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.

2. Results

2.1. Human chromosome 20, [RY-3-YR, YR-3-RY] motif

We partition DNA sequence of chromosome 20 into 62,965 non-overlapping 1 kb windows. Windows with less than 90% sequencing rate are discarded, leaving 59,502 windows, or 94.5% of the original number. For each window, densities of various NPMs are calculated, as well as the density of repetitive sequences. These densities at the length scale of 1 kb are the basis for similar calculation at larger length scales. The first NMP we examined is [RY-3-YR, YR-3-RY] [8], whose density in chromosome 20 is 0.067 copies per base if overlapping motif is prohibited (0.10 if overlapping is allowed). Note that YR-3-RY is not only a 5-base shift in a RY-3-YR tandem repeat, but also a reverse complement pattern of RY-3-YR.

2.1.1. YR-3-RY and RY-3-YR density is negatively correlated with the repetitive sequence density

We consider both heptamers YR-3-RY and RY-3-YR, so the definition of the motif is independent from which strand is used, and whether the 5-base move is from major to minor groove or from minor to major groove. At the 1 kb window level, the repetitive sequence density and R/Y-based heptamer density is not significantly correlated (Pearson correlation coefficient (cc) is −0.0016 with p-value 0.69, and non-parametric Pearson correlation coefficient −0.0090 with p-value 0.029).

At larger window sizes, however, it is increasingly clear that the two are negatively correlated, as summarized in Table 1(A) (left columns) (Note: the notation (e.g.) 5E-7 means 5×10^{-7}). We combine the two consecutive windows into one to move to the next length scale, from 1 kb window size to 2 kb, then to 4 kb, etc. The magnitude of the negative correlation, for both Pearson and Spearman correlation, gradually increases. Despite the loss of sample size (number of windows), the statistical significance still increases from p-value $\sim10^{-2}$ at 2 kb to p-value $\sim10^{-15}$ to 10^{-17} at 32 kb to 64 kb. Then for even larger window sizes, the significance is reduced as the number of samples is reduced, though the magnitude of negative correlation increases.

2.1.2. Regional variation of the correlation

Fig. 1(A) shows the scatter plot of number of repetitive sequence bases per kb (x-axis) and number of copies of [YR-3-RY, RY-3-YR] motif per kb for chromosome 20. There are 232 points (a point is a 256 kb window) in Fig. 1(A). The points/windows in the left, middle, or right 1/3 of the sequence are labeled by green, blue, and red colors, respectively. Linear regression lines for all points and for the three groups of points are shown. Although all three groups show negative regression slopes with somewhat comparable significance, the third group spans a wider range of repetitive sequence densities (with more windows at low repetitive sequence densities). This indicates a spatial heterogeneity between different chromosome regions.

Both the sign of the correlation and its regional variation along the chromosome has been confirmed by an independent wavelet analysis (see Table S1 of the Supplementary material).

2.1.3. Calculating [YR-3-RY, RY-3-YR] density in repeat-filtered sequence

To further understand the source of the negative correlation, we calculated the [YR-3-RY, RY-3-YR] motif density in the unique sequence,

