BMC Bioinformatics

RESEARCH Open Access

MethylSeqLogo: DNA methylation smart sequence logos

Fei-Man Hsu[1] and Paul Horton[2*]

From International Conference on Genome informatics GIW XXXI/ISCB-Asia V 2022
Tainan, Taiwan. 12-14 December 2022.

*Correspondence:
paulh@iscb.org

[1] Department of Molecular Cell and Developmental Biology, University of California, Los Angeles, USA
[2] Department of Computer Science and Information Engineering, National Cheng Kung University, 1 University Road, Tainan 70101, Taiwan

Abstract

Background: Some transcription factors, MYC for example, bind sites of potentially methylated DNA. This may increase binding specificity as such sites are (1) highly under-represented in the genome, and (2) offer additional, tissue specific information in the form of hypo- or hyper-methylation. Fortunately, bisulfite sequencing data can be used to investigate this phenomenon.

Method: We developed MethylSeqLogo, an extension of sequence logos which includes new elements to indicate DNA methylation and under-represented dimers in each position of a set binding sites. Our method displays information from both DNA strands, and takes into account the sequence context (`CpG` or other) and genome region (promoter versus whole genome) appropriate to properly assess the expected background dimer frequency and level of methylation. MethylSeqLogo preserves sequence logo semantics—the relative height of nucleotides within a column represents their proportion in the binding sites, while the absolute height of each column represents information (relative entropy) and the height of all columns added together represents total information

Results: We present figures illustrating the utility of using MethylSeqLogo to summarize data from several `CpG` binding transcription factors. The logos show that unmethylated `CpG` binding sites are a feature of transcription factors such as MYC and ZBTB33, while some other `CpG` binding transcription factors, such as CEBPB, appear methylation neutral.

Conclusions: Our software enables users to explore bisulfite and ChIP sequencing data sets—and in the process obtain publication quality figures.

Keywords: DNA methylation, Transcription Factor Binding Sites, Visualization

Background

Transcription Factors (TFs) are proteins which bind genomic DNA at specific sites (Transcription Factor Binding Sites: TFBSs) to regulate gene expression and thereby enable Eukaryotic cells to appropriately express genes according to: cell type, the cell cycle, the developmental stage of the organism, external conditions, etc. [1–3]. Moreover, perturbation of TF function plays major roles in the etiology of diseases such as cancer [4] and diabetes [5]. In humans these effects are realized by an ensemble of approximately 1600 TFs, each with distinct and often cell-type specific TFBSs [1, 6].

Given this importance and complexity, the study of TF function is a long-standing and on-going topic in molecular biology. One of the early successes in this endeavor was the invention of "sequence logos" [7], an effective way to visualize the position specific base preferences which partially characterize TFBSs. Sequence logos consist of columns of the letters ({A, C, G, T} for a DNA motif) at each position, with the total column height of each position proportional to the information content of the distribution of bases in that position. Their popularity attests to their utility in visually summarizing binding sites, which in turn facilitates communication (as figures in papers, etc.), and comparison between the binding preferences of distinct TFs. Indeed sequence logos have been extended in several ways; for example to improve the resolution of enriched/depleted components, e.g. Seq2logo [7] and EDlogo [8] or to show higher order sequence motifs [9] or inter-positional correlations in binding sites [10].

Sequence logos help biologists understand the sequence preference of TFs; but the local DNA sequence is only one factor determining binding site selection, and cannot explain cell type specific TFBS selection. Evidently, a more complete understanding of TF function requires the integration of local DNA sequence with epigenetic marks [11].

DNA methylation is particularly interesting because it can affect the binding of many transcription factors [12–16]; and is easily cast as DNA sequence information, since 5-methylcytosine can be viewed as a fifth DNA base [17, 18]. Moreover, technologies such as bisulfite sequencing can measure tissue specific genome-wide DNA methylation levels at single-base resolution, and such data is already available for many cell types and conditions [19, 20].

Here we present MethylSeqLogo; a method which naturally extends classical sequence logos to visualize the methylation of a collection of TFBSs relative to an appropriate background. For user convenience we provide a software implementation prepackaged with methylation data for several cell lines from human, mouse, *Arabidopsis* and maize. The software also includes MethylScape, a companion method to MethylSeqLogo, which displays the methylation level of TFBS flanking regions.

Visualization method

Here we describe the design rationale and details of the MethylSeqLogo display; schematically presented in Fig. 1.

