Frontiers in Bioscience Scholar

Front. Biosci. (Schol Ed) **2024**; 16(1): 2
https://doi.org/10.31083/j.fbs1601002

Short Communication

CpG Islands, Gene Expression and Pseudogenization: A Case for a Potential Trilogy

Ammad Aslam Khan[1,*], Anees Fatima[1]

[1]Department of Biological Sciences, Virtual University of Pakistan, 54000 Lahore, Pakistan
*Correspondence: ammad.aslam@vu.edu.pk; ammadaslamkhan@gmail.com (Ammad Aslam Khan)
Academic Editor: George Garinis
Submitted: 6 December 2023 Revised: 25 December 2023 Accepted: 16 January 2024 Published: 29 February 2024

Abstract

Background: The promoters of mammalian genes contain clusters of CG dinucleotides known as CpG islands. Most mammalian housekeeping genes predominantly contain CpG islands (CGIs), facilitating gene transcription. Numerous studies have explored the physiological implications of the relationship between CGIs and gene expression. However, the evolutionary implications of this relationship remain largely unexplored. Pseudogenes, in contrast, are genomic remnants that have lost their function over evolutionary time. **Methods**: In our current research, we employed comparative genomic techniques to demonstrate a correlation between the absence of gene expression due to a lack of CGIs in the gene promoters and pseudogenization. **Results**: We showed that there is a significant enrichment of tissue-specific genes in the functional orthologs of pseudogenes. We also found a significant correlation between the lack of CGIs and enriched tissue specificity in these functional orthologs of pseudogenes. **Conclusions**: We inferred that perhaps tissue-specific genes are more prone to the process of pseudogenization. In this way, because of their impact on gene expression, CGIs may affect the fate of a gene. To our knowledge, this is the first study to propose a connection between CGIs, gene expression, and the pseudogenization process and discuss the evolutionary implications of this potential trilogy.

Keywords: CpG islands; pseudogenes; gene expression; pseudogenization; tissue specificity; genome evolution

1. Introduction

Pseudogenes are non-coding and generally non-functional copies of the genome. Despite being physiological dead ends, they are universally prevalent in all metazoan genomes. In vertebrates, three types of pseudogenes are commonly present: Duplicated pseudogenes (DPGs), which arise from the duplication of a parent gene [1]. The second type is processed pseudogenes (PPGs), which originate from the retro-transposition of a gene. Both DPGs and PPGs have functional paralogs within a given species, referred to as parent genes. The third type is a distinct type of pseudogenes, termed Unitary pseudogenes (UPGs), which emerge due to the accumulation of disabling mutations, although they have functional orthologues in other species [2]. All these pseudogenes arise from the accumulation of disabling mutations due to the relaxed functional constraints these genes face for various reasons. For instance, the presence of multiple copies in the case of DPGs and PPGs and species-specific physiological variations in the case of UPGs [2–4].

CpG islands (CGIs) are clusters of cytosine and guanine dinucleotide aggregates in the promoters of most mammals and other warm-blooded animals. In other vertebrates, the genome is generally heavily methylated. This indiscriminate methylation process leads to a general depletion of CG dinucleotides due to the mutagenicity of methylcytosine [5]. However, although the genome is heavily methylated in mammals, CGIs generally remain unmethylated compared to other vertebrates. It has long been observed that DNA methylation generally suppresses the transcription process [6,7]. Therefore, the fact that CGIs remain unmethylated suggests that their presence promotes gene transcription. This comprehension is reinforced by the observation that nearly all housekeeping genes have CGIs in their promoters [8–11]. Furthermore, the existence of CGIs in the promoters of nearly all mammalian housekeeping genes and a portion of tissue-specific genes suggests their possible involvement in regulating transcription [12]. Genome-wide *in situ* chromatin immunoprecipitation (ChIP) and transcriptome analysis has also shown the recruitment of RNA polymerase II (RNAPII), specificity protein 1 (SP1), nuclear respiratory factor 1 (NRF-1), E2 promoter binding factor (E2F), and other important transcription factors, to promote the transcriptional process [13–15]. This recruitment of transcription factors is also considered one of the reasons for keeping the CGIs in their unmethylated state [16–18]. However, it is noteworthy that not all CGIs remain unmethylated. CGIs are methylated in certain cases where gene methylation is indispensable, such as x-chromosomal inactivation and genetic imprinting [19,20]. Similarly, in particular cancer tissues, CGIs are hypermethylated, leading to the transcriptional suppression of certain tumor suppressor genes [21]. The methylated CGIs recruit different proteins that methylate these sites and, ultimately, promote gene silencing. The main proteins that bind with methylated CGIs are the methyl-CGI-binding do-

