Mammalian Genome (2021) 32:443–447
https://doi.org/10.1007/s00335-021-09893-1

Lack of CpG islands in human unitary pseudogenes and its implication

Ammad Aslam Khan[1] · Muhammad Shahryar Ali[1] · Farah Babar[1] · Anees Fatima[1] · Muhammad Awais Shafqat[1] · Bisma Asghar[1] · Nimra Ilyas[1] · Maheen Fatima[1] · Ayesha Liaqat[1] · Muhammad Aslam Gondal[2]

Received: 31 March 2021 / Accepted: 7 July 2021 / Published online: 16 July 2021

Abstract

CpG islands (CGIs) are aggregation of CpG dinucleotides in the promoters of mammalian genes. These CGIs are present in almost all the housekeeping genes and some tissue-specific genes in the mammalian genome. Extensive research has been done on the prevalence and role of CGIs in protein-coding genes. However, little is known about CGIs in pseudogenes. In the current research project, we focused on CGIs in three main classes of pseudogenes e.g., duplicated pseudogenes (DPGs), processed pseudogenes (PPGs), and unitary pseudogenes (UPGs). We discovered a predominant absence of CGIs in the promoters of all three pseudogenes. We also compared the CGI profile of these pseudogenes with their parent genes and found that unitary pseudogenes (UPGs) differ from the DPGs and PPGs in the sense that in the latter, lack of CGIs is a consequential event while in UPGs, this lack of CGIs in their promoters is not a result of pseudogenization process. We also discussed the implication of the results obtained from this comparison. To our knowledge, this is the first-ever study highlighting this aspect of UPGs throwing new insights into the evolution of genome in general and especially in the context of pseudogenes.

Introduction

Vertebrate genomes are methylated predominantly at the dinucleotide CpG, resulting in a deficiency in CpGs because of the mutagenicity of methylcytosine (Bird 1980). Interestingly, there is an accumulation of CpGs in the promoter region of genes in mammals and other warm-blooded animals (Illingworth and Bird 2009; Sharif et al. 2010). These methylation-resistant clustered CpGs are called CpG islands (CGIs). In mammals, more than 40% of the genes and almost all the housekeeping genes contain CGIs in their promoters (Fatemi et al. 2005; Saxonov et al. 2006; Alberts et al. 2007). In humans, 60–70% of genes have CGIs in their promoters. The presence of CGIs in the promoters of almost all the mammalian housekeeping genes and fraction of the tissue-specific genes hints towards their potential role in transcription regulation (Deaton and Bird 2011).

Pseudogenes are a special class of genetic elements. Pseudogenes show sequence homology to other genes but do not get encoded into a functional protein (Mighell et al. 2000). There are three types of pseudogenes: processed pseudogenes (PPGs), unprocessed or duplicated pseudogenes (DPGs); and unitary pseudogenes (UPGs). DPGs are derived from the duplication of genes (Sen et al. 2010) while the PPGs are a result of the reintegration of a gene due to retrotransposition (Vanin 1985). UPGs are distinctive from the other two classes of pseudogenes in the sense that they are a special type of unprocessed pseudogenes that lack any functional counterpart in the genome of the same species nevertheless, they have their functional orthologues on the same locus in other species (Zhang et al. 2010). Some work has been done on CGIs in the first two classes of pseudogenes (Bird 1987; Antequera 2003) but almost nothing is known about their prevalence in UPGs. In this work, we have explored the status of CGIs in all three classes of pseudogenes with special focus on the prevalence of CGIs in UPGs and discussed its implications.

Material and method

The pseudogenes data for all three classes of human pseudogenes and mouse UPGs were extracted by using the BioMart tool, Ensembl 94 (Kinsella et al. 2011; Aken et al. 2016; Cunningham et al. 2019). The orthologous

✉ Ammad Aslam Khan
ammad.aslam@vu.edu.pk; ammadaslamkhan@gmail.com

[1] Department of Bioinformatics and Computational Biology, Virtual University, Lahore 547 92, Pakistan