Design goals

1 Keep the advantages of sequence logos; including familiarity.
2 For methylation, clearly display:

- Strand (+/−) of the binding site
- Trinucleotide context (CG, CHG or CHH)
- Comparison relative to a background model
- Dimer enrichment/depletion in the motif

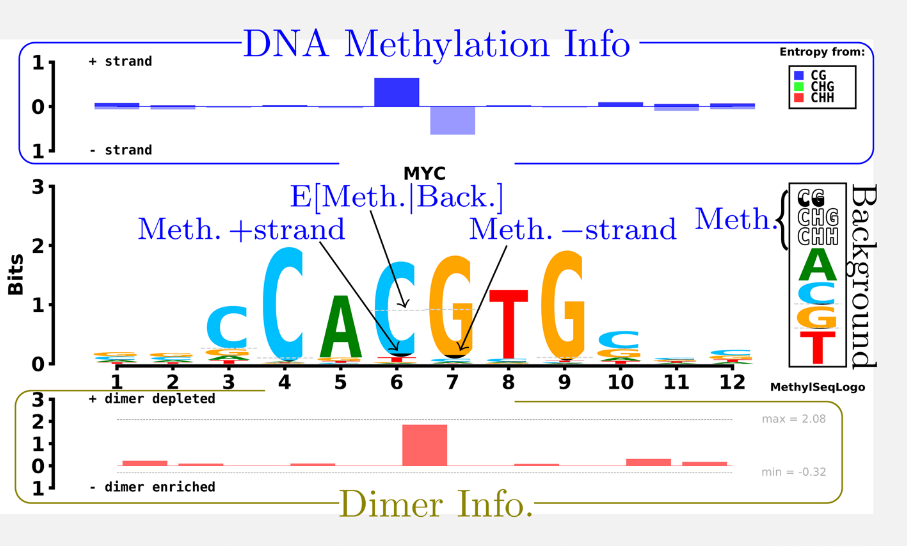

Fig. 1 Design of MethylSeqLogo. Proportional shading of C's and G's indicates the methylation level of TFBS cytosines on the forward and reverse strands respectively; while a dashed line indicates the expected level of methylation based on the background distribution. The methylation key at lower right of the logo shows background methylation probabilities of CG, CHG and CHH, respectively; and the four single nucleotide background probabilities. The top track shows the relative entropy contributed by methylation in each context/strand combination, with information associated with cytosines on the reverse strand displayed downward. In the bottom track positive height indicating the presence of under-represented dimers (typically CpG), and negative height (not seen in this example) indicating the presence of over-represented dimers. For reference, the theoretical maximum and minimum possible dimer relative entropy contribution achievable for the given background are also shown

We achieve the first goal by respecting two expectations viewers familiar with sequence logos will have: first, the relative height of an element (e.g. "A") within a column represents the frequency of the corresponding element; and second, that the height of a column represents an information theoretic measure (relative entropy) of the degree to which that position in the binding sites differs from background [7].

We achieve the second goal by adding several intuitive elements to the plot:

- Partial shading of C's and G's
- Dashed line indicating expected methylation level
- Box at right showing background frequencies
- Context Colored Methylation info track at top
- Dimer enrichment/depletion info track at bottom

The height of the shading of C's and G's is proportional to the methylation level of cytosines on the forward and reverse strands respectively.

In order to give users a clear image of hyper- or hypo-methylation, we added a dashed line showing the methylation level which would be expected based on the background distribution (taking the trinucleotide context {CG, CHG, CHH} in each binding site into account).

We designed a methylation info track showing (for each position in the binding site) the contribution of each context to the methylation information; and a box at right to show the background distribution of bases and methylation used for the relative entropy computation (Fig. 1).

Column heights

This section describes how column height is determined for MethylSeqLogo's three tracks so that the total information in a set of binding sites can be estimated by visually adding up the height of all elements in a MethylSeqLogo display.

Column heights indicate relative entropy

Sequence logos often employ a background model fit to a set of background sequences, such as the whole genome or promoter regions etc. The background model is used to compute how "typical" the binding sequences are, with the idea that atypical binding site sequences should be emphasized visually (given taller column height) to reflect their statistical distance from background. For example, binding sites abundant in C and G should be emphasized more against an AT-poor background than against an AT-poor background. Quantitatively, the column heights are made proportional to the *relative entropy*; also known as the *Kullback–Leibler directed divergence* [21], and equivalent to *information content* [22] under a uniform distribution background.

Sequence background models In explaining the MethylSeqLogo sequence logo and dimer information tracks, we will refer to zero order Markov model and first order Markov model background models. Zero order models generate each nucleotide of a DNA sequence independently, but first order models condition the nucleotide probabilities on the previous nucleotide.

Relative entropy formula To facilitate describing the column heights of the MethylSeqLogo tracks in the following sections, we state the definition of relative entropy:

$$\mathrm{D}(M||B) \overset{\text{def}}{=} \mathrm{E}\left[\lg\left(\frac{P[\,s|\text{ Motif Model } M]}{P[\,s|\text{ Background model } B]}\right)\right]$$

using lg to denote $\log_2$.

