An Isochore-Like Structure in the Genome of the Flatworm *Schistosoma mansoni*

Guillermo Lamolle[1], Anna V. Protasio[2], Andrés Iriarte[1,3], Eugenio Jara[1], Diego Simón[1], and Héctor Musto[1,*]

[1]Laboratorio de Organización y Evolución del Genoma, Facultad de Ciencias, Udelar, Montevideo, Uruguay

[2]Wellcome Trust Genome Campus, Wellcome Trust Sanger Institute, Cambridge, United Kingdom

[3]Dpto. de Desarrollo Biotecnológico, Facultad de Medicina, Instituto de Higiene, Udelar, Montevideo, Uruguay

*Corresponding author: E-mail: hmusto@gmail.com.

Accepted: July 6, 2016

Abstract

Eukaryotic genomes are compositionally heterogeneous, that is, composed by regions that differ in guanine–cytosine (GC) content (isochores). The most well documented case is that of vertebrates (mainly mammals) although it has been also noted among unicellular eukaryotes and invertebrates. In the human genome, regarded as a typical mammal, this heterogeneity is associated with several features. Specifically, genes located in GC-richest regions are the GC3-richest, display CpG islands and have shorter introns. Furthermore, these genes are more heavily expressed and tend to be located at the extremes of the chromosomes. Although the compositional heterogeneity seems to be widespread among eukaryotes, the associated properties noted in the human genome and other mammals have not been investigated in depth in other *taxa*. Here we provide evidence that the genome of the parasitic flatworm *Schistosoma mansoni* is compositionally heterogeneous and exhibits an isochore-like structure, displaying some features associated, until now, only with the human and other vertebrate genomes, with the exception of gene concentration.

Key words: *Schistosoma mansoni*, genome evolution, isochores, compositional correlations.

Introduction

Vertebrate genomes, mainly mammalian and human in particular, are the most well characterized and studied. In terms of its guanine–cytosine (GC) content, the human genome is composed of a series of long (> 300 kb) blocks whose boundaries are determined by a significant change in the GC content. According to Bernardi's group (Costantini et al. 2009) the human genome (which is not different from the majority of mammalian genomes) is composed by five families of isochores, namely L1, L2, H1, H2, and H3, ranging from GC-poor (L1) to GC-rich (H3) isochores. However, isochores do not differ among each other only at the GC level. Indeed, each isochore family displays its own phenotype, such as the distribution of oligonucleotides (Costantini and Bernardi, 2008a), its localization within the cell nucleus (Saccone et al. 2002), the distribution of repeated sequences like SINES and LINES (Pavlícek et al. 2001), the replication time (Schmegner et al. 2007), gene density (D'Onofrio et al. 1999), distribution of CpG islands and the presence of 5mC (Bird, 1980; Varriale and Bernardi, 2010), and the coding sequences (CDS) and introns length (Comeron, 2004). Therefore, the isochore structure exhibits a crucial role not only in the phenotype of the human genome but also in its physiology.

Compositional heterogeneity is widespread in evolution (Cammarano et al, 2009), even in prokaryotes (Bernaola-Galván et al. 2004; Bohlin et al. 2010). But to the best of our knowledge, most of the physiological features associated with the compositional heterogeneity had almost exclusively been found in mammals. Therefore, a crucial question arises. Given that all these features are well known for mammalian genomes, and they are not present in prokaryotes, when did they emerge in the evolution of eukaryotes?

To answer this crucial question in the evolution of the eukaryotic genome, we have analyzed the genome of *Schistosoma mansoni*. Previous results from our group, published around 20 years ago (Musto et al. 1994; 1995; 1998)

suggested that the nuclear genome of *S. mansoni* might be compositionally compartmentalized, in other words, organized in an "isochore-like" structure. Although the number of sequences at that time was scarce, and therefore our analyses were prone to be biased, three different pieces of evidence pointed toward that interpretation. Firstly, the frequency of the dinucleotide CpG was highly variable among CDS and correlated with the GC content of the flanking regions (Musto et al. 1994). Secondly, we detected significant correlations 1) between the GC3 of each gene with the corresponding GC of the surrounding regions, 2) between the frequencies of CpG dinucleotide with the respective frequencies in flanking regions, and 3) between the GC content of 5′ and 3′ flanking CDS regions (Musto et al. 1995). Third, although there is a remarkable variation in codon usage among different genes, this does not seem to be related to translational selection, but is strongly dependent of the physical location of each gene (Musto et al. 1998). All three points, as mentioned, are characteristic of strongly compositional compartmentalized genomes, as those of mammals (for a review on this topic, see Bernardi, 2015) and birds (Musto et al. 1999).

