Jurnal Teknologi

Full paper

The Breakdown of Symmetry in Word Pairs in 1,092 Human Genomes

Vera Afreixo[a*], Sara P. Garcia[b], João M. O. S. Rodrigues[b]

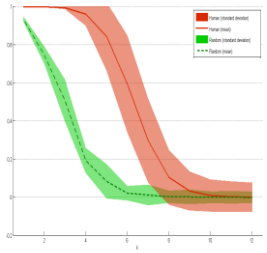
[a]Department of Mathematics & CIDMA, University of Aveiro, 3810-193 Aveiro, Portugal
[b]Department of Electronics, Telecommunications and Informatics & Signal Processing Lab, IEETA, University of Aveiro, 3810-193 Aveiro, Portugal

*Corresponding author: vera@ua.pt

Article history

Received :11 September 2012
Received in revised form :
21 February 2013
Accepted :15 April 2013

Graphical abstract



Abstract

Single strand symmetry has been observed in several genomes, and some authors have associated this phenomenon to genome evolution. However, it is still not clear how strong and exceptional this phenomenon is. We use next-generation sequencing data from a sample of 1,092 human individuals made available by the 1000 Genomes Project. To evaluate the phenomenon of symmetry of single-strand human genomic DNA, we explore and analyze these 1,092 human genomes and 1,092 randomly generated sequences, each forced to mimic the nucleotide frequency distribution of their real counterpart. Our methodology is based on measurements, traditional and equivalence statistical tests using different parameters. By statistical testing we find that the global symmetries phenomenon is significant for word lengths smaller than 8. When we evaluate the global symmetry scores, we obtain strong values for all word lengths and both types of sequences under study. However, the symmetry scores in human genomes reach higher values and have lower dispersion than those in random sequences. We also find that human and random symmetry scores are significantly different. We conclude that in the human genome, the differences between symmetric words are higher than in random sequences, but the correlation between symmetric words in human genomes is higher.

Keywords: Human genomes; single strand symmetry; equivalence testing; symmetry score; 1000 genomes project

1.0 INTRODUCTION

Chargaff's second parity rule asserts that the percentage of complementary nucleotides should be similar in each of the two strands of a DNA sequence [11, 5, 12]. Different authors have described similarities between the frequencies of words and their inverted complements (which we call symmetry phenomenon) even for longer word lengths (e.g. [10, 4, 3, 7, 14]). However, to the best of our knowledge, no previous work used genomes of several individuals from the same species to characterize the significance of this symmetry phenomenon within the species. Here, we explore and characterize the significance of the symmetry phenomenon in the human genome using data from multiple genomes made available by the 1,000 Genomes Project [2], the first to sequence the genomes of a large number of individuals.

We present novel methodologies to explore similarities between symmetric words using sequencing data obtained with next-generation methodologies, and explore the symmetry phenomenon in word lengths of 1 to 12 nucleotides. Furthermore, to add further support to our findings, we compare results obtained for each human genome to a simulated genome that mimics the nucleotide distribution of the former.

2.0 MATERIALS AND EXPERIMENTALS PROCEDURES

We evaluate the symmetry phenomenon using word frequency counts in human genomes and random DNA sequences. Words are interchangeably called k-mers. We study word lengths $\in \{1, 2, \dots, 12\}$. Our sample has $n = 1092$ human genomes and their corresponding 1092 random sequences counterpart. For each individual, all words of length k were counted in both the real and simulated genomes. For each word length, the word (w) and its corresponding symmetric word (w') counts are paired to obtain symmetric pair counts $(N_w ; N_{w'})$.

Note that, the number of distinct k-mers is 4^k. For $\in \{1, 2, \dots, n\}$, N_w^i is the number of times the word w appears in the genome sequence of individual i and

$$\sum_w N_w^i = \sum_{w'} N_{w'}^i \equiv S^i.$$

The corresponding relative frequencies are represented by $f_w^i = \frac{N_w^i}{S^i}$ and $f_{w'}^i = \frac{N_{w'}^i}{S^i}$.

2.1 Materials

We use the GRCh37.1 reference human genome assembly [2] and version 3 (March 16, 2012) of a Phase 1 integrated variant call set based on both low coverage and exome whole genome sequencing data from 1,092 individuals [1]. The VCF files contain the alterations necessary to incorporate in the reference human genome in order to obtain a different, individual human genome. We developed a package of custom-made C programs to generate alternate FASTA genomes from population sequencing VCF data, and to count occurrences of words from these individual genomes. Our data processing pipeline is schematically represented in Figure 1.

