C. R. Mecanique 340 (2012) 745–757

Contents lists available at SciVerse ScienceDirect

Comptes Rendus Mecanique

www.sciencedirect.com

Out of Equilibrium Dynamics

Gene organization inside replication domains in mammalian genomes

Lamia Zaghloul [a,b], Antoine Baker [a,b], Benjamin Audit [a,b], Alain Arneodo [a,b,*]

[a] *Université de Lyon, 69000 Lyon, France*
[b] *Laboratoire de Physique, ENS de Lyon, CNRS, 46, allée d'Italie, 69007 Lyon, France*

ARTICLE INFO

Article history:
Available online 16 November 2012

Keywords:
Gene organization
Replication domains
Gene length
Chromatin structure

ABSTRACT

We investigate the large-scale organization of human genes with respect to "master" replication origins that were previously identified as bordering nucleotide compositional skew domains. We separate genes in two categories depending on their CpG enrichment at the promoter which can be considered as a marker of germline DNA methylation. Using expression data in mouse, we confirm that CpG-rich genes are highly expressed in germline whereas CpG-poor genes are in a silent state. We further show that, whether tissue-specific or broadly expressed (housekeeping genes), the CpG-rich genes are over-represented close to the replication skew domain borders suggesting some coordination of replication and transcription. We also reveal that the transcription of the longest CpG-rich genes is co-oriented with replication fork progression so that the promoter of these transcriptionally active genes be located into the accessible open chromatin environment surrounding the master replication origins that border the replication skew domains. The observation of a similar gene organization in the mouse genome confirms the interplay of replication, transcription and chromatin structure as the cornerstone of mammalian genome architecture.

1. Introduction

The organization of genes in the genomes of a wide range of species is non-random [1]. First in bacterial genomes that are circular and very compact, mostly composed of genes, the distribution of genes is non-random with respect to the replication origin. Highly expressed genes tend to be close to the origin [2] and essential genes tend to be co-oriented with the replication fork [3,4], possibly as the result of a selective pressure. In human and mouse, genes only account for ∼30–40% of the genome content, while protein-coding DNA accounts for less than 5% [5,6]. In a sea of non-coding DNA, where are the genes? Are they randomly distributed or do they tend to cluster together? Are genes with similar expression patterns clustered? Are the genes still organized around replication origins? How is this gene organization related to chromatin environment?

For many years, mammalian genome organization has been linked to the so-called isochore structure [7]. As compared to GC-poor isochores, GC-rich isochores tend to have a high gene density, short genes and to be more transcriptionally active [5,8–10]. The recent discovery of replication skew domains in the human and mouse genomes from the analysis of nucleotide compositional asymmetries [11–13] has provided a new perspective in the understanding of gene organization in mammalian genomes (Fig. 1) [14,15]. Because skew domain borders were shown to be "master" origins that are likely to play a key role in the regulation of the replication spatio-temporal program [15–20] by the chromatin tertiary structure [14,18,21,22], it is of fundamental interest to study the particular organization of genes around them. Indeed, these master origins

[*] Corresponding author at: Laboratoire de Physique, ENS de Lyon, CNRS, 46, allée d'Italie, 69007 Lyon, France.
E-mail address: alain.arneodo@ens-lyon.fr (A. Arneodo).

1631-0721/$ – see front matter © 2012 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.
http://dx.doi.org/10.1016/j.crme.2012.10.023

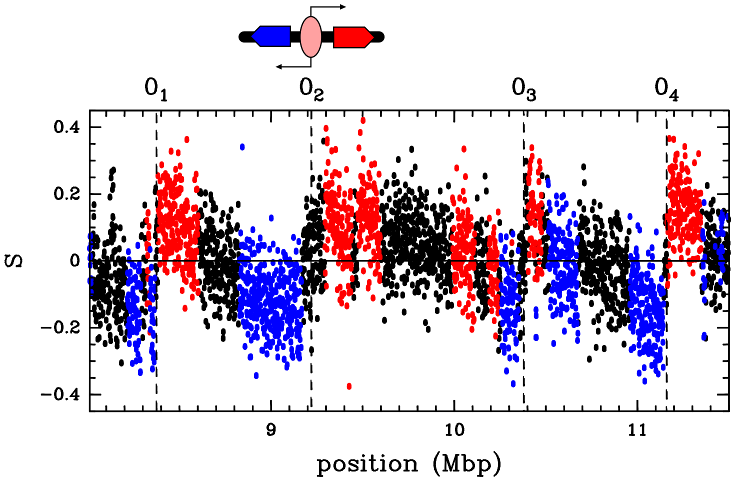

Fig. 1. Illustration of the gene organization around 4 putative replication origins O_1, O_2, O_3 and O_4 bordering 3 adjacent skew N-domains in a fragment of the human chromosome 9 [14]. The skew $S = S_{TA} + S_{GC}$ was computed on repeat-masked sequences; each point corresponds to a 1 kbp window: red, (+) genes (coding strand identical to the Watson strand); blue, (−) genes (coding strand opposite the Watson strand); black, intergenic regions (the color is defined by majority rule). In most cases, genes are oriented in the same direction as replication fork progressing from the putative replication origins (vertical dashed lines).

were shown to be specified by a region of open chromatin (∼300 kbp) in an otherwise heterochromatin environment [22] and that they tend to be replicating early in the S phase [15,16,18]. These early replication initiation zones were further shown to be permissive to transcription. As illustrated in Fig. 1, around these master origins, genes are abundant and broadly expressed and their transcription is predominantly co-oriented with replication fork progression [15]. At the center of the replication skew domains, genes are rare and expressed in few tissues. In this manuscript, we will address some issues that these preliminary observations raise. In particular we will try to bring some answer to the following questions. Do genes close to replication skew domain borders tend to be expressed in the germline? What is behind the observed decrease of average gene expression breadth with the distance to these domain borders in terms of the individual status of genes? Does the presence of an accessible chromatin in these early initiation zones contribute to the positive regulation of gene expression?

The manuscript is organized as follows. Section 2 is devoted to materials and methods. In Section 3, we show that human and mouse genes can be separated in two categories depending on their CpG enrichment at the promoter (either CpG-rich or CpG-poor). Using a precise germline expression dataset in mouse, we further demonstrate that CpG enrichment at the promoter is a marker of gene expression in the germline and more precisely in spermatogonia, which is one of the stages of spermatogenesis. In Section 4, we show that the over-representation of broadly expressed housekeeping genes close to replication domain borders can mainly be explained by their belonging to the CpG-rich category. Indeed, it is CpG-rich genes, whether tissue-specific or broadly expressed, that are over-represented close to the bordering master replication origins, whereas CpG-poor genes appear to be homogeneously distributed in replication skew domains. In Section 5, we investigate the orientation bias of genes nearby replication domain borders. We show that the preferential divergent orientation is mostly true for the longest CpG-rich genes in relation to the local accessible open chromatin environment. We conclude in Section 6 by discussing the robustness of the reported results in mammalian germline cells as well as their possible generalization to other cell types.

2. Materials and methods

2.1. Sequence data

Sequence data were retrieved from the Genome Browser of the University of California Santa Cruz (UCSC) [23]. We used human genome assembly of May 2004 (NCBI35 or hg17) and mouse assembly of February 2006 (NCBIm36 or mm8).

Compositional skew Compositional skew (or skew) was computed as

$$S = S_{TA} + S_{GC} = \frac{n_T - n_A}{n_T + n_A} + \frac{n_G - n_C}{n_G + n_C}$$

where n_T, n_A, n_G and n_C are the number of T, A, G and C counted along the sequence masked for repeats, using the RepeatMasker annotation. The S_{TA} and S_{GC} were shown to be positively correlated [11,12,24–28], so that we considered the sum of the two.

GC content GC content was computed as

$$GC = \frac{n_G + n_C}{n_A + n_T + n_G + n_C}$$

where n_A, n_T, n_G and n_C are the number of A, T, G and C counted along the sequence masked for repeats, using the RepeatMasker annotation.

CpG enrichment CpG enrichment (also called CpG observed/expected ratio) was computed as

$$\frac{n_{CpG}.(n_A + n_T + n_G + n_C)}{n_C n_G}$$

where n_A, n_T, n_G and n_C are the number of A, T, G and C and n_{CpG} is the number of dinucleotides CG counted along the native sequence.

2.2. Skew domains data

In this work, we considered for our analyses the 678 N-domains detected by Huvet et al. [15] in addition to the 135 N-domains and 113 Split-N-domains [14] that were detected in Zaghloul's thesis [29]. This resulted in a final data set of 926 skew domains.

2.3. Genes and transcription data

CpG-rich and CpG-poor genes We defined as CpG-rich genes, genes with a CpG enrichment at the Transcription Start Site (TSS) (calculated in a 1 kbp window centered on the TSS) > 0.48 and as CpG-poor genes with a CpG enrichment at the TSS < 0.48.

