summary of the results obtained when analyzing the 33 available histone marks with 35 and 8 attributes

HPTMs	Hmel-noTum		Hmel-Tum		Pmel-Tum	
	35 attr	8 attr	35 attr	8 attr	35 attr	8 attr
H2BK5ac	0.97	0.97	0.94	0.94	0.93	0.93
H3K4me2	0.87	0.86	0.79	0.79	0.69	0.68
H3K79me1	0.95	0.95	0.96	0.96	0.92	0.91
H4K20me2	0.77	0.77	0.73	0.73	0.71	0.71
H4K20me3	0.75	0.75	0.75	0.76	0.74	0.74
H3K36me1	0.86	0.87	0.83	0.83	0.82	0.83

These models were trained using the non-tumorigenic cell line Pmel with 35 attributes (32 PTMs and annotation data) and 8 attributes with greater importance according to the RFs trained for each mark. The models were tested using the data from the non-tumorigenic cell lines Hmel (Hmel-noTum) and the tumorigenic cell lines Hmel and Pmel (Hmel-Tum and Pmel-Tum). The performance of the models was analyzed using Pearson's correlation.

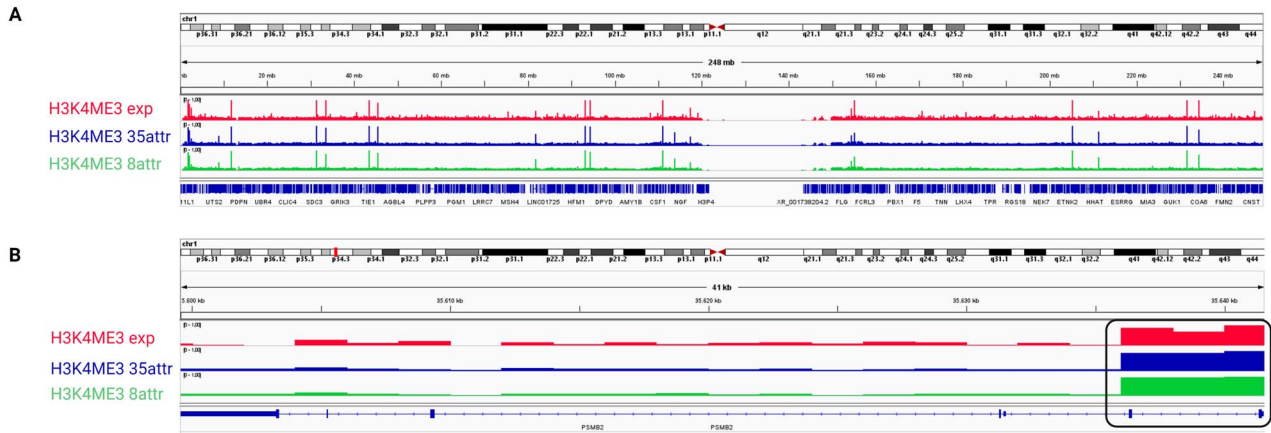


Figure 3. Comparison of H3K4me3 predicted and actual peaks with IGV genome browser. (A) Whole chromosome 1 enrichment peaks, the top track displays experimental data, the middle track contains those predicted with the 35 attributes model, and the bottom track shows those predicted with the model generated with the 8 most important attributes. (B) Visualization of the PSMB2 gene, the square indicates regions of greater enrichment of H3K4me3, corresponding to the promoter region of PSMB2.

Table 2. Groups of HPTMs that share attributes

Groups	HPTMs	Shared attributes
1	H3K23ac, H3K4ac	H3K27ac, H3K14ac, H4K5ac
2	H3K9me1, H3K36me3, H4K8ac, H3K27me1	H4K12ac, H3K4ac, H4K16ac
3	H3K4me1, H2AK5ac, H4K91ac, H4K5ac, H2BK5ac, H3K36ac	H3K4ac, H3K27ac, H2BK120ac
4	H3K79me2, H3K27me3, H4K20me1	H3K79me1, H3K79me3, H4K12ac
5	H3K36me2, H3K36me1	H3K14ac, H3K27me3, H4K20me1
6	H3K27ac, H3K9ac, H3K14ac	H3K4ac, H3K23ac, H3K4me2
7	H2BK120ac, H3K18ac	H2BK5ac, H3K4ac, H3K4me1
8	H3K4me2, H3K4me3	H3K9me3, H3K4me1, H3K9ac
9	H4K20me3, H4K20me2, H3K79me3	H3K4ac, H3K9me3, H4K20me1
10	H3K79me1, H3K9me3	H4K16ac, H4K20me1, H3K36me2
11	H2BK15ac, H4TETRAac	H3K4ac, H3K18ac
12	H4K16ac, H4K12ac	H4K8ac, H3K79me1, H4K5ac