main proteins (MDB1, MDB2, MDB3, MDB4, etc.) and another class of several structurally unrelated methyl-CpG-binding zinc-finger proteins in the Kaiso family (ZBTB33, ZBTB4, ZBTB38, etc.) [22]. In this way, proteins binding with the unmethylated CGIs differ from those binding with the methylated CGIs, where the former promotes transcription, and the latter suppresses it. Nevertheless, even though there are exceptional cases of CGI methylation, CGIs generally remain unmethylated and, consequently, are implicated in activating the transcription process.

Although much has been written on the effect of CGIs on gene expression, the evolutionary implications of this relationship have been thoroughly ignored. For instance, to our knowledge, no study has previously investigated the impact of CGI-based gene expression on gene evolution. In our previous work, we showed that there is a lack of CGIs in pseudogenes and emphasized that this lack of CGIs may have led to the pseudogenization process, promoting the formation of UPGs [23]. So, an inevitable question arises: How could the CGI profiles affect the fate of genes? The current research project is a step forward in determining this missing link between CGIs and pseudogenization. To the best of our knowledge, this is the first study that focuses on the potential way through which CGIs might be implicated in gene pseudogenization.

2. Materials and Methods

The identification of pseudogenes, their orthologs, and the prediction of CGIs was conducted as described by Khan *et al.* (2021) [23]. Briefly, pseudogenes were extracted using the Ensembl–BioMart tool [24]. Functional orthologs of the UPGs were identified using the Ensembl–BLAST tool [25,26]. We classified genes as orthologous to pseudogenes based on three criteria: (i) they were the highest scoring results in the BLASTN search, (ii) the associated human UPGs were the top-ranked when the orthologous gene was utilized as a query sequence in a reverse-BLAST search, and (iii) the list of human orthologs in the Ensemble genome browser did not contain a functional ortholog for the mouse/primate gene. For the gene expression data, we utilized three gene expression databases: (i) gene expression data from the fantom-5 project, focusing on 35 adult mice tissues [27], (ii) gene expression data from a study by Huntley *et al.* [28], focusing on gene expression in 9 organs in adult mice, and (iii) gene expression data from a study by Brawand *et al.* [29], focusing on gene expression in 6 organs in adult mice. Using data from three different sources rather than one minimized the chances of bias and increased the coverage of our target genes. A gene was considered expressed in a given tissue if its expression level was equal to or greater than 1 read per kilobase per million mapped reads, or RPKM [30]. Genes that were expressed in two-thirds or more of the total tissues for which gene expression data was taken were considered broadly expressed, while genes below this threshold were considered tissue-specific [31]. CGIs were predicted using the CpGprod tool [32]. We selected a 1200 bp DNA segment for each gene, which included 1000 bp upstream and 200 bp downstream from the transcription start site (TSS), to identify the CGIs in the gene promoters. CpGProD employs rigorous standards for detecting CpG islands, specifically, DNA segments exceeding 500 bp with an average G + C content over 0.5 and a CpG observed/expected ratio above 0.6. The significance levels were calculated using the chi-squared test to determine the association between two categorical values [33].