The 1000 genomes project provides next-generation sequencing data for a sample of 1,092 individuals. Table 1 shows chromosome length statistics for this sample. Note that, in this study, we have not considered chromosome Y, as not all individuals sequenced are males.

We also simulated a random genomes sample with the same sample size of human genomes sample. For each human genome and for each chromosome, we simulated a random sequence with the same length and the same nucleotide distribution as the real chromosome. Interruptions (bursts of N symbols) were also simulated with the same probability as that found on the real chromosome. We call this sample the random sequences sample.

Table 1 Mean and standard deviation of chromosome lengths (in base pairs), in 1,092 individual human genomes

Chromosome	Mean	Standard deviation
1	225,063,092	59,317
2	238,034,379	47,799
3	194,663,337	34,075
4	187,364,675	72,534
5	177,560,300	43,132
6	167,226,737	47,983
7	155,218,216	54,607
8	142,793,178	37,803
9	120,065,958	48,495
10	131,249,905	22,593
11	131,005,311	60,435
12	130,387,034	29,357
13	95,513,403	31,000
14	88,197,641	74,202
15	81,623,244	40,677
16	78,836,816	19,823
17	77,738,095	44,005
18	74,614,983	21,612
19	55,698,510	81,773
20	59,448,802	12,640
21	35,088,565	13,938
22	34,800,723	62,786
X	151,020,669	268,971
Sum	2,8E+09	410,718

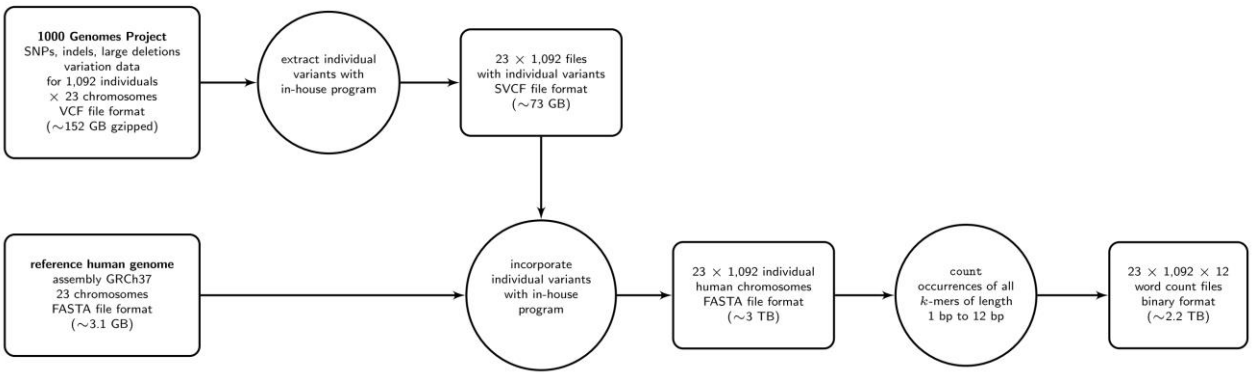


Figure 1 Data processing pipeline. A package of custom-made C programs was developed for efficiently generating the alternate FASTA genomes from population sequencing VCF data, by incorporating individual variation data into the reference human genome, hence generating a surrogate individual human genome

2.2 Statistical Hypothesis Testing

Traditional statistical hypothesis testing may be used to assess differences. However, it is well known that when traditional hypothesis tests are applied to large data sets, any small effect is always deemed significant [9, 6, 8]. Furthermore, we want to evaluate if there are similarities, not differences, between the occurrences of symmetric words. To overcome this drawback, we use equivalence tests for testing the equivalence between symmetric words. Let μ_{R_w} denote the (population) mean of the ratio of the w word frequency and its corresponding reversed complement word frequency (ratio of the frequency of the symmetric pair). We studied the equivalence between pairs of symmetric words (w, w') using the ratio of the frequency of the symmetric pair R_w and a practical tolerance δ (> 1), and concluding the equivalence when $\frac{1}{\delta} < \mu_{R_w} < \delta$. Let $\overline{R_w}$ denote the corresponding sample mean and for each individual i the ratio is given by $R_w^i = \frac{f_w^i}{f_{w'}^i}$.