Human genes annotation We used the RefSeq annotation table from the UCSC Genome Browser. We preferred the RefSeq annotation over the Known Genes annotation used in the other chapters because the RefSeq annotation is more precise and is thus more appropriate for analyses at the TSS of genes. When several genes presenting the same orientation overlapped, they were merged into one gene whose coordinates corresponded to the union of all the overlapping gene coordinates. This resulted in a data set of 18 351 genes.

When analyzing genes within skew domains, we only considered genes that are entirely comprised inside of a domain. The distance of the gene to its closest border corresponds to the distance between the border and the closest end of the gene (which corresponds to the TSS if the gene is oriented divergently from the border and to its transcription stop site if the gene is oriented convergently towards the border). This way of measuring the distance enables to avoid possible biases due to the variability in gene sizes. We considered 6875 genes in the 926 N- and Split-N-domains.

Human intergenes Intergenes are defined as the complementary on the genome of genes taken from the Known Genes annotation of the UCSC Genome Browser, and that were extended by 2 kbp on each side. We chose to take the complementary of the Known Genes annotation, which contains more transcripts than the RefSeq annotation, and to extend genes by 2 kbp on each side in order to define intergenes in a conservative way.

Human expression breadth Expression breadth was determined from EST data as described in [15]. EST data were obtained from Sémon et al. [30,31]. The expression breadth of a gene corresponds to the number of tissues where it is expressed. Each gene was associated to an expression value in a set of 50 non-cancerous tissues. When a tissue was associated to expression data both at an adult and an early developmental stage, the expression level was averaged between the two stages. Expression breadths vary between 0 and 46 tissues, with a median value of 11.

The gene annotation we used was different from the one in the EST expression data. We therefore associated genes to an expression breadth value when we could find an overlapping transcript with the same orientation. Out of the 6875 genes in N- and Split-N-domains, 5495 could be assigned an expression breadth value.

Human genes expression level Expression level of genes was determined using microarray data from the GNF Atlas of human gene expression [32], where probes were associated to an expression value in 79 different tissues. We used in our analyses expression data determined in testis germ cells. We associated genes to the mean expression value of the probes that were strictly comprised in the gene and that had the same orientation. We ended up with expression level data for 4810 genes out of the 6875 genes we considered.

Mouse genes annotation In order to determine CpG enrichment at the TSS of all mouse genes, we used the RefSeq annotation table of the UCSC Genome Browser. When several genes presenting the same orientation overlapped, they were merged into one gene whose coordinates corresponded to the union of all the overlapping gene coordinates. This resulted in 18 976 distinct genes over the 21 mouse chromosomes.

Mouse genes germline expression Expression level was determined from microarray data from [33], where probes were associated to an expression value in spermatogonia, spermatocyte, spermatid and whole testis. The original data set, available in the GermOnline database http://www.germonline.org, comprises 6000 genes that were found to be differentially expressed between the different stages analyzed (spermatogonia, spermatocyte, spermatid and whole testis). These genes correspond to the Ensembl genes (ensGene) annotation table of the UCSC Genome Browser. Among these 6000 genes, 258 were present twice or three times in the data set. We associated to these genes the average expression value of the corresponding entries, and ended up with a data set of expression level for 5483 genes.

2.4. Statistical tests

For the analysis of correlations, we reported the Pearson product moment correlation coefficient r and the associated P-value for no association ($r = 0$).

We used the Wilcoxon rank sum test to assess if two distributions are statistically different or if they cannot be differentiated and correspond instead to two samples of the same distribution (Hypothesis H0). The Wilcoxon rank sum test is a non-parametric statistical test that makes no assumptions on the form of the distributions analyzed. We also used the Kruskal–Wallis rank sum test which is an extension of the Wilcoxon rank sum test to the comparison of more than two distributions.

All statistical computations were performed using the R software (http://www.r-project.org/).

3. Sorting the genes according to their expression level in the germline

3.1. CpG islands and DNA methylation

The methylation of DNA consists in the addition of a methyl group on cytosine nucleotides. DNA methylation, as an efficient pathway for gene silencing, plays a great role in the regulation of genomes. It is interesting to note that there are great differences in the methylation landscapes of genomes across species. In contrast with invertebrates or certain plants like *Arabidopsis thaliana* that exhibit a mosaic pattern of methylation, mammalian genomes tend to be globally methylated except for small CpG-rich regions, that are called CpG islands (for review, [34]). These CpG islands (CGIs) have attracted a considerable amount of interest since they were first described [35–37].

Methylated cytosines become hypermutable, turning into thymines, so that CpG deficiency is an indicator of the level of germline DNA methylation. It was early proposed that CGIs escaped germline DNA methylation [37,38]. CGIs were linked to the presence of a promoter active in the germline in a single-gene analysis in mouse [39] and genes with a CGI-promoter were shown to be expressed in early embryo using a data set of ∼400 human genes with a mouse ortholog [40]. Indeed, CpG islands were shown to be associated to 60% of human gene promoters [41] and also to replication origins in mammals [42,43]. A recent study, that mapped 283 replication origins in the ENCODE regions of the human genome, showed that indeed they were associated to CGIs but that this was only true for half of them [44]. In mouse, a study recently mapped 97 replication origins in 0.4% of the genome, among which 39 map to a CGI and showed that origins that map on a CGI correspond to the strongest origins in ES cells [45]. It was also reported that more than 65% of skew domains borders from [15] match a CGI [46].

3.2. Two classes of promoters: CpG-rich and CpG-poor promoters

There are several methods to detect CpG islands on the genome using DNA sequence analysis. The standard definition of CGIs are regions of length >200 bp, with a GC content $\geqslant 50\%$ and a CpG enrichment >0.6 [37]. The CpG enrichment (see Section 2) is also sometimes called the CpG observed/expected ratio. A variant definition proposed by [47] uses a more stringent set of parameters to define an island (length >500 bp, with a GC content $\geqslant 55\%$ and a CpG enrichment >0.65). Initially, when using the CGI annotation of the UCSC Genome Browser (based on the first definition), we found several promoters that were not overlapping a CGI yet had an enrichment in CpG content at the promoter. This led us to adopt a different method for assessing the CpG status of promoters. We estimated, as noted by [48], that it is clearer to simply consider the CpG enrichment at the promoter calculated in a 1 kbp window centered on the TSS (see Section 2), rather than its association to a CpG island, whose definition is based on *ad hoc* criteria.

The distribution of CpG enrichment of human genes, as noted by others [49–51], is bimodal (Fig. 2(a)), which is also the case in other mammalian genomes, including the mouse genome (Fig. 2(b)). We could thus easily separate genes in two categories depending on their CpG enrichment at the promoter. We used a threshold value of 0.48 so that genes with a CpG enrichment >0.48 were considered CpG-rich and those with a CpG enrichment <0.48 CpG-poor (Fig. 2). The CpG-poor category has a normal distribution corresponding to the average CpG enrichment of the globally methylated genome, whereas the CpG-rich category has been less subject to DNA methylation in the germline and therefore to CpG loss [49]. We found that our criteria define 65% of the genes as being CpG-rich in good agreement with the previous estimate (60%) of Antequera [41].

Let us point out that these two classes of promoters have different regulations and present different characteristics. Whereas CpG-poor genes have a specific initiation site, usually a TATA-box, CpG-rich genes have a broad initiation site [52].

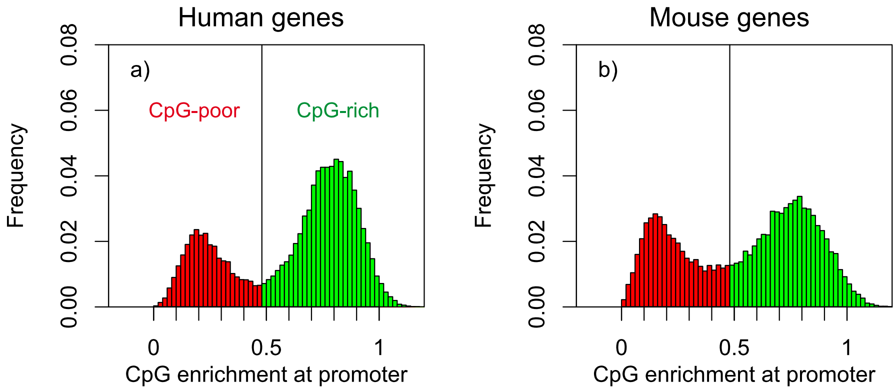

Fig. 2. Histogram of the CpG enrichment calculated in a 1 kbp window centered on the TSS of all 18 351 human genes (a) and of all 18 976 mouse genes (b) (see Section 2). A vertical bar is drawn at the threshold value (0.48) chosen to distinguish between CpG-rich genes (green, on the right of the line) and CpG-poor genes (red, on the left of the line).