3. Results and Discussion

Gene pseudogenization is a universal feature of all genomes. Among the three pseudogene classes, unitary pseudogenes (UPGs) are unique since they are non-functional in one species but functional in many others. Gene expression is a hallmark of all protein-coding genes. Depending on the extent of the expression, genes can be categorized as either tissue-specific or broadly expressed. In this research project, we sought to explore if any relationship exists between gene expression and gene pseudogenization. To this end, we compared the expression profile of the functional orthologs of human UPGs (O-UPGs) in mice. Intriguingly, the majority of O-UPGs were tissue-specific in all our datasets that were used for the gene expression analysis, i.e., 46 out of 56 (82%), 41 out of 50 (82%), and 28 out of 32 (88%) genes were tissue-specific in the studies by Huntley *et al.* [28], Brawand *et al.* [29], and Fantom-5, respectively (Fig. 1A–C and **Supplementary file-1**). All these studies showed that a significantly dominant portion of O-UPGs was tissue-specific compared to all other protein-coding genes, with $p > 0.00001$ at a 0.01 significance level (Fig. 1D). Thus, the question arises as to the reason for these O-UPG expression profiles. In this direction, we examined these gene promoters since promoters are well known for modulating gene expression in eukaryotes [34]. Interestingly, the majority of the tissue-specific genes lacked CGIs in all three datasets, i.e., from the 5058, 5444, and 7016 tissue-specific genes used in our datasets from the studies by Huntley *et al.* [28], Brawand *et al.* [29], and Fantom-5, respectively, 4198 (83%), 4456 (66%), and 2406 (66%), respectively, lacked CGIs (Fig. 1E–G and **Supplementary file-2**). This shows a strongly significant correlation between tissue specificity and the lack of CGIs, with $p > 0.00001$ at a significance level of 0.01 (Fig. 1H). As expected, this trend also persisted in the O-UPGs, with the majority of tissue-specific O-UPGs from studies by Huntley *et al.* [28], Brawand *et al.* [29], and the Fantom-5 project, i.e., 46 out of 56 (82%), 41 out of 50 (82%), and 28 out of 32 (88%) respectively, lacking CGIs in their promoters (Fig. 2A–C). This reinforces already known observation that a significant portion of tissue-specific genes lack CGIs [9,11]. It has been shown in various studies that CGIs resist DNA methylation, hence, enhancing gene transcription

