Genomics 101 (2013) 125–133

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Human repetitive sequence densities are mostly negatively correlated with R/Y-based nucleosome-positioning motifs and positively correlated with W/S-based motifs

Wentian Li [a,*], Daniela Sosa [b,c], Marco V. Jose [d]

[a] The Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institute for Medical Research, North Shore LIJ Health System, Manhasset, 350 Community Drive, NY 11030, USA
[b] Facultad de Ciencias, Universidad Nacional Autónoma de México, México 04510 DF, Mexico
[c] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, México 04510 DF, Mexico
[d] Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Apdo Postal 70228, México 04510 DF, Mexico

ARTICLE INFO

Article history:
Received 9 July 2012
Accepted 29 October 2012
Available online 5 November 2012

Keywords:
Nucleosome positioning
Repetitive sequences
DNA motifs
Wavelet transformation

ABSTRACT

We examined statistical correlations between the frequencies of seven proposed nucleosome positioning motifs and the densities of repetitive sequences in the human genome. For both parametric and non-parametric measures of statistical correlations there is a tendency for repetitive sequence density to be negatively correlated with the density of R/Y-based nucleosome positioning motifs, while being positively correlated with that of W/S-based motifs. These results largely hold even when motifs are examined only within repeat-filtered sequences. The RRRRRYYYYY motif and its 5-base shift YYYYYRRRRR, in particular, is over-represented in the human genome; and its negative correlation is consistently present at different regions and at different length scales. For some other nucleosome positioning motifs, the relationship with repeats can be regional or length scale dependent. Considering the importance of nucleosome formation in epigenetic regulations, these results may provide new insight to the evolution of repetitive sequences.

1. Introduction

After double helix, nucleosome provides the next level of physical structure for DNA molecules (the chromatin structure) that play an important role in gene regulation [1–3]. With the chromatin being accessible at the promoter region, sequence is well positioned with nucleosome downstream from the promoter [4,5]. It has been long recognized that some DNA segments have a higher affinity to the nucleosome core histones, perhaps due to their own intrinsic bending, than other segments [6]. This observation led to many proposals of the nucleosome positioning motifs (NPM) (other names are also used, such as nucleosome core sequence pattern, nucleosome positioning code, etc.) which presumably cause certain DNA sequences to be located in the nucleosome core (as versus linker), at specific positions with respect to the central "dyad" region of the two-round wrapping of DNA around histone octamer. These motifs only increase the nucleosome positioning probability, and do not necessarily dictate absolute presence of them (or absolute absence of others) in the nucleosome cores. To cite from ref. [7], "you can position all of the nucleosomes some of the time and some of the nucleosomes all the time, but you can't position all the nucleosomes all of the time".

A major focus of NPM is to examine what sequences are preferred in the major and in the minor groove. This would define a sequence pattern which spans 5 basepair positions. Two types of these spacing-of-5-base motifs were proposed. One is the R/Y-based (R for purine: A or G, Y for pyrimidine: C or T), carving two segments from the … YRNNNRYNNNYR… sequence [8] around the two grooves: YRNNNRY and RYNNNYR (N for any nucleotide base). In this paper, these two patterns are written as the motif [YR-3-RY, RY-3-YR]. The motif YR-3-RY reads: a YR dinucleotide followed by any three bases, then followed by a RY dinucleotide. Another motif is the W/S-based (W for weak: A or T, S for strong: C or G), written as [WW-3-SS, SS-3-WW] [9]. The WW-3-SS is actually a more general motif than the originally observed [AA,TT,TA]NNNGC [10], i.e., either AA, TT, or TA dinucleotide followed by any three bases, then followed by the GC dinucleotide.

One extension of the above two types of short motifs (5 bases spacing or 7-mer or heptamer) is by a tandem repeat of them, leading to a periodicity of ten. For example, a tandem repeat of the W/S-based motif would lead to [WW-8-WW, SS-8-SS]; these two motifs are out of phase by 5 bases. In fact, the [AA,TT]NNNNNNNN[AA,TT] pattern is a main result in ref. [9], though the peak-to-peak distance does not always stay at 10 bases. Trifonov and Sussman uncovered the periodicity of 10.5 bases for dinucleotides [AA,TT], [GG,CC], TA, and TG [11], with the first three belonging to the W/S-type.

