Workshop

Carlos A. C. Bastos*, Vera Afreixo, João M. O. S. Rodrigues and Armando J. Pinho

Concentration of inverted repeats along human DNA

https://doi.org/10.1515/jib-2022-0052

Received October 24, 2022; accepted February 27, 2023; published online July 25, 2023

Abstract: This work aims to describe the observed enrichment of inverted repeats in the human genome; and to identify and describe, with detailed length profiles, the regions with significant and relevant enriched occurrence of inverted repeats. The enrichment is assessed and tested with a recently proposed measure (*z*-scores based measure). We simulate a genome using an order 7 Markov model trained with the data from the real genome. The simulated genome is used to establish the critical values which are used as decision thresholds to identify the regions with significant enriched concentrations. Several human genome regions are highly enriched in the occurrence of inverted repeats. This is observed in all the human chromosomes. The distribution of inverted repeat lengths varies along the genome. The majority of the regions with severely exaggerated enrichment contain mainly short length inverted repeats. There are also regions with regular peaks along the inverted repeats lengths distribution (periodic regularities) and other regions with exaggerated enrichment for long lengths (less frequent). However, adjacent regions tend to have similar distributions.

Keywords: distance distribution; human genome; inverted repeats; Markov model

1 Introduction

The non-B DNA structures play important roles for biological processes (see for example, [1–6]). This work focus on the study of the lengths of potential hairpins/cruciforms non-B DNA structures. Various studies in the literature argue that cruciform structures are a common DNA feature important for regulating biological processes and that the genomes contain a remarkable number of inverted repeats in a non-random distribution (see [7] for a review).

Cruciform structures are formed by inverted repeats, which are composed by a single stranded sequence of nucleotides followed downstream by its reverse complement. There are several procedures and computational approaches to study inverted repeats and to describe the regional variation of inverted repeat lengths [8–10].

***Corresponding author: Carlos A. C. Bastos**, DETI – Department of Electronics, Telecommunications and Informatics, IEETA – Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, 3810-193 Aveiro, Portugal; and LASI – Intelligent Systems Associate Laboratory, Aveiro, Portugal, E-mail: cbastos@ua.pt. https://orcid.org/0000-0001-6869-4713

Vera Afreixo, CIDMA – Center for Research and Development in Mathematics and Applications, DMAT – Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal

João M. O. S. Rodrigues and Armando J. Pinho, DETI – Department of Electronics, Telecommunications and Informatics, IEETA – Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, 3810-193 Aveiro, Portugal; and LASI – Intelligent Systems Associate Laboratory, Aveiro, Portugal. https://orcid.org/0000-0001-9187-8094 (J.M.O.S. Rodrigues). https://orcid.org/0000-0002-9164-0016 (A.J. Pinho)

In this work we have developed a simulation scenario in order to highlight/identify the regions with exaggerated enrichment globally (of all lengths) and by lengths sections. We also present an exploratory analysis of all human chromosomes inverted repeat enrichment globally (in all genome) and in each chromosome separately.

2 Methods

This work uses the human genome (GRCh38 assembly sequences) and searches for features beyond the already well-known repetition structures published in the literature. Thus, the pre-masked sequences available from the UCSC Genome Browser webpage [11] with repeats reported by RepeatMasker [12] and Tandem Repeats Finder [13] masked with N symbols were used. The unknown or ambiguous nucleotides are usually coded with N symbols. All these ambiguous nucleotides are considered separators that split the sequence into a set of unambiguous subsequences.

Inverted repeats are nucleotide sequences that can form self complementary pairings between their two halves e.g. *CCTTACGnnnnnnCGTAAGG*, where $\{A, C, G, T\}$ represent the DNA alphabet and *nnnnnn* represent a sequence of nucleotides with a known length.

In this work, we analyse the distribution of the distances between reverse complement sequences with $k = 7$ nucleotides, which is similar to the distribution of the lengths of the inverted repeats of 7 nucleotides. We study this distribution along the genome by dividing the complete genome in successive windows containing 10^5 nucleotides.

For all words of length k, we compute the frequency distributions of each distance, $m(d)$, between occurrences of each word and all succeeding reversed complements at distances between k and 4000.

2.1 Measuring the concentration of inverted repeats

In order to evaluate the behaviour of the observed values of the $m(d)$, inverted repeat cumulative frequencies are compared to the corresponding expected values obtained from a Markov chain reference model of order 7.