Besides, CpG-rich promoters evolve more rapidly than CpG-poor ones. A hypothesis on the origin of the two categories – CpG-rich and CpG-poor – was proposed in [51] but not investigated further: these two categories could have a different evolutionary history, with CpG-rich genes being the 'oldest' ones, present before the global methylation appeared in vertebrate genomes, and CpG-poor genes being more 'recent'.

3.3. *CpG enrichment at the promoter is a marker of spermatogonia expression*

The relation between CpG enrichment at the promoter and germline expression has already been shown for a small subset of genes [40]. We wanted to confirm this result for the whole set of human genes. We used data from the GNF Atlas of human gene expression [32], where expression levels were assessed for all human genes in 79 tissues using microarray. Using the GNF Atlas data set, we found that CpG enrichment at the promoter did not relate well to expression in 'testis germ cells' as many CpG-poor genes are expressed in these cells, and the overall correlation between CpG enrichment and expression level only reaches 0.12. This poor correspondence could be due to an improper estimation of germline expression. We therefore performed the same analyses this time using more precise germline expression data [33] in mouse where each stage of spermatogenesis was distinguished. To our knowledge, there was not a similar data set in human. Note that this data set of mouse germline expression was only available for a subset of 5483 mouse genes that were found to be differentially expressed in between the stages analyzed (spermatogonia, spermatocyte, spermatid and whole testis). The bimodal distribution of CpG enrichment at promoters that we observed in human can also be observed in other mammals. In mouse, less genes are associated to CGIs than in human [38], but the distribution of CpG enrichment at promoters remains comparable (Fig. 2(b)), so that we could use the same threshold (0.48) to distinguish between CpG-rich and CpG-poor genes. We also verified that the histogram of CpG enrichment at the promoter was also comparable for the subset of 5483 mouse genes for which germline expression was available [29].

We were then able to compare expression in spermatogonia, spermatocyte, spermatid and whole testis and how they each correlated to CpG enrichment. The correlations between CpG enrichment and expression level are respectively $r = 0.33$; $P < 2 \times 10^{-16}$, $r = 0$; $P = 0.09$, $r = -0.15$; $P < 2 \times 10^{-16}$ and $r = -0.08$; $P = 8 \times 10^{-10}$ for spermatogonia, spermatocyte, spermatid, and the whole testis. Besides, we obtained a clearly distinct distribution of the expression levels of CpG-rich and CpG-poor genes only when considering the spermatogonia data set (Fig. 3). This shows that taking into account all stages of spermatogenesis to measure germline expression can be misleading when trying to relate expression to CpG content. One explanation for the difference between spermatogonia and other stages is that spermatogonia undergo several mitotic divisions, whereas differentiation into spermatocytes and spermatids happens through meiosis, so that one can expect that it is mainly in spermatogonia that DNA substitutions are accumulated. Using the GNF Atlas data set, it was reported that 17% of CpG-rich genes that undergo *de novo* methylation correspond to testis specific genes [49]. It would also be interesting to reiterate this analysis using spermatogonia expression data and see if this proportion increases.

4. Organization of human genes in replication skew domains

4.1. *CpG-rich genes are over-represented close to the skew domain borders*

The distribution of promoter CpG enrichment of the 6875 genes found in the 926 replication skew domains previously delineated in the human genome, is quite similar to the one observed genome-wide (Fig. 2(a)). Yet, when separating genes according to their distance from their closest border, we found that mostly one population, the CpG-rich one, is present close to borders, this tendency weakening as the distance increases (Fig. 4). The proportion of genes close to borders that are CpG-rich reaches 80% (Fig. 5(a)). The average density in CpG-rich genes reaches ~7 genes per Mbp close to borders and

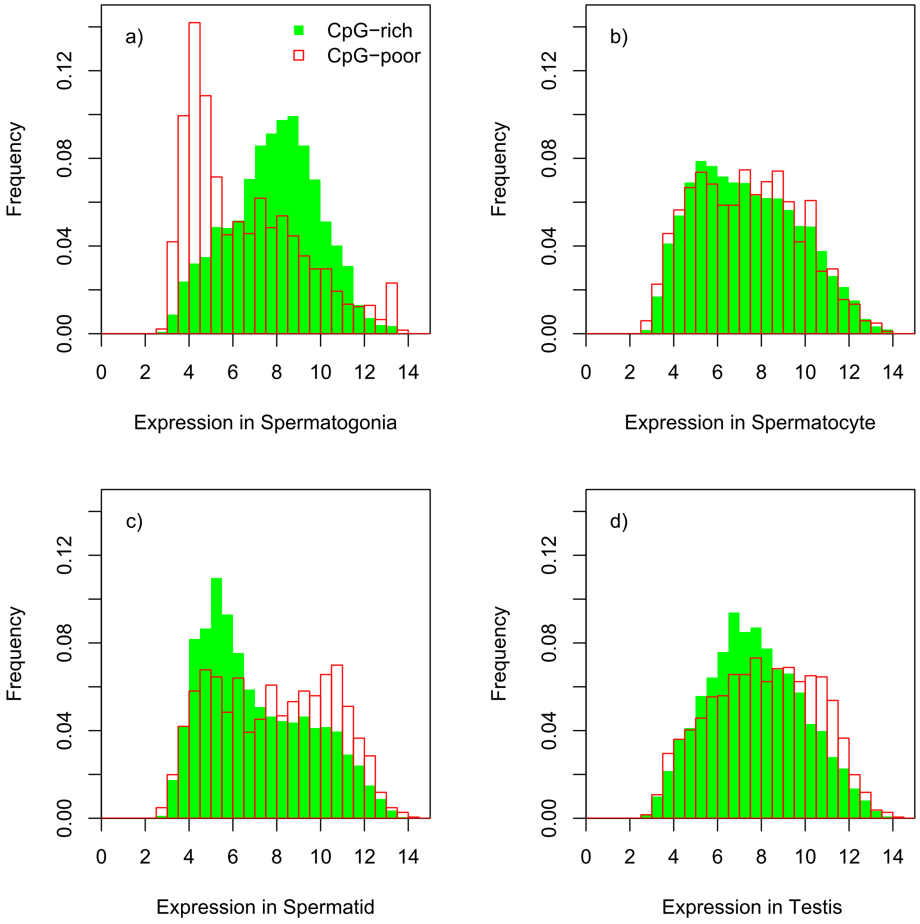

Fig. 3. Histogram of expression for the 5483 mouse genes for which germline expression was available (see Section 2), with CpG-rich genes shown in green (solid bars) and CpG-poor genes in red (open bars), in (a) spermatogonia, (b) spermatocyte, (c) spermatid and (d) whole testis.

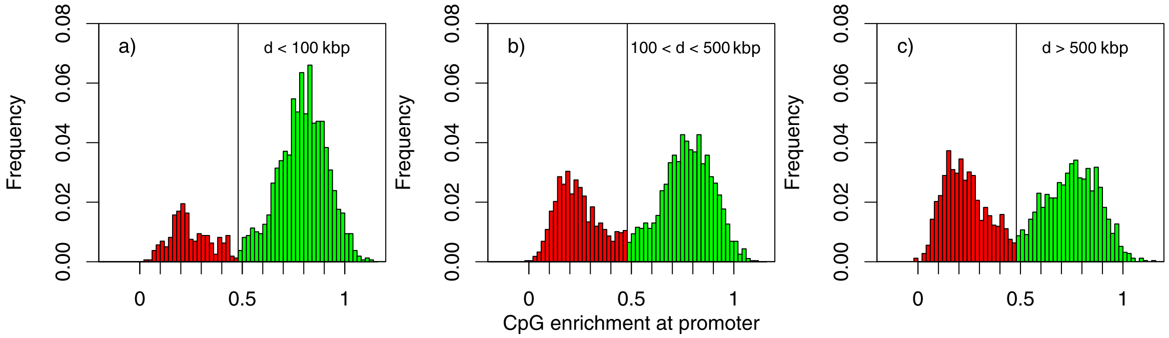

Fig. 4. Histogram of the CpG enrichment calculated in a 1 kbp window centered on the TSS of the 6875 human genes found in skew domains (see Section 2). Genes were separated in three subsets depending on their distance d from their closest domain border: (a) $d < 100$ kbp, (b) $100 < d < 500$ kbp, and (c) $d > 500$ kbp. A vertical bar is drawn at the threshold value (0.48) chosen to distinguish between CpG-rich genes (green, on the right of the bar) and CpG-poor genes (red, on the left of the bar).