[2] Department of Biology, Lahore University of Management Sciences, Lahore, Pakistan

of genes to human UPGs were extracted from genome sequences of mice and different primate species by making use of the BLASTN tool employed in the Ensembl genome browser. The genes were considered as orthologous when (i) they showed as the top score in the BLASTN search results, (ii) corresponding human UPGs came at the top in the results when the orthologous gene was used as a query sequence in reverse-BLAST, and (iii) there was no functional ortholog of mice/primate gene in human orthologs list of the Ensemble genome browser. During the study, we observed that the genome sequence of many primates is not available in a fully annotated form. As a result, we decided not to focus on a specific species of primate but to select a species (among the well-known primates e.g., chimpanzee, bushbaby, mice lemur, Orangutan, Marmoset, etc.) for which we could find the corresponding orthologous gene. To further confirm the authenticity of our approach for finding the orthologs of unitary pseudogenes, we compared our data with the data from Zheng et al. (2010). Of the 76 mouse orthologs, Zhang et al. included in their analysis, 21 are present in our data. The human-mice ortholog comparison of these 21 genes showed 100% match between our and Zhang's data, confirming the accuracy of our approach for finding the orthologs of human CGIs. The reason for finding only 21 mouse genes from Zhang's data in our dataset might be the fact that Zhang's pioneer work on UPGs was published more than a decade ago while Ensembl database is updated routinely and hence is based upon most up-to-date gene annotations (Supplementary Table 3). Therefore, we preferred to make use of a more updated and well-annotated ENSEMBL database for collecting UPGs and for the prediction of their mouse and primate orthologs. The parent genes of PPGs and DPGs were obtained from Psicube (Karro et al. 2007; Sisu et al. 2014). These parent genes are the functional homologs of PPGs and DPGs which did not accumulate deleterious mutations because of functional constraint and hence did not get pseudogenized. These parent genes are identified based on their sequence similarity with the pseudogenes. The pseudogenes in Piscube are annotated based on their structural features. So, the DPGs, like their parent genes, have intron–exon-like genomic structures and may still maintain the upstream regulatory elements. In contrast, PPGs, having lost their introns, contain only exonic sequence and do not retain the upstream regulatory regions (Li et al. 1981; Pei et al. 2012).

For each gene, a 1200 bp region of DNA composing of 1000 bp upstream and 200 bp downstream of TSS was selected to find the CGIs in promoters of these genes. The status of CGIs in all the genes was determined by using CGI finding program CpGProD (Ponger and Mouchiroud 2002). CpGProD uses quite stringent criteria for CpG island detection, i.e., DNA regions longer than 500 bp with G + C average of more than 0.5 and CpG observed/expected ratio of more than 0.6. GraphPad prism was used for conducting statistical analysis (Swift 1997).

Result and discussion

In the current research project, we looked for CGIs in the promoters of human UPGs. We extracted all the human UPGs (91 in total) from the ensemble genome browser and examined for the presence or absence of CGIs in their promoter. To our surprise, a predominant number of these UPGs, e.g., 86 of 91 (94.5%) lack CGIs in their promoter (Fig. 1a and Supplementary Table 1). To see whether this trend is specific to humans, we also studied the status of CGIs in mice genome. Indeed, we found the same pattern in UPGs in mice as well, e.g., 49 out of 52 (94%) UPGs lacked CGIs (Fig. 1a and Supplementary Table 2). Our findings that most of the UPGs lack CGIs in their promoters posed an intriguing question; is the absence of CGIs in UPGs a consequential event of the process that leads to their pseudogenization or is it one of the causal events that contribute to the process of pseudogenization of these genes? To address this question, we tried to find the genes in mice and primates which were orthologous to human UPGs. If the orthologous genes contain CGIs but these were lost in corresponding human UPGs, the absence of CGIs in human UPGs might be considered as a consequential event of the process of pseudogenization. However, if the presence (or absence) of these CGIs in mice and chimp orthologous genes is maintained in human UPGs, it would hint towards a potential role of lack of CGIs in the pseudogenization of genes in human. We observed that a majority of mice and chimpanzee genes are orthologous to human UPGs exhibiting CGIs profile similar to human UPGs. For instance, of the 75 orthologous genes in mice, 66 (88%, with a z-score of 17.2 and p value < 0.00001 at 0.05 significance level, with a null hypothesis that CGIs profile is not the same in human and their mice orthologs, i.e., $H_0 \neq H_A$) showed CGIs profile similar to corresponding UPGs in human. Similarly in primates, 50 out of 58 (86%, with a z-score of 11.8 and p value < 0.00001 at 0.05 significance level, with a null hypothesis that CGIs profile is not the same in human and their primate orthologs, i.e., $H_0 \neq H_A$) orthologous genes showed CGIs profile similar to human UPGs (Fig. 1b and Supplementary Table 1). So, the lack of CGIs in human UPGs is most probably not a consequence of the process of pseudogenization of these genes but instead, the genes which lacked CGIs were perhaps more prone to the process of pseudogenization in human as compared to the genes which contained CGIs.