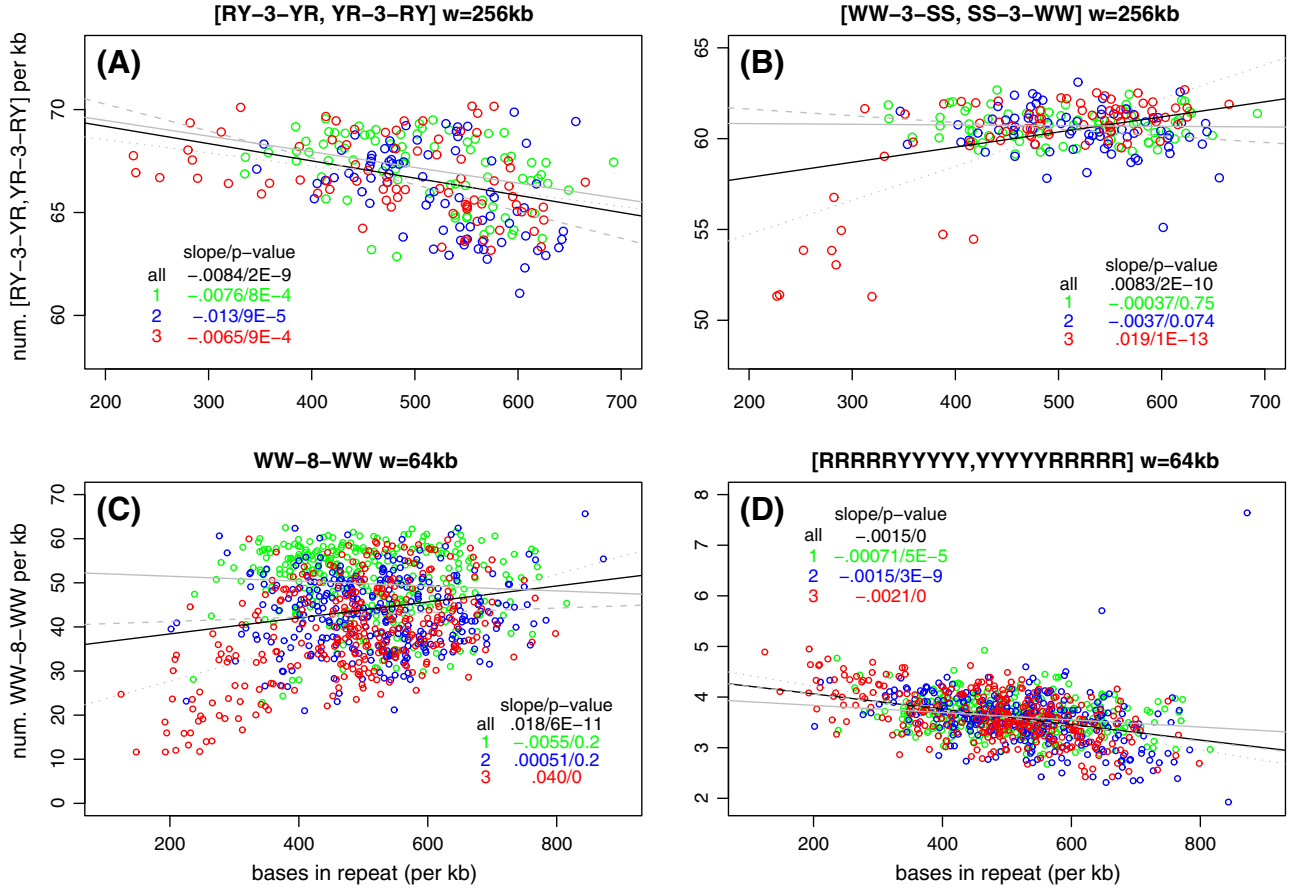

Fig. 1. Scatter plot of number of copies of NPMs in non-overlapping windows versus number of bases in repetitive sequence (in the same window) for human chromosome 20. The windows from the left, middle, and right 1/3 of the chromosome are labeled by green, blue, and red colors, respectively. Regression lines for all points, and for points from the three regions are shown. (A) [RY-3-YR, RY-3-YR], window size is 256 kb; (B) [WW-3-SS, SS-3-WW], window size is 256 kb; (C) WW-8-WW, window size is 64 kb; (D) [RRRRRYYYYY, YYYYYRRRRR], window size is 64 kb.

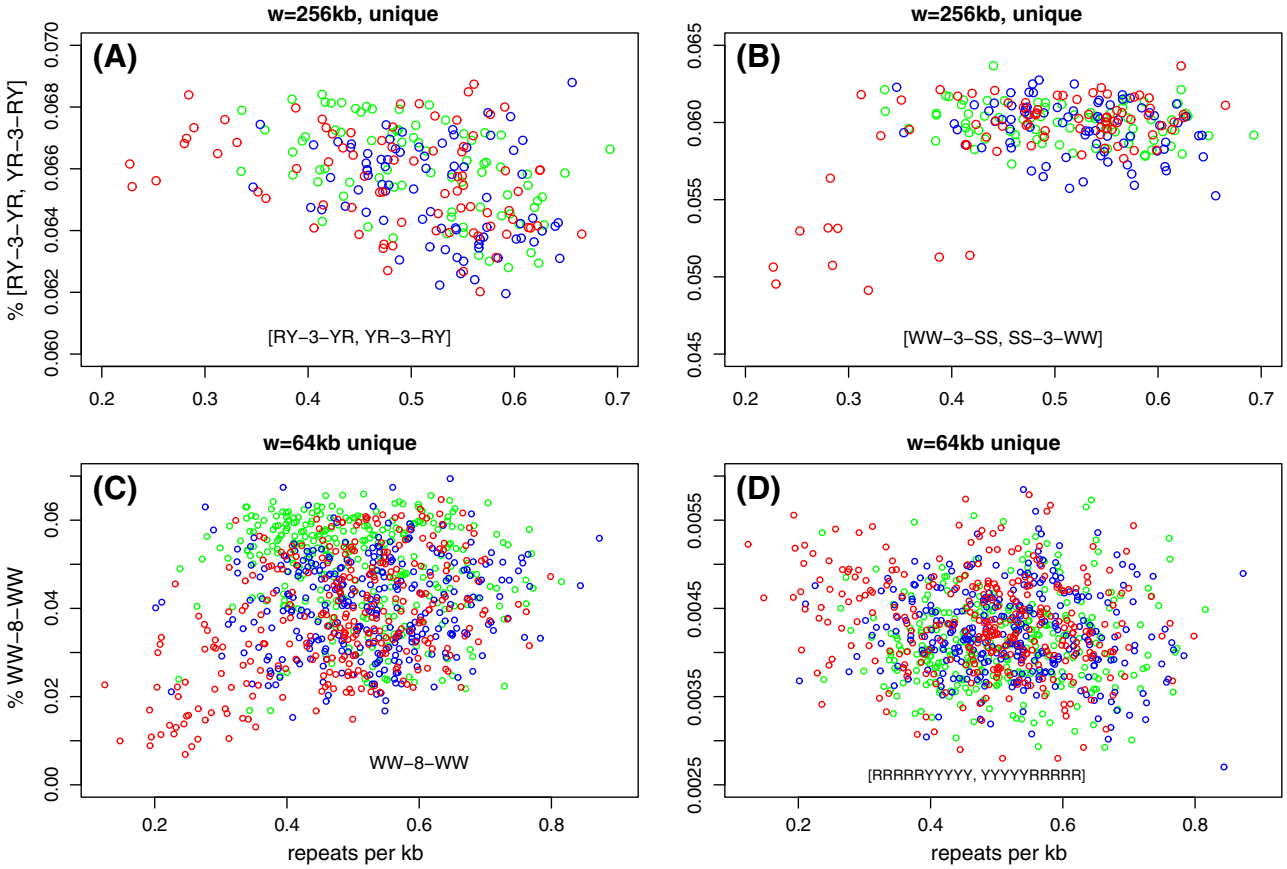

Fig. 2. Scatter plot of NPM densities within unique sequences (repeat-filtered sequences) in non-overlapping windows versus repetitive sequence density (in the same window) for human chromosome 20. The windows from the left, middle, and right 1/3 of the chromosome are labeled by green, blue, and red colors, respectively. (A) [RY-3-YR, RY-3-YR], window size is 256 kb; (B) [WW-3-SS, SS-3-WW], window size is 256 kb; (C) WW-8-WW, window size is 64 kb; (D) [RRRRRYYYYY, YYYYYRRRRR], window size is 64 kb.

i.e., in the sequence after the repetitive sequences are filtered/removed. Windows with 100% repetitive sequences are discarded.

Table 1(A) (right columns) shows that the correlation between NPM density and repetitive sequence density remains negative, with comparable magnitude of cc and *p*-values. The scatter plot in Fig. 2(A) shows this trend more directly at the 256 kb window size.

2.2. Human chromosome 20, [WW-3-SS, SS-3-WW] motif

The second NPM we examine is the [WW-3-SS, SS-3-WW] [9] whose density in chromosome 20 is 0.06 copies per base if the next motif is at least 7 base away from the current one, but 0.085 if overlapping is allowed. Note that SS-3-WW is not only a 5-base shift of the SS-3-WW tandem repeat, but also a reverse complement pattern of WW-3-SS.

