Predicting the human epigenome from DNA motifs

John W Whitaker[1,3], Zhao Chen[1] & Wei Wang[1,2]

The epigenome is established and maintained by the site-specific recruitment of chromatin-modifying enzymes and their cofactors. Identifying the *cis* elements that regulate epigenomic modification is critical for understanding the regulatory mechanisms that control gene expression patterns. We present Epigram, an analysis pipeline that predicts histone modification and DNA methylation patterns from DNA motifs. The identified *cis* elements represent interactions with the site-specific DNA-binding factors that establish and maintain epigenomic modifications. We cataloged the *cis* elements in embryonic stem cells and four derived lineages and found numerous motifs that have location preference, such as at the center of H3K27ac or at the edges of H3K4me3 and H3K9me3, which provides mechanistic insight about the shaping of the epigenome. The Epigram pipeline and predictive motifs are at http://wanglab.ucsd.edu/star/epigram/.

Epigenomic modifications, including histone modifications and DNA methylation, play critical roles in development and other key biological processes. The establishment and maintenance of specific epigenomic patterns are regulated by many factors. But how exactly these mechanisms collectively regulate the epigenome, and the importance of *cis*-regulatory motifs, remains unclear.

The genome sequence is unchanged between an individual's different cell types; but the expression and activity of chromatin-modifying enzymes and their cofactors vary between cell types and cellular conditions (**Fig. 1a**). Epigenomic regulatory mechanisms use combinations of enzymes and cofactors to read a *cis*-regulatory code that defines locus-specific modification patterns. Therefore, given a particular epigenomic state, it is possible to identify the *cis* elements that interact with epigenomic modifications and are responsible for their establishment and/or maintenance (**Fig. 1b**). A global picture of the *cis*-regulatory code that regulates the epigenome may emerge from surveying a diversity of cell types and conditions. Indeed, evidence supporting the importance of *cis*-regulatory code in shaping the epigenome is rapidly accumulating[1]. For example: (G+C)-rich sequences are strongly correlated with trimethylation of histone 3 lysine 27 (H3K27me3; ref. 2) and H3K4me3 (ref. 3); (G+C)-rich motifs establish H3K27me3 by recruiting the Polycomb repressive complex 2 (PRC2) through interaction with long noncoding RNAs[4,5]; the CpG-binding protein, CFP1, recruits the H3K4 methyltransferase SETD1 to (G+C)-rich motifs[3]; and another H3K4me3 methyltransferase, PRDM9, has a sequence-specific binding motif that directs it to meiotic recombination hotspots[6].

Despite these suggestive observations, methods to systematically catalog the epigenome's *cis*-regulatory program are lacking. Studies of nucleosome positioning[7] have identified an ~10-bp periodicity of A+T dinucleotides that oscillates out of phase with the dinucleotide G+C[8–10] and poly(dA-dT) tracks that inhibit nucleosome formation[11–13]. However, the involvement of DNA sequence in nucleosome positioning remains controversial[14]. Nevertheless, these studies did not intend to predict histone modifications from DNA sequence. Enrichment of transcription factor (TF) binding and sequence features in various chromatin states have been examined[15], but DNA motifs were not used to predict epigenomic modification. Recently, DNA 6-mers were used to predict the presence of H3K4me3 with reasonable accuracy but failed to find sequence features associated with other histone modifications[16]; notably, this study did not focus on DNA motifs, which are recognized by DNA-binding factors.

Herein, we used our analysis pipeline, Epigram, to capture the *cis* elements that interact with the dynamic regulatory program to shape the epigenome (**Fig. 1b**). By surveying various cell types, we revealed mark-specific motifs, which may be universally recognized by chromatin-modifying enzymes, and motifs with cell type–specific interplay, which may be recognized by cell type–specific cofactors. We applied this approach to predicting the placement of six histone modifications and DNA methylation valleys (DMVs) in five cell types[17]: human embryonic stem cells (H1), neural progenitor cells (NPC), trophoblast-like cells (TBL), mesendoderm cells (ME) and mesenchymal stem cells (MSC) (**Fig. 1c**). To tease out the *cis* elements that are recognized by epigenomic regulatory factors, we removed simple sequence biases such as G+C content during analysis. We observed that motifs have location preferences within modified regions, such as the center of H3K27ac or the edge of H3K4me3 or H3K9me3. Furthermore, we demonstrated the importance of Epigram motifs in the regulation of histone modification through the significant correlation between their disruption and inter-individual H3K27ac variation. Our study provides a catalog of *cis* elements that play important roles in shaping the epigenomic modifications, which is useful for designing new epigenome-editing tools.

[1]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California, USA. [2]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, California, USA. [3]Present address: Research & Development IT, Janssen Pharmaceutical of Johnson & Johnson, San Diego, California, USA. Correspondence should be addressed to W.W. (wei-wang@ucsd.edu).

RECEIVED 10 FEBRUARY; ACCEPTED 23 JUNE; PUBLISHED ONLINE 21 SEPTEMBER 2014; DOI:10.1038/NMETH.3065

RESULTS

Predicting the epigenome from DNA motifs

We first examined whether DNA motifs could distinguish genomic regions that possess modified histones from regions that do not possess any modified histones. For the sake of discussion, we refer to this as the 'single-mark analysis'. We started by correcting a potential bias in the chromatin immunoprecipitation–sequencing (ChIP-seq) data that can be caused by the preferential sequencing of (G+C)-rich genomic fragments[18,19] (**Fig. 1d**). To identify regions that were enriched with a histone modification from ChIP-seq, we called two types of peaks: tight for H3K27ac, H3K4me1and H3K4me3; broad for H3K27me3, H3K36me3 and H3K9me3. The genome sequence of peaks from a specific modification formed the foreground for *de novo* motif finding. The background sequences consisted of genomic regions not covered by any histone modification peak. Next we performed sequence-set balancing (SSB; **Fig. 1e**, Online Methods, **Supplementary Figs. 1** and **2** and **Supplementary Note**) to ensure that discrimination was coming from complex motifs and not G+C biases that are linked to modification placement. Identifying motifs that are enriched within the peaks is challenging, as the methodology must be able to efficiently analyze tens of thousands of variable-length regions. Thus, we employed two *de novo* motif-finding methods: Homer[20] and Epigram's own algorithm, as we found that the combination of both was more effective at predicting modification than either alone. In particular, Epigram is able to identify

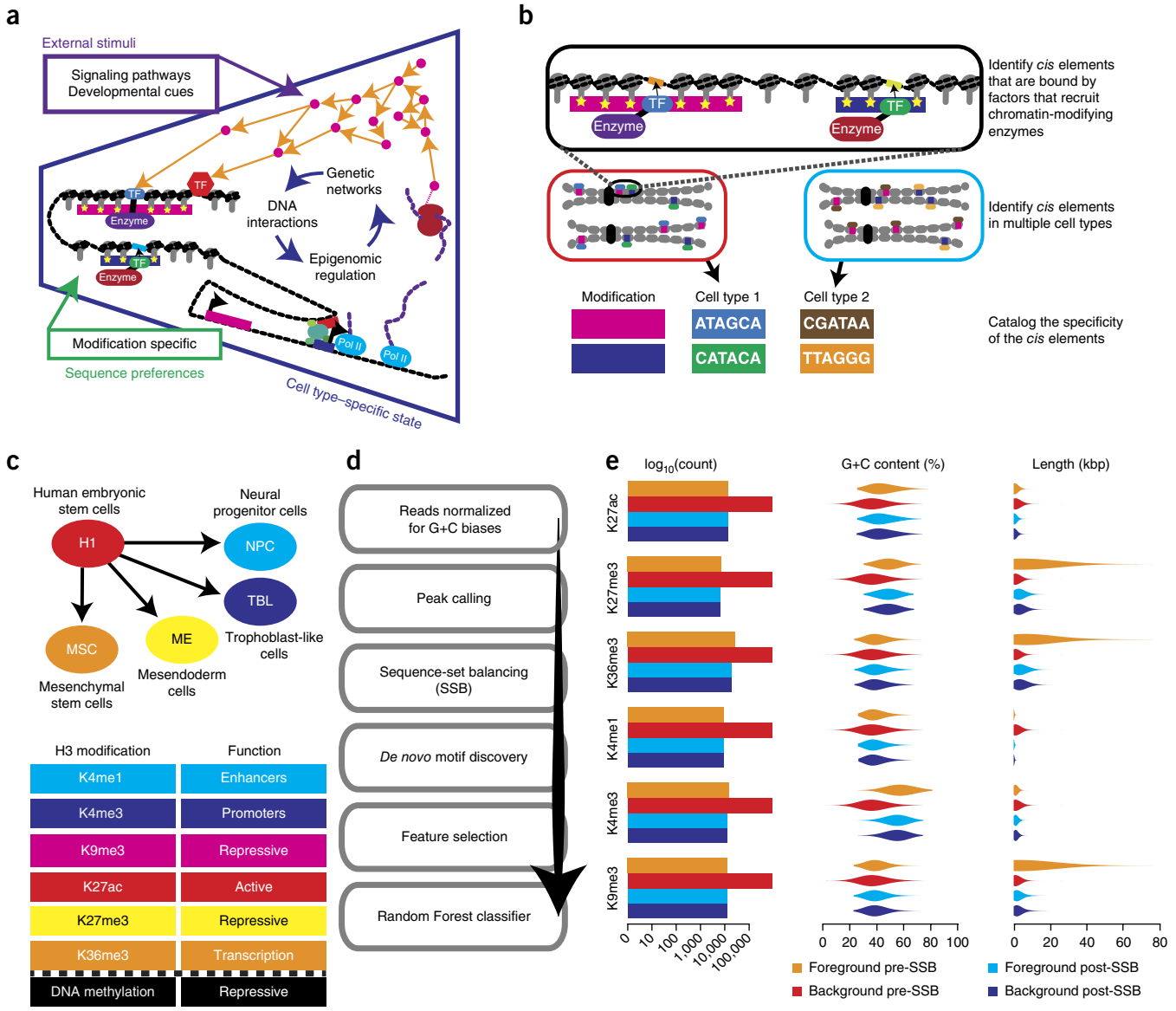

Figure 1 | Identifying motifs that are predicative of epigenomic modifications. (**a**) Site-specific DNA-binding factors regulate the epigenome. The blue section shows three regulatory levels of the cell type–specific state: (i) gene regulatory network, (ii) site-specific DNA-binding factors and (iii) epigenomic regulation of gene expression. The green square represents non-cell-type–specific DNA sequence regulatory influences over the epigenome. The purple box shows stimuli that influence the cell type–specific state. (**b**) Overview of the *cis*-element cataloging process. (**c**) Schematic showing H1 human embryonic stem cells and the other four cell types that were derived through *in vitro* differentiation. The table lists the analyzed epigenomic modifications. (**d**) Flow chart of the key stages in our analysis pipeline. (**e**) Effect of sequence-set balancing (SSB) on sequences sets. The bar plot shows the number of regions in a set before and after SSB. Violin plots show the distribution of region G+C content and length before and after SSB.