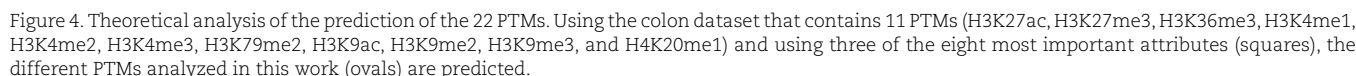
First column indicates the groups based on the shared predictive marks; second column contains PTMs belonging to each group and the third column the shared attributes for each group.

Comparison with other methods

We compared our approach with two methodologies for predicting PTMs from other PTMs with CHROMIMPUTE, based on

regression-type decision trees, and AVOCADO, which uses a tensor factorization approach. For both comparisons, Pmel-noTum and Hmel-noTum on chromosome 1 were used. For each compared model, the same histone marks used in our method were used as attributes to predict the marks in the Hmel-noTum cell line. CHROMIMPUTE accepts the same BEDgraphs used to test our method as input, so the comparison is straightforward. For AVOCADO, the data needed to be reanalyzed using 250-base fragments because of the input format required by this tool. We transformed the output of AVOCADO as follows: the average predicted signal was calculated for every eight fragments and used as the label for the equivalent 2kb BEDgraphs.

The performance of each method was evaluated by calculating the correlation between the predicted data and the experimental data. We estimated statistical significance using t-tests and 100 random subsamplings of 80% for each mark (Table S7). Our approach shows better performance than CHROMIMPUTE for the 33 PTMs analyzed, reaching a very large difference in PTMs, such as H3K27me3, where our approach has a correlation coefficient of 0.81 and CHROMIMPUTE of only 0.32. Compared with AVOCADO, the latter performed better in three of the 33 PTMs analyzed (H3K27me1, H3K27me3, and H4K20me2). Regarding the 8-attribute models, the same scenario was repeated when comparing our method against CHROMIMPUTE showed worse performance for all marks. When compared to AVOCADO, our RF approach is only worse with H3K27me3 using the 8-attributes models. Finally, CHROMIMPUTE also fails to outperform our approach when using models with fewer attributes. AVOCADO



Test in colorectal cancer

Additionally, the generalization capability of the models was evaluated using a different cell line, which had no relation to any cancerous cell line. For this purpose, an astrocyte cell line was used, which possesses the same HPTMs available for HCT116. The

Prediction cascade

Chromatin state assignment using predictions

A gold standard was first generated using the ChIP-seq data from the 33 PTMs of Hmel-noTum and the ChromHMM tool to assign 8 chromatin states ('poised enhancer', 'active enhancer', 'weak promoter', 'active promoter', 'poised promoter', 'transcriptional elongation', 'repressive state', and 'low signal state') which were assigned to chromatin fragments based on [36]. Because the generated RF models predict peaks in the BEDgraph format, it is impossible to use these data directly with ChromHMM software.

Table 3. Chromatin states assigned with predictions from the cascade tool

State	HPTMs cascade		10 exp HPTMs	
	P	R	P	R
Poised promoter	0.84	0.81	0.81	0.74
Transcriptional elongation	0.78	0.7	0.77	0.68
Active enhancer	0.89	0.81	0.89	0.79
Active promoter	0.85	0.78	0.83	0.75
Poised enhancer	0.71	0.57	0.69	0.49
Low signal	0.87	0.9	0.84	0.89
Weak promoter	0.9	0.8	0.85	0.73
Repressive	0.91	0.86	0.9	0.82

Chromatin states assignment using cascade predictors and experimental data. 8 chromatin states were assigned by means of a classifier RF-based predictor, which uses predictions obtained by our random forest models cascade and labels obtained with ChromHMM as input data. A second assignment was made using only experimental data of 10 histone marks. State labels were assigned according to the enrichment of 33 histone marks experimentally determined. The performance of the models was measured using to the Precision (P) and Recall (R) metrics.

