RESEARCH

Open Access

Evolution of retrocopies in the context of HUSH silencing

Joanna Kozłowska-Masłoń[1,2], Joanna Ciomborowska-Basheer[1,3], Magdalena Regina Kubiak[1] and Izabela Makałowska[1*]

Abstract

Retrotransposition is one of the main factors responsible for gene duplication and thus genome evolution. However, the sequences that undergo this process are not only an excellent source of biological diversity, but in certain cases also pose a threat to the integrity of the DNA. One of the mechanisms that protects against the incorporation of mobile elements is the HUSH complex, which is responsible for silencing long, intronless, transcriptionally active transposed sequences that are rich in adenine on the sense strand. In this study, broad sets of human and porcine retrocopies were analysed with respect to the above factors, taking into account evolution of these molecules. Analysis of expression pattern, genomic structure, transcript length, and nucleotide substitution frequency showed the strong relationship between the expression level and exon length as well as the protective nature of introns. The results of the studies also showed that there is no direct correlation between the expression level and adenine content. However, protein-coding retrocopies, which have a lower adenine content, have a significantly higher expression level than the adenine-rich non-coding but expressed retrocopies. Therefore, although the mechanism of HUSH silencing may be an important part of the regulation of retrocopy expression, it is one component of a more complex molecular network that remains to be elucidated.

Keywords Retrocopy, Retroposed genes expression, Retrotransposition, HUSH complex, HUSH silencing

Introduction

Genome evolution is a major driver of biological diversity. The mechanisms of these changes in both coding and non-coding sequences and their impact on the species evolution, have been extensively studied [1–4]. One of these processes is the emergence of novel genes which can occur *de novo* from non-coding DNA, as a result of different sequence rearrangements producing genes with new functionality, and by duplication of existing genes. The importance of gene duplication has been emphasized since publications by Nei [1] and Ohno [2]. Furthermore, studies by de Koning et al. indicated that approximately 70% of the human genome consists of repetitive sequences, the vast majority of which are transposable elements [5], underlining the importance of studying these once underestimated components.

Gene duplication can occur by retrotransposition, a process of reverse transcription of messenger RNA (mRNA) and the subsequent integration of the resulting complementary DNA (cDNA) into the genome. The proteins required for this process are provided by several retrotransposable elements, e.g., long interspersed nuclear elements 1 (LINE1) [6, 7]. These proteins bind to

*Correspondence:
Izabela Makałowska
izabela.makalowska@amu.edu.pl
[1] Institute of Human Biology and Evolution, Faculty of Biology, Adam Mickiewicz University, Uniwersytetu Poznańskiego 6, Poznań, Poland
[2] Laboratory of Cancer Genetics, Greater Poland Cancer Centre, Garbary 15, Poznań, Poland
[3] Present address: Laboratory of Nature Education and Conservation, Faculty of Biology, Adam Mickiewicz University, Uniwersytetu Poznańskiego 6, Poznań, Poland

the mRNA, forming a complex that is transported back to the nucleus, where it anneals to double-strand breaks, undergoes reverse transcription, and is incorporated into the genome [3, 4, 8]. The resulting replicas (retrocopies) are characterized by the presence of poly(A) tracts, the absence of introns and regulatory components, and the repetitive sequences flanking the inserted sequence [9].

Retrocopies, which generally lack promoters, are regarded as 'dead on arrival', i.e., non-functional copies of their parents [10]. To become functional, the retrocopies have to be expressed and therefore have to acquire regulatory elements. One way to obtain this is to 'hitchhike' on the regulatory elements of other genes [11]. Indeed, many retrocopies are found nearby or within other transcribed genes [8]. A retrocopy may be also inserted downstream of pre-promoters that have evolved into functional elements over time, or it can acquire a distant promoter by gaining a new 5' exon from the vicinity of the insertion site [11, 12]. In some cases, a retrogene may also obtain a promoter from its progenitor if the parental gene is transcribed from the site upstream of the canonical transcription start site [13].

Many of these retroposed and transcriptionally active copies evolve neutrally because they do not encode proteins. Other, with intact coding sequence, may encode a protein that is beneficial to the organism and fulfil functions comparable to parental genes (subfunctionalization) [2, 14, 15]. However, the acquisition of a new role through evolution (neofunctionalization) is also very common [4, 16, 17]. As some studies have shown, retrocopies can also occasionally functionally replace their progenitors [18, 19].

Similarly to other retroelements, such retrotransposons and retroviruses, retrocopies provide genetic material that may bring an adaptive benefit and contribute to intra- and interspecies differences [20–22]. On the other hand, retroposition also pose a threat to genome integrity. An inserted retroelement may disrupt exonic sequence, interfere with splicing, affect transcriptional machinery [23]. In addition, not only transposable elements but also viral genetic material can be incorporated into the DNA of cells. Therefore, controlling this constant threat of RNA-derived elements invasion is fundamental to genome integrity. Developed defense strategies are usually based on chromatin silencing factors, such as small RNAs that bind to their targets or sequence-specific DNA-binding proteins [24]. The germline and pluripotent stem cells are primarily protected by PIWI-interacting RNAs (piRNAs) [25] and KRAB-containing zinc-finger proteins (KRAB-ZNFs) [26], whereas in differentiated cells, the human silencing hub (HUSH) complex is the most active one [27, 28]. HUSH is composed of transgene activation suppressor (TASOR), M-phase phosphoprotein 8 (MPP8), and Periphilin (PPHLN1, isoform 2) and has the ability to successfully silence LINE1s as well as retroviruses through the chromatin modification and histone H3 lysine 9 trimethylation (H3K9me3) [29–31]. Seczynska et al. found that sequences repressed by the HUSH complex can often be characterized as long, intronless, transcriptionally active transposable elements with a high level of adenine on the sense strand [27]. This critical genome defence strategy and ability of HUSH to target retroposed cellular mRNAs could have a significant impact on the evolution and expression of functional retrocopies. Retroposition of cellular mRNA is a primary mechanism of the new gene formation, and therefore, HUSH-mediated repression may play a key role in the functional evolution of these new genetic materials. The aim of this study is to examine the evolution of different classes of protein-coding genes' retrocopies in the context of the HUSH regulation.

