Le et al. BMC Genomics 2010, 11(Suppl 4):S3
http://www.biomedcentral.com/1471-2164/11/S4/S3

BMC Genomics

PROCEEDINGS Open Access

Sequence-dependent histone variant positioning signatures

Ngoc Tu Le[1,2*], Tu Bao Ho[1,3], Bich Hai Ho[1,3]

From Asia Pacific Bioinformatics Network (APBioNet) Ninth International Conference on Bioinformatics (InCoB2010)
Tokyo, Japan. 26-28 September 2010

Abstract

Background: Nucleosome, the fundamental unit of chromatin, is formed by wrapping nearly 147bp of DNA around an octamer of histone proteins. This histone core has many variants that are different from each other by their biochemical compositions as well as biological functions. Although the deposition of histone variants onto chromatin has been implicated in many important biological processes, such as transcription and replication, the mechanisms of how they are deposited on target sites are still obscure.

Results: By analyzing genomic sequences of nucleosomes bearing different histone variants from human, including H2A.Z, H3.3 and both (H3.3/H2A.Z, so-called double variant histones), we found that genomic sequence contributes in part to determining target sites for different histone variants. Moreover, dinucleotides CA/TG are remarkably important in distinguishing target sites of H2A.Z-only nucleosomes with those of H3.3-containing (both H3.3-only and double variant) nucleosomes.

Conclusions: There exists a DNA-related mechanism regulating the deposition of different histone variants onto chromatin and biological outcomes thereof. This provides additional insights into epigenetic regulatory mechanisms of many important cellular processes.

Background

Eukaryotic genomes are packaged into chromatin, a highly condensed structure like a bead-on-string fiber with fundamental repeating units, the nucleosomes. Each nucleosome is formed by wrapping 147bp of DNA around a histone core, an octamer of proteins that contains a central $(H3-H4)_2$ tetramer flanked on both side by two $H2A-H2B$ dimers [1]. Biological evidences have increasingly shown that, far beyond simple DNA compaction chromatin imposes ubiquitous and profound effects on many important DNA-based processes, such as transcription, DNA replication and repair [2-5]. During those processes, the structure of chromatin must be dynamically and reversibly altered to enable or repress the access of cellular machineries to underlying genomic sequences.

To alter histone-DNA and histone-histone interactions the cell has developed diverse and complicated pathways, such as post-translational modifications (PTMs) of histone proteins, ATP-dependent chromatin remodeling or replacement of canonical histones by nonallelic histone variants [6]. For example, changes in the net charge of nucleosomes by lysine acetylation would result in increasing nucleosome mobility, making acetylated nucleosomes easier to displace from their translational positions [7-9]; in yeast, nucleosomes can be moved from their default positions to energetically unfavorable positions by ISW2 ATP-dependent chromatin remodeling complex or even evicted from chromatin by orchestrated action of RSC chromatin remodeling complex and nucleosome-assembly protein 1 (Nap1) histone chaperone [10,11]; in human, nucleosomes bearing both H2A.Z and H3.3 histone variants are observed to be

* Correspondence: ngoctule@jaist.ac.jp
[1]School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
Full list of author information is available at the end of the article

BioMed Central
© 2010 Le et al; licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

prone to eviction [12]. Despite their important roles in regulating chromatin structure and consequently chromatin-based processes, not until recently have histone variants received considerable attention. Among the most extensively studied histone variants are H3.3 and H2A.Z, the isoforms of histones H3 and H2A, respectively. H3.3 has been implicated in many biological pathways in which it may function as neutral replacement histone at chromatin regions where histones have been displaced upon transcriptional activities [13], or it may help to transmit epigenetic memory of active gene states [14]. Recent studies have also revealed unrecognized roles of this variant in chromatin remodeling mechanisms during sexual reproduction [15] and in forming open chromatin regions by inhibiting the binding of linker histone H1 [16]. While studies on H3.3 highly agree about its functions, controversial characteristics have been observed for H2A.Z [17,18]. In some cases, H2A.Z is shown to increase the stability of nucleosomes bearing it [19] while in others it appears that H2A.Z is easier to displace from chromatin than is H2A [20,21]; H2A.Z has also been claimed to involve in gene inactivation [22] and activation (or both) [23]. Nevertheless, most of previous works came into the same conclusion that H2A.Z and H3.3 are non-randomly distributed along the genome and those distributions may reflect their biological functions [24]. So understanding how H3.3 and H2A.Z are distributed over the genome and its biological implication gives us deeper insights into epigenetic mechanisms of many important processes.

Previous works have shown that, histone variants differ from their canonical counterparts not only by their biochemical compositions, ranging from a few amino-acid positions to large protein domains, but also by their incorporations into chromatin. While canonical histones are expressed and incorporated into chromatin during DNA replication in S phase, most histone variants are synthesized throughout the cell cycle and available to nucleosome assembly pathways that occur in a Replication-Independent (RI) manner [13,25]. For example, it has been known in [26-28] that Swi2/Snf2-related ATP-dependent chromatin remodeling complex SWR1, histone chaperones Nap1 and Chz1 involve in the assembly of Htz1 (H2A.Z in yeast), while the deposition of H3.3 involves with the activity of histone chaperone HIRA [29]. It still remains elusive, however, that how histone variants H2A.Z and H3.3 are targeted to the deposited sites [24,25,30].