For this purpose, a new predictive model based on a multi-label RF classifier was developed. Two models were trained using the predicted data for all 33 histone marks and only 10 marks (H3K27ac, H3K27me3, H3K36me3, H3K4me1/me2/me3, H3K79me2, H3K9ac, H3K9me3, and H4K20me1). As target, we employed chromatin states generated with ChromHMM using the experimental data from 33 PTMs. This subset of 10 marks was selected because they are among the most characterized in databases and some are of clinical interest because of their association with cancer [38, 39]. These models can generate robust predictions of chromatin states using only the predicted data. Precision (P) and Recall (R) values $\geq 70\%$ are obtained for both models on the genome not used to train the predictors (chromosomes 2–22 and X), with the worst performance was for the state ‘Poised enhancer’ state (Table S12).

We also generated a new model to predict chromatin states; however, we used the 23 PTMs predicted marks of Hmel-noTum obtained with the cascade of predictors and the experimental data of the remaining 10 PTMs. This new model shows P and R values $\geq 70\%$ in seven states. Only the state ‘Poised enhancer’ obtained an R-value of 57%. To test whether the use of predictions and experimental data improves the performance of chromatin assignment, a new model was generated that uses only the experimental data from the 10 PTMs used for the previous model. When comparing the performance of both models, including the predicted data improved chromatin assignment in all states (Table 3).

Discussion

Epigenetic marks are usually characterized as determinants of the activity of *cis* and *trans* regulatory elements in the genome. Several marks have been associated with active or repressive promoter activity [2], similar to the activity of enhancers [40] or for the binding of TFs to chromatin [41, 42]. Another application of these marks is to determine chromatin states, a more detailed definition that subdivides traditional euchromatin/heterochromatin into several functional states, known to require many different PTMs for consistency [43]. One of the less studied aspects of epigenetic marks is the existence of relationships between different PTMs, where the effect of a single PTM is altered by the presence of other nearby marks [44, 45]. Even more important is demonstrating the existence of an epigenetic code where different patterns of epigenetic modifications of histones linked to specific regions are associated with specific effects [6, 7].

In this work, we describe a new tool for predicting epigenetic marks from other epigenetic PTMs. Our approach allows us to increase the amount of data available for studying different tissues or cell lines for which there is not much data available, aiming to reduce the cost of analyses that require many histone marks. Moreover, our methodology allowed us to establish several interdependency and redundancy relationships between different epigenetic marks, thus providing evidence for the establishment of the human epigenetic code. We chose RFs as our predictive algorithm because, in addition to accurate predictions they report a ranking of importance for each attribute used as input [35].

When comparing our predictions with other methods, the most relevant difference is that the RF predictor generates interpretative relationships that, for example, neural networks cannot generate. In absolute terms, CHROMIMPUTE is worse than our tool for all marks in our experiments. It should be noted, that CHROMIMPUTE uses information from other marks in the same sample and from the same mark in other samples. Only intra sample data is employed in our tests and thus, CHROMIMPUTE performance is assessed in sub-optimal conditions, even so if in our tests performance is not that far away from the performance reported by the authors of CHROMIMPUTE. In the case of AVOCADO, it is better for only a few marks independently of the number of marks employed as the input. These results indicate that despite the algorithm used, our RF models outperform other existing methods, and that the modest hardware requirements make our models more readily usable. The ability to determine the relevance of the input attributes is a very important difference from other algorithms because, in this application, it is important to understand the relationships reported when training RFs to obtain good predictions.

Given the output format of our methodology, which consists of BEDgraph files, it was not possible to directly use tools such as ChromHMM to assign states to chromatin. This incompatibility is something that can be improved in the future for our method and for the state assignment methodologies because many of the epigenetic data available in public databases are in this format, making it impossible to use them for tasks of this type. For this reason, to verify if the use of predictions obtained with our method, the assignment of chromatin states was possible, a state predictor was made using the states generated by ChromHMM and predicted BEDgraphs for histone marks. This model achieved robust state assignments that were very similar to those obtained with the experimental data (Tables 3 and S12).