Materials and methods

Data source

Analyses were performed based on the human and pig sets of retrocopies deposited in RetrogeneDB2, a database of retrocopy annotations in eucaryotic genomes developed in our laboratory [32]. Parental genes sequences were downloaded from the Ensembl database (release 105) [33]. Ensembl annotations were also used to identify protein-coding retrogenes.

Retrocopies deposited in RetrogeneDB were identified based on similarities between the reference genomic sequence and proteins encoded by multiexon genes. Several criteria were applied to filter the results and increase accuracy. It was required that at least two introns were lost and the alignment had at least 150 bp, at least 50% identity and covered at least 50% of the parental protein [32].

RNA-seq data analysis

For human we utilized genes expression estimation from previous studies performed in our laboratory based on 818 ENCODE RNA-seq libraries [16, 34]. 205 samples representing normal tissue were selected from this set. Raw reads from 15 porcine RNA-seq experiments were downloaded from publicly available databases, such as SRA NCBI [35], ENA EBI [36], or ENCODE [37]. It was required that selected RNA-seq datasets were composed of pair-end reads with 50 bp minimum length and originate from normal tissues or organs. A list of 205 human and 15 porcine analyzed libraries is shown in Table S1. The processing of RNA-seq reads was the same as previously for human data [16, 34]. First, reads went through quality control steps using FastQC [https://www.bioinformatics.babraham.ac.uk/projects/fastqc/] followed by quality filtering, quality trimming, and adapter clipping utilizing BBDuk2 from BBTools package (Joint Genome

Institute; https://jgi.doe.gov). The following parameters were set up for this step: qtrim=w, trimq=20, maq=10, rref=adapters.fa (a built-in set of Illumina adapters), k=23, mink=11, hdist=1, tbo, tpe, minlength=2/3 of raw read length, removeifeitherbad=t, which are thoroughly described on the tool's website (https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbduk-guide/). The reads originating from ribosomal RNA (rRNA) were filtered based on mapping with a set of human and porcine rRNA sequences obtained from Ensembl [33] and Refseq [38]. This step was performed using Bowtie 2 [39]. To establish a particular type of RNA-seq library and to ensure that only pair-end sets are further analysed, we used Bowtie and infer_experiment.py from the RSeQC package [40]. After downloading and preparing he porcine transcriptome from Ensembl (release 105), the expression levels for transcripts were estimated with Salmon v0.7.2 [41] using most of the default parameters, except for: --seqBias and --gcBias. The TPM (transcripts per million) values obtained for all the transcripts assigned to each gene were then summed using a Python script and combined with the RetrogeneDB2 annotations. Retrogenes annotated as known protein-coding genes were considered to be expressed. Retrogenes annotated as pseudogenes had to meet the following criterion to be counted as expressed: expression level≥1 TPM in at least three or two RNA-seq libraries for human and pig, respectively.

Analysis of protein-coding retrogenes origin

The group of human protein-coding retrogenes was analysed using the GenTree database to determine time of their origin [42]. Retrocopies were assigned to the branches of the phylogenetic tree based on their Ensemble ID. All retrogenes that originated after the Simiiformes branch split were recognized as young and specific for primates.

Identification of orthologs of HUSH complex components

Porcine orthologues of all components of the human HUSH complex - TASOR, MPP8, MPHOSPH8, and PPHLN1 - were identified based on the NCBI HomoloGene database (HomoloGene, https://www.ncbi.nlm.nih.gov/homologene). Orthology was also confirmed by reciprocal blastp search [43].

Calculation of nucleotide content

To calculate the adenine, thymine and GC content (A-content, T-content, GC-content) in the analyzed retrogenes and their parental gene sequences, the corresponding FASTA files for human, and pig retrocopies were downloaded from the RetrogeneDB2 database [32] and from the Ensemble database in the case of the parental genes. To calculate the GC-content in the retrocopy flanking regions, 5000 nt downstream and upstream of the retrocopy site were extracted from the human genome (Ensemble release 105). Subsequently, the computation step was conducted using the seq.kit script developed by W. Shen et al. [44].

Substitution analysis

To evaluate the rate of codon substitution at different codon position, first sequences of a retrocopy and a parental gene were aligned with tblastn to generate alignments at amino acid level. When the similarity between retrogene and parental gene was relatively low, retrogene nucleotide sequence was aligned to parental protein using blastx. Next, cognate coding sequences were aligned guided by amino acid alignment. This ensured that codons were aligned properly. Finally, the number of substitutions at each codon position was counted and the substitution rate was calculated. This has been done using a custom perl script. The type of the substitution was examined using in house Python script.

