BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 18 2013, pages 2231–2237
doi:10.1093/bioinformatics/btt382

Genome analysis

Advance Access publication July 2, 2013

An integrative approach to understanding the combinatorial histone code at functional elements

William K. M. Lai and Michael J. Buck*

Department of Biochemistry, Center of Excellence in Bioinformatics and Life Sciences, State University of New York at Buffalo, Buffalo, NY 14203, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: The rapid advancement of genomic technology has revealed the enormous complexity and combinatorial nature of chromatin modifications. To facilitate interpretation of the combinatorial nature of chromatin, we have developed a novel method to integrate all chromatin datasets into distinct nucleosome types (nucleosome alphabet). We have applied this approach to *Saccharomyces cerevisiae*, generating a nucleosome alphabet, which forms chromatin motifs when mapped back to the genome. By applying novel chromatin alignment and global word search approaches, we have defined distinctive chromatin motifs for introns, origins of replication, tRNAs, antisense transcripts, double-strand-break hotspots and DNase hypersensitive sites, and can distinguish genes by expression level. We have also uncovered strong associations between transcription factor binding and specific types of nucleosomes. Our results demonstrate the uses and functionality of defining a chromatin alphabet and provide a unique and novel framework for exploring chromatin architecture.

Contact: mjbuck@buffalo.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on April 17, 2013; revised on June 10, 2013; accepted on June 27, 2013

1 INTRODUCTION

The DNA of all eukaryotic organisms is organized and structured into chromatin, a nucleoprotein complex through which genetic material is structured and manoeuvred to elicit cellular processes such as transcription, cellular division, differentiation and DNA repair The core component of eukaryotic chromatin is the nucleosome, which is composed of 147 base pairs of DNA wrapped around a histone octamer of H2A, H2B, H3 and H4. The amino and carboxyl termini of histones are susceptible to an enormous number of post-translational covalent modifications, of which 59 distinct modifications have currently been identified. These modifications have been proposed to function by directly altering the intrinsic chromatin structure and/or by presenting or alternatively binding of chromatin-modifying complexes (Rando, 2012).

Chromatin immunoprecipitation (ChIP) in combination with high-throughput sequencing (NGS) and microarrays (ChIP-seq and ChIP-chip) has been used to measure the occupancy of specific histone modifications and histone variants at a genomic scale. For *Saccharomyces cerevisiae*, a broad spectrum of histone modifications, including but not limited to H3K14Ac, H3K9Ac and H3K79me3, have been investigated with low-resolution (200–300 bp) microarrays (Pokholok *et al*., 2005). Higher-resolution (50–100 bp) microarrays have been used to interrogate the location of modifications such as H3K4me3, H3R2me2a and H3K4Ac (Guillemette *et al*., 2011; Kirmizis *et al*., 2007). ChIP-seq has recently been used to map H2A.Z, H2BK123ub, H3K36me2 and H3K36me3 (Albert *et al*., 2007; Batta *et al*., 2011). In addition to ChIP, micrococcal nuclease (MNase) digestion assays have been used to map nucleosome positioning and occupancy at near-bp resolution across the yeast genome (Rizzo *et al*., 2011, 2012).

The analysis of these datasets has been instrumental in expanding our understanding of chromatin's regulatory potential. However, each of these genome-wide datasets is only a small portion of an extremely complex system. Emerging evidence suggests that chromatin modifications function in a combinatorial code that extends across several neighbouring nucleosomes (Rando, 2012). It is therefore advantageous to examine the pattern and spacing of nucleosomes with their histone variants and modifications, i.e. the chromatin 'architecture' of genomic functional elements (Givens *et al*., 2012; Lai and Buck, 2010).

We have integrated numerous genomic datasets and identified the dominant nucleosome types (nucleosome alphabet) by their post-translational modifications and histone variants. Our nucleosome alphabet has allowed us to identify chromatin motifs for genomic functional elements, including novel motifs for introns, ARS consensus sequences (ACS), tRNAs, DNase hypersensitivity sites (DHS), double-strand-break hotspots (DSB Hotspots) and antisense transcripts. We have also identified significant associations between transcription factor (TF) binding and specific types of nucleosomes. Overall, our approach allows us to uncover the shared chromatin architecture across all similar functional elements, and apply traditional sequence alignment and motif-finding algorithms to chromatin for the first time.

2 METHODS

For detailed descriptions of procedures, please see Supplementary Methods.

2.1 Nucleosome calling and genomic datasets

The locations of 68 288 nucleosomes were determined from an MNase-Seq dataset using the template-filtering algorithm (Rizzo *et al*., 2011;

*To whom correspondence should be addressed.