Regarding the marks for which our models showed worse prediction performance, H3K27me3 and H3K4me3, these two marks have been associated to bivalent promoters, i.e. to co-occur at the same promoter during development and cell differentiation [46, 47]. Importantly, these two marks are known to follow cell specific patterns and to be found only at reduced number of loci, and that specially the presence of H3K4me3 is very dynamic compared to other marks [48, 49]. Given the limited information we provide our ML models about the region in which the histone marks are found and the higher specificity of these two marks with respect to their specific locations marks more difficult to predict them than the other marks. It is also possible to hypothesize that given the relative strong relationship between these two marks, their relationships with the other marks employed in this work are weaker or in other words, they follow a relationship where the type of loci is more important than the patterns of other neighboring marks.

Generally, it is thought that greater precision is expected by having greater amount of data because more information about

the problem to be solved would be available. However, our results show that it is possible to generate predictions of equal or better quality using fewer experiments (Tables 1 and S5). Importantly, this also supports the existence of dependency and redundancy relationships among the PTMs. For example, when avoiding the use of a certain mark but using other marks to make predictions, if the results are not altered, we can assume that there is redundancy between the marks used and the one not used to train the predictor. When the use of a mark worsens the performance of the prediction of another mark, we can assume that there is no relationship between them or that it is counter-informative. Similarly, if more than one mark is required to accurately predict another mark and without these marks, or if using them separately, the predictions are much worse, we assume that there is an interdependence relationship. This type of relationship is observed, for example, with the analysis of H4K16ac, a mark related to breast, prostate, and colorectal cancer [38, 39], the results of which show a relationship of the interdependence of H4K12ac and H4K8ac (Additional File 1: Tables S3 and S6). Importantly, these relationships support the existence of an epigenetic code [6, 7].

Another type of relationship is defined by grouping PTMs using shared attributes from those that are among the most important for their prediction. In this way, redundancy-type relationships were identified. For example, we could say that H3K23ac and H3K4ac, which belong to group 1 (Table 2), would have a redundancy relationship with H3K27ac, H3K14ac, and H4K5ac because they all share these three marks as predicting attributes. Importantly, we generated robust models using only these three marks to predict H3K23ac and H3K4ac (Additional file 1: Table S5). Thus, given that it is possible to determine where two marks are present by determining the other three, our predictions provide evidence that the epigenetic code is redundant.

We also observed a slight drop in the performance of the models trained on non-tumor cells when tested on tumor cells. This phenomenon may be because there is a known alteration in the patterns of certain marks in tumors [22]. Therefore, the cascade of alterations caused at many levels in tumors affects the manner in which epigenetic marks are related to each other and could affect our predictions [50, 51]. Our results also indicate that, at least in the cell lines employed in this study, there are some marks with prediction results that are not different between tumoral and non-tumoral cells, as indicated by the good generalization capabilities of our models (Table 1). These findings suggest that, at least in the samples we used, only the levels and locations of certain marks are affected, and that these changes only partially affect the relationships found using our RF approach.

Next, we analyzed if our models were biased to the cell lines used in their training in which it was being analyzed. For this test, we employed a dataset of 10 histone marks from the HCT116 cell line, which corresponds to colon carcinoma. These analyses indicated good results for models trained with Pmel-noTum data on HCT116 cells (Table S9). The only exception was observed for H3K4me3 when testing on Pmel-Tum, which could be because the marks with the most significant relationships with it are not in this dataset. Importantly, this test on a different cell line validates the generalization capabilities of our approach, because it works largely independently of the cell line used.

Conclusion

Here, we describe a new approach based on RFs to predict histone marks from other histone marks. First, we proved that our

approach is robust and outperforms other available tools that perform the same task. Moreover, it is even more important that by using our tool, it is possible to accurately determine many histone marks. Increasing in this way the number of PTMs available from the same sample, thus allowing more analysis from the same experimental data. For instance, by combining our approach with 10 experimentally determined marks, we were able to accurately assign states to chromatin in a way that allowed to determine GRN based on active promoters (explained in Supplementary files 2 and 3). This GRN analysis, applied to a colorectal cancer cell line, was validated by genes whose expression is linked to this type of cancer.