The advancement of high-throughput profiling technologies such as ChIP-Chip and ChIP-Seq makes it possbile to map genome-wide distributions of nucleosomes bearing different histone variants [21,31]. This offers an unprecedented opportunity to investigate the effects of histone variant distribution on cellular processes (e.g. transcription) as well as its regulatory factors. More recently, Jin et al. [32] have investigated genome-wide distributions of nucleosomes containing different histone variants, including H2A.Z, H3.3 and both (H3.3/H2A.Z, so-called double variant histones) in human genome. The work has shown that nucleosomes bearing double variant histones mainly account for the enriched patterns of histone variants observed at promoters and other important regulatory regions of active genes. Using the data available from Jin et al. [32], we investigated whether there exist genomic features that may help to distinguish target sites of nucleosomes bearing different histone variants in the vicinity of promoter regions. We found that genomic sequence contributes partially to determining the target sites of these variants. Moreover, dinucleotides CA/TG are remarkably important in distinguishing target sites of H2A.Z-only nucleosomes with those of H3.3-containing (both H3.3-only and double variant) nucleosomes. These results give additional insights into epigenetic regulatory mechanisms of many important cellular processes.

Results and discussion

Genomic sequence partially distinguishes target sites of double variant nucleosomes with those of nucleosomes bearing only H3.3 or H2A.Z

High-resolution studies on crystal structure of the nucleosome core particle have revealed that the DNA is wrapped around the histone octamer in a flat, left-handed superhelix [1]. Based on this finding, the notion of sequence-dependent affinity of the histone core is defined as the energetic cost to bend the DNA to accommodate the superhelical path. The observation that nucleosomes show higher affinity for particular DNA sequences [33] has fostered many efforts to find genomic signatures related to genome-wide nucleosome positioning [34-37]. Although the role of genomic sequence on nucleosome distribution is still a controversial topic, it has been confirmed, both *in vivo* and *in vitro,* that DNA sequence imparts nucleosome distribution and there actually exists genomic code for nucleosome positioning [38-40]. Regarding important functions of H2A.Z, many efforts have been spent on identifying genomic signatures that may affect its genome-wide distribution. Tolstorukov et al. [41] compared genomic sequences of H2A.Z-containing nucleosomes from yeast and human and concluded that human nucleosomal sequences do not show the pattern of 10-bp periodicity as observed in yeast. Despite trying a handful of approaches, Gervais et al. [42] could not find any specific DNA motif that can help to distinguish H2A.Z-containing nucleosomal sequences from those of nucleosomes containing canonical histones. Other

computational approaches also failed to distinguish H2A.Z-containing nucleosomal sequences from linker sequences [37]. Recently, the work of Jin et al. [32] gives the evidence that most of H2A.Z-containing nucleosomes observed at important genomic regions, such as promoters, CTCF binding sites, etc. are actually double variant nucleosomes. Taken together, we speculate that at important genomic regions (e.g. promoters) although histone variant-containing nucleosomes do not occupy thermodynamically favorite locations, DNA sequence may help to distinguish target sites of double variant nucleosomes from those of nucleosomes bearing only H2A.Z or H3.3.

To verify this hypothesis, we applied the computational procedure proposed by Peckham et al. [36] to the problem of discriminating double variant nucleosomal sequences from H2A.Z-only or H3.3-only nucleosomal sequences. This method has shown competitive performance for the task of discriminating "nucleosome forming" sequences from "nucleosome inhibiting" sequences and was applied sucessfully on human data [43]. The SVM classifiers in our work were built using *Radial Basis Function (RBF)* kernel (see *Methods)* instead of linear (dot product) kernel as used in original work. Sequence datasets, called *double-h2az* and *double-h33* for training SVM classifiers were created as described in *Methods.* Each sequence in the training sets was represented as a 2, 772-entry vector, in which each entry is a normalized count of the occurences of a particular *k-mer* or its reverse complement, with $k=1$ up to 6. These vectors were used to train SVM classifiers.

To evaluate the performances of the resultant classifiers we used 10-fold cross-validation procedure. According to this procedure, each dataset was divided randomly into 10 subsets. The classifiers were trained on 9 subsets and tested on the remaining one. This training-testing procedure was repeated 10 times using a different hold-out set at each time. To measure the performances of the classifiers, we utilized the *receive-operator-characteristic* (ROC) curve. The quality of the classifier can be evaluated by calculating the *area-under-the-curve* (AUC) (the "ROC score"), in which a random classifier achieves the ROC score of 0.5 and a perfect classifier achieves the ROC score of 1.0. The average ROC scores from 10-fold cross validation on two sequence datasets, *double-h2az* and *double-h33,* were 0.62 ($SD \approx 0.02$) and 0.63 ($SD \approx 0.03$), respectively. These are significantly higher than the performance of the random classifier ($p = 0.3 \times 10^{-5}$ and $p = 0.15 \times 10^{-4}$ (t-test), correspondingly). This result shows that, genomic sequence contributes in part to targeting double variant nucleosomes to sites different from those of H2A.Z-only and H3.3-only nucleosomes.

Characteristics of sequences wrapping diffenrent histone variants

The result above suggests us to search for sequence features characterizing target sites of different histone variants. Basically, there are two kinds of such features [44]: one is compositional discriminative motif, which may help to distinguish double variant nucleosomal sequences from H2A.Z-only and H3.3-only ones; and the other is periodic pattern, which may appear in the set of nucleosomal sequences.