decreases with the distance to borders while the average density in CpG-poor genes is almost constant in domains at a value of ∼1.5 genes per Mbp (Fig. 5(b)). These observations suggest that regions around borders are transcriptionally active in the germline. This is particularly patent in Fig. 6 where the 926 skew domains were centered and ordered vertically from the smallest (top) to the largest (bottom) and only gene promoters were represented. By a simple visual inspection, we recognize the edges of the skew domains from the local enrichment of GpG-rich gene promoters (Fig. 6(a)), whereas the CpG-poor gene promoters seem to be spatially distributed without preferential positioning relative to the replication domains (Fig. 6(b)). Similar results were obtained in the mouse genome in Figs. 6(c) and 6(d) which strongly suggest that

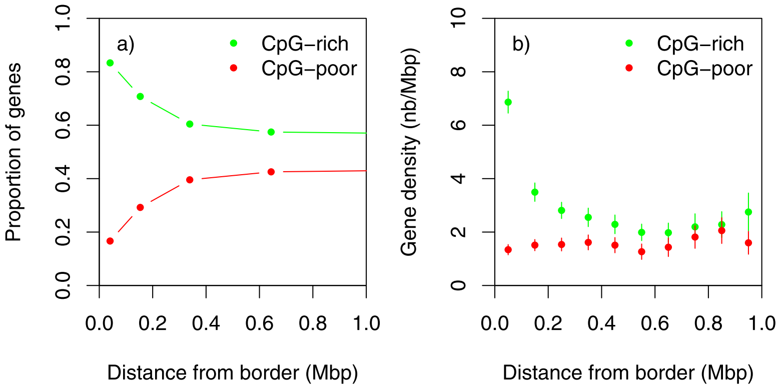

Fig. 5. (a) Proportion of genes in skew domains that belong to the CpG-rich (green) and to the CpG-poor (red) category versus the distance to the closest domain border. At each distance, the sum of the two curves is 1, so that an equiprobable repartition between the two categories gives a value of 0.5. (b) Average gene density (nb genes per Mbp) versus the distance to the closest border for CpG-rich genes (green) and CpG-poor genes (red).

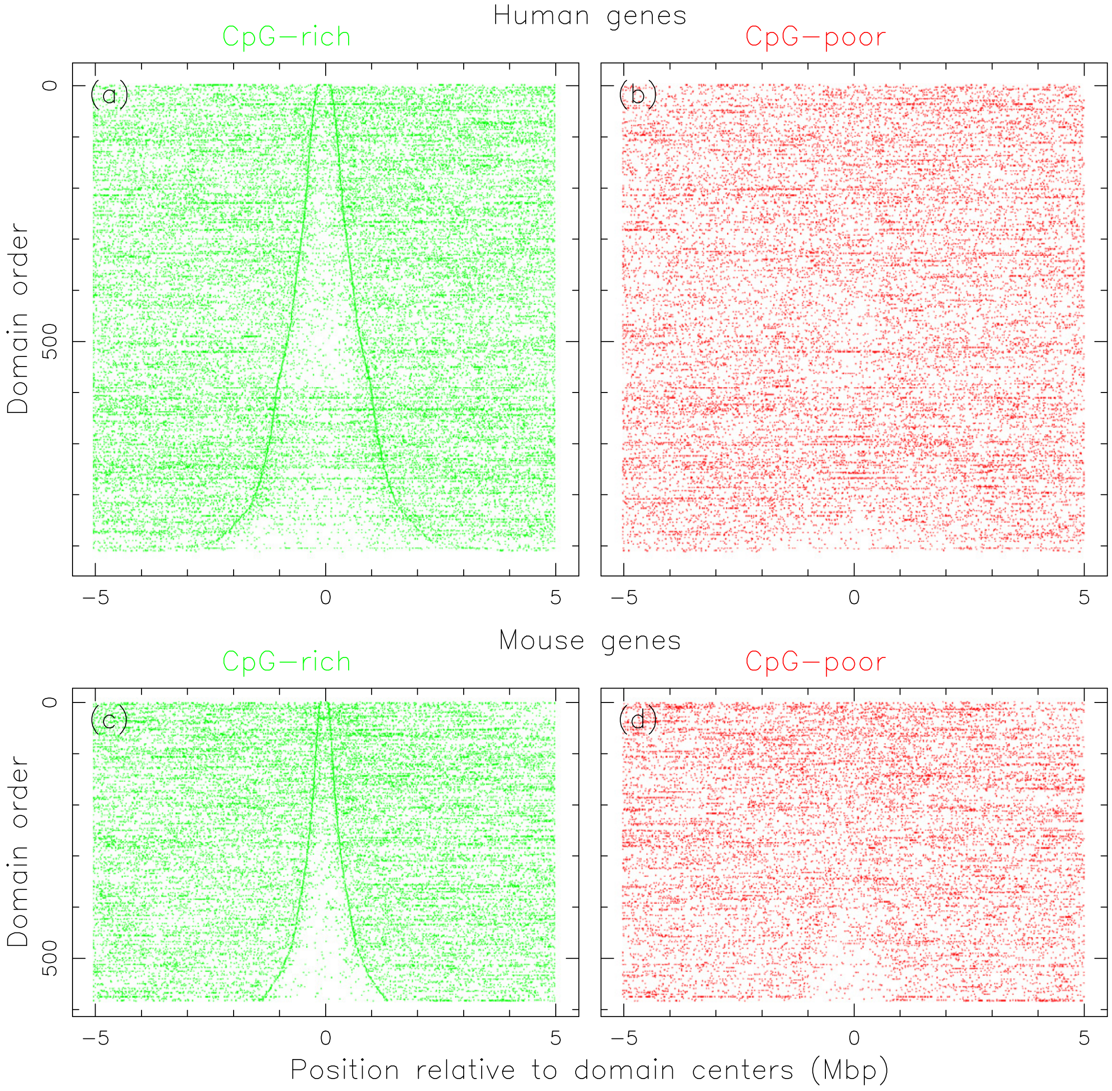

Fig. 6. Organization of CpG-rich (left panels) and CpG-poor (right panels) genes inside replication skew domains. Human genome (top panels): the 926 skew domains were centered and ordered vertically from the smallest (top) to the longest (bottom); the position of the gene promoters in each domains are figures as a dot along the corresponding horizontal line. Mouse genome (bottom panels): same representation of the positioning of mouse gene promoters inside the 585 replication skew N-domains identified in the mouse genome [13].

Table 1
Repartition of the 6875 human genes in skew domains into categories of CpG enrichment and expression breadth. Percents in exponent correspond to vertical frequencies and percents in indices correspond to horizontal frequencies. A gene is considered tissue-specific if it is expressed in 10 tissues or less and it is considered broadly expressed if it is expressed in 25 tissues or more (see Section 2).

	All	Broadly expressed	Tissue-specific
CpG-rich	$4481^{(65\%)}_{(100\%)}$	$405^{(91\%)}_{(9\%)}$	$1637^{(54\%)}_{(37\%)}$
CpG-poor	$2394^{(35\%)}_{(100\%)}$	$39^{(9\%)}_{(2\%)}$	$1396^{(46\%)}_{(58\%)}$
All	$6875^{(100\%)}_{(100\%)}$	$444^{(100\%)}_{(6\%)}$	$3033^{(100\%)}_{(44\%)}$

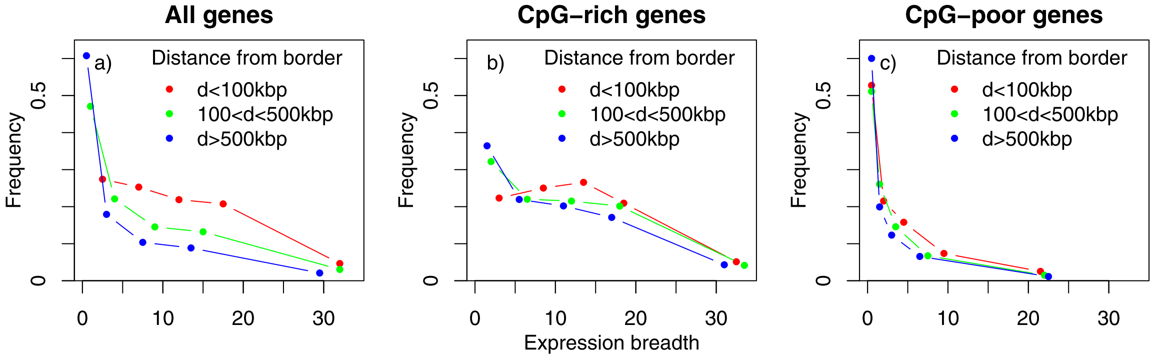

Fig. 7. Histogram of the expression breadths of (a) all human genes, (b) CpG-rich genes and (c) CpG-poor genes found in skew domains for three classes of distance d from their closest border: in red $d < 100$ kbp, in green $100 < d < 500$ kbp, and in blue $d > 500$ kbp. Each histogram has 5 bins with approximately the same number of genes.

the enrichment of genes expressed in the germline nearby master origins that border the replication skew domains is a general feature common to mammalian genomes.