Statistical analysis

Adenine and thymine content and expression values were examined using GraphPad Prism 5 (GraphPad, San Diego, CA, USA, www.graphpad.com) and R [45] with the "ggsignif" [46], "ggplot2" [47], "ggh4x" [48], and "smplot2" [49] packages. First, the Shapiro-Wilk normality test was used to determine whether the data had a normal distribution. The Kruskal-Wallis test was with Dunn's multiple comparison test was then used to compare differences between adenine content and expression levels in protein-coding, expressed, and non-expressed retrocopies, as well as parental genes, in both species studied. The relationship between retrocopies adenine content and expression level was studied using the Spearman correlation test. Finally, to determine how adenine content and expression levels vary between retrocopies and their parental genes, the Mann-Whitney U test or the unpaired t-test with Welch correction was conducted. In all analyses, $p<0.05$ was considered statistically significant.

Visualisation

The graphs were prepared using R [45] with the four packages mentioned above [46–49], as well as the "tidyverse" package [50]. For transparency and to improve the quality of the graphs, approximately 0.03% of the outliers with the highest values were removed from each of the expression datasets. The operation did not affect the significance of the presented data. The phylogenetic tree was made using an online graphic design tool, Canva.

Results

Expression of retrocopies and parental genes

The 4611 human retrocopies were downloaded from RetrogeneDB2. Retrocopies recently retired from the Ensembl database [33] were excluded from the analyses, resulting in a final set of 4463 retrocopies [32]. Nevertheless, the number of retrogenes is likely to be underestimated due to the stringent requirements that were applied to the retrogene identification process in RetrogeneDB2. These retrocopies originated from 1503 parental genes. As many as 1340 retrocopies originated from RPL and RPS ribosomal proteins, which is not surprising [51]. The genes with the highest number of retrocopies include: *RPL21* (108), *PPIA* (88), *RPL23A* (68), *KRT18* (67), *HNRNPA1* (66), *RPL7* (55), *HMGN2* (55), *RPS2* (48), *RPL31* (46), and *RPL12* (43).

The expression level of retrogenes and their progenitors was estimated based on publicly available RNA-seq data. Based on the annotation and expression data, retrocopies were divided into three categories: protein-coding retrogenes and non-coding retrocopies, which were further divided into expressed and non-expressed retrocopy subgroups. Throughout the manuscript, these groups are referred to as protein-coding, expressed and non-expressed. The first subgroup (protein-coding) was the least numerous, accounting for only 2.38% of all retrocopies. Expressed and non-expressed retrocopies accounted for 42.93% and 54.69% respectively. Taken together, nearly 50% of human retrocopies demonstrated transcriptional activity.

Similarly to the retrocopies, their progenitors were also divided into three groups according to the category of retrocopy they produced. Some genes were placed in two or all three groups because they produce multiple retrocopies with different statuses. Comparison of expression levels showed that retrocopies have on average lower expression than their progenitors, and that protein-coding retrogenes have significantly higher expression than the subgroup of expressed retrocopies. However, the three groups of parental genes did not differ (Fig. 1).

The low level of expression of the retrogenes may indicate that they are under HUSH control similar to other retroposed sequences. Parental genes have multiple introns that protect them from the repressive HUSH effect. However, retrocopies have a much simpler structure than their cognate genes, in most cases only a relatively long single exon. To investigate this, we checked if there was a correlation between the length of the retrocopy genomic sequence and expression. In the group of protein-coding retrogenes, there was no correlation (Fig. 2A). However, in the case of expressed retrocopies, a significant negative correlation was found (Fig. 2A). This is in agreement with the work of Seczynska et al. [27], which showed that longer but intronless sequences are more susceptible to HUSH silencing. To clarify the lack of correlation in the group of protein-coding retrogenes, the expression of single-exon and multiexon molecules was then compared and significant differences were found. Single exon protein-coding retrogenes had significantly lower expression levels than those containing at

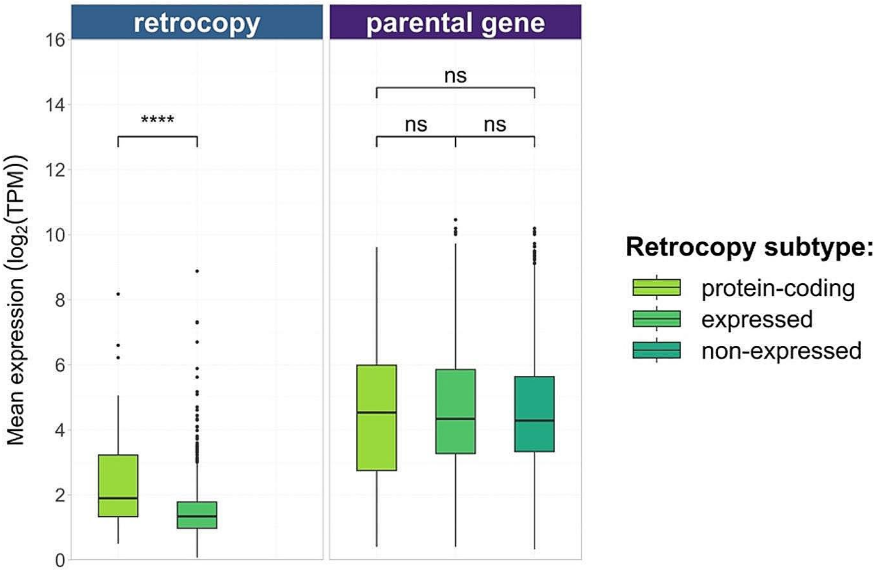

Fig. 1 The expression of retrocopies and their progenitors. Values were transformed to $\log_2$ for visualization purposes. **** $p \leq 0.0001$, ns – not significant