There are several approaches to find compositional discriminative motifs. For example, Peckham et al. [36] used a simple word counting method to compute the frequencies of different DNA motifs and evaluated discriminative power of each motif in separating "nucleosome forming" sequences from "nucleosome inhibiting" ones based on its ROC score. Gupta et al. [43] computed the percentages of dinucleotides from "nucleosome forming" and "nucleosome inhibiting" sequence sets and found an overrepresentation of several dinucleotides in "nucleosome inhibitory" and "nucleosome favorable" sequences, such as AC/GT and CC/GG, correspondingly. In our work, we used feature selection with Fisher criterion, a simple yet effective method (see *Methods),* for the task. Two sets of normalized count vectors corresponding to *double-h2az* and *double-h33* datasets were used to evaluate the discriminative power of different DNA motifs. Analyzing the numbers of occurences of 20 strongest discriminative motifs in *double-h2a.z* (Table 1, ranked by F-score), we found that H2A.Z-only nucleosomal sequences are richer in AT-related motifs (e.g. AAT/ATT, TA, AA/TT, AAA/TTT), which are known to be nucleosome inhibitory signals, while double variant nucleosomal sequences are richer in such motifs as CAG/CTG, CA/TG, C/G, which are known to be nucleosome favoring signals [36]. For 20 strongest discriminative motifs in *double-h3.3* (Table 2, ranked by F-score), H3.3-only nucleosomal sequences are richer in highly flexible, nucleosome favoring motifs, such as CA/TG, ACA/TGT, CACA/TGTG, while double variant nuclesomal sequences are richer in nucleosome favoring but less flexible motifs, such as CC/GG.

The periodicity has been known as one of the fundamental features that may appear in a set of DNA sequences. For example, 3bp sequence period is known to characterize coding sequences [45], while ~10bp sequence period is known to affect the curvature, bendability [46] and establish rotational setting on the histone surface [31] of the DNA sequences. To identify periodic patterns that may appear in nuclesomal sequences wrapping different histone variants, we employed autocorrelation analysis, which was sucessfully applied to detect hidden sequence periodicities in sets of

Table 1 Strongest discriminative motifs corresponding to double-h2az dataset ranked by F-scores

Order	Motifs	F-score	Richer in H3.3/H2A.Z (+) or H2A.Z-only (-) nucleosomal sequences
1	CAG	0.0961057	+
2	AAT	0.079221	-
3	TA	0.0781893	-
4	CA	0.0715567	+
5	CAGG	0.0713273	+
6	AA	0.0709853	-
7	TAA	0.0662379	-
8	AATA	0.0655541	-
9	C	0.0649125	+
10	AAAT	0.0630749	-
11	ATA	0.0595751	-
12	ATAA	0.0559362	-
13	AAA	0.0554242	-
14	AGG	0.0551005	+
15	AG	0.0541048	+
16	AAAAT	0.0516458	-
17	TAAA	0.051629	-
18	CC	0.0491037	+
19	AAATA	0.0459333	-
20	AT	0.044603	-

DNA sequences [47,48](see *Methods).* Investigating the periodograms of DNA motifs which show strongest discriminative power, we found clear periods of 2, 4, 6 of dinucleotides CA/TG in the set of 2169 double variant nucleosomal sequences (Figure 1); and clear periods of 2, 4, 6, 8 in the set of 348 H3.3-only nucleosomal sequences (Figure 2) while the set of 894 H2A.Z-only nucleosomal sequences does not exhibit this feature (Figure 3).

The enrichment (and possibly periodic patterns) of CA/TG dinucleotides in the genomes of many eukaryotic organisms can be resulted from CpG-methylation-

Table 2 Strongest discriminative motifs corresponding to double-h33 dataset ranked by F-scores

Order	Motifs	F-score	Richer in H3.3/H2A.Z (+) or H3.3-only (-) nucleosomal sequences
1	GGA	0.08384	+
2	CA	0.0709705	-
3	GGAA	0.0655294	+
4	ACTCCC	0.0649791	+
5	ACA	0.0625949	-
6	AC	0.0595409	-
7	AGGA	0.0571402	+
8	GCTCC	0.0509788	+
9	ACAT	0.0499685	-
10	ATG	0.0499393	-
11	CTCCC	0.0469626	+
12	CC	0.0445543	+
13	CAC	0.0441645	-
14	CTCC	0.0428833	+
15	GGAAA	0.0427223	+
16	ACACA	0.0406986	-
17	GGGA	0.039697	+
18	CCCAGG	0.0386637	+
19	TGGAAA	0.0379053	+
20	CACA	0.0375203	-