Remarks. It was already noted that CpG-rich genes tend to be seated in high GC isochores [40,53]. Indeed, skew domains borders do not correspond to high GC isochores, with the average GC content around borders being ∼41% which corresponds to the genome average [22,54].

4.2. *Distribution of expression breadth inside replication skew domains*

As reported in [15], the average expression breadth around skew domains borders is high and decreases towards the domain center. Indeed, the relation between a high expression breadth and the presence of a CGI-promoter is non-reciprocal. If housekeeping genes are almost always CpG-rich, the opposite is not true as all expression breadths are represented in the CpG-rich category (Table 1). We defined as broadly expressed genes that are expressed in 25 tissues or more (see Section 2). We found that 90% of these broadly expressed genes are CpG-rich, which is consistent with other analyses that reported proportions of CpG-rich housekeeping genes between 90% and 100% [38,40,55].

In the CpG-rich category itself, do broadly expressed genes and tissue-specific genes have similar CpG enrichment? In the CpG-rich category, broadly expressed genes tend to be more CpG-rich than tissue-specific ones. The difference in CpG enrichment, even though much smaller than the one between CpG-rich and CpG-poor genes, is statistically significant (Wilcoxon rank sum test $P = 2 \times 10^{-12}$). Hence, we will be cautious when relating our observations to either CpG enrichment at the promoter or expression breadth, as the two are not completely independent.

Given the link we found between CpG enrichment at the promoter and distance to the border (Figs. 5 and 6), we wished to verify that the higher expression breadth close to the border was not in fact subsequent to the presence of more genes of the CpG-rich type. We therefore estimated how the distribution of expression breadths varies with the distance to the border (i) for all genes, (ii) for CpG-rich genes and (iii) for CpG-poor genes. If the over-representation of housekeeping genes close to borders is only a consequence of the over-representation of CpG-rich genes, we expect to find the same distribution of breadths at all distances in both the CpG-rich and the CpG-poor categories. This is close to what we found: the variation in the distribution of breadths with the distance that is observed for all genes (Fig. 7(a)) is greatly reduced once separating genes into a CpG-rich and a CpG-poor category (Figs. 7(b) and 7(c)).

It seems nevertheless that CpG-rich genes found at less than 100 kbp from their closest border (shown in red in Fig. 7(b)) have a higher expression breadth than other CpG-rich genes found further away. Indeed, Wilcoxon rank sum tests of the distributions of expression breadths of CpG-rich genes comparing the three classes of distances from the border (Fig. 7(b))

Table 2
For the 4355 human genes in skew domains found at a distance $d \leqslant 500$ kbp from their closest border: repartition into 12 categories according to their orientation with respect to their closest border, their CpG enrichment at the promoter and their expression breadth. Percents in exponent correspond to vertical frequencies and percents in indices correspond to horizontal frequencies. A gene is considered tissue-specific if it is expressed in 10 tissues or less and it is considered broadly expressed if it is expressed in 25 tissues or more (see Section 2). B-div. stands for border-divergent and B-conv. stands for border-convergent. BE stands for broadly expressed and TS stands for tissue-specific.

	All	CpG-rich			CpG-poor		
		All	BE	TS	All	BE	TS
B-div.	$2866^{(66\%)}_{(100\%)}$	$2162^{(70\%)}_{(75\%)}$	$208^{(72\%)}_{(7\%)}$	$727^{(66\%)}_{(25\%)}$	$704^{(56\%)}_{(25\%)}$	$13^{(59\%)}_{(0\%)}$	$422^{(57\%)}_{(15\%)}$
B-conv.	$1489^{(34\%)}_{(100\%)}$	$927^{(30\%)}_{(62\%)}$	$82^{(28\%)}_{(6\%)}$	$380^{(34\%)}_{(26\%)}$	$562^{(44\%)}_{(38\%)}$	$9^{(41\%)}_{(0\%)}$	$324^{(43\%)}_{(22\%)}$
All	$4355^{(100\%)}_{(100\%)}$	$3089^{(100\%)}_{(71\%)}$	$290^{(100\%)}_{(7\%)}$	$1107^{(100\%)}_{(25\%)}$	$1266^{(100\%)}_{(29\%)}$	$22^{(100\%)}_{(0\%)}$	$746^{(100\%)}_{(17\%)}$

show that the three classes distributions are significantly different. CpG-rich genes at a distance $d < 100$ kbp have a different distribution of expression breadths than CpG-rich genes at a distance $100 < d < 500$ kbp ($P = 6.9 \times 10^{-4}$) and than CpG-rich genes at a distance $d > 500$ kbp ($P = 4.7 \times 10^{-8}$). Even though a residual effect of the distance from the border can be observed, our analysis nevertheless shows that the average decrease in the expression breadth from borders is in a great part explained by the differential distribution of CpG-rich and CpG-poor genes around skew domains borders.

Whereas broadly expressed genes are almost exclusively CpG-rich, tissue-specific genes can be either CpG-rich or CpG-poor (Table 1). It appears that all CpG-rich genes, regardless of their expression breadth, are over-represented close to borders. This result can be linked to the observation that the frequency of tissue-specific genes that are associated to a CGI is higher in regions of high GC content, whereas the frequency of housekeeping genes that are associated to a CGI presents no link with isochores [40]. As previously noted, skew domains borders do not correspond to GC-rich isochores, yet they share some characteristics with these regions in terms of open chromatin, early replication timing, high gene density and over-representation of CpG-rich genes [16,22,54].

The separation of genes in these two CpG-rich and CpG-poor categories helped us making better sense of the previous results and our observations put forward the CpG enrichment at the promoter as a key parameter to understand the distribution of genes around skew domains borders. We will see in the following analyses that it is also a key parameter to explain other aspects of gene organization around skew domains borders such as orientation bias and gene length.

5. Gene orientation and length inside replication skew domains

5.1. CpG-rich genes have the strongest orientation bias

It was reported that close to borders genes are majoritarily oriented divergently from the border [15]. Given the importance of separating between CpG-rich genes and CpG-poor genes, we asked if the orientation bias of genes was true for both categories or not.

It appears again that it is mainly CpG-rich genes that have an orientation bias, with the proportion of border-divergent genes reaching 80% close to the border, whereas in CpG-poor genes this proportion reaches only 60%. A summary table (Table 2) with the number of border-divergent and border-convergent genes clearly shows that the orientation bias is mainly found in the CpG-rich category with an average ratio of 70% for CpG-rich genes and 56% for CpG-poor genes. Inversely, when looking at how border-divergent genes are distributed among the two categories of genes, it appears that 75% belong to the CpG-rich category, whereas the bias is less pronounced for border-convergent genes (62% vs 38%). We also observed that the proportion of broadly expressed or tissue-specific genes is similar for border-divergent genes and border-convergent genes when considering either CpG-rich or CpG-poor genes.

In contrast, we found that the expression breadth parameter is not related to the orientation bias. Indeed, when separating the CpG-rich category between broadly expressed genes and tissue-specific genes, both subsets have comparable orientation bias (data not shown).

In summary, our analyses show that (i) genes close to skew domain borders are mainly CpG-rich and border-divergent, and that (ii) these two properties are not independent as, close to borders, CpG-rich genes have a high probability to be border-divergent which is not the case for CpG-poor genes.

5.2. Border-divergent CpG-rich genes are longer than expected

CpG enrichment at the promoter is a key parameter to describe the distribution, activity and regulation of human genes. We have found that it is also the case when describing gene organization around skew domains borders in terms of gene density and orientation. Is it also true in terms of gene length? It was reported that, close to skew domains borders, border-divergent genes are on average three times longer that border-convergent genes, and that the average gene length decreases with the distance to the border [15].

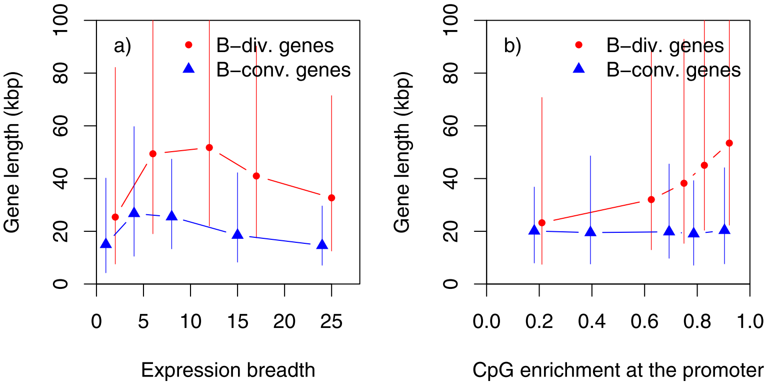

Fig. 8. For human genes in skew domains at less than 500 kbp from their closest border, median gene length in kbp vs (a) expression breadth and (b) CpG enrichment at the promoter for border-divergent genes (red dot) and border-convergent ones (blue triangle). Genes were ordered according to increasing values of expression breadth in (a) and CpG enrichment at the promoter in (b) and separated in 7 even groups for which median values of gene length and of expression breadth in (a) and CpG enrichment at the promoter in (b) were calculated. Vertical bars extend from the first quartile (25% of the distribution) to the third quartile (75% of the distribution). B-div. stands for border-divergent and B-conv. stands for border-convergent.