2.1.1 Expected values under higher order Markov chain for DNA sequences: Let $M(d)$ be the random variable that represents the total number of inverted repeats occurrence at distance d in a genomic region of length L and $n(d)$ the corresponding total number of possible word pairs at distance d, where $d = k, k+1, \ldots, 4000$ and $d = k$ means that the two words (of length k) are in adjacent positions.

Let $p(d)$ be the probability of occurrence of inverted repeats at distance d. If we assume the independence between trials, $M(d)$ follows a binomial distribution, $M(d) \frown B(n(d), p(d))$ with expected value $n(d)\,p(d)$ and standard deviation $\sqrt{n(d)\,p(d)(1-p(d))}$. We assumed a Markov model of order k to estimate $p(d)$ since the occurrence of reversed complements cannot be considered independent.

We use a z-score, as proposed recently [14], as a measure between the observed values and the expected values obtained from the Markov model,

$$Z(d) = \frac{M(d) - n(d)\,p(d)}{\sqrt{n(d)\,p(d)(1-p(d))}}. \tag{1}$$

In order to measure the concentration of inverted repeats for a set of successive distances we compute the sum of all T values between two bounds ($d1$ and $d2$)

$$S_{[d1,d2]} = \sum_{d\in\{d1,\ldots,d2\}} T(d), \tag{2}$$

with $T(d)$ a z-score adjusted to account for the effect of the presence of ambiguous symbols in the sequence [14].

2.1.2 Simulation study: A control scenario simulation was developed and run to evaluate the results of the S measure under controlled conditions. Control scenario conditions:

- 24 sequences with the same size of each of the human chromosomes;
- the same number and in the same positions of the ambiguous symbols (Ns) in human chromosomes;
- the DNA sequences were generated by a 7-order Markovian model;
- the probabilities of the words and the transition matrices of the Markov model, were estimated from each chromosome sequence.

We use the results of the simulation procedure to obtain a critical value for the S statistic (Equation (2)) under the assumption that the DNA sequences were generated by a 7-order Markovian model. An empirical distribution, of the S measure from the simulated sequences, was generated for each chromosome combining the contribution of all the windows in each chromosome.

We compute the critical values on the empirical distribution of the simulated genome, assuming a significance level of 5 %. The critical values are the 0.95 quantiles (cv) of the S values of all windows in each chromosome (or globally). Windows with S values surpassing the critical value are considered significantly enriched.

2.2 Data analysis

The data analysis is based on $M(d)$ and $S_{[d1,d2]}$. It is divided into 3 parts:
- comparison of inverted repeats enrichment between chromosomes;
- analysis of inverted repeats enrichment as a function of inverted repeats length;
- analysis of the inverted repeats enrichment as a function of position along each chromosome.

Inverted repeat lengths were grouped into nine classes: $I_t = S_{[7,4000]}$, $I_1 = S_{[7,500]}$, $I_2 = S_{[501,1000]}$, $I_3 = S_{[1001,1500]}$, $I_4 = S_{[1501,2000]}$, $I_5 = S_{[2001,2500]}$, $I_6 = S_{[2501,3000]}$, $I_7 = S_{[3001,3500]}$, $I_8 = S_{[3501,4000]}$.

3 Results and discussion

3.1 Comparison of inverted repeats enrichment between chromosomes

Figure 1 shows the boxplots of $S_{[7,4000]}$ for all chromosomes and all considered inverted repeat lengths (length class I_t), both for the human genome and the control scenario. The distributions of the S values in the human genome are reasonably similar for the various chromosomes: all distributions show positive mean, positive skew and similar dispersion. The distributions of the control scenario are clearly different from those of the human genome: all distributions have null mean, are symmetric and have a comparatively much lower dispersion.

Table 1 shows the percentage of windows that were considered to have significantly enriched concentration of inverted repeats in each chromosome and for each inverted repeat length class. The percentage of enriched windows for all considered inverted repeats lengths (class I_t) in the complete genome is quite large (66.3 %). This confirms the strong positive skew of the S distribution previously observed.

3.2 Analysis of inverted repeats enrichment as a function of inverted repeats length

Table 1 also shows the percentage of windows with significantly enriched concentration of inverted repeats in each chromosome, for each length class (I_1 through I_8). The percentage decreases with the increase of inverted repeat length.