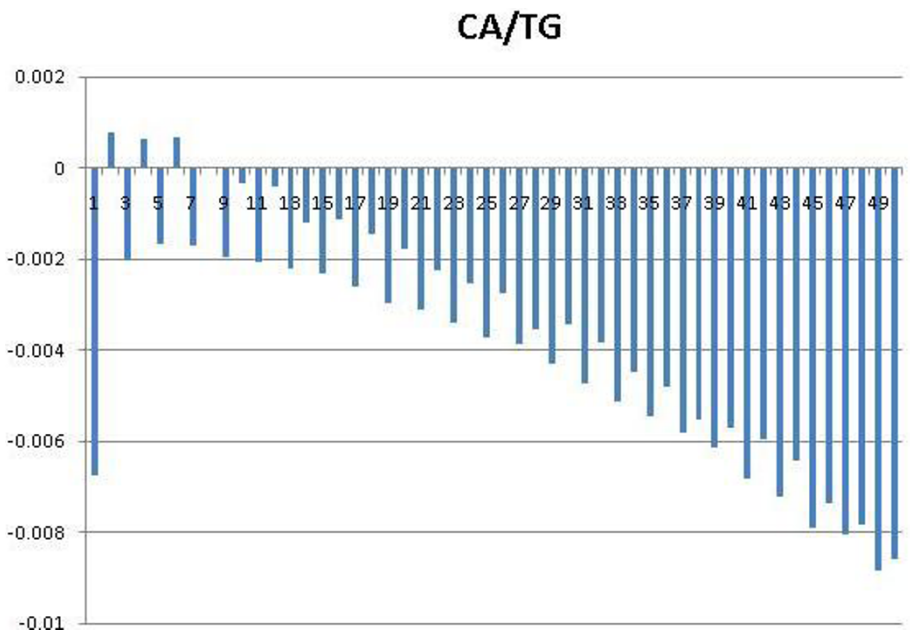

Figure 1 Periodograms of dinucleotides CA/TG with *x*-axis representing distance *k*, *y*-axis representing average covariance value C_{XX} (*k*). CA/TG periodicities corresponding to the set of double variant nucleosomal sequences.

deamination process, which dominates point substitutions in vertebrates [49], in which highly-methylated CpG dinucleotide is deaminated and then mutated to CA or TG [50,51]. This may lead to the possibility that the observed periodicities of CA/TG are caused by random deposition of H3.3-containing nucleosomes to CA/TG-enriched regions generated by this process. However, a recent work [52] has reported that H3.3 variant preferentially targets CpG-rich promoters and a large number of H3.3 promoters has low DNA

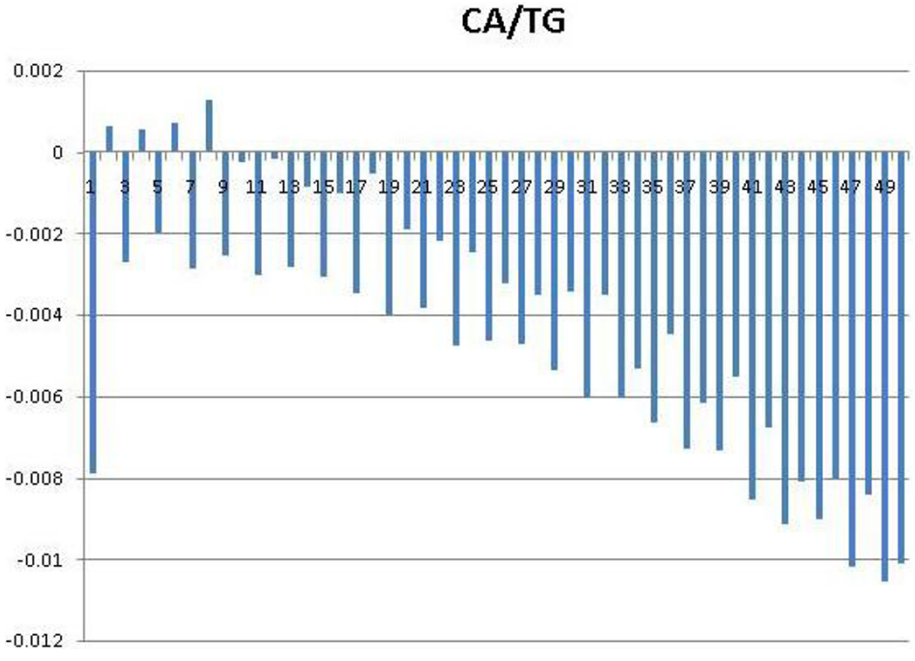

Figure 2 Periodograms of dinucleotides CA/TG with *x*-axis representing distance *k*, *y*-axis representing average covariance value C_{XX} (*k*). CA/TG periodicities corresponding to the set of H3.3-only nucleosomal sequences.

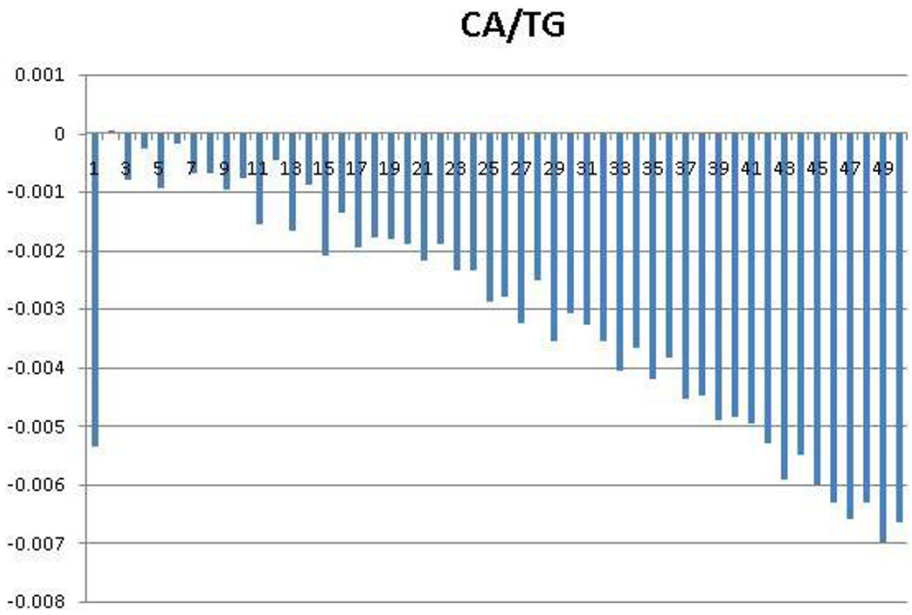

Figure 3 Periodograms of dinucleotides CA/TG with *x*-axis representing distance *k*, *y*-axis representing average covariance value C_{XX} (*k*). CA/TG show no specific periodicity among the set of H2A.Z-only nucleosomal sequences.