5.2.1. Genome-wide analysis

Genome-wide studies have related the length of human genes to expression breadth and expression level. It was shown that highly expressed genes and housekeeping genes are shorter than average [56–58]. It was proposed that their shorter length results from a selective pressure to save time and energy during transcription [56]. Actually, genes tend to be shorter in GC-rich regions than in GC-poor regions [5,59], and the probability for genes to be CpG-rich is higher in GC-rich regions [53]. As previously mentioned, housekeeping genes, which are almost always CpG-rich, are also known to be short. One could therefore think that CpG-rich genes are shorter than CpG-poor genes. It is the contrary however that is observed. It was reported in [50] that intron length was higher for CpG-rich genes. We similarly found that gene length increases with CpG enrichment at the promoter. The median gene length reaches ∼15 kbp for CpG-poor genes, whereas in the CpG-rich category, it increases from ∼20 kbp for genes with a CpG enrichment at the promoter of ∼0.6 to ∼35 kbp for genes with a CpG enrichment at the promoter of ∼0.9. The main difference in gene length is found when comparing the CpG-rich and the CpG-poor categories, but there is also, inside the CpG-rich category, some increase in gene length with CpG enrichment.

Indeed, when considering expression breadth, it turns out that CpG enrichment at the promoter is a key parameter to describe the length of genes belonging to the “tissue-specific” category and that are not broadly expressed. Importantly, the length of broadly expressed genes is definitely smaller than what is expected given their belonging to the CpG-rich category, which is compatible with an effect of selection for economy in transcriptional cost [56] that would be more pronounced in this subset of genes.

5.2.2. Analysis inside skew domains

When concentrating our analysis of gene length inside replication skew domains only, we could take into account the orientation and the distance of genes with respect to their closest domain border. Indeed, it was reported that close to borders, border-divergent genes are longer that border-convergent ones [15]. Can the greater length of border-divergent genes be accounted for by the over-representation of CpG-rich genes?

As shown in Fig. 8(a), we actually recovered the relation between gene length and expression breadth observed genome-wide, both for border-divergent and border-convergent genes. For a given expression breadth, border-convergent genes tend to be smaller than border-divergent ones (Fig. 8(a)). However, we recovered the positive relation between CpG enrichment at the promoter and gene length only for border-divergent genes (Fig. 8(b)). The difference in length between border-divergent and border-convergent genes does not thus simply reflect a difference in their CpG enrichment at the promoter. Border-divergent genes are not only longer than border-convergent ones, but when comparing them to the genome average they also appear longer than what is expected given their CpG enrichment at the promoter. For a CpG enrichment of 0.85–0.95, the median gene length in the genome is ∼30–35 kbp whereas for border-divergent genes, it reaches ∼45–50 kbp, suggesting that they constitute a particular subset of long human genes. We performed a Wilcoxon rank sum test to compare the distribution of gene lengths between all human CpG-rich genes, and CpG-rich border-divergent genes and CpG-rich border-convergent genes found in skew domains at less than 500 kbp from their closest border. We found in all two by two comparisons that the length distributions presented statistically significant differences. CpG-rich border-divergent genes have a different length distribution than all CpG-rich genes ($P < 2 \times 10^{-16}$), CpG-rich border-convergent genes have a different length distribution than all CpG-rich genes ($P = 1.7 \times 10^{-9}$), and CpG-rich border-divergent genes have a different length distribution than CpG-rich border-convergent genes ($P < 2 \times 10^{-16}$). The median gene length of CpG-rich genes decreases from ∼60 kbp at the border to ∼20 kbp at a distance of 1 Mbp from the border, while CpG-poor genes have a median length of ∼25 kbp close to the border and ∼10 kbp at a distance of 1 Mbp. CpG-rich genes close to borders are longer than what is expected given their CpG enrichment and are also majoritarily border-divergent. When further separating genes into four categories depending both on their CpG enrichment at the promoter and on their orientation with respect to their closest border, we found that close to skew domains borders, the orientation factor is determinant, as border-convergent genes are short, whether CpG-rich or CpG-poor, whereas border-divergent genes are long. The CpG enrichment also matters for border-divergent genes as CpG-rich border-divergent genes are significantly longer than CpG-poor border-divergent genes close to borders (Fig. 8(b)). If the CpG enrichment at the promoter appears as a key factor, our observations show that it does not explain the difference in gene length between border-divergent and border-convergent genes (Fig. 8(a)) and put forward the orientation of the gene with respect to its closest border as an additional factor to describe variations in gene length.

It was proposed that the predominant orientation of genes divergently from skew domains borders and the smaller gene length observed in border-convergent genes could result from a selection to minimize collisions between the replication fork emanating from the border and the RNA polymerases transcribing genes [15]. This hypothesis was however challenged by the observation that genes that are transcriptionally active during the S phase, which is when the collisions can occur, and genes that are inactive during the S phase present comparable orientation biases with respect to their closest border [46]. An analogous explanation involving transcription could also be considered [60]. If initiating transcription can transcribe neighboring genes, it is an advantage for genes (i) to be clustered according to their expression pattern, (ii) to be arranged with a coordinated orientation on the neighborhood of highly used transcription initiation sites. The presence of a long CpG-rich border-divergent gene close to skew domains borders could result from a selection to favor the complete transcription of this gene (i) by placing its promoter in the burst of open and accessible chromatin structure surrounding the master replication origins at skew domain borders [14,22] and (ii) by not placing other long genes next to it particularly oriented convergently. The absence of long border-convergent genes could minimize the pervasive antisense transcription of border-divergent genes [29].

6. Conclusion and perspectives

Quite often, large expression breadth is taken as proxy to define essentiality and broadly expressed genes are named “housekeeping”. Housekeeping genes are expected to be under selective pressure to be expressed the right way. It was for example proposed that the clustering of housekeeping genes in regions of open chromatin could be due to selection to reduce transcriptional noise [61]. The fact that we find that all CpG-rich genes, whether tissue-specific or broadly expressed, are over-represented close to domains borders (Figs. 5 and 6) raises the question of whether (i) there was selection involved to have an over-representation of housekeeping genes close to skew domains borders, (ii) selection favored the over-representation of CpG-rich genes expressed in the germline, or (iii) there was no selection involved and for instance, what played a role in defining skew domains borders is the presence of CpG-rich genes. Genes expressed in the germline, some of which become repressed in somatic cells with *de novo* methylation [49,62,63], are not expected to be under the same kind of selective pressure as housekeeping genes. Yet, this is difficult to assert as (i) what happens in the germline is determinant for what will happen next in the life of the organism and (ii) germ cells undergo many cell divisions thus enabling some cell selection. Therefore germline expression could be under a stronger selective constraint than expression in the liver or in the brain for example. For a discussion on the patterns of germline mutations, see [64]. It would be thus interesting to distinguish between skew domain borders that are surrounded by housekeeping genes and those that are surrounded by tissue-specific CpG-rich genes and compare how they differ in terms of markers of open chromatin and in terms of replication timing. An additional difference between CpG-rich tissue-specific genes and housekeeping genes is that the promoters of CpG-rich tissue-specific genes are likely to be methylated in many tissues, where they are not expressed. Could this have an effect on skew domains borders around which CpG-rich genes are tissue-specific rather than broadly expressed? It seems unlikely that the methylation of CpG-rich tissue-specific promoters could have an effect on the overall chromatin opening of the region. Indeed, the proportion of CGIs that undergo *de novo* methylation is small, with only 6–8% of human CGIs becoming methylated in somatic tissues, and preferably those that are intergenic or intragenic [65].

The recent availability of replication timing data in different cell types in human [66–69] and mouse [70–73] genomes has provided the opportunity to study the possibility of coordinated changes in replication timing and gene expression during differentiation. In a previous work [18], we have shown that for seven different human cell types including ES cells, somatic and HeLa cells, about half of the human genome can be divided in replication domains where the mean replication timing has a characteristic U-shape with early initiation zones at the borders and late replication at centers. Remarkably, a significant overlap is observed between these replication U-domains in different cell lines and also with germline skew N-domains. By further demonstrating that the average replication fork polarity is directly reflected by both the compositional skew and the derivative of the replication timing profile [18–20,74,75], we have argued that the experimental observation that this derivative displays an N-shape in replication U-domains sustains the existence of large-scale gradients of replication fork polarity in somatic and germline cells. The analysis of chromatin interaction (4C, Hi-C) and chromatin marker data has further revealed that these replication U-domains indeed correspond to high-order self-interacting chromatin units [18,76]. The compartmentalization of the human genome into these structural and functional domains provides new perspectives in the modeling of the spatio-temporal replication timing program in relation with transcription and chromatin environment [18,77,78]. Further analyses of gene expression data inside replication U-domains in different cell types are on the way with the specific goal to confirm that the enrichment of expressed genes nearby the early initiation zones bordering these domains is a general signature of the coordinated regulation of transcription and replication by the chromatin tertiary structure.