To investigate if an isochore-like structure with its associated features is a primitive characteristic that appeared early in evolution of the genome in multicellular species, we analyzed the genome of the flatworm *S. mansoni*. This species is ideal for this goal since 1) it occupies a basal position in the evolution of metazoan species (Halanych, 2004) and 2) its complete genome is available and although fragmented it is by far the most contiguous and better annotated of all schistosome species (Protasio et al. 2012).

Finally, given its importance as the causative agent of schistosomiasis, an important and neglected human disease (Steinmann et al. 2006), the understanding of its basic genomic features can help to elucidate its basic biology, and therefore, may contribute to develop strategies to control its influence on human health.

The analyses reported here show that the genome of *S. mansoni* 1) is indeed compositionally heterogeneous and 2) shows an isochore-like structure. Most importantly, we demonstrate that the genes localized at the GC-richest regions are the most highly expressed in any given life-cycle stage analyzed, and display features such as shorter introns and CDS, namely the typical signature of isochoric organization similar to that of the human genome (Bernardi, 1995). The results are discussed within an evolutionary framework.

Materials and Methods

Genomic, CDS and annotation data were obtained from ftp://ftp.ebi.ac.uk/pub/. The *S. mansoni* reference genome was sequenced by a consortium led by the Wellcome Trust Sanger Institute, as described by Berriman et al. (2009). The assembly version used here is the v5.0 freeze from GeneDB (December 2013), published in Protasio et al. (2012) and which corresponds with INSDC assembly ASM23792v2. Expression data Reads per kilobase per million mapped reads (RPKM) were taken from Protasio et al. (2012).

Although there are more than 10,000 annotated CDS, the original set was filtered removing CDS without initiation and/or stop codons or with frameshifts, as well as those shorter than 100 bp. For some analyses, the number of CDS that met these conditions and for which, in addition, data expression is available, was reduced to less than 8,000 CDS. This number is sufficient to extract statistical conclusions.

Sliding windows analyses at the chromosomic level were performed using the R (R Development Core Team, 2008) package Zoo (Zeileis & Grothendieck 2005), and compositional analyses were carried out with R package seqinR (Charif and Lobry, 2007). Several complementary processes and calculations were performed using home-made Perl and R scripts.

For certain analyses, chromosomes were divided into non-overlapping stretches. The compositional properties of several window sizes were studied (from 25 kb to 300 kb), and the most informative length was 100 kb. Furthermore, in short chromosomes, the variation among contiguous stretches vanishes when using longer windows. Therefore, this length was chosen. These fragments were classified into three regions according to their GC content: A ($GC\% < 33.5$), B ($33.5 \leq GC\% < 36.5$), and C ($GC\% \geq 36.5$). These limits are essentially arbitrary and represent a compromise between the division into thirds according to the distribution of GC content, and the number of windows falling in each region. Other divisions tested yielded similar results. Further studies may allow refine the methodology. The number of windows of A, B, and C were, respectively, 589, 1,399, and 595. In some cases, the same approach was used for 1) the correlation between GC contents of CDS and their 100 kb windows and 2) the location of each CDS in the corresponding window defined by GC content, but masking the CDS before calculating the GC content of the windows. This modification was applied in order to avoid self-correlations and other distortions. With the latter strategy, regions A, B, and C displayed 597, 1,373, and 613 windows.

When calculating the average value of RPKM per window, only those windows with one or more genes were considered. In this case, A, B, and C contained 551, 1,317, and 607 windows.

Correlations were generally calculated according to the Pearson method, but in cases where one of the variables had a nonnormal distribution and/or there were many outliers, the nonparametric method of Spearman was used. In this case, the significance was evaluated using multiple permutations-based tests, and this is noted by writing ρ instead r. Unless specified, the P values were always < 0.0001.

In all figures, bands A, B, and C are represented in blue, orange, and red, respectively.