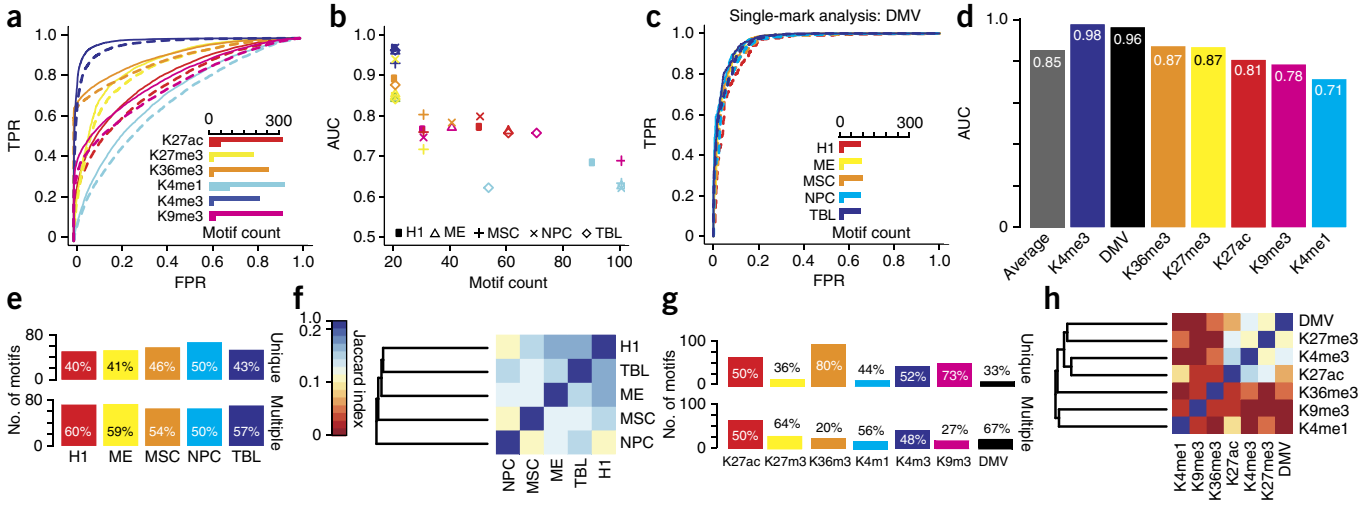

Figure 2 | Predicting epigenomic modification from DNA motifs. (**a**) Receiver operating characteristic (ROC) curve showing the prediction performance in H1. Solid and dashed lines show the full and reduced models, respectively. Inset, number of motifs used in the full and reduced models. (**b**) Performance summarized across all cell types. TPR, true positive rate; FPR, false positive rate; AUC, area under the ROC curve. The same color scheme is used to represent the H3 marks in the bar chart and the ROC curve in **a** and the scatter plot in **b**. (**c**) ROC curves showing the DNA methylation valley (DMV) predictions performance. (**d**) Averaged results across five cell types for the single-mark analysis (full model). (**e**) Number of motifs from each cell type that are predictive of modification in only that cell type (unique) or that are also predictive of modification in other cell types (multiple), calculated using 589 motif groups. We excluded motifs from the cell type–specific comparison, as H1 is featured in multiple comparisons, and motifs enriched in the background, as they are not specifically predictive of modifications. (**f**) Heat map showing the proportion of shared motifs between each pair of cell types. The Jaccard index was used to measure overlap, and clustering was performed using the complete linkage method. (**g**,**h**) Results as in **e**,**f** but with respect to modification specificity.

predictive motifs in very large sets of sequences. For example, Epigram could identify predictive motifs in 980,465 sequences with a mean length of 1,640 bp, whereas Homer could not. For the purpose of feature selection, we next exploited a LASSO[21] logistic regression to classify the foreground and background using the found motifs. Only the motifs with nonzero coefficients were kept to create the full set of motifs, which were then input to a Random Forest classifier. To improve interpretability, we clustered the motifs and retained, from each cluster, a single motif: the one with the best AUC, that is, the area under the receiver operating characteristic (ROC) curve. The reduced model motif set was the lowest number of motifs that could achieve an AUC of >95% of the full model's AUC. We assessed our method's performance through fivefold cross-validation, and, to avoid a biased inflation of predictability, we performed *de novo* motif discovery and feature selection using only the training data[22].

The selected motifs could successfully discriminate modified and unmodified regions: the average full-model accuracy across all the peaks in the genome was 79%. This performance is excellent in light of the prediction challenges: (i) there was a large number of sequences in each set, (ii) the region sizes were variable, (iii) the sequence sets were greatly unbalanced for G+C content and region size, and (iv) prediction requires the identification and combined predictive power of motif combinations. The excellent performance was also reflected by the average AUC in cell line H1 of 0.85 for the full model (270 motifs) and 0.82 for the reduced (38 motifs; **Fig. 2a**,**b**). When all cell types were averaged, the AUC was 0.84 for the full model (227 motifs) and 0.80 for the reduced (43 motifs), which shows that the total motifs can be reduced greatly while the majority of the prediction performance is maintained. Among the six marks, H3K4me3 was the most predictable in all cell types (average AUC = 0.96 for reduced models). To investigate the possible factors limiting the prediction performance, we compared the level of reads in the background for each of the modifications (**Supplementary Fig. 3**). The least predictable modification, H3K4me1, had the highest level of reads in its background, which reduces the distinction between foreground and background. The consistent prediction performance for each mark demonstrates the robustness of our model.

Next we conducted additional control analyses (Online Methods, **Supplementary Figs. 4** and **5** and **Supplementary Note**) to further demonstrate that DNA motifs are predictive of histone modification and that the prediction power does not result from sequence features involved in general nucleosome positioning or those that are associated with regions in which a modification commonly occurs, such as H3K4me3 in promoters. Furthermore, we found that DNA motifs were much less discriminative of cell type–specific modified regions for the same histone mark, which indicates the existence of mark-specific motifs. The remaining discrimination may come from the binding of cell type–specific factors, which are regulated by cell type–specific patterns of gene expression and open chromatin. These control analyses illustrate that our method can detect DNA motifs that are recognized by mark-specific chromatin-modifying enzymes and regulatory cofactors.

Motifs are predictive of DNA methylation

To further demonstrate the ability of DNA motifs to predict epigenomic modification, we applied the Epigram pipeline to DMVs, which are defined as large genomic domains (>5 kbp) that are devoid of DNA methylation[17]. DMVs have been shown to be enriched for early developmental regulatory genes and gain methylation in cancer cells, suggesting their biological importance. DMVs are relatively few in number (639–1,004 per cell type shown in **Fig. 1c**) and show substantial overlap between cell types

(461 are common to these five cell types). Therefore, we conducted only single-mark analysis on DMVs. The average AUC for the DMVs was 0.96 for the full model (95 motifs; accuracy = 0.91) and 0.95 for the reduced (**Fig. 2c**,**d**). The prediction performance remained high in all cell types, and the models all reduced down to 20 motifs, which is the lowest number of motifs assessed. Furthermore, in a separate report, the Epigram pipeline predicted the methylation status at tissue-specific differentially methylated regions from 18 human tissues (M.D. Schultz, Y. He, J.W.W., M. Hariharan, E. Mukamel *et al.*, unpublished data). Notably, the overlap between predictive motifs and single-nucleotide polymorphisms (SNPs) that 'break' motifs was compared between genotypes with DNA methylation concordance and discordance. This analysis identified a 2.6-fold enrichment of motif-breaking SNPs, a result that further supports the association between Epigram's predictive motifs and epigenomic modifications.

Comparison of DNA motif specificities

The landscape of interplay between *cis* elements and epigenomic modifications is complex (**Supplementary Fig. 6**). To pinpoint the *cis* elements that are recognized by specific factors, we conducted comparative analyses to identify mark– or cell type–specific and independent DNA motifs. The five cell types had similar proportions of cell type–specific (unique) motifs (**Fig. 2e**). The degree of motif overlap between the cell types was consistent with the known similarity between the five cell types: for example, H1 is most related to TBL but most distinct from NPC (**Fig. 2f**). The number of motifs per modification varied considerably (**Fig. 2g**). The active enhancer mark H3K27ac[23] had the most motifs, which is expected as chromatin at enhancers is dynamic across cell types[24] and thus requires more cell type–specific regulation. DMVs had the fewest motifs, which may reflect the stability of these large domains. The transcriptional activity mark H3K36me3 had the highest proportion of unique motifs, which suggests distinct motif-based regulation. H3K4me3 and H3K27ac, both enriched at active promoters, shared the most motifs (**Fig. 2h**). They formed a larger cluster with H3K27me3 and DMV because H3K4me3, H3K27me3 and DMV shared (G+C)-rich motif regulation, which is consistent with the lowest proportion of unique motifs that DMV and H3K27me3 have. Although the two active enhancer marks, H3K27ac and H3K4me1, did not cluster adjacently, their proportion of overlap was similar to that of H3K27ac and H3K4me3 (**Fig. 2h**).

To identify cell type– or mark–specific motifs, we separately clustered the motifs by cell type and modification specificity (**Fig. 3a**). The clusters contain motifs whose gene expression patterns matched their interplay with H3K27ac (**Fig. 3b**) and that had known associations with particular epigenomic modifications and cell types. For example, the SOX2 monomer motif was found to be associated with H3K27ac in H1 and NPC, whereas the OCT4-SOX2 heterodimer motif was found in only H1. This observation is consistent with the functional roles of OCT4 in H1 and SOX2 in both H1 and NPC[25]. The motif that is recognized by the four TEAD family members was associated with H3K27ac in all cell types, which is consistent with loss of H3 acetylation following deletion of a TEAD binding site[26] (**Supplementary Note**).