Acknowledgements

We thank C.-L. Chen, A. Goldar, O. Hyrien, Y. d'Aubenton-Carafa and C. Thermes for interesting discussions. This work was supported by the ANR (REFOPOL, ANR 10 BLAN 1615).

References

[1] B. Alberts, Essential Cell Biology: An Introduction to the Molecular Biology of the Cell, Garland Publishing, 1998.

[2] E. Couturier, E.P.C. Rocha, Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes, Mol. Microbiol. 59 (2006) 1506–1518.

[3] E.P.C. Rocha, A. Danchin, Gene essentiality determines chromosome organisation in bacteria, Nucleic Acids Res. 31 (2003) 6570–6577.

[4] E.P.C. Rocha, A. Danchin, Essentiality, not expressiveness, drives gene-strand bias in bacteria, Nat. Genet. 34 (2003) 377–378.

[5] E.S. Lander, et al., Initial sequencing and analysis of the human genomes, Nature 409 (2001) 860–921.

[6] Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse genome, Nature 420 (2002) 520–562.

[7] G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, F. Rodier, The mosaic genome of warm-blooded vertebrates, Science 228 (1985) 953–958.

[8] D. Mouchiroud, G. D'Onofrio, B. Aïssani, G. Macaya, C. Gautier, G. Bernardi, The distribution of genes in the human genome, Gene 100 (1991) 181–187.

[9] G. Bernardi, Isochores and the evolutionary genomics of vertebrates, Gene 241 (2000) 3–17.

[10] M.J. Lercher, A.O. Urrutia, A. Pavlicek, L.D. Hurst, A unification of mosaic structures in the human genome, Hum. Mol. Genet. 12 (2003) 2411–2415.

[11] E.-B. Brodie of Brodie, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, From DNA sequence analysis to modeling replication in the human genome, Phys. Rev. Lett. 94 (2005) 248103.

[12] M. Touchon, S. Nicolay, B. Audit, E.-B. Brodie of Brodie, Y. d'Aubenton-Carafa, A. Arneodo, C. Thermes, Replication-associated strand asymmetries in mammalian genomes: Toward detection of replication origins, Proc. Natl. Acad. Sci. USA 102 (2005) 9836–9841.

[13] A. Baker, S. Nicolay, L. Zaghloul, Y. d'Aubenton-Carafa, C. Thermes, B. Audit, A. Arneodo, Wavelet-based method to disentangle transcription- and replication-associated strand asymmetries in mammalian genomes, Appl. Comput. Harmon. Anal. 28 (2010) 150–170.

[14] A. Arneodo, C. Vaillant, B. Audit, F. Argoul, Y. d'Aubenton-Carafa, C. Thermes, Multi-scale coding of genomic information: From DNA sequence to genome structure and function, Phys. Rep. 498 (2011) 45–188.

[15] M. Huvet, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, A. Arneodo, C. Thermes, Human gene organization driven by the coordination of replication and transcription, Genome Res. 17 (2007) 1278–1285.

[16] B. Audit, S. Nicolay, M. Huvet, M. Touchon, Y. d'Aubenton Carafa, C. Thermes, A. Arneodo, DNA replication timing data corroborate in silico human replication origin predictions, Phys. Rev. Lett. 99 (2007) 248102.

[17] C.-L. Chen, L. Duquenne, B. Audit, G. Guilbaud, A. Rappailles, A. Baker, M. Huvet, Y. d'Aubenton Carafa, O. Hyrien, A. Arneodo, C. Thermes, Replication-associated mutational asymmetry in the human genome, Mol. Biol. Evol. 28 (2011) 2327–2337.

[18] A. Baker, B. Audit, C.-L. Chen, B. Moindrot, A. Leleu, G. Guilbaud, A. Rappailles, C. Vaillant, A. Goldar, F. Mongelard, Y. d'Aubenton-Carafa, O. Hyrien, C. Thermes, A. Arneodo, Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines, PLoS Comput. Biol. 8 (2012) e1002443.

[19] A. Baker, H. Julienne, C.-L. Chen, B. Audit, Y. d'Aubenton Carafa, C. Thermes, A. Arneodo, Linking the DNA strand asymmetry to the spatio-temporal replication program. I. About the role of the replication fork polarity in genome evolution, Eur. Phys. E 35 (2012) 92.

[20] A. Baker, C.-L. Chen, H. Julienne, B. Audit, Y. d'Aubenton Carafa, C. Thermes, A. Arneodo, Linking the DNA strand asymmetry to the spatio-temporal replication program. II. Accounting for neighbor-dependent substitution rates, Eur. Phys. E (2012), in press.

[21] P. St-Jean, C. Vaillant, B. Audit, A. Arneodo, Spontaneous emergence of sequence-dependent rosettelike folding of chromatin fiber, Phys. Rev. E 77 (2008) 061923.

[22] B. Audit, L. Zaghloul, C. Vaillant, G. Chevereau, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Open chromatin encoded in DNA sequence is the signature of "master" replication origins in human cells, Nucleic Acids Res. 37 (2009) 6064–6075.

[23] D. Karolchik, R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, W.J. Kent, The UCSC genome browser database, Nucleic Acids Res. 31 (2003) 51–54.

[24] M. Touchon, S. Nicolay, A. Arneodo, Y. d'Aubenton-Carafa, C. Thermes, Transcription-coupled TA and GC strand asymmetries in the human genome, FEBS Lett. 555 (2003) 579–582.

[25] M. Touchon, A. Arneodo, Y. d'Aubenton-Carafa, C. Thermes, Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes, Nucleic Acids Res. 32 (2004) 4969–4978.

[26] S. Nicolay, E.-B. Brodie of Brodie, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, From scale invariance to deterministic chaos in DNA sequences: Towards a deterministic description of gene organization in the human genome, Physica A 342 (2004) 270–280.

[27] S. Nicolay, F. Argoul, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Low frequency rhythms in human DNA sequences: A key to the organization of gene location and orientation? Phys. Rev. Lett. 93 (2004) 108101.

[28] S. Nicolay, E.B. Brodie of Brodie, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Bifractality of human DNA strand-asymmetry profiles results from transcription, Phys. Rev. E 75 (2007) 032902.

[29] L. Zaghloul, Transcriptional activity, chromatin state and replication timing in domains of compositional skew in the human genome, Ph.D. thesis, Université de Lyon, Ecole Normale Supérieure de Lyon, 2009.

[30] M. Sémon, J.R. Lobry, L. Duret, No evidence for tissue-specific adaptation of synonymous codon usage in humans, Mol. Biol. Evol. 23 (2006) 523–529.

[31] M. Sémon, L. Duret, Evolutionary origin and maintenance of coexpressed gene clusters in mammals, Mol. Biol. Evol. 23 (2006) 1715–1723.

[32] A.I. Su, T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M.P. Cooke, J.R. Walker, J.B. Hogenesch, A gene atlas of the mouse and human protein-encoding transcriptomes, Proc. Natl. Acad. Sci. USA 101 (2004) 6062–6067.

[33] F. Chalmel, A.D. Rolland, C. Niederhauser-Wiederkehr, S.S.W. Chung, P. Demougin, A. Gattiker, J. Moore, J.-J. Patard, D.J. Wolgemuth, B. Jégou, M. Primig, The conserved transcriptome in human and rodent male gametogenesis, Proc. Natl. Acad. Sci. USA 104 (2007) 8346–8351.

[34] M.M. Suzuki, A. Bird, DNA methylation landscapes: Provocative insights from epigenomics, Nat. Rev. Genet. 9 (2008) 465–476.

[35] D.N. Cooper, M.H. Taggart, A.P. Bird, Unmethylated domains in vertebrate DNA, Nucleic Acids Res. 11 (1983) 647–658.

[36] A. Bird, M. Taggart, M. Frommer, O.J. Miller, D. Macleod, A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA, Cell 40 (1985) 91–99.

[37] M. Gardiner-Garden, M. Frommer, CpG islands in vertebrate genomes, J. Mol. Biol. 196 (1987) 261–282.

[38] F. Antequera, A. Bird, Number of CpG islands and genes in human and mouse, Proc. Natl. Acad. Sci. USA 90 (1993) 11995–11999.

[39] D. Macleod, R.R. Ali, A. Bird, An alternative promoter in the mouse major histocompatibility complex class II I-Abeta gene: Implications for the origin of CpG islands, Mol. Cell. Biol. 18 (1998) 4433–4443.