methylation level. Taken together with this, our result suggests that the periodicities of CA/TG dinucleotides may play important roles in determining the target sites of nucleosomes containing H3.3 (both H3.3-only and double) variant. Also, from Table 2 we can see that the enrichment of CA/TG dinucleotides is one of the strongest discriminative signals distinguishing target sites of double variant nucleosomes from those of H3.3-only nucleosomes. No DNA motif, however, shows dominant effect on this target site selection. This suggests that the role of DNA sequence may be modest compared with that of other factors (such as chromatin remodeling complexes, PTMs) in identifying exact locations for double variant nucleosomes.

Previous work has shown that in both *D. melanogaster* and human cells, H3.3 is deposited into transcribed genes, promoters and gene regulatory elements [6]. It is incorporated into genes upon transcription induction and is associated with transcriptional elongation [24]. Recent *in vitro* study about the effects of histone variants on transcription reported, however, that the presence of H3.3 is transparent to the transcription and the effect of "hybrid" particles containing both H2A.Z and H3.3 (i.e. double variant nucleosomes) was entirely dominated by the presence of H2A.Z [53]. These results, together with what was reported in Jin et al. [32], have raised important questions that: why double variant nucleosomes are present at important regulatory regions and whether or not H3.3 simply marks sites of nucleosome destabilization or is itself important for mediating the process [54]. Our result here provides a complementary answer to these questions: the presence of H3.3 in double variant nucleosomes can help to position the particles at specific locations in promoters; based on that, H2A.Z, which is also included in the particles, can execute its functions on transcription process. This is consistent with the hypothesis that the process incorporating H3.3 into nucleosomes, which involves histone chaperones and nucleosome remodelers, also facilitates H2A.Z incorporation [54]; and with the *in vitro* result that H3.3 may be just a marker of chromatin regions in flux while the presence of H2A.Z directly affects the transcriptional properties of the particles [53].

Conclusions

Histone variants, such as H2A.Z and H3.3, play significant biological roles in regulating chromatin structure and chromatin-based processes thereof. So it is important to understand how they are targeted to the deposited sites. Although it has been known that there are several biological pathways related to the deposition of these variants onto chromatin, the mechanisms are still unclear. Our work here shows that, there may exist a DNA-related mechanism regulating the deposition of different histone variants onto chromatin. In that, highly flexible dinucleotides, such as CA/TG, play remarkable role in the selection of deposited sites for H3.3-containing (both H3.3-only or double variant) nucleosomes. Moreover, recent works have also shown that H3.3-containing nucleosomes are deposited to sites marked with

active PTMs (such as H3K4Me3) [55] and this deposition is controlled by different factors depending on specific genomic regions [52]. Future studies therefore should seek to understand how those factors coordinate to regulate the deposition of H3.3 variant onto chromatin and its relationship with transcription process.

Methods

Data preparation

Nucleosome positioning analysis

Experimental ChIP-Seq data (BED files) corresponding to "H2A.Z only", "H3.3 only" and Double (H2A.Z/H3.3) histone variants from human H3.3 HeLa S3 cell were received from [32]. Nucleosome positioning profiles of those histone variants were identified using NPS [56] with default settings except that the parameter "minimum nucleosome length" was set to 100bp. We received totally 14565 double variant, 4876 H2A.Z-only and 2799 H3.3-only nucleosomes. The lengths of these nucleosomes were then truncated/extended, centered on the nucleosome, to 150bp if the resultant lengths were longer/shorter than 150bp. Genomic sequences corresponding to these nucleosomes were extracted from UCSC Genome Browser [57], human genome Build 36.1 (hg18 assembly). Repeats from RepeatMasker and Tandem Repeat Finder were excluded in our analysis.

The gene set

UCSC Old Known Genes were extracted and then mapped to Affymetrix U133P2 probe IDs using the tables provided in the UCSC Genome Browser [58]. Genes without corresponding U133P2 IDs were removed. If multiple genes map to the same U133P2 ID, only one was retained. We also removed genes from chromosomal regions marked with "random" or genes from haplotype regions. The final set contained 18285 genes.

Training datasets

All nucleosomes identified above were mapped to the vicinity of the transcription start sites (TSSs) (10000bp upstream and 2000bp downstream) of the genes in the gene set. Only nucleosomes belonging to these regions were used for further analysis. After this step, we received 2169 double variant, 894 H2A.Z-only and 348 H3.3-only nucleosomes. We then created two sequence datasets, namely *doulbe-h2az* and *double-h33,* for training purpose. *double-h2az* dataset contained 800 double variant nucleosomal sequences, selected randomly from 2169 double variant nucleosomes, and 800 H2A.Z-only nucleosomal sequences, selected randomly from 894 H2A.Z-only nucleosomes. Similarly, *double-h33* dataset contained 300 double variant nucleosomal sequences, selected randomly from 2169 double variant nucleosomes, and 300 H3.3-only nucleosomal sequences, selected randomly from 348 H3.3-only nucleosomes above.