To systematically identify motifs that may be involved in the placement of specific epigenomic modifications, we identified those that were selected in more than one analysis but associated with only one modification. We found 56 of these motifs (**Fig. 3c**) that may represent the binding preferences of modification-specific chromatin-modifying enzymes or their cofactors. These motifs include matches to three known TF motifs that associate with H3K27ac (motif groups 457, 125 and 127 in **Fig. 4a** respectively match RUNX, GATA and HNRNPH3) and two that interplay with H3K36me3 (motif groups 142 and 240 respectively match ELSPBP1-MYOD1-MYOG and PSMD9). Two motifs matched with families of TFs (RUNX[27] and GATA[28]) that are known to be involved in embryonic development.

Predictive motifs have location preferences

The identified *cis* elements may play various roles in shaping the epigenome, such as setting the boundary of a histone modification domain or opening chromatin to allow remodeling enzymes to bind DNA. These roles may restrain the relative location (edge or center) of a motif within the modified regions (**Fig. 4a**). Although the majority of the motifs fell into the 'neutral' category (no location preference), numerous motifs showed biased location distributions (**Fig. 4b** and **Supplementary Fig. 7**). The heterochromatin mark H3K9me3 was associated with edge and neutral motifs but not with any central motifs, suggesting that the edge motifs may help set the boundary of the large H3K9me3 domain. Concentrated marks of H3K4me1, H3K4me3 and H3K27ac were associated with central motifs, which may guide the recruitment of the chromatin-modifying enzymes to initiate, or other factors to maintain, the modifications[29,30]. Interestingly, whereas the enhancer marks H3K4me1 and H3K27ac had no edge motifs, the promoter marker H3K4me3 was associated with several edge motifs, which may help define the promoter boundary. In contrast to H3K9me3, the widespread histone mark H3K27me3 and the DMV were largely associated with central motifs, which suggests different regulatory mechanisms. The transcriptional activity mark H3K36me3 almost exclusively associated with neutral motifs.

The majority (81%) of the H3K9me3 edge motifs were found in H1, and these matched the known motifs of KLF12, Rel homology domain and YY1 (**Fig. 4c**,**d**). Multiple lines of evidence support these associations. KLF12 mediates transcriptional repression through interaction with phosphoprotein CtBP[31], which forms a complex with histone methyltransferase and DNA-binding proteins to target H3K9 for methylation[32]. NFKB1, a member of the Rel homology domain TFs, is known to function with deacetylase SIRT6 to repress gene expression via H3K9 deacetylation[33], which clears the site for methylation. YY1 is a transcriptional regulator that directs localization of histone acetyltransferases, deacetylases and members of the PRC2 complex[34], which directs the placement of H3K9me3 and H3K37me3 (ref. 35). Furthermore, YY1 knockdown during mouse spermatogenesis results in global decrease of H3K9me3 (ref. 36). Further analysis of the H1 H3K9me3 edge motifs suggested that they may represent a regulatory system present in human embryonic stem cells for establishing regions of heterochromatin and repressing gene expression (**Supplementary Note** and **Supplementary Data 1**). In light of findings that show H3K9me3 as a primary epigenomic determinant during induced pluripotent stem cell reprogramming[37], we speculate that these interactions may be important in establishing and maintaining the pluripotent state.

We also observed that many of the H3K4me3, H3K27me3 and DMV central motifs had high G+C content. We defined

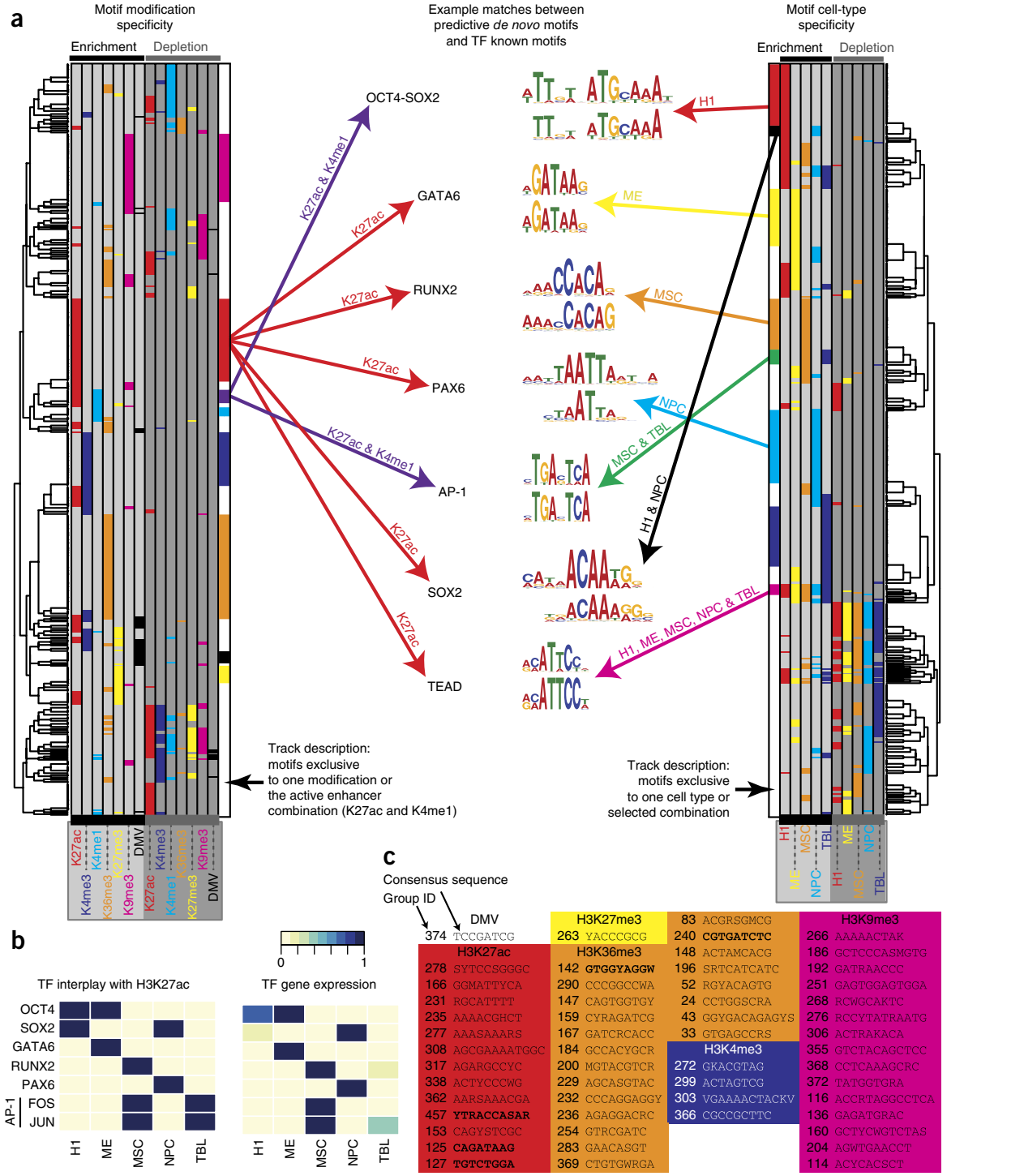

Figure 3 | The specificities of interplay between DNA motifs and the epigenome. (**a**) Left, 589 motif groups hierarchically clustered by their interplay with epigenomic modification. Each row represents a different motif, and the positions are colored if the motif associates with the modification. The first six columns show positive interplay (when a motif is enriched within a modification peaks), and the last six columns show negative interplay (when a motif is depleted in the modification peaks). The rightmost bar (and the leftmost bar on the right subpanel) indicates groups of motifs that are specific to certain modifications or combinations thereof. These bars follow the same color scheme as the heat map. Additionally, purple represents H3K4me1 and H3K27ac, which corresponds to active enhancers. Right, groups clustered by cell-type specificity. Here additional colors represent the following combinations of cell types: black, positive interplay with both H1 and NPC; green, positive interplay with both MSC and TBL; magenta, positive interplay with all cell types. Center, example motifs: the known motif (top) and the identified *de novo* motif (bottom). (**b**) Left, positive interplay between H3K27ac and TFs. Right, normalized expression values of the genes. Gene expression values were taken from ref. 17 and normalized for each gene separately. The low expression levels of FOS in TBL can be explained by the ability of JUN to bind the AP-1 binding site as a homodimer[48]. (**c**) Modification-specific motifs. The motif group numbers and consensus sequences are given. Motifs in bold match known motifs (see text).