Finally, it is important to highlight that in the process of creating a cascade of predictors, we found evidence to support the existence of a robust and redundant epigenetic code. In this code, we have shown how certain marks are accurately predicted by other marks, i.e. there are redundancies, and some other marks are indispensable to predict the existence of others, which indicates an interdependence between marks that also explains the large alterations in epigenetic profiles that are usually observed in complex diseases.

Our approach, based on a prediction algorithm, would benefit from the availability of more data from other samples or other types of information such as gene expression profiles or DNA accessibility. Nonetheless, our new approach is useful on its own, and will help other scientists in their work.

Key Points

- Our Random Forest-based approach identified relationships between histone post translational modifications, reducing the number of experimentally determined data required to assign chromatin states.
- The existence of different relationships between epigenetic marks provides evidence to support the existence of a redundant epigenetic code, even if specific variations are still significant.
- Our approach performance outperforms state-of-the-art methodologies for the predictions of histone PTMs.
- Analysis of GRNs based on chromatin states assigned from predicted histone marks on a cancer cell line confirmed the validity of our results.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of Interest

None declared

Author contributions statement

Formulation of research objectives and goals: L.M., G.P., A.M., and M.S.; Data curation and model training: L.M.; Result interpretation and analysis: L.M., E.R., A.M., and M.S.; Writing of manuscript, proofreading, and editing was carried out by all authors. All authors read and approved the final manuscript.

Funding

This work was funded by ANID PhD fellowship [21201856] to LM, FONDECYT Regular Projects [1181089, 1191526, 1231629], FONDECYT Inicio [11171015], Funding from Universidad Católica de Temuco Vice-Rectorate for Research [2023FIAS-MS-02, 2023FEQUIP-RB-01, 2024GI-AH-03] and Centro Ciencia & Vida, FB210008, Financiamiento Basal para Centros Científicos y Tecnológicos de Excelencia de ANID. PoweredNLHPC: this research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02).