The intriguing absence of CGIs in UPGs also compelled us to see the status of CGIs in the other two classes

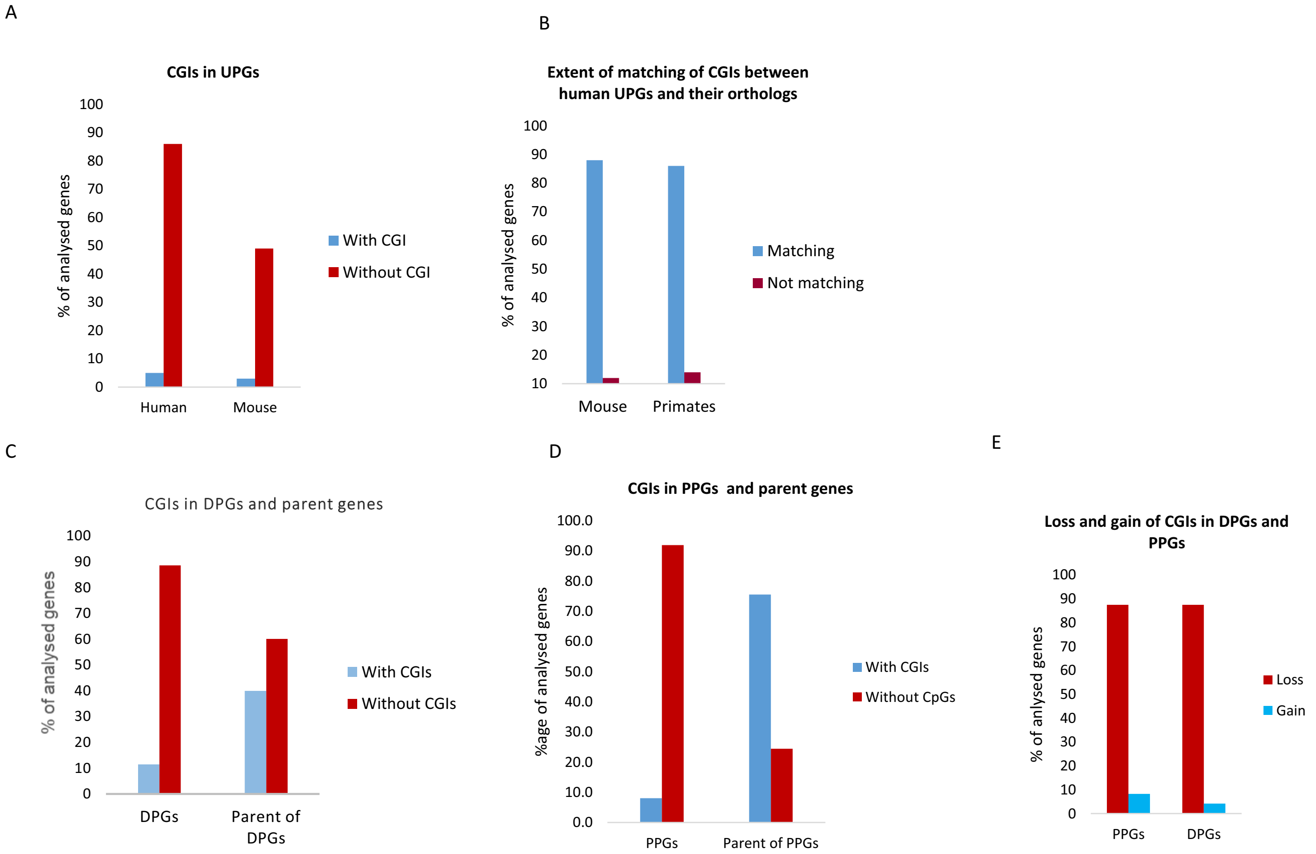

Fig. 1 CGIs in pseudogenes **a** Frequency of CGIs in unitary pseudogenes in human and mice. **b** Frequency of mice/primate genes (orthologous to human UPGs) which maintain or differ from CGI's profile in human UPGs. **c** Frequency of CGIs in DPGs and their parent genes. **d** Frequency of CGIs in PPGs and their parent genes. **e** Frequency of DPGs and PPGs which gained or lost CGIs with reference to their parent genes

