bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finding *De novo* methylated DNA motifs

Vu Ngo and Wei Wang
Bioinformatics and Systems Biology Program, Department of Chemistry and Biochemistry, University of California, San Diego, CA 92093-0359

Correspondence: wei-wang@ucsd.edu

Abstract

There is increasing evidence that posttranslational modifications (PTMs) such as methylation and hydroxymethylation on cytosine would greatly impact the binding of transcription factors (TFs). We modified a motif finding method Epigram to discover methylated motifs and other motifs containing other PTMs. When applied to TF ChIP-seq and DNA methylome data in H1 and GM12878, our method successfully identified novel methylated motifs that can be recognized by the TFs or their co-factors. We also observed spacing constraint between the canonical motif of the TF of interest and the newly discovered methylated motifs, which suggests operative recognition of these cis-elements by collaborative proteins.

Introduction

DNA methylation is a major epigenetic mark that plays crucial roles in many key biological processes[1]. Promoters with high DNA methylation levels are correlated with repressed genes, whereas promoters with low levels of methylation generally observe high levels of gene expression. Previously, it was understood that DNA methylation correlates with low levels of TF-DNA binding and subsequently gene repression. New studies have shown that certain TFs do, in fact, bind to methylated sequences[2] and these sequences are often different from the canonical motifs. For example, Hu *et. al* has discovered that Kruppel-like factor 4 (KLF4) can bind to CCmCGCC sequence, which is different from its canonical motif (CACACC)[2]. This shows that DNA methylation is not always involved in gene repression but can potentially be an important element in facilitating transcription.

Currently there is no computational method to identify methylated motifs. Here we present a new version of our published motif finding method Epigram[3] to identify methylated motifs. We added new bases representing such as methylated cytosine (mC) and modified the original Epigram program to allow for handling modified bases. This new algorithm provides a new tool that can identify DNA motifs with PTMs, which would be useful for analyzing the epigenomic data mapping different PTMs such as non-CpG methyl cytosine, carboxy cytosine, formyl cytosine, hydroxymethyl cytosine.

Methods and Results:

mEpigram:

Previously, we developed a de novo motif finding algorithm called Epigram, which was designed to identify ungapped motifs in very large sets of sequences. Epigram can identify motifs in as many as 980,465 sequences, with a mean length of 1640bp[3]. Such a large set of sequences

bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

would be impractical for other motif finding programs to process. For example, HOMER would simply crash given such a data set[3,4]; MEME only accepts input size of ≤ 60000 characters with sequence lengths of ≤ 1000 base pairs[5,6]. Epigram's scaling efficiency is comparable to that of DREME[7] but it can find motifs longer than 8 base pairs, which is DREME's motif length[7]. The program discovers motifs by building position-specific weight matrices (PWMs) from the most enriched k-mers in the positive sequences over the negative sequences as "seeds" and extending the motifs to both directions[3]. In order to incorporate methylation information into Epigram, instead of having the conventional 4 bases (A, C, G, T), we introduce a fifth base E to represent methylated cytosines (mC) and refer to the new version of Epigram as mEpigram. Since both mC and C pair with G, the algorithm can lose information if it attempts to find the reverse compliment of mC. As the vast majority of CpG methylations are symmetrical[8] and all of the known TF-binding m-motifs have been shown to contain mCpGs[2], the current version of mEpigram has an option to only consider mC's in CpG dinucleotides, which greatly reduces running time; in this mode, the reverse compliment of mCpG is also mCpG.

mEpigram is also capable of finding motifs for other cytosine modifications (non-CpG methyl cytosine, carboxy cytosine, formyl cytosine, hydroxymethyl cytosine). Since these modifications are not symmetrical, we introduced another character, F, to mark the guanine that pairs with the modified cytosine. This approach creates a more complex alphabet thus increasing the running time of the algorithm; therefore we only allow for one type of modification at a time. The default seed length for this mode of mEpigram is 7 instead of 8 for the CG methylated mode.

