Guanine and Cytosine Exhibit a Consistent Linear Gradient Skew Across Each Chromosome in the Malaria Genome

Genshiro Esumi
Department of Pediatric Surgery, Hospital of the University of Occupational and Environmental Health, Kitakyushu, Japan

Abstract

It is empirically known that the numbers of thymine and adenine, as well as guanine and cytosine, are nearly equal on a single strand of each chromosome in genomes, a phenomenon referred to as Chargaff's Second Parity Rule (CSPR). An investigation of how this CSPR is achieved in the *Plasmodium* genome, which is a eukaryotic genome with an extremely low GC content, revealed that although the numbers of guanine and cytosine are nearly equal on each chromosome, their local ratio is skewed in a linear gradient depending on the position along each chromosome. This suggests local deviations from CSPR as a function of chromosome position. While CSPR has traditionally been explained by the equivalence of base substitution ratios, the consistent gradient skew of guanine and cytosine across all 14 chromosomes in malaria suggests a more complex underlying mechanism shaping CSPR.

Keywords: Chargaff's Second Parity Rule (CSPR), guanine-cytosine skew (GC skew), *plasmodium falciparum*, nucleotide composition, genomics.

Email: esumi@clnc.uoeh-u.ac.jp

Background

All living organisms encode the structure of their proteins as nucleotide sequences within their genomes. While mutations are generally considered to be random, genome analyses of various organisms have empirically shown that the numbers of thymine and adenine, as well as guanine and cytosine, on one strand of each chromosome are nearly equal to each other in almost all organisms. This phenomenon is known as Chargaff's Second Parity Rule (CSPR) [1].

Several hypotheses have been proposed regarding the underlying mechanisms of CSPR [2]. One of these hypotheses is based on the equivalence of reciprocal substitution rates in single nucleotide changes [3]. However, it remains uncertain whether this alone can fully explain CSPR. Thus, the question of how CSPR is maintained and why it is conserved has not yet been definitively explained [4].

Previous analyses by the author have suggested that in organisms adhering to CSPR, the local balance of thymine and adenine (TA skew) and guanine and cytosine (GC skew) distributes symmetrically along a single strand of the chromosome [5]. In this study, an analysis was performed to determine how this symmetry is preserved in the malaria parasite *Plasmodium*, which has a genome with one of the lowest GC contents.

Materials and Methods

The genome of *Plasmodium falciparum 3D7*, consisting of 14 chromosomes, was used in this study [6]. Nucleotide sequence data was downloaded from the NCBI website as RefSeq database data in FASTA format.

First, to confirm CSPR, the number of each nucleotide (T, A, G, C) was counted for each chromosome. Next, for each chromosome, sequences were extracted using a sliding window of 5000 base pairs (bp) with a 1 bp step over the entire length. TA skew, GC skew, and GC content were then calculated for each window.

> TA skew, GC skew, and GC content were calculated using the following formulas, where T, A, G, and C represent, respectively, the number of thymine, adenine, guanine, and cytosine nucleotides within each window:
>
> - TA skew = (T - A) / (T + A)
> - GC skew = (G - C) / (G + C)
> - GC content = (G + C) / (T + A + G + C)

The calculated TA skew, GC skew, and GC content were plotted with the Y-axis representing these values and the X-axis representing the starting position (bp) of each window along the chromosome. These plots were generated for each chromosome. In addition, for GC skew, plots were also generated with the X-axis representing the relative starting position (ratio) on each chromosome for reasons discussed later.

For the TA skew and GC skew plots, a regression line calculated based on the data was superimposed to clarify the gradient in each chromosome. Microsoft® Excel for Mac was used to process the data, and JMP® PRO 18 was used to generate the plots.

Results

The nucleotide counts for each of the 14 chromosomes of *Plasmodium falciparum 3D7* are shown (Table 1). As an indicator of the nucleotide composition of each chromosome, the TA skew, GC skew, and GC content for each chromosome were also calculated and listed. The numbers of thymine and adenine are almost equal, and the numbers of guanine and cytosine are also almost equal (Table 1).

The results of plotting TA skew, GC skew, and GC content for each chromosome by the position of each 5000 bp window are shown. Distributions of TA skew and GC content remained nearly constant regardless of the window position, although GC content exhibited a somewhat distinctive distribution pattern. In contrast, GC skew showed consistent gradient distributions, with lower values at the 5' end and higher values at the 3' end depending on the window position (Figure 1a, 1b, 1c).

For GC skew, plots were also generated using the relative position (ratio) of each window, with the 5' end as 0 and the 3' end as 1, instead of the absolute window position (bp). Despite differences in chromosome size, all chromosomes showed a nearly uniform linear gradient in GC skew in these plots with relative positions (Figure 1d).

Discussion

In the first part of this study, the number of nucleotides on each chromosome in the *Plasmodium* genome was counted. The approximate parity between thymine and adenine, as well as guanine and cytosine, known as Chargaff's Second Parity Rule (CSPR), was also observed in *Plasmodium*, indicating that CSPR is indeed applicable to this organism.