With this notation, the difference in relative entropy when employing different background models $\mathbf{B}_1$ versus $\mathbf{B}_0$ is:

$$\mathrm{D}(M||\mathbf{B}_1) - \mathrm{D}(M||\mathbf{B}_0) \;=\; \mathrm{E}\big[\lg(P[s|\mathbf{B}_0])\big] - \mathrm{E}\big[\lg(P[s|\mathbf{B}_1])\big]$$

Where the expectation is the average over the individual binding site sequences *s* in a set of binding sites.

Sequence logo track column height

Standard sequence logos typically display columns with a height proportional to relative entropy using a PWM (Position Weight Matrix) based motif model which assigns distinct probabilities to the nucleotides {A,C,G,T} at each position but assumes independence between positions. A zero order Markov model, which also assumes positional independence, is usually employed as a background model. In this case the relative entropy of the binding sites is easily decomposed into a sum with one term for each position; and therefore can be conveniently displayed via the height of the column representing each position. MethylSeqLogo adopts these conventions for its sequence logo track.

Dimer information track

Although convenient, a zero order background model is unable to represent the striking (sometimes > 4x) depletion of `CpG` (relative to `CpC`, `GpC`, and `GpG` dinucleotides) in mammalian genomes. Admittedly, `CpG`'s are much less depleted in promoter regions, but there is still discrepancy between actual dimer frequencies versus what would be predicted by a zero order model. Therefore a first order model should provide a substantially more useful measure of how statistically distinct a set of binding sites is from background.

Given the potential size of this effect and the fact that methylation occurs at `CpG` dimers, we decided MethylSeqLogo should display information based on a first order Markov model background. We did not want to change the sequence logo track, so instead of directly displaying relative entropy against a first order Markov model, we chose to display the *difference* between that relative entropy and the zero order background relative entropy in a separate track. Fortunately, this difference can easily be decomposed into the sum of a set of terms; one term for each pair of adjacent positions (see supplementary text for a mathematical derivation). Since these terms represent pairs of adjacent nucleotide positions, MethylSeqLogo displays them as vertical bars between the two positions. In theory, column heights in this track can be negative if the binding sites contain many over-represented dimers (for example homodimers `XpX` may be somewhat over-represented).

Methylation track column height

Hyper- or hypo-methylation of TF bindings sites (relative to a background) may help distinguish those binding sites from background. To allow users to see this effect, MethylSeqLogo presents a methylation information track above the main sequence logo track. Informally, the height of bars in the methylation information track represent the amount of additional surprise experienced when observing the methylation value at position i from one of the TFBSs; *after* having observed the primary sequences, since that information is already accounted for in the other tracks. The propensity of genomic cytosines to be methylated differs strongly depending on the following base or two (i.e. `CG`, `CHG`, or `CHH` trinucleotide context), so we separate these cases in our computation. For a background distribution these three cases are enough; while for binding sites, position and strand must also be considered. Thus altogether we separate the methylation data for each position in a collection of TFBSs into 6 strand specific contexts: 3 trinucleotide contexts × 2 strands (Supplementary Fig. 2).

Formally, let $P_{\text{context}\,|\,i}$ denote the probability that a binding site will have a cytosine matching the given context at position i and $P_{m\,|\,\text{context},i}$ denote the probability that such a cytosine will be methylated or not; while $P_{m\,|\,\text{context,BG}}$ denotes the background probability of a cytosine in that given context being methylated or not. We can write the contribution of methylation information to the height of column i as:

$$\mathsf{MH}_i = \sum_{\substack{m \in \\ \left\{\substack{\text{methylated} \\ \text{or not}}\right\}}} \sum_{\substack{\mathsf{context} \in \\ \left\{\substack{\texttt{CG},\ \texttt{CHG},\ \texttt{CHH}, \\ \reflectbox{\texttt{CG}},\ \reflectbox{\texttt{CHG}},\ \reflectbox{\texttt{CHH}}}\right\}}} P_{\mathsf{context}\,|\,i} \lg\left(\frac{P_{m\,|\,\mathsf{context},\,i}}{P_{m\,|\,\mathsf{context},\,\mathsf{BG}}}\right)$$

Note that relative entropy is inherently robust to small sample estimation error in $P_{m\,|\,\mathsf{context},\,i}$ since it includes a multiplicative term $P_{\mathsf{context}\,|\,i}$ in the contribution of that context to column height. Thus rare contexts cannot make large contributions to column height.

Data and software

DNA methylation and TFBS data

MethylSeqLogo requires binding sites and methylation information, preferably specific to a given tissue or cell-type. To gather this information we built a computational pipeline to process ChIP-seq data for TFBSs and WGBS to calculate the methylation probability of each position in the aligned TFBSs, as well as the background probabilities (Supplementary Fig. 1).