Figure 2 shows the boxplots of S values for each length class in the human genome and in the control scenario.

The human genome reveals some regions with very high and significant enrichment in all length classes. The control scenario shows that the dispersion of S decreases with the increase of inverted repeats length. This reveals that the S measure is sensitive to the inverted repeat length. This limitation of the measure does not compromise our analysis, since critical values were obtained from the control scenario in each length class.

Figure 3 shows the variations of S in the different length classes along chromosome X.

Windows with similar enrichment seem to form regional clusters along the genome.

3.3 Analysis of the inverted repeats enrichment as a function of position along each chromosome

Figure 4 shows the absolute frequencies of occurrence of each inverted repeat length in three different windows of chromosome X. The top/bottom plots pertain to the windows with the highest/lowest $S_{[7,4000]}$ values in that chromosome. The middle plot pertains to the window with the highest $S_{[2001,2500]}$ values.

The Supplementary Material contains the absolute frequency plots for windows selected according to the same criteria in every chromosome.

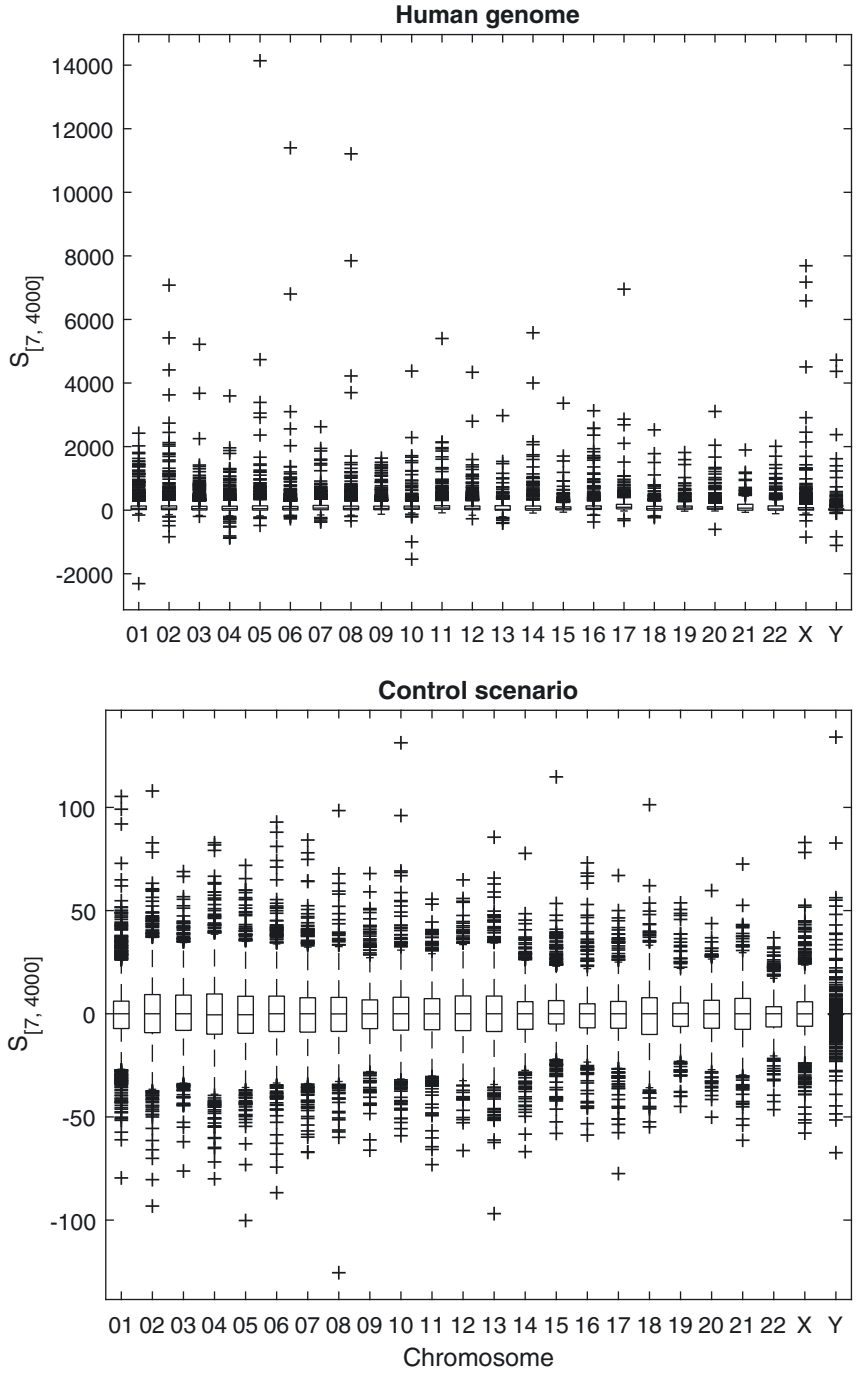

Figure 1: Boxplot of $S_{[7,4000]}$ values for each chromosome: top, human genome; bottom, control scenario.