2.2.1. WW-3-SS and SS-3-WW density is positively correlated with the repetitive sequence density

Direct calculation of correlation coefficient, both Pearson's and non-parametric Spearman's, at different window sizes, shows that [WW-3-SS, SS-3-WW] density is positively correlated with the repetitive sequence density (Table 1(B)). The statistical significance is the best (*p*-value is indistinguishable from zero) at smaller window sizes, mainly because there are more samples. However, the magnitude of the correlation coefficient increases with the window size. This simultaneous increase of correlation coefficient and decrease of statistical significance with the increase of length scale has been previously observed in other applications [25].

2.2.2. Regional variation still exists

We show the scatter plot for window size 256 kb in Fig. 1(B). The points from the first, second, and last 1/3 of the chromosome are labeled by green, blue, and red colors, respectively. Linear regression of motif density over repetitive sequence density in the three non-overlapping subsets show that the positive correlation mainly originates from the third subset, which contains windows with low repetitive sequence density (and these low repetitive sequence density windows have low motif density). Similar conclusion by wavelet analysis can be found in Table S2 of the Supplementary material.

2.2.3. Density of [WW-3-SS, SS-3-WW] motif in repeat-filtered sequence is not consistently correlated with the repetitive sequence density

When [WW-3-SS, SS-3-WW] motif is obtained from the unique sequence (repeat-filtered/removed sequence), its density becomes negatively correlated with the repetitive sequence density at smaller window sizes (2 kb–16 kb), as shown in Table 1(B). This reversal from positive to negative correlation at these length scales hints that repetitive sequence itself contains the relevant NPMs. However, at larger window sizes, the correlation is back to positive (though less significant) (Table 1(B)).

The scatter plot in Fig. 2(B) shows that the situation is more complicated. Compared to Fig. 1(B), the points in region-3 are still low-repeat-density and low-NPM-density in unique sequence. However, the flat trend for the remaining points in Fig. 1(B) begin to have a negative trend in Fig. 2(B). A single correlation coefficient value cannot describe the nonlinear relationship between the two densities, and the sign of the correlation may depend on the repetitive sequence density.

2.3. Human chromosome 20, WW-8-WW motif

2.3.1. Periodicity-10 of WW dinucleotides is positively correlated with the repetitive sequence density

The results in Table 1(C) shows a very strong positive correlation between WW-8-WW and repetitive sequence density at small window sizes. The corresponding wavelet-based analysis is in Table S3 of the Supplementary material. However, it does not mean lack of heterogeneity. Fig. 1(C) shows that there are more low-repeat-density and low-motif-density windows in the last 1/3 of windows (at window size of 64 kb). Without these windows, the strength of the positive correlation between WW-8-WW and repetitive sequence density would be weaker.

As an [A,T]-rich motif, WW-8-WW is expected to be less common in [G,C]-rich regions. We would like to check whether the repetitive sequences tend to be more [G,C] rich. In chromosome 20, the [A,T]-content in unique sequences is 0.553 which is indeed lower than that in repetitive sequence, 0.564. But this is a very small difference, and its expected effect on WW-8-WW density is only by a ratio of $0.564^4/0.553^4 = 1.08$. This ratio is too small to account for the drop of WW-8-WW density in low-repeat-density regions. The relationship between [G,C]-contents of unique and repetitive sequence at the 100 kb window level was plotted in Fig. 3 of ref. [26], and besides systematic deviation between the two, the [G,C]-contents in the two types of sequences are generally matched.

2.3.2. Positive correlation remains when WW-8-WW motif density is calculated from the unique sequence

When WW-8-WW density is determined from the repeat-filtered sequence, its correlation with the repetitive sequence density remains positive (Table 1(C), Fig. 2(C)). Both the magnitude of cc and *p*-value do not seem to be altered very much.

2.4. Human chromosome 20, [RRRRRYYYYY, YYYYYRRRRR] motif

The [RRRRRYYYYY, YYYYYRRRRR] motif is a more recently proposed NPM whose density in chromosome 20 is 0.0035 copies per base. Interestingly, this observed density is much higher than the expected by the random sequence model (see Table S8 of the Supplementary material). Note that the reverse complement of RRRRRYYYYY is itself (i.e., palindromic).

2.4.1. Decamer [RRRRRYYYYY,YYYYYRRRRR] density is negatively correlated with the repetitive sequence density

A tandem repeat of [RRRRRYYYYY, YYYYYRRRRR] contains the [RY-3-YR, YR-3-RY] motif with NNN replaced by [YYY, RRR]. One may consider [RRRRRYYYYY, YYYYYRRRRR] as a longer, but more specific example of [RY-3-YR, YR-3-RY].

The result in Table 1(D) shows a very strong and statistically significant negative correlation between [RRRRRYYYYY, YYYYYRRRRR] and repetitive sequence density, at almost all length scales examined. Fig. 1(D) shows the scatter plot between the two at window size of 64 kb, marked by whether the window is from the first 1/3, middle 1/3, or the last 1/3 of the chromosome. And Fig. S1 shows the spatial fluctuation of both densities along chromosome 20, as well as the position-scale heatmap by the wavelet transformation.

Different from the similar scatter plots in Figs. 1(A–C), the negative correlation in Fig. 1(D) is consistently observed in all regions (there are exceptions, however, such as an outlier, visible in both Fig. 1(D) and Fig. S1, where a very high motif density appears in a high repetitive sequence density window). The negative slopes of linear regression in the three segments have similar magnitude and similar *p*-values. The consistent negative correlation is also observed in a wavelet-based correlation calculation (Table S4 of the Supplementary material).

2.4.2. Negative correlation remains when [RRRRRYYYYY, YYYYYRRRRR] motif density is calculated from the unique sequence

Table 1(D) shows that neither the magnitude nor the *p*-value of correlation between [RRRRRYYYYY, YYYYYRRRRR] density and repetitive sequence density are much affected, when the NPM density is calculated from the repeat-filtered sequences. Fig. 2(D) shows a scatter plot at the 64 kb window size. When it is compared with the similar plot in Fig. 1(D), the correlation is weaker and much less significant.