Whole Genome Bisulfite Sequencing data

We downloaded Human reference genome GRCh37 (hg19) and GRCh38 (hg38) from the Illumina iGenomes website and Human WGBS (Whole Genome Bisulfite Sequencing) from the ENCODE [23] website. The figures in this publication reflect data from ENCODE IDs: (`086MMC`, `379ZXG`, `417VRB`, `524BMX`, `601NBW`, `918PML`) and (`030LDK`, `086KJC`, `300GSM`, `390OZB`, `624VFJ`, `847OWL`) for H1-hESC and HepG2 cell lines respectively (all IDs start with `ENCFF`).

We merged the methylation calling BED files of two replicates for each cell type, by averaging the methylation levels of cytosine sites (on either strand) with read depth greater than four.

Methylation of TFBSs

We collected TFBS coordinates from the JASPAR database [24]; and tissue-specific ChIP-seq data from ReMap [25] for TFs (EGR1, MYC, SP1, USF1 and ZBTB33) to generate the figures in this text, and CEBPB for a supplementary text figure. To obtain tissue-specific TFBS coordinates, we used the `bedtools intersect` function. Based on those coordinates, one can generate the intermediate input files needed by MethylSeqLogo to generate MethylSeqLogo images (Fig. 2).

Promoter regions

Promoter regions have special significance for most transcription factors, but the distribution of both `CpG`'s and their methylation differs sharply between promoters regions and the genome as a whole. Thus we provide predefined promoter regions defined as 1000bp upstream to 200bp downstream of annotated major transcription start sites [26].

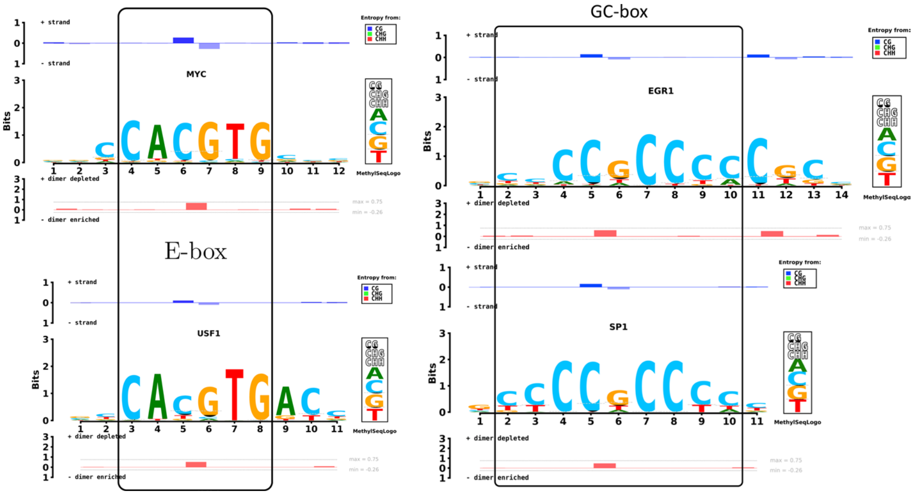

Fig. 2 MethylSeqLogo facilitates comparison of the DNA methylation of transcription factors with similar binding preferences. The E-box elements binding TFs MYC and USF1 (at left) and the GC-box elements binding TFs SP1 and EGR1 (at right) are compared using promoter region binding sites and background model. Data from H1-hESC cells

MethylSeqLogo program

We provide an open source implementation of the MethylSeqLogo visualization method and a companion program MethylScape described in below.

MethylSeqLogo comes with precomputed probability models for the examples discussed in this paper and many other tissues that have published WGBS data. Users can also calculate the methylation probabilities from their own WGBS datasets with a script provided in the MethylSeqLogo package and generate logos reflecting their data.

Example MethylSeqLogos

MYC binding sites

MYC transcription factors (data shown here is for c-Myc) are oncogenic transcription factors that bind DNA as a heterodimer with MAX [27]. Figure 3 (left) shows MethylSeqLogos of MYC using data from H1-hESC cells (numerical data shown in Supplementary Table 1). From these images it is apparent that when looking at the entire genome, MYC binding sites are statistically characterized by hypo-methylation and the occurrence of the under-represented dimer `CpG`. On the other hand, the promoter region based MethylSeqLogo's (Fig. 3 (top)) shows greatly reduced information from hypo-methylation and `CpG`; but it is "cleaner" in the sense that the methylation information is concentrated at positions 6 and 7, consistent with reports that methylation in the center `CpG` site of MYC binding sites reduces binding efficacy [28].