Support vector machine (SVM) classifiers

Given a training set containing instance-class pairs (x_i, y_i), i = 1, 2, ..., l where $x_i \in R^l$ and $y_i \in \{-1, 1\}$ is a class label, an SVM classifier is a hyperplane $w^T\varphi(x_i) + b$, where $\varphi(x_i)$ is a function mapping x_i into a higher (maybe infinite) dimensional space, that best separates the two classes. The hyperplane is obtained by solving the following primal optimization problem:

$$\begin{aligned} &Minimize: \tfrac{w^T w}{2} + C\sum_{i=1}^{l} \xi_i \\ &\text{Subject to}: y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i \\ &\xi_i \geq 0 \; i = 1, 2, \ldots, l \end{aligned} \tag{1}$$

Its dual is a quadratic optimization problem:

$$\begin{aligned} &Minimize: \tfrac{\alpha^T Q\alpha}{2} - e^T\alpha \\ &\text{Subject to: } C \geq \alpha_i \geq 0 \; i = 1, 2, \ldots, l \\ &\gamma^T\alpha = 0 \end{aligned} \tag{2}$$

where e is an unit vector, $C > 0$ is an error penalty parameter, $Q_{ij} = y_i y_j K(x_i, x_j)$, $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$ is a kernel function. In our work, we employed *Radial Basis Function (RBF)* kernel $K(x_i, x_j) = exp(-(x_i - x_i)^2)$ to build SVM classifiers. Discriminant value for a testing instance x given by a trained classifier is: $f(x) = \Sigma_i \alpha_i y_i K(x, x_i) + b$. Gist software package [59] was utilized for the tasks of training and testing SVM classifiers.

Feature selection with Fisher criterion

Feature selection is a process of selecting a subset of relevant features available from the data that most contribute to distinguishing instances from different classes. In our method, sequence features related to two kinds of histone variants, double variant and H3.3 only or double variant and H2A.Z only, were identified and ranked by their Fisher scores (or F-score in short). This is one of statistical criteria that is simple, effective and independent of the choice of classification method. The discriminative strength of each feature is defined as following:

Given a dataset X with two classes, denote instances in class 1 as X^1, and those in class 2 as X^2. Assume $\overline{X}_j^k$ is the average of the *jth* feature in X^k, the F-score of the *jth* feature is:

$$F(j) = \frac{\left(\bar{x}_j^1 - \bar{x}_j^2\right)^2}{\left(s_j^1\right)^2 + \left(s_j^2\right)^2} \tag{3}$$

where

$$\left(S_j^k\right)^2 = \sum_{x \in X^k} \left(x_j - \bar{x}_j^k\right)^2 \quad (4)$$

The numerator indicates the discrimination between two classes, and the denominator indicates the scatter within each class. The larger the F-score is, the more likely this feature is more discriminative. Gist software package [59] was used to calculate F-scores for different DNA motifs.

Autocorrelation analysis

Correlation functions measure the enrichment of certain pairs of motifs at a distance of k bp. To calculate XX-autocorrelation fuction we followed the method described in [48]. Given a motif X and a DNA sequence S, we count in the entire S the number $N_{XX}^S(k)$ of pairs of two identical motifs X and X separated by k base pairs. There are $L - k - l + 1$ pairs in a sequence of length L, where l is the length of the motif. Consequently, the probability to find the pair $X - X$ at the distance k can be estimated as:

$$P_{XX}(k) = N_{XX}^S(k) \,/\, (L - k - l + 1) \quad (5)$$

The probability to find a single motif X, denoted by $P_X(k)$, can be estimated as:

$$P_X(k) = N_X^S \,/\, (L - k - l + 1) \quad (6)$$

where N_X^S is the number of motif X in the sequence S. If the pairs at a distance k are statistically independent we have: $P_{XX}(k) = P_X(k) * P_X(k)$. Thus the difference, $C_{XX}(k) = P_{XX}(k) - P_X(k) * P_X(k)$, measures the correlation at a distance of k base pairs. A positive peak of the covariance $C_{XX}(k)$ implies that there are more $X - X$ pairs at a distance of k than expected by chance. The mean covariance function for a set of sequences was calculated by averaging individual functions over all the sequences in the set.

Acknowledgements

We would like to gratefully thank Mr. Chongzhi Zang and his colleagues for kindly sharing their experiment data. The first and the third authors have been supported by Japanese Government Scholarship (Monbukagakusho) to study in Japan. The second author is supported by NAFOSTED (Vietnam's National Foundation for Science and Technology Development).
This article has been published as part of *BMC Genomics* Volume 11 Supplement 4, 2010: Ninth International Conference on Bioinformatics (InCoB2010): Computational Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2164/11?issue=S4.

Author details

[1]School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan. [2]Hanoi National University of Education, 136 Xuan Thuy, Cau Giay, Hanoi, Vietnam. [3]Vietnamese Academy of Science and Technology, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam.