Weiner *et al*., 2010). Histone modification and RNA-Seq datasets were downloaded from the Saccharomyces Genome Database and the Sequence Read Archive (Supplementary Table S1). All sequencing data were aligned to the r64 build of *S.cerevisiae* using the short-read alignment algorithm bowtie, allowing up to one mismatch (Langmead *et al*., 2009). All histone modification datasets were normalized, transformed into $\log_2$ ratios and used to score a 147-bp window around the dyad of each nucleosome for that modification. Fragments per kilobase of exon per million fragments mapped were calculated using cufflinks for all coding sequencing (CDS) and for genes with transcription starts and stops (Trapnell *et al*., 2010).

2.2 Alphabetizing and annotating the nucleosomes

The data matrix containing the $\log_2$ ratio scores for each experiment at each nucleosome was analysed using the Weka Java program's implementation of the X-means algorithm to identify the optimal number of clusters in the matrix (Hall *et al*., 2009; Pelleg and Moore, 2000). Eighteen distinct clusters were identified with a peak Bayesian Information Criterion of 325 867; see Supplementary File 1 with location of all nucleosomes.

2.3 Chromatin alignment

Nucleosomes letters for gene alignment were extracted between their known transcriptional starts and stops and binned according to the size of each gene (Xu *et al*., 2009). Chromatin motifs were generated after alignment with MUSCLE (Edgar, 2004).

2.4 Genome-wide chromatin word search

The list of 18 nucleosome letters was permutated to generate the list of all possible five-letter combinations (words) to a total number of 1 889 568 possible words. All words in the *S.cerevisiae* genome were then annotated to known functional elements and enrichment was determined by permutation. Similar words were combined and motifs generated at significantly enriched functional elements.

2.5 Testing for nucleosome enrichment

Nucleosomes were assigned to conserved TF binding sites for 118 TFs from MacIssac *et al*. downloaded from Saccharomyces Genome Database (MacIsaac *et al*., 2006). Hyper-geometric testing with multiple testing corrections was performed to identify which, if any, nucleosome letters enriched for a TF.

3 RESULTS

3.1 Correlation reveals histone code relationships

To examine the relationships between histone modifications, we calculated the pairwise correlations between each histone modification at the 68 228 nucleosomes in the yeast genome (Fig. 1A). In general, the correlation structure revealed two distinct classes of histone modifications. The first subset of correlated marks includes H3K9Ac, H3K4Ac, H3K4me3, H3K14Ac, H4Ac, H2A.Z and H3K4me2. Most of these modifications have been independently associated with promoter activity and transcriptionally active regions (Liu *et al*., 2005; Rando, 2012). Our findings suggest that these modifications may function in tandem and possibility cooperatively with each other.

The second subset of correlated modifications includes H2BK123ub, H3K36me3, H3K36me2, H3, H3K79me3, H3K4me1 and H3R2me2a. These modifications have been independently associated with diverse biological function, including, but not limited to, transcriptional repression (H3K79me3, H3R2me2a) and exon-coding sequence (H3K36me2, H3K36me3) (Batta *et al*., 2011; Kirmizis *et al*., 2007; Liu *et al*., 2005; Pokholok *et al*., 2005). Known anti-correlation relationships were also recapitulated, such as the previously characterized anti-correlation between H3K4me3 and H3R2me2a (Guillemette *et al*., 2011; Kirmizis *et al*., 2007). The strong correlation between H3R2me2a and H3K4me1 and the strong anti-correlations between H3K4me1 and the promoter-associated histone modifications suggest a broader role for H3K4me1 than previously suggested in *S.cerevisiae*.

3.2 Eighteen distinct classes of nucleosomes define biological functions

The 68 228 by 14 nucleosome histone modification matrix was clustered to identify 18 dominant nucleosome letters (Supplementary Table S2). We then compared each nucleosome type with each other by examining the correlation coefficients between their respective histone modifications (Fig. 1B). After 2-D hierarchical clustering, the resulting correlation structure reveals three blocks of similar nucleosome types with partially overlapping relationships. Visual inspection of the nucleosome types revealed that there was preferential positioning of each type with respect to the underlying genomic feature (Fig. 1C). To develop a clearer understanding of the relationships among the nucleosome types, we clustered them by their histone modifications (Fig. 1D). Many of the nucleosome clusters were similar, with subtle, yet interesting, differences, with approximately six sub-groups of nucleosome clusters representing the majority of the diversity of nucleosome clusters. Comparison of nucleosome classes with a variety of genomic features revealed shared biological function (Fig. 1E and Supplementary Fig. S1).

Group 1 nucleosomes represented 13% of all yeast nucleosomes and contained the types A, B, C and D. These nucleosomes all possessed high occupancy scores for promoter regions, the 5′ end of CDS and transcriptional start site (TSS)-associated nucleosomes across all genes, including both high and lowly expressed CDS. The histone marks ranked highly at these nucleosomes were H3K9Ac, H3K4Ac, H3K14Ac, H4Ac, H3K4me3 and H2A.Z, which are all histone modifications that have been shown in separate studies to associate with active transcription (Liu *et al*., 2005; Pokholok *et al*., 2005). However, to our knowledge, this is the first analysis to demonstrate that these modifications may function in tandem with one another. This group of nucleosomes, specifically nucleosome C, also possessed high occupancy scores for DHS and DSB Hotspots, possibly due to this nucleosome's lower H3 occupancy (Hesselberth *et al*., 2009; Pan *et al*., 2011). In addition, genes with a +1 A or B nucleosome are enriched for the Gene Ontology (GO) processes DNA-dependent transcription, DNA metabolism and RNA metabolic processes (Supplementary Table S3).