The recent genome-scale sequencing of nucleosome core DNA has generated large amount of data and provided fertile ground for testing ideas on NPM [12–16]. In particular, Trifonov's group suggested

* Corresponding author at: The Robert S Boas Center for Genomics and Human Genetics, Feinstein Institute for Medical Research, North Shore LIJ Health System, 350 Community Drive, Manhasset, NY 11030, USA. Fax: +1 516 562 1153.
E-mail addresses: wtli2012@gmail.com, wli@nshs.edu (W. Li).

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2012.10.005

Table 1
Correlation coefficients between NPM densities and repetitive sequence density for human chromosome 20. The densities are calculated from non-overlapping windows. Window sizes are doubled consecutively, starting from 1 kb to 2048 kb (2.048 Mb). The first column is the window size; columns 2-4 are the number of windows, Pearson correlation coefficients and the corresponding *p*-values for testing zero correlation, Spearman correlation coefficient and the corresponding *p*-value; The next three columns are similar for NPM densities calculated from the unique (repeat-filtered) sequence only. (A) [RY-3-YR, RY-3-YR]; (B) [WW-3-SS, SS-3-WW]; (C) WW-8-WW; and (D) [RRRRRYYYYY, YYYYYRRRRR].

W size (kb)	All seq			Unique seq		
	No. W	Pearson/pv	Spearman/pv	No. W	Pearson/pv	Spearman/pv
A						
2	29751	−0.013/0.03	−0.019/*E*−3	28686	−0.25/0	−0.15/0
4	14875	−0.041/5*E*−7	−0.045/5*E*−8	14692	−0.23/2*E*−179	−0.14/4*E*−63
8	7437	−0.80/4*E*−12	−0.084/4*E*−13	7427	−0.20/1*E*−69	−0.15/2*E*−36
16	3718	−0.13/4*E*−16	−0.13/2*E*−16	3716	−0.19/1*E*−30	−0.16/1*E*−21
32	1859	−0.19/8*E*−17	−0.18/2*E*−15	1859	−0.21/3*E*−19	−0.19/8*E*−17
64	929	−0.27/2*E*−16	−0.26/7*E*−16	929	−0.26/9*E*−16	−0.27/1*E*−16
128	464	−0.33/*E*−13	−0.33/2*E*−13	464	−0.31/4*E*−12	−0.34/7*E*−14
256	232	−0.38/2*E*−9	−0.40/*E*−10	232	−0.36/2*E*−8	−0.37/6*E*−9
512	116	−0.48/7*E*−8	−0.53/8*E*−10	116	−0.45/3*E*−7	−0.48/9*E*−8
1024	58	−0.51/4*E*−5	−0.60/*E*−6	58	−0.51/5*E*−5	−0.52/4*E*−5
2048	29	−0.58/9*E*−4	−0.75/6*E*−6	29	−0.61/5*E*−4	−0.70/4*E*−5
B						
2	29751	0.057/0	0.049/2*E*−17	28686	−0.29/0	−0.18/5*E*−205
4	14875	0.081/0	0.060/3*E*−13	14692	−0.25/7*E*−205	−0.14/4*E*−64
8	7437	0.11/0	0.070/2*E*−9	7427	−0.16/3*E*−46	−0.10/2*E*−16
16	3718	0.16/0	0.091/3*E*−8	3716	−0.043/8*E*−3	−0.057/5*E*−4
32	1859	0.20/0	0.095/4*E*−5	1859	0.013/.6	−0.052/0.02
64	929	0.25/4*E*−15	0.11/8*E*−4	929	0.12/2*E*−4	−0.046/0.2
128	464	0.34/*E*−13	0.13/6*E*−3	464	0.20/1*E*−5	−0.026/0.6
256	232	0.40/2*E*−10	0.12/.07	232	0.27/4*E*−5	−0.040/0.5
512	116	0.45/2*E*−7	0.11/0.2	116	0.33/3*E*−4	−0.093/0.3
1024	58	0.52/2*E*−5	0.12/.4	58	0.40/2*E*−3	−0.068/0.6
2048	29	0.58/*E*−3	0.070/.7	29	0.44/0.02	−0.25/0.2
C						
2	29751	0.25/0	0.20/2*E*−258	28686	0.038/0	0.058/1*E*−22
4	14875	0.28/0	0.21/3*E*−142	14692	0.11/0	0.10/3*E*−37
8	7437	0.28/0	0.20/2*E*−67	7427	0.17/0	0.13/5*E*−30
16	3718	0.27/0	0.17/4*E*−25	3716	0.18/0	0.12/3*E*−13
32	1859	0.24/0	0.13/8*E*−9	1859	0.17/4*E*−14	0.10/2*E*−5
64	929	0.21/6*E*−11	0.089/.007	929	0.16/1*E*−6	0.069/0.04
128	464	0.21/6*E*−6	0.060/.2	464	0.18/6*E*−5	0.065/0.2
256	232	0.20/.002	0.012/0.9	232	0.19/5*E*−3	0.046/0.5
512	116	0.16/.08	−0.10/.3	116	0.16/0.1	−0.060/0.5
1024	58	0.17/.2	−0.15/.3	58	0.16/0.2	−0.081/0.5
2048	29	0.11/0.6	−0.31/.1	29	0.085/0.6	−0.25/0.2
D						
2	29751	−0.21/4*E*−284	−0.21/3*E*−289	28686	−0.13/3*E*−113	−0.23/0
4	14875	−0.23/2*E*−174	−0.24/*E*−189	14692	−0.13/2*E*−55	−0.16/4*E*−83
8	7437	−0.26/8*E*−115	−0.28/2*E*−131	7427	−0.10/4*E*−19	−0.10/1*E*−19
16	3718	−0.29/*E*−74	−0.31/9*E*−86	3716	−0.10/5*E*−9	−0.072/1*E*−5
32	1859	−0.35/*E*−54	−0.36/*E*−58	1859	−0.094/5*E*−5	−0.077/9*E*−4
64	929	−0.40/4*E*−36	−0.40/5*E*−36	929	−0.094/4*E*−3	−0.061/0.06
128	464	−0.46/3*E*−25	−0.42/7*E*−21	464	−0.10/0.03	−0.051/0.3
256	232	−0.50/9*E*−16	−0.50/5*E*−16	232	−0.10/0.1	−0.041/0.5
512	116	−0.58/*E*−11	−0.56/4*E*−11	116	−0.094/0.3	0.042/0.6
1024	58	−0.63/*E*−7	−0.62/4*E*−7	58	−0.068/0.6	0.11/0.4
2048	29	−0.85/.02	−0.86/.02	29	−0.055/0.8	0.15/0.4