2.5. Human chromosome 20, other NPMs

Besides the four proposed NPMs analyzed so far: [YR-3-RY, RY-3-YR], [WW-3-SS, SS-3-WW], WW-8-WW, [RRRRRYYYYY, YYYYYRRRRR], there are other extensions and/or specific proposed NPMs. For example, an extension of [YR-3-RY, RY-3-YR] from spacing-of-5 to spacing-of-10 leads to the [YR-8-YR, RY-8-RY] motif. Another recent proposal of decamer NPM is [GRAAATTTYC, TTTYCGRAAA] [19]. Besides [RRRRRYYYYY, YYYYYRRRRR], the other parental degenerate of [GRAAATTTYC, TTTYCGRAAA] is [SSWWWWWWSS, WWWSSSSWWW]. All these NPMs are palindromic.

We found, generally speaking, densities of [SSWWWWWWSS, WWWSSSSWWW], [GRAAATTTYC, TTTYCGRAAA], and [YR-8-YR, RY-8-RY] to be positively correlated with the repetitive sequence density (see Tables S5,S6,S7 of the Supplementary material). This summary cannot characterize the whole range of complexity of the correlation analyses, as the results may differ at different length scales, between parametric and non-parametric correlation, and between the magnitude of correlation and statistical significance.

The positive correlation between [YR-8-YR, RY-8-RY] and repetitive sequence density is intriguing, as it provides an exception to negative correlation between densities of repetitive sequences and that of R/Y-base NPMs. However, if the NPM density is calculated only within unique sequences, the correlation with the repetitive sequence density becomes negative (and the correlation is statistically very significant). When we take a close look of the correlation by a scatter plot in Fig. 3 (at window size of 64 kb), the [YR-8-YR, RY-8-RY] density is essentially independent of the repetitive sequence density for windows in the first region. For points in the second region, removing an outlier changes the $cc = 0.005$ (p-value $= 2\times10^{-5}$) to $cc = 0.0029$ (p-value $= 2\times10^{-3}$). Both are very weak correlations. Only for the last 1/3 of the chromosome is the positive correlation more significant ($cc = 0.0038$, p-value $= 7\times10^{-6}$). These observations show that regional heterogeneity may affect the sign of the correlation.

2.6. Correlation between repetitive sequence density and proposed nucleosome positioning motifs in other chromosomes

2.6.1. The correlation pattern observed in chromosome 20 is consistently observed in all other chromosomes in the human genome

Calculations carried out on chromosome 20 are extended to all autosomal chromosomes. Table 2 is intended to summarize a large number of results which are all based on consecutively doubling of window sizes from 1 kb to 8.192 Mb. Note that very high percentage of all windows are used (second column in Table 2) in the correlation analysis (the filtering criterion being that 90% of bases within the window are typed), with the only low percentages being in acrocentric chromosomes (13, 14, 15, 21, 22) due to untyped heterochromatin regions.

The [SSWWWWWWSS, WWWSSSSWWW] motif (3 million copies) and the more specific [GRAAATTTYC, TTTYCGRAAA] motif (a low-count of only 37,000 copies) are positively correlated with the repetitive sequence in all chromosomes and almost all window sizes. The [SS-3-WW, WW-3-SS] (190 million copies), WW-8-WW (249 million copies), [RY-8-RY, YR-8-YR] (269 million copies) motifs are positively correlated with repetitive sequence density for most chromosomes, though the correlation could become negative at

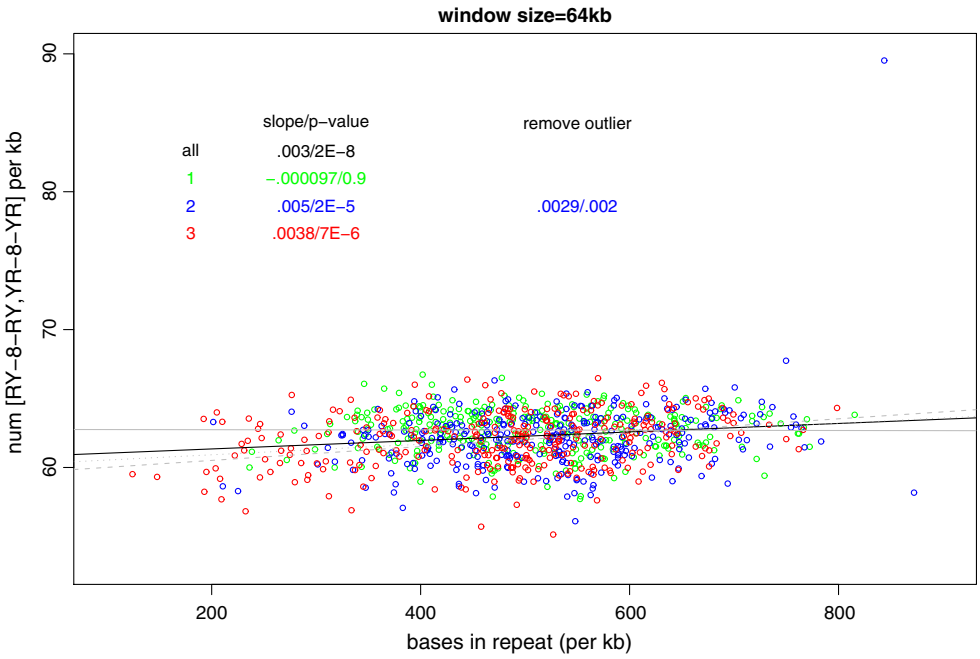

Fig. 3. Scatter plot of number of copies of [RY-8-RY, YR-8-YR] versus number of bases in repetitive sequence at window size of 64 kb. Data are from chromosome 20 only.

larger window sizes. These inconsistency may be caused by spatial heterogeneities in the correlation. The [RRRRRYYYYY, YYYYYRRRRR] (9.5 million copies) and [RY-3-YR, YR-3-RY] (274 million copies) are negatively correlated with repetitive sequence density for most chromosomes and for most window sizes, though the correlation may become positive at larger window sizes in some chromosomes.