The statistical hypotheses for the equivalence test are:

$$H_{0_w}: \mu_{R_w} \geq \delta \quad or \quad \mu_{R_w} \leq \frac{1}{\delta} \quad vs \quad H_{1_w}: \frac{1}{\delta} < \mu_{R_w} < \delta$$

The ratio between two frequencies, $r_w^i = \frac{f_w^i}{f_{w'}^i}$, is an effect size measure. As in many studies, e.g. [13], we consider the effect to be weak when it assumes values between 1.1 and 1.3 and we explore these lower effect size values as a tolerance to conclude practical equivalence. When the sample size is high, by the central limit theorem, we use the z interval for the unknown true value of μ_{R_w}, which is,

$$\left(\overline{R_w} \mp \mathrm{z} * \mathrm{SE}(R_w)\right)$$

Where $\mathrm{SE}(R_w) = S_{R_w}/\sqrt{n}$ with S_{R_w} denoting the sample standard deviation of R_w.

In this case, the equivalence tests procedure consists of obtaining the confidence interval for the parameter and checking if it is contained in the interval $(1/\delta, \delta)$. If so, H_{0_w} is rejected and for the (w, w') pair, the equivalence can be assumed.

For each type of data and for each word length k, we construct 4^k equivalence tests. When we reject all of the 4^k null hypotheses, we consider that the symmetry phenomenon is present, as all symmetric pairs are equivalent in a global way.

For testing the non equivalence between symmetric words we use the following statistical hypotheses:

$$H_{0_w}: \frac{1}{\delta} \leq \mu_{R_w} \leq \delta \quad vs \quad H_{1_w}: \mu_{R_w} > \delta \quad or \quad \mu_{R_w} < \frac{1}{\delta}$$

For both types of data and for each word length k, we construct 4^k non equivalence tests. When we reject one of the 4^k null hypotheses, we consider that the global symmetry phenomenon is not present.

Since we conduct simultaneous tests, we apply the Bonferroni correction.

2.3 Symmetry Score

We use Pearson's correlation to measure the global agreement between symmetric words in each individual. In particular, we use the coefficient as a score of symmetry in each individual

$$SS_i = \frac{\sum_w \left(N_w^i - \overline{N^i}\right)\left(N_{w'}^i - \overline{N^i}\right)}{\sum_w \left(N_w^i - \overline{N^i}\right)^2}, \qquad (1)$$

$i \in \{1,2,\ldots,n\}$.

2.4 Correlation

To evaluate the significance of correlation between a pair of symmetric words, we apply the two tailed Pearson correlation test. Considering ρ the Pearson correlation parameter, the tests hypotheses are:

$$H_0: \rho = 0 \; vs \; H_1: \rho \neq 0,$$

with $T = c\sqrt{\frac{n-2}{1-c^2}} \sim_{under\ H_0} t_{\{n-2\}}$ and c the sample Pearson correlation coefficient.

If all genomes have the same size and verify Chargaff's second parity rule we expect a null correlation for each symmetric pair,. However, for different genomes sizes, we expect a positive correlation.