Group 2 (G, H and I, 10% of all nucleosomes) nucleosomes are similar to Group 1 nucleosomes but lacked H2A.Z as well as the PolII-associated modifications H3K36me3 and H3K79me3 (Venters and Pugh, 2009). In addition to preferential placement at the promoters and 5′ regions of highly expressed genes, these nucleosomes are also highly associated with the

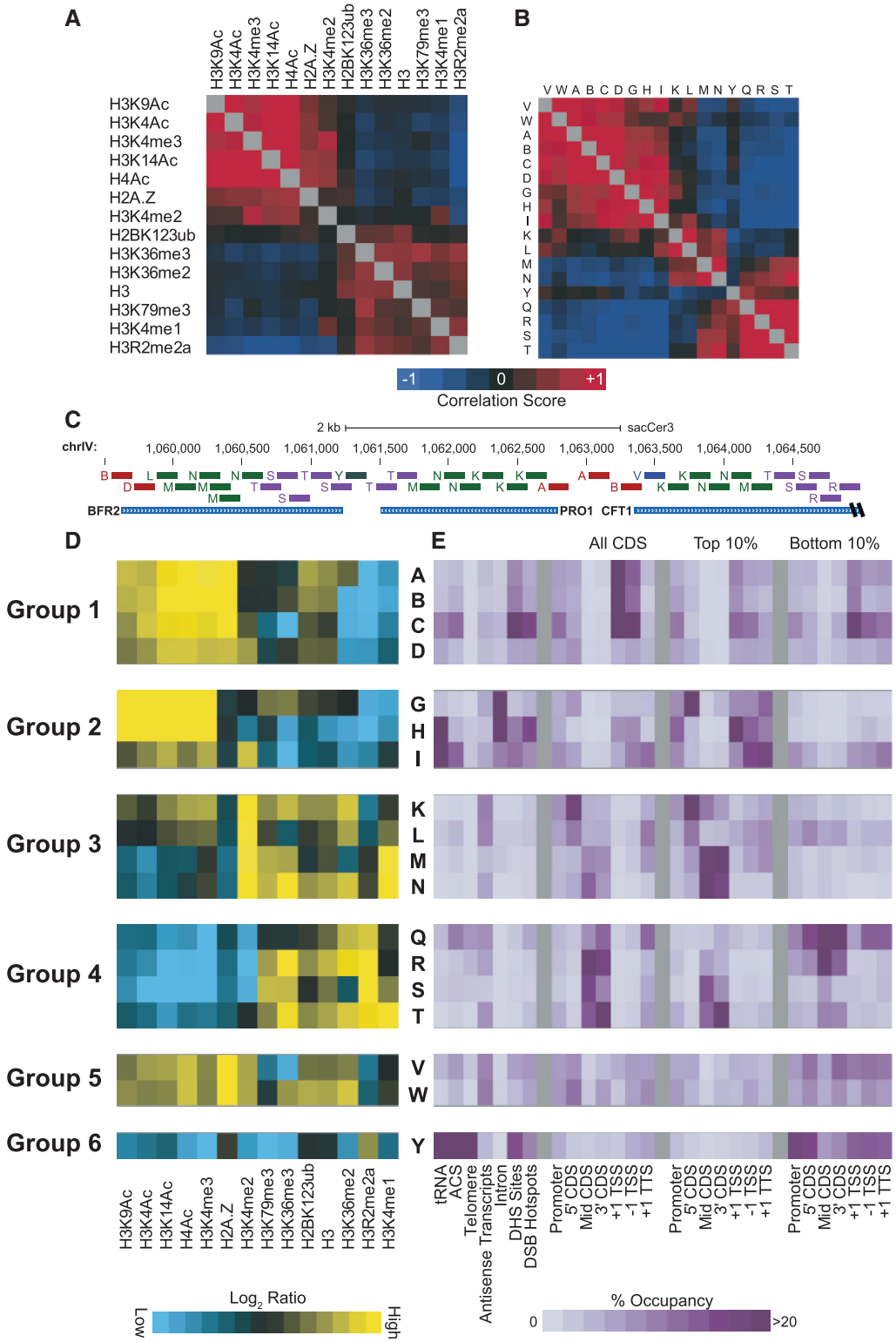

Fig. 1. Defining a nucleosome alphabet. (**A**) Pearson correlation between histone modification scores for each called nucleosome. (**B**) Pearson correlations between each nucleosome type. (**C**) UCSC Genome Browser view of mapped nucleosomes. (**D**) Hierarchical clustering of average modification scores for each nucleosome type. The 18 nucleosome types were subdivided into six primary groups based on their clustering distance. (**E**) Occupancy of each nucleosome type at different genomic locations. ACS, ARC consensus Sequences; DHS, DNase hypersensitivity sites; DSB, double-strand break; CDS, coding sequencing; TSS, transcriptional start site; TTS, transcriptional termination site

PolIII-transcribed tRNA and may represent a nucleosome specific for RNA PolIII transcription. Genes with a +1 H and G nucleosome are enriched for translation and ribosome biogenesis (Supplementary Table S3).