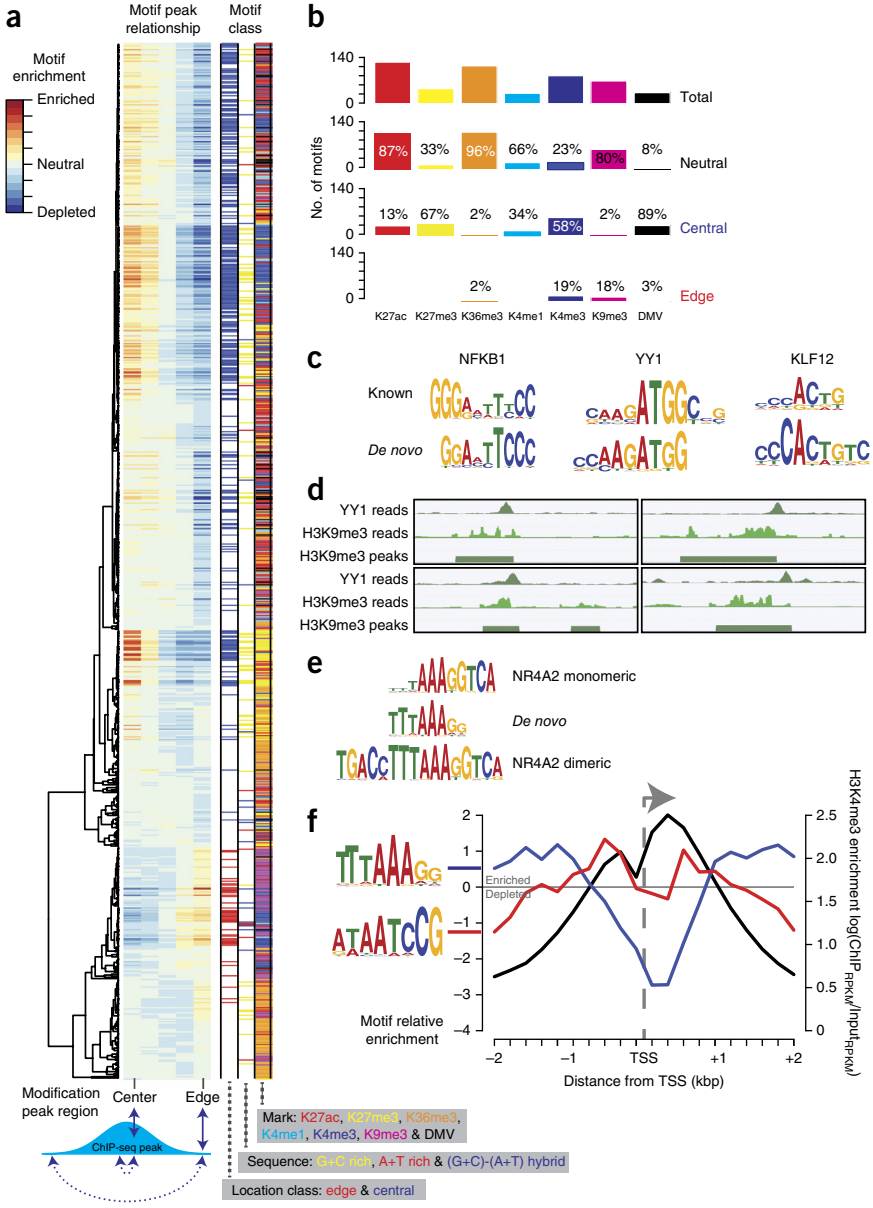

Figure 4 | Predictive motifs have location preferences. (**a**) Hierarchical clustering of 812 motifs showing positive interplay in the 'single-mark' analysis by their location preferences. The motifs were scanned against their corresponding modification peaks. The scores were then summed in five bins that represent different regions of the peaks (**Supplementary Fig. 8**). The bin scores for each motif were then hierarchically clustered. Motifs with edge or central preferences were classified by comparing edge and center bin scores and by using a χ^2 test *P*-value cutoff of 10^{-10} (Online Methods). (**b**) Summary of the location preference of motifs by mark specificity. (**c**) Motifs that associate with H3K9me3 edges in H1. NFKB1 is given as an example of a Rel homology domain. (**d**) Four screen shots[49] showing YY1 ChIP-seq reads at the edge of a region of H3K9me3. Clockwise from the top left, the four YY1 sites start at chromosome 2 (chr2) 17515620, chr6 16069456, chr12 14424514 and chr2 626745 (genome version: hg18). (**e**) *De novo* (G+C)-(A+T) hybrid motif aligned to the NR4A2 monomeric and dimeric motifs. (**f**) Average profile of two (G+C)-(A+T) hybrid motifs (red and blue lines) and H3K4me3 (black line) at 13,962 TSSs (Online Methods).

(G+C)-rich motifs as those containing >80% of positions in which (i) high-probability positions (>0.5) were G or C and (ii) the combined probability of G and C was >0.75. In total, we identified 150 such motifs (**Fig. 4a**), which were found in all cell types and enriched in all modifications. The association of TF binding and (G+C)-rich regions may explain the general abundance of (G+C)-rich motifs[38]. However, the (G+C)-rich motifs never showed negative interplay (depleted from the modification peaks) with H3K4me3, H3K27me3 and DMV, thereby suggesting a more specific association. In mouse embryonic stem cells, an artificial, promoterless and CpG-rich sequence bound by CFP1 results in H3K4me3 establishment[3]. Promoters with high G+C content tend to be repressed by H3K27me3, whereas other promoters tend to be repressed by DNA methylation[17]. Our results were consistent with these previous reports and systematically pinpointed the DNA motifs in these (G+C)-rich sequences that are responsible for forming the specific modifications. When the same criteria were reversed to identify (A+T)-rich motifs (**Fig. 4a**), only 22 motifs were found and no overall trend was observed.

When examining the (G+C)-(A+T) hybrid motifs (motifs made up of a continuous stretch of G and/or C followed by a continuous stretch of A and/or T, or vice versa), we found no overall trend of their interplay with the epigenomic modifications. However, (G+C)-(A+T) hybrid motifs whose G+C portion occupied three or fewer positions were found to prefer the edge of H3K4me3 and DMV (**Fig. 4a**). One of the (G+C)-(A+T) hybrid motifs matched the motif of the nuclear receptor NR4A2 (**Fig. 4e**). The NR4A family contains two members (NR4A1 and NR4A3) that have highly similar DNA-binding domains and are constitutively active[39]. Moreover, NR4A2 has been shown to mediate gene expression by inducing H3K4me3 and histone acetylation at the promoter of FOXP3 (ref. 40). (G+C)-(A+T) hybrid-motif enrichment at H3K4me3-enriched transcription start sites (TSSs) had two patterns: (i) TTTAAAGG was enriched ~1 kbp from the TSS on either side; (ii) ATAATCCG was enriched ~0.5 kbp from the TSS on either side (**Fig. 4f**). Consecutive runs of 3–5 pyrimidines are believed to narrow the minor groove of DNA[41] and have been shown to flank the binding sites of TFs[42]. Furthermore, poly(dA-dT) controls nucleosome positioning by forming nucleosome-depleted regions (because of their structural stiffness) around which nucleosomes are positioned[11,12]. Moreover, poly(dA-dT) tracks capped with a single G residue on the same strand as the poly(dA) have been shown to flank well-positioned nucleosomes at promoters in yeast[13]. The adjacency of the G and A nucleotides is consistent with our findings in human. Taken together, (G+C)-(A+T) hybrid motifs may define the boundary of H3K4me3- and DMV-modified regions through a combination of several possible

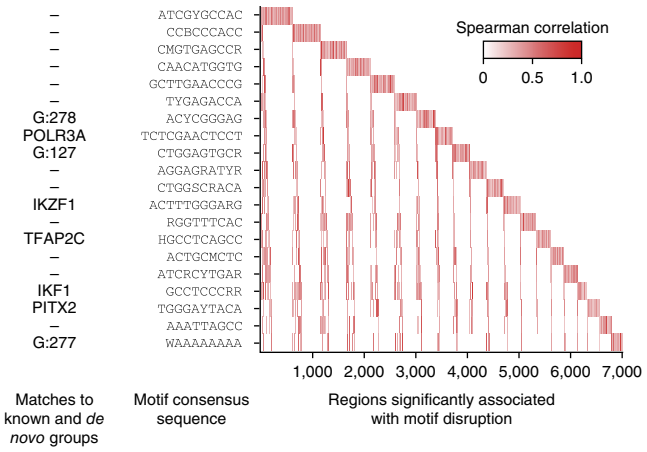

Figure 5 | *De novo* motif disruption and H3K27ac levels are correlated. The disruption of *de novo* motifs was correlated with variation in H3K27ac levels[43]. Motifs are sorted by their number of significantly correlated peaks; peaks are sorted by their associated motifs. Matches with known TFs and motif groups (from the analysis of H1 and the four derived cell types) are shown on the left. Motif groups start with 'G:'.

mechanisms: (i) by creating a nucleosome-free region around the TSS, (ii) by providing a stretch of G- and C-free sequence that cannot be bound by the factors preferring G+C regions and (iii) by being bound by TFs, such as NR4A2, that in turn recruit chromatin-modifying enzymes.

Motif disruption is correlated with H3K27ac variation

A recent study of 19 individuals correlated sequence variation at known TF motif sites with variation in H3K27ac levels at overlapping peaks[43]. Kasowski *et al.* found that H3K27ac variation in 32,886 peaks correlated with disruption of 662 known motifs by SNPs among the 19 individuals, and significant association was found in 32% of regions (significance determined using Spearman's rank and label permutation[43]). To demonstrate the power of the Epigram pipeline, we repeated the analyses done by Kasowski *et al.* by first running Epigram on H3K27ac, which resulted in a full model featuring 133 motifs that are predictive of H3K27ac (Online Methods). Epigrams motifs were significantly correlated in 62% of regions using a motif set that is ~20% the size of those used by Kasowski *et al.* (662 known motifs). Thus, Epigram discovers motifs that are significantly correlated with H3K27ac variation in 30% more regions and that represent the novel binding patterns for regulators of H3K27ac. Furthermore, Kasowski *et al.*[43] showed 20 TFs that are significantly correlated within ~4,500 variable regions, whereas the motifs from Epigram's 20-motif model are significantly correlated within 7,006 variable regions (**Fig. 5**). One of the Epigram's 20 motifs matches the known IKZF1 motif, which has been shown to target chromatin remodeling and deacetylation complexes during lymphocyte differentiation[44]. In addition, we also found that 3 of these 20 motifs match motif groups identified to be associated with H3K27ac in H1, NPC, MSC and TBL. Taken together, Epigram is able to explain significantly more variants while using fewer motifs than the Kasowski *et al.* analysis.

DISCUSSION

Although the mechanisms by which the identified motifs orchestrate the epigenome are largely unknown, these mechanisms are ultimately mediated by DNA-specific factor binding to establish locus-specific modifications, and our study represents the first step toward unveiling the enigmatic *cis* regulation of the human epigenome. In light of genome-editing technologies[45], such as transcription activator–like effector nucleases (TALENs)[46] and the clustered, regularly interspaced, short palindromic repeats (CRISPR) system[47], our study provides the first comprehensive catalog of DNA motifs to guide locus-specific epigenome editing through alteration of regulatory *cis* elements. Ultimately, the effects of sequence variation on the epigenome will be linked to phenotypic variation and disease.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was partially supported by the US National Institutes of Health (U01 ES017166 to W.W., principal investigator, B. Ren). The authors wish to thank B. Ren, D.R. Westhead and M.H. Sherman for discussion of this work. We are grateful to M. Snyder for providing the SNP data of the 19 individuals.