The selected windows display very different distributions. The distributions on the top and middle plots both present frequency values much higher than the expected values. The bottom plot represents a rare example of a window with negative S values. The periodic regularities seen in the middle plot were previously identified and studied in [9].

Table 1: Percentage of windows with significantly enriched concentration of inverted repeats for each length class. Column “wins” shows the total number of windows in each chromosome. Row “Global %” shows the percentages for the complete genome. Row “Mean cv” shows the weighted mean of the critical values.

Chr	Wins	I_t	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
1	2490	70.6	75.4	45.9	31.0	23.8	15.9	12.2	9.8	8.8
2	2422	67.7	72.9	44.2	30.0	19.2	13.7	11.5	9.0	7.0
3	1983	61.9	67.3	40.0	27.8	20.1	14.3	11.0	8.3	7.4
4	1903	60.4	64.1	41.3	25.3	19.3	11.0	8.8	6.5	5.6
5	1816	59.9	65.5	42.2	29.6	20.6	11.8	9.3	7.2	6.7
6	1709	60.8	68.5	38.2	27.5	18.4	13.6	10.6	8.4	6.7
7	1594	68.3	73.4	46.0	33.6	21.8	16.4	13.2	9.3	7.5
8	1452	67.6	72.2	42.4	29.8	19.8	14.3	11.0	9.6	6.3
9	1384	71.6	73.9	42.6	30.8	21.2	14.9	10.6	8.5	8.1
10	1338	78.7	82.7	44.3	30.2	19.1	12.8	8.8	8.1	7.5
11	1351	79.8	82.9	51.2	33.0	22.6	16.4	14.1	10.9	8.6
12	1333	66.1	71.1	42.6	31.7	20.0	14.1	12.0	8.9	7.4
13	1144	48.9	51.8	36.5	28.1	20.3	12.0	9.3	8.0	5.7
14	1071	61.7	66.3	42.7	29.8	22.2	16.4	11.8	9.1	9.0
15	1020	65.0	69.8	37.1	26.3	17.2	15.6	11.5	10.2	7.5
16	904	78.9	80.3	42.1	31.4	23.7	18.7	14.8	11.3	10.6
17	833	85.7	89.3	59.4	40.6	30.3	23.9	18.4	14.6	11.0
18	804	60.2	64.7	36.9	25.1	17.8	11.6	9.3	6.1	6.7
19	587	83.3	84.7	52.8	40.7	27.6	22.8	18.1	14.8	10.9
20	645	85.0	86.5	46.4	28.1	16.1	16.6	9.6	8.4	6.7
21	468	62.2	65.6	51.3	40.8	22.6	15.4	13.5	11.1	10.7
22	509	62.5	63.5	40.3	34.6	23.8	18.3	16.3	13.2	10.8
X	1561	56.6	60.8	31.3	20.2	14.0	9.8	7.5	5.6	6.1
Y	573	33.2	34.0	23.7	17.8	10.8	8.9	6.3	6.1	3.5
Global %		66.3	70.6	42.4	29.6	20.4	14.5	11.3	8.9	7.5
Mean cv		23.8	18.9	8.0	3.8	1.9	1.1	0.6	0.3	0.1