The nucleosomes K, L, M and N composed Group 3 nucleosomes (27% of all nucleosomes), and were characterized by higher occupancy scores at mid-coding regions of the CDS and the 3′ end of CDS of all genes. Although the nucleosomes of Group 3 appeared, on average, at all mid-coding and the 3′ ends of CDS, they displayed a preference for highly expressed genes. These nucleosomes contained high levels of the coding-associated histone modifications H3K36me3, H3K36me2 and H3K79me3, as well as possessing higher occupancy scores for H3 (Batta *et al*., 2011; Pokholok *et al*., 2005). The high occupancy levels at active coding regions in combination with higher levels of H3K36me3 may indicate that these nucleosomes represent a specific class of nucleosomes that depend on PolII for transcription.

Representing 36% of all nucleosomes, Group 4 was the largest group and consisted of Q, R, S and T nucleosomes. Similar to Group 3 nucleosomes, these nucleosomes possessed high levels of occupancy in mid-coding regions and the 3′ end CDS for all genes. Also like Group 3 nucleosomes, they possessed moderate enrichment at antisense transcripts. However, in contrast to Group 3 nucleosomes, Group 4 nucleosomes occurred preferentially at lowly expressed genes. Although possessing similar chromatin modifications to Group 3, Group 4 possessed negligible levels of H3K4me2. Group 4 nucleosomes were also characterized with increased enrichment for H3R2me2a, a known transcriptional silencer that correlates with the 3′ end of CDS (Kirmizis *et al*., 2007). These data along with Group 4 nucleosomes' preferential placement suggest that this group may represent a distinct nucleosome for lowly expressed coding regions.

The nucleosomes assigned the letters V and W belonged to Group 5 (10% of total nucleosomes) and possessed moderate scores across most modifications. In particular, they possessed high scores for H2A.Z, a histone variant associated with transcriptional activity. H2A.Z has also been implicated in transcriptional repression of genes that are transiently repressed (Brickner *et al*., 2007). Their presence at lowly expressed CDS would imply that this group of nucleosome may be involved in the repression of genes that may need to be quickly activated.

The Y nucleosome (4% of total nucleosomes) was unique in that it was the only member of its group as well as being characterized by lacking almost all modifications, with the exception of low levels of H3R2me2a. The Y nucleosome has a strong presence at known ACS, tRNAs, telomeres, DHS and DSB Hotspots. This would indicate that nucleosomes at these genomic features contain overall low nucleosome occupancy with a minimum of histone modifications. In addition, the Y nucleosomes appeared prevalently throughout the CDS of the bottom 10% of expressed genes, indicating that this specific nucleosome may be indicative of non-expressed genes.

3.3 Alignment of alphabetized chromatin define chromatin motifs

With every nucleosome in the genome assigned a letter, we then examined how the arrangements of these nucleosomes define genes. Genes were binned by size, and the intervening nucleosomes between transcription start and stop were concatenated into a chromatin sequence. Analogous to protein and DNA sequences, these chromatin sequences can be further interpreted using sequence alignment algorithms to identify the conserved structural patterns at locations of interest. However, as no chromatin sequence alignment algorithm currently exists, we adapted the MUSCLE (MUltiple Sequence Comparison by Log-Expectation) algorithm, which has been previously used to perform nucleotide and amino acid alignments, to also perform alignments of chromatin sequences (Edgar, 2004).

This novel application of sequence alignment allowed us to identify a chromatin motif for genes (see Methods). The resulting chromatin motif is similar across all genes regardless of size (Fig. 2A). In these chromatin motifs, the +1 nucleosome is predominately an A, B or C nucleosome (Group 1). Position +2 is predominately a D or V nucleosome (Group 1/5). Within the middle of genes are K, M or N nucleosomes (Group 3), followed by Q, R, S or T nucleosomes (Group 4).

Another defined functional element in the genome, origins of replication were also aligned to determine whether a distinct chromatin structure surrounds the origin of replications (Eaton *et al*., 2010). Nucleosome types within 1 kb upstream and downstream of replication origins were aligned and a unique motif was generated that was characterized by the absence of almost all nucleosomes throughout the motif (Fig. 2C). This chromatin motif did feature a preference for the Y nucleosome, which is a lowly occupied nucleosome, lacking all histone modifications except H3R2me2a.