ZBTB33 binding sites

Figure 3 (right) shows MethylSeqLogos for ZBTB33 in HepG2 cells. ZBTB33, also named Kaiso [29], is a homodimeric transcription factor associated with several types of cancer [30]. ZBTB33 has been reported to bind methylated `CpG`'s and the sequence motif `TCCTGCNA` [31], especially `TCTCGCGAGA` [32]; with *in vitro* data indicating a

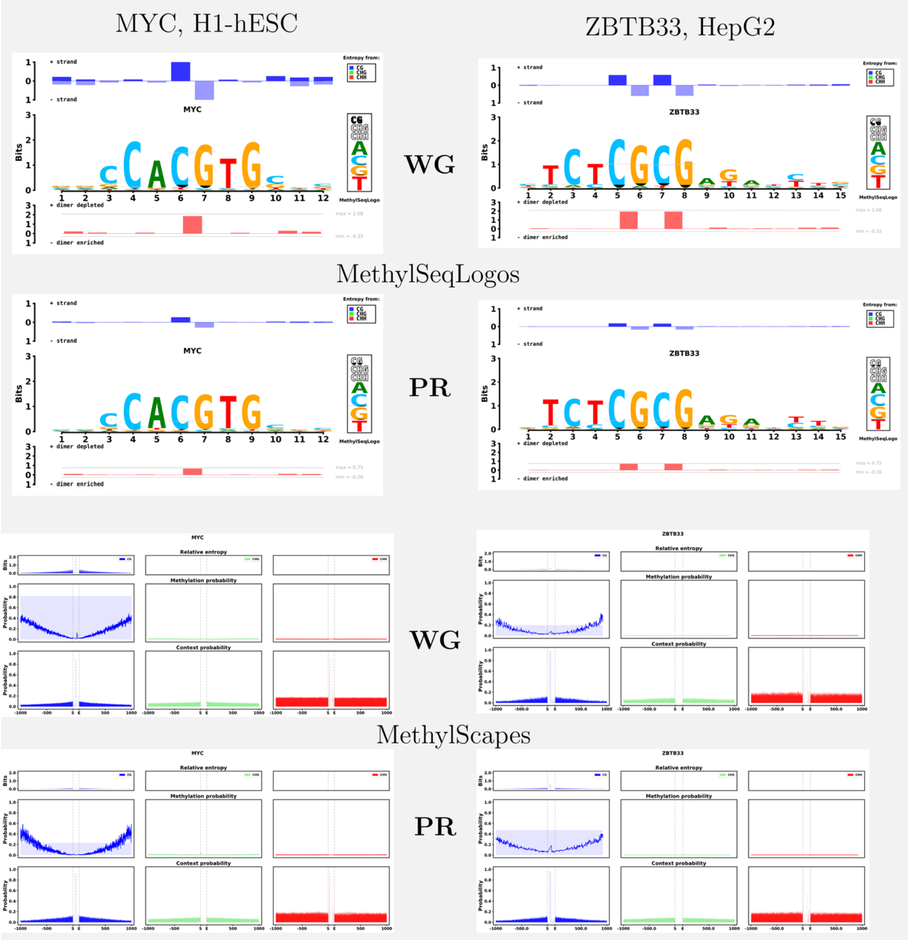

Fig. 3 MethylSeqLogos (top) and MethylScape(s) (bottom) of c-Myc binding sites in H1-hESC cells (left) and ZBTB33 binding sites in HepG2 cells (right). Logos in rows marked with **WG** show information for all binding sites relative to a whole genome (WG) background model, while logos in rows marked with **PR** show information for promoter region binding sites relative to a promoter region (PR) background model. The three columns in a MethylScape logo represent the contexts: `CpG`, `CHG`, and `CHH`; with faint background color in the middle row representing the background model methylation probability for each respective context

much higher affinity for this motif when methylated. Comparing ChIP-Seq and bisulfite sequencing data, Blattler et al. [33] were able to confirm the `TCTCGCGAGA` motif, but found that very few ZBTB33 binding sites are methylated *in vivo*. The visual impression given by MethylSeqLogo is in line with their conclusions.

Contrasting TF binding motifs

Transcription factors can be grouped by structural features of their DNA-binding domains. Often TFs with the same type of DNA-binding domains will bind to similar DNA sequences, which are sometimes called *response elements*. For example, an E-box (enhancer element) is a response element with palindromic general pattern `CANNTG` (`N` denotes any base) and canonical sequence `CACGTG`.

MYC and USF1 both have bHLH (basic helix-loop-helix) DNA-binding domains which bind to canonical E-box response elements. Comparing the methylation track of their MethylSeqLogos in figure 2 (left), USF1 appears more tolerant of methylation of the cytosines in the central CpG. Interestingly, comparing the sequence logo tracks, one can see that USF1 binding sites also exhibit more frequent substitution of 5-methyluracil (i.e. thymine) for cytosine as well. This example illustrates the utility of MethylSeqLogo in simultaneously comparing the primary sequence and methylation preferences of DNA binding motifs. Figure 2 (right) shows another example, comparing the GC-box element transcription factors SP1 and EGR1.