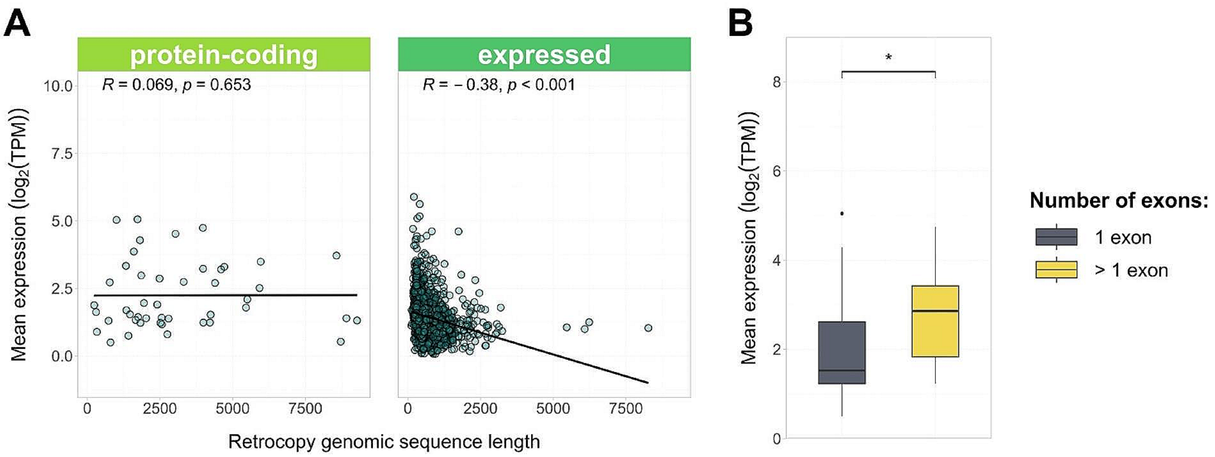

Fig. 2 Retrocopy genomic sequence length and expression level. (**A**) Expression correlation for protein-coding retrogenes and expressed retrocopies. (**B**) Comparison of the mean expression level between single- and multiexon protein-coding retrocopies. Values were transformed to $\log_2$ for visualization purposes. * $p \leq 0.05$

least one intron (Fig. 2B), confirming the protective role of introns.

TPM values do not take into account differences in sequencing depth. To ensure that our results are not biased by this issue, we repeated some analyses at the individual sample level. We compared protein-coding retrogenes with expressed retrocopies in all individual libraries. In each sample, protein-coding retrogenes had higher expression level (not shown). Similarly, we calculated the correlation coefficient between expression and the length of the expressed retrocopies. The correlation was always negative and in the vast majority statistically significant (not shown).

Retrocopies sequence composition and expression

It was noted that the adenine content (A-content) of the gene sense strand was positively correlated with the silencing by HUSH [27]. Therefore, the A-content was calculated in all groups of retrocopies and parental genes. It was determined that protein-coding retrogenes have the lowest level of adenine compared to the remaining two types, expressed and non-expressed. The mean adenine content was 26.98%, 30.33%, and 29.75%, respectively, and the differences observed between all groups were statistically significant (Fig. 3A). In accordance with the expression level, there was no variation in the average A-content between the three groups of parental genes (not shown). Interestingly, protein-coding retrogenes did not differ from their progenitors in this respect, but in the case of other two categories, the parental genes had a significantly lower fraction of adenine than their retrocopies (Fig. 3B).

In the context of HUSH silencing, these results appear to be in concordance with the expression level analysis. Protein-coding retrogenes have not acquired as many adenines as other recopies and may be therefore less susceptible to the influence of HUSH, and consequently achieve a higher level of expression, although still not as abundant as their precursors. However, if the high A-content on the sense strand would be selected for by HUSH silencing, no differences should be observed on the opposite DNA strand. To check this, we also analyzed the T-content on the sense strand, which reflects A-content on the other strand. Interestingly, there are no differences in a T-content on the sense strand between coding and expressed retrocopies, and their parental genes. The non-expressed retrocopies have a higher amount of T than their progenitors, although the difference is less significant than in the case of A-content. However, there are no significant differences between retrocopies categories (Fig. 3C-D). Consequently, with changes in the A-content, the GC-content of non-coding retrocopies is significantly lower compared to their progenitors (Fig. 4A). Despite this, all retrocopies have a higher GC-content than their surroundings, regardless of whether they are in the intergenic region or in the intron of another gene. The latter is quite common in retrocopies (Fig. 4B) [8]. This is because they inherit GC-content from their progenitors. Protein coding sequences are known to have high GC-content sequences compared to introns and intergenic sequences [52]. In addition, retrocopies tend to arise from genes with even higher GC-content, especially at their 5' ends [8].