3.4 Chromatin motif varies by gene expression

Expression data from RNA-Seq were used to identify the top and bottom 10% of expressed genes (Yassour *et al*., 2009). The highly and lowly expressed genes were binned and aligned as previously described (Fig. 2B). The resulting aligned chromatin motifs revealed that the chromatin architectures of highly and lowly expressed genes are fundamentally different from one another. The 5′ CDS regions of highly expressed genes contained a higher presence of the H nucleosome than the other motifs. This nucleosome contained higher scores for histone modifications related to active transcription, such as H3K9Ac and H3K4Ac, and possessed the lowest H3. The mid-coding regions of highly expressed genes followed the structural motifs of all CDS with a high presence of Group 3 nucleosomes. However, the 3′ CDS region was distinct in that it possessed much lower levels of Group 4 nucleosomes compared with the other motifs and in the case of short genes, completely absent. As the Group 4 nucleosomes possess higher occupancy and the presence of the previously described repressive marks, it is expected that they should be virtually absent from highly expressed genes. The motif for highly expressed long genes was noticeably less defined than the motif for smaller genes. This contrasts with the same size motif for lowly expressed genes, implying that highly expressed genes possess less defined nucleosome positioning and organization relative to lower expressed genes.

The chromatin motifs for lowly expressed genes were distinct from all CDS and highly expressed CDS. The 5′ CDS region of lowly expressed genes contained notably higher levels of nucleosomes from Groups 4, 5 and 6. Interestingly, the chromatin

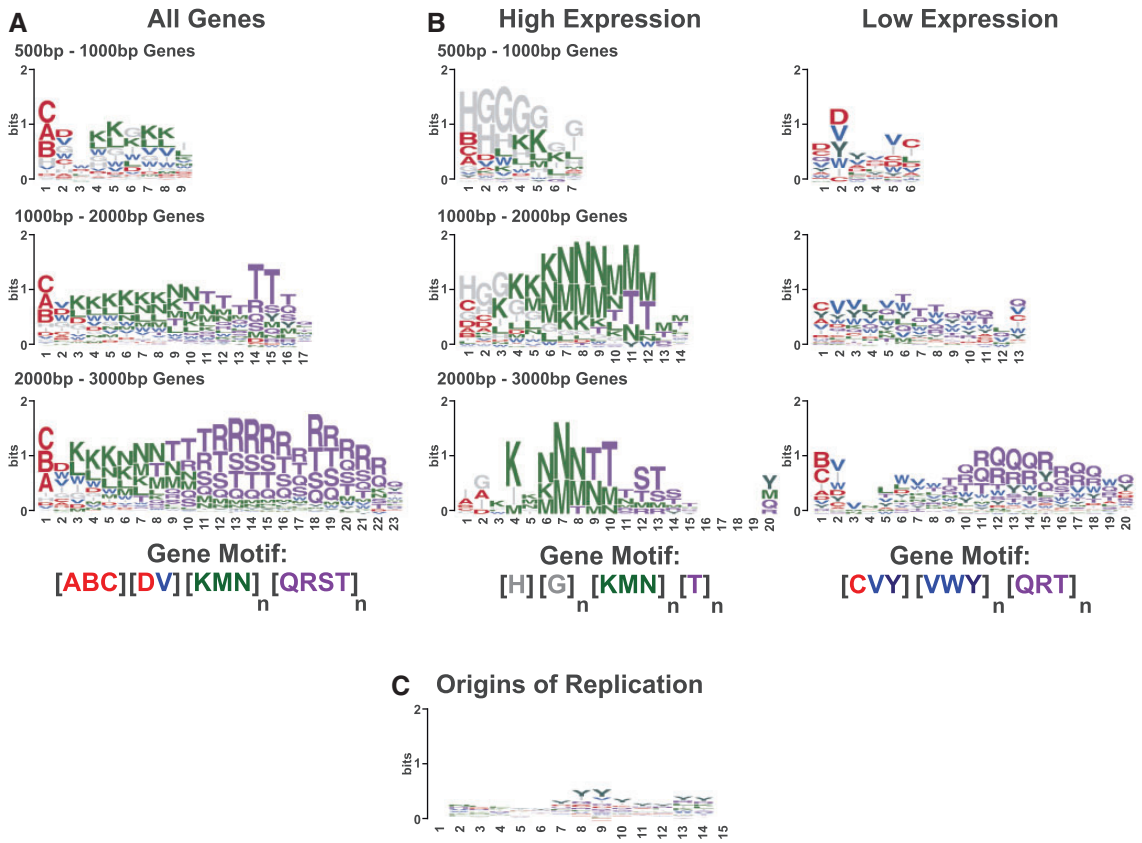

Fig. 2. Chromatin motifs for genes and origins of replication. (**A**) Chromatin structural sequences were separated by gene size: 500–1000 bp, 1–2 kb and 2–3 kb. Chromatin sequences were then aligned with MUSCLE (Crooks *et al.*, 2004; Edgar, 2004). Gene motif shows the average nucleosome sequence motif for all genes in the *S.cerevisiae* genome. (**B**) Genes were sorted by top and bottom 10% by expression, separated by size and aligned independently using MUSCLE. (**C**) Chromatin sequences for *S.cerevisiae* origins of replication were extracted and aligned using the MUSCLE algorithm

motifs of lowly expressed genes were distinct from other motifs, in that they possessed a large number of the V and W nucleosomes across the entire motif. The nucleosomes of Group 5 are characterized by higher occupancy and high levels of H2A.Z, a histone variant believed to play a role in reactivation of expression after repression (Brickner *et al.*, 2007).