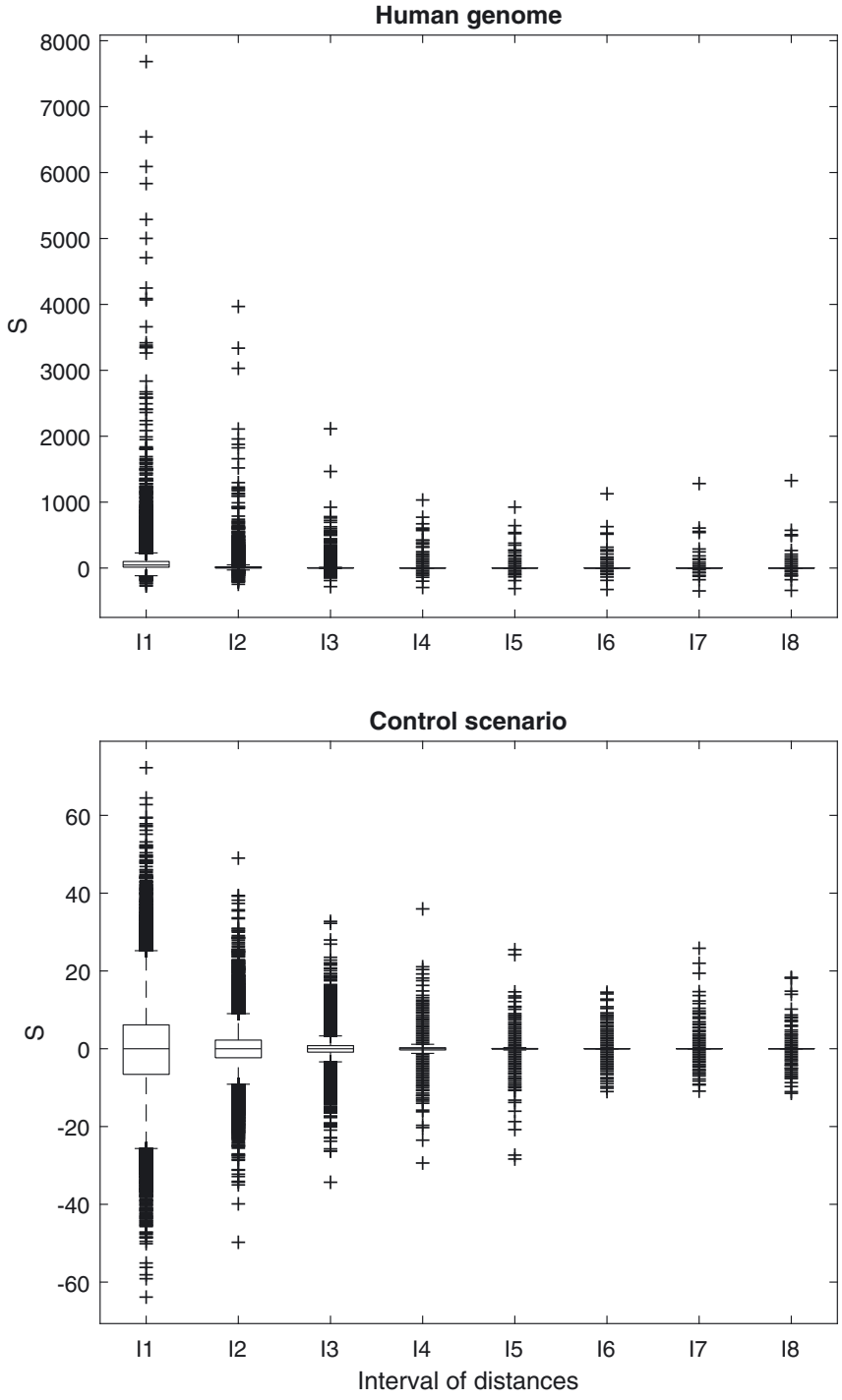

Figure 2: Boxplot of $S_{[d1,d2]}$ values for 8 length classes: top, human genome; bottom, control scenario.