MethylScape shows methylation relative entropy in a wider window

The cytosine methylation levels around TFBSs may relate to TF binding [34]. Therefore we developed MethylScape, a companion program to MethylSeqLogo, that can display methylation entropy, methylation probability and context probability trends around TFBSs. Figure 3 (bottom) shows MethylScape plots of MYC and ZBTB33 whole genome and promoter region binding sites. Compared to the flanking regions, MYC binding sites are hypo-methylated, in both promoter and whole genome (middle MethylScape panel), even though some CpG's can be seen near the binding sites (bottom MethylScape panel, left columns). Since hypo-methylation is somewhat less surprising in promoter regions, the CG in the center of the MYC binding site is more prominent in the whole genome MethylScape than in the promoter regions MethylScape.

Related visualization tools

Some other methods have been proposed to extend sequence logos to include DNA methylation information. MeDReaders [35] is a database summarizing methylation level with TFBS coordinates. MethMotif [36] is a database organizing tissue-specific data. Both of these resources provide methylation aware sequence logos for the convenience of their users. While Meth-eLogo [37] extends affinity (energy) sequence logos to include DNA methylation. The visual design of these tools is completely different than MethylSeqLogo (see supplementary material for a comparison).

Discussion

Caveats

When viewing MethylSeqLogos one must keep in mind the choice of background. In particular, many TFs tend to bind promoter regions, and promoter regions tend to be hypo-methylated. Thus, when using a whole genome background, MethylSeqLogos will tend to show some amount of hypo-methylation for any CpG binding TF. This effect can be seen in the logos shown in Fig. 3; under a whole genome background the MYC logo shows methylation information distributed across many positions, but under a promoter region background only some methylation information in the central binding motif CpG remains. In a narrow sense both logos faithfully depict statistical differences between binding sites and the respective background; but in terms of the impression given, a whole genome background may seem to exaggerate the importance of methylation.

On the other hand, the MethylSeqLogos displayed here may also understate the importance of methylation on TF binding. The methylation and TF binding data used here are the average of many cells from two samples (of the same cell line, but not the same cells), so if TF binding and methylation vary between cells or samples, the correlation between them will be under-estimated. Measurement noise (unless systematically biased) will also tend to decrease correlations. Therefore the correlations presented in the logos here may be reduced in magnitude.

Future work

Tailored background models

The particular definition of promoter regions we used here seems to work well, but may not always be the most appropriate. Certainly more choices could be offered, perhaps: core promoter, extended promoter, promoter + known enhancer regions, etc. Going one step further, background regions could be tailored for a given set of TFBSs, by using regions within some distance (say 50bp) of each binding site. Ideally this would be done independently for each binding site (so that a genome position near x binding sites would be included x times in the background model statistics). Thus ensuring the statistical differences depicted in MethylSeqLogo logos would be due to the binding sites (or at most their immediately flanking bases), rather than larger scale trends in methylation and/or CpG frequency across the genome.

Displaying more information

Other cytosine modifications 5-hydroxymethyl cytosine (5hmC) is an intermediate in the demethylation pathway from 5mC to unmethylated cytosine [38]. These three forms of cytosine have distinct chemical structures and may provide distinct binding affinities for DNA binding proteins [39]. But the data presented in this manuscript lumps 5mC and 5hmC together, as standard bisulfite sequencing cannot distinguish between them [40, 41]. Fortunately, data specific for 5hmC is becoming available [42] and extending MethylSeqLogo to visualize that data should be relatively manageable; perhaps modeling the distinct between 5mC and 5hmC as an additional piece of information gained after learning that a cytosine is modified in some way (is either 5mC and 5hmC). Conveniently, like 5mC, 5hmC also occurs primarily in CpG context [42]; which MethylSeqLogo already treats specially.

Other epigenetic information We briefly considered the display of other forms of DNA modification—or, more ambitiously, histone modification. We are aware of one attempt to display histone modification in a sequence logo type display, but only at a very broad resolution of introns, exons, etc. [8]. Indeed, since histone marks are not associated with single DNA residues, and in general may be positioned differently at each binding site of a TF, it is not clear where a 'column' in a histone mark sequence logo should begin and end. Thus other approaches such as juxtaposing [43] or averaging [44] heatmaps or 'wiggles', may turn out to be better suited than sequence logos for this task.

One concept from sequence logos which might be applicable would be to try making the area of histone mark wiggles proportional to some measure of their information relative to a background model. In any case, visualizing histone modification is beyond the scope of this work.