The mid-coding regions of lowly expressed genes again differed from all CDS in their preference for the L nucleosome compared with the average CDS preference for the K, M and N nucleosomes. This L nucleosome differs from the K, M and N in that it almost completely lacks H3K36me3 and H3K79me3, two covalent modifications dependent on PolII elongation. This supports the idea of the L nucleosome being a mid-coding nucleosome for lowly transcribed genes. The 3′ CDS region of lowly expressed genes also showed a much more dominant Q nucleosome.

3.5 Identification of novel chromatin architectures at functional elements

To identify novel chromatin motifs, we examined all possible five-letter nucleosome arrangements and tested for significant associations with functional elements (see Methods). We identified three chromatin motifs that enriched for intronic regions (Fig. 3A). These regions were characterized by motifs composed

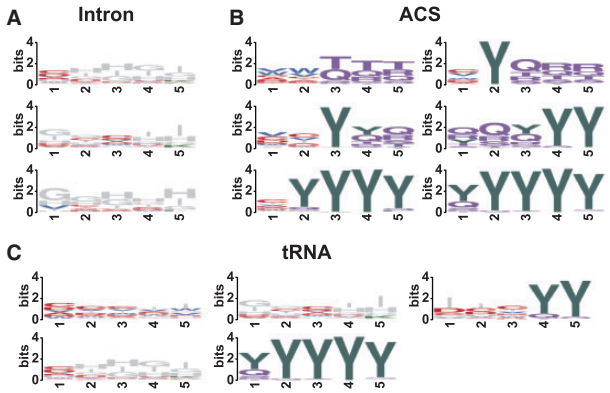

Fig. 3. Chromatin motifs at introns, ACS and tRNA sites. Enriched chromatin motifs identified by a global unguided word search. Only significantly enriched motifs that possess a $P < 0.001$ are shown. Enriched motifs are shown for (**A**) introns, (**B**) ACS and (**C**) tRNA. Additional enrichment motifs for DHS, DSB hotspots and antisense transcripts are shown in Supplementary Figure S3

primarily of Group 2 nucleosomes. This is to our knowledge the first described chromatin pattern for introns in *S.cerevisiae*. We also tested for chromatin motif enrichment at ACS and identified six chromatin motifs that associated significantly with known

ACS. These chromatin motifs mirrored the previous chromatin motifs generated by multiple sequence alignment (Fig. 3B). However, these motifs were far more clearly defined than the previous motifs and imply a chromatin polarity with asymmetrical distributions of Y and Q nucleosomes, reminiscent of the asymmetric chromatin architecture observed at *Schizosaccharomyces pombe* origins of replication (Givens *et al*., 2012).

tRNA-enriched chromatin motifs were composed primarily of the transcriptionally active Group 1 and 2 nucleosomes. This supports their role as highly transcribed genes in the *S.cerevisiae* genome (Fig. 3C). However, the presence of the Y nucleosome, which is characterized by exceptionally low occupancy of almost all nucleosome modifications, was intriguing. The frequency of nucleosome subtypes at each of the unique codons was calculated, and on average, 25% of tRNA associated nucleosomes were the Y nucleosome for each unique codon (Supplementary Fig. S2). Of great interest was the finding that the TTG and TTA codons, which both code for leucine, possessed considerably different compositions of nucleosome subtypes. TTG is composed primarily of the Y nucleosome, whereas TTA is composed primarily of the active Group 2 nucleosomes. The TTG tRNA has recently been shown to be preferentially transcribed in oxidative stress conditions (Chan *et al*., 2012). This finding provides evidence that chromatin may play a role in differential tRNAs' regulation in the *S.cerevisiae* genome. DHS and DSB hotspots were both enriched for similar chromatin motifs that were composed of high levels of nucleosome acetylation and nucleosomes containing low H3 occupancy (Supplementary Fig. S3A and B). Antisense transcripts, on the other hand, were enriched for chromatin motifs composed predominately of Group 5 nucleosomes (Supplementary Fig. S3C).

3.6 TFs enrich for specific nucleosomes

To determine if nearby binding by specific TFs was associated and possibly directing certain histone modification patterns, we examined our nucleosome letters in relation (<100 bp from called dyad) to known TF binding sites (Table 1 and Supplementary Table S4). The majority of TFs were associated with nucleosomes B, C, H or I. These nucleosomes are enriched for promoters and the ±1 TSS nucleosome. These lowly occupied nucleosomes have the active histone modifications H3K9ac, H3K4ac, H3K14ac, H4ac and H3K4me3. The enrichment of these nucleosomes at promoters was largely expected because most TFs bind directly upstream of TSSs, where nucleosome occupancy is the lowest (Harbison *et al*., 2004). However, there are clear distinctions with certain TFs targeting to the Y nucleosome. The Y nucleosome lacks all histone modifications, except H3R2me2a, and is strongly associated with the bottom 10% expressed gene promoters.