Results and Discussion

The haploid genome of *S. mansoni* (360 MB) is organized into eight chromosomes, seven autosomes, and one pair of sexual chromosomes with males carrying ZZ and females carrying ZW chromosomes, and 10,852 protein-coding genes have been annotated (Protasio et al. 2012).

Compositional Heterogeneity

The haploid genome of *S. mansoni* (360 MB) is organized into eight chromosomes, seven autosomes, and one pair of sexual chromosomes with males carrying ZZ and females carrying ZW chromosomes, and 10,852 protein-coding genes have been annotated (Protasio et al. 2012).

Figure 1 shows a histogram of the GC content of the genome of *S. mansoni* cut in pieces of 100 kb. The distribution is normal, with a mean of 35% GC (standard deviation = 1.9%). The distribution is skewed toward low-GC values, and there is a higher frequency of some windows characterized by certain GC content (e.g., 32%). This result suggests that, to some extent, compositional heterogeneity is present in this genome as suggested by Musto et al. (1999).

GC Content Chromosome by Chromosome

Figure 2 shows the GC content of nonoverlapping 100 kb windows for each chromosome. Some conclusions can be derived from these figures. First, the compositional heterogeneity, although strongly influenced by the bias toward AT, is evident for each chromosome, second, and perhaps more important, the GC content distribution is not uniform along the chromosomes, since in the majority of them, the GC-richest fragments tend to be located at the ends of the chromosomes. A similar outcome, although with highest GC values, has been described for human chromosomes (Costantini et al. 2006). Third, although there are stretches of the same GC category windows, many "jumps" are evident (e.g., GC-rich fragments in the mid region of Chr3 and Chr6 and two "islands" in the ZW Chr). This clearly suggests that different and contiguous regions of the genome of *S. mansoni* are submitted to different factors leading to changes in GC content. Again, this resembles the structure of the human genome (Costantini et al. 2006).

Compositional Correlations

The results described above, led us to investigate compositional correlations, that is, the correlations that hold between the GC levels of coding and noncoding regions: GC3, introns, CDS, flanking regions, etc. This is a crucial point for two reasons. First, we have previously postulated, albeit with very few sequences available at that time, that an "isochore-like structure" was present in the genome of *S. mansoni* (Musto et al. 1994; 1995) and if this study was correct, then the same correlations would be observed when using the now available complete genome. Second, in all well-studied genomes displaying isochores, the compositional correlations do hold (Costantini and Bernardi, 2008b).

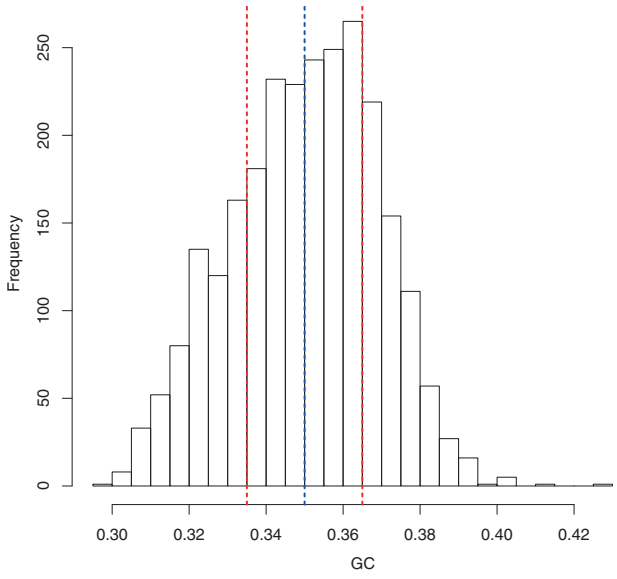

Fig. 1.—Distribution of GC content of nonoverlapping 100 kb fragments of the whole genome of *S. mansoni*. Blue dashed line represents the average GC content. Red dashed lines represent the limits between regions A, B, and C, as defined in Materials and Methods.

We found significant correlations between the GC levels of: exons versus flanking regions, upstream versus downstream regions, exons versus introns, introns versus flanking regions, etc. We should note that the *r* values were usually low (never higher than 0.41). This could be explained by the fact that the variation in GC content is low, and therefore minor variations due to other factors (codon usage, selection for RNA secondary structure, conserved regions in introns, random effects in noncoding regions, etc.), might be acting and therefore blurring the compositional associated effect. But in spite of these effects, the compositional correlations described herein reinforce our hypothesis that the genome of *S. mansoni* is characterized by an isochore-like structure.