Authors' contributions

NTL and TBH defined the research problem. NTL and BHH designed the experiment. NTL, BHH and TBH drafted the manuscript. All authors contributed to and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 2 December 2010

References

1. Luger K, Mader AW, Richmond AK, Sargent DF, Richmond TJ: **Crystal structure of the nucleosome core particle at 2.8 A resolution.** *Nature* 1997, **389**:251-260.
2. Groth A, Rocha W, Verreault A, Almouzni G: **Chromatin Challenges during DNA Replication and Repair.** *Cell* 2007, **128(4)**:721-733.
3. Li B, Carey M, Workman JL: **The Role of Chromatin during Transcription.** *Cell* 2007, **128(4)**:707-719.
4. Corpet A, Almouzni G: **Making copies of chromatin: the challenge of nucleosomal organization and epigenetic information.** *Trends in Cell Biology* 2008, **19**:29-41.
5. Probst AV, Dunleavy E, Almouzni G: **Epigenetic inheritance during the cell cycle.** *Nature Reviews Molecular Cell Biology* 2009, **10**:192-206.
6. Henikoff S: **Nucleosome destabilization in the epigenetic regulation of gene expression.** *Nature Reviews Genetics* 2008, **9**:15-26.
7. Waterborg JH: **Dynamics of histone acetylation in vivo. A function for acetylation turnover?** *Biochemistry and cell biology* 2002, **80(3)**:363-378.
8. Reinke H, Horz W: **Histones Are First Hyperacetylated and Then Lose Contact with the Activated PHO5 Promoter.** *Molecular Cell* 2003, **11(6)**:1599-1607.
9. Zhao J, Diaz JH, Gross DS: **Domain-Wide Displacement of Histones by Activated Heat Shock Factor Occurs Independently of Swi/Snf and Is Not Correlated with RNA Polymerase II Density.** *Molecular and Cellular Biology* 2005, **25(20)**:8985-8999.
10. Whitehouse I, Tsukiyama T: **Antagonistic forces that position nucleosomes in vivo.** *Nat. Struct. Mol. Biol.* 2006, **13(7)**:633-640.
11. Lorch Y, Maier-Davis B, Kornberg RD: **Chromatin remodeling by nucleosome disassembly in vitro.** *Proc. Natl. Acad. Sci. U.S.A.* 2006, **103(9)**:3090-3093.
12. Jin C, Felsenfeld G: **Nucleosome stability mediated by histone variants H3.3 and H2A.Z.** *Genes and Development* 2007, **21(12)**:1519-1529.
13. Henikoff S, Ahmad K: **Assembly of variant histones into chromatin.** *Annual review of cell and developmental biology* 2005, **21**:133-153.
14. Ng RK, Gurdon JB: **Epigenetic memory of active gene transcription is inherited through somatic cell nuclear transfer.** *Proc. Natl. Acad. Sci. U.S.A.* 2005, **102(6)**:1957-1962.
15. Ooi SL, Henikoff S: **Germline histone dynamics and epigenetics.** *Current opinion in cell biology* 2007, **19(3)**:257-265.
16. Braunschweig U, Hogan GJ, Pagie L, van Steensel B: **Histone H1 binding is inhibited by histone variant H3.** *The EMBO journal* 2009, **28(23)**:3635-3645.
17. Guillemette B, Gaudreau L: **Reuniting the contrasting functions of H2A.Z.** *Biochemistry and cell biology* 2006, **84(4)**:528-535.
18. Zlatanova J, Thakar A: **H2A.Z: view from the top.** *Structure* 2008, **16(2)**:166-179.
19. Park YJ, Dyer PN, Tremethick DJ, Luger K: **A new fluorescence resonance energy transfer approach demonstrates that the histone variant H2AZ stabilizes the histone octamer within the nucleosome.** *The Journal of biological chemistry* 2004, **279(23)**:24274-24282.
20. Zhang H, Roberts DN, Cairns BR: **Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss.** *Cell* 2005, **123(2)**:219-231.
21. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K: **Dynamic regulation of nucleosome positioning in the human genome.** *Cell* 2008, **132(5)**:887-898.
22. Guillemette B, Bataille AR, Gevry N, Adam M, Blanchette M, Robert F, Gaudreau L: **Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning.** *PLoS biology* 2005, **3(12)**:e384.