<u>mEpigram's motif scanner</u>:
Because of the introduction of the new bases, a new motif-scanning tool is needed in Epigram to search for matching k-mers using the modified motifs (m-motifs). To scan for the occurrences of a motif of interest in a set of DNA sequences, the program first simulates a score distribution for the motif by dinucleotide-shuffling the input sequences and calculates the scores for all of the k-mers inside the shuffled sequences using the motif's PWM. Only sequences longer than 20bps were used in this step to ensure that they can be dinucleotide-shuffled. The shuffling is repeated until an adequate number of scores is achieved, which was chosen to be at least 1 million. Matches were called based on their p-values. The score threshold is estimated so that only a fraction equaling to the p-value can pass. For example, the score threshold for p-value of 0.01 is the lowest score in the top 1% of the k-mers from the shuffled sequences. The score of a k-mer given a motif (represented as a PWM) is calculated as $S = log\ (\frac{\prod_{i=1}^{w} P_i(x_i)}{\prod_{i=1}^{w} P_b(x_i)})$), where *w* is the motif width, $P_i(x_i)$ and $P_b(x_i)$ are the probabilities of observing nucleotide x_i at position *i* from the motif and the background distributions, respectively.

<u>*1. Testing via simulation:*</u>

Retrieving inserted methylated motifs:
We first evaluated whether our motif finder can pick up methylated motifs inserted into the sequences. We inserted motifs by randomly generating k-mers from their PWMs. We chose the H1 EGR1 peaks data since its size is about the median of the datasets we retrieved from

bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ENCODE (8743 sequences at 260bps per sequence on average). The k-mers were inserted into 5%, 2%, 1% and 0.5% of the EGR1 randomly selected peaks. The m-motifs inserted have information content ranging from 2 to 0.229 per position **(Table 1).** The matching k-mers were then collected to reconstruct the inserted motif. To calculate the similarity between the original m-motif and the reconstructed one, we aligned them by sliding one over another and calculated the Pearson correlation for each alignment. The minimum of overlap of two PWMs was at 4 bps.

Depending on the information content (IC) of the motifs, mEpigram can identify m-motifs that are present in at low as 0.5% of the sequences (in the case of completely conserved motif) or not able to find motif even at 5% abundances (in the case of least conserved motifs) (**Table 1**). Since most motifs have average IC between 0.5 and 1.0, we expect the program to be able to find novel m-motifs. The simulated data shows that mEpigram can correctly identify methylated motifs if the motifs are adequately conserved and abundant.

Choosing p-value threshold for the motif scanner:
We found that the lower p-value cutoffs gave higher precision but lower sensitivity while the higher cutoffs gave lower precision and higher sensitivity. Among the p-value cutoffs tested, 0.0001 is the most appropriate value: for motifs with IC per position of 0.6-1.0, which is the amount of information we expect from real motifs, the sensitivity is about 0.6 to 0.93 (**Table S2**) while the precision stays above 0.5 for motifs inserted into more than 2% of the sequences. Thus, the default p-value cutoff was set at 0.0001 for scanning motifs.

Choosing enrichment threshold:
To choose an enrichment cutoff for the motifs, we generate a distribution of possible enrichment scores given shuffled sets of sequences. The distribution is approximately normal. The enrichment score of 1.5 is approximately 3 standard deviations from the mean (**Figure 2**). Therefore, we chose 1.5 as the enrichment cutoff for our motifs.

2. Analysis of H1 and GM12878 data:

mEpigram was used to discover motifs including both non-methylated and methylated motifs in 55 TF ChIP-seq datasets of H1 and 44 datasets of GM12878. For H1, the processed whole genome bisulfite sequencing (WGBS) data was obtained from the Roadmap Epigenomics project[9], for GM12878, the raw WGBS data was obtained from the ENCODE consortium[10]. The ChIP-seq datasets contain from 1,113 to 75,690 peaks each, with an average of 13,980; peaks are defined by the ENCODE consortium[11]. The average length of each peak ranges from 152 to 1,426bp. In the mEpigram runs, the maximum number of motifs to be reported was set at the default 200. Motifs from the same run were aligned to each other and redundant motifs were removed. 63 to 123 motifs were reported for each dataset in the end (**Table S1.2**).

mEpigram was able to identify canonical motifs for most of the TF's (**Table S1.1).** Motifs found are compared to known motifs using TomTom[12]. A TF's canonical motif is considered found if the motif appears in the top 5 most enriched of all the found motifs. The enrichment for a motif was defined as the number of peak sequences containing the motif divided by the number of

bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

shuffled sequences containing the motif. Our customized motif scanner was used to identify matches in these sequences. The p-value cutoff was set at 0.0001.