Higher Concentration of Genes in GC-Rich Regions

In a highly compositionally compartmentalized genome like the human genome, genes are not distributed equally in each isochore family (Bernardi, 1995). Indeed, genes are significantly more concentrated in GC-rich isochores (H2 and H3) than in the GC-poor (L1, L2, and H1) (Bernardi, 1995). Although the difference in GC composition of isochores in *S. mansoni* is less intense than in the human genome, we

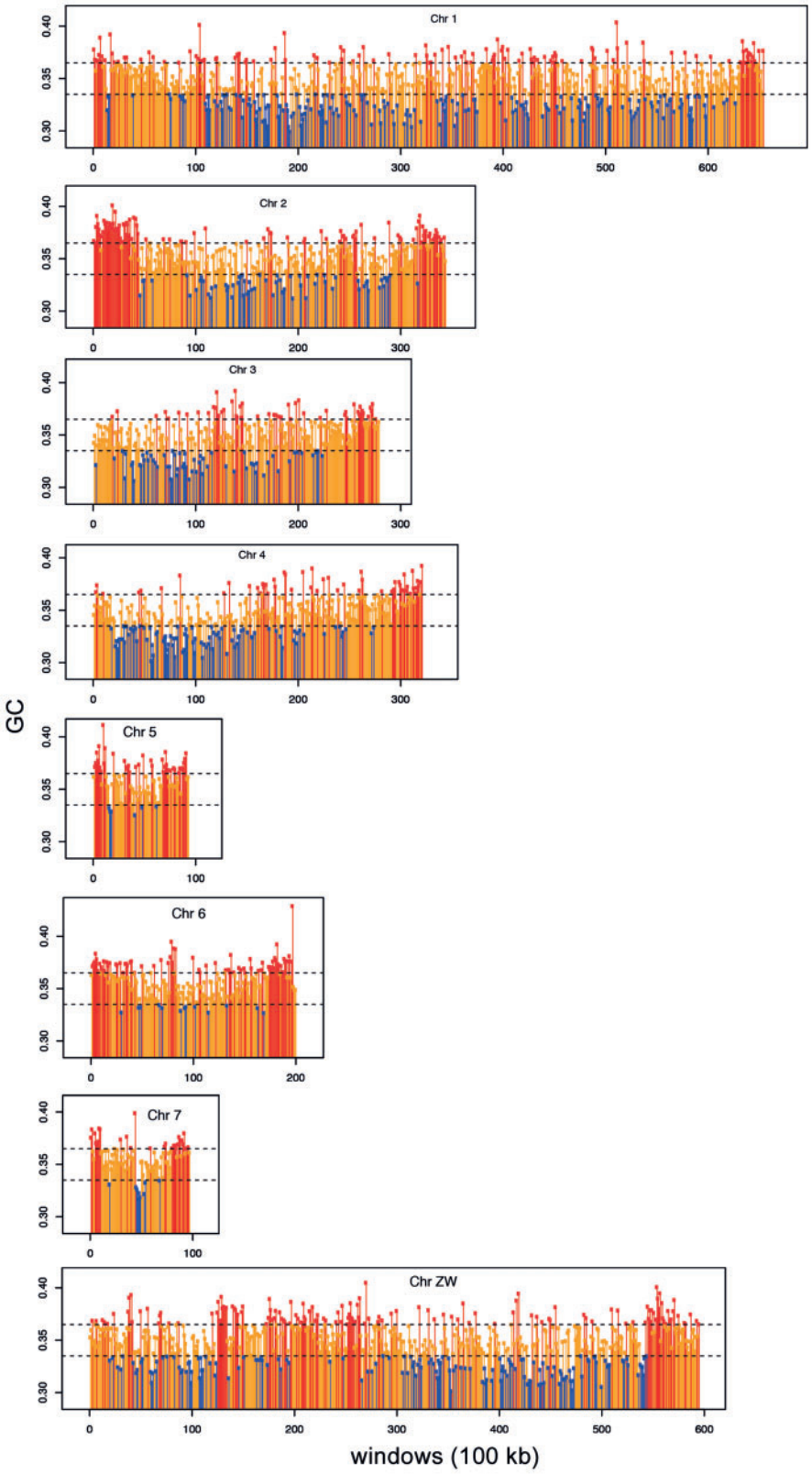

Fig. 2.—Spatial distribution of windows of regions A, B, and C as defined in Materials and Methods for each chromosome. These regions are colored in blue, orange, and red, respectively.