Dimer track information could be displayed as letters

Currently MethylSeqLogo displays the dimer track simply as bars indicating total column height. One could imagine using sequence-logo-like letters in this track instead of bars. So for example, "`CG`" could be drawn with height proportional to the contribution of `CpG` to the dimer information. In the examples shown in this manuscript, `CpG` is in fact responsible for the bulk of the information in the dimer track, so if rendered as "`CG`" it should be tall enough to be legible in some cases. Nevertheless, when designing the display we felt that a lettered dimer track would overall be more distracting than informative. The idea might be worth exploring in the future however, especially since the concept of a background model based dimer track is not specific to methylation and could be added to any sequence logo, even protein sequence logos.

Higher order background models

Finally we note that in principle a "trimer information track" (or even higher order tracks) could be added to the display, with each level showing the change in relative entropy resulting in incrementing the background model order. This approach might make sense in applications where the background sequence has significantly under/over-represented trimers (e.g. DNA sequences coding for proteins).

Conclusions

Sequence logos are the method of choice to visualize the nature and strength of the local primary DNA sequence contribution to TFBS selection. DNA methylation also contributes significantly to binding site selection for some transcription factors and DNA methylation data is conveniently analogous to the primary sequence data used for traditional sequence logos. Thus it is natural and desirable to extend sequence logos to include DNA methylation. We believe MethylSeqLogo has accomplished this and will prove useful.

MethylSeqLogo comes with precomputed probability models for many tissues that have published WGBS data. Users can also calculate the methylation probabilities of their own WGBS dataset with a script provided in the MethylSeqLogo package and plot on the basis of that background. Complementing MethylSeqLogo, MethylScape gives a wider view around TFBSs.

Abbreviations

H1-hESC	Human Embryonic Stem Cell line H1
TF	Transcription factor
TFBS	Transcription Factor Binding Sites
WGBS	Whole Genome Bisulfite Sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05896-2.

Supplementary file 1.

Acknowledgements

We thank Dr. Martin Frith, Dr. Toshikazu Ushijima, Dr. Naoko Ida, Dr. Tony Kuo and Dr. Matteo Pellegrini for kindly providing comments on the manuscript.

About this supplement
This article has been published as part of BMC Bioinformatics, Volume 25 Supplement 2, 2024: The Applications of Bioinformatics in Genome Research. The full contents of the supplement are available at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-25-supplement-2.

Author contributions
P.H. conceived of the study. FM.H. and P.H. designed the display. FM.H. implemented the software and selected the test cases shown. FM.H. and P.H. wrote the manuscript.

Funding
This work was supported by a Taiwanese Government Ministry of Science and Technology [Grant 108-2218-E-006-057-MY3 to P.H.].

Availability of data and materials
Source code: https://gitlab.com/fmhsu0114/methylseqlogov2. Gallery: https://feimanh.shinyapps.io/MethylSeqLogo. Data: The figures in this manuscript were generated based on publicly available data. The sources are listed in the main text.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare they have no Competing interests.

Received: 10 November 2022 Accepted: 8 August 2024
Published online: 09 October 2024

References

1. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009;10(4):252–63.
2. Spitz F, Furlong EE. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13(9):613–26.
3. Liu Y, Chen S, Wang S, Soares F, Fischer M, Meng F, et al. Transcriptional landscape of the human cell cycle. Proc Natl Acad Sci USA. 2017;114(13):3473–8.
4. Bhagwat AS, Vakoc CR. Targeting transcription factors in cancer. Trends Cancer. 2015;1(1):53–65.
5. Mitchell SM, Frayling TM. The role of transcription factors in maturity-onset diabetes of the young. Mol Genet Metab. 2002;77(1–2):35–43.
6. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The human transcription factors. Cell. 2018;172(4):650–65.
7. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990;18(20):6097–100.
8. Dey KK, Xie D, Stephens M. A new sequence logo plot to highlight enrichment and depletion. BMC Bioinform. 2018;19(1):473.
9. Kiesel A, Roth C, Ge WW, Wess M, Meier M, Söding J. The BaMM web server for de-novo motif discovery and regulatory sequence analysis. Nucleic Acids Res. 2018;46(W1):W215-20.
10. Siebert M, Söding J. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. Nucleic Acids Res. 2016;44(13):6055–69.
11. Xin B, Rohs R. Relationship between histone modifications and transcription factor binding is protein family specific. Genome Res. 2018;28:321–33.
12. Tate PH, Bird AP. Effects of DNA methylation on DNA-binding proteins and gene expression. Curr Opin Genet Dev. 1993;3(2):226–31.
13. Zhu H, Wang G, Qian J. Transcription factors as readers and effectors of DNA methylation. Nat Rev Genet. 2016;17(9):551–65.
14. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science. 2017;356(6337):eaaj2239.
15. Hérberlé É, Bardet AF. Sensitivity of transcription factors to DNA methylation. Essays Biochem. 2019;63(6):727–41.
16. Kribelbauer JF, Lu XJ, Rohs R, Mann RS, Bussemaker HJ. Toward a mechanistic understanding of DNA methylation readout by transcription factors. J Mol Biol. 2019. https://doi.org/10.1016/j.jmb.2019.10.021.
17. Lister R, Ecker JR. Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res. 2009;19(6):959–66.
18. Viner C, Johnson J, Walker N, Shi H, Sjöberg M, Adams DJ, et al. Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet. bioRxiv. 2016.
19. Guo F, Yan L, Guo H, Li L, Hu B, Zhao Y, et al. The transcriptome and DNA methylome landscapes of human primordial germ cells. Cell. 2015;161(6):1437–52.