Then, in this study, the distribution of TA skew, GC skew, and GC content along each chromosome was examined. The results showed that the distribution of TA skew was consistent with that observed in other organisms adhering to CSPR. Surprisingly, however, the GC skew exhibited a common linear gradient depending on their chromosomal position in all 14 chromosomes. Furthermore, this gradient was consistent across each chromosome and appeared to correlate with the relative position along the entire length of the chromosomes, regardless of their size.

The simplest hypothesis to explain CSPR is thought to be that it arises from symmetry in the rates of single nucleotide substitutions. However, the findings of this study suggest that deviations from CSPR in *Plasmodium* are influenced by the relative position on the chromosome. This behavior cannot be fully explained by the symmetry of single nucleotide substitutions alone, suggesting that actual adherence to CSPR involves a more complex underlying mechanism.

Conclusion

In the *Plasmodium* genome, local GC skews have been shown to exhibit a uniform gradient distribution pattern by their relative positions along each chromosome. This suggests that CSPR cannot be fully explained by single nucleotide substitutions alone.

Reference

1. Rudner, R., Karkas, J. D., & Chargaff, E. (1968). Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. In Proceedings of the National Academy of Sciences (Vol. 60, Issue 3, pp. 921–922). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.60.3.921

2. Jain, S., Raviv, N., & Bruck, J. (2018). Attaining the 2nd Chargaff Rule by Tandem Duplications. In 2018 IEEE International Symposium on Information Theory (ISIT). 2018 IEEE International Symposium on Information Theory (ISIT). IEEE. https://doi.org/10.1109/isit.2018.8437526

3. Pflughaupt, P., & Sahakyan, A. B. (2023). Generalised interrelations among mutation rates drive the genomic compliance of Chargaff's second parity rule. In Nucleic Acids Research (Vol. 51, Issue 14, pp. 7409–7423). Oxford University Press (OUP). https://doi.org/10.1093/nar/gkad477

4. Forsdyke, D. R. (2021). Neutralism versus selectionism: Chargaff's second parity rule, revisited. In Genetica (Vol. 149, Issue 2, pp. 81–88). Springer Science and Business Media LLC. https://doi.org/10.1007/s10709-021-00119-5

5. Esumi, G. (2023). The Nucleic Acid Sequences of the Genome Are Highly Structured on a Genome-Wide Scale in Terms of Nucleic Acid Composition Indices Such as TA Skew and GC Skew. Jxiv. https://doi.org/10.51094/jxiv.436

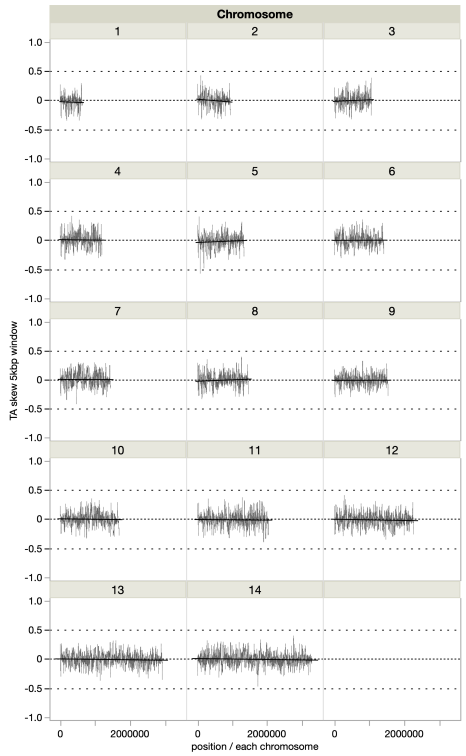
6. National Center for Biotechnology Information. Genome assembly GCA_000002765, *Plasmodium falciparum 3D7*. (April 7, 2016). Retrieved from https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002765.6/ Accessed August 30, 2022.