The number of m-motifs found for each TF range from 0 to 41 (**Table S1.2**). The program can identify motifs in both CG-rich (e.g NRF1 66.7%) or non-CG-rich (e.g USF1 52.6%) peaks **(Figure 1).**

Out of 55 ChIP-seq datasets in H1, 31 show enrichment for m-motifs at higher than 1.5 (**Table S1.2**). For GM12878, 24 out of 44 datasets have significantly enriched motifs. In most of these cases, the m-motif is significantly different from the canonical one or the most enriched motif present **(Table 2**).

In general, fewer m-motifs were found for each TF in GM12878. When found, they tend to contain higher E and G probabilities and have lower enrichments. (**Table S1.2**). From the GM12878 WGBS data, there are 14109049 CpG loci with beta-value of having mC being greater than or equal to 0.5, which is significantly lower than that of the H1 data (roughly 25 millions loci). This, couples with the different peak locations, results in a lot of the ChIP-seq peaks for GM12828 not being as methylated as compared to those in H1. The most significant difference is that of CEBPB. Between the two cell types, the number of peaks differs significantly (15557 for H1 and 5786 for GM12878); only 2.14% of the base pairs are shared with CEBPB peaks in GM12878. The amount of mC in the peaks is also drastically different: 33484 for H1 and 1653 for GM12878

Some of the TFs in Table 2 have been previously shown to either have interactions with DNA methylation or bind with specific methylated DNA sequences. For instance, CTCF is known to bind to DNA in a methylation specific manner and CTCF binding is regulated partly by differential DNA methylation[13]. The top m-motifs for two replicates of CTCF are similar to each other (Pearson correlation is 0.931 when aligned together), this increases our confidence that this is a real m-motif. The other dataset, which didn't have the motif, used a different CTCF antibody for the ChIP-seq experiment. For CEBPB, the top motif is the methylated canonical CEBPB motif. This is explained by CEBPB's strong binding with methylated motif; it was shown that the TF binds to 39% of the methylated canonical sequence[14]. We discovered a strong m-motif for NRF1, presented in 3.68% of the peaks. It is the canonical motif methylated at its CG dinucleotide. As a comparison, the canonical motif is present in 25% of the sequences. This finding is consistent with the observation that NRF1 TF exhibits binding with methylated sequences[2].

Spatial constraint analysis: The SpaMo algorithm was employed to identify pairs of motifs that have significant distance constrains[15]. We found several significant spatial relationships between methylated motifs and other motifs within a set of ChIP-seq peaks. In **Table 3**, we can see that some m-motifs exhibit highly significant spacing constraints with other motifs, most notably are motifs from CTCF. These CTCF motifs have enrichments over 1.5 and the SpaMo analyses gave the adjusted p-values of less than 10^-3.

Cross-scanning between H1 and GM12878 m-motifs:

To identified m-motifs that are present in both cell types and ones that are only present in one, we scanned motifs found in H1 peaks against GM12878 peaks and vice versa. The enrichments

bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of the m-motifs in the new cell type were calculated and compared with the enrichment in the cell type where they were discovered.

a/ Constant enrichment in both cell types:
For CTCF, ERG1, and NRF1, several m-motifs are enriched in both GM12878 and H1 data. These motifs have enrichments of over 1.5 and appear in at least 2% of the peaks in both of the datasets. The top NRF1 m-motifs found in H1 and GM12878 are very similar to each other **(Table 4.1),** and thus have similar enrichments. The top m-motifs for EGR1, on the other hand, are different from one another.

b/ Differential enrichment between the two cell types:
The top CEBPB m-motifs found in in H1 were enriched in GM12878. In general, the motifs found in GM12878 for CEBPB appear significantly less often, the maximum enrichment is 4.7 compared to 24.38 in H1 **(Table 4.2)**. This is unsurprising given the differences between the H1 and GM12878 data.