2.6.2. Combining all chromosomes into one dataset for correlation analysis

When windows from all chromosomes are combined to one analysis, due to the increase of sample size, statistical significance for testing zero correlation is expected to improve (smaller *p*-values). Since there is only one single correlation calculation between a NPM and repetitive sequence density, heterogeneity between chromosomes will be a factor. All the signs of correlation obtained in chromosome 20 data are confirmed in the combined genome-wide data (see Tables S9–S15 of the Supplementary material).

Fig. 4 shows the scatter plot of four NPM densities versus repetitive sequence densities at the window size of 256 kb, with the genome-wide data. Comparing Fig. 4 with Fig. 1, the negative correlation with the two R/Y-based NPMs (Figs. 4(A, D)) and positive correlation with the two W/S-based NPMs (Figs. 4(B, C)) are confirmed. The scatter plot for [RRRRRYYYYY, YYYYYRRRRR] is particularly interesting: despite the negative trend followed by the majority of the points, there is a minority trend for high repetitive sequence densities and high NPM densities.

3. Discussion

In principle, there could be three types of NPMs using binary symbols: those based on R/Y, W/S, and on M/K (M for amino, C or A, K for keto, G or T). The first two types have been studied in this paper, but not the M/K-based ones. One simple explanation is that

Table 2
Correlation between the seven NPMs and repetitive sequence in 22 human autosomal chromosomes. The "+" ("−") mean positive (negative) correlation; "−/+" means negative correlation at smaller window sizes but positive correlation at larger window sizes; "[ns]" means the correlation is not statistically significant (p-value > 0.01).

chromosome	%w used	RRRRRYYYYY	SSWWWWWWSS	GRAAATTTYC	RY-3-YR	SS-3-WW	WW-8-WW	RY-8-RY
1	90.4	−	+	+	+/− [ns]	+	+	+
2	97.9	−	+	+	− [ns]	+/− [ns]	+	+
3	98.4	−	+	+	−	+	+	+
4	98.2	−	+	+	−/+[ns]	+	−/+	−/+
5	98.2	−	+	+	−	+	+	+
6	97.8	−	+	+	+[ns]	+	+	+
7	97.6	−/+[ns]	+	+	−	+	+/− [ns]	+/− [ns]
8	97.6	−/+	+	+	−	+	+	+/− [ns]
9	85.1	−	+	+	− [ns]/+[ns]	+	+	+
10	96.9	−	+	+	−	+	+	+
11	97.1	−/+	+	+	+	+	+	+
12	97.5	−/+	+	+	− [ns]	+	+	+
13	83.0	−	+	+	−	+/-[ns]	+	−/+
14	82.2	−	+	+	− [ns]/+[ns]	+/-[ns]	+	+
15	79.7	−	+	+	+	+/−	+	+
16	87.3	−	+	+	−	+	+	+
17	95.8	−	+	+	+	+	+	+
18	95.6	−	+	+	−	+/− [ns]	+	+
19	94.4	−	+	+	−	+	+	+
20	94.4	−	+	+	−	+	+	+
21	72.9	−	+	+	−	+/− [ns]	+	+
22	68.0	−	+	+	−	+	+	+

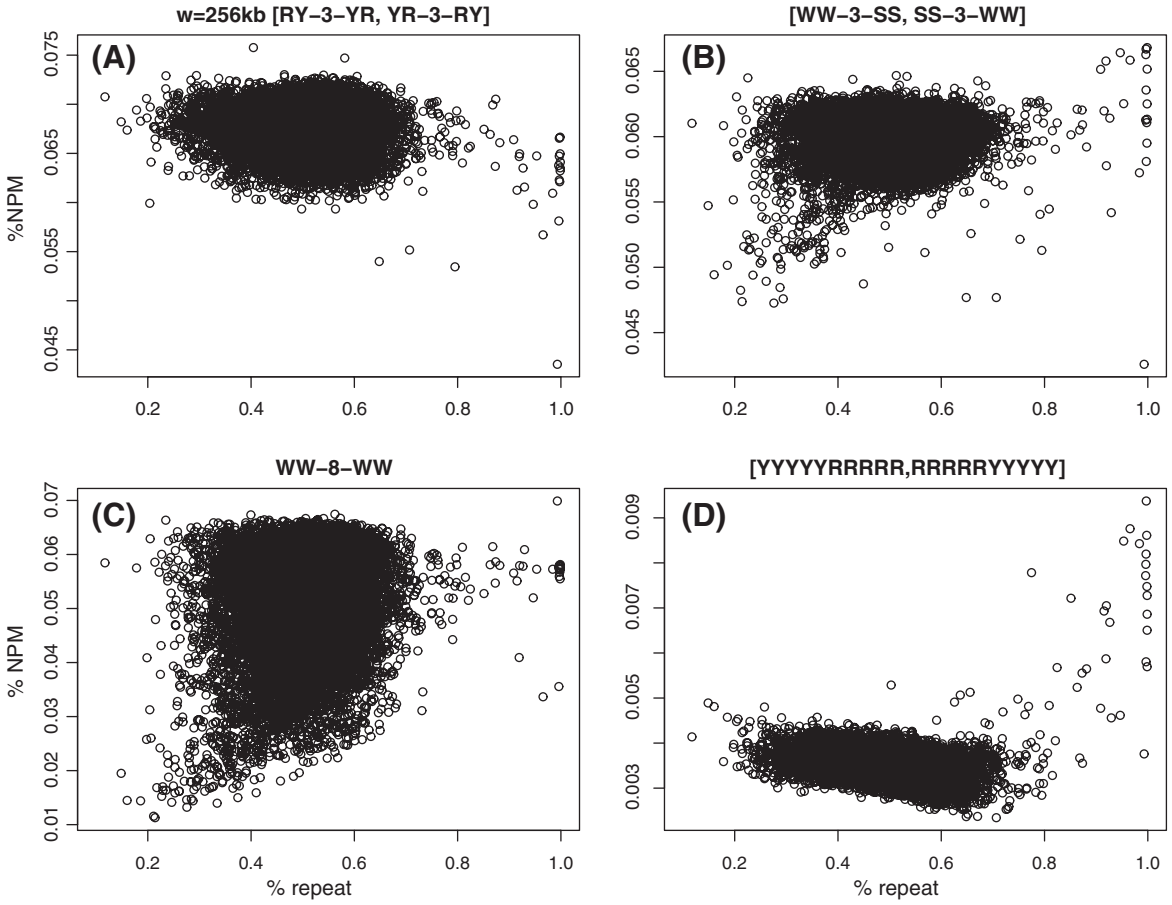

Fig. 4. Scatter plot of genome-wide (chromosomes 1-22) NPM densities versus repetitive sequence density at the level of 256 kb windows. (A) [RY-3-YR, RY-3-YR]; (B) [WW-3-SS, SS-3-WW]; (C) WW-8-WW; and (D) [RRRRRYYYYY, YYYYYRRRRR].