References

- Phillips T, Shaw K. Chromatin remodeling in eukaryotes. *Nat Educ* 2008; **1**:209.
- Barth T, Imhof A. Fast signals and slow marks: the dynamics of histone modifications. *Trends Biochem Sci* 2010; **35**:618–26. <https://doi.org/10.1016/j.tibs.2010.05.006>.
- Chambeyron S, Bickmore W. Chromatin decondensation and nuclear reorganization of the HOXB locus upon induction of transcription. *Genes Dev* 2004; **18**:1119–30. <https://doi.org/10.1101/gad.292104>.
- Bartova E, Krejci J, Harnicarova A. et al. Histone modifications and nuclear architecture: a review. *J Histochem Cytochem* 2008; **56**:711–21. <https://doi.org/10.1369/jhc.2008.951251>.
- Strahl BD, Allis CD. The language of covalent histone modifications. *Nature* 2000; **403**:41–5. <https://doi.org/10.1038/47412>.
- Jenuwein T, Allis CD. Translating the histone code. *Science* 2001; **293**:1074–80. <https://doi.org/10.1126/science.1063127>.
- Turner BM. Defining an epigenetic code. *Science* 2007; **9**:2–6. <https://doi.org/10.1038/ncb0107-2>.
- Kharchenko PV, Alekseyenko AA, Schwartz YB. et al. Comprehensive analysis of the chromatin landscape in drosophila melanogaster. *Nature* 2011; **471**:480–5. <https://doi.org/10.1038/nature09725>.
- Grewal SIS, Moazed D. Heterochromatin and epigenetic control of gene expression. *Science* 2003; **301**:798–802. <https://doi.org/10.1126/science.1086887>.
- Baker M. Making sense of chromatin states. *Nat Methods* 2011; **8**:717–22. <https://doi.org/10.1038/nmeth.1673>.
- Filion GJ, van Bommel JG, Braunschweig U. et al. Systematic protein location mapping reveals five principal chromatin types in drosophila cells. *Cell* 2010; **143**:212–24. <https://doi.org/10.1016/j.cell.2010.09.009>.
- Julienne H, Zoufir A, Audit B. et al. Human genome replication proceeds through four chromatin states. *PLoS Comput Biol* 2013; **9**:e1003233. <https://doi.org/10.1371/journal.pcbi.1003233>.
- Marco E, Meuleman W, Huang J. et al. Multi-scale chromatin state annotation using a hierarchical hidden Markov model. *Nat Commun* 2017; **8**:15011. <https://doi.org/10.1038/ncomms15011>.
- Ernst J, Kellis M. ChromHMM: Automating chromatin-state discovery and characterization. *Nat Methods* 2012; **9**:215–6. <https://doi.org/10.1038/nmeth.1906>.
- Ernst J, Kellis M. Chromatin-state discovery and genome annotation with chromHMM. *Nat Protoc* 2017; **12**:2478–92. <https://doi.org/10.1038/nprot.2017.124>.
- Mirabella AC, Foster BM, Bartke T. Chromatin deregulation in disease. *Chromosoma* 2016; **125**:75–93. <https://doi.org/10.1007/s00412-015-0530-0>.
- Moosavi A, Motevalizadeh A, Ardekani. Role of epigenetics in biology and human diseases. *Iran Biomed J* 2016; **20**:246–58.
- Soler-Botija C, Galvez-Montan C, Bayas-Genas. Epigenetic biomarkers in cardiovascular diseases. *Front Genet* 2019; **10**:950. <https://doi.org/10.3389/fgene.2019.00950>.
- Zhang W, Song M, Qu J. et al. Epigenetic modifications in cardiovascular aging and diseases. *Circ Res* 2018; **123**:773–86. <https://doi.org/10.1161/CIRCRESAHA.118.312497>.
- Grady WM, Carethers JM. Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology* 2008; **135**:1079–99. <https://doi.org/10.1053/j.gastro.2008.07.076>.
- Benard A, Goossens-Beumer IJ, van Hoesel AQ. et al. Histone trimethylation at H3K4, H3K9 and H4K20 correlates with patient survival and tumor recurrence in early-stage colon cancer. *BMC Cancer* 2014; **14**:531. <https://doi.org/10.1186/1471-2407-14-531>.
- Fiziev P, Akdemir KC, Miller JP. et al. Systematic epigenomic analysis reveals chromatin states associated with melanoma progression. *Cell Rep* 2017; **19**:875–89. <https://doi.org/10.1016/j.celrep.2017.03.078>.
- Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet* 2015; **16**:321–32. <https://doi.org/10.1038/nrg3920>.
- Ohler U, Liao G, Niemann H. et al. Computational analysis of core promoters in the drosophila genome. *Genome Biol* 2002; **3**:RESEARCH0087. <https://doi.org/10.1186/gb-2002-3-12-research0087>.
- Chen Y, Li Y, Narayan R. et al. Gene expression inference with deep learning. *Bioinformatics* 2016; **32**:1832–9. <https://doi.org/10.1093/bioinformatics/btw074>.
- Chen C, Hou J, Shi X. et al. DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks. *BMC Bioinformatics* 2021; **22**:38. <https://doi.org/10.1186/s12859-020-03952-1>.
- Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* 2015; **33**:364–76. <https://doi.org/10.1038/nbt.3157>.
- Schreiber J, Durham T, Bilmes J. et al. Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome Biol* 2020; **21**:81. <https://doi.org/10.1186/s13059-020-01977-6>.
- Stelzer G, Rosen N, Plaschkes I. et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr Protoc Bioinform* 2016; **54**:1.30.1–1.30.33. <https://doi.org/10.1002/cpbi.5>.
- Howe KL, Achuthan P, Allen J. et al. Ensembl 2021. *Nucleic Acids Res* 2020; **49**:D884–91. <https://doi.org/10.1093/nar/gkaa942>.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**:357–9. <https://doi.org/10.1038/nmeth.1923>.
- Li H, Handsaker B, Wysoker A. et al. The sequence alignment/map format and samtools. *Bioinformatics* 2009; **25**:2078–9. <https://doi.org/10.1093/bioinformatics/btp352>.
- Ramírez F, Ryan DP, Grüning B. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 2016; **44**:W160–5. <https://doi.org/10.1093/nar/gkw257>.
- Davis CA, Hitz BC, Sloan CA. et al. The Encyclopedia of DNA Elements (ENCODE): data portal update. *Nucleic Acids Res* 2017; **46**:D794–801.
- Pedregosa F, Varoquaux G, Gramfort A. et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; **12**:2825–30.
- Jiang S, Mortazavi A. Integrating chip-seq with other functional genomics data. *Brief Funct Genomics* 2018; **17**:104–15. <https://doi.org/10.1093/bfpg/ely002>.