of pseudogenes e.g., processed pseudogenes (PPGs) and unprocessed or duplicated pseudogenes (DPGs) and their parent genes. Of the 6940 PPGs and 2850 DPGs studied, we found a predominant absence of CGIs in 90% of genes in both PPGs and DPGs. However, in contrast to PPGs and DPGs, their parent genes contained CGIs in their promoters to a much larger extent, e.g., 40% and 75% of the parents of DPGs and PPGs have CGIs in their promoters, respectively (Fig. 1c, d and Supplementary Tables 4, 5). Accordingly, a stark contrast could be seen in terms of CGIs loss and gain in this comparison with 87% of the parent genes losing their CGIs in corresponding DPGs and PPGs while only a slim minority e.g., 4% and 8%, respectively, of the daughter pseudogenes gained CGIs (Fig. 1e). This predominant loss of CGIs of parent genes in their pseudogenized counterpart shows that the process of pseudogenization results in loss of CGIs in DPGs and PPGs. However, this must not be very surprising, as it is known that the PPG and DPGs tend to lack CGIs. This is because PPGs are derived because of the transposition of a gene from one part of the genome to another part. As PPGs depend upon reverse transcription for transposition, they lack any promoter (with or without CGIs) upstream of their transcription start site (Esnault et al. 2000). Similarly, in the case of DPGs, reduction in functional constraint is because of the presence of a duplicated copy of the gene that leads to pseudogenization of duplicated genes and depletion of CGIs (Lynch and Conery 2000; Subramanian and Kumar 2003).

Our findings that (i) UPGs predominantly lack CGIs; (ii) they retain the CGI profile of their mice and primate orthologous UPGs; and (iii) relatively a bigger proportion of the parent genes of DPGs and PPGs contain CGIs which get lost in the two classes of pseudogenes, strengthen our holding that (i) lack of CGIs in UPGs correlates with their pseudogenization, (ii) lack of CGI is not a consequence of the process of pseudogenization, and (iii) lack of CGI might be one of the factors among others (other factors include gene redundancy, loss of gene promoter because of transposition, etc.), which contribute in the pseudogenization of the genes. One way through which this lack of CGIs might have contributed to the process of pseudogenization is by giving some survival advantage to a gene which puts it under a greater functional constraint via CGIs. Hence, those genes which do not have CGIs might be under lesser functional constraint. After evolving under lesser functional constraint, such genes can easily accumulate disabling mutations (e.g., nucleotide insertions, deletions, and/or substitutions), ultimately leading to their pseudogenization (Li et al. 1981; Zhang and Gerstein 2003). It is well known that all three types of pseudogenes, DPGs, PPGs, and UPGs gets pseudogenized because of the accumulation of disabling mutations in their coding regions which either disrupts the reading frame or leads to the replacement of one or more than one functionally indispensable residue. In the case of DPGs and PPGs, the process of natural selection selects these mutations because of the redundancy of these genes as they arise from gene duplication and, therefore, evolve under lesser functional constraints, allowing accumulation of mutations and ultimately leading to their pseudogenization (Balakirev and Ayala 2003; Torrents et al. 2003). As PPGs arise from retrotransposition, they lack promoters and other regulatory elements which further contribute to their pseudogenization. The UPGs share the same fate as PPGs and DPGs, i.e., pseudogenization because of the accumulation of disabling mutations. But unlike PPGs and DPGs where we know the factors that contribute to the decrease in functional constraint, not much is known about the factors altering the functional constraints in the case of UPGs. Some of the studies have argued that UPGs arise because of organism-specific pseudogenization of genes as genes may be under different functional constraints in different organisms (Wu et al. 1989; Gilbert and Ziony 2001). So, maybe the lack of CGIs in UPGs along with the aforementioned organism-specific lack in functional constraint facilitates the process of pseudogenization. The organisms where the orthologs of human UPGs are still functional even if they lack CGIs may be because organism-specific functional constraints are quite stronger in them, and the absence of CGIs alone may not be too strong a factor to prevent purifying selection for these genes. Intriguingly, our findings also hint towards the fate of these functional orthologs of human UPGs. These orthologs are currently functional but a scenario may emerge in the future where organism-specific strong functional constraints are relaxed, which along with the lack of CGIs leads to pseudogenization of these genes, quicker than those which contain CGIs. This also leads to another intriguing evolutionary implication, i.e., maybe the introduction of CGIs in mammals (as non-mammal vertebrates generally lack CGIs) leads to some sort of functional stabilization of genes because of reduction in pseudogenization of CGI-containing genes and, thus, playing an important role in the overall evolution of the mammalian genome. Further studies are required, however, to explore the mechanistic details to confirm if the lack of CGIs is really playing some role in the pseudogenization of genes.

Our findings also point towards a need for giving special attention to UPGs rather than just considering them merely as a class of pseudogenes like DPGs and PPGs. With the genomes of increased species being sequenced, one can hope for better cataloging of UPGs in other species in near future. A comprehensive catalog of UPGs across vertebrates and characterization of their genetic signatures, i.e., the presence or absence of CGIs, will contribute in answering the questions raised in our study and our understanding of the process of genome evolution in general and the process of pseudogenization in special.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00335-021-09893-1.