23. Raisner RM, Hartley PD, Meneghini MD, Bao MZ, Liu CL, Schreiber SL, Rando OJ, Madhani HD: **Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin.** *Cell* 2005, **123(2)**:233-248.
24. Talbert PB, Henikoff S: **Histone variants—ancient wrap artists of the epigenome.** *Nature Reviews Molecular Cell Biology* 2010, **11(4)**:264-275.
25. Jin J, Cai Y, Li B, Conaway RC, Workman JL, Conaway JW, Kusch T: **In and out: histone variant exchange in chromatin.** *Trends in biochemical sciences* 2005, **30(12)**:680-687.
26. Mizuguchi G, Shen X, Landry J, Wu WH, Sen S, Wu C: **ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex.** *Science* 2004, **303(5656)**:343-348.
27. Luk E, Vu ND, Patteson K, Mizuguchi G, Wu WH, Ranjan A, Backus J, Sen S, Lewis M, Bai Y, Wu C: **Chz1, a nuclear chaperone for histone H2AZ.** *Molecular cell* 2007, **25(3)**:357-368.
28. Straube K, Blackwell JJS, Pemberton LF: **Nap1 and Chz1 have Separate Htz1 Nuclear Import and Assembly Functions.** *Traffic* 2010, **11(2)**:185-197.
29. Tagami H, Ray-Gallet D, Almouzni G, Nakatani Y: **Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis.** *Cell* 2004, **116**:51-61.
30. Orsi GA, Couble P, Loppin B: **Epigenetic and replacement roles of histone variant H3.3 in reproduction and development.** *The International journal of developmental biology* 2009, **53(2-3)**:231-243.
31. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF: **Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome.** *Nature* 2007, **446(7135)**:572-576.
32. Jin C, Zang C, Wei G, Cui K, Peng W, Zhao K, Felsenfeld G: **H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions.** *Nature Genetics* 2009, **41(8)**:941-945.
33. Thastrom A, Lowary PT, Widlund HR, Cao H, Kubista M, Widom J: **Sequence Motifs and Free Energies of Selected Natural and Non-natural Nucleosome Positioning DNA Sequences.** *Journal of molecular biology* 1999, **288(2)**:213-229.
34. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JPZ, Widom J: **A genomic code for nucleosome positioning.** *Nature* 2006, **442(7104)**:772-778.
35. Ioshikhes IP, Albert I, Zanton SJ, Pugh BF: **Nucleosome positions predicted through comparative genomics.** *Nature Genetics* 2006, **38(10)**:1210-1215.
36. Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z: **Nucleosome positioning signals in genomic DNA.** *Genome Research* 2007, **17(8)**:1170-1177.
37. Yuan GC, Liu JS: **Genomic sequence is highly predictive of local nucleosome depletion.** *PLoS computational biology* 2008, **4**:e13.
38. Chung HR, Vingron M: **Sequence-dependent nucleosome positioning.** *Journal of molecular biology* 2008, **386(5)**:1411-1422.
39. Fraser RM, Keszenman-Pereyra D, Simmen MW, Allan J: **High-resolution mapping of sequence-directed nucleosome positioning on genomic DNA.** *Journal of molecular biology* 2009, **390(2)**:292-305.
40. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, Segal E: **The DNA-encoded nucleosome organization of a eukaryotic genome.** *Nature* 2009, **458(7236)**:362-366.
41. Tolstorukov MY, Kharchenko PV, Goldman JA, Kingston RE, Park PJ: **Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes.** *Genome Research* 2009, **19(6)**:967-977.
42. Gervais AL, Gaudreau L: **Discriminating nucleosomes containing histone H2A.Z or H2A based on genetic and epigenetic information.** *BMC Molecular Biology* 2009, **10(18)**.
43. Gupta S, Dennis J, Thurman RE, Kingston R, Stamatoyannopoulos JA, Noble WS: **Predicting human nucleosome occupancy from primary sequence.** *PLoS computational biology* 2008, **4(8)**:e1000134.
44. Bolshoy A: **Revisiting the relationship between compositional sequence complexity and periodicity.** *Computational biology and chemistry* 2007, **32**:17-28.
45. Trifonov EN: **Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences.** *Journal of molecular biology* 1987, **194(4)**:643-652.
46. Herzel H, Weiss O, Trifonov EN: **10-11 bp periodicities in complete genomes reflect protein structure and DNA folding.** *Bioinformatics* 1999, **15**:187-193.
47. Tolstorukov MY, Virnik KM, Adhya S, Zhurkin VB: **A-tract clusters may facilitate DNA packaging in bacterial nucleoid.** *Nucleic Acids Research* 2005, **33(12)**:3907-3918.
48. Kumar L, Futschik M, Herzel H: **DNA Motifs and Sequence Periodicities.** *In Silico Biology* 2006, **6(1-2)**:71-78.
49. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, TJ TJH, Lipshutz R, Chee M, Lander ES: **Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.** *Science* 1998, **280(5366)**:1077-1082.
50. Coulondre C, Farabaugh JHMPJ, Gilbert W: **Molecular basis of base substitution hotspots in Escherichia coli.** *Nature* 1978, **274(5673)**:775-780.
51. Razin A, Riggs AD: **DNA methylation and gene function.** *Science* 1980, **210(4470)**:604-610.
52. Goldberg AD, Banaszynski LA, Noh KM, Lewis PW, Elsaesser SJ, Stadler S, Dewell S, Law M, Guo X, Li X, Wen D, Chapgier A, DeKelver RC, Miller JC, Lee YL, Boydston EA, Holmes MC, Gregory PD, Greally JM, Rafii S, Yang C, Scambler PJ, Garrick D, Gibbons RJ, Higgs DR, Cristea IM, Urnov FD, Zheng D, Allis CD: **Distinct factors control histone variant H3.3 localization at specific genomic regions.** *Cell* 2010, **140(5)**:678-691.
53. Thakar A, Gupta P, McAllister WT, Zlatanova J: **Histone Variant H2A.Z Inhibits Transcription in Reconstituted Nucleosomes.** *Biochemistry* 2010, **49(19)**:4018-4026.
54. Henikoff S: **Labile H3.3+H2A.Z nucleosomes mark 'nucleosome-free regions'.** *Nature Genetics* 2009, **41(8)**:865-866.
55. Delbarre E, Jacobsen BM, Reiner AH, AL ALS, Kuntziger T, Collas P: **Chromatin Environment of Histone Variant H3.3 Revealed by Quantitative Imaging and Genome-scale Chromatin and DNA Immunoprecipitation.** *Molecular biology of the cell* 2010, **21(11)**:1872-1884.
56. Zhang Y, Shin H, Song JS, Lei Y, Liu XS: **Identifying Positioned Nucleosomes with Epigenetic Marks in Human from ChIP-Seq.** *BMC Genomics* 2008, **9(537)**.
57. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Research* 2002, **12(6)**:996-1006.
58. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool.** *Nucleic Acids Research* 2004, **32**:D493-D496.
59. Pavlidis P, Wapinski I, Noble WS: **Support vector machine classification on the web.** *Bioinformatics* 2004, **20**:586-587.

doi:10.1186/1471-2164-11-S4-S3
Cite this article as: Le *et al.*: Sequence-dependent histone variant positioning signatures. *BMC Genomics* 2010 **11**(Suppl 4):S3.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central