To further clarify the phenomenon of high A-content in non-protein-coding retrocopies, we calculated the

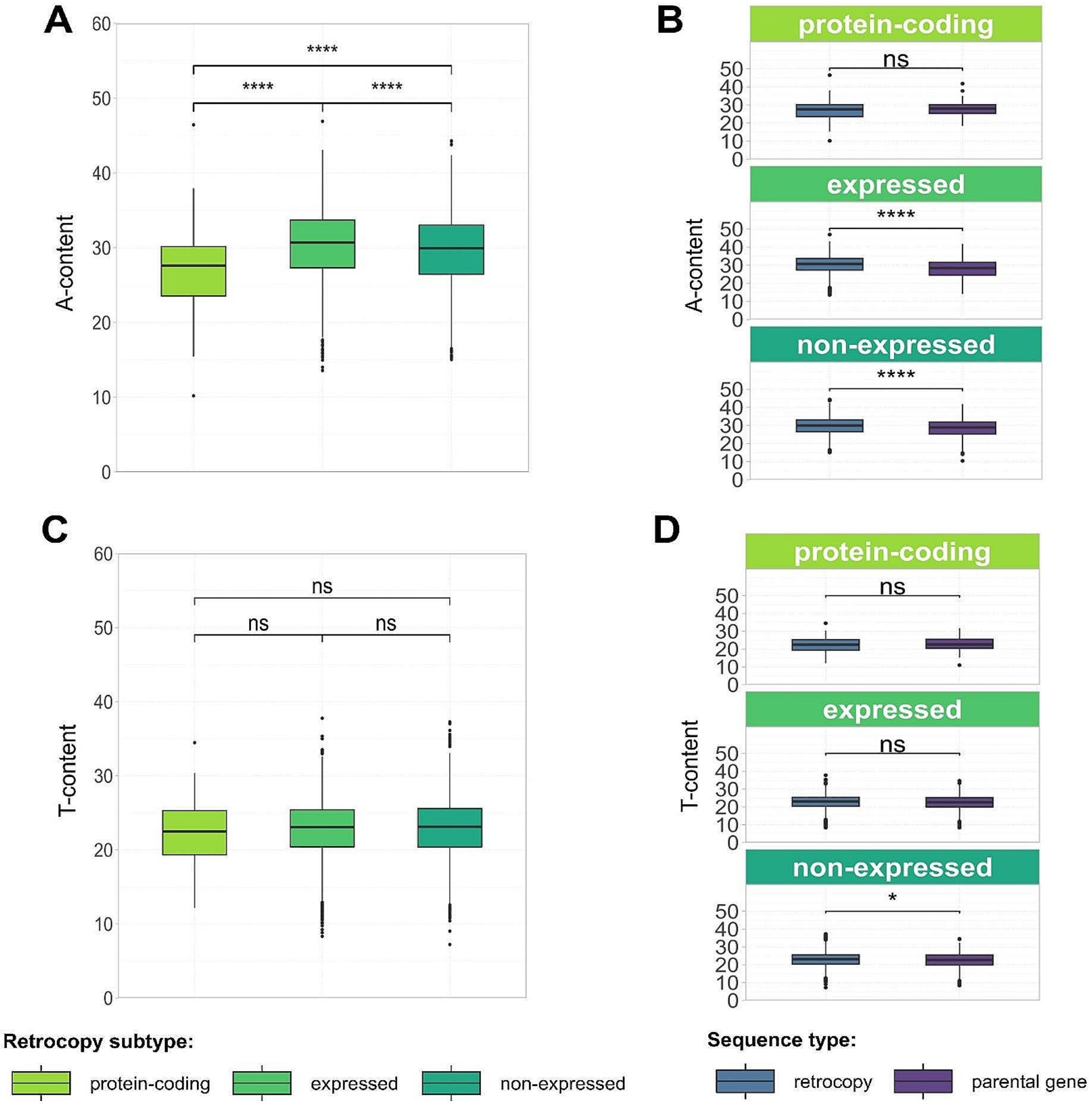

Fig. 3 Adenine content (A-content) and thymine content (T-content) in human retrocopies. (**A**) Comparison of the A-content in three groups of retrocopies, (**B**) between retrocopies and their progenitors, (**C**) comparison of the T-content in three groups of retrocopies and (**D**) between retrocopies and their progenitors, **** $p \leq 0.0001$, ns – not significant

Spearman correlation coefficient between A-content and retrocopy expression and found no correlation in either group - protein-coding retrogenes and expressed retrocopies (not shown). Nevertheless, it is plausible that there is no direct correlation, and that some threshold level of adenine must be reached to render a retrocopy susceptible to HUSH silencing. Therefore, to investigate the relationships between A-content, expression, transcript length, and gene structure, we divided the expressed retrocopies into six groups. The classification was made according to the genomic structure (single or multi-exon) and content of adenine (low – below the 25th percentile, medium – between the 25th and 75th percentile, and high – above the 75th percentile of A-content values) (Fig. 5). In each group, we calculated the correlation between the length of the transcript sequence and the expression. Interestingly, in the group of single exon retrocopies, there was a negative correlation between gene expression and transcript length regardless of adenine content (Fig. 5A-C). However, retrocopies with introns showed no significant correlation between expression and transcript length in any of the groups analyzed