none such motif has been proposed. W/S-based motifs are closely related to the thermodynamic stability of the double helix DNA molecule. R/Y-based motifs, due to the difference of nucleotide sizes (R is larger than Y) and the limitation of physical space, are intrinsically related to the bending and rigidity of the DNA molecule. If a M/K-based motif exists, it could be either due to interactions between keto bases (G or T) and the histones, or related to the fact that keto bases have two alternative forms of the structure (keto vs. enol form).

The total number of copies of a NPM in the genome provides important information on how much this NPM may contribute to the nucleosome positioning. It is estimated that 20% of the human genome [15], around 600 Mb, or even more [27], are occupied by nucleosome with stable positioning. In order for repeating [RRRRRYYYYY, YYYYYRRRRR] motif to cover the 600 Mb region, 60 to 120 million copies of them are needed. When only 9 million copies (or 3% of the genome) are actually observed (Table S8), several consequences can be expected.

One is that it is less likely to observe the periodicity-10 signal as there would not be enough copies of the motif to repeat tandemly (D. Sosa, P. Miramontes, W. Li, V. Mireles, J.R. Bobadilla, M.V. José, unpublished results). Another consequence of lower number of copies of a NPM is that instead of a densely packing of the NPM in nucleosome regions, we may only need a few NPM per nucleosome, while other factors contribute to the positioning. For example, it is suggested in ref. [28] that barriers near a gene's promoter region may help the positioning of nucleosomes. It raises the question of the importance of not only a particular proposed NMP in nucleosome positioning, but also of roles played by DNA sequence in general.

We are not aware of previous studies on the correlation between NPMs and repetitive sequences. In ref. [29], an experimentally obtained nucleosome signal is plotted within and around (up to 1 kb) the *Alu* element. The goal of this experimental study is very different from ours as it is centered around the *Alu* element and within a much smaller length scale (the peak-to-trough distance is 200 bp). Even if we know where the nucleosome signal is located within an *Alu* element, we still do not know whether the presence of *Alu* sequence increases the nucleosome positioning probability in that region, though *Alu* elements were claimed to confer nucleosome positioning *in vitro* [30].

Besides treating *Alu* as a subgroup, there are also subgroups within *Alu*. There are roughly 40 different *Alu* sequences such as *Alu*Jb, *Alu*Sx, *Alu*Y, *Alu*Sx1, each of which with more than 100,000 copies. The *Alu*Y sequence is in a relatively younger group [31]. There are also human-specific branches of *Alu*, i.e., Yc1, Ya5a2, Yb9 [32], with much lower frequencies. Preliminary analyses show that most of our results between NPM densities and repetitive sequence densities hold true for *Alu* or *Alu*Y densities also. However, the correlation with densities of W/S-based NPMs may become negative.

In ref. [33], the autocorrelation function of CG dinucleotide is calculated for the original and the repeat-masked sequence. Peaks at distances of 31 and 62 bps disappear in the repeat-masked sequence, but at the same time, new peaks at distances 10 and 21 appear. In that paper, any peak at a multiple of 10 bp is considered to be a nucleosome positioning signal, then such signal is present in both *Alu* and non-repetitive sequences. The authors of ref. [33] suggested that *Alu* elements might play a role of “anchor” for nucleosomes, which is reminiscent of the barrier idea in ref. [15]. If both positive and negative correlations exist between NPMs and repetitive sequence density, it indicates nucleosome positioning information can be enriched either in repetitive sequences or in unique sequences.

Whether more repetitive sequences in a genome increase or decrease the probability for nucleosome positioning may provide insight on the evolution of repetitive sequences [34]. Most repetitive sequences are transposable elements caused by at least three mechanisms [35] and are particularly abundant in sexual organisms [36]. As a major force in expanding the higher organisms' genome including the human's [37], it must have an effect on the genome function [38–44]. But most of the focus concerning impact of repetitive sequences is on the genomic instability introduced, genetic innovation accompanied by the extra DNA sequences [41], and gene expression or regulatory networks [45]. Discussion on repetitive sequences' impact via nucleosome formation was mostly in promoter region [46,47].

The results in this paper hint that repetitive sequences can also have subtle and complicated impact to nucleosome-forming potential by either increasing or decreasing NPM density in repetitive sequence regions. This effect can be small, and may be detectable only in local regions with extreme densities of repetitive sequence.

Detection of sequence signal or statistical correlation between any two sequence measures can always be more complex than the apparent calculations. First, stratifying sequence data by controlling other quantities can much weaken a signal or a correlation. For example, the periodicity of R-7-R [48] or WW-8-NWW motifs [49] around a CpG dinucleotide can be absent if the CpG dinucleotide is located in a [G,C]-rich and unmethylated CpG island. Second, when multiple sequence measures are pairwisely correlated, the cause-effect relationship between these measures intrinsically affect conditional correlation result [50]. The idea that repetitive sequences have relevant evolutionary impact on higher organisms only under certain conditions was discussed in ref. [51]. Whether the correlation between NPM and repetitive sequence densities discussed here can disappear by conditionaling on other sequence measures worth future studies.

4. Materials and methods

4.1. Human DNA sequence data

The GRCh37/hg19 (Feb. 2009) version of the human genome sequence is downloaded from UCSC's genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/). Repeat sequences are marked as lowercase in the file as versus the uppercase letters for unique sequences. For specific repetitive sequence family, we use the rmsk.txt file from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/ which lists the starting and ending positions of 5.298 millions occurrence of more than 1300 different types of repetitive sequences.