AUTHOR CONTRIBUTIONS

J.W.W. and W.W. conceived of and designed the project, J.W.W. performed all the analyses, Z.C. contributed to data analysis, W.W. analyzed the data, and J.W.W. and W.W. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Yuan, G.C. Linking genome to epigenome. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **4**, 297–309 (2012).
2. Mendenhall, E.M. *et al.* GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genet.* **6**, e1001244 (2010).
3. Thomson, J.P. *et al.* CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464**, 1082–1086 (2010).
4. Klattenhoff, C.A. *et al.* Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell* **152**, 570–583 (2013).
5. Tsai, M.C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689–693 (2010).
6. Baudat, F. *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327**, 836–840 (2010).
7. Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* **20**, 267–273 (2013).
8. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).
9. Yuan, G.C. & Liu, J.S. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.* **4**, e13 (2008).
10. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366 (2009).
11. Iyer, V. & Struhl, K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* **14**, 2570–2579 (1995).
12. Segal, E. & Widom, J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.* **19**, 65–71 (2009).
13. Wu, R. & Li, H. Positioned and G/C-capped poly(dA:dT) tracts associate with the centers of nucleosome-free regions in yeast promoters. *Genome Res.* **20**, 473–484 (2010).
14. Zhang, Y. *et al.* Evidence against a genomic code for nucleosome positioning. *Nat. Struct. Mol. Biol.* **17**, 920–923 (2010).
15. Ernst, J. & Kellis, M. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome Res.* **23**, 1142–1154 (2013).
16. Ha, M., Hong, S. & Li, W.H. Predicting the probability of H3K4me3 occupation at a base pair from the genome sequence context. *Bioinformatics* **29**, 1199–1205 (2013).

17. Xie, W. *et al.* Epigenomic analysis of multi-lineage differentiation of human embryonic stem cell. *Cell* **153**, 1134–1148 (2013).
18. Benjamini, Y. & Speed, T.P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72 (2012).
19. Cheung, M.S., Down, T.A., Latorre, I. & Ahringer, J. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res.* **39**, e103 (2011).
20. 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).
21. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
22. Yuan, Y., Guo, L., Shen, L. & Liu, J.S. Predicting gene expression from sequence: a reexamination. *PLoS Comput. Biol.* **3**, e243 (2007).
23. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA* **107**, 21931–21936 (2010).
24. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
25. Graham, V., Khudyakov, J., Ellis, P. & Pevny, L. SOX2 functions to maintain neural progenitor identity. *Neuron* **39**, 749–765 (2003).
26. Mauvieux, L., Villey, I. & de Villartay, J.P. TEA regulates local TCR-Jα accessibility through histone acetylation. *Eur. J. Immunol.* **33**, 2216–2222 (2003).
27. Choi, J.Y. *et al.* Subnuclear targeting of Runx/Cbfa/AML factors is essential for tissue-specific differentiation during embryonic development. *Proc. Natl. Acad. Sci. USA* **98**, 8650–8655 (2001).
28. Morrisey, E.E., Ip, H.S., Tang, Z., Lu, M.M. & Parmacek, M.S. GATA-5: a transcriptional activator expressed in a novel temporally and spatially-restricted pattern during embryonic development. *Dev. Biol.* **183**, 21–36 (1997).
29. Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**, 958–970 (2008).
30. He, H.H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nat. Genet.* **42**, 343–347 (2010).
31. Schuierer, M. *et al.* Induction of AP-2α expression by adenoviral infection involves inactivation of the AP-2rep transcriptional corepressor CtBP1. *J. Biol. Chem.* **276**, 27944–27949 (2001).
32. Shi, Y. *et al.* Coordinated histone modifications mediated by a CtBP co-repressor complex. *Nature* **422**, 735–738 (2003).
33. Kawahara, T.L. *et al.* SIRT6 links histone H3 lysine 9 deacetylation to NF-κB-dependent gene expression and organismal life span. *Cell* **136**, 62–74 (2009).
34. Woo, C.J., Kharchenko, P.V., Daheron, L., Park, P.J. & Kingston, R.E. Variable requirements for DNA-binding proteins at Polycomb-dependent repressive regions in human HOX clusters. *Mol. Cell. Biol.* **33**, 3274–3285 (2013).
35. de la Cruz, C.C. *et al.* The Polycomb group protein SUZ12 regulates histone H3 lysine 9 methylation and HP1α distribution. *Chromosome Res.* **15**, 299–314 (2007).
36. Wu, S., Hu, Y.C., Liu, H. & Shi, Y. Loss of YY1 impacts the heterochromatic state and meiotic double-strand breaks during mouse spermatogenesis. *Mol. Cell. Biol.* **29**, 6245–6256 (2009).
37. Chen, J. *et al.* H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. *Nat. Genet.* **45**, 34–42 (2013).
38. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012).
39. Wang, Z. *et al.* Structure and function of Nurr1 identifies a class of ligand-independent nuclear receptors. *Nature* **423**, 555–560 (2003).
40. Sekiya, T. *et al.* The nuclear orphan receptor Nr4a2 induces Foxp3 and regulates differentiation of $CD4^+$ T cells. *Nat. Commun.* **2**, 269 (2011).
41. Rohs, R. *et al.* Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **79**, 233–269 (2010).
42. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
43. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342**, 750–752 (2013).
44. Kim, J. *et al.* Ikaros DNA-binding proteins direct formation of chromatin remodeling complexes in lymphocytes. *Immunity* **10**, 345–355 (1999).
45. Hathaway, N.A. *et al.* Dynamics and memory of heterochromatin in living cells. *Cell* **149**, 1447–1460 (2012).
46. Miller, J.C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.* **29**, 143–148 (2011).
47. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
48. Chinenov, Y. & Kerppola, T.K. Close encounters of many kinds: Fos-Jun interactions that mediate transcription regulatory specificity. *Oncogene* **20**, 2438–2452 (2001).
49. Wang, T. *et al.* STAR: an integrated solution to management and visualization of sequencing data. *Bioinformatics* **29**, 3204–3210 (2013).

ONLINE METHODS

The data set. ChIP-seq experiments using antibodies for six different histone modifications, in five different cell types, were used to assess the predictability of histone modification from DNA motifs[17,50]. The six histone modifications are H3K4me1, H3K4me3, H3K27me3, H3K27ac and H3K36me3; and the five cell types are human embryonic stem cells (H1), trophoblast-like (TBL), neural progenitor cell (NPC), mesendoderm (ME) and mesenchymal cells (MSC). Each of the ChIP-seq experiments had at least two replicates, and input control samples are also provided. Mapped reads were made monoclonal using Homer[20].

The IMR90 data obtained from the Human Epigenome Atlas (http://www.genboree.org/epigenomeatlas/) and was originally published in ref. 50. The data for the other seven cell types (A549, CD14+, GM12878, HeLa, HepG2, HUVEC and K562) were obtained from ENCODE (a complete list of files is available in **Supplementary Data 2**). These cell types were picked as they were all the ENCODE tier 1 and 2 cell types for which the six histone marks being considered were available. The ChIP-seq data used in the further analysis of YY1 were also obtained from ENCODE (a complete list of files is available in **Supplementary Data 2**).

The UCSC Known Gene[51] and RefSeq[52] gene models were downloaded from UCSC in June of 2012 and correspond to build hg18 of the human genome.

(G+C)-content correction in mapping reads and peak calling. It has been shown that technical biases can influence ChIP-seq, resulting in a sequencing preference for (G+C)-rich fragments[18,19]. When considering the relationship between epigenome and DNA sequences, such biases could influence predictive performance. Thus, during our analysis, the G+C content of all read sets (ChIP and control) was normalized to the genome before peak calling. The G+C content bias removal was performed by Homer[20], which started by extending the sequencing reads to their fragment length and calculating their G+C content. As the fragment length of individual reads is unknown, we use the median of the range of fragment lengths. The fragment length selection range was 180–400 bp, so we used a fragment length of 290 bp (ref. 50). Then the distribution of the fragments' G+C content is calculated for a ChIP-seq experiment. This is compared to the expected distribution, which is based on the distribution of G+C content across the entire genome. The normalization works by taking each range of G+C content and comparing the expected and observed distributions. If they do not match, the fragments are assigned fractional values to adjust for the difference. For example, if the sample is more G+C rich that expected, reads in (G+C)-rich regions will be reduced to a fractional value beneath 1. Homer uses the fractional read values during peak calling.

To identify regions that are enriched with a histone modification, we called two styles of peaks: tight and broad. The tight peaks are typically <1 kbp and best represent the tightly localized modifications (H3K27ac, H3K4me1 and H3K4me3). Tight peaks were called using the Homer program "findPeaks" with the style "histone"[20]. Homer identifies 500-bp peaks that significantly enriched with reads in comparison to input control. Then peaks within 1,000 bp of each other are stitched together to form a set of peaks with variable length. The broad peaks can be as long as 80 kbp and best represent modifications that form domains (H3K27me3, H3K36me3 and H3K9me3). Broad peaks were called using the Homer program "findPeaks" with the following options "-region -size 1000 -minDist 2500". When Homer is run with these options, the initial sets of peaks are 1,000 bp wide, and peaks within 2,500 bp of each other are stitched together. Both styles of peaks were called for the domain-forming modifications, and the results were merged. Furthermore, to ensure only regions of high confidence are considered, we take the intersect (regions present in both sets) of two biological replicates as being our final set of high-confidence regions. The peaks were merged and intersect extraction was made with BedTools[53].

Making the sets of sequences for the prediction of histone modification. The 'single-mark' analysis compares regions that are enriched with an epigenomic modification to regions that do not possess any of the modifications being considered. The enriched regions, or foreground, were the high-confidence regions that were identified as the intersect of two or more replicates. To establish a background, we took all the continuous stretches in the genome that are 100% mappable but do not overlap with any of the histone modifications peaks. Regions of the genome are not 100% mappable if the DNA sequence is replicated elsewhere in the genome. This replication of DNA sequences reduces mappability, as it is a requirement of the mapping procedure that reads map uniquely. To measure regions' mappability, we used a precomputed data set that considered 35-bp reads mapping uniquely within the human genome[54]. When considering overlap between 100% mappable regions and histone modification peaks, the union of all peaks was used rather than the high-confidence regions (the intersect of two or more replicates).

The 'mark-specific' analysis compares regions that are enriched with only one specific modification to regions that are enriched with any other modification but not the modification being considered. As both the foreground and background sequence sets harbor nucleosomes, it is the enrichment of a specific histone modification that is being predicted and not a sequences propensity to contain nucleosomes. To define the enriched regions, or foreground, we used high-confidence regions that were identified as the intersect of two or more replicates. Then foreground regions that overlap the high-confidence peaks of any other modifications were removed to produce the final foreground set. To establish the background set, the high-confidence peaks of all other modifications were merged, and then any region overlapping the union of peaks from the modification of interest were removed to give the final background set.