Six TFs (Ume6, Phd1, Sum1, Msn2, Sok2 and Swi4) enriched for the Y nucleosome. The majority of these TFs have been associated with repression. Sum1 was unique in that it enriched for the Y nucleosome without any Group 1 or 2 nucleosome enrichments. The Y nucleosomes appear predominately at the nucleosomes immediately flanking TSS (+1/−1 positions) of lowly expressed CDS, which would indicate that this TF binds preferentially at regions possessing a chromatin profile of low genomic activity. Sum1 is a known transcriptional repressor that has been shown to recruit histone deacetylases, which is consistent with the low acetylation levels of the Y nucleosomes (Irlbacher *et al*., 2005). In addition, Phd1 has been shown to directly interact with the Tup1 co-repressor complex and recruit it to its targets (Hanlon *et al*., 2011). Swi4, Phd1 and Sok2 enriched simultaneously for the C, H and Y nucleosomes, which represent +1/−1 nucleosomes for highly, lowly and all expressed CDS. These TFs have all been shown to interact with each other and are involved in chromatin remodelling (Pan and Heitman, 2000; Yeang *et al*., 2005). This would indicate that these TFs may bind to promoters independent of chromatin state. Also of interest were TFs that enriched solely for the B nucleosome that included Adr1, Tye7, Rpn4 and Fkh1. Although these TFs possessed a wide variety of functions, it is interesting to note that they specifically enriched for the B nucleosome that differs from the other Group 1 nucleosomes by having a higher level of H3K36me3.

Table 1. Transcription factors enrich for specific nucleosomes

Type	Enriched transcription factor[a]
A	GCN4, DAL82
B	CBF1, TYE7, STE12, REB1, FKH2, ADR1, GCN4, RPN4, FKH1, AFT2, RCS1
C	CBF1, ABF1, PHD1, REB1, SWI5, MSN2, GCN4, SKN7, SOK2, NRG1, SWI6, PHO2, STB4, SWI4, RCS1, GLN3, GAT1
H	BAS1, CBF1, ABF1, PHD1, ROX1, STE12, RAP1, REB1, FKH2, SWI5, MET31, DIG1, TEC1, GCN4, HAP5, SKN7, FHL1, YAP6, RTG3, SUT1, SOK2, NRG1, SWI6, SKO1, XBP1, PHO2, SWI4, AFT2, NDD1, CST6, ACE2, RLM1, RCS1, PHO4, GLN3, ARG80, ASH1, MOT3, MAC1
I	UME6, STE12, CIN5, SPT2, HSF1, MSN4, DIG1, TEC1, YAP6, SUT1, PHO2, NDD1, DAL82, SPT23
V	MSN4
Y	UME6, PHD1, SUM1, MSN2, SOK2, SWI4

[a]Significance determined by hyper-geometric testing to conserved TF binding sites (MacIsaac *et al*., 2006)

4 DISCUSSION

As high-throughput chromatin data are continually generated through efforts such as the ENCODE and modENCODE projects, it is becoming increasingly important to be able to visualize and understand how these histone modifications interact with each other (ENCODE, 2011). The importance of the combinatorial nature of chromatin can be seen in the NuA4 complex in human, a large multimeric complex that is involved in regulation of transcription, cell-cycle and DNA repair. This complex has seven chromatin binding domains, including two bromodomains, two SANT domains, two chromodomains and a PHD domain (Doyon *et al*., 2004). These domains could allow this complex to bind simultaneously to histone tails containing methylated lysines and arginines, acetylated lysines as well as, deacetylated histone tails. This suggests that the targeting of

this complex would require a special arrangement of modifications across several nucleosomes, and demonstrates the need for studying chromatin architecture across numerous covalent modifications. To address the difficulties involved in interpreting multidimensional datasets, we have implemented a novel method that integrates numerous chromatin datasets across a variety of platforms into a simple chromatin alphabet. Using our chromatin alphabet, we have been able to identify chromatin motifs for functional regions, define TF–nucleosome associations and uncover novel chromatin architectures for introns, ACS, tRNAs, DHS, DSB Hotspots and antisense transcripts.

Current approaches addressing interpretation of combinatorial covalent modifications in the human genome, such as ChomHMM and Seqway, show that genome-wide analysis of multiple modifications is feasible (Ernst and Kellis, 2012; Hoffman *et al*., 2012). The advantage of our approach is the identification of similar functional elements by their chromatin architecture, the eventual development of BLAST-like algorithms and identification of biologically important chromatin attributes by sequence alignment. As the number of genome-wide datasets increases, our technique can be readily applied while simultaneously incorporating older datasets from a variety of platforms. Future applications of our chromatin sequence alphabet will allow for the development of a human chromatin alphabet using the substantial amount of data generated by the ENCODE project (ENCODE, 2011).