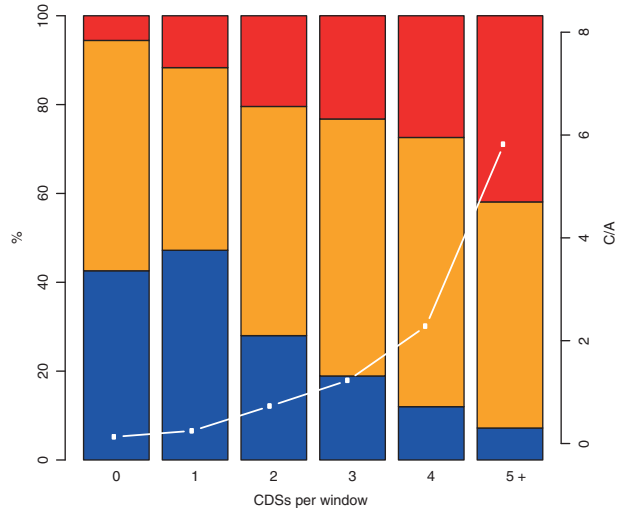

Fig. 3.—Distribution of CDS for regions A, B, or C. Genomic regions with higher GC content (C) tend to have more CDS than regions with lower GC (A). Windows were categorized into six groups according to the number of CDS present in them (x-axis). Colored bars represent the percentage of windows from each region (A, B, and C) in each of the six groups. The white line represents the ratio C/A (right y-axis).

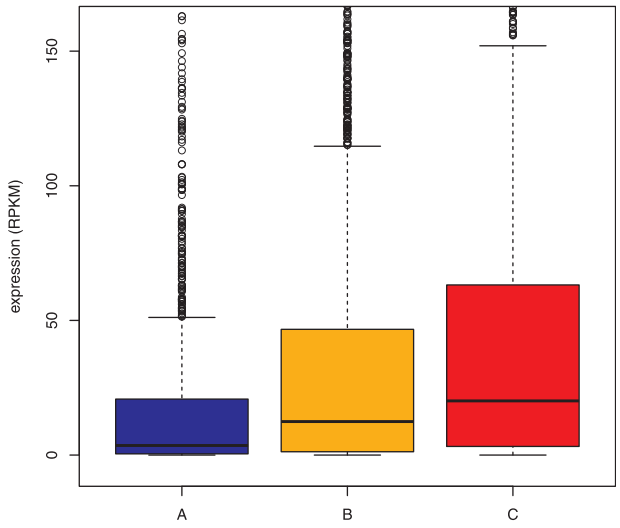

Fig. 4.—Box plots of average expression level (RPKM) calculated for each window, for regions A, B, and C.

studied the distribution of genes according to the following division (see Materials and Methods): region A (GC% < 33.5), region B ($33.5 \leq$ GC% < 36.5), and region C (GC% ≥ 36.5). As can be seen in figure 3, the concentration of genes increases from A to B to C. Again, in spite of the low variation in GC content, this result mirrors the results found in the human genome. These calculations were done excluding coding regions, therefore avoiding biases arising from including CDS. A similar pattern was previously described in some invertebrates (Cammarano et al. 2009).

Gene Expression Correlates with GC Content

For the human genome, it has been demonstrated that genes expressed at higher levels are localized in the GC-richest isochores (Bernardi, 2007). Therefore, taking advantage of the available expression data, we decided to investigate if the same is true for *S. mansoni*. Figure 4 shows the mean RKPM values of genes located in the three different isochore-like categories. These results suggest that, like in the human genome, higher expressed genes are located in GC-richer regions. Genes located in C (see section above and Materials and Methods) are significantly more expressed than genes located in A or B. This result goes in the direction that the genome of *S. mansoni*, in spite of its lowest variation in GC content, displays features at the genomic level that, in a way, make it similar to the human genome, and very probably to all mammalian genomes. We note that the results shown in figure 4 are from intramammalian stages. When data from intramolluscan and free-living stages were analyzed, the results were qualitatively identical (data not shown).