The 'typical background' analysis compares the high-confidence peaks of a modification to a background of regions that typically possess the modification, i.e., promoters for H3K4me3, H3K27ac, H3K27me3 and H3K4me1, and gene bodies for H3K36me3 and H3K9me3. In this comparison both the foreground and background sequence sets contain sequence features that are typical of the genomic regions enriched with the modification: for example, G+C islands in promoters for H3K4me3. Thus, it is the enrichment of a modification and not its typical genomic content that is being predicted. The enriched regions, or foreground, were the high-confidence regions that were identified as the intersect of two or more replicates. The background sets were taken from the promoter or gene-body regions in the UCSC knowngene.txt file.

Promoters were extracted from −2,500 bp to +500 bp at all transcription start sites (TSSs). Any regions in the background set overlapping with a region in the foreground set were removed. For this, the union of all peaks was used rather than the high-confidence regions.

The 'cell type–specific' analysis predicts the origin of cell type–specific modification peaks. In this comparison both sets of sequences are enriched with the same histone modification but in a cell type–specific manner. For this purpose, we compared H1 to each of the other four cell types, as these were derived from H1. To compare two cell types, A and B, we first established the positive sets of sequences for cell type A by taking its set of high-confidence peaks, and we then subtracted from it the union of replicate of the same modification in cell type B. These were compared to a set that was made by reversing the roles of A and B. A caveat of this comparison is that different cell types introduce additional complexity and the identified motifs may be responsible for cell-type specificity but not necessarily related to histone modifications (see **Supplementary Note**).

$$\mathrm{CEF} = \frac{(\mathrm{Peak\,Count}/\mathrm{Peak\,Total})}{\left((\mathrm{Whole\,Genome\,Count}/\mathrm{Whole\,Genome\,Total})+(\mathrm{Shuffle\,Count}/\mathrm{Shuffle\,Total})\right)\times 0.5}$$

Sequence-set balancing. During sequence-set balancing (SSB), we adjusted the distributions of sequence G+C content and length in the foreground and background sequence sets so that they match. This procedure is crucial to avoid such differences from biasing the Random Forest predictions. To do this, we bin the sequences into a matrix that describes their length and G+C content (**Supplementary Fig. 9**). Length is measured to the nearest 100 bp, and G+C content is measured to the nearest 1%. Then the matrices that describe the foreground and background sets of sequences are compared. If the counts at a position in the two matrices are not equal, entries are randomly removed to make the counts equal. This was repeated for every entry in the matrices. At the end of the procedure, the counts in every position in the two matrices are equal.

***De novo* motif discovery.** To identify DNA motifs that are overrepresented within peak sequence, we used two *de novo* motif-finding programs: Homer[20] and our own, Epigram (**Supplementary Fig. 10a**). To identify motifs with Homer, we used the program "findMotifsGenome.pl". As Homer may miss some motifs, we developed a motif-finding algorithm to search for motifs in large numbers of long genomic regions, such as those regions covered by histone modifications. For example, Epigram was found to identify predictive motifs in 980,465 sequences with a mean length of 1,640 bp, whereas Homer crashes without producing any output. Epigram starts with identification of enriched 9-mers in the foreground and background. Similar 9-mers are merged to generate position-weight matrix (PWM) to represent a sequence motif.

1. Identification of enriched 9-mers. Epigram calculates the enrichment (EF) of 9-mers within the foreground (for example, peaks of histone modifications) in comparison to two backgrounds: (i) the entire genome (EF_{wg}) and (ii) the shuffled foreground sequences (EF_{sh}) that have the same G+C content as the foreground sequences (peaks)

$$\mathrm{EF} = \frac{(\mathrm{Peak\,Count}/\mathrm{Peak\,Total})}{(\mathrm{Background\,Count}/\mathrm{Background\,Total})}$$

The 9-mers that have both EF_{wg} >1 and EF_{sh} >1 are considered enriched and taken forward in the motif discovery process.

For a 9-mer to initiate a new motif, or PWM, it must have a combined enrichment factor, CEF, greater than an enrichment threshold ET (ET was set to 1.5 in this study), which determines the degree of enrichment a motif must have to be potentially discovered by Epigram. This level of enrichment is modest but is suitable for our purpose as it allows identification of 200–300 candidate motifs for the predicting epigenomic modifications. Furthermore, a feature selection procedure (described below) removes motifs with little predictive power to facilitate interpretation of the results

9-mers that have EF_{wg} >1 and EF_{sh} >1 but CEF <ET can be incorporated into an existing alignment but cannot be used to initiate an alignment.

2. Construction of motif PWMs. Each of the 9-mers that is taken forward is given a weight, *W*, that reflects its enrichment

$$W = \log(\mathrm{PP})\left(\log(\mathrm{EF_{wg}}+\log(\mathrm{EF_{sh}})\right)$$

where PP is the proportion of foreground sequences that possess at least one copy of the 9-mer. The top-weighted 9-mer is taken as a seed for motif construction, and all the enriched 9-mers, which differ from the seed 9-mer by one or two mismatches when perfectly aligned, are aligned to the seed 9-mer (**Supplementary Fig. 10b**). When a 9-mer alignment is converted into a PWM, the *W* scores weight the 9-mer contributes. Thus, the most enriched 9-mer would have the greatest influence on the PWM. PWMs are made only for alignments that contain five or more 9-mers.

3. Refinement of motif PWMs. The overall enrichment of a PWM is ensured by calculating the motif score, MS

$$\mathrm{MS} = \sum_{\mathrm{NNN}} \mathrm{MM}\left(\log(\mathrm{EF_{wg}}) + \log(\mathrm{EF_{sh}})\right)$$

MS is a weighted sum of all non-neutral 9-mers, NNN. The NNN are those that have a CEF >ET or CEF <1/ET. These cutoffs were used as they select the 9-mers that are over or underrepresented in the peaks at the ET being considered. For each PWM the summation of the NNN EF scores is weighted by MM, which is the best score of NNN evaluated by the PWM

$$\mathrm{MM} = \prod_{1\ldots 9} P_{ij}$$

doi:10.1038/nmeth.3065

Where P_{ij} is the probability of the nucleotide at position i of NNN_k in the PWM being evaluated. For a PWM to be taken forward, it must have an MS score greater than the ET.

4. Widening of the motif. The alignments are then widened by adding 9-mers. A new 9-mer is added to the alignment if it has 8 consecutive bases aligned to any 9-mer currently included in the alignment up to one mismatch and is offset by one position (**Supplementary Fig. 10c**). PWMs that are wider than a 9-mer have an MS calculated for each 9-mer fragment. Expansions that do not result in a decrease in MS in the existing 9-mer fragments, and where the expanded 9-mer fragment has an MS greater than ET, are accepted. Expansion will happen for a maximum for four rounds in either direction, i.e., the maximum possible length of a PWM is 17 bp. Only motifs that include 20 or more 9-mers are reported in the final output. Once a motif PWM is completed, then the process starts again with the next most enriched 9-mer being used as the seed. To prevent redundant motifs being produced, we use an enriched 9-mer in only one motif.

Note that each sequence is parsed only once by Epigram. For comparison, other methods, such as Homer[20] and DREME[55], work through an iterative procedure that masks the locations of identified motifs before the next motifs are identified. This means that Epigram is better enabled to identify motifs in some of the large sets of sequences being considered in this study owing to its efficiency. Epigram can be downloaded at http://wanglab.ucsd.edu/star/epigram/.

Selecting the most discriminative motifs while maintaining diversity of predictive motifs. Owing to dual motif discovery strategy, the found motif sets could be redundant. To reduce such redundancy, we implemented a two-stage feature selection procedure. The first stage is a LASSO penalized logistic regression to classify the foreground and background using the motifs[21]. LASSO sets the coefficients of less informative motifs to 0. These motifs are then ignored during further analysis. To perform the LASSO we use the R function "cv.glmnet" from the package "glmnet". To select the optimum parameters for the LASSO, we first run it on all the training data to obtain a lambda sequence. Lambda is the LASSO's regularization parameter, and the sequence represents the values that will be tested during each of the cross-validations. Then tenfold cross-validation is performed within the training data, during which the precomputed lambda sequence is used to train the LASSO model while ignoring one-tenth of the data. The average error at each lambda value is calculated using the combined set of results from the tenfold cross-validation. Finally, the value of lambda that gives the least-squares error is selected. The value of alpha is always kept at the default value, which is 1. This nested cross-validation avoids our models becoming over-fitted. The LASSO is performed by scanning the motifs against both foreground and background sets of sequences. The motifs are scored as S, which is the log odds between the motif and a background model

$$S=\log\frac{\prod_{k=1}^{w}P_k(x_k)}{\prod_{k=1}^{w}P_b(x_k)}$$

where w is the motif width, $P_k(x_k)$ and $P_b(x_k)$ are the probabilities of observing nucleotide x_k at position k from the motif and the background distributions, respectively. The motifs are scored against every position in a sequence, and the best single score is recorded. The frequencies of single nucleotides in the human genome were used as a background model. The scores for each motif in both sets of sequences are input to LASSO. The set of motifs that are selected by LASSO are considered as our full set of motifs.

In the second stage, we further reduced the number of motifs by hierarchical clustering using the distance D between two motif matrices, a and b, as the metric[56]

$$D(a,b)=\frac{1}{w}\sum_{i=1}^{w}\frac{1}{\sqrt{2}}\sum_{L\in\{\mathrm{ACGT}\}}\left(a_{i,L}-b_{i,L}\right)^2$$

D is the distance metric between two motifs that is normalized by the motif length, w, and adjusted by $\sqrt{2}$ to give the scores that are approximately proportional to the fraction of similarity, i.e., if D = 0.5 then the two motifs are approximately 50% the same. This adjustment makes the motif similarity values suitable for distance matrix construction, which is used during the hierarchical clustering of the motifs. As the optimal alignment of two motifs is not known, multiple alignments are measured and the minimum D is retained. All alignments are at least seven positions, or, when either motif is shorter than nine positions, by two positions fewer than the shortest motif. The distance between the motif reverse complements is also measured, and the shortest D is used for hierarchical clustering. Next the motif distance matrix is hierarchically clustered and cut multiple times, forming 20, 30, 40 … 100 groups. Within each group only the motif with the best AUC was retained. The AUC for each motif was calculated according to the motif's ability to discriminate the foreground and background sets of sequences in the training set using the motif score S and a variable cutoff. Then the Random Forest was used to evaluate the prediction performance of the subset of motifs. The lowest number of motifs whose AUC was greater than 95% of the full model's AUC was taken as the reduced model. If none of the tested motif subsets had an AUC >95% of the full model AUC, then the 100-motif subset was used. All feature selection was carried out using only the training data.