4.2. Motif countings

For some NMPs, whether overlapping motifs are counted as more than one copy or not will change the counting value. For example, WW-3-SS may overlap with another WW-3-SS shifted by one base position, while we may consider both to contribute to one NPM. The countings for the NMP density calculation in Table S8, however, are all obtained by shifting one position.

4.3. Motif density in unique sequences

For a window with N bases (e.g. $N=1000$), N_{not}, N_{rep}, N_{uniq} are the number of bases that are not sequenced, part of a repetitive sequence, or not part of the repetitive sequence (thus part of the unique sequence), and $N=N_{\text{not}}+N_{\text{rep}}+N_{\text{uniq}}$. During the quality control stage, windows with sequencing rate ($N_{\text{not}}/\text{N}\leq 0.9$) are discarded, so for almost all windows used, $N_{\text{not}}=0$. Denote n and n_{uniq} as the number of copies of a NPM in the window and in the unique sequence within the window, respectively, then $n/(N_{\text{rep}}+N_{\text{uniq}})$ is the NPM density, and $n_{\text{uniq}}/N_{\text{uniq}}$ is the NPM density in the unique sequence.

4.4. Statistical methods

Pearson's and Spearman's statistical correlation and the corresponding tests were carried out by the *cor.test* function in *R* (http://www.r-project.org/), with the option *method* = *"pearson"* (default) or *method* = *"spearman"*. Spearman's correlation is simply a Pearson's correlation by replacing the raw data with its ranking values. Kendall's correlation coefficient is, like Spearman's correlation, another non-parametric measure of correlation, defined as $(\#\text{ concordant pairs}-\#\text{ discordant pairs})/(n(n-1)/2)$, and can be calculated by the above *R* function with *method* = *"kendall"*.

4.5. Wavelet analysis

Wavelet transformation [52] provides an alternative way in dealing with correlation analysis at different length scales. We adopt the *R* routines *plot.pair.wavelet* used in ref. [25] with the Haar wavelet basis, which requires the installation of two *R* packages: *Rwave* [53] and *wavethresh* [54].

Acknowledgments

We thank Pedro Miramontes, Jan Freudenberg, Victor Mireles for discussions, and W.L. acknowledges the support from the Robert S Boas Center for Genomics and Human Genetics. MVJ acknowledges support from PAPIIT UNAM, project IN107112.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2012.10.005.

References

[1] A. Wolffe, Chromatin: Structure and Function, 3rd edition Academic Press, 1999.
[2] C.L. Woodcock, R.P. Ghosh, Chromatin higher-order structure and dynamics, Cold Spring Harb. Perspect. Biol. 2 (2010) a000596.
[3] G. Li, D. Reinberg, Chromatin higher-order structures and gene regulation, Curr. Opin. Genet. Dev. 21 (2011) 175–186.
[4] A. Kundaje, S. Kyriazopoulou-Panagiotopoulou, M. Libbrecht, C.L. Smith, D. Raha, E.E. Winters, S.M. Johnson, M. Snyder, S. Batzoglou, A. Sidow, Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements, Genome Res. 22 (2012) 1735–1747.
[5] R.E. Thurman, et al., The accessible chromatin landscape of the human genome, Nature 489 (2012) 75–82.
[6] C.R. Calladine, H. Drew, B. Luisi, A. Travers, Understanding DNA: The Molecule and How it Works, 3rd edition Academic Press, 2004.
[7] T.C. Bishop, Chromatin in 1, 2 and 3 dimensions. Comment on 'Cracking the chromatin code: precise rule of nucleosome positioning' by E.N. Trifonov, Phys. Life Rev. 8 (2011) 56–58.
[8] V.B. Zhurkin, Specific alignment of nucleosomes on DNA correlates with periodic distribution of purine–pyrimidine and pyrimidine–purine dimers, FEBS Lett. 158 (1983) 293–297.
[9] S.C. Satchwell, H.R. Drew, A.A. Travers, Sequence periodicities in chicken nucleosome core DNA, J. Mol. Biol. 191 (1986) 659–675.
[10] X. Wang, G.O. Bryant, M. Floer, D. Spagna, M. Ptashne, An effect of DNA sequence on nucleosome occupancy and removal, Nat. Struct. Mol. Biol. 18 (2011) 507–509.
[11] E.N. Trifonov, J.L. Sussman, The pitch of chromatin DNA is reflected in its nucleotide sequence, Proc. Natl. Acad. Sci. 77 (1980) 3816–3820.
[12] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I.K. Moore, J.Z. Wang, J. Widom, A genomic code for nucleosome positioning, Nature 442 (2006) 772–778.
[13] S.M. Johnson, F.J. Tan, H.L. McCullough, D.P. Riordan, A.Z. Fire, Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin, Genome Res. 16 (2006) 1505–1516.
[14] A. Valouev, J. Ichikawam, T. Tonthat, J. Stuart, S. Ranade, H. Packham, K. Zeng, J.A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, S.M. Johnson, A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning, Genome Res. 18 (2008) 1051–1063.
[15] A. Valouev, S.M. Johnson, S.D. Boyd, C.L. Smith, A.Z. Fire, A. Sidow, Determinants of nucleosome organization in primary human cells, Nature 474 (2011) 516–520.