20. Gkountela S, Zhang KX, Shafiq TA, Liao WW, Hargan-Calvopina J, Chen PY, et al. DNA demethylation dynamics in the human prenatal germline. Cell. 2015;161(6):1425–36.
21. Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat. 1951;22(1):79–86.
22. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. J Mol Biol. 1986;188(3):415–31.
23. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
24. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2018;46(D1):D1284.
25. Chèneby J, Gheorghe M, Artufel M, Mathelier A, Ballester B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. Nucleic Acids Res. 2018;46(D1):D267-75.
26. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. Genom Hum Genet. 2006;7(1):29–59.
27. Beaulieu ME, Castillo F, Soucek L. Structural and biophysical insights into the function of the intrinsically disordered Myc oncoprotein. Cells. 2020;9(4):1038.
28. Perini G, Diolaiti D, Porro A, Della VG. In vivo transcriptional regulation of N-Myc target genes is controlled by E-box methylation. Proc Natl Acad Sci USA. 2005;102(34):12117–22.
29. Daniel JM, Reynolds AB. The catenin p120^{ctn} interacts with kaiso, a novel BTB/POZ domain zinc finger transcription factor. Mol Cell Biol. 1999;19(5):3614–23.
30. Pierrea CC, Hercules SM, Yates C, Daniel JM. Dancing from bottoms up: roles of the POZ-ZF transcription factor Kaiso in cancer. Biochim Biophys Acta Rev Cancer. 2019;1871(1):64–74.
31. Daniel JM, Spring CM, Crawford HC, Reynolds AB, Baig A. The p120^{ctn}-binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides. Nucleic Acids Res. 2002;30(13):2911–9.
32. Raghav SK, Waszak SM, Krier I, Gubelmann C, Isakova A, Mikkelsen TS, et al. Integrative genomics identifies the corepressor SMRT as a gatekeeper of adipogenesis through the transcription factors C/EBPb and KAISO. Mol Cell. 2012;46(3):335–50.
33. Blattler A, Yao L, Wang Y, Ye Z, Jin VX, Farnham PJ. ZBTB33 binds unmethylated regions of the genome associated with actively expressed genes. Epigenet Chromatin. 2013;6(13):1–18.
34. Siegfried Z, Eden S, Mendelsohn M, Feng X, Tsuberi BZ, Cedar H. DNA methylation represses transcription in vivo. Nat Genet. 1999;22(2):203–6.
35. Wang G, Luo X, Wang J, Wan J, Xia S, Zhu H, et al. MeDReaders: a database for transcription factors that bind to methylated DNA. Nucleic Acids Res. 2018;46(D1):D146-51.
36. Xuan Lin QX, Sian S, An O, Thieffry D, Jha S, Benoukraf T. MethMotif: an integrative cell specific database of transcription factor binding motifs coupled with DNA methylation profiles. Nucleic Acids Res. 2019;47(D1):D145-54.
37. Zuo Z, Roy B, Chang YK, Granas D, Stormo GD. Measuring quantitative effects of methylation on transcription factor-DNA binding affinity. Epigenet Chromatin. 2014;3(35):eaao1799.
38. Shi DQ, Ali I, Tang J, Yang WC. New insights into 5hmC DNA modification: generation, distribution and function. Front Genet. 2017;8:100.
39. Sayeed SK, Zhao J, Sathyanarayana BK, Golla JP, Vinson C. C/EBPβ (CEBPB) protein binding to the C/EBP|CRE DNA 8-mer TTGC|GTCA is inhibited by 5hmC and enhanced by 5mC, 5fC, and 5caC in the CG dinucleotide. Biochim Biophys Acta. 2015;1849(6):583–9.
40. Jin SG, Kadam S, Pfeifer GP. Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. Nucleic Acids Res. 2010;38(11): e125.
41. Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. PLoS ONE. 2010;5(1): e8888.
42. Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A, et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell. 2012;149(6):1368–80.
43. Zhou X, Maricque B, Xie M, Li D, Sundaram V, Martin EA, et al. The human epigenome browser at Washington University. Nat Methods. 2011;8(12):989–90.
44. Shen L, Shao N, Liu X, Nestler E. ngs.plot: quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genom. 2014;15(284):1–14.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.