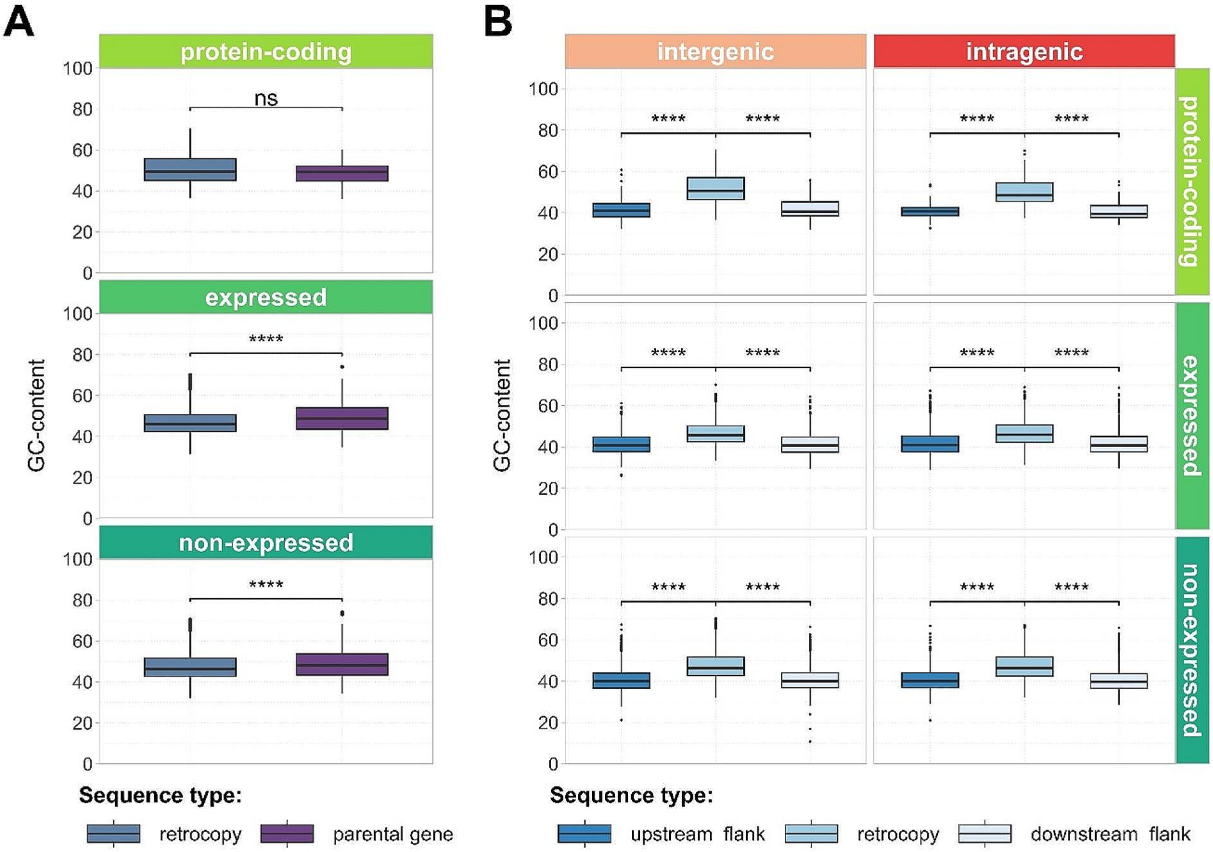

Fig. 4 Comparison of the GC-content (**A**) Between retrocopies and their progenitors and (**B**) Between retrocopies and their surroundings, **** $p \leq 0.0001$, ns – not significant

(Fig. 5D-F). Based on these results, it could be concluded that the exon length may be the factor that makes retrocopies susceptible to HUSH silencing and the presence of the intron may have some protective effect. However, they do not confirm a direct relationship between A-content and the expression level and suggest that the A-content, although significantly higher than in parental genes, is not an important factor.

Substitution pattern

Retrocopies are known to be 'dead on arrival', i.e. they are transcriptionally inactive after retroposition due to the lack of regulatory elements. They are therefore not under evolutionary pressure and accumulate mutations. The elevated levels of adenine in both groups of non-coding retrocopies (expressed and non-expressed) may indicate that these duplicates have evolved freely without any evolutionary pressure. To determine which substitutions contributed the most to adenine accumulation and to identify differences between protein-coding retrogenes and other retrocopies, all types of nucleotide changes were counted. In protein-coding retrogenes, T>C; A>G substitutions are the most common, followed by G>A; C>T changes (Fig. 6). Interestingly, it is opposite in both expressed and non-expressed retrocopies, the dominant substitutions are G>A; C>T and they are followed by T>C; A>G changes (Fig. 6). This result is similar to other nucleotide substitution studies in pseudogenes and the pattern was found to be the same regardless of the background GC composition [53].

We also checked the frequency of substitutions at different codon positions in the protein-coding retrocopies. The total length of the aligned amino acid sequences was 82,998 amino acids (excluding gaps), or 248,994 nucleotides respectively. As expected, the highest substitution rate (0.034) was at the third codon position, followed by first codon position (0.016) and the second one (0.013). This is a typical behavior of genes evolving under negative selection and consequently implying functionality of the analyzed genes [54, 55].