GRAAATTTYC as a most recent "finale" of the long-searched "chromatin code" [17–19]. This decamer motif and its two degenerate parental motifs, RRRRRYYYYY and SSWWWWWWSS (also the derived ones from tandem repeat followed by shift) are all mergers of the R/Y-based and W/S-based spacing-of-5 motifs mentioned early.

Human genomes are full of repetitive sequences [20] which occupy at least 50% (e.g., [21]) of the genome (it is even suggested that they may occupy as much as 2/3 of the genome [22]). It is natural to ask whether a relationship exists, if any, between NPM and repetitive sequences [23]. In an ongoing work, we examine the effect of repetitive sequences on the observed periodicities of [RRRRRYYYYY, YYYYYRRRRR] (D. Sosa, P. Miramontes, W. Li, V. Mireles, J.R. Bobadilla, M.V. José, unpublished results). Here we analyze the statistical correlations between the density of NPMs and the density of repetitive sequence directly. Obviously, there are only three possible relationships between the two: negative correlation, positive correlation, and no correlation (or statistically insignificant correlations).

The main technical obstacle in answering the posed question is that composition/density of any sequence type/motif may depend on the length scale at which the density is calculated. In a simple form, even base composition may depend on window size such that a [G,C]-rich domain can contain [G,C]-poor subdomains [24]. We will deal with this problem by directly testing correlations at different length scales, as well as by a more systematic approach of wavelet transformation, particularly useful for capturing multiple scales at once. Due to the large number of calculations and tests, we will start by examining one human chromosome (chromosome 20) in more detail. Then these analyses will be extended to the whole genome.