Enrichments of De-methylated m-motifs:
To evaluate the effect of cytosine methylation, for each sample, we de-methylated the m-motifs identified by mEpigram and then scanned them against the sample's de-methylated ChIP-Seq peak regions. To do this, in each PWM, at each position *i*, the probability of E, $P_i(x_i = E)$, were added to $P_i(x_i = C)$ and then $P_i(x_i = E)$ were set to zero. The results PWMs were then scanned against their respective peak regions without the methylation information (containing only A, C, T, G). We then compared the enrichment found in this test with the ones calculated previously for each m-motif. We found m-motifs that have their enrichments changed significantly and m-motifs that are equally enriched in both cases.

a/ Unchanged enrichment after de-methylation:
Some of the m-motifs, when scanned after de-methylation, retain their enrichment (**Table 5.1**). This is often because of these m-motifs are the methylated canonical motifs. For example, CEBPB and NRF1's top m-motifs are both methylated canonical motifs. Their enrichments remain relatively unchanged after de-methylation.

b/ Reduction of enrichment after de-methylation:
Some m-motifs have their enrichment significantly reduced after de-methylation (**Table 5.2**). These motifs generally contain more than 1 methylated cytosine in their sequences. Thus, removing the methylation significantly changes their enrichment.

c/ Increase of enrichment after de-methylation:
For some m-motifs, removing the methylation can actually boost their enrichment significantly (**Table 5.3**). In this case, like the first case, the methylation motif is essentially the methylated canonical motif, therefore, removing the methylation doesn't reduce enrichment.
Also, a lot of these, when de-methylated, become repeated sequences that appear often in the genome.

bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Comparison with Hu et al's study:

To compare our results with Hu et al, we scanned for their m-motifs found NRF1 and RXRA against the data we obtained. These motifs did not show enrichment in the data we used. This is likely because of the differences in the data, which can also explains why mEpigram didn't find the novel m-motifs reported in their study.

Methods:

Bisulfite sequencing data processing:

We used Bismark[16] to process the GM12878 whole genome bisulfite sequencing data from the ENCODE project[10]. We chose bowtie2 as the option for the process. In addition, we ignore the first 3 nucleotides on read number 2 of each pair-end reads since they have the tendency to be biased. For extracting methylation data, we only consider loci with coverage of at least 2.

Spatial Analysis:

We carried out functional analyses of the m-motifs using our custom scripts. The approach is based on SpaMo[15]. The peak sequences were reformatted into 500bp-long sequences centered at the center of each original peak. RepeatMasker[17] was used to mask repeated sequences with chains of "N" to reduce false positives. We first used our motif scanner to find matches for each of our novel m-motifs in TF peak sequences. The p-value cutoff chosen was 0.001, as the stricter cutoff 0.0001 did not result in many significant matches for our spatial analysis. Our primary interest is the novel methylated motifs; therefore, we used the m-motifs as the primary motifs and look for their significant binding partners. The algorithm assumes that every spacing between the primary and secondary motifs is equally probable if there is no spatial constrain between the two motifs. Then, the number of co-occurrences of the two motifs at a given displacement will follow a binomial distribution with the number of trials as the total number of co-occurrences. The P-values are calculated using this model. Bonferoni's correction was used to adjust the p-values. The p-value threshold for identifying a significant spacing was chosen at 0.001.

Cross-scan between H1 and GM12878:

The motifs found for each TF in H1 were scanned against the peaks for the same TF in GM12878, vice versa. For each TF with replicates, the peak sequences were pooled together to calculate an average enrichment. An m-motif is considered to have consistent enrichment in both cell types if it has enrichment of at least 1.5 in its original cell type, appears in at least 1% of the peaks, and it also has enrichment of no less than 1/1.5 of the original enrichment. For differentially enriched m-motifs, they have to have enrichment of at least 1.5 in their original cell type, and have ratios of enrichments of higher than 1.5 between the two cell types.

Scanning after de-methylation :

Similarly to above, the m-motif has to have original enrichment of higher than 1.5 and appear in at least 1% of the peaks

bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion:

Several novel m-motifs were identified by mEpigram. Significant spatial constraints were observed for several m-motifs, suggesting that they are biologically relevant.