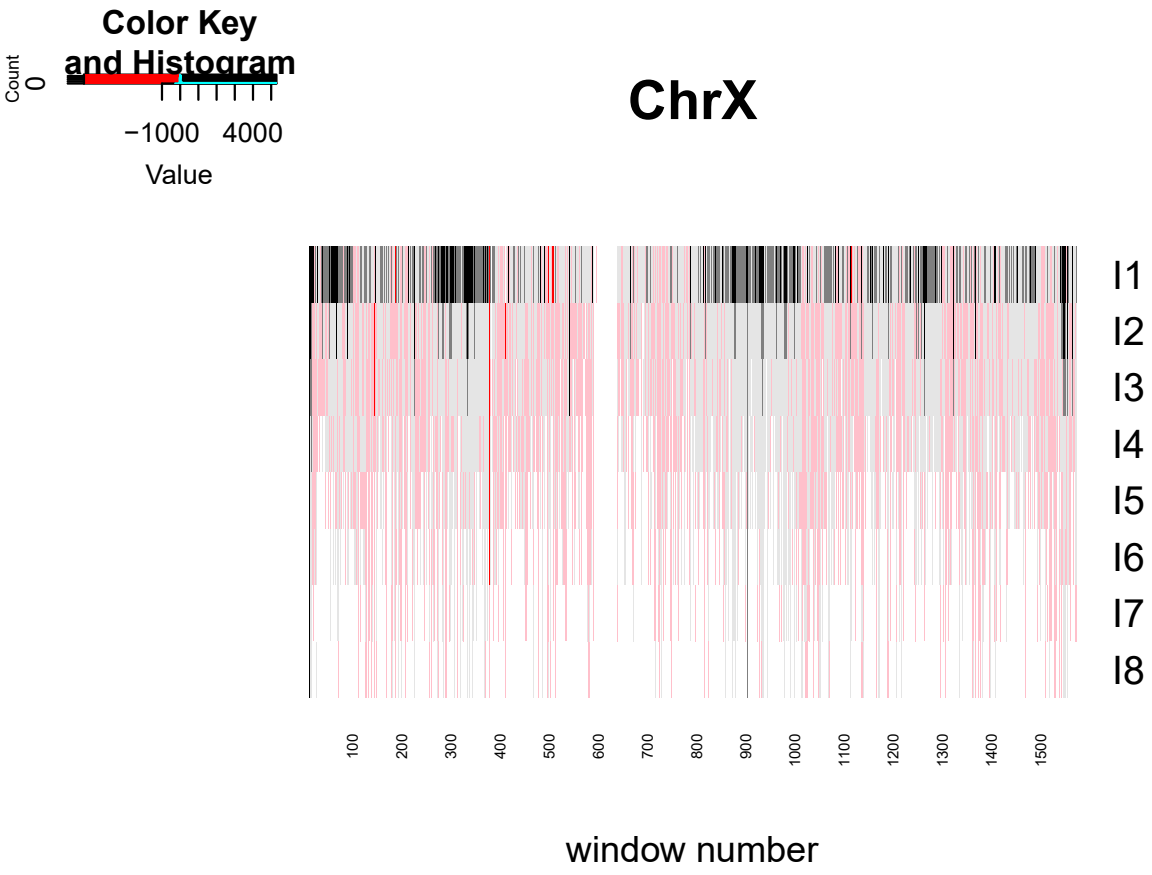

Figure 3: Heatmap of the S values for the inverted repeats length classes $I_1, I_2, \ldots, I_8$ in chromosome X. Black shows enrichment and red shows reduction of the frequency of inverted repeats.