Authors contribution AAK conceived the idea. AAK, MSA, FB, AF, MAS, BA, NA, MF, AU, and MAG collected the data. AAK, MSA, FB, AF, and MAS analyzed the data. AAK wrote the article.

Declarations

Conflict of interest The authors have no conflict of interest to disclose.

References

Aken BL, Ayling S, Barrell D et al (2016) The ensembl gene annotation system. Database. https://doi.org/10.1093/database/baw093

Alberts B, Johnson A, Lewis J, et al (2007) Molecular biology of the cell

Antequera F (2003) Structure, function and evolution of CpG island promoters. Cell Mol Life Sci

Balakirev ES, Ayala FJ (2003) Pseudogenes: are they "Junk" or functional DNA? Annu Rev Genet 37

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. https://doi.org/10.1093/nar/8.7.1499

Bird AP (1987) CpG islands as gene markers in the vertebrate nucleus. Trends Genet. https://doi.org/10.1016/0168-9525(87)90294-0

Cunningham F, Achuthan P, Akanni W et al (2019) Ensembl 2019. Nucleic Acids Res. https://doi.org/10.1093/nar/gky1113

Deaton AM, Bird A (2011) CpG islands and the regulation of transcription. Genes Dev. https://doi.org/10.1101/gad.2037511

Esnault C, Maestre J, Heidmann T (2000) Human LINE retrotransposons generate processed pseudogenes. Nat Genet. https://doi.org/10.1038/74184

Fatemi M, Pao MM, Jeong S et al (2005) Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. Nucleic Acids Res. https://doi.org/10.1093/nar/gni180

Gilbert SF, Ziony Z (2001) Congenital human baculum deficiency. Am J Med Genet 101

Illingworth RS, Bird AP (2009) CpG islands—"A rough guide." FEBS Lett

Karro JE, Yan Y, Zheng D et al (2007) Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res. https://doi.org/10.1093/nar/gkl851

Kinsella RJ, Kähäri A, Haider S et al (2011) Ensembl biomarts: a hub for data retrieval across taxonomic space. Database. https://doi.org/10.1093/database/bar030

Li WH, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature. https://doi.org/10.1038/292237a0

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science. https://doi.org/10.1126/science.290.5494.1151

Mighell AJ, Smith NR, Robinson PA, Markham AF (2000) Vertebrate pseudogenes. FEBS Lett

Pei B, Sisu C, Frankish A et al (2012) The GENCODE pseudogene resource. Genome Biol. https://doi.org/10.1186/gb-2012-13-9-r51

Ponger L, Mouchiroud D (2002) CpGProD: Identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics. https://doi.org/10.1093/bioinformatics/18.4.631

Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci USA. https://doi.org/10.1073/pnas.0510310103

Sen K, Podder S, Ghosh TC (2010) Insights into the genomic features and evolutionary impact of the genes configuring duplicated pseudogenes in human. FEBS Lett. https://doi.org/10.1016/j.febslet.2010.08.012

Sharif J, Endo TA, Toyoda T, Koseki H (2010) Divergence of CpG Island promoters: a consequence or cause of evolution? Dev Growth Differ. https://doi.org/10.1111/j.1440-169X.2010.01193.x

Sisu C, Pei B, Leng J et al (2014) Comparative analysis of pseudogenes across three phyla. Proc Natl Acad Sci USA. https://doi.org/10.1073/pnas.1407293111

Subramanian S, Kumar S (2003) Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res. https://doi.org/10.1101/gr.1152803

Swift ML (1997) GraphPad prism, data analysis, and scientific graphing. J Chem Inf Comput Sci

Torrents D, Suyama M, Zdobnov E, Bork P (2003) A genome-wide survey of human pseudogenes. Genome Res. https://doi.org/10.1101/gr.1455503

Vanin EF (1985) Processed pseudogenes: characteristics and evolution. Annu Rev Genet

Wu X, Lee CC, Muzny DM, Caskey CT (1989) Urate oxidase: primary structure and evolutionary implications. Proc Natl Acad Sci USA. https://doi.org/10.1073/pnas.86.23.9412

Zhang Z, Gerstein M (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res. https://doi.org/10.1093/nar/gkg745

Zhang ZD, Frankish A, Hunt T et al (2010) Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol. https://doi.org/10.1186/gb-2010-11-3-r26

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.