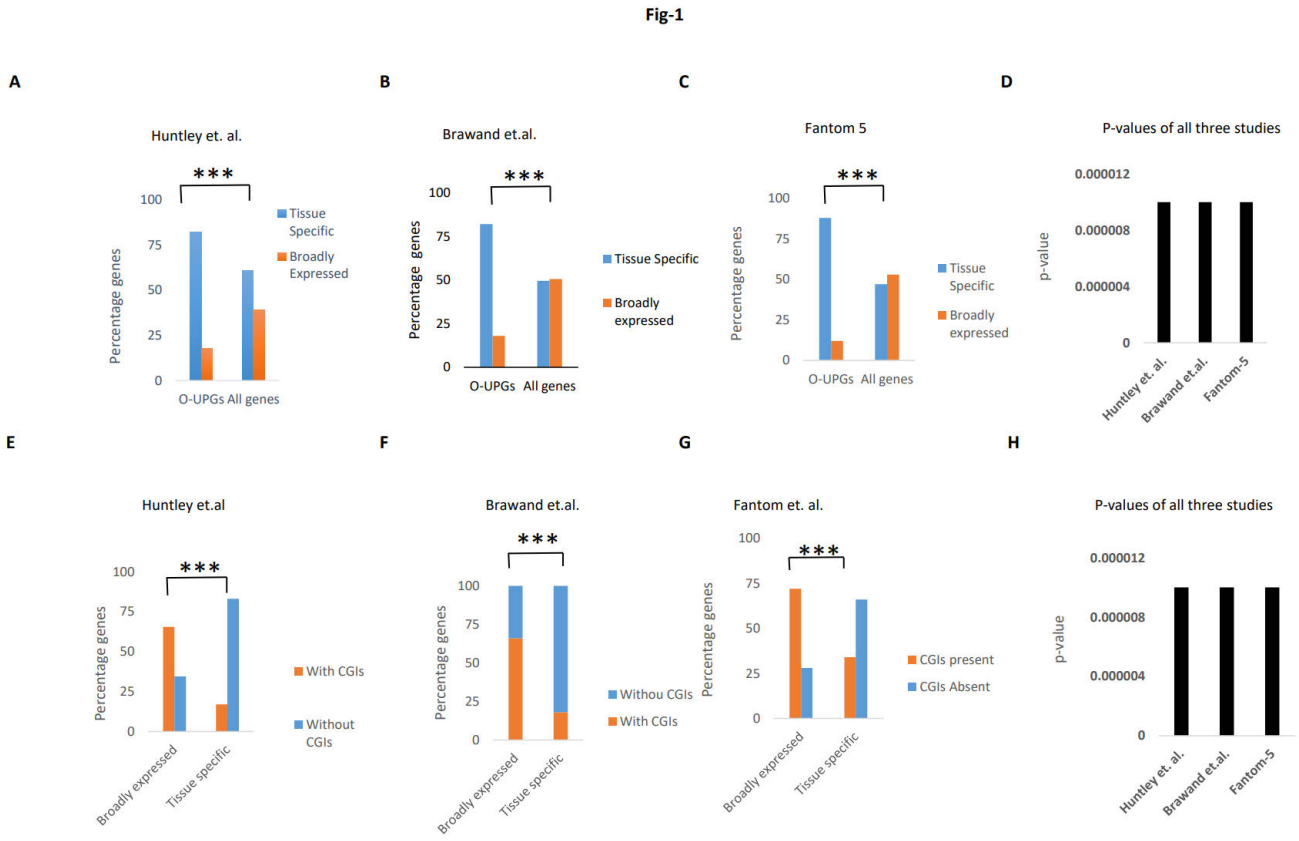

Fig. 1. Enrichment of tissue-specific genes. (A–C) The frequency of broadly expressed and tissue-specific genes in O-UPGs and all the protein-coding genes from three gene expression datasets. (D) Comparison of *p*-values for our results in all three datasets. X^2 test was employed to calculate the *p*-value at a 0.01 significance level. All three studies clearly show a highly significant correlation between tissue specificity and gene pseudogenization. (E–G) Distribution of CGIs in broadly expressed and tissue-specific genes. Gene expression data from three different studies was used in this analysis. (H) Comparison of *p*-values for our results in all three datasets. X^2 test was employed to calculate the *p*-value at a 0.01 significance level. All three studies clearly show a highly significant correlation between the lack of CGIs and tissue specificity at the 0.01 significance level. The asterisk (***) shows the significance level, with one asterisk representing very low significance and three asterisks representing highly significant correlations. O-UPGs, orthologs of human UPGs; CGIs, CpG islands.

[12,35]. This is one of the reasons that almost all mammalian housekeeping genes contain CGIs in quite contrast to tissue-specific genes, where the frequency of the CGIs is lower [36].

Our previous work demonstrated a predominant absence of CGIs in the PGs [23]. We were particularly intrigued by the fact that this lack of CGIs was also present in the mouse and primate genes orthologous to human UPGs. This led us to hypothesize that the genes lacking CGIs are more susceptible to pseudogenization. However, how the lack of CGIs could have led to gene pseudogenization was unclear. Our findings in the current research project provide the missing link between the lack of CGIs and gene pseudogenization, i.e., perhaps the lack of CGIs leads to an absence of gene expression, which results in tissue-specific gene expression. These tissue-specific genes are more prone to gene pseudogenization than highly active, broadly expressed genes. In this way, we can clearly see a correlation between the lack of CGIs, gene expression, and the gene pseudogenization process. Subsequently, the inevitable question arises: Why does nature tend to discount broadly expressed genes from the pseudogenization process and instead prefer tissue-specific genes for this fateful decay? One possible and intuitive explanation could be that genes expressed in a broad spectrum of tissues are perhaps under more functional constraints than genes expressed in only a few tissues. A gene, which is required by more tissues, will undoubtedly be quite crucial to the functioning of different body parts and will contribute immensely to the survival of the organism. However, a tissue-specific gene may be under relatively lesser functional constraint because its pseudogenization may not be a survival issue. Thus, compared to broadly expressed genes, tissue-specific genes will harbor more mutations, meaning a relatively relaxed functional constraint will ultimately lead to an ever-increasing accumulation of disabling mutations. Further-