[16] Z. Zhang, C.H. Wippo, M. Wal, E. Ward, P. Korber, B.F. Pugh, A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome, Science 332 (2011) 977–980.
[17] I. Gabdank, D. Barash, E.N. Trifonov, Nucleosome DNA bendability matrix (*C. elegans*), J. Biomol. Struct. Dyn. 26 (2009) 403–411.
[18] E.N. Trifonov, Nucleosome positioning by sequence, state of the art and apparent finale, J. Biomol. Struct. Dyn. 27 (2010) 741–746.
[19] E.N. Trifonov, Cracking the chromatin code: precise rule of nucleosome positioning, Phys. Life Rev. 8 (2011) 39–50.
[20] R.J. Britten, DE Kohne, Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms, Science 161 (1968) 529–540.
[21] W. Li, On parameters of the human genome, J. Theol. Biol. 288 (2011) 92–104.
[22] A.P.J. de Koning, W. Gu, T.A. Castoe, M.A. Batzer, D.D. Pollock, Repetitive elements may comprise over two-thirds of the human genome, PLoS Genet. 7 (2011) e1002384.
[23] H. Takata, K. Maeshima, Irregular folding of nucleosomes in the cell. Comment on 'Cracking the chromatin code: precise rule of nucleosome positioning' by Edward N. Trifonov, Phys. Life Rev. 8 (2011) 51–52.
[24] W. Li, Are isochore sequences homogeneous? Gene 300 (2002) 129–139.
[25] C.C.A. Spencer, P. Deloukas, S. Hunt, J. Mullikin, S. Myers, B. Silverman, P. Donnelly, D. Bentley, G. McVean, The influence of recombination on human genetic diversity, PLoS Genet. 2 (2006) e148.
[26] J. Paces, R. Zíka, V. Paces, A. Pavlícek, O. Clay, G. Bernardi, Representing GC variation along eukaryotic chromosomes, Gene 333 (2004) 135–141.
[27] D.J. Gaffney, G. McVicker, Y. Fondufe-Mittendorf, J. Widom, Y. Gilad, J.K. Pritchard, Most nucleosomes in the human genome are consistently positioned, The Biology of Genome (May 8-12, 2012, Cold Spring Harbor Laboratory). http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36979.
[28] Y. Zhang, Z. Moqtaderi, B.P. Rattner, G. Euskirchen, M. Snyder, J.T. Kadonaga, X.S. Liu, K. Struh, Intrinsic histone–DNA interactions are not the major determinant of nucleosome positions in vivo, Nat. Struct. Mol. Biol. 16 (2009) 847–852.
[29] Y. Tanaka, R. Yamashita, Y. Suzuki, K. Nakai, Effects of Alu elements on global nucleosome positioning in the human genome, BMC Genomics 11 (2010) 309.
[30] E.W. Englander, B.H. Howard, Nucleosome positioning by human Alu elements in chromatin, J. Biol. Chem. 270 (1995) 10091–10096.
[31] M. Costantini, F. Auletta, G. Bernardi, The distribution of 'new' and 'old' Alu sequences in the human genome: the solution of a 'mystery', Mol. Biol. Evol. 29 (2012) 421–427.
[32] M.A. Batzer, P.L. Deininger, Alu repeats and human genome diversity, Nat. Rev. Genet. 3 (2002) 370–379.
[33] T. Bettecken, Z.M. Frenkel, E.N. Trifonov, Human nucleosomes: special role of CG dinucleotides and Alu-nucleosomes, BMC Genomics 12 (2011) 273.
[34] J. Jurka, V.V. Kapitonov, O. Kohany, M.V. Jurka, Repetitive sequences in complex genomes: structure and evolution, Ann. Rev. Genomics Hum. Genet. 8 (2007) 241–259.
[35] A.F.A. Smit, The origin of interspersed repeats in the human genome, Curr. Opin. Genet. Dev. 6 (1996) 743–748.
[36] D.A. Hickey, Evolutionary dynamics of transposable elements in prokaryotes and eukaryotes, Genetica 86 (1992) 269–274.
[37] H.H. Kazazian Jr., Mobile elements: drivers of genome evolution, Science 303 (2004) 1626–1632.
[38] M. Syvanen, The evolutionary implications of mobile genetic elements, Annu. Rev. Genet. 18 (1984) 271–293.
[39] D.J. Finnegan, Eukaryotic transposable elements and genome evolution, Trends Genet. 5 (1989) 103–107.
[40] C. Feschotte, E.J. Pritham, DNA transposons and the evolution of eukaryotic genomes, Annu. Rev. Genet. 41 (2007) 331–368.
[41] R. Cordaux, M.A. Batzer, The impact of retrotransposons on human genome evolution, Nat. Rev. Genet. 10 (2009) 691–703.
[42] C. Biémont, A brief history of the status of transposable elements: from junk DNA to major players in evolution, Genetics 186 (2010) 1085–1093.
[43] J.A. Shapiro, Mobile DNA and evolution in the 21st century, Mob. DNA 1 (2010) 4.
[44] A. Hua-Van, A. le Rouiz, T.S. Boutin, J. Filée, P. Capy, The struggle for life of the genome's selfish architects, Biol. Direct 6 (2011) 19.
[45] C. Feschotte, Transposable elements and the evolution of regulatory networks, Nat. Rev. Genet. 9 (2008) 397–405.
[46] A. Huda, L. Mariño-Ramirez, D. Landsman, I.K. Jordan, Repetitive DNA elements, nucleosome binding and human gene expression, Gene 436 (2009) 12–22.
[47] A. Huda, Epigenetic Regulation of the Human Genome by Transposable Elements, Ph.D Thesis, Georgia Institute of Technology, 2010.
[48] O. Clay, W. Schaffner, K. Matsuo, Periodicity of eight nucleotides in purine distribution around human genomic CpG dinucleotides, Somat. Cell Mol. Genet. 21 (1995) 91–98.
[49] A. Tanay, A.H. O'Donnell, M. Damelin, T.H. Bestor, Hyperconserved CpG domains underlie polycomb-binding sites, Proc. Natl. Acad. Sci. 104 (2007) 5521–5526.
[50] J. Freudenberg, M. Wang, Y. Yang, W. Li, Partial correlation analysis indicates causal relationships between GC-content, exon density and recombination rate variation in the human genome, BMC Bioinform. 10 (Suppl. 1) (2009) S66.
[51] E. Zuckerkandl, G. Cavalli, Combinatorial epigenetics, "junk DNA", and the evolution of complex organisms, Gene 390 (2007) 232–242.
[52] D.B. Percival, A.T. Walden, Wavelet Methods for Time Series Analysis, Cambridge University Press, 2005.
[53] R. Carmona, W.L. Hwang, B. Torresani, Practical Time–Frequency Analysis: Gabor and Wavelet Transformations With an Implementation in S, Academic Press, 1998.
[54] G. Nason, Wavelet Methods in Statistics with R, Springer, 2008.