ACKNOWLEDGEMENTS

M.J.B. conceived and supervised the study. W.K.M.L. designed and implemented the clustering and annotations. We thank Jason Rizzo for discussions on concept design and implementation.

Funding: National Science Foundation (IIS1016929, to M.J.B.), and NY State Department of Health (C026714 to M.J.B.).

Conflict of Interest: none declared.

REFERENCES

Albert,I. *et al.* (2007) Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature*, **446**, 572–576.

Batta,K. *et al.* (2011) Genome-wide function of H2B ubiquitylation in promoter and genic regions. *Genes Dev.*, **25**, 2254–2265.

Brickner,D.G. *et al.* (2007) H2A.Z-mediated localization of genes at the nuclear periphery confers epigenetic memory of previous transcriptional state. *PLoS Biol.*, **5**, e81.

Chan,C.T. *et al.* (2012) Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins. *Nat. Commun.*, **3**, 937.

Crooks,G.E. *et al.* (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Doyon,Y. *et al.* (2004) Structural and functional conservation of the NuA4 histone acetyltransferase complex from yeast to humans. *Mol. Cell. Biol.*, **24**, 1884–1896.

Eaton,M.L. *et al.* (2010) Conserved nucleosome positioning defines replication origins. *Genes Dev.*, **24**, 748–753.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

ENCODE. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.

Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.

Givens,R.M. *et al.* (2012) Chromatin architectures at fission yeast transcriptional promoters and replication origins. *Nucleic Acids Res.*, **40**, 7176–7189.

Guillemette,B. *et al.* (2011) H3 lysine 4 is acetylated at active gene promoters and is regulated by H3 lysine 4 methylation. *PLoS Genet.*, **7**, e1001354.

Hall,M. *et al.* (2009) The WEKA Data Mining Software: an update. *SIGKDD Explor.*, **11**, 10–18.

Hanlon,S.E. *et al.* (2011) The stress response factors Yap6, Cin5, Phd1, and Skn7 direct targeting of the conserved co-repressor Tup1-Ssn6 in S. cerevisiae. *PLoS One*, **6**, e19060.

Harbison,C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hesselberth,J.R. *et al.* (2009) Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.

Hoffman,M.M. *et al.* (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.

Irlbacher,H. *et al.* (2005) Control of replication initiation and heterochromatin formation in *Saccharomyces cerevisiae* by a regulator of meiotic gene expression. *Genes Dev.*, **19**, 1811–1822.

Kirmizis,A. *et al.* (2007) Arginine methylation at histone H3R2 controls deposition of H3K4 trimethylation. *Nature*, **449**, 928–932.

Lai,W.K. and Buck,M.J. (2010) ArchAlign: coordinate-free chromatin alignment reveals novel architectures. *Genome Biol.*, **11**, R126.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Liu,C.L. *et al.* (2005) Single-nucleosome mapping of histone modifications in *S.cerevisiae*. *PLoS Biol.*, **3**, e328.

MacIsaac,K.D. *et al.* (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics*, **7**, 113.

Pan,J. *et al.* (2011) A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell*, **144**, 719–731.

Pan,X. and Heitman,J. (2000) Sok2 regulates yeast pseudohyphal differentiation via a transcription factor cascade that regulates cell-cell adhesion. *Mol. Cell. Biol.*, **20**, 8364–8372.

Pelleg,D. and Moore,A. (2000) X-means: extending K-means with efficient estimation of the number of clusters. In: *Proceedings of the Seventeenth International Conference on Machine Learning*. Stanford, CA, USA, pp. 727–734.

Pokholok,D.K. *et al.* (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, **122**, 517–527.

Rando,O.J. (2012) Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr. Opin. Genet. Dev.*, **22**, 148–155.

Rizzo,J.M. *et al.* (2011) Tup1 stabilizes promoter nucleosome positioning and occupancy at transcriptionally plastic genes. *Nucleic Acids Res.*, **39**, 8803–8819.

Rizzo,J.M. *et al.* (2012) Standardized collection of MNase-seq experiments enables unbiased dataset comparisons. *BMC Mol. Biol.*, **13**, 15.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Venters,B.J. and Pugh,B.F. (2009) How eukaryotic genes are transcribed. *Crit. Rev. Biochem. Mol. Biol.*, **44**, 117–141.

Weiner,A. *et al.* (2010) High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res.*, **20**, 90–100.

Xu,Z. *et al.* (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature*, **457**, 1033–1037.

Yassour,M. *et al.* (2009) Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc. Natl Acad. Sci. USA*, **106**, 3264–3269.

Yeang,C.H. *et al.* (2005) Validation and refinement of gene-regulatory pathways on a network of physical interactions. *Genome Biol.*, **6**, R62.