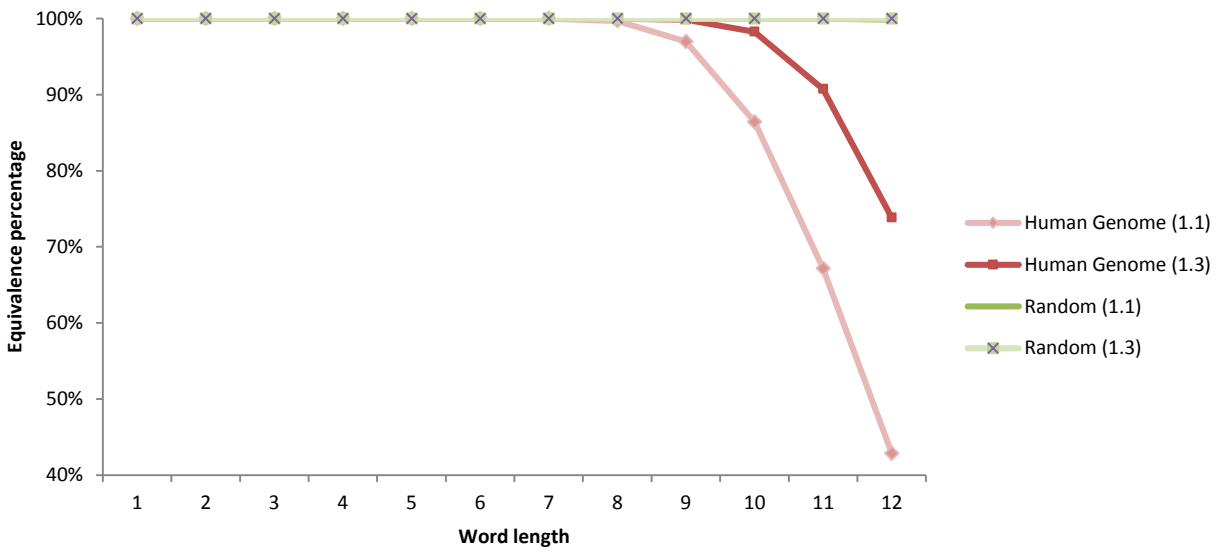


Figure 2 Percentage of equivalence tests that reject the null hypothesis for different word lengths

3.0 RESULTS AND DISCUSSION

3.1 Statistical Hypothesis Testing

Figure 2 displays the percentage of equivalent pairs (in the sense of what has been described previously) for each k-mer length and each tolerance value (δ). We verify equivalence between symmetric pairs for $k \leq 8$ for both tolerance values $\delta = 1.1$ and $\delta = 1.3$. For $k > 8$, we identify some pairs where the non equivalence is significant using non equivalence tests (see Figure 3).

A random sequence generated subject only to Chargaff´s second parity rule should exhibit the symmetry phenomenon (similar frequencies for all symmetric pairs), regardless of the word length. This is confirmed in all equivalence and non equivalence tests, as shown in Figures 2 and 3.

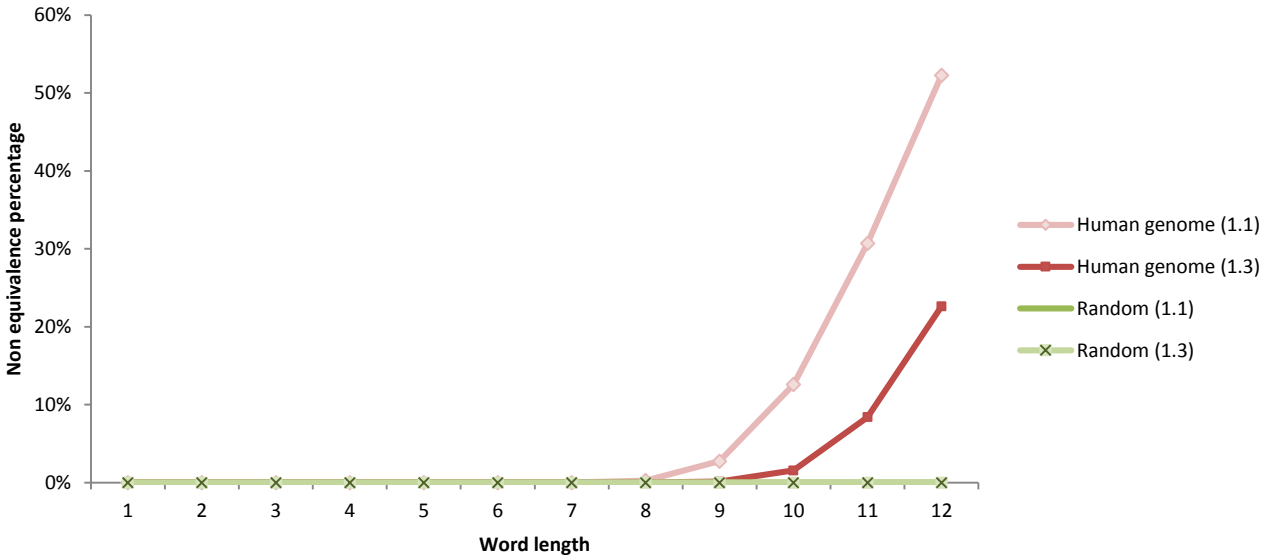


Figure 3 Percentage of non equivalence tests that reject the null hypothesis for different word lengths

3.2 Score of Symmetry

Figure 4 displays an error bar plot of the global scores of symmetry (SS, equation 1). We observe high score values (close to 1) for all word lengths $k \in \{1, 2, \ldots, 12\}$ and for all sequence types. However, this score has a tendency to decrease as the word length increases. Note that, though all global scores of symmetry have high values, these might be attributable to the contribution of a few outliers. In this figure, we observe a high association between k and the scores (approximately parabolic behavior, concavity down with inflection point in k = 4).

The global symmetry score has higher values in the human genome than in random sequences, and the random results have higher dispersion than in human sequences.

Table 2 presents the results of two-tailed t tests for mean differences. As expected, for nucleotides the difference is not significant, but for the other k-mers, we obtain significant differences between the symmetry scores of human and random sequences.

Table 2 Pair sample test results for the mean differences of symmetry scores between human genomes and correspondent random sequence

Word length	Mean	p-value
1	0,0000001	0,255
2	0,0000074	0,000
3	0,0000093	0,000
4	0,0000099	0,000
5	0,0000108	0,000
6	0,0000136	0,000
7	0,0000247	0,000
8	0,0000684	0,000
9	0,0002405	0,000
10	0,0008907	0,000
11	0,0032779	0,000
12	0,0118712	0,000

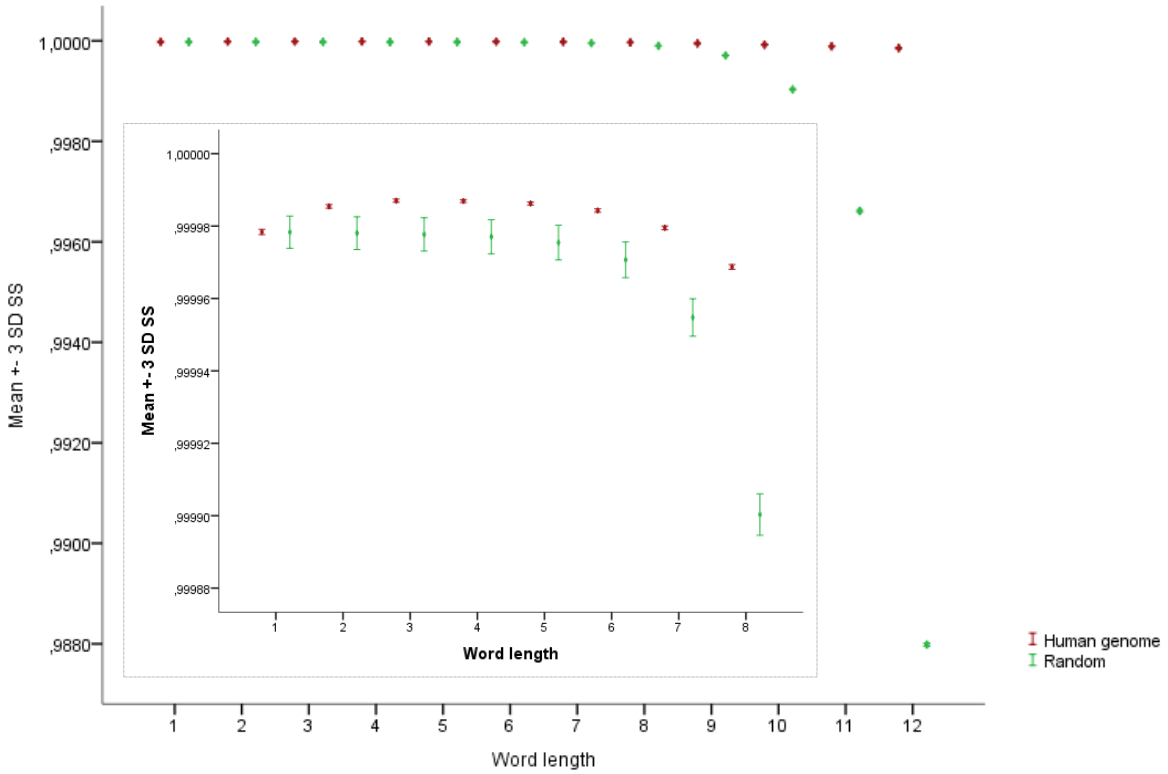


Figure 4 Error bar of the scores of symmetry (SS) in 1,092 human genomes and their corresponding random sequences. The inset plot highlights the 8 shorter word lengths

3.3 Association by Symmetric Word Pair

We study the association behavior of symmetric word pairs by analyzing the correlation coefficient between symmetric pair counts (N_w, N'_w). For each word length k, there are 4^k pairs of symmetric words. However, some words are their own inverted complement $(w = w')$, and so the symmetric pair counts are necessarily equal $(N_w = N'_w)$. To avoid any bias, we exclude such self-symmetric words (SSW) from this study. (Note that there are 2^k SSW when k is even and none when k is odd.)

The correlation coefficient and the corresponding statistical test p-value are obtained for each symmetric word pair based on a sample of 1,092 individuals. Table 3 displays the frequency table of the correlation coefficients, highlighting the corresponding conclusion of t correlation tests. Figure 5 shows statistics of the correlation coefficients between pairs of

symmetric words in 1,092 human genomes (red) and their corresponding random sequences (green).

For both types of data, as expected, we observe significant positive correlation and in some cases non significant correlation (see Table 3). However, the dispersion of correlations in human genomes is lower than in random sequences, and for k>8 we observe some significantly negative correlations.

Table 3 Percentage of word pairs (excluding SSWs) with correlation coefficients in each class of effect size. *the p-value of one tailed Pearson correlation test is <0.05

Human Genome

Correlation	*k=1*	*k=2*	*k=3*	*k=4*	*k=5*	*k=6*	*k=7*	*k=8*	*k=9*	*k=10*	*k=11*	*k=12*
[-1; -0.50)*	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0,00**	**0,00**
[-0.50; -0.30)*	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0,04**	**0,04**	**0,06**	**0,07**	**0,08**
[-0.30; -0.10)*	0	0	0	0	0	0,10	1,48	3,14	5,80	7,70	8,28	8,03
[-0.10; -0.05)*	0	0	0	0	0	0,40	2,33	5,34	10,12	12,77	13,70	13,59
[-0.05; 0.05)	0	0	0	0	0,20	3,57	11,84	29,24	46,97	52,44	53,76	55,85
[0.05; 0.10)*	0	0	0	0	0	3,17	6,02	17,81	18,48	15,75	14,62	13,90
[0.10; 0.30)*	0	0	0	0	3,13	10,47	29,08	36,93	17,06	10,89	9,33	8,43
[0.30; 0.50)*	0	0	0	0	5,27	7,94	32,59	5,68	1,11	0,29	0,19	0,11
[0.50; 1]*	100	100	100	100	91,41	74,36	16,66	1,81	0,42	0,09	0,05	0,01

Random

Correlation	*k=1*	*k=2*	*k=3*	*k=4*	*k=5*	*k=6*	*k=7*	*k=8*	*k=9*	*k=10*	*k=11*	*k=12*
[-1; -0.50)*	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
[-0.50; -0.30)*	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
[-0.30; -0.10)*	0	0	0	0	0	0	0,04	0,04	0,04	0,05	0,05	0,05
[-0.10; -0.05)*	0	0	0	0	0	1,64	3,60	4,52	4,58	4,79	4,85	4,87
[-0.05; 0.05)	0	0	0	0	40,82	78,52	86,55	88,22	89,36	89,70	89,92	90,03
[0.05; 0.10)*	0	0	0	5	38,09	16,87	6,56	6,66	5,18	5,29	4,94	4,98
[0.10; 0.30)*	0	0	3	90	16,21	2,98	2,56	0,56	0,67	0,16	0,20	0,07
[0.30; 0.50)*	0	0	50	5	4,69	0	0,68	0	0,17	0	0,04	0
[0.50; 1]*	100	100	47	0	0,20	0	0,01	0	0,00	0	0,00	0

In Figure 5, for both sequence types, we observe a curious tendency: as k increases, the mean of the correlation tends to zero. Moreover, for $k \leq 8$, the human genome results have higher association (positive correlation) values than their random counterparts. For $k > 8$, both types of data have means of correlation within the range (-0.05;0.05).

For $k = 5$, the symmetric pair (CGTTA, TAACG) is the single pair responsible for the hypothesis test not rejecting the null hypothesis. Moreover, there are 8.6% of pairs where the correlation is not strong (Table 3, $k = 5$). For $k > 5$, there are many more pairs responsible for the non rejection of the null hypothesis. The percentage of not strongly correlated pairs also increases. The left panel of Figure 6 shows a scatter plot for the symmetric word pair (CGTTA, TAACG), which does not present significant positive correlation.

However, for all $\in \{1, 2, \ldots, 12\}$, there are several pairs of symmetric words where the correlation is significant and strong. An example is displayed in the right panel of Figure 6, for the symmetric word pair (AAAAAAA, TTTTTTT) with positive correlation.

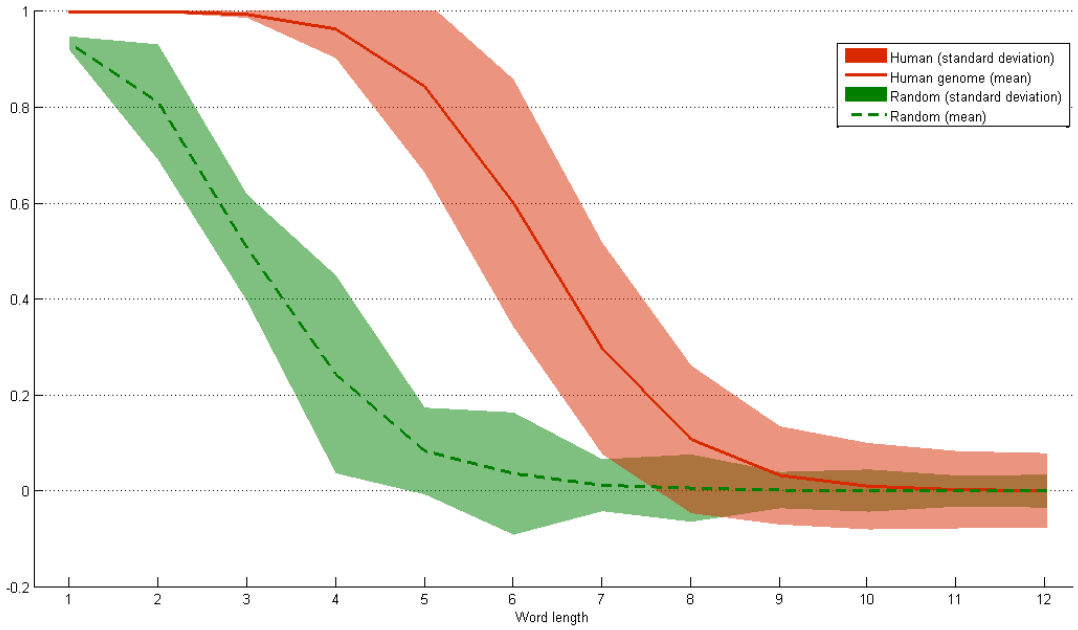


Figure 5 Summary of statistics of the correlation coefficients between pairs of symmetry words in 1,092 human genomes (red) and their corresponding random sequences (green). The line represents the mean and the shaded region represents the standard deviation around the mean (mean $\pm$ standard deviation)

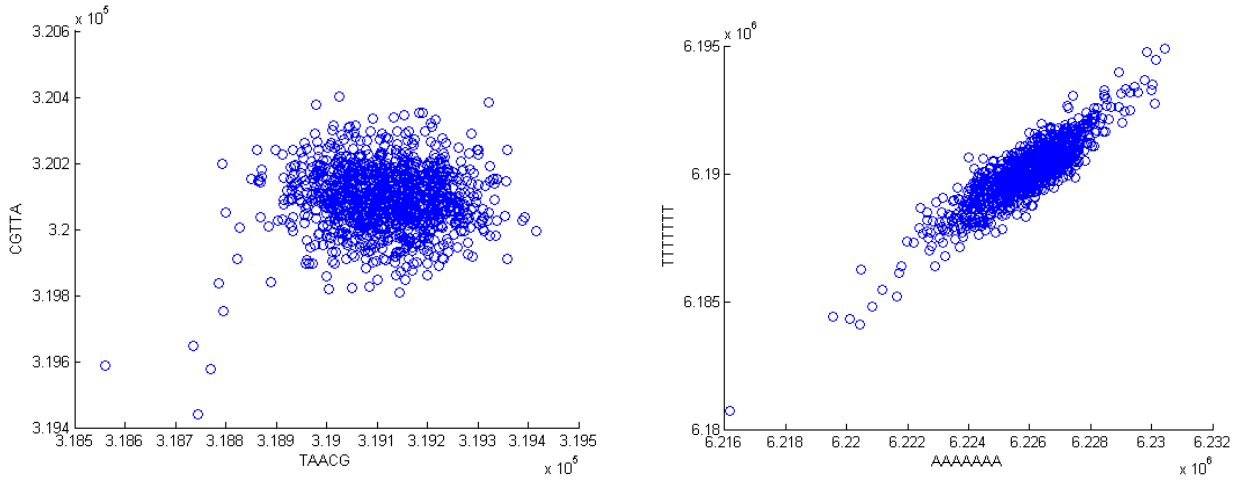


Figure 6 Left: Scatter plot of the frequencies of the (CGTTA, TAACG) symmetric pair, with $r = -0.008$ and p-value 0.790
Right: Scatter plot of the frequencies of the (AAAAAAA, TTTTTTTT) symmetric pair, with $r = 0{,}870$ and p-value <0.001

4.0 CONCLUSIONS

Here, we studied the word symmetry phenomenon, characterized through word frequencies, in 1,092 human genomes. We confirmed the global tendency of the symmetry phenomenon using equivalence tests and a global score of symmetry. We identified an interval of word lengths where the global symmetry phenomenon tendency starts to be non significant. Whereas the global score of symmetry has high values for all word lengths investigated, the equivalence tests show a breakdown of symmetry for $k > 8$.

In the human genome we identified several words with unexpected association with their corresponding inverted complement. Moreover, several symmetric pairs have significantly strong negative correlation.

We conclude that the symmetry phenomenon is less prevalent in human genomes than previously thought. It will be interesting to investigate this symmetry phenomenon for selected genomic regions.

Acknowledgements

This work was supported by FEDER funds through COMPETE-Operational Programme Factors of Competitiveness and by Portuguese funds through the Center for Research and Development in Mathematics and Applications (University of Aveiro) and the Portuguese Foundation for Science and Technology within project PEst-C/MAT/UI4106/2011 with COMPETE number FCOMP-01-0124-FEDER-022690.

References

[1] The 1000 genomes project data release: Integrated variant call set for phase 1, version 3. http://www.1000genomes.org/

[2] Grch37 Reference human genome assembly. ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/

[3] Guenter Albrecht-Buehler. 2007. Inversions and Inverted Transpositions as the Basis for an Almost Universal "Format" of Genome Sequences. *Genomics*. 90: 297–305.

[4] Pierre-François Baisnée, Steve Hampson, and Pierre Baldi. 2002. Why are Complementary DNA Strands Symmetric? *Bioinformatics*. 18(8): 1021–1033.

[5] J. D. Karkas, R. Rudner, and E. Chargaff. 1968. Separation of B. Subtilis DNA into Complementary Strands. II. Template Functions and Composition as Determined by Transcription with RNA Polymerase. Proceedings of the National Academy of Sciences of the United States of America. 60(3): 915–920.

[6] R.B. Kline. 2004. *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. Washington, DC: American Psychological Association.

[7] Sing-Guan Kong, Wen-Lang Fan, Hong-Da Chen, Zi-Ting Hsu, Nengji Zhou, Bo Zheng, and Hoong-Chien Lee. 2009. Inverse Symmetry in Complete Genomes and Whole-genome Inverse Duplication. *PLoS ONE*. 4(11): e7553.

[8] Sonia Migliorati and Andrea Ongaro. 2010. Adjusting p-values when n is Large in the Presence of Nuisance Parameters. In Statistics for Industry and Technology. 305–318.

[9] D. S. Moore. 1997. *Statistics: Concepts and Controversies*. 4th edition. New York: WH Freeman & Co.

[10] Dong Qi and A. Jamie Cuticchia. 2001. Compositional Symmetries in Complete Genomes. *Bioinformatics*. 17(6): 557–559.

[11] R. Rudner, J. D. Karkas, and E. Chargaff. 1968. Separation of B. Subtilis DNA into Complementary Strands, I. Biological Properties. Proceedings of the National Academy of Sciences of the United States of America. 60(2): 630–635.

[12] R. Rudner, J. D. Karkas, and E. Chargaff. 1968. Separation of B. Subtilis DNA into Complementary Strands. III. Direct Analysis. Proceedings of the National Academy of Sciences of the United States of America. 60(3): 921–922.

[13] George Thanassoulis and Ramachandran S. Vasan. 2010. Genetic Cardiovascular Risk Prediction—Will We Get There? *Circulation*. 122(22): 23232334.

[14] Shang-Hong Zhang and Ya-Zhi Huang. 2010. Limited Contribution of Stem-loop Potential to Symmetry of Single-stranded Genomic DNA. *Bioinformatics*. 26(4): 478–485.