[40] L. Ponger, L. Duret, D. Mouchiroud, Determinants of CpG islands: Expression in early embryo and isochore structure, Genome Res. 11 (2001) 1854–1860.

[41] F. Antequera, Structure, function and evolution of CpG island promoters, Cell. Mol. Life Sci. 60 (2003) 1647–1658.

[42] S. Delgado, M. Gómez, A. Bird, F. Antequera, Initiation of DNA replication at CpG islands in mammalian chromosomes, EMBO J. 17 (1998) 2426–2435.

[43] F. Antequera, A. Bird, CpG islands as genomic footprints of promoters that are associated with replication origins, Curr. Biol. 9 (1999) R661–R667.

[44] J.-C. Cadoret, F. Meisch, V. Hassan-Zadeh, I. Luyten, C. Guillet, L. Duret, H. Quesneville, M.-N. Prioleau, Genome-wide studies highlight indirect links between human replication origins and gene regulation, Proc. Natl. Acad. Sci. USA 105 (2008) 15837–15842.

[45] J. Sequeira-Mendes, R. Diaz-Uriarte, A. Apedaile, D. Huntley, N. Brockdorff, M. Gomez, Transcription initiation activity sets replication origin efficiency in mammalian cells, PLoS Genet. 5 (2009) e1000446.

[46] A. Necsulea, C. Guillet, J.-C. Cadoret, M.-N. Prioleau, L. Duret, The relationship between DNA replication and human genome organization, Mol. Biol. Evol. 26 (2009) 729–741.

[47] D. Takai, P.A. Jones, Comprehensive analysis of CpG islands in human chromosomes 21 and 22, Proc. Natl. Acad. Sci. USA 99 (2002) 3740–3745.

[48] S. Saxonov, P. Berg, D.L. Brutlag, A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters, Proc. Natl. Acad. Sci. USA 103 (2006) 1412–1417.

[49] M. Weber, I. Hellmann, M.B. Stadler, L. Ramos, S. Pääbo, M. Rebhan, D. Schübeler, Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome, Nat. Genet. 39 (2007) 457–466.

[50] C.S.M. Tang, R.J. Epstein, A structural split in the human genome, PLoS One 2 (2007) e603.

[51] F. Mohn, D. Schübeler, Genetics and epigenetics: Stability and plasticity during cellular differentiation, Trends Genet. 25 (2009) 129–136.

[52] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C.A.M. Semple, M.S. Taylor, P.G. Engström, M.C. Frith, A.R.R. Forrest, W.B. Alkema, S.L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S.M. Grimmond, C.A. Wells, V. Orlando, C. Wahlestedt, E.T. Liu, M. Harbers, J. Kawai, V.B. Bajic, D.A. Hume, Y. Hayashizaki, Genome-wide analysis of mammalian promoter architecture and evolution, Nat. Genet. 38 (2006) 626–635.

[53] B. Aïssani, G. Bernardi, CpG islands, genes and isochores in the genomes of vertebrates, Gene 106 (1991) 185–195.

[54] C. Lemaitre, L. Zaghloul, M.-F. Sagot, C. Gautier, A. Arneodo, E. Tannier, B. Audit, Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation, BMC Genomics 10 (2009) 335.

[55] F. Larsen, G. Gundersen, R. Lopez, H. Prydz, CpG islands as gene markers in the human genome, Genomics 13 (1992) 1095–1107.

[56] C.I. Castillo-Davis, S.L. Mekhedov, D.L. Hartl, E.V. Koonin, F.A. Kondrashov, Selection for short introns in highly expressed genes, Nat. Genet. 31 (2002) 415–418.

[57] E. Eisenberg, E.Y. Levanon, Human housekeeping genes are compact, Trends Genet. 19 (2003) 362–365.

[58] A.O. Urrutia, L.D. Hurst, The signature of selection mediated by expression on human genes, Genome Res. 13 (2003) 2260–2264.

[59] L. Duret, D. Mouchiroud, C. Gautier, Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores, J. Mol. Evol. 40 (1995) 308–317.

[60] M. Ebisuya, T. Yamamoto, M. Nakajima, E. Nishida, Ripples from neighbouring transcription, Nat. Cell Biol. 10 (2008) 1106–1113.

[61] N.N. Batada, L.D. Hurst, Evolution of chromosome organization driven by selection for reduced gene expression noise, Nat. Genet. 39 (2007) 945–949.

[62] F. Mohn, M. Weber, M. Rebhan, T.C. Roloff, J. Richter, M.B. Stadler, M. Bibel, D. Schübeler, Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors, Mol. Cell 30 (2008) 755–766.

[63] C.R. Farthing, G. Ficz, R.K. Ng, C.-F. Chan, S. Andrews, W. Dean, M. Hemberger, W. Reik, Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes, PLoS Genet. 4 (2008) e1000116.

[64] N. Arnheim, P. Calabrese, Understanding what determines the frequency and pattern of human germline mutations, Nat. Rev. Genet. 10 (2009) 478–488.

[65] R. Illingworth, A. Kerr, D. Desousa, H. Jorgensen, P. Ellis, J. Stalker, D. Jackson, C. Clee, R. Plumb, J. Rogers, S. Humphray, T. Cox, C. Langford, A. Bird, A novel CpG island set identifies tissue-specific methylation at developmental gene loci, PLoS Biol. 6 (2008) e22.

[66] K. Woodfine, D.M. Beare, K. Ichimura, S. Debernardi, A.J. Mungall, H. Fiegler, V.P. Collins, N.P. Carter, I. Dunham, Replication timing of human chromosome 6, Cell Cycle 4 (2005) 172–176.

[67] R. Desprat, D. Thierry-Mieg, N. Lailler, J. Lajugie, C. Schildkraut, J. Thierry-Mieg, E.E. Bouhassira, Predictable dynamic program of timing of DNA replication in human cells, Genome Res. 19 (2009) 2288–2299.

[68] C.-L. Chen, A. Rappailles, L. Duquenne, M. Huvet, G. Guilbaud, L. Farinelli, B. Audit, Y. d'Aubenton-Carafa, A. Arneodo, O. Hyrien, C. Thermes, Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes, Genome Res. 4 (2010) 447–457.

[69] R.S. Hansen, S. Thomas, R. Sandstrom, T.K. Canfield, R.E. Thurman, M. Weaver, M.O. Dorschner, S.M. Gartler, J.A. Stamatoyannopoulos, Sequencing newly replicated DNA reveals widespread plasticity in human replication timing, Proc. Natl. Acad. Sci. USA 107 (2010) 139–144.

[70] S. Farkash-Amar, D. Lipson, A. Polten, A. Goren, C. Helmstetter, Z. Yakhini, I. Simon, Global organization of replication time zones of the mouse genome, Genome Res. 18 (2008) 1562–1570.

[71] I. Hiratani, T. Ryba, M. Itoh, T. Yokochi, M. Schwaiger, C.-W. Chang, Y. Lyou, T.M. Townes, D. Schubeler, D.M. Gilbert, Global reorganization of replication domains during embryonic stem cell differentiation, PLoS Biol. 6 (2008) e245.

[72] T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T.C. Schulz, A.J. Robins, S. Dalton, D.M. Gilbert, Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types, Genome Res. 20 (2010) 761–770.

[73] I. Hiratani, T. Ryba, M. Itoh, J. Rathjen, M. Kulik, B. Papp, E. Fussner, D.P. Bazett-Jones, K. Plath, S. Dalton, P.D. Rathjen, D.M. Gilbert, Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis, Genome Res. 20 (2010) 155–169.

[74] A. Baker, B. Audit, S.C.-H. Yang, J. Bechhoefer, A. Arneodo, Inferring where and when replication initiates from genome-wide replication timing data, Phys. Rev. Lett. 108 (2012) 268101.

[75] B. Audit, A. Baker, C.-L. Chen, A. Rappailles, G. Guilbaud, H. Julienne, A. Goldar, Y. d'Aubenton-Carafa, O. Hyrien, C. Thermes, A. Arneodo, Multi-scale analysis of genome wide replication timing profiles using a wavelet-based signal-processing algorithm, Nat. Protoc. (2012), in press.

[76] B. Moindrot, B. Audit, P. Klous, A. Baker, C. Thermes, W. de Laat, P. Bouvet, F. Mongelard, A. Arneodo, 3D chromatin conformation correlates with replication timing and is conserved in resting cells, Nucleic Acids Res. 40 (2012) 9470–9481.

[77] O. Hyrien, A. Goldar, Mathematical modelling of eukaryotic DNA replication, Chromosome Res. 18 (2010) 147–161.

[78] G. Guilbaud, A. Rappailles, A. Baker, C.-L. Chen, A. Arneodo, A. Goldar, Y. d'Aubenton-Carafa, C. Thermes, B. Audit, O. Hyrien, Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome, PLoS Comput. Biol. 7 (2011) e1002322.