37. Robinson JT, Thorvaldsdóttir T, Winckler W. et al. Integrative genomics viewer. *Nature* 2011; **29**:24–6. <https://doi.org/10.1038/nbt.1754>.
38. Audia JE, Campbell RM. Histone modifications and cancer. *Cold Spring Harb Perspect Biol* 2016; **8**:a019521–1. <https://doi.org/10.1101/cshperspect.a019521>.
39. Kimura H. Histone modifications for human epigenome analysis. *Pathol Oncol Res* 2020; **26**:2023–33. <https://doi.org/10.1007/s12253-019-00663-8>.
40. Qin J, Wen B, Liang Y. et al. Histone modifications and their role in colorectal cancer (review). *J Hum Genet* 2013; **58**:439–45.
41. Shogren-Knaak M, Ishii H, Sun JM. et al. Histone h4-k16 acetylation controls chromatin structure and protein interactions. *Science* 2006; **311**:844–7. <https://doi.org/10.1126/science.1124000>.
42. Liu L, Jin G, Zhou X. Modeling the relationship of epigenetic modifications to transcription factor binding. *Nucleic Acids Res* 2015; **43**:3873–85. <https://doi.org/10.1093/nar/gkv255>.
43. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* 2010; **28**:817–25. <https://doi.org/10.1038/nbt.1662>.
44. Lee J, Smith E, Shilatifard A. The language of histone crosstalk. *Cell* 2010; **142**:682–5. <https://doi.org/10.1016/j.cell.2010.08.011>.
45. Zhang T, Cooper S, Brockdorff N. The interplay of histone modifications - writers that read. *EMBO Rep* 2015; **16**:1467–81. <https://doi.org/10.15252/embr.201540945>.
46. Bernstein BE, Mikkelsen TS, Xie X. et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 2006; **125**:315–26. <https://doi.org/10.1016/j.cell.2006.02.041>.
47. Iwagawa T, Watanabe S. Molecular mechanisms of H3K27me3 and H3K4me3 in retinal development. *Neurosci Res* 2019; **138**:43–8. <https://doi.org/10.1016/j.neures.2018.09.010>.
48. Liu X, Wang C, Liu W. et al. Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. *Nature* 2016; **537**:558–62.
49. Yue Y, Li X, Jiao R. et al. H3K27me3-H3K4me1 transition at bivalent promoters instructs lineage specification in development. *Cell Biosci* 2023; **13**:66–20. <https://doi.org/10.1186/s13578-023-01017-3>.
50. Paul D. The systemic hallmarks of cancer. *J Cancer Metastasis Treat* 2020; **2020**:29. <https://doi.org/10.20517/2394-4722.2020.63>.
51. Garnis C, Buys TPH, Lam WL. Genetic alteration and gene expression modulation during cancer progression. *Mol Cancer* 2004; **3**:9. <https://doi.org/10.1186/1476-4598-3-9>.

A. Additional file 1—Supplementary tables

Multi-page table in excel format containing Supplementary Tables S1–S14.

B. Additional file 2—GRN analysis of colorectal cancer and Astrocyte samples

This file contains our analysis of the colorectal cancer and astrocyte cell line that has 10 PTMs available previously used as test set for the cascade predictor.

C. Additional file 3—GRN whole genome

GRN whole genome of the colorectal cancer cell line of case of study. This file is a Cytoscape session, to view is necessary load in a Cytoscape 3.0. The GRN was created using the whole genome data, which consists of 2248 nodes and 61 526 connections. Of the total number of nodes, 64 correspond to TFs. Additionally, those target genes that are related to colorectal cancer were identified. The CancerGeneticsWeb and DisGeNet databases were used for this. In this way, 114 genes were obtained, which according to the literature, are associated with colorectal cancer (Table S14).

D. Additional file 4—GRN chr1 Astrocytes

GRN chromosome 1 cell line of Astrocytes for case of study. This file is a Cytoscape session, to view is necessary load in a Cytoscape 3.0. The GRN was created using the chr1 data, which consists of 221 nodes and 2581 connections. Of the total number of nodes, 36 correspond to TFs.