The results of our analyses suggest that some m-motifs are enriched because of their similarity with other canonical motifs (e.g. NRF1, CEBPB), whereas others can be different from the canonical ones but still exhibit significant enrichments (e.g. CTCF). Interestingly, some novel m-motifs that we found, when de-methylated, become significantly less enriched. This suggests that DNA methylation plays a role in the enrichment of these motifs, taking these methyl groups away from the sequences thus disrupts these sequences.

mEpigram has shown that it can detect m-motifs in practice. This is helpful since these motifs can be later tested experimentally to verify their existences and functions. Further improvements of mEpigram will be made to optimize its performance.

bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures and Tables:

Figure 1: Some of the canonical motifs found by mEpigram. Comparisons were made using Tomtom. P-value cutoff was chosen at 10^-6. All the motifs found are shown in Table S1.

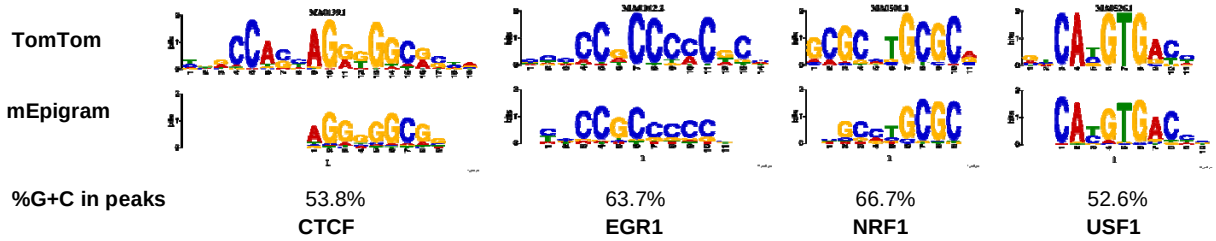
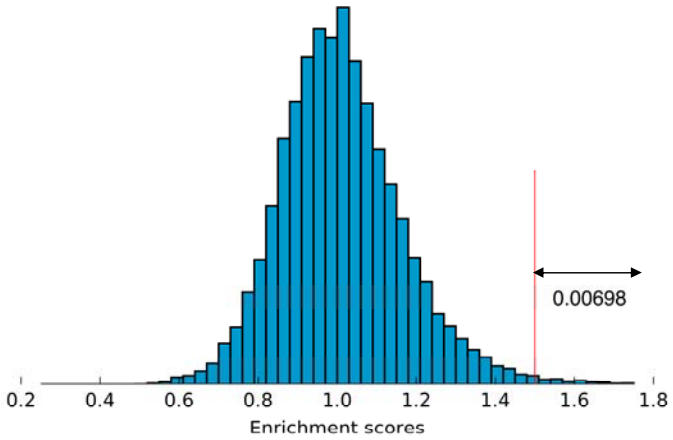
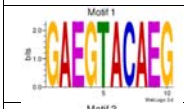
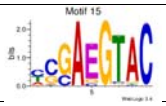
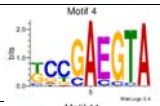
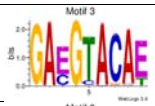
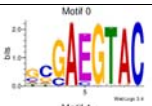
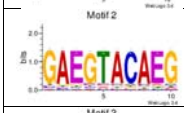
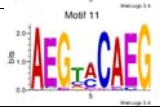
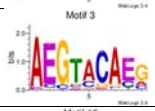
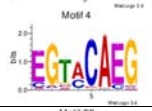
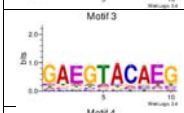
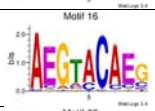
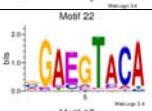
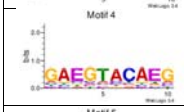
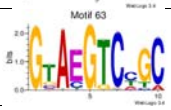
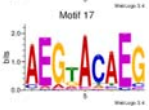
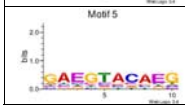


Figure 2: The enrichment of motifs in dinucleotide-shuffled sequences: Enrichment of 1.5 has p-value less than 0.01