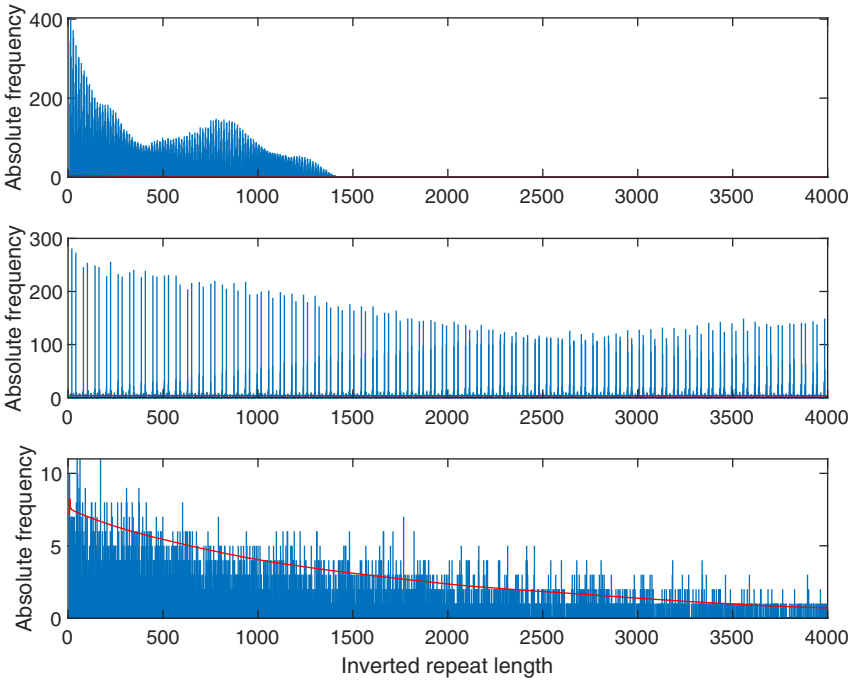

Figure 4: Inverted repeat length frequencies of three windows in chromosome X: top, window with the highest $S_{[7,4000]}$, chrX:53000001:53100000; middle, window with the highest $S_{[2001,2500]}$, chrX:100001:200000; bottom, window with the lowest $S_{[7,4000]}$ chrX:36700001:36800000. The red solid line represents the expected value for the absolute frequency.

4 Conclusions

The analysis carried out in this work revealed several human genome regions with highly enriched occurrence of inverted repeats in all human chromosomes. The enrichment in inverted repeats concentration is not uniform along the genome and it depends on the repeats lengths, being more prominent for short lengths.

Even though we removed well known repetitive sequences, we still found regions with atypically enriched concentration of inverted repeats. Further studies for understanding the reasons for this phenomenon are needed, which may imply analysing the genomic word composition of these regions.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: This work was supported by the Institute of Electronics and Informatics Engineering of Aveiro (IEETA) and Center for Research and Development in Mathematics and Applications (CIDMA) though the Portuguese Foundation for Science and Technology (FCT-Fundação para a Ciência e a Tecnologia) references UIDB/00127/2020, UIDB/04106/2020 and UIDP/04106/2020.

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

1. Du Y, Zhou X. Targeting non-B-form DNA in living cells. Chem Rec 2013;13:371—84.
2. Bacolla A, Wells RD. Non-B DNA conformations, genomic rearrangements, and human disease. J Biol Chem 2004;279:47411—4.
3. Bowater RP, Bohálová N, Brázda V. Interaction of proteins with inverted repeats and cruciform structures in nucleic acids. Int J Mol Sci 2022;23:6171.
4. Cer RZ, Bruce KH, Mudunuri US, Yi M, Volfovsky N, Luke BT, et al. Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. Nucleic Acids Res 2010;39(suppl 1):D383—91.

5. Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starner NJ, et al. Non-B DB v2. 0: a database of predicted non-B DNA-forming motifs and its associated tools. Nucleic Acids Res 2012;41:D94—100.
6. Guiblet WM, Cremona MA, Harris RS, Chen D, Eckert KA, Chiaromonte F, et al. Non-B DNA: a major contributor to small-and large-scale variation in nucleotide substitution frequencies across the genome. Nucleic Acids Res 2021;49:1497—516.
7. Brázda V, Laister RC, Jagelská EB, Arrowsmith C. Cruciform structures are a common DNA feature important for regulating biological processes. BMC Mol Biol 2011;12:1—16.
8. Brázda V, Kolomazník J, Lỳsek J, Hároníková L, Coufal J, Šťastnỳ J. Palindrome analyser—a new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. Biochem Biophys Res Commun 2016;478:1739—45.
9. Bastos CAC, Afreixo V, Rodrigues JMOS, Pinho AJ. Detection and characterization of local inverted repeats regularities. In: Fdez-Riverola F, Rocha M, Mohamad MS, Zaki N, Castellanos-Garzón JA, editors. Practical applications of computational biology and bioinformatics, 13th international conference. Cham: Springer International Publishing; 2020:113—20 pp.
10. Tavares AH, Pinho AJ, Silva RM, Rodrigues JM, Bastos CA, Ferreira PJ, et al. DNA word analysis based on the distribution of the distances between symmetric words. Sci Rep 2017;7:728.
11. Kent W, Sugnet C, Furey T, Roskin K, Pringle T, Zahler A, et al. The human genome browser at UCSC. Genome Res 2002;12:996—1006.
12. Smit AFA, Hubley R, Green P. RepeatMasker open-4.0; 2013—2015. Available from: http://www.repeatmasker.org.
13. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999;27:573.
14. Bastos CAC, Afreixo V, Rodrigues JMOS, Pinho AJ. Genomic regions with atypical concentration of inverted repeats. In: Fdez-Riverola F, Rocha M, Mohamad MS, Caraiman S, Gil-González AB, editors. Practical applications of computational biology and bioinformatics, 16th international conference (PACBB 2022). Cham: Springer International Publishing; 2023:89—99 pp.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/jib-2022-0052).