The Length of Introns and CDS

Figures 5 and 6 show the relationship between GC-content of regions A, B, and C with intron and CDS lengths, respectively. Both the Kruskal–Wallys test and the paired Mann–Whitney test give significant values ($P < 0.01$), except for Mann–Whitney test for the comparison of the pair AB in the case of the lengths of the CDS ($P = 0.29$). These results show that both noncoding and coding regions of genes, tend to display a compaction process associated with increased GC content of the regions surrounding them, as occurs in the human genome (Bernardi, 1995).

Conclusions

Several years ago, we have postulated that the genome of *S. mansoni* was compositionally compartmentalized, displaying an isochore-like structure. Those studies were performed with a small and hence less reliable data set. In this communication, we show that indeed, the genome of this important human parasite is not only composed by an isochore-like structure, but displays features that make it similar to the human (and other mammalian) genomes. We conclude that the features observed, such as the higher concentration of highly expressed genes at the GC-richest regions, and their tendency to be located at the extremes of the chromosomes are not random and we postulate that they are at the very basis of what makes a multicellular eukaryote. The relative role of purifying selection on this organization and the effect of this structure constraining the synonymous codon usage, the

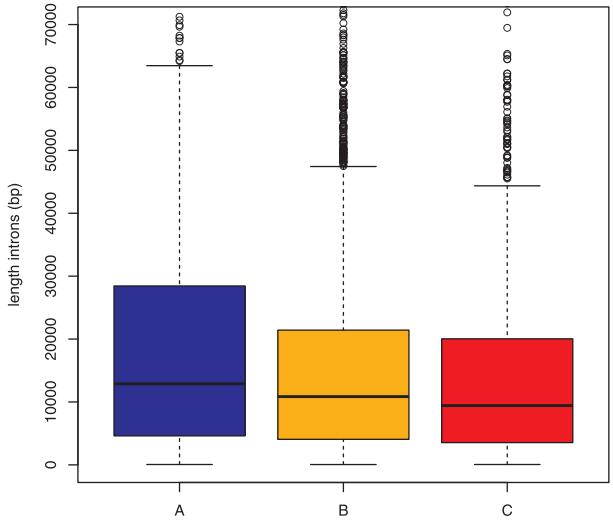

Fig. 5.—Box plots of average intron length for each window, for regions A, B, and C.

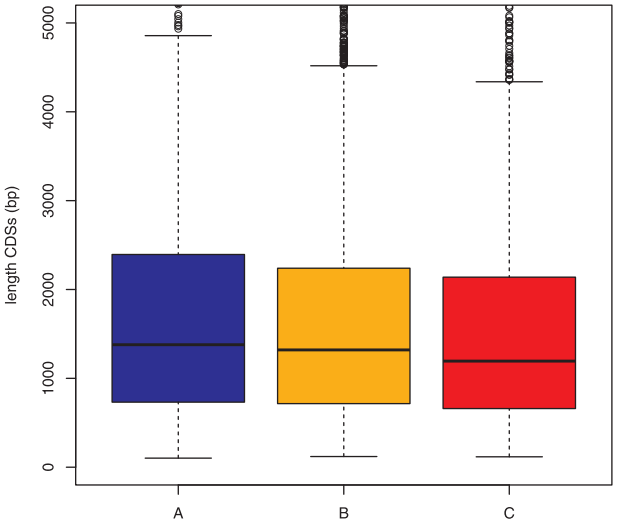

Fig. 6.—Box plots of average CDS lengths for each window, for regions A, B, and C.

regulation mechanisms and/or the rate of rearrangements is something that deserves more analyses.

Platyhelminths are basal in the tree of life, in relation to vertebrates. And, at least *S. mansoni*, as we show, does display features that in spite of its strong compositional bias toward AT, is similar with the human genome. Therefore, we postulate that an isochore-like structure, with its associated idiosyncratic features, is intrinsic to the genome evolution of multicellular eukaryotes. Why and how the primordial isochores which we call A, B, and C, evolved to the families L1, L2, H1, H2, and H3 in mammals, and even H4 in birds, remains to be established. But analyzing new complete and well-annotated genomes from organisms belonging to different multicellular *phyla* will certainly shed light on this crucial aspect of molecular evolution. This paper is a step on this way.

Acknowledgments

We thank Gabriel Rinaldi and Fernando Álvarez-Valin for critically reading the manuscript. We thank PEDECIBA and ANII for finantial support.