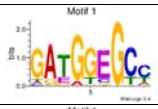
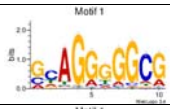
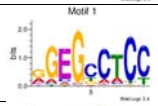
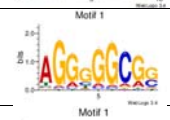
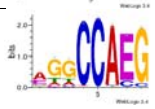
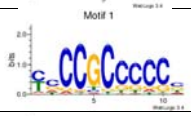
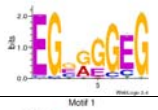
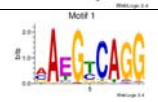
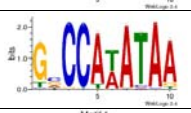
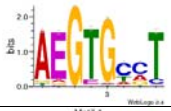
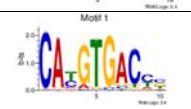
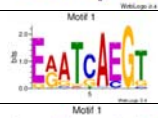
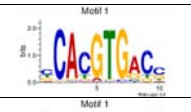
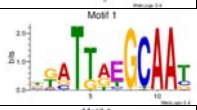
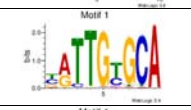
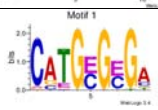
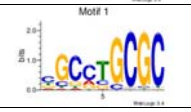
bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1: Retrieval of inserted motifs: Motifs that are abundant or highly conserved can be identified more easily.

Motif	IC/position	% of peaks			
		0.5%	1%	2%	5%
	2.0				
	1.046	Not found			
	0.626	Not found	Not found		
	0.384	Not found	Not found		
	0.229	Not found	Not found	Not found	Not found

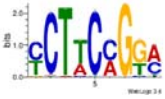
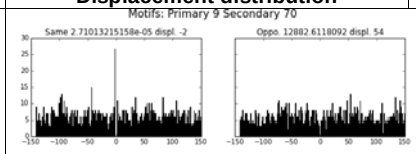
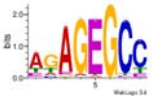
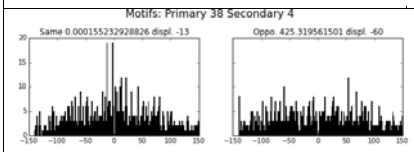
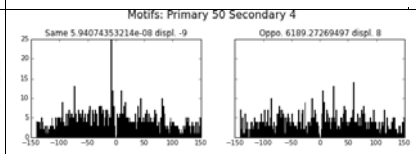
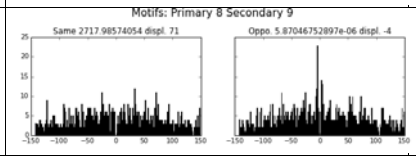
bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2: Datasets with significant m-motifs: Number of motifs and m-motifs of some of the samples with the most enriched m-motifs. See Supplemental table S1.2 for the complete list.

Sample	TF	No. Unmethyl.. motif	No. m-motifs	Top M-motif	Enrichment	Canonical	Enrichment
H1_CDS	CTCF	85	38		1.945		6.318
H1_DAR	CTCF	80	37		2.268		7.223
H1_CIV	EGR1	56	9		2.275		5.492
GM12878_CGW	EGR1	60	8		2.204		5.513
H1_CJN	SRF	99	22		2.613		16.906
H1_CJS	USF1	87	33		2.858.		27.29
H1_CRI	USF2	76	40		2.388		21.61
H1_CQR	CEBPB	89	33		24.38		6.312
H1_CRC	NRF1	58	9		3.772		15.70

bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 3: Most significant spacing pairs between m-motifs and other motifs. The motifs were scanned against Tomtom's database to identify, only matches with E-value less than 0.1 were accepted.

Sample	TF	Primary	Secondary	Displacement distribution	P-value
H1_CDS	CTCF	E2F4 H1_CDS_9	 ELK1 H1_CDS_70		2.7e-05 Displ. -2 Same strand
H1_DAR	CTCF	 Unknown H1_DAR_38	Unknown H1_DAR_4		1.55e-04 Displ.-13 Same strand
H1_DAR	CTCF	Notsure H1_DAR_50	Unknown H1_DAR_4		5.94e-08 Displ. -9 Same strand
H1_CJR	TEAD4	H1_CJR_8 Unknown	H1_CJR_9 Unknown		5.87e-06 Displ. -4 Opposite strand

bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 4.1: Some of the consistently enriched m-motifs between H1 and GM12878: The enrichments remain significant when scanned in both H1 and GM12878 data. See **Table S4.1** for the complete list.