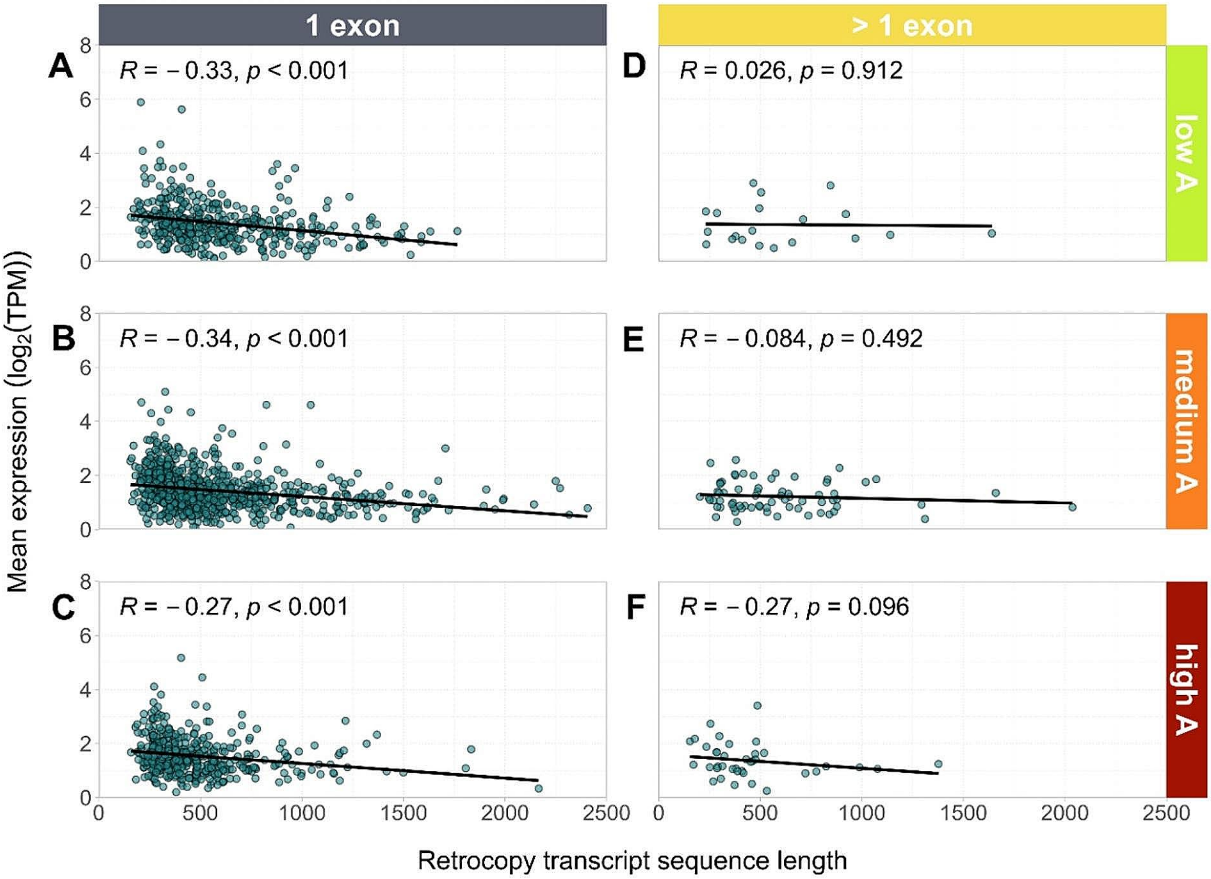

Fig. 5 Transcript length and expression correlation in six groups of expressed retrocopies: single exon with (**A**) low (below 25th percentile), (**B**) medium (between 25th and 75th percentile), and (**C**) high (above 75th percentile) A-content and multi-exon with (**D**) low, (**E**) medium, and (**F**) high A-content. The expression values were transformed to $\log_2$ for visualization purposes

Retrogenes in the pig genome

The HUSH complex is conserved from fish to mammals, so we investigated whether similar observations could be made in the case of different mammalian species. Although there is a wealth of data available for the mouse, we deliberately chose the pig for comparison as it is a more distant species. The 1026 retrocopies were downloaded from RetrogeneDB2. It is a significantly lower number than for humans, partly due to the burst of retroposition in primates [56] and partly due to gaps in the annotation of the pigs' genome. Nevertheless, the number is quite similar to other estimates [57]. The fractions of protein-coding and expressed retrocopies are higher compared to human, reaching 6.77% and 54.45% respectively (Fig. 6A). Two major factors contributed to these differences. First, the burst of retroposition in primates resulted in a large number of young and inactive retrocopies. Second, protein-coding retrogenes are old and mainly shared between mammals. Therefore, they make up a larger proportion of a smaller set of retrogenes. Analysed retrocopies originated from 508 parental genes. Ribosomal protein genes yield 278 retrocopies however, the gene with a highest number of retrocopies is *FTL* (33). It is followed by *RPL17* (30), *RPLP1* (16), *RPL9* (15), *SUMO2* (15), *RPL11* (14), *RPS25* (12), *RPS20* (12), *FTH1* (9), and *RPL32* (9).

RNA-seq data analysis revealed that the expression level of protein-coding retrogenes is elevated compared to the group of expressed retrocopies (Fig. 6B). These results are consistent with previous conclusions based on human data. In addition, as in humans, the expression of parental genes does not vary between genes that produce different types of retrocopies (Fig. 7B). We then examined the expression pattern of single and multi-exon protein-coding retrogenes. In pigs, as in humans, more complex retrogenes were more abundantly expressed, although the differences in the mean expression levels were not statistically significant (Fig. 7C).