The Random Forest classifier. We chose a Random Forest classifier to make our predictions, as they have been shown to be a powerful machine learning method[57]. The Random Forest model was constructed with 1,000 trees and using the square root of the total number of features at each split. The Weka implementation of the Random Forest was used[58].

License and prerequisites. The Epigram pipeline is available for Linux. It is open access and free for not-for-profit use. If a for-profit wishes to use it, a request must be submitted to the UCSD Technology Transfer Office. The pipeline requires R and the R package "glmnet"[21] to be installed. We also recommend installing Homer[20] so that two *de novo* motifs discovery methods can be used in combination.

Run times for the Epigram pipeline. We ran the Epigram pipeline (with Epigram and Homer for motif discovery) on all H1 H3K4me3 peaks (15,229 peaks in total) compared to all 100% mappable regions of the genome that contain overlap any of the six core modifications (890,005 regions). Running the pipeline fully (with the model reduction stage) took 28 h and 20 min. This was reduced to 19 h and

10 min if the reduce stage was ignored. This was done on an Intel quad-core Xeon 2.53GHz and used less than 8 GB of RAM.

Quantitating prediction performance. To measure the performance of a model, we first count true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs). TPs are defined as modified regions that are predicted to possess a modification. TNs are defined as unmodified regions that are predicted to be unmodified. FPs are defined as unmodified regions that are predicted to be modified. FNs are modified regions that are predicted to be unmodified. We then determine accuracy as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

We also use receiver operating characteristic (ROC) curve analysis to calculate the area under the curve (AUC). This is done by varying the number of Random Forest votes that are required to make a positive prediction.

Parameterizing the value of the enrichment threshold ET. Throughout this work, Epigram's enrichment threshold, ET, was set to 1.5. To demonstrate the appropriateness of this value, we ran the Epigram pipeline on all H1 H3K4me3 peaks from chromosome 1 (1,515 peaks in total) compared to 100% mappable regions of the H1 genome chromosome 1 that do not overlap the peaks (68,616 regions in total). Homer was not used so that the effect of varying ET could be better judged. We ran the pipeline with three different values of ET: 1.75, 1.5 and 1.25. We recorded accuracy and AUC using in the full model. Varying ET did not have any significant effect on the results as all values of ET resulted in the same AUC (0.95). However, the accuracy was slightly higher when ET was 1.5 (89 vs. 88 with the other two values). Thus, ET remained set at 1.5 throughout this work.

Comparison-specific approaches to prediction. Above we have outlined the general approach that we have taken to the prediction of epigenome from DNA motifs; however, each of the comparisons has its own particular specificities that we will outline here.

For the single-mark analysis, there are many more background sequences than foreground sequences. This means that during SSB, few foreground sequences will be lost but many background sequences will be lost. For example, during the single-mark analysis of H3K4me1 in H1, the pre-SSB foreground sequence count is 9,086 and the background sequence count is 890,005; but post-SSB, both counts are 9,086. This means that only 1.02% of the background sequences are included in the post-SSB set but all foreground sequences are included. For the purpose of *de novo* motif discovery, ~1% of a set of sequences is not representative. Thus, for identifying motifs in the background set, the sequences that were removed from the background during SSB are used. As these sequences are never used for testing the Random Forest model, this does not violate the fivefold cross-validation assessment procedure. As Homer was unable to run on these very large sets of sequences, only Epigram was used at this stage.

For the mark-specific analysis and the cell type–specific analysis, the background sets of sequences are derived from histone peaks that do not overlap with the foreground set. Consequently, the proportion of foreground sequences removed during SSB is far greater in these two comparisons. For example, during the mark-specific analysis of H3K4me1 in H1, only 46% of the foreground sequences are retained in the post-SSB set. To allow the sequences that were removed during SSB to be tested by our model, we make an extension to our fivefold cross-validation procedure. For testing these sequences, a Random Forest is trained using motifs that were identified on the complete post-SSB set of sequences (**Supplementary Fig. 1**). Through this extension all the sequences that were in the original pre-SSB set are tested without ever using the same sequence in both training and testing.

During the typical background analysis, promoters or gene bodies are used as the background. For promoters, a 3-kbp-long region was taken around the TSS (from −2,500 bp to +500 bp). The gene bodies are the transcribed regions. In both cases, the length distribution of the background sequences differs considerably from the foreground sequences. Thus, many sequences are removed during standard SSB. To reduce this, we extended the standard SSB procedure by trimming the background sequences such that it allows removed sequences to be added back. To do this, the background sequences that have been removed but are longer than removed foreground sequences are trimmed to the length of a foreground sequence. Then G+C content of the trimmed sequences is measured. If it matches the foreground sequence, then the pair of sequences can be added back into the final post-SSB set. The background sequences are trimmed so that three different sections are considered: the 5′ end, the 3′ end and the middle. For example, when the 5′-end section is considered, the sequence is trimmed at its 3′ end.

During the mark-specific analysis, regions with only a single specific modification are compared to regions with any other modification. Thus, in this comparison the background sequence set for a particular modifications predication is a mixture of the foreground sets from other modification predictions. Rather than performing *de novo* motif prediction in the background sets, we select the motifs from other modification foreground sets to help the Random Forest better predict a sequence as belonging to the background set. To do this, we first identify a set of motifs in each of the modification foreground sets and perform the LASSO penalized logistic regression. Then all the selected motifs are combined into one set of motifs for that cell type, and LASSO is run again. Finally, the set of motifs that are selected in the second LASSO is used in the full model and then reduced to form the reduced model. To maintain variation in our fivefold cross-validation testing, the motifs from the same fold are combined with each other. For example, the six sets (one for each mark) that were produced in the first fold of the cross-validation are all combined and used in only the first fold of the cross-validation.

Data quality influences prediction performance. We sought to better understand how the quality of the data and peak calls might be affecting the marks predictions performance. We plotted the normalized read counts ($ChIP_{RPKM}$ – $input_{RPKM}$, where RPKM is the reads per kilobase per million mapped) from the peaks called for H1 mark-specific analysis and compared them to the normalized read counts in the background and the marks prediction performance (**Supplementary Fig. 3**). H3K4me3 has the highest enrichment of reads within its peaks. This strong of enrichment is probably a result of both biological and technical specificity: for example, the mark is very specific to promoters

doi:10.1038/nmeth.3065

and the antibody is very specific to only this modification[50]. This quality of data and peaks is likely important for H3K4me3 being the best-performing modification in the two intercomparisons. The maximum normalized read count in the background of H3K4me1 is 7.3, whereas the second highest is 3.5. Moreover, the third quartile of H3K4me1 is 0.03, whereas it is 0 for all other marks. From manual inspection of the marks through the genome browser, it can be observed that H3K4me1 is more diffusive than the other nondomain modifications. The higher read-count levels in the background of H3K4me1 reduces the distinction between foreground and background, which is likely responsible for its reduced performance.

Comparison to known motifs and motif clustering. To identify known motifs that match the *de novo* motifs, we ran Tomtom[59,60]. The minimum overlap was set as 5, and an *E*-value cutoff of 0.1 was used to define similarity. When Tomtom was run, a library of known motifs was constructed from the following five databases: Transfac[61], Jaspar[62], Uniprobe[63], hPDI[64] and Taipale[42]. When the library of known motifs was constructed, Jaspar motifs with IDs starting with "LM," "PF," "PH" or "PB" were excluded as the interacting partners of these motifs are unknown or because they were also present in Uniprobe.

To calculate motif-enrichment *P* values, a motif score cutoff was taken. Then the number of foreground and background sequences above this cutoff was counted. These counts where then combined with counts of total sizes of the foreground and background sequence sets to produce a 2×2 contingency table. Then Fisher's exact test was used to calculate a *P* value. For each motif, 1,000 cutoffs were used and the lowest *P* value was retained. The highest cutoff attempted was the motif score that resulted in at least 5% of the foreground sequences as being classified correctly. The lowest cutoff attempted was the motif score that resulted in at least 5% of the background sequences being classified at negative. The remaining 998 cutoffs were evenly spaced between the highest and lowest cutoffs.

The most significantly enriched motifs (Fisher's exact test *P* value $< 1.0 \times 10^{-20}$; 2,034 motifs) from the reduced sets were taken forward for further analysis. For organization of the motifs into groups, they were clustered as described in the motif-selection stage of the pipeline. The hierarchy was then cut to form $2.(N-1)$ groups, where *N* is the total number of motifs. The number of groups with the greatest enrichment of known motifs within the groups was used in the final analysis. To determine this, we used the hypergeometric distribution to calculate a *P* value for the enrichment of each of the known motifs within the groups of *de novo* motifs. A known motif was counted as being within a *de novo* motif cluster if the Tomtom *E*-value was <0.1. The enrichment *P* value was calculated for each group the known motif matched, and the lowest was used in the overall calculation. A final *P* value was calculated by using Fisher's method to combine the *P* values from each of the known motifs. The number of groups with the lowest combined *P* value was used for the final grouping. The 2,034 motifs were organized into 589 groups with a combined log(*P* values) = −10,883.64. The groups of *de novo* motifs and the known motifs that they match can be viewed at http://wanglab.ucsd.edu/star/epigram/. The motif (PWM) with the greatest information content from each of the 589 groups is available in **Supplementary Data 3**.