Table 1

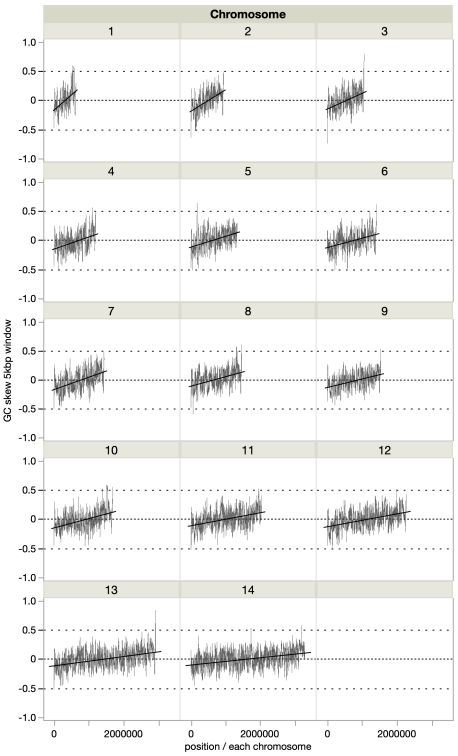
Chromosome	T	A	G	C	total	TA skew	GC skew	GC content
1	249026	260441	66352	65032	640851	-0.022	0.010	0.205
2	380683	379375	93073	93971	947102	0.002	-0.005	0.197
3	428650	425041	107768	106512	1067971	0.004	0.006	0.201
4	482862	471349	120580	125699	1200490	0.012	-0.021	0.205
5	534957	548922	131765	127913	1343557	-0.013	0.015	0.193
6	572337	565274	139998	140633	1418242	0.006	-0.002	0.198
7	584842	573501	141941	144923	1445207	0.010	-0.010	0.198
8	592172	592412	146632	141589	1472805	0.000	0.017	0.196
9	622441	626123	144862	148309	1541735	-0.003	-0.012	0.190
10	682693	673445	166037	165481	1687656	0.007	0.002	0.196
11	822933	828383	195330	191694	2038340	-0.003	0.009	0.190
12	911053	921969	219985	218487	2271494	-0.006	0.003	0.193
13	1184698	1186161	280394	273983	2925236	-0.001	0.012	0.190
14	1346276	1338745	305646	301269	3291936	0.003	0.007	0.184

Table 1. Nucleotide numbers and calculated indices for each chromosome of *Plasmodium falciparum 3D7*.
The table shows the count of thymine (T), adenine (A), guanine (G), and cytosine (C) nucleotides in each chromosome. The total column represents the sum of the four nucleotides, which corresponds to the size of each chromosome. For TA skew, GC skew, and GC content, bar graphs were created to allow visual comparison. The first two indices (TA skew and GC skew) are represented by bars ranging from -1 to 1 to indicate deviation from zero, while the last one (GC content) is shown with bars ranging from 0 to 1. The graphs for the first two indices, TA skew and GC skew, show minimal deviation from zero, with very short bars. The GC content is almost identical across all chromosomes.

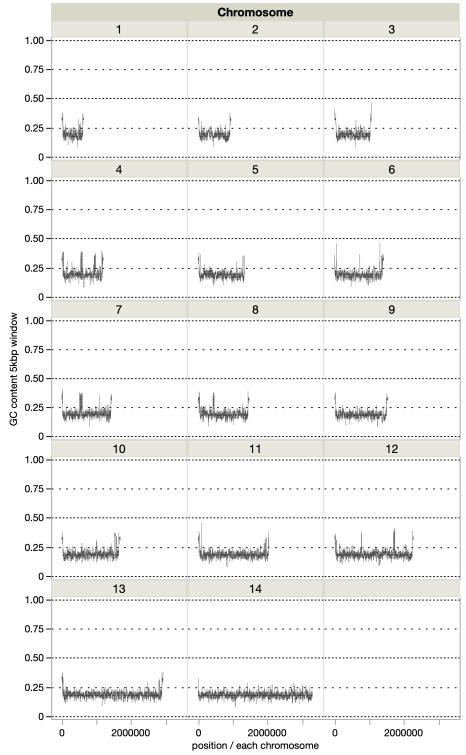
Figure 1(a)



(b)



(c)



(d)

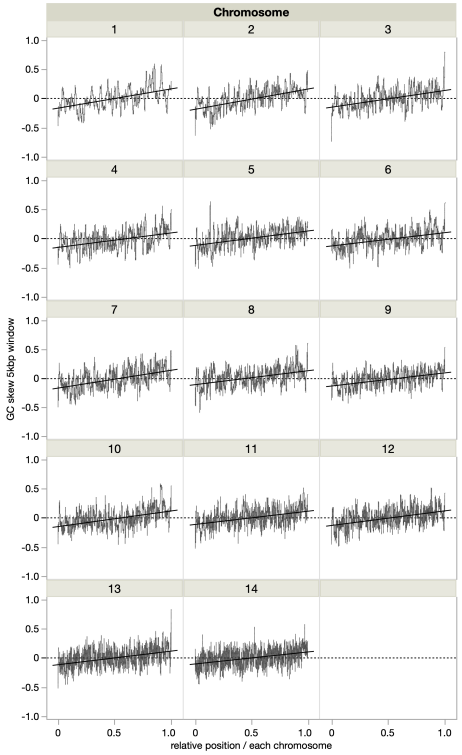


Figure 1. Distribution of TA skew, GC skew, and GC content across the 14 chromosomes of *Plasmodium falciparum 3D7*

(a) TA skew plotted along the position on each chromosome. The skew remains relatively constant along the chromosomes.

(b) GC skew plotted along the position on each chromosome. A linear gradient is observed, with skew values increasing from the 5' end to the 3' end.

(c) GC content plotted along the position on each chromosome. The content shows minor variation but is largely consistent.

(d) GC skew plotted along the relative position on each chromosome, with the 5' end set to 0 and the 3' end set to 1. A consistent linear gradient is observed across all chromosomes, regardless of their length.