Literature Cited

Bernaola-Galván P, Oliver J, Carpena P, Clay O, Bernardi G. 2004. Quantifying intrachromosomal GC heterogeneity in prokaryotic genomes. Gene 333:121–133.

Bernardi G. 1995. The human genome: organization and evolutionary history. Annu Rev Genet. 29:445–476.

Bernardi G. 2007. The neoselectionist theory of genome evolution. Proc Natl Acad Sci U S A. 104(20):8385–8390.

Bernardi G. 2015. Chromosome architecture and genome organization. PLoS ONE 10(11):e0143739. doi:10.1371/journal.pone.0143739.

Berriman M, et al. 2009. The genome of the blood fluke *Schistosoma mansoni*. Nature 460(7253):352–358.

Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 8(7):1499–1504.

Bohlin J, et al. 2010. Analysis of intra-genomic GC content homogeneity within prokaryotes. BMC Genomics 11:464. doi: 10.1186/1471-2164-11-464.

Cammarano R, Costantini M, Bernardi G. 2009. The isochore patterns of invertebrate genomes. BMC Genomics 14:538. doi: 10.1186/1471-2164-10-538

Charif D, Lobry JR. 2007. A contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U., Porto M., Roman H.E., Vendruscolo M., editors. Molecules, networks, populations; Berlin (Heidelberg): Springer. pp. 207–232.

Comeron J. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics 167(3):1293–1304.

Costantini M, Bernardi G. 2008a. The short-sequence designs of isochores from the human genome. Proc Natl Acad Sci U S A. 105(37):13971–13976.

Costantini M, Bernardi G. 2008b. Correlations between coding and contiguous non-coding sequences in isochore families from vertebrate genomes. Gene 410(2):241–248.

Costantini M, Cammarano R, Bernardi G. 2009. The evolution of isochore patterns in vertebrate genomes. BMC Genomics 10:146. doi: 10.1186/1471-2164-10-146.

Costantini M, Clay O, Auletta F, Bernardi G. 2006. An isochore map of human chromosomes. Genome Res. 16:536–541.

D'Onofrio G, et al. 1999. Evolutionary genomics of vertebrates and its implications. Ann N Y Acad Sci. 870:81–94.

Halanych, K. 2004. The new view of animal phylogeny. Annu. Rev. Ecol. Evol. Syst. 35:229–256

Musto H, Rodríguez-Maseda H, Alvarez F, Tort J. 1994. Possible implications of CpG avoidance in the flatworm *Schistosoma mansoni*. J Mol Evol. 38(1):36–40.

Musto H, Rodríguez-Maseda H, Alvarez F. 1995. Compositional correlations in the nuclear genes of the flatworm *Schistosoma mansoni*. J Mol Evol. 40(3):343–346.

Musto H, Romero H, Rodríguez-Maseda H. 1998. Heterogeneity in codon usage in the flatworm *Schistosoma mansoni*. J Mol Evol. 46(2):159–167.

Musto H, Romero H, Zavala A, Bernardi G. 1999. Compositional correlations in the chicken genome. J Mol Evol. 49(3):325–329.
Pavlícek A, et al. 2001. Similar integration but different stability of Alus and LINEs in the human genome. Gene 276(1–2):39–45.
Protasio AV, et al. 2012. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. PLoS Negl Trop Dis. 6:e1455. doi: 10.1371/journal.pntd.0001455.
R Development Core Team (2008). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org (2008).
Saccone S, Federico C, Bernardi G. 2002. Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. Gene 300(1–2):169–178.
Schmegner C, Hameister H, Vogel W, Assum G. 2007. Isochores and replication time zones: a perfect match. Cytogenet Genome Res. 116(3):167–172.
Steinmann P, Keiser J, Bos R, Tanner M, Utzinger J. 2006. Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. Lancet Infect Dis. 6:411–425. doi: 10.1016/s1473-3099(06)70521-7
Varriale A, Bernardi G. 2010. Distribution of DNA methylation. CpGs, and CpG islands in human isochores 95(1):25–28. doi: 10.1016/j.ygeno.2009.09.006.
Zeileis A, Grothendieck G. 2005. zoo: S3 Infrastructure for Regular and Irregular Time Series. J Stat Softw. 14(6):1–27. doi:10.18637/jss.v014.i06

Associate editor: Maria Costantini