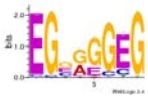
Motif	TF	Motif Logo	Original Enrichment	Cross-scanned Enrichment
H1_DAR_10	CTCF		2.269	1.6083
H1_CDS_9	CTCF		1.9458	1.5832
GM12878_CGW_41	EGR1		2.2037	1.6242
H1_CIV_13	EGR1		1.7633	1.614
GM12878_CPD_23	NRF1		2.1645	2.8571
H1_CRC_3	NRF1		3.7727	2.1428

Table 4.2: **Some of the differentially enriched m-motifs between H1 and GM12878.** See **Table S4.2** for the complete list.

Motif	TF	Motif Logo	Original Enrichment	Cross-scanned Enrichment
H1_CQR_0	CEBPB		24.3824	1.473
H1_CJN_98	SRF		2.6129	1.5679
H1_CJH_63	RXRA		2.4499	0.8621
GM12878_CHH_57	REST		2.1875	1.202

bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 5.1: Some of the m-motifs that retain high enrichments after de-methylation. See **Table S5.1** for the complete list.

Motif	TF	Motif Logo	Original Enrichment	Post-Demethyl. Enrichment
H1_CQR_0	CEBPB		24.3824	21.0314
H1_CRC_3	NRF1		3.7727	4.037
H1_DAR_10	CTCF		2.269	2.6724

Table 5.2: **Some of the m-motifs that are enriched in their methylated form but not after de-methylation.** See **Table S5.2** for the complete list.

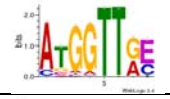
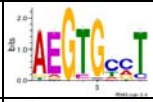
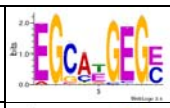
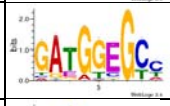
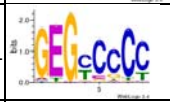
Motif	TF	Motif Logo	Original Enrichment	Post-Demethyl. Enrichment
H1_CRC_54	NRF1		2.2553	1.2424
H1_CQR_10	CEBPB		2.4566	1.2179
GM12878_CIC_47	YY1		1.8495	1.0801
GM12878_CIB_51	USF1		1.5786	1.0144

Table 5.3: Some m-motifs that have increased enrichment after de-methylation. See **Table S5.3** for the complete list.

Motif	TF	Motif Logo	Original Enrichment	Post-Demethyl. Enrichment
H1_CJS_40	USF1		2.8579	6.802
H1_CRI_13	USF2		2.3882	6.6078

bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

GM12878_CPD_23	NRF1		2.1645	13.8749
H1_CDS_9	CTCF		1.9458	3.7433
H1_CRD_36	RAD21		1.8677	3.8426

bioRxiv preprint doi: https://doi.org/10.1101/043810; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14,** 204–20 (2013).
2. Hu, S. *et al.* DNA methylation presents distinct binding sites for human transcription factors. *Elife* **2013,** 1–16 (2013).
3. Whitaker, J. W., Chen, Z. & Wang, W. Predicting the human epigenome from DNA motifs. *Nat Methods* **12,** 265–272 (2015).
4. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38,** 576–589 (2010).
5. Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34,** (2006).
6. Tran, N. T. L. & Huang, C.-H. A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. *Biol. Direct* **9,** 4 (2014).
7. Bailey, T. L. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27,** 1653–1659 (2011).
8. Guo, W., Chung, W. Y., Qian, M., Pellegrini, M. & Zhang, M. Q. Characterizing the strand-specific distribution of non-CpG methylation in human pluripotent cells. *Nucleic Acids Res.* **42,** 3009–3016 (2014).
9. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28,** 1045–1048 (2010).
10. Material, S. O., Web, S., Press, H., York, N. & Nw, A. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640 (2004).
11. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22,** 1813–31 (2012).
12. Bailey, T. L. *et al.* MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* **37,** 202–208 (2009).
13. Wang, H. *et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* **22,** 1680–1688 (2012).
14. Mann, I. K. *et al.* CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. *Genome Res.* **23,** 988–997 (2013).
15. Whitington, T., Frith, M. C., Johnson, J. & Bailey, T. L. Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res.* **39,** 1–11 (2011).
16. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27,** 1571–2 (2011).
17. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-3.0. *RepeatMasker Open-3.0* www.repeatmasker.org (1996).