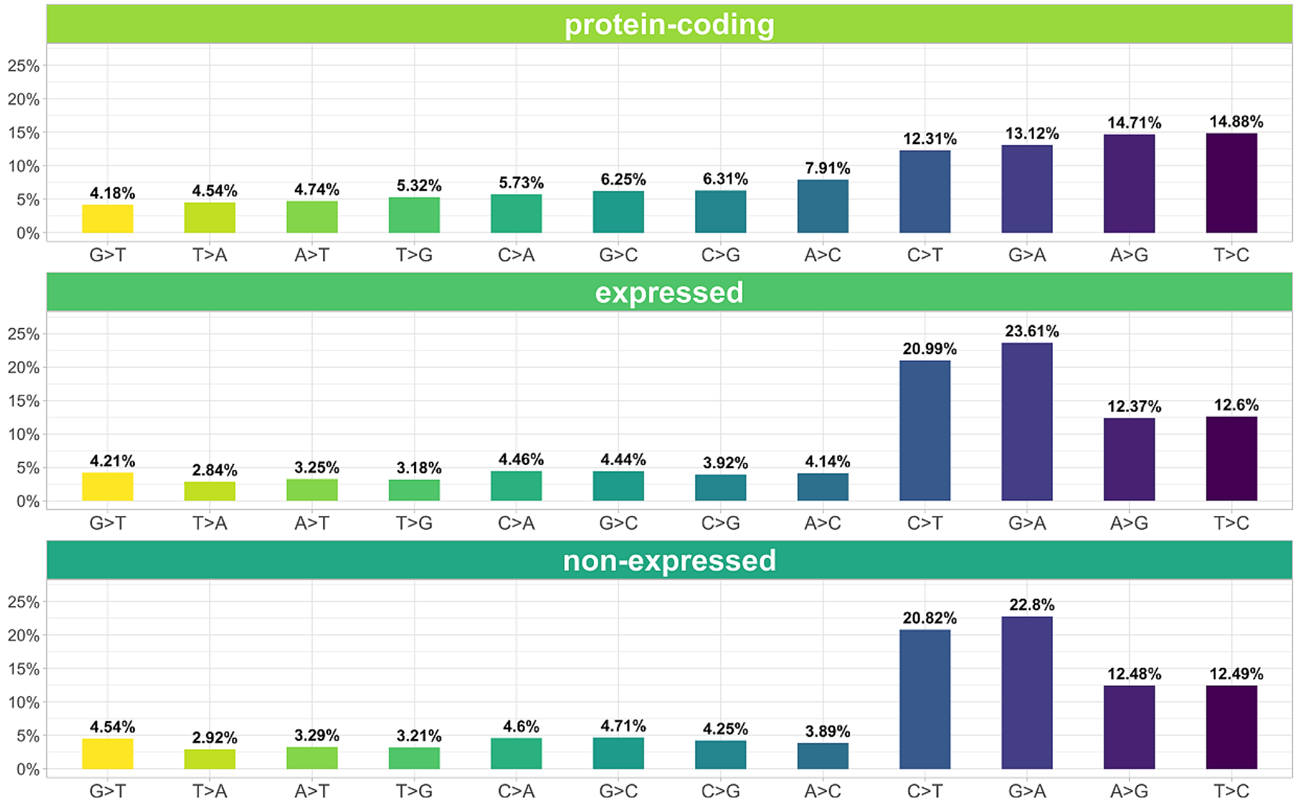

Fig. 6 Directions of nucleotide substitutions in all analyzed groups of retrocopies

Adenine content analysis also gave results consistent with the analysis of human retrocopies. The A-content is significantly lower in protein-coding retrogenes and there is no difference between the two other groups of retrocopies. Also, while protein-coding retrogenes do not differ from their progenitors in the average adenine content, there is a significant difference in the case of the remaining two groups of retrocopies (Fig. 7D-E). The substitution patterns resemble those observed in humans (Fig. 7F).

Discussion

The high level of retrotransposition, accompanied by complex mechanisms of the development of new functions, confirms the impact of RNA processing and RNA-directed rewriting of DNA on the evolution and phenotypic diversity of organisms. Retrocopies have been shown to significantly influence the diversification of transcriptomes and proteomes, earning them the title of ‘seeds of evolution’ [58, 59]. Studies of young retrogenes have shown that these sequences played a substantial role in, e.g., evolution of brain in primates [60] and *Drosophila melanogaster* [61]. Also, these new additions developed unique spatial expression patterns compared to the parental genes, and molecules derived from these retrogenes gained novel biochemical properties [60, 62, 63], and/or different subcellular localization patterns [60, 62]. This subcellular adaptation or relocalization process represents a new evolutionary pathway for the development of new gene functions [8, 64].

Retrogenes play a crucial role in genome evolution by providing novel genetic material, but they also pose a threat to genome integrity. As products of reverse transcription, they can be recognized as genomic ‘parasites’ and are therefore susceptible to repression by the HUSH complex, as determined by Seczynska et al. [27]. The researchers showed that the HUSH complex represses the products of reverse transcription inserted into the genome. They also showed that HUSH targets long, intronless, and transcriptionally active sequences in which the sense strand is rich in adenine [27]. The HUSH complex, TASOR, MPP8, and periphilin regulate the expression of retroposed sequences in an H3K9me3-dependent manner, meaning that transcription is required for the H3K9me3 initiation and propagation. Targets are localized by periphilin, which binds to RNA and enables HUSH to respond to increased transcription. Consequently, an increased amount of target RNA leads to further periphilin binding and intensified HUSH occupancy. This in turn recruits more SETDB1, a histone methyltransferase, and MORC2, an ATP-dependent chromatin remodeler that compacts chromatin [29, 65].

The HUSH complex recognizes evolutionarily young retroelements and provides an immediate defense