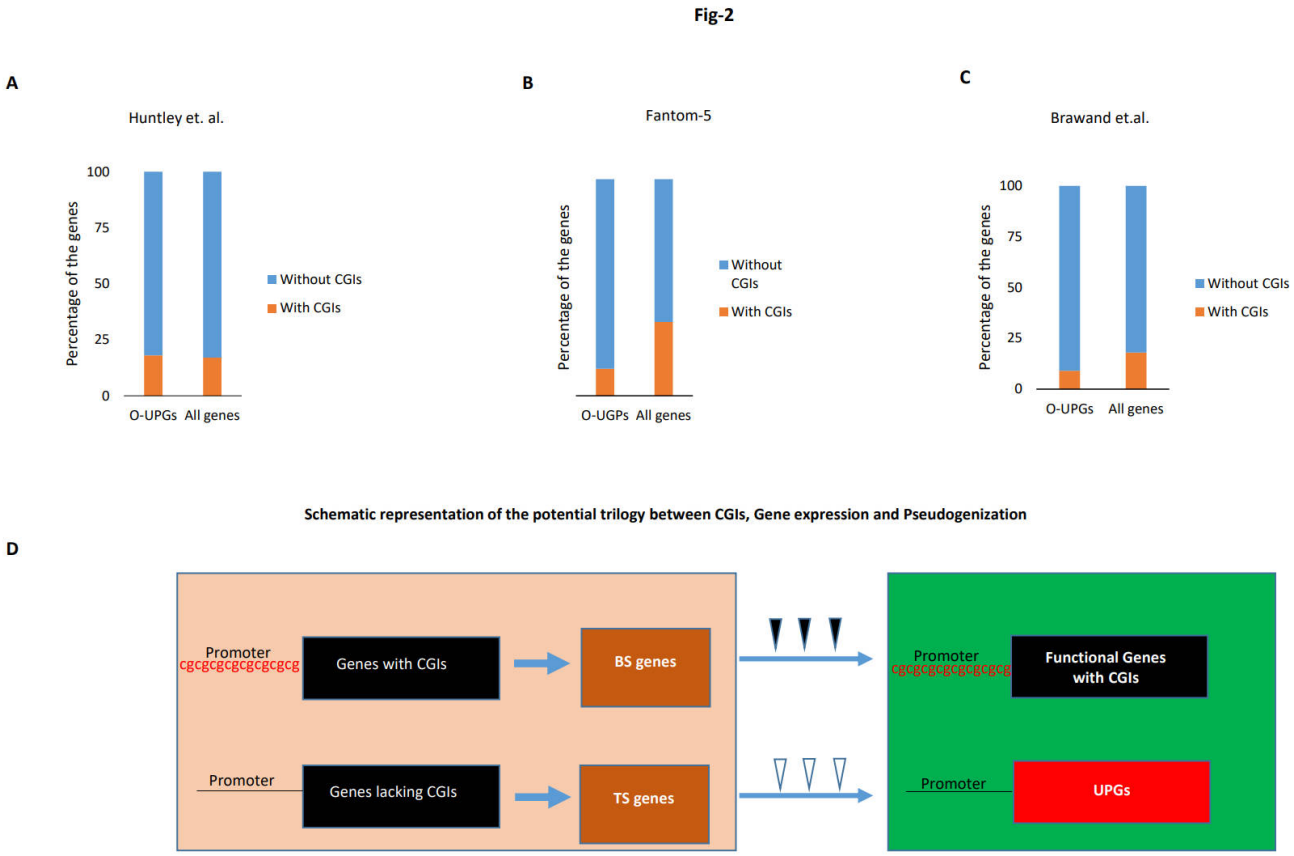

Fig. 2. Tissue specificity and pseudogenization. (A–C) Lack of CGIs in tissue-specific O-UPGs. Distribution of CGIs in the tissue-specific O-UPGs compared to their distribution in the tissue-specific gene in all the mouse protein-coding genes. The gene expression data from three different studies was used in this analysis. (D) A schematic representation of the potential trilogy. Impact of the status of the CGIs on gene expression and their fate. The left-hand side (light brown background) represents proto-pseudogene. The lack of CGIs leads to tissue specificity, which paves the way for UPGs to form in certain species, as shown on the right-hand side (green background). The triangles indicate functional constraints, with solid ones for strong and hollow ones for weak functional constraints.

more, the accumulation of disabling mutations ultimately leads to pseudogenization [37–39]. Hence, there are indeed studies hinting toward a relatively more conserved evolution of broadly expressed genes compared to tissue-specific genes [31,40]. One obvious consequence of this faster evolution of tissue-specific genes is that a possible lesser functional constraint, compared to broadly expressed genes in certain species, may lead to their relatively easier pseudogenization (Fig. 2D). In this way, our assertion that the lack of CGIs may be one of the reasons for the pseudogenization of UPGs is reinforced here; the lack of CGIs in gene promoters leads to the expression of genes in only a limited number of tissues, which in turn relaxes the functional constraint on these genes, ultimately, leading to pseudogenization in certain species.

4. Conclusions and Future Prospects

CGIs are a significant feature of mammalian gene promoters, yet CGIs are predominantly absent from pseudogene promoters. The absence of CGIs in most O-UPGs may have contributed to these genes becoming tissue-specific. These tissue-specific genes, in turn, are under lesser functional constraints and might be an easy target for pseudogenization. Therefore, the introduction of CGIs into the mammalian genome has not only affected the gene expression profile of genes, it may also have influenced their fate, with genes lacking CGIs becoming tissue-specific and, ultimately, pseudogenized in certain species, owing to relatively reduced functional constraint. This study underscores the need to further explore the control of gene expression through CGIs and how this process influences the evolution of vertebrate genomes in general and the mammalian genome in particular. The trilogy between CGIs, gene expression, and gene pseudogenization highlights the importance of introducing CGIs into the mammalian genome and underscores the need for a deeper and more comprehensive exploration of this all-important event in the history of mammalian evolution.

Availability of Data and Materials

All the Data is provided as supplementary material. In addition, the data utilized and/or examined in the present study can be obtained from the corresponding author upon a reasonable request.

Author Contributions

AAK conceived the idea. AAK planned and supervised the research. AAK and AF extracted the data. AAK and AF analyzed the data. AAK wrote the manuscript. In addition, both authors have participated sufficiently in the work to take public responsibility for appropriate portions of the content and agreed to be accountable for all aspects of the work in ensuring that questions related to its accuracy or integrity. Both authors read and approved the final manuscript. Both authors contributed to editorial changes in the manuscript.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

We are thankful to HEC for providing the funding (HEC-NRPU-17093) for this project.

Funding

This research is funded by Higher Education Commission of Paksitan (HEC) under the project HEC-NRPU-17093.

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/j.fbs1601002

References

[1] Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science (New York, N.Y.). 2000; 290: 1151–1155.

[2] Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biology. 2010; 11: R26.

[3] Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Carter DRF. Pseudogenes: Pseudo-functional or key regulators in health and disease? Rna. 2011; 17: 792–798.

[4] Sen K, Ghosh TC. Pseudogenes and their composers: delving in the 'debris' of human genome. Briefings in Functional Genomics. 2013; 12: 536–547.

[5] Bird AP. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Research. 1980; 8: 1499–1504.

[6] Yisraeli J, Frank D, Razin A, Cedar H. Effect of in vitro DNA methylation on beta-globin gene expression. Proceedings of the National Academy of Sciences of the United States of America. 1988; 85: 4638–4642.

[7] Boyes J, Bird A. DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. Cell. 1991; 64: 1123–1134.

[8] Fatemi M, Pao MM, Jeong S, Gal-Yam EN, Egger G, Weisenberger DJ, *et al*. Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. Nucleic Acids Research. 2005; 33: e176.

[9] Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103: 1412–1417.

[10] Crow M, Lim N, Ballouz S, Pavlidis P, Gillis J. Predictability of human differential gene expression. Proceedings of the National Academy of Sciences of the United States of America. 2019; 116: 6491–6500.

[11] Yang H, Li D, Cheng C. Relating gene expression evolution with CpG content changes. BMC Genomics. 2014; 15: 693.

[12] Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes & Development. 2011; 25: 1010–1022.

[13] Butler JEF, Kadonaga JT. The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes & Development. 2002; 16: 2583–2592.

[14] Landolin JM, Johnson DS, Trinklein ND, Aldred SF, Medina C, Shulha H, *et al*. Sequence features that drive human promoter function and tissue specificity. Genome Research. 2010; 20: 890–898.

[15] Rozenberg JM, Shlyakhtenko A, Glass K, Rishi V, Myakishev MV, FitzGerald PC, *et al*. All and only CpG containing sequences are enriched in promoters abundantly bound by RNA polymerase II in multiple tissues. BMC Genomics. 2008; 9: 67.

[16] Macleod D, Charlton J, Mullins J, Bird AP. Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes & Development. 1994; 8: 2282–2292.

[17] Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, *et al*. Sp1 elements protect a CpG island from de novo methylation. Nature. 1994; 371: 435–438.

[18] Takeshima H, Yamashita S, Shimazu T, Niwa T, Ushijima T. The presence of RNA polymerase II, active or stalled, predicts epigenetic fate of promoter CpG islands. Genome Research. 2009; 19: 1974–1982.

[19] Payer B, Lee JT. X chromosome dosage compensation: how mammals keep the balance. Annual Review of Genetics. 2008; 42: 733–772.

[20] Edwards CA, Ferguson-Smith AC. Mechanisms regulating imprinted genes in clusters. Current Opinion in Cell Biology. 2007; 19: 281–289.

[21] Moosavi A, Motevalizadeh Ardekani A. Role of Epigenetics in Biology and Human Diseases. Iranian Biomedical Journal. 2016; 20: 246–258.

[22] Bogdanović O, Veenstra GJC. DNA methylation and methyl-CpG binding proteins: developmental requirements and function. Chromosoma. 2009; 118: 549–565.

[23] Khan AA, Ali MS, Babar F, Fatima A, Shafqat MA, Asghar B, *et al*. Lack of CpG islands in human unitary pseudogenes and its implication. Mammalian Genome: Official Journal of the International Mammalian Genome Society. 2021; 32: 443–447.

[24] Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, *et al*. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database: the Journal of Biological Databases and Curation. 2011; 2011: bar030.

[25] Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, *et al*. Ensembl 2023. Nucleic Acids Research. 2023; 51: D933–D941.

[26] Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SMJ, *et al*. The Ensembl analysis pipeline. Genome Research. 2004; 14: 934–941.
[27] Noguchi S, Arakawa T, Fukuda S, Furuno M, Hasegawa A, Hori F. FANTOM5 CAGE profiles of human and mouse samples. Scientific data. 2017; 4: 1–10.
[28] Huntley MA, Lou M, Goldstein LD, Lawrence M, Dijkgraaf GJP, Kaminker JS, *et al*. Complex regulation of ADAR-mediated RNA-editing across tissues. BMC Genomics. 2016; 17: 61.
[29] Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, *et al*. The evolution of gene expression levels in mammalian organs. Nature. 2011; 478: 343–348.
[30] Hounkpe BW, Chenou F, de Lima F, De Paula EV. HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. Nucleic Acids Research. 2021; 49: D947–D955.
[31] Yang J, Su AI, Li WH. Gene expression evolves faster in narrowly than in broadly expressed mammalian genes. Molecular Biology and Evolution. 2005; 22: 2113–2118.
[32] Ponger L, Mouchiroud D. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics (Oxford, England). 2002; 18: 631–633.
[33] Tang DL. Chi-squared test. Zhonghua Yi Xue Za Zhi. 1984; 64: 50–53.
[34] Phillips T. Regulation of Transcription and Gene Expression in Eukaryotes. Nature Education. 2014; 1: 199.
[35] Vavouri T, Lehner B. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. Genome Biology. 2012; 13: R110.
[36] Zhu J, He F, Hu S, Yu J. On the nature of human housekeeping genes. Trends in Genetics: TIG. 2008; 24: 481–484.
[37] Chandrasekaran C, Betrán E. Origins of New Genes and Pseudogenes. Nature Education. 2008, 1: 181.
[38] Cheetham SW, Faulkner GJ, Dinger ME. Overcoming challenges and dogmas to understand the functions of pseudogenes. Nature Reviews. Genetics. 2020; 21: 191–201.
[39] Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000; 154: 459–473.
[40] Zhang L, Li WH. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Molecular Biology and Evolution. 2004; 21: 236–239.