When a group of *de novo* motifs is assigned to a known motif, great care must be taken, as there are many families for TFs that possess similar DNA-binding motifs. For example, many basic helix-loop-helix TFs (for example, MYC) bind at the E-box. Thus, if an E-box motif is found as being enriched, it cannot be taken as evidence as any individual E-box binding factor binding. To handle these situations, we annotated groups of *de novo* motifs with a label that describes the common binding properties rather than any individual member. However, in certain situations additional information such as literature and expression data can be used to increase the resolution of the *de novo* motif and known motif assignments. For example, in NPC a group of homeobox-containing motifs is identified as being predictive of H3K27ac. This group was labeled PAX6, which is a known master regulator of this lineage and is highly, and specifically, expressed in NPC. As our library of known motifs included motifs from nonhumans, we must take care to make sure that matches are of relevance to humans. To do this, the nonhuman protein sequence was compared to a database of human protein sequences using BLAST. If the protein matched a human protein with >40% sequence identity over its DNA-binding domain, then the assignment was accepted. Of the 589 groups, 117 could be reliably matched to a known motif recognized by a human factor or family of factors. Details of the *de novo* motif curation with known motifs are given in **Supplementary Data 4**.

We observed that motifs were enriched in the foreground (positively interplay) of a modification in one analysis were possibly also enriched in the background (negatively interplay) of the modification in another comparison (**Supplementary Fig. 6**). For example, motifs matching sequences involved in transcriptional regulation were enriched in the H3K36me3-marked regions in the single-mark analysis but also in the background (gene body) in the typical background analysis. This indicates that these motifs are associated with the genomic region but not with the modification *per se*. Thus, motifs were removed from future analysis if the group to which they belonged contained both positive and negative motifs for the modification. This resulted in the following number of groups ignored for each modification: H3K36me3 = 55, H3K27ac = 15, H3K9me3 = 11, H3K27me3 = 1 and H3K4me1 = 1.

Clustering motif groups by mark and cell-type specificity. We wished to organize the identified motifs by their interplay with epigenomic modifications and cell types. We first organized the motifs by their interplay with modification by taking the *de novo* motifs from each group and summarizing the specificities as a vector of 1s and 0s. Each position in the vector represents whether the motif group was found to have interplay with a modification (1) or not (0). When doing this, we treated motif enrichment and depletion separately. Thus, the vector has a length of 14 as there are seven modifications but enrichment and depletion are treated separately. To cluster the motifs by the vectors, we created a distance matrix by taking the Manhattan distance between each vector. The distance matrix was then used to hierarchically cluster the motifs. The same procedure was used to cluster the motifs by cell-type specificities but using a vector of length 10 as there are five cell types but enrichment and depletion were treated separately.

For the heat maps in **Figure 3a**, gray is used to represent an absence (0 in the vector) of specificity, whereas specificity is

represented by a bright nongray color. Different colors are used in each of the columns to make it easier for the reader.

Location analysis. The peaks were split into ten equally sized bins, where the first and last bins represent the edges, and the fifth and sixth bin represents the center (**Supplementary Fig. 8**). Then we took a count of the motif occurrences within the bins for each set of modification peak sequence. As the peaks have no directionality, the bins from opposite positions but equal distance from the center are combined. For example, bin 1 and 10 are combined, and bin 2 and 9 are combined. This results in five bins where bin 1 is the edge and bin 5 is the center. The count in each bin is divided by the total count of the five bins to give the proportion of the counts in any one bin. If a motif has no binding preference, then one would expect all bins to score 0.2. If the score in the edge bin is 0.05 greater than either of the two middle bins, then a motif is categorized as 'edge'. Conversely, if either of the middle bins scores as 0.05 greater than the edge, then it is categorized as 'central'. Motifs that are neither edge nor peak are categorized as neutral. As an additional criterion, a χ^2 test is used to measure the significance of the variation of the five bin scores. A P value of $<1.0 \times 10^{-10}$ is necessary for categorizing a motifs as being central or edge.

To make **Figure 4a**, we made a distance matrix by taking the Pearson correlation coefficient between the five bin scores of each of the motifs. This distance matrix was then used to cluster the motifs using Ward's hierarchical clustering method[65].

To look at the relationship between the (G+C)-(A+T) hybrid motifs, H3K4me3 and TSSs, we identified a set of TSSs from RefSeq that overlapped H3K4me3 peaks in MSC and were at least 5 kbp from another TSS. We excluded TSSs that were within 5 kbp of another TSS, as we did not want signals from adjacent TSSs confusing the analysis. Furthermore, RefSeq genes were used instead of UCSC known genes as RefSeq represent a smaller and higher-confidence set of genes. This is advantageous as it provides a cleaner view of H3K4me3 signal in relation to motifs at TSSs. In total this identified 13,962 TSSs. We took a 4-kbp window around each TSS and split it into 200-bp bins. To look at the H3K4me3 signal in each of the bins, we counted the number of reads from the MSC H3K4me3 ChIP-seq in each of the bins. We combined both replicates to increase the overall resolution. We repeated this process using the input (control). Then the ChIP and input signals were normalized by the total read counts in each set to get an RPKM score. The final score for a bin was calculated by taking the log of the ratio between the chip and input RPKM values. When taking the log ratio a pseudocount of 0.05 was added to both scores to avoid division by very small numbers or 0. To create the motif profiles, we scanned the motifs against the TSS sequences. The score and location of the best match in each sequence were retained. Then, to get an overall profile, we totaled the scores in each of the bins. Next, each of the bin scores was divided by the sum of all bin scores and multiplied by 100. If a motif has no location preference then we would expect all bins to have a score of 5 (as we are considering 20 bins). Thus, to represent motif enrichment and depletion, we subtracted 5 from each of the bin scores.

When looking further into the relationship between YY1 and H3K9me3 in H1, we chose to focus on YY1 sites that were within 200 bp of an H3K9me3 peak star/end. This is different from what is shown in **Figure 4a**, in which edge bins correspond to the two-fifths of the H3K9me3 peaks that are closest to the start and end of its peaks. When taking a closer look at the relationship between YY1 and H3K9me3 peaks, we chose to consider within 200 bp from the end of H3K9me3 as it looks more specifically at the starts/ends of H3K9me3 peaks while allowing for some inaccuracy (brought about by noise in the assay) in the ChIP-seq peaks.

The YY1 and H3K9me3 examples shown in **Figure 4f** were picked from the 874 YY1 peaks that were within 200 bp of the start/end boundary of an H3K9me3 peak region (see **Supplementary Note**). To pick this subset of sites, we first ranked the sites by degree of difference in H3K9me3 levels between 1-kbp windows on either side of the YY1 motifs. Degree of difference was measured using a χ^2 statistic that expects the levels to be the same on either side of the YY1 motif site. The four sites shown in **Figure 4d** were chosen manually from the top 50 ranked sites.

Comparison of YY1 peaks to other factors peaks. To identify potential cofactors of YY1, we compared the ChIP-seq peaks of all the factors that were available in H1 from ENCODE (list of files available in **Supplementary Data 2**) to the two sets YY1 peaks: (i) those that overlap with an H3K4me3 peak but not an H3K9me3 peak (14,337 sites) and (ii) those that overlap with an H3K9me3 peak but not an H3K4me3 peak (624 sites). The overlap score was calculated as the proportion of the YY1 peaks that overlap the TF peak by at least 1 bp. This was done separately for H3K9me3 (column B) and H3K4me3 (column C). To identify the most differential TFs, the final ranking subtracted the H3K4me3 from the H3K9me3 score. If replicates of a TF were available, then the intersect of the two files was taken. Then the analysis was carried out using the intersect and each of the two replicates in isolation. The results of this analysis are given in **Supplementary Data 1**.

Motif disruption is correlated with H3K27ac variation between individuals. To produce sets of predictive DNA motifs, we ran the Epigram pipeline on H3K27ac peaks from each of the 19 individuals[43]. As a background, mappable regions of the genome that did not overlap any H3K27ac peaks were used. The average full-model AUC was 0.87 with an average of 285 motifs. The set of motifs with the best prediction performance (AUC = 0.88) was used to correlate with H3K27ac in variable regions. The correlation of H3K27ac variation and motif disruption was conducted as previously described[43]. However, the genome sequences of four individuals from the San population could not be included, as the necessary genetic data could not be transferred in time for preparation of this manuscript. Therefore, only 15 individuals were used in the analysis. This discrepancy may have resulted in a reduced percentage of peaks being correlated with motif disruption as any peaks with variation that is specific to the San individuals is unlikely to be explained.

To aid comparison with Kasowski *et al.*, we presented our results in the same way as they did (**Fig. 5**). Each of the rows corresponds to a motif, and each column corresponds to an H3K27ac peak. Only peaks that possess motif sites that are significantly distributed by at least one motif are shown. The rectangles are colored such that they represent the Spearman correlation coefficient between a motifs disruption and inter-individual variation in H3K27ac levels. The motifs are ordered such that the motif with

doi:10.1038/nmeth.3065

the greatest number of significant correlations is at the top of the figure and the motif with the least is at the bottom. The peaks are sorted such that motifs that are associated with a greater number of peaks have their significantly associated peaks shown earlier as one reads the figure from left to right.

The motifs from the 20-motif model were compared to other sets of motifs using Tomtom. The known motif set was the same as was used in the section "Comparison to known motifs and motif clustering." To identify matching motifs groups, we used Tomtom to compare them to the motifs that were predictive of H3K27ac in the single-mark analysis.

50. Hawkins, R.D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479–491 (2010).
51. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42**, D764–D770 (2014).
52. Pruitt, K.D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756–D763 (2014).
53. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
54. Koehler, R., Issac, H., Cloonan, N. & Grimmond, S.M. The uniqueome: a mappability resource for short-tag sequencing. *Bioinformatics* **27**, 272–274 (2011).
55. Bailey, T.L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659 (2011).
56. Harbison, C.T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
57. Caruana, R., Karampatziakis, N. & Yessenalina, A. in *Proc. 25th Int. Conf. Mach. Learn.* 96–103 (ACM, 2008).
58. Hall, M. *et al.* The WEKA data mining software: an update. *SIGKDD Explor.* **11**, 10–18 (2009).
59. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. & Noble, W.S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).
60. Tanaka, E., Bailey, T., Grant, C.E., Noble, W.S. & Keich, U. Improved similarity scores for comparing motifs. *Bioinformatics* **27**, 1603–1609 (2011).
61. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
62. Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38**, D105–D110 (2010).
63. Robasky, K. & Bulyk, M.L. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **39**, D124–D128 (2011).
64. Xie, Z., Hu, S., Blackshaw, S., Zhu, H. & Qian, J. hPDI: a database of experimental human protein-DNA interactions. *Bioinformatics* **26**, 287–289 (2010).
65. Ward, J.H. Jr. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).