Acta Scientific MICROBIOLOGY (ISSN: 2581-3226)

Volume 4 Issue 4 April 2021

Research Article

Significant Differences in Nucleotide and Peptide Features Between Chromosomes Suggesting Sequence Non-randomness Across Chromosomes

Gabriel ZK Lim[1,2], Hykal H Azmi[1,2], Mariia Dolmatova[1,2] and Maurice HT Ling[1,2,3,4]*

[1]*Department of Applied Sciences, Northumbria University, United Kingdom*

[2]*School of Life Sciences, Management Development Institute of Singapore, Singapore*

[3]*HOHY PTE LTD, Singapore*

[4]*School of Data Sciences, Perdana University, Malaysia*

***Corresponding Author:** Maurice HT Ling, Department of Applied Sciences, Northumbria University, United Kingdom.

Received: February 20, 2021

Published: March 16, 2021

Abstract

Eukaryotic genomes are organized into multiple chromosomes. Several studies have suggested that chromosomes are organized functionally and spatially, indicative of selective pressure in chromosomal organization. However, question remains as to whether chromosomes of the same organism are significantly different based on nucleotide and peptide features. Here, we examine three eukaryotic species across kingdoms; animalia (*Sarcophilus harrisii*), plantae (*Prunus dulcis*), and SAR supergroup (*Plasmodium falciparum*); to identify whether chromosomes of the same organism are significantly different based on nucleotide and peptide features. Our results show that the average GC contents in coding sequences are significantly different (p-value ≤ 3.30E-09) to their chromosomal GC content in all 30 chromosomes across 3 organisms. Our results also show that 38 out of 45 (15 features by 3 organisms) the nucleotide and peptide features are significantly different (p-value ≤ 0.044) between chromosomes. These results imply the presence of selective pressure in chromosomal organization.

Keywords: Chromosomal Non-randomness; Chromosomal Organization; Selective Pressure; *Sarcophilus harrisii*; *Prunus dulcis*; *Plasmodium falciparum*

Introduction

Eukaryotic genomes are organized into linear chromosomes. Due to chromosomal rearrangements, gene replication and mutations, the length of the longest chromosome arm of a given karyotype, becomes affected by reciprocal translocation or chromosomal fusion [1]. The shape and size of chromosomes can be altered by reciprocal translocation, which exchanges unequal parts between the chromosomes involved by loss of dispensable parts e.g. deletion, by insertion e.g. transposition, or by sequence amplification or loss: (i) via unequal sister chromatid exchange e.g. mitotically, (ii) via unequal crossover e.g. meiotically, or (iii) via replication slipping when microscopically detectable amounts of chromatin are involved. Except for replication slipping, these alterations represent primary chromosome rearrangements, reflecting mis-repair of DNA damage, in particular by non-homologous end-joining of double-strand breaks [2].

Ferrai., *et al.* [3] suggest that gene position within the genome is evolutionarily conserved and plays a functional role in genome activity. Functional organization of chromosomes includes different types of chromatin and reflects the average folded state of the chromosome [4]. Eukaryotic genomes are spatially organized within the nucleus by chromosome folding, interchromosomal contacts, and interaction with nuclear structures. This phenomenon was observed in various organisms and contributes well to gene expression and differentiation [5]. Sáez-Vásquez and Gadal [6] discovered that chromosomes were able to achieve an efficient and coordinated gene expression through the formation of specialized structures such as nuclear bodies, where genes, RNA, and protein factors are clustered together; hence, indicating a preferential organization among chromosomes. This preferential spatial organization is an active by-product of emergent properties and mechanisms of nuclear biology, which is determined by the relationship between DNA-binding proteins and chromatin changes. Thus, the chromosomal organization gives rise to differences in spatial arrangement among each chromosome; however, the question remains as to whether chromosomes of the same organism are significantly different based on nucleotide and peptide features.

In this study, we examine whether chromosomes of the same organism are significantly different based on nucleotide and peptide features by studying three eukaryotic species across kingdoms – animalia [*Sarcophilus harrisii*, commonly known as Tasmanian devil [7]], plantae (*Prunus dulcis*, commonly known as almond [8]], and SAR supergroup (*Plasmodium falciparum*, a human protozoan parasite that causes malaria [9,10]]. Our findings suggest that a majority of the nucleotide and peptide features are significantly different (p-value ≤ 0.044) between chromosomes, implying the presence of selective pressure in chromosomal organization.

Materials and Methods

Nucleotide and peptide sequences, representing the coding sequences, of each chromosome in the three species were obtained from NCBI; namely, *S. harrisii* assembly mSarHar1.11 [GenBank assembly accession GCA_902635505.1 [7]], *P. falciparum* 3D7 (GenBank assembly accession GCA_000002765.3 [9,10]], and *P. dulcis* assembly ALMONDv2 (GenBank assembly accession GCA_902201215.1 [8]]. Average nucleotide compositions (CDS length, %GC, %A, %T, %G, and %C) and peptide properties (molecular weight, isoelectric point, aromaticity, hydropathy, secondary structure proportions, and molar coefficient extinction) for each coding sequence were determined using Biopython [11] via SeqProperties [12]. One-way ANOVA was used to evaluate each nucleotide and peptide feature under the null hypothesis of no difference in average feature across the chromosomes where a probability of more than 5% was considered insignificant. Differences in average features between any two chromosomes were determined using 2-samples t-test assuming heteroscedasticity.

Results and Discussion

Nucleotide and peptide properties of *P. dulcis*, *S. harrisii* and *P. falciparum* were examined. Our results show that the gene density (= Average CDS Length x Number of Genes/Chromosomal Length) is significantly different (F = 489.05, p-value = 6.22E-22) across the three species and between any two species (T > 6.36, p-value < 3.82E-4), with *S. harrisii* and *P. falciparum* having the lowest and highest gene density respectively (Table 1). This is consistent with Gardner., *et al*. [10] showing higher coding density in *P. falciparum* compared to plants. In addition, %GC in CDSes are consistently (between 4.9% to 16%) and significantly higher (p-value ≤ 3.30E-09) in coding sequences compared to the entire chromosome, which is supported by previous studies [13-15]. However, there is a significant correlation between chromosomal %GC and average %GC of CDS (R-square = 0.938, F = 420.81, p-value = 2.08E-18).

By evaluating whether the average features are equal across the chromosomes within the same organism, our results (Table 2) suggest that the 8 chromosomes of *S. harrisii* are unequal in all features (F ≥ 6.21, p-value ≤ 2.70E-07) whereas the chromosomes of *P. dulcis* are unequal in all features (F ≥ 2.06, p-value ≤ 0.044) except isoelectric point (F = 1.01, p-value = 0.423). On the other hand, the chromosomes of *P. falciparum* are unequal (F ≥ 1.90, p-value ≤ 0.025) in 9 of the 15 evaluated features. A possible reason for more randomness in the chromosomes of *P. falciparum* is that most of the proteins of *P. falciparum* are dominated by low-complexity elements as the number of low-complexity elements is significantly higher in *P. falciparum* compared to other eukaryotes [16,17].

Nevertheless, these results suggest that chromosomes of the same organism are significantly different based on nucleotide and peptide features; thus, nucleotide and peptide features are not random across chromosomes of the same organism. Non-randomness may suggest the presence of selective pressure in chromosomal or-

Organism	Chromosome	Accession Number	Number of Genes	Chromosome Length (megabases)	Average CDS Length (bases)	Gene Density (%)	Chromosomal %GC	Average %GC from CDS	p-value (Null: Average %GC from CDS = Chromosomal %GC)
P. dulcis (Almond)	1	NC_047650.1	5478	44.00	1511.2	18.8	37.7	45.5	1.42E-76
	2	NC_047651.1	3097	26.14	1534.2	18.2	37.8	45.1	3.11E-70
	3	NC_047652.1	2906	24.07	1508.2	18.2	37.8	45.1	3.05E-70
	4	NC_047653.1	2862	24.38	1458.0	17.1	37.9	44.0	1.34E-62
	5	NC_047654.1	2404	18.23	1480.3	19.5	37.7	45.5	6.82E-79
	6	NC_047655.1	3571	29.60	1471.6	17.8	37.7	45.4	1.18E-67
	7	NC_047656.1	2597	21.34	1479.3	18.0	37.8	45.3	2.25E-68
	8	NC_047657.1	2631	20.43	1442.6	18.6	37.7	45.4	1.74E-72
S. harrisii (Tasmanian Devil)	1	NC_045426.1	5422	716.41	2227.0	1.69	35.7	48.8	<1E-240
	2	NC_045427.1	5683	662.75	2132.9	1.83	36.3	48.3	<1E-240
	3	NC_045428.1	5160	611.35	2194.1	1.85	35.6	48.4	<1E-240
	4	NC_045429.1	4674	464.9	2152.6	2.16	36.4	50.0	<1E-240
	5	NC_045430.1	2526	288.12	2024.6	1.78	36.4	48.5	<1E-240
	6	NC_045431.1	2487	254.9	2108.9	2.06	36.3	49.2	<1E-240
	X	NC_045432.1	843	83.08	1892.4	1.92	40.5	52.1	<1E-240
	Y	NC_045433.1	6	0.13	4238.3	19.6	33.7	44.9	3.30E-09
P. falciparium (Malaria Parasite)	1	NC_004325.2	158	0.64	2007.6	49.6	20.5	36.5	<1E-240
	2	NC_037280.1	234	0.95	2120.5	52.2	19.7	25.6	<1E-240
	3	NC_000521.4	249	1.07	2312.6	53.8	20.1	25.1	<1E-240
	4	NC_004318.2	261	1.20	2640.6	57.4	20.5	25.5	<1E-240
	5	NC_004326.2	332	1.34	2326.7	57.6	19.3	25.1	<1E-240
	6	NC_004327.3	332	1.42	2386.1	55.8	19.8	24.7	<1E-240
	7	NC_004328.3	331	1.45	2704.1	61.7	19.8	25.2	<1E-240
	8	NC_004329.3	341	1.47	2345.2	54.4	19.6	25.2	<1E-240
	9	NC_004330.2	382	1.54	2036.6	50.5	19.0	25.0	<1E-240
	10	NC_037281.1	408	1.69	2125.3	51.3	19.6	25.4	<1E-240
	11	NC_037282.1	503	2.04	2240.0	55.2	19.0	24.9	<1E-240
	12	NC_037284.1	552	2.27	2304.0	56.0	19.3	24.8	<1E-240
	13	NC_004331.3	730	2.93	2270.1	56.6	19.0	24.7	<1E-240
	14	NC_037283.1	802	3.29	2344.3	57.1	18.4	24.6	<1E-240

Table 1: Summary of Chromosomes.

ganization [18-24]. However, the role of selective pressure in shaping nucleotide and peptide features remains unknown.

It has been known that spatial organization within genome is not random [3,6,25]. However, it is not known whether chromosomes of the same organism are significantly different based on

Feature Type	Feature	*P. dulcis* (Almond)	*S. harrisii* (Tasmanian Devil)	*P. falciparum* (Malaria Parasite)
Length	CDS Length	F = 3.12 P = 0.003	F = 8.04 P = 8.53E-10	F = 1.57 P = 0.085 NS
Nucleotide Features	%GC	F = 10.33 P = 5.27E-13	F = 62.86 P = 1.69E-90	F = 82.85 P = 1.18E-200
	%A	F = 2.58 P = 0.012	F = 59.47 P = 1.90E-85	F = 0.83 P = 0.631 NS
	%T	F = 6.77 P = 4.74E-08	F = 52.53 P = 4.01E-75	F = 99.05 P = 9.49E-239
	%G	F = 8.46 P = 2.23E-10	F = 56.11 P = 1.94E-80	F = 2.50 P = 0.002
	%C	F = 2.06 P = 0.044	F = 53.40 P = 2.070E-76	F = 233.41 P <1E-240
Peptide Features	Molecular Weight	F = 2.98 P = 0.004	F = 8.01 P = 9.30E-10	F = 1.57 P = 0.085 NS
	Isoelectric Point (pI)	F = 1.01 P = 0.423 NS	F = 12.47 P = 4.59E-16	F = 0.47 P = 0.939 NS
	Hdropathy (GRAVY)	F = 2.72 P = 0.008	F = 49.14 P = 4.47E-70	F = 1.9 P = 0.025
	Aromaticity	F = 2.98 P = 0.004	F = 24.37 P = 2.19E-33	F = 1.63 P = 0.070 NS
	Secondary Structure: Helix Fraction	F = 4.49 P = 5.26E-05	F = 37.36 P = 1.41E-52	F = 2.86 P = 3.90E-02
	Secondary Structure: Turn Fraction	F = 2.20 P = 0.031	F = 6.27 P = 2.24E-07	F = 1.88 P = 0.027
	Secondary Structure: Sheet Fraction	F = 2.13 P = 0.037	F = 15.86 P = 5.99E-21	F = 0.87 P = 0.581 NS
	Extinction Coefficient (Reduced)	F = 3.44 P = 0.001	F = 6.21 P = 2.70E-07	F = 1.9 P = 0.025
	Extinction Coefficient (Unreduced)	F = 3.45 P = 0.001	F = 6.21 P = 2.72E-07	F = 1.94 P = 0.021

Table 2: P-values obtained from 1-Way ANOVA. Statistically non-significant features are labelled as NS. Distributions of each feature are provided in supplementary materials.

nucleotide and peptide features. Using three eukaryotic species across kingdoms, our findings suggest that a majority of the nucleotide and peptide features are significantly different (p-value ≤ 0.044) between chromosomes.

Conclusion

Our results suggest that chromosomes are not in uniform in features for *S. harrisii* as well as certain features for *P. falciparum* and *P. dulcis*. The non-randomness was consistent across certain nucleotide and peptide features in *P. falciparum* and *P. dulcis*. Finally, we can conclude that the nucleotide and peptide features are not equally distributed among all 3 eukaryotic organisms, reinforcing the presence of selective pressure in chromosomal organization.

Supplementary Materials

Supplementary tables and figures are available at http://bit.ly/ChrFeatures.

Data Availability

Sequence and result files are available as a multi-volume ZIP file at http://bit.ly/ChrFeaturesD1 (volume 1 of 3), http://bit.ly/ChrFeaturesD2 (volume 2 of 3), and http://bit.ly/ChrFeaturesD3 (volume 3 of 3).

Conflict of Interest

The authors declare no conflict of interest.

Bibliography

1. Schubert I and Oud JL. "There is an Upper Limit of Chromosome Size for Normal Development of an Organism". *Cell* 88.4 (1997): 515-520.

2. Schubert I. "Chromosome Evolution". C*urrent Opinion in Plant Biology* 10.2 (2007): 109-115.

3. Ferrai C., *et al*. "Gene Positioning". *Cold Spring Harbor Perspectives in Biology* 2.6 (2010): a000588.

4. Ridgway P., *et al*. "Functional Organization of the Genome: Chromatin". *Atlas of Genetics and Cytogenetics in Oncology and Haematology* 3 (2011).

5. Brickner J. "Genetic and Epigenetic Control of the Spatial Organization of the Genome". *Molecular Biology of the Cell* 28.3 (2017): 364-369.

6. Sáez-Vásquez J and Gadal O. "Genome Organization and Function: A View from Yeast and Arabidopsis". *Molecular Plant* 3.4

(2010): 678-690.

7. Miller W., *et al*. "Genetic Diversity and Population Structure of the Endangered Marsupial Sarcophilus harrisii (Tasmanian devil)". *Proceedings of the National Academy of Sciences of the United States of America* 108.30 (2011): 12348-12353.

8. Alioto T., *et al*. "Transposons Played a Major Role in the Diversification Between the Closely Related Almond and Peach Genomes: Results from the Almond Genome Sequence". *Plant Journal* 101.2 (2020): 455-472.

9. Böhme U., *et al*. "Progression of the Canonical Reference Malaria Parasite Genome from 2002-2019". *Wellcome Open Research* 4 (2019): 58.

10. Gardner MJ., *et al*. "Genome Sequence of the Human Malaria Parasite Plasmodium falciparum". *Nature* 419.6906 (2002): 498-511.

11. Cock PJA., *et al*. "Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics". *Bioinformatics* 25.11 (2009): 1422-1423.

12. Ling MHT. "SeqProperties: A Python Command-Line Tool for Basic Sequence Analysis". *Acta Scientific Microbiology* 3.6 (2020): 103-106.

13. Corrochano LM., *et al*. "Nucleotide Composition in Protein-Coding and Non-Coding DNA in the Zygomycete Phycomyces blakesleeanus". *Mycological Research* 108.8 (2004): 858-863.

14. Bohlin J., *et al*. "Investigations of Oligonucleotide Usage Variance Within and Between Prokaryotes". *PLOS Computational Biology* 4.4 (2008): e1000057.

15. Bohlin J., *et al*. "The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes". *BMC Genomics* 18.1 (2017): 151-151.

16. Chanda I., *et al*. "Proteome Composition in Plasmodium falciparum: Higher Usage of GC-Rich Nonsynonymous Codons in Highly Expressed Genes". *Journal of Molecular Evolution* 61.4 (2005): 513-523.

17. Singh GP., *et al*. "Hyper-Expansion of Asparagines Correlates with an Abundance of Proteins with Prion-Like Domains in Plasmodium falciparum". *Molecular and Biochemical Parasitology* 137.2 (2004): 307-319.

18. Keng BM., *et al*. "Codon Usage Bias is Evolutionarily Conserved". *Asia-Pacific Journal of Life Sciences* 7.3 (2014): 233-242.

19. Ling MHT., *et al*. "Conserved Expression of Natural Antisense Transcripts in Mammals". *BMC Genomics* 14 (2013): 243.

20. Pelletier F and Coltman DW. "Will Human Influences on Evolutionary Dynamics in the Wild Pervade the Anthropocene?" *BMC Biology* 16.1 (2018): 7.

21. Ometto L., *et al*. "Rates of Evolution in Stress-Related Genes are Associated with Habitat Preference in Two Cardamine Lineages". *BMC Evolutionary Biology* 12.1 (2012): 7.

22. De La Torre AR., *et al*. "Genome-Wide Analysis Reveals Diverged Patterns of Codon Bias, Gene Expression, and Rates of Sequence Evolution in Picea Gene Families". *Genome Biology and Evolution* 7.4 (2015): 1002-1015.

23. Bochkareva OO., *et al*. "Genome Rearrangements and Selection in Multi-Chromosome Bacteria Burkholderia Spp". *BMC Genomics* 19.1 (2018): 965.

24. Maitra A and Ling MH. "Codon Usage Bias and Peptide Properties of Pseudomonas balearica DSM 6083T". *MOJ Proteomics Bioinformation* 8.2 (2019): 27-39.

25. Diament A., *et al*. "Three-Dimensional Eukaryotic Genomic Organization is Strongly Correlated with Codon Usage Expression and Function". *Nature Communication* 5 (2014): 5876.