HebbPlot: An intelligent tool for learning and visualizing chromatin mark signatures

Hani Z. Girgis[1], Alfredo Velasco II[1]

1 Tandy School of Computer Science, University of Tulsa, Tulsa, OK, USA

* hani-girgis@utulsa.edu

Abstract

Histone modifications play important roles in gene regulation, heredity, imprinting, and many human diseases including diabetes, obesity, and cancer. The histone code is complex and consists of more than 100 marks. Therefore, biologists need computational tools to characterize general signatures representing the distributions of tens of chromatin marks around thousands of regions. To this end, we developed a software tool called HebbPlot, which utilizes a Hebb neural network in learning a general chromatin signature from regions with a common function. Hebb networks can learn the associations between tens of marks and thousands of regions. This is the first application of Hebb networks in the epigenetics field. HebbPlot presents a signature as a digitized image, in which a bright pixel indicates the presence of a mark around a part of the genetic element, and a black pixel indicates the absence of the mark. A row of pixels represents one mark. Similar rows are clustered in the image. We validated HebbPlot on synthetic data and on 111 epigenomes provided by the Roadmap Epigenomics Project. HebbPlot was able to retrieve distinct chromatin signatures for promoters, enhancers, and genes active in each of the 111 cell types. Our analysis reveals that active promoters have a directional signature; marks such as H3K79(me1/me2), H3K4(me1,me2,me3), and H3K9ac stretch toward coding regions. The plots of inactive promoters show that H3K27me3 is consistently present around them. Further, the signatures of enhancers that are fully included in repetitive regions are almost identical to those located outside repeats, indicating that transposons have an enhancer-like function in the human genome. Furthermore, the chromatin signature of active elements consists of the presence of H3K79me1 and the absence of H3K9me3 and H3K27me3. In sum, HebbPlot is a general tool that can be applied to wide array of studies, facilitating the deciphering of the histone code.

Author summary

Chromatin marks have gained much attention because of their important roles in gene regulation, cell differentiation, Lamarckian inheritance, and imprinting. A chromatin signature of a genetic element, such as genes or enhancers, consists of multiple marks and may differ from a tissue to a tissue. Currently, tens of histone modifications are known. Several marks of more than 100 human cell types have been determined. Many epigenomes of other normal and pathological cell types will be available soon. Extracting a chromatin signature representing the distributions of tens of marks around thousands of regions is a challenging task. Hebb networks are a special type of artificial neural networks known for their ability to learn associations. We developed a software tool called HebbPlot. The tool uses a Hebb network to learn how a mark is distributed around a set of regions that have the same function, e.g. promoters active in the same tissue. HebbPlot produces a pattern representing mark distributions around all of the

regions. Mark patterns are clustered based on their similarity to one another. Then a digitized image representing the learned pattern is generated. HebbPlot will help biologist with characterizing and visualizing chromatin signatures in numerous studies.

Introduction

Understanding the effects of histone modifications will provide answers to important questions in biology and will help with finding cures to several diseases including cancer. Carey highlights several functions of epigenetic factors including Cytosine methylation and histone modifications [1]. It was reported that methylation of CpG islands inhibit transcription [2], whereas the complex histone code has a wide range of regulatory functions [3, 4]. Additionally, epigenetic marks may affect body weight and metabolism [5]. Interestingly, chromatin marks may explain how some acquired traits, such as obesity and exposure to some toxins, are passed from one generation to the next (Lamarckian inheritance) [6–9]. Further, epigenetics may explain how two identical twins have different disease susceptibilities [10]. Epigenetic factors play a role in imprinting, in which a chromosome, or a part of it, carries a maternal or a paternal mark(s) [11, 12]. Defects in the imprinting process may lead to several disorders [13–18], and may increase the "birth defects" rate of assisted reproduction [19]. Furthermore, chromatin marks play a role in cell differentiation by selectively activating and deactivating certain genes [20, 21]. Some chromatin marks take part in deactivating one of the X chromosomes [22]. It has been observed in multiple types of cancer that some tumor suppressor genes were deactivated by hypermethylating their promoters [23–25], the removal of activating chromatin marks [26, 27], or adding repressive chromatin marks [28]. Anti-cancer drugs that target the epigenome [1] have been designed. Two compounds are used in these drugs. One compound inhibits DNA methylation [29, 30], whereas the other compound inhibits histone deacetylation [31] (histone acetylation is an activating mark).

Pioneering computational and statistical methods for deciphering the histone code have been developed. Some tools are designed for profiling and visualizing the distribution of a chromatin mark(s) around multiple regions [32, 33]. Additionally, a tool for clustering and visualizing genomic regions based on their chromatin marks has been developed [34]. Several systems are available for characterizing histone codes/states in an epigenome [35–43]. Further, an alphabet system for histone codes was proposed [44]. Other tools can recognize and classify the chromatin signature associated with a specific genetic element [**?**, 45–54]. Furthermore, methods that compare the chromatin signature of healthy and sick individuals have been proposed [55].

Scientists have identified about 100 histone marks [37]. Additionally, there is a near infinite number of future studies, in which scientists need to characterize the pattern of chromatin marks around a set of regions in the genome. Therefore, there is a definite need for an automated framework that enables scientists to (i) automatically characterize the chromatin signature of a set of sequences that have a common function, e.g. exons, promoters, or enhancers; and (ii) visualize the identified signature in a simple intuitive form. To meet this need, we designed and developed a software tool called HebbPlot. This tool allows average users, without extensive computational knowledge, to characterize and visualize the chromatin signature associated with a genetic element automatically.

HebbPlot includes the following four innovative approaches in an area that has become the frontier of medicine and biology:

- **HebbPlot can learn the chromatin signature of a set of regions automatically.** Sequences that have the same function in a specific cell type, e.g.

exons, promoters, or enhancers, are expected to have similar marks. The learned signature represents these marks around all of the regions. *HebbPlot differs from the other tools in its ability to learn one signature representing the distributions of all available chromatin marks around thousands of regions.*

- **This is the first application of Hebb neural networks in the epigenetics field.** These networks are capable of learning associations; therefore, they are well suited for learning the associations among tens of marks and genetic elements.
- **The framework enables average users to train artificial neural networks *automatically*.** Users are not burdened with the training process. Self-trained systems for analyzing protein structures and sequence data have been proposed [56–58]. HebbPlot is the analogous system for analyzing chromatin marks.
- **HebbPlot is the first system that integrates the tasks of learning and visualizing a chromatin signature.** Once the signature is learned, the marks are clustered and displayed as a digitized image. This image shows one pattern representing thousands of regions. To illustrate, the distributions of the marks appear around one region; however, they are learned from all input regions.

We have applied our tool to learning and visualizing the chromatin signatures of several active and inactive genetic elements in the 111 consolidated epigenomes provided by the Roadmap Epigenomics Project. These case studies demonstrate the applicability of HebbPlot to many interesting problems in molecular biology, facilitating the deciphering of the histone code.

Materials and methods

Methods

In this section, we describe the computational principles of our software tool, HebbPlot. The core of the tool is an unsupervised neural network known as Hebb network.

Region representation

To represent a group of histone marks overlapping a region, these marks are arranged according to their genomic locations on top of each other and the region. Then equally-spaced vertical lines are superimposed on the stack of the marks and the region. The numerical representation of this group of marks is a matrix. A row of the matrix represents a mark. A column of the matrix represents a vertical line. If the i^{th} mark intersects the j^{th} vertical line, the entry i and j in the matrix is 1, otherwise it is -1. The first vertical line is at the beginning of the region; the last vertical line is at the end of the region. The rest of the lines are spread out evenly. Fig 1 shows the graphical and the numerical representations of a region and the overlapping marks. Finally, the two-dimensional matrix is converted to a one dimensional vector called the epigenetic vector. The number of vertical lines is determined experimentally. We used 41 and 101 lines in our experiments. This number should be adjusted according to the average size of a region.

Data preprocessing

Preprocessing input data is a standard procedure in machine learning. During this procedure, the noise in the input data is reduced. Each epigenetics vector is compared to two other vectors selected randomly from the same set. The value of an entry in the

Fig 1. Representations of a group of chromatin marks overlapping a region. (a) Horizontal double lines represent a region of interest. Horizontal single lines represent the marks. Vertical lines are spaced equally and bounded by the region. (b) The intersections between the marks and the vertical lines are encoded as a matrix where rows represent the marks and columns represent the vertical lines. If a vertical line intersects a mark, the corresponding element in the matrix is 1, otherwise it is -1.

vector is kept if it is the same in the three vectors, otherwise it is set to zero. For example, consider the vector [1 1 -1]. Suppose that the vectors [1 -1 -1] and [1 -1 -1] were selected randomly. The preprocessed vector would be [1 0 -1] because the first and the third elements are the same in the three vectors, but the second element is not.

Hebb recall network

Associative learning, also known as Hebbian learning, is inspired by biology. "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased" [59]. In behavioral psychology, Ivan Pavlov conducted a famous experiment, which demonstrated learning by association. In this experiment, a dog was trained to associate the sound of a bell with food; this dog salivated when it heard the bell whether or not food was present. The presence of food is referred to as the unconditioned stimulus, p^0, and the sound of the bell is referred to as the conditioned stimulus, p. Associating these two stimuli together is the goal. After training, the response to either the conditioned stimulus or the unconditioned one is the same as the response to both stimuli combined [60].

In the context of epigenetics, a Hebb network can be viewed as the dog in Pavlov's experiment. The unconditioned stimulus, p^0, is a one-dimensional vector representing the distributions of histone marks over a sequence e.g. one tissue-specific enhancer. This vector is referred to as the epigenetic vector; it is obtained as outlined above. The conditioned stimulus is always the one vector, which include ones in all entries. We would like to train the network to give a response, analogous to the salivation of the dog, when it is given the ones vector, whether or not the epigenetic vector is provided. The response of the network is a prototype/signature representing the distributions of histone marks over the entire set of genomic locations, e.g. all enhancers of a specific tissue.

Eq 1 and Eq 2 define how the response of a Hebb network is calculated. The training of the network is given by Eq 3 [60].

$$satlins(x) = \begin{cases} +1 & \text{if } x \geq 1 \\ x & \text{if } -1 < x < 1 \\ -1 & \text{if } x \leq -1 \end{cases} \quad (1)$$

Eq 1 defines a transformation function. This function ensures that the response of the network is similar to the unconditioned stimulus, i.e. each element of the response is between 1 and -1. If x is a vector, the function is applied component wise.

$$a(p^0, w, p) = satlins(p^0 + w \odot p) \quad (2)$$

Eq 2 describes how a Hebb network responds to the two stimuli. The response of the network is transformed using Eq 1. In Eq 2, p^0 is the unconditioned stimulus, e.g. presence of food or an epigenetic vector; w is the weights vector, which is the prototype/signature learned so far; and p is the conditioned stimulus, e.g. sound of a bell or the one vector. The operator $\odot$ represents the component-wise multiplication of two vectors. In the current adaptation, if the network is presented with an epigenetic

vector and the one vector, the response is the sum of the prototype learned so far and the epigenetic vector. In the absence of the epigenetic vector, i.e. all-zeros p^0, the response of the network is the prototype, demonstrating the ability of the network to learn associations.

$$w_i = w_{i-1} + \alpha(a(p_i^0, w_{i-1}, p_i) - w_{i-1}) \odot p_i \tag{3}$$

Eq 3 defines Hebb's unsupervised learning rule. Here, w_i and w_{i-1} are the prototype vectors learned in iterations i and $i-1$. The i^{th} pair of unconditioned and conditioned stimuli is p_i^0 and p_i. Learning occurs, i.e. the prototype changes, only when the i^{th} conditioned stimulus, p_i, has non-zero components. This is the case in our adaptation because p_i is always the one vector. Due to a small α, which represents the learning and the decay rates, the prototype vector changes a little bit in each iteration when learning occurs; it moves closer to the response of the network to the i^{th} pair of stimuli.

Comparing two signatures

Two signatures can be compared numerically. The dot product of two victors indicates how close they are to each other in space. When these vectors are normalized, i.e. each element is divided by the vector norm, the dot product is between 1 and -1. The dotsim function (Eq 4) normalizes the vectors and calculates their dot product.

$$dotsim(x, y) = \frac{x}{\|x\|} \cdot \frac{y}{\|y\|} \tag{4}$$

Here, x and y are vector; $\|x\|$ and $\|y\|$ are the norms of these vectors; the $\cdot$ symbol is the dot product operator.

It is easy to interpret the meaning of the dot product of two normalized vectors. If the two vectors are very similar to each other, the value of the dotsim function approaches 1. If the values at the same index of the two vectors are opposite of each other, i.e. 1 and -1, the value of dotsim approaches -1. The dotsim function can be applied to the whole epigenetic vector or to the part representing a specific chromatin mark. When comparing the chromatin signatures of two sets of regions, a mark with a dotsim value approaching 1 is common in the two signatures. A mark with a dotsim value approaching -1 has opposite distributions, distinguishing the signatures. Marks with dotsim values approaching zero do not have consistent distribution(s) in one or both sets; these marks should not be considered while comparing the two signatures.

Visualizing a chromatin signature

Row vectors representing different marks are clustered according to their similarity to each other. We used hierarchical clustering in grouping marks with similar distributions. Hierarchical clustering is an iterative bottom-up approach, in which the closest two items/groups are merged at each iteration. The algorithm requires a pair-wise distance function and a cluster-wise distance function. For the pair-wise distance function, we utilized the city block function to determine the distance between two vectors representing marks. For the group-wise distance function, we applied the weighted pair group method with arithmetic mean [61]. To determine the group-wise distance between a cluster A, and another cluster consisting of two sub-clusters B and C, add the distance between A and B to the distance between A and C; then divide the sum by 2. We utilized the implementation of hierarchical clustering provided in the Statistics and Machine Learning Toolbox of Matlab (R2017A) by MathWorks.

A digitized image represents the chromatin signature of a genetic element. A one-unit-by-one-unit square in the image represents an entry in the matrix representing

the signature. A row of these squares represents one mark. The color of a square is a shade of gray if the entry value is less than 1 and greater than -1; the closer the value to 1 (-1), the closer its color to white (black).

Up to this point, we illustrated the computational principles of our software tool, HebbPlot. Next, we provide the details of the data used in validating the tool.

Data

We used HebbPlot in extracting and visualizing chromatin signatures characterizing multiple genetic elements of the 111 consolidated epigenomes of the Roadmap Epigenomics Project [62]. Specifically, we applied HebbPlot to:

1. Active promoters.
2. Active promoters on the positive strand.
3. Active promoters on the negative strand.
4. Inactive promoters.
5. Active enhancers.
6. Active repetitive enhancers.
7. Active non-repetitive enhancers.
8. Inactive enhancers.
9. Coding regions of active genes.
10. Coding regions of inactive genes.
11. Random genomic locations.

We obtained the genomic locations of the putative promoters specific to each of the 111 consolidated epigenomes from the Roadmap Epigenomics Project (`http://egg2.wustl.edu/roadmap/data/byDataType/dnase/BED_files_prom/`). These promoters were predicted using DNase I hypersensitive sites and chromatin states characterizing active promoters. To obtain the inactive promoters, we performed the following two steps: (i) all tissue-specific promoters are collected and merged if overlapping and (ii) all promoters are compared to the tissue-specific promoters; for each tissue, promoters that do not overlap with the tissue-specific promoters are considered inactive in this tissue. To compare the chromatin signatures of promoters on the positive and the negative strands, we separated the promoters according to the strand. If a putative promoter overlaps a transcription start site on the positive strand only, it is considered positive and vice versa. Each group was sorted and overlapping regions, if any, were merged.

The putative enhancers were obtained from the Roadmap Epigenomics Project (`http://egg2.wustl.edu/roadmap/data/byDataType/dnase/BED_files_enh/`). The inactive enhancers were obtained using the same procedure applied in obtaining the inactive promoters. Later in this paper, we compare the chromatin signature of putative enhancers overlapping with repeats to that of the non-overlapping ones. The hg19 human assembly repeats (`http://www.repeatmasker.org/species/hg.html`), including transposons and simple tandem repeats, were used for determining repetitive enhancers. In order for an enhancer to be considered repetitive, it must be entirely included in a repetitive region. In another experiment, we considered an enhancer to be repetitive if at least half of its sequence overlaps a repetitive region.

The coding regions were obtained from the University of California Santa Cruz Genome Browser (`http://genome.ucsc.edu`). The Ensemble genes for the hg19 human genome assembly were used in this study. Active genes in a tissue are defined as those that their transcription start sites overlap with the tissue-specific putative promoters. Otherwise, they are considered inactive. After that, coding regions of the active (or the inactive) genes in a tissue are collected and merged if overlapping.

Regarding the random genomic locations, we sampled uniformly 500 regions from each chromosome of the human genome. Each region is 1000 base pairs (bp) long. For each of the 111 consolidated epigenomes, chromatin marks overlapping with the random locations were obtained.

If the number of the regions, e.g. tissue-specific enhancers, was more than 10,000 regions, we sampled uniformly 500 regions from each chromosome.

In this section, we discussed the computational method and the data used in the validation experiments. In the next section, we validate HebbPlot on synthetic and real data.

Results and Discussion

HebbPlot

We invented a new software tool called HebbPlot. HebbPlot has the following two specific aims: (i) learning automatically the chromatin signature of a group of genomic locations that have a common function, and (ii) representing this signature as a digitized image that is easily interpreted. The core of HebbPlot is a Hebb neural network. Hebb networks are known for their ability to learn associations, making them well suited for learning the chromatin signatures of genetic elements. *To the best of our knowledge, this is the first application of Hebb networks in the field of epigenetics.* The training process of the neural network is fully automated, enabling biologists without extensive computational knowledge to take advantage of advanced machine learning algorithms such as Hebb networks. The tool is general and can be applied to any set of genomic locations. HebbPlot is freely available to the academic community. It can be found at Software S1.

Results on synthetic data

Consider a step-pyramidal shape (Figure 2). One thousand noisy instances of this shape were generated by randomly shifting a step of the pyramid to the right or to the left by at most 200 units. A step may be deleted with a probability of 0.2. Each shape is represented by a matrix, in which an entry has a value of 1 (white) or -1 (black). To obtain this matrix, a group of evenly-spaced vertical lines are superimposed on the shape. If a line intersects a step of the pyramid, the corresponding entry in the matrix is 1. Otherwise, it is -1. More details about representing a shape are given under the Materials and Methods Section.

As a baseline, the original shape was retrieved from the noisy instances by a simple majority voting scheme. In this scheme, an entry of the prototype matrix is assigned 1 if the majority of the values stored in same entry of the 1000 matrices are 1; otherwise, it is assigned -1. The prototype due to this method is similar to the original shape; however, its boundaries are inaccurate, whereas the prototype retrieved by the Hebb network looks very similar to the original shape. The boundaries of the steps are accurate; however, they are fuzzy. Similar results were obtained when this experiment was repeated multiple times using higher and lower mutation rates (shift amount: 0-300,

step-deletion probability: 0-0.3), demonstrating the ability of Hebb networks to retrieve the original shape successfully.

Fig 2. A controlled experiment demonstrating the ability of Hebb networks to retrieve a prototype from noisy shapes. (a) The real shape is a two-dimensional step pyramid. (b-d) Examples of noisy shapes produced by corrupting the real shape. The start and the end of each horizontal white segment may be shifted to the right or the left by up to 200 unites; additionally, a horizontal segment may be deleted with a probability of 0.2. (e) The shape retrieved by a majority-voting scheme. (f) The shape retrieved by the network.

Results on real data

Next, we studied multiple enhancers potentially active in the H1 cell line (embryonic stem cell) obtained from the Roadmap Epigenomics Project. These enhancers were predicted using DNase I Hypersensitive sites and chromatin states associated with enhancers. This data set contains 11,369 putative H1-specific enhancers and 27 chromatin marks. Each enhancer region was expanded by 10% on each end to study how chromatin marks differ from/resemble the surrounding regions. To begin, 41 uniform samples/points were obtained from each region. Then for each point, it was determined whether or not it falls in a mark region overlapping the putative enhancer.

Next, we plotted the results as shown in Fig. 3. No clear signature appears in these plots. After that, we used the majority-voting scheme described earlier and HebbPlot in generating the signature of the H1-specific enhancers. The figure generated by HebbPlot shows more information than the majority plot.

The Hebb plot shows four distinct zones representing the absent marks, and the present ones with different confidence levels. For example, the top zone shows marks that are absent from the H1-specific enhancers. These marks include H2A.Z, H4K8ac, H3K9me3, H3K4me3, and H3K36me3. The bottom zone shows the marks that present around these enhancers with the highest confidence level. These marks include H3K4(me1,me2), H3K79(me1,me2), and many acetylation marks. In contrast, the plot due to the majority-voting scheme shows only two zones representing the absent and the present marks without confidence information.

Further, because the enhancer regions were expanded on each end by 10%, a present mark is expected to be brighter around the center of an enhancer than its peripheries. The Hebb plot shows such information, whereas the brightness of the present marks is uniform around almost all marks shown in the majority plot. These results show that a Hebb plot is more accurate and shows more information than a plot generated by the majority-voting scheme.

Fig 3. Retrieving the chromatin signature of the H1-specific enhancers. The data set consists of 11,369 putative enhancers active in H1. Four examples of potential enhancers are shown in parts a-d. It is hard to see a common pattern in these four examples. The signature retrieved by the majority-voting scheme has two zones (part e), whereas the signature learned by the Hebb network is characterized by four zones (part f). The top most zone represents chromatin marks that are absent from the enhancer regions, whereas the next three zones represent the present marks with increasing certainty. The signature retrieved by the majority-voting scheme does not show confidence indicators of the absence or the presence of a mark. For example the third zone from the top of the Hebb plot is not as strong as the forth zone. In contrast, such information does not appear in the signature retrieved by the majority-voting scheme. Enhancer regions were expanded by 10% on each end. Therefore, the intensity of the signal is expected to be weaker at the peripheries than around the center of a signature. Again, the Hebb plot shows such information, whereas the signature retrieved by the majority-voting scheme does not.

The distinct chromatin signatures of different active elements

Twenty eight chromatin marks of the IMR-90 (fetal lung fibroblasts cell Line) epigenome are available through the Roadmap Epigenomics Project. The project provides access to predicted enhancers and promoters specific to IMR-90. We sampled 11,268 enhancers, 13,226 promoters, and 11,390 coding regions of active genes in IMR-90. About 500 regions were uniformly sampled from each chromosome. In addition, we selected 10,000 locations sampled uniformly from all chromosomes of the human genome. Then we trained four Hebb networks to learn the chromatin signature of each genetic element.

Fig 4 shows the four Hebb plots. The promoter signature is characterized by a bright box that is clearly different from the surrounding regions. The center, where the transcription start sites are located, of the upper part of the box is less bright than its peripheries. With regard to the chromatin signature of the enhancers, it is characterized by multiple zones. Each zone has consistent brightness. The brightest zone at the bottom of the Hebb plot is the widest. Similarly, the coding regions signature is multi-zonal; however, the brightest zone is the narrowest and the middle gray zone is the widest. Chromatin marks should not be distributed in a consistent manner around regions that do not have a common function. As expected, the Hebb plot representing the random genomic locations displays a black box, indicating that no chromatin mark is distributed consistently around these regions.

After that, we repeated the same experiment on each of the 111 epigenomes of the Roadmap Epigenomics Project. The Hebb plots of the promoters, the enhancers, and the coding regions of active genes are available through Data set S1, Data set S2, and Data set S3. The four distinct signatures are consistent across all tissue types.

These plots demonstrate that HebbPlot is able to learn the chromatin signature from a group of regions with the same function. In addition, the chromatin signatures of the promoters, the enhancers, and the coding regions are clearly distinct.

Fig 4. IMR-90 chromatin signatures representing (a) active promoters, (b) active enhancers, (c) coding regions of active genes, and (d) random genomic locations. The boundaries of these elements were expanded by 10% on each end to show a chromatin signature in contrast to the surrounding regions. Active promoters have a unique signature characterized by a bright box that clearly differs from the background. In addition, the center of the upper half of the box has a less bright area around the transcription start sites than its peripheries. Enhancers specific to IMR-90 has a zonal signature, where each zone has consistent brightness. The enhancer signature has a wide bright zone. Genes active in IMR-90 has a zonal signature as well. However, the middle gray zone is the widest, and the brightest zone at the bottom is the narrowest. The random locations do not have a common function; therefore, chromatin marks around them should not be distributed consistently. As expected, HebbPlot did not retrieve any pattern as displayed by a black box.

The directional signature of active promoters

Because promoters are upstream from their genes, some marks may indicate the direction of the transcription. To determine whether or not marks have direction, tissue-specific putative promoters were separated according to the positive and the negative strands into two groups. Then the promoter region was expanded to include two equal-size regions upstream and downstream from the promoter region. Thus, the expanded region has these three equal-size parts: (i) the region upstream from the promoter, (ii) the promoter region itself, and (iii) the region downstream from the promoter. We trained two Hebb networks to learn the chromatin signatures of

tissue-specific promoters on the positive and the negative strands. Fig. 5 shows the Hebb plots of the positive and the negative promoters active in H1 and male skeletal muscle. The two plots of the promoters on the positive and the negative strands are mirror images of each other, indicating that multiple marks are distributed in a directional manner; some marks tend to stretch more downstream (bright) than upstream (dark).

Fig 5. Hebb plots of active promoters on the positive and the negative strands. Multiple chromatin marks are distributed in a direction specific way. These marks tend to stretch downstream from the promoters toward the coding regions. Examples are H3K4(me1,me2,me3) and H3K79(me1,me2). The two Hebb plots of the promoters on the positive and the negative strands are mirror images of each other.

Next, we generated Hebb plots for the positive (Data set S4) and the negative (Data set S5) promoters of all tissues available through the Roadmap Epigenomics Project. This phenomenon was very consistent in all tissues.

Recall that two vectors pointing in opposite directions have a dotsim value of -1. The closer the value to -1 is, the closer the angle between the two vectors to 180° is. To determine directional marks, the learned prototype of a mark over the upstream part of the expanded promoter region was compared to the prototype of the same mark over the downstream part. If the dotsim value between the two prototypes is -0.5 or lower, this mark is considered directional.

We list the number of times a chromatin mark was determined for a tissue and the number of times it showed directional preference in Table 1. The Roadmap Epigenomics Project did not determine all marks for the 111 tissues. We found that H3K79(me1/me2), H3K4(me1,me2,me3), and H3K9ac are extended toward the coding regions in 50% or more of the tissues, in which they are known. *These results show that active promoters have a directional chromatin signature.*

Table 1. Directional chromatin marks. Six chromatin marks around the promoters extend toward coding regions.

	Positive Strand			**Negative Strand**		
Mark	Known	Directional	Ratio	Known	Directional	Ratio
H3K79me2	5	5	1	5	5	1
H3K79me1	7	5	0.71	7	5	0.71
H3K4me1	111	75	0.68	111	76	0.68
H3K4me2	8	5	0.63	8	5	0.63
H3K4me3	111	51	0.46	111	53	0.48
H3K9ac	47	22	0.47	47	22	0.47

Promoters were separated according to the strand to positive and negative groups. Then the region of a promoter was expanded 100% on each end. Mark vectors over the upstream and the downstream thirds of the expanded regions were compared. A mark is considered directional if these two vectors are opposite to one another (a dotsim value of -0.5 or lower). Not all marks were determined for all tissues. The number of tissues, for which a mark was determined, is listed under the column titled "Known." The number of tissues, in which a mark has directional preference around the promoter regions, is listed under the column titled "Directional." The ratio of these two numbers are listed under the column labeled with "Ratio."

The chromatin signatures of repetitive and non-repetitive enhancers

It has been reported that transposon subfamilies have an enhancer-like function in the human genome [63]. Further, transposons are known to act as enhancers in plant genomes [64–68]. Given the availability of the putative enhancers of more than a hundred cell types, we asked two questions.

First, what is the percentage of enhancers that are located within repeat sequences, e.g. transposons? To answer this question, we calculated the percentage of the tissue-specific enhancers that are included entirely in repetitive regions. Interestingly, up to 25% of the tissue-specific enhancers are repetitive. The highest percentage of 25% was observed in the primary T helper cells PMA-I stimulated, and the lowest percentage of 12% was observed in the female fetal brain. If the overlap percentage between enhancers and repeats is lowered to 50% instead of 100%, the percentages of the repetitive tissue-specific enhancers range between 22% and 37% (see Table S1). These results indicate that a large portion of enhancers are repetitive.

Second, how similar/different are the chromatin signatures of the repetitive enhancers and the non-repetitive ones? To answer this question, we obtained two chromatin signatures by training a Hebb network on the repetitive enhancers (Data set S6) and another network on the non-repetitive enhancers (Data set S7) active in each tissue. Then, we compared the two chromatin signatures using the dotsim function. The two signatures are almost identical (mean = 0.98, standard deviation = 0.03, maximum=0.99, minimum=0.83); recall that the dotsim value obtained by comparing a signature to itself is 1 (see Table S2). As an example, Fig 6 shows the two Hebb plots of the repetitive and the non-repetitive enhancers active in IMR-90. The two Hebb plots are almost identical. *These results prove that the chromatin signature of the repetitive tissue-specific enhancers is identical to the signature of the non-repetitive enhancers, further supporting the enhancer-like function of transposons in the human genome.*

Fig 6. The repetitive and the non-repetitive enhancers active in the IMR-90 cell line have identical chromatin signatures. Enhancers that are fully included within repeat sequences, e.g. transposons, are considered repetitive. Enhancers that do not overlap with repeats are considered non-repetitive. The Hebb plots representing the chromatin signatures of the repetitive and the non-repetitive enhancers are almost identical. Further, a dotsim value of 0.9976 was obtained by comparing these signatures. Recall that a dotsim value of 1 is the result of comparing a signature to itself.

The signature of active elements

Next, we asked if there is a common code among active genetic elements. Specifically, what is the combination of marks absent or present around active promoters, active enhancers, and coding regions of active genes? To answer this question, we applied our software tool, HebbPlot, to three active elements in the 111 consolidated epigenomes. A mark is included in our analysis if it is known in at least 5 of the 111 epigenomes. We compared the distributions of the same mark around two active genetic elements using the dotsim function (see the Materials and Methods Section). Two distributions of a mark are considered similar if they have a dotsim value of 0.5 or higher in at least 50% of the tissues, in which this mark is known.

Table 2 shows the similar marks between (i) active promoters and active enhancers; (ii) active promoters and coding regions of active genes; and (iii) active enhancers and coding regions of active genes. These comparisons show that H3K79me1 is present with similar distributions around the three elements. Further, H3K9me3 and H3K27me3 are

absent from these elements. Previously, H3K79me1 is reported to be associated with gene expression [47], whereas the two absent marks are known to be repressive marks [69]. *These results imply that the chromatin signature of active elements consists of the presence of H3K79me1 and the absence of H3K9me3 and H3K27me3.* These three marks represent a basic signature, which may be expanded by studying other active elements and additional chromatin marks when they become available.

Table 2. Marks with similar distributions around two genetic elements.

	Mark	Known	Similar	Ratio
Promoters and enhancers	H2BK15ac	5	5	1.00
	H2BK5ac	7	7	1.00
	H3K14ac	6	6	1.00
	H3K18ac	7	7	1.00
	H4K8ac	7	7	1.00
	H3K9me3 (absent)	111	108	0.97
	H3K27me3 (absent)	111	105	0.95
	H3K4me2	8	7	0.88
	H3K36me3 (absent)	111	97	0.87
	H3K23ac	7	6	0.86
	H3K4ac	7	6	0.86
	H2BK12ac	6	5	0.83
	H4K91ac	6	5	0.83
	H3K79me2	5	4	0.80
	H3K4me1	111	84	0.76
	H2A.Z	7	5	0.71
	H2BK120ac	7	5	0.71
	H3K79me1	7	5	0.71
	H3K9ac	47	30	0.64
	H3K27ac	82	52	0.63
	H2AK5ac	7	4	0.57
Promoters and coding regions	H3K9me3 (absent)	111	110	0.99
	H3K27me3 (absent)	111	104	0.94
	H3K79me1	7	6	0.86
	H3K23ac	7	5	0.71
	H4K8ac	7	4	0.57
Enhancers and coding regions	H3K27me3 (absent)	111	110	0.99
	H3K9me3 (absent)	111	108	0.97
	H3K4me3 (absent)	111	95	0.86
	H3K79me1	7	6	0.86
	H2A.Z	7	4	0.57

The distributions of known marks in each of the 111 tissues were compared between (i) active promoters and active enhancers; (ii) active promoters and coding regions of active genes; and (iii) active enhancers and coding regions of active genes. The distributions of a mark over two genetic elements are considered similar if they have a *dotsim* value of 0.5 or higher. Recall that the *dotsim* values range between -1 and 1. The number of tissues, for which a mark was determined, is listed under the column titled "Known." The number of tissues, in which a mark has similar distributions around two genetic elements, is listed under the column titled "Similar." The ratio of these two numbers are listed under the column labeled with "Ratio."

Differences among the signatures of active elements

The figures generated by HebbPlot show that the signatures of active promoters, active enhancers, and coding regions of active genes are distinct. Additionally, the figures of the promoters and the enhancers appear more similar to one another than to the figure representing coding regions. In this analysis, we wanted to quantify the similarity/difference among these three elements by determining marks that are distributed differently.

We applied HebbPlot to the 111 epigenomes. Then we compared the distributions of the same mark around two genetic elements. The distributions of a mark around two genetic elements are considered opposite if they have a dotsim value of -0.5 or lower in at least 50% of the tissues, in which this mark is known.

Table 3 shows marks with different distributions between (i) active promoters and active enhancers; (ii) active promoters and coding regions of active genes; and (iii)

Table 3. Marks with opposite distributions around two genetic elements.

	Mark	Known	Opposite	Ratio
Promoters vs. Enhancers	H3K4me3	111	93	0.84
Promoters vs. Coding regions	H3K4me3	111	111	1.00
	H3K36me3	111	88	0.79
	H3K9ac	47	37	0.79
	H3K4me1	111	84	0.76
	H3K4me2	8	6	0.75
	H3K18ac	7	5	0.71
	H3K27ac	82	55	0.67
	H2A.Z	7	4	0.57
Enhancers vs. Coding regions	H3K14ac	6	6	1.00
	H3K4me1	111	111	1.00
	H4K91ac	6	5	0.83
	H3K27ac	82	61	0.74
	H2AK5ac	7	5	0.71
	H2BK120ac	7	5	0.71
	H3K18ac	7	5	0.71
	H3K36me3	111	76	0.68
	H2BK12ac	6	4	0.67
	H3K4me2	8	5	0.63
	H2BK15ac	5	3	0.60
	H3K79me2	5	3	0.60
	H3K9ac	47	27	0.57
	H3K4ac	7	4	0.57

The distributions of known marks in each of the 111 tissues were compared between (i) active promoters and active enhancers; (ii) active promoters and coding regions of active genes; and (iii) active enhancers and coding regions of active genes. The distributions of a mark around two genetic elements are considered opposite if they have a *dotsim* value of -0.5 or lower. Recall that the dotsim values range between -1 and 1. Not all marks were determined for all tissues. The number of tissues, for which a mark was determined, is listed under the column titled "Known." The number of tissues, in which a mark has opposite distributions over two genetic elements, is listed under the column titled "Opposite." The ratio of these two numbers are listed under the column labeled with "Ratio."

active enhancers and coding regions of active genes. These comparisons reveal that the signatures of active enhancers and active promoters are very similar; only one mark, H3K4me3, has different distributions around them. In contrast, the signature of active promoters differs in 8 marks from that of coding regions of active genes; these marks are H3K4(me1,me2,me3), H3K(9,18,27)ac, and H2A.Z. The signature of active enhancers differs in 14 marks from that of coding regions of active genes. These marks are H3K4(me1,me2), H3K36me3, H3K79me2, and 10 acetylation marks including H3K(9,18,27)ac. Interestingly, H3K14ac has opposite distributions around active enhancers and coding regions of active genes in all of the six tissues, in which it is known.

Clearly, the distributions of these marks can be used for distinguishing the signatures of the three active elements from each other. *These results show that active enhancers and active promoters have similar signatures which markedly differ from the signature of coding regions of active genes.*

Signature of inactive elements

We conducted the following experiment in search of a chromatin signature for inactive elements. Specifically, we aimed at studying the chromatin signatures of inactive promoters, inactive enhancers, and inactive genes. To determine promoters that are inactive in a specific tissue, we merged all putative promoters of all tissues. A promoter is considered inactive in a tissue if it does not overlap with any of the promoters active in this tissue. Inactive enhancers were determined in the same way. A gene that its transcription start site does not overlap with any of the putative tissue-specific promoters is considered inactive in this tissue. Next, we sampled about 500 elements from each chromosome of the human genome, totaling 11,000-13,000 elements. Then three Hebb networks were trained on the inactive promoters, the inactive enhancers, and the inactive genes of each tissue. After that, Hebb plots were generated from the signatures learned by these networks (Data set S8, Data set S9, and Data set S10). Upon examining the Hebb plots generated for the 111 tissues, we found the following:

- Promoters and enhancers that are inactive in stem cells have chromatin signatures consisting of many marks. The intensities of these marks are weaker (less bright) than their counterparts in the signatures of promoters and enhancers active in stem cells (Fig 7 and Fig 8).
- Out of the 111 tissues, the inactive promoters of 84 tissues were marked by H3K27me3, which is a repressive mark [69]. The H3K27me3 shows a moderate signal around inactive promoters of the steam cells and the differentiated cells alike.
- No mark of the available ones was present consistently around inactive enhancers in the differentiated cells (Fig 8).
- No mark of the available ones was present consistently around coding regions of genes that are inactive in the stem and the differentiated cells (Fig 9).

There are more than 100 chromatin marks [37]. Therefore, it is possible that other marks may repress promoters, enhancers, or genes. However, *the currently available data indicate that only H3K27me3 is consistently present around inactive promoters.*

Fig 7. Active (top row) and inactive (bottom row) promoters of two stem cell types, H1 and H9, and two differentiated cell types, IMR-90 and liver. The inactive promoters of the stem cells show some pattern, whereas those of the differentiated cells do not show any.

Fig 8. Active (top row) and inactive (bottom row) enhancers of two stem cell types, H1 and H9, and two differentiated cell types, IMR-90 and liver. Many marks are present around the inactive enhancers of the stem cells.

Fig 9. Coding regions of active (top row) and inactive (bottom row) genes of two stem cell types, H1 and H9, and two differentiated cell types, IMR-90 and liver. No marks is consistently present around the coding regions of inactive genes.

Online resource

We generated Hebb plots for multiple genetic elements, which are active and inactive in the 111 consolidated epigenomes provided by the Roadmap Epigenomics Project. Specifically, Hebb plots were generated for the following elements:

- Active promoters.
- Active promoters on the positive strand.
- Active promoters on the negative strand.
- Inactive promoters.
- Active enhancers.
- Active repetitive enhancers.
- Active non-repetitive enhancers.
- Inactive enhancers.
- Coding regions of active genes.
- Coding regions of inactive genes.

These Hebb plots are available in Data set S1-Data set S10. All of these regions were expanded by 10% on each end, except the active promoters on the positive and the negative strands were expanded by 100% on each end. The HebbPlot program is provided in Software S1.

Conclusion

Identifying a complex chromatin signature consisting of tens of marks distributed around thousands of regions is a challenging task. In this article, we described the first application of Hebb networks to learning the chromatin signature of a genetic element, e.g. promoters active in a specific tissue. These networks are known for their ability to learn associations. Therefore, they are well suited for learning the association between chromatin marks and thousands of sequences. We have developed a software tool called HebbPlot. The core of this tool is a Hebb network. Additionally, HebbPlot generates a digitized image representing the learned signature. The brightness level of a pixel indicates the confidence with which a mark is present or absent. For example, a white pixel indicates the presence of a mark around a part of the genetic element, and a black pixel indicates the absence of the mark. A row of pixels represents one mark. Similar rows are clustered and displayed together.

The Roadmap Epigenomics Project determined tens of chromatin marks for 111 cell types. We used HebbPlot in driving the chromatin signatures of multiple genetic

elements including: (1) active promoters, (2) active promoters on the positive strand, (3) active promoters on the negative strand, (4) inactive promoters, (5) active enhancers, (6) active enhancers within repetitive regions, (7) active enhancers outside repetitive regions, (8) inactive enhancers, (9) active genes, and (10) inactive genes. By analyzing these plots, we drove the following conclusions:

- Active promoters, active enhancers, and active genes have distinct chromatin signatures.
- The promoter signature is directional; multiple marks around the promoters are stretched toward coding regions.
- Enhancers within and outside repeats have almost identical chromatin signatures, supporting the enhancer-like functionality of transposons in the human genome.
- H3K79me1 is distributed similarly around the three active elements. Additionally, H3K9me3 and H3K27me3 are absent from the three genetic elements. These three marks represent a basic signature of elements active in almost all of the 111 cell types.
- The signatures of active promoters and active enhancers are more similar to one another than to the signature of coding regions of active genes.
- H3K27me3, which is a repressive mark, is consistently present around inactive promoters.

The software and the signature plots of all elements of the 111 epigenomes have been made available.

In sum, HebbPlot is a general software tool that can learn and represent visually the chromatin signature of thousands of regions having the same function. HebbPlot can be applied to the currently available epigenomes and the ones that will be available in the near future.

Supporting information

Software S1 The source code of our software tool, HebbPlot.

Data set S1 Hebb plots of potential promoters of the 111 tissues.

Data set S2 Hebb plots of potential enhancers of the 111 tissues.

Data set S3 Hebb plots of coding regions of active genes of the 111 tissues.

Data set S4 Hebb plots of potential promoters, on the positive strand, of the 111 tissues.

Data set S5 Hebb plots of potential promoters, on the negative strand, of the 111 tissues.

Data set S6 Hebb plots of repetitive enhancers of the 111 tissues.

Data set S7 Hebb plots of non-repetitive enhancers of the 111 tissues.

Data set S8 Hebb plots of inactive promoters of the 111 tissues.

Data set S9 Hebb plots of inactive enhancers of the 111 tissues.

Data set S10 Hebb plots of coding regions of inactive genes of the 111 tissues.

Table S1 Percentages of repetitive enhancers in the 111 tissues. The percentages of the tissue-specific enhancers overlapping simple and interspersed repeats are listed in this file. Enhancers that do not overlap repeats are listed under the column "0%." Under column "50%," we list the percentages of enhancers that at least 50% of their nucleotides overlap repeats. The percentages of enhancers fully included within repetitive regions are listed under the column titled "100%." (XLS)

Table S2 Comparisons between the signatures of repetitive and non-repetitive enhancers of the 111 tissues. The two signatures were compared using the dotsim function. (XLS)

Acknowledgments

This research was supported by internal funds provided by the College of Engineering and Natural Sciences and the Faculty Research Grant Program at the University of Tulsa.

References

1. Carey N. The epigenetics revolution: how modern biology is rewriting our understanding of genetics, disease, and inheritance. New York Chichester, West Sussex: Columbia University Press; 2012.
2. Lewis JD, Meehan RR, Henzel WJ, Maurer-Fogy I, Jeppesen P, Klein F, et al. Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. Cell. 1992;69(6):905–914.
3. Jenuwein T, Allis CD. Translating the histone code. Science. 2001;293(5532):1074–1080.
4. Kouzarides T. Chromatin modifications and their function. Cell. 2007;128(4):693 – 705.
5. Whitelaw NC, Chong S, Morgan DK, Nestor C, Bruxner TJ, Ashe A, et al. Reduced levels of two modifiers of epigenetic gene silencing, Dnmt3a and Trim28, cause increased phenotypic noise. Genome Biol. 2010;11(11):R111.
6. Carone BR, Fauquier L, Habib N, Shea JM, Hart CE, Li R, et al. Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. Cell. 2010;143(7):1084–1096.
7. Ng SF, Lin RCY, Laybutt DR, Barres R, Owens JA, Morris MJ. Chronic high-fat diet in fathers programs [bgr]-cell dysfunction in female rat offspring. Nature. 2010;467(7318):963–966.

8. Anway MD, Cupp AS, Uzumcu M, Skinner MK. Epigenetic transgenerational actions of endocrine disruptors and male fertility. Science. 2005;308(5727):1466–1469.

9. Guerrero-Bosagna C, Settles M, Lucker B, Skinner MK. Epigenetic transgenerational actions of Vinclozolin on promoter regions of the sperm epigenome. PLoS One. 2010;5(9):1–17.

10. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, et al. Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A. 2005;102(30):10604–10609.

11. Hammoud SS, Nix DA, Zhang H, Purwar J, Carrell DT, Cairns BR. Distinctive chromatin in human sperm packages genes for embryo development. Nature. 2009;460(7254):473–478.

12. Ooi SKT, Qiu C, Bernstein E, Li K, Jia D, Yang Z, et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature. 2007;448(7154):714–717.

13. Prader A, Labhart A, Willi H. A syndrome characterized by obesity, small stature, cryptorchidism and oligophrenia following a myotonia-like status in infancy. Schweiz Med Wochenschr. 1956;86:1260—-1261.

14. Angelman H. 'Puppet' children: a report on three cases. Dev Med Child Neurol. 1965;7(6):681–688.

15. Wiedemann HR. Familial malformation complex with umbilical hernia and macroglossia–a "new syndrome"? J Genet Hum. 1964;13:223–232.

16. Beckwith JB. Macroglossia, omphalocele, adrenal cytomegaly, gigantism and hyperplastic visceromegaly. Birth Defects. 1969;5:188–196.

17. Silver H, Kiyasu W, George J, Deamer W. Syndrome of congenital hemihypertrophy, shortness of stature and elevated urinary gonadotropins. Pediatrics. 1953;12:368–376.

18. Russell A. A syndrome of intra-uterine dwarfism recognizable at birth with cranio-facial dysostosis, disproportionately short arms, and other anomalies (5 examples). Proc R Soc Med. 1954;47:1040–1044.

19. Bukulmez O. Does assisted reproductive technology cause birth defects? Curr Opin Obstet Gynecol. 2009;21(3).

20. Shi Y, Lan F, Matson C, Mulligan P, Whetstine JR, Cole PA, et al. Histone demethylation mediated by the Nuclear Amine Oxidase Homolog LSD1. Cell. 2204;119(7):941–953.

21. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell. 2006;125(2):315–326.

22. Lee JT. The X as model for RNA's niche in epigenomic regulation. Cold Spring Harb Perspect Biol. 2010;2(9).

23. Herman JG, Latif F, Weng Y, Lerman MI, Zbar B, Liu S, et al. Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. Proc Natl Acad Sci U S A. 1994;91(21):9700–9704.

24. Esteller M, Silva JM, Dominguez G, Bonilla F, Matias-Guiu X, Lerma E, et al. Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. J Natl Cancer Inst. 2000;92(7):564.

25. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JPJ. CpG island methylator phenotype in colorectal cancer. Proc Natl Acad Sci U S A. 1999;96(15):8681–8686.

26. Lu Z, Luo RZ, Peng H, Huang M, Nishmoto A, Hunt KK, et al. E2F–HDAC complexes negatively regulate the tumor suppressor gene ARHI in breast cancer. Oncogene. 2006;25:230–239.

27. Gery S, Komatsu N, Kawamata N, Miller CW, Desmond J, Virk RK, et al. Epigenetic silencing of the candidate tumor suppressor gene Per1 in non–small cell lung cancer. Clin Cancer Res. 2007;13(5):1399–1404.

28. Kondo Y, Shen L, Cheng AS, Ahmed S, Boumber Y, Charo C, et al. Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation. Nat Genet. 2008;40(6):741–750.

29. Jones PA, Taylor SM. Cellular differentiation, cytidine analogs and DNA methylation. Cell. 1980;20(1):85–93.

30. Santi DV, Garrett CE, Barr PJ. On the mechanism of inhibition of DNA-cytosine methyltransferases by cytosine analogs. Cell. 1983;33(1):9–10.

31. Marks PA, Breslow R. Dimethyl sulfoxide to vorinostat: development of this histone deacetylase inhibitor as an anticancer drug. Nature Biotechnol. 2007;25:84–90.

32. Hentrich T, Schulze JM, Emberly E, Kobor MSa. CHROMATRA: a Galaxy tool for visualizing genome-wide chromatin signatures. Bioinformatics. 2012;28(5):717–718.

33. Younesy H, Nielsen CB, Lorincz MC, Jones SJM, Karimi MM, Möller T. ChAsE: chromatin analysis and exploration tool. Bioinformatics. 2016;32(21):3324.

34. Lukauskas S, Visintainer R, Sanguinetti G, Schweikert GB. DGW: an exploratory data analysis tool for clustering and visualisation of epigenomic marks. BMC Bioinformatics. 2016;17(16):53–63.

35. Hon G, Ren B, Wang W. ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS Comput Biol. 2008;4(10):e1000201.

36. Ucar D, Hu Q, Tan K. Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering. Nucleic Acids Res. 2011;39(10):4063–4075.

37. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature Methods. 2012;9(3):215–216.

38. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012;9(5):473–476.

39. Wang J, Lunyak VV, Jordan IK. Chromatin signature discovery via histone modification profile alignments. Nucleic Acids Res. 2012;40(21):10642–10656.

40. Lai WKM, Buck MJ. An integrative approach to understanding the combinatorial histone code at functional elements. Bioinformatics. 2013;29(18):2231–2237.

41. Zhou J, Troyanskaya OG. Global quantitative modeling of chromatin factor interactions. PLoS Comput Biol. 2014;10(3):1–13.

42. Hamada M, Ono Y, Fujimaki R, Asai Ka. Learning chromatin states with factorized information criteria. Bioinformatics. 2015;31(15):2426–2433.

43. Song J, Chen KC. Spectacle: fast chromatin state annotation using spectral learning. Genome Biol. 2015;16(1):33.

44. Lai WK, Buck MJ. ArchAlign: coordinate-free chromatin alignment reveals novel architectures. Genome Biol. 2010;11(12):R126.

45. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007;39(3):311–318.

46. Won KJ, Chepelev I, Ren B, Wang W. Prediction of regulatory elements in mammalian genomes using chromatin signatures. BMC Bioinformatics. 2008;9(1):547.

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

48. Firpi HA, Ucar D, Tan K. Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics. 2010;26(13):1579–1586.

49. Cheng C, Yan KK, Yip KY, Rozowsky J, Alexander R, Shou C, et al. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biol. 2011;12(2):R15.

50. Cheng C, Shou C, Yip KY, Gerstein MB. Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors. Genome Biol. 2011;12(11):R111.

51. Zhang Z, Zhang MQ. Histone modification profiles are predictive for tissue/cell-type specific expression of both protein-coding and microRNA genes. BMC Bioinformatics. 2011;12:155.

52. Fernández M, Miranda-Saavedra D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. Nucleic Acids Res. 2012;40(10):e77–e77.

53. Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, et al. RFECS: A random-forest based algorithm for enhancer identification from chromatin state. PLoS Comput Biol. 2013;9(3):e1002968.

54. Kumar S, Bucher P. Predicting transcription factor site occupancy using DNA sequence intrinsic and cell-type specific chromatin features. BMC Bioinformatics. 2016;17(Suppl 1):S4.

55. Park SH, Lee SM, Kim YJ, Kim S. ChARM: Discovery of combinatorial chromatin modification patterns in hepatitis B virus X-transformed mouse liver cancer using association rule mining. BMC Bioinformatics. 2016;7:1307.

56. Girgis HZ, Corso JJ, Fischer D. On-line hierarchy of general linear models for selecting and ranking the best predicted protein structures. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2009. p. 4949–4953.

57. Girgis HZ, Ovcharenko I. Predicting tissue specific cis-regulatory modules in the human genome using pairs of co-occurring motifs. BMC Bioinformatics. 2012;13(1):25.

58. Girgis HZ. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics. 2015;16(1).

59. Hebb DO. The organization of behavior: a neuropsychological theory. new ed. Lawrence Erlbaum Associates, Inc., Publishers; 2002.

60. Hagan MT, Demuth HB, Beale MH, De Jesús O. Neural network design. 2nd ed.; 2014.

61. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin. 1958;38:1409–1438.

62. Consortium RE, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–330.

63. Xie M, Hong C, Zhang B, Lowdon R, Xing X, Li D, et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat Genet. 2013;45(7):836–841.

64. Hayashi K, Yoshida H. Refunctionalization of the ancient rice blast disease resistance gene Pit by the recruitment of a retrotransposon as a promoter. Plant J. 2009;3:413–425.

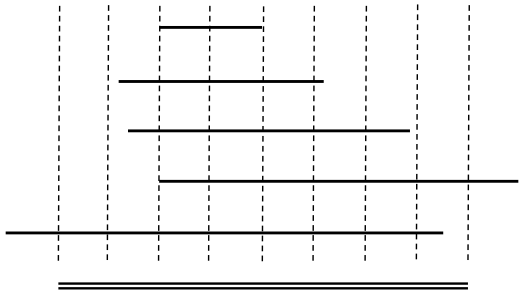
65. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008;9:397–405.

66. Fernandez L, Torregrosa L, Segura V, Bouquet A, Martinez-Zapater JM. Transposon-induced gene activation as a mechanism generating cluster shape somatic variation in grapevine. Plant J. 2010;61(4):545–557.

67. Rebollo R, Romanish MT, Mager DL. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. Annu Rev Genet. 2012;46:21–42.

68. Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol. 2014;65:505–530.

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

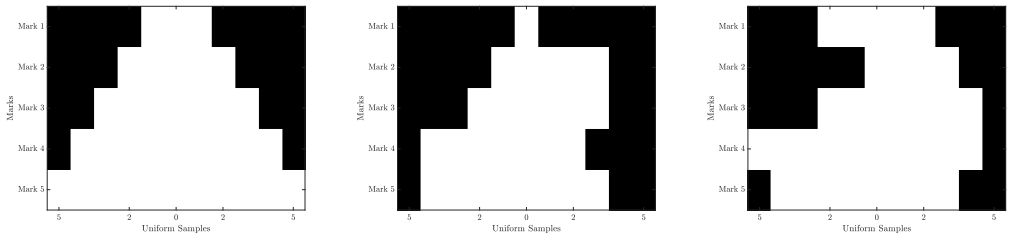


(a) Visual representation

$$\begin{bmatrix} -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 \\ -1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 \end{bmatrix}$$

(b) Numerical representation

Fig 1. Representations of a group of chromatin marks overlapping a region. (a) Horizontal double lines represent a region of interest. Horizontal single lines represent the marks. Vertical lines are spaced equally and bounded by the region. (b) The intersections between the marks and the vertical lines are encoded as a matrix where rows represent the marks and columns represent the vertical lines. If a vertical line intersects a mark, the corresponding element in the matrix is 1, otherwise it is -1.



(a) The real shape

(b) A noisy shape example

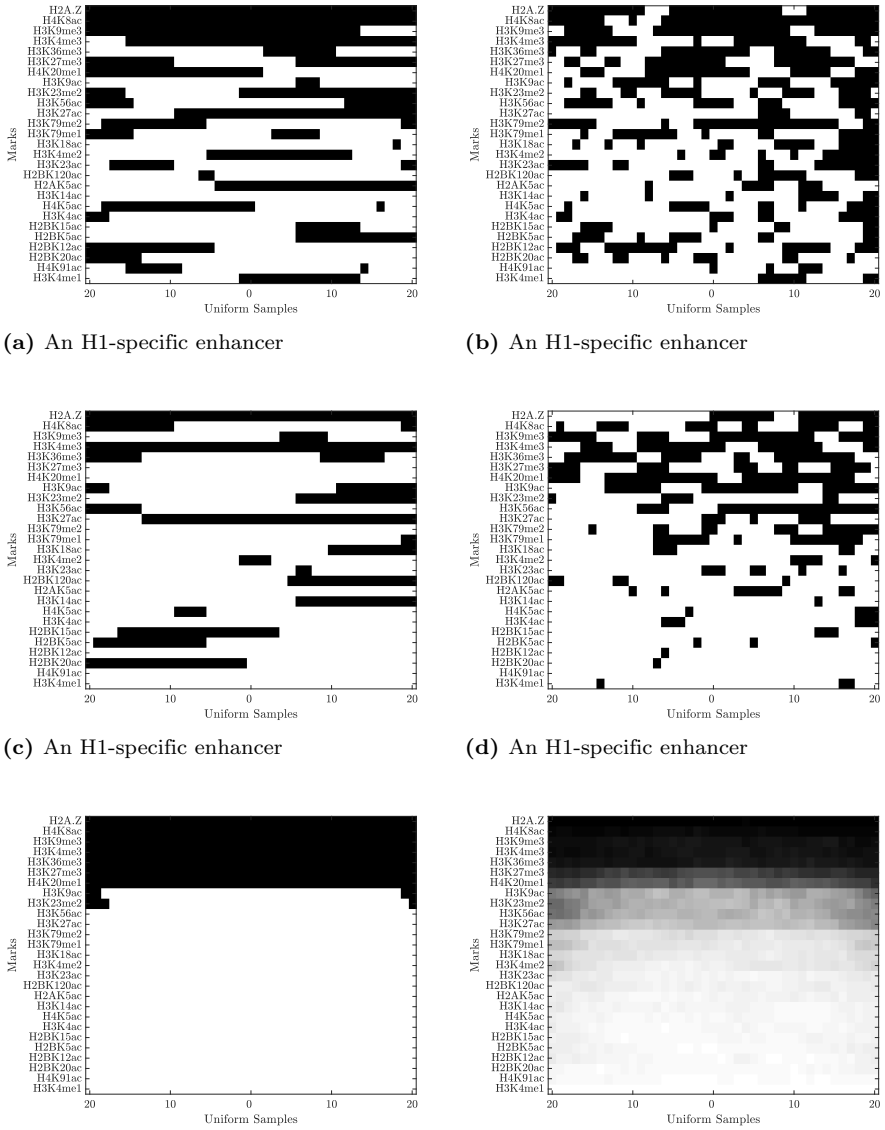
(c) A noisy shape example

(d) A noisy shape example

(e) The shape retrieved by a majority vote

(f) The shape retrieved by Hebb network

Fig 2. A controlled experiment demonstrating the ability of Hebb networks to retrieve a prototype from noisy shapes. (a) The real shape is a two-dimensional step pyramid. (b-d) Examples of noisy shapes produced by corrupting the real shape. The start and the end of each horizontal white segment may be shifted to the right or the left by up to 200 unites; additionally, a horizontal segment may be deleted with a probability of 0.2. (e) The shape retrieved by a majority-voting scheme. (f) The shape retrieved by the network.



(a) An H1-specific enhancer

(b) An H1-specific enhancer

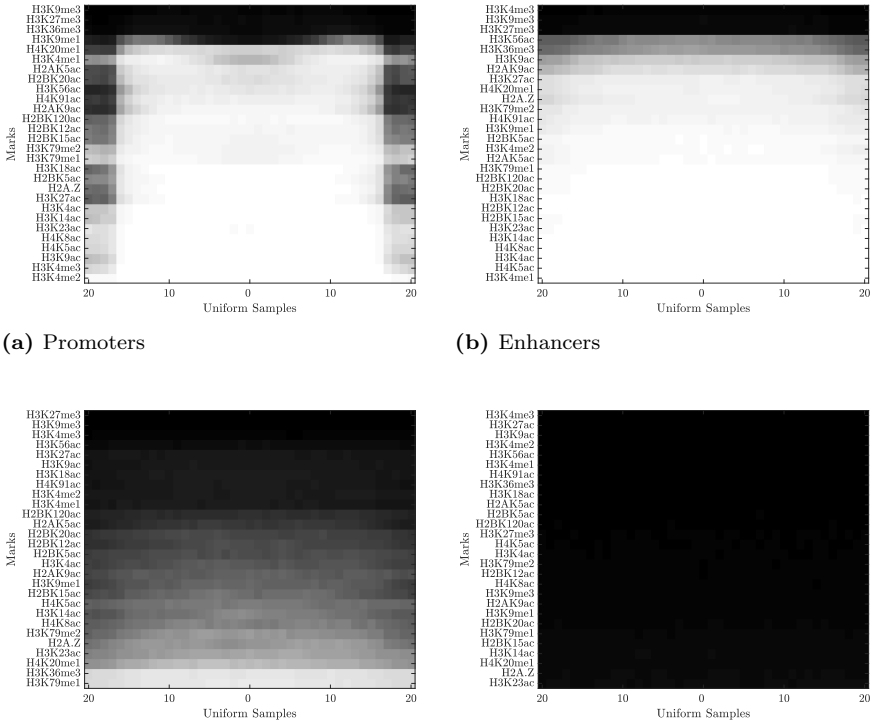
(c) An H1-specific enhancer

(d) An H1-specific enhancer

(e) The shape retrieved by a majority vote

(f) The shape retrieved by HebbPlot

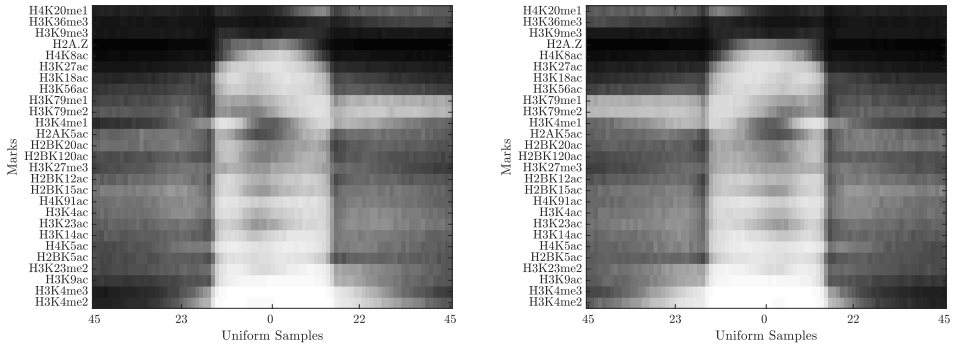
Fig 3. Retrieving the chromatin signature of the H1-specific enhancers. The data set consists of 11,369 putative enhancers active in H1. Four examples of potential enhancers are shown in parts a-d. It is hard to see a common pattern in these four examples. The signature retrieved by the majority-voting scheme has two zones (part e), whereas the signature learned by the Hebb network is characterized by four zones (part f). The top most zone represents chromatin marks that are absent from the enhancer regions, whereas the next three zones represent the present marks with increasing certainty. The signature retrieved by the majority-voting scheme does not show confidence indicators of the absence or the presence of a mark. For example the third zone from the top of the Hebb plot is not as strong as the forth zone. In contrast, such information does not appear in the signature retrieved by the majority-voting scheme. Enhancer regions were expanded by 10% on each end. Therefore, the intensity of the signal is expected to be weaker at the peripheries than around the center of a signature. Again, the Hebb plot shows such information, whereas the signature retrieved by the majority-voting scheme does not.



(a) Promoters **(b)** Enhancers

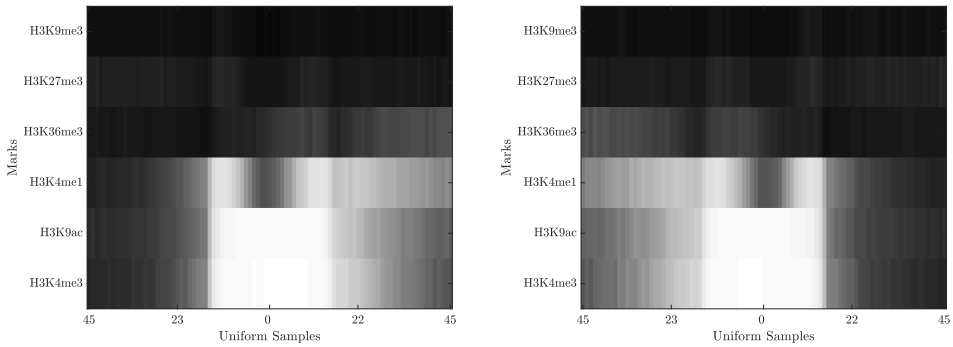
(c) Coding regions **(d)** Random regions

Fig 4. IMR-90 chromatin signatures representing (a) active promoters, (b) active enhancers, (c) coding regions of active genes, and (d) random genomic locations. The boundaries of these elements were expanded by 10% on each end to show a chromatin signature in contrast to the surrounding regions. Active promoters have a unique signature characterized by a bright box that clearly differs from the background. In addition, the center of the upper half of the box has a less bright area around the transcription start sites than its peripheries. Enhancers specific to IMR-90 has a zonal signature, where each zone has consistent brightness. The enhancer signature has a wide bright zone. Genes active in IMR-90 has a zonal signature as well. However, the middle gray zone is the widest, and the brightest zone at the bottom is the narrowest. The random locations do not have a common function; therefore, chromatin marks around them should not be distributed consistently. As expected, HebbPlot did not retrieve any pattern as displayed by a black box.



(a) + H1 Promoters

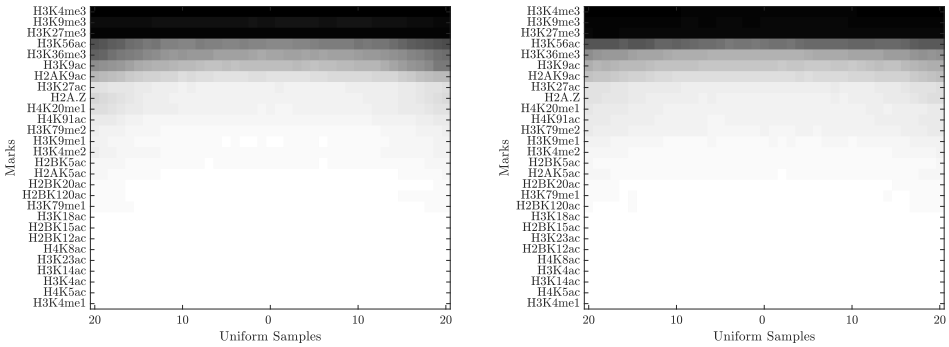
(b) - H1 Promoters



(c) + Skeletal Muscle Male Promoters

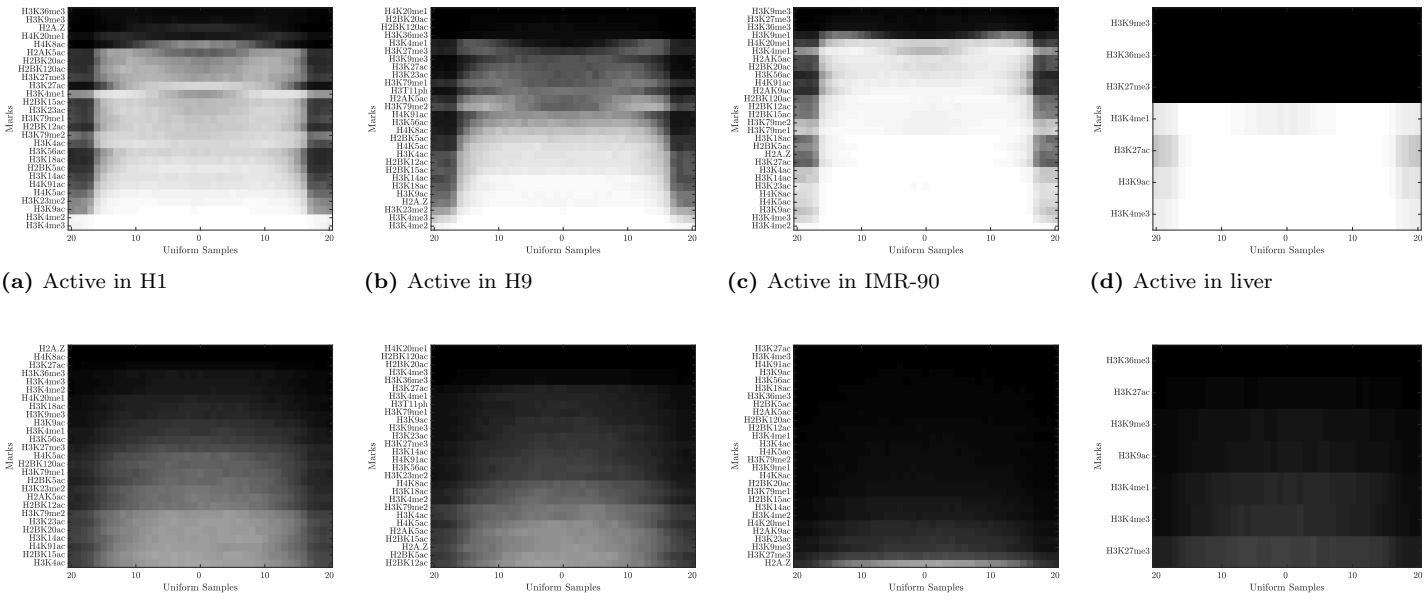
(d) - Skeletal Muscle Male Promoters

Fig 5. Hebb plots of active promoters on the positive and the negative strands. Multiple chromatin marks are distributed in a direction specific way. These marks tend to stretch downstream from the promoters toward the coding regions. Examples are H3K4(me1,me2,me3) and H3K79(me1,me2). The two Hebb plots of the promoters on the positive and the negative strands are mirror images of each other.



(a) Repetitive **(b)** Non-repetitive

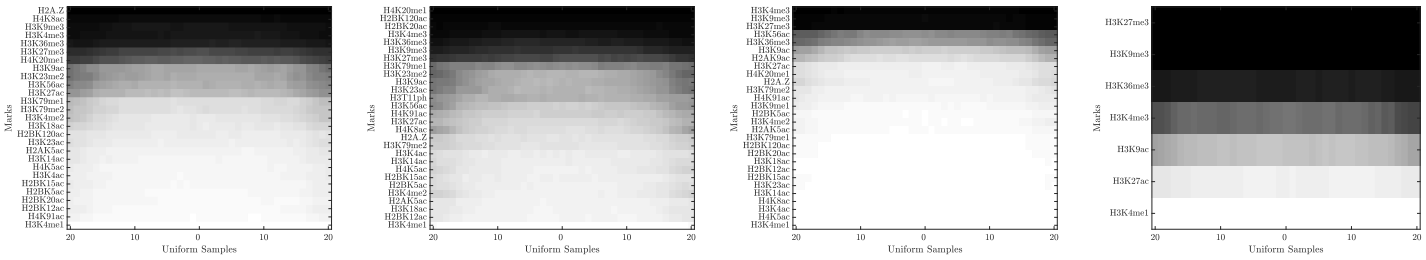
Fig 6. The repetitive and the non-repetitive enhancers active in the IMR-90 cell line have identical chromatin signatures. Enhancers that are fully included within repeat sequences, e.g. transposons, are considered repetitive. Enhancers that do not overlap with repeats are considered non-repetitive. The Hebb plots representing the chromatin signatures of the repetitive and the non-repetitive enhancers are almost identical. Further, a dotsim value of 0.9976 was obtained by comparing these signatures. Recall that a dotsim value of 1 is the result of comparing a signature to itself.



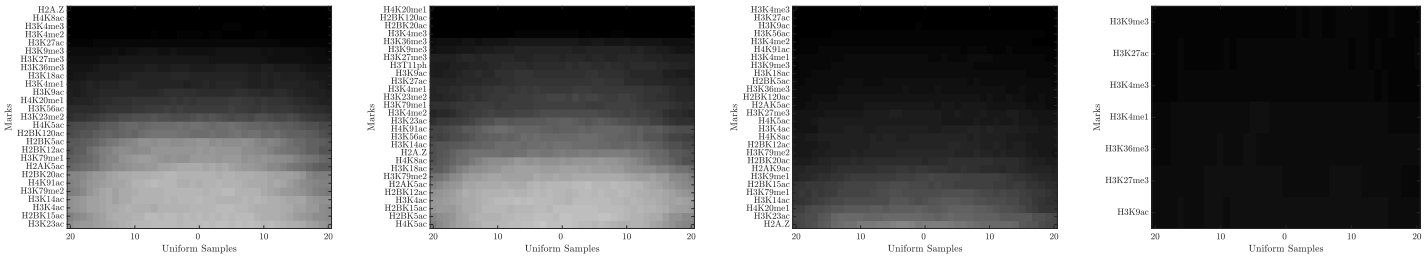
(a) Active in H1 **(b)** Active in H9 **(c)** Active in IMR-90 **(d)** Active in liver

(e) Inactive in H1 **(f)** Inactive in H9 **(g)** Inactive in IMR-90 **(h)** Inactive in liver

Fig 7. Active (top row) and inactive (bottom row) promoters of two stem cell types, H1 and H9, and two differentiated cell types, IMR-90 and liver. The inactive promoters of the stem cells show some pattern, whereas those of the differentiated cells do not show any.

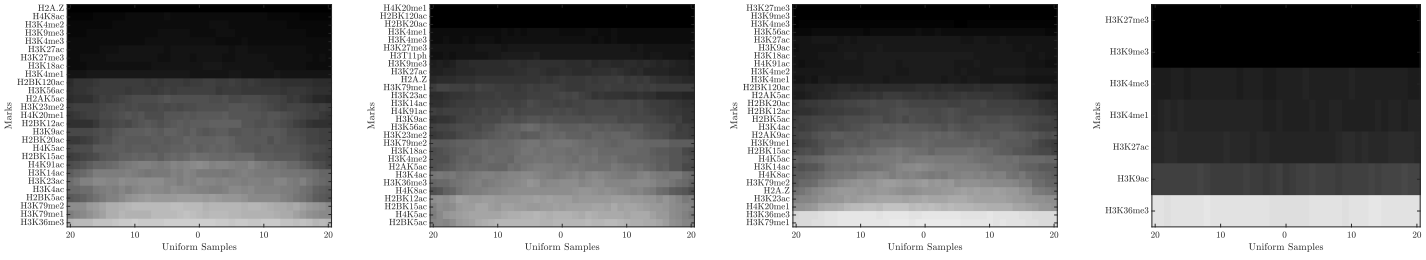


(a) Active in H1 **(b)** Active in H9 **(c)** Active in IMR-90 **(d)** Active in liver

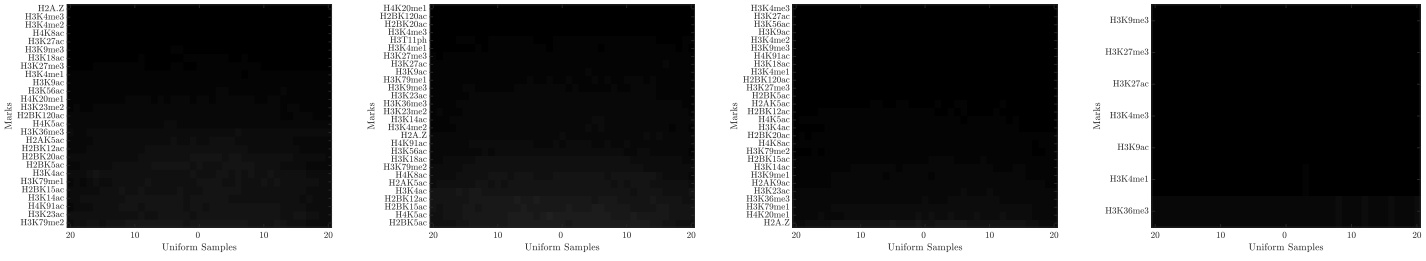


(e) Inactive in H1 **(f)** Inactive in H9 **(g)** Inactive in IMR-90 **(h)** Inactive in liver

Fig 8. Active (top row) and inactive (bottom row) enhancers of two stem cell types, H1 and H9, and two differentiated cell types, IMR-90 and liver. Many marks are present around the inactive enhancers of the stem cells.



(a) Active in H1 **(b)** Active in H9 **(c)** Active in IMR-90 **(d)** Active in liver



(e) Inactive in H1 **(f)** Inactive in H9 **(g)** Inactive in IMR-90 **(h)** Inactive in liver

Fig 9. Coding regions of active (top row) and inactive (bottom row) genes of two stem cell types, H1 and H9, and two differentiated cell types, IMR-90 and liver. No marks is consistently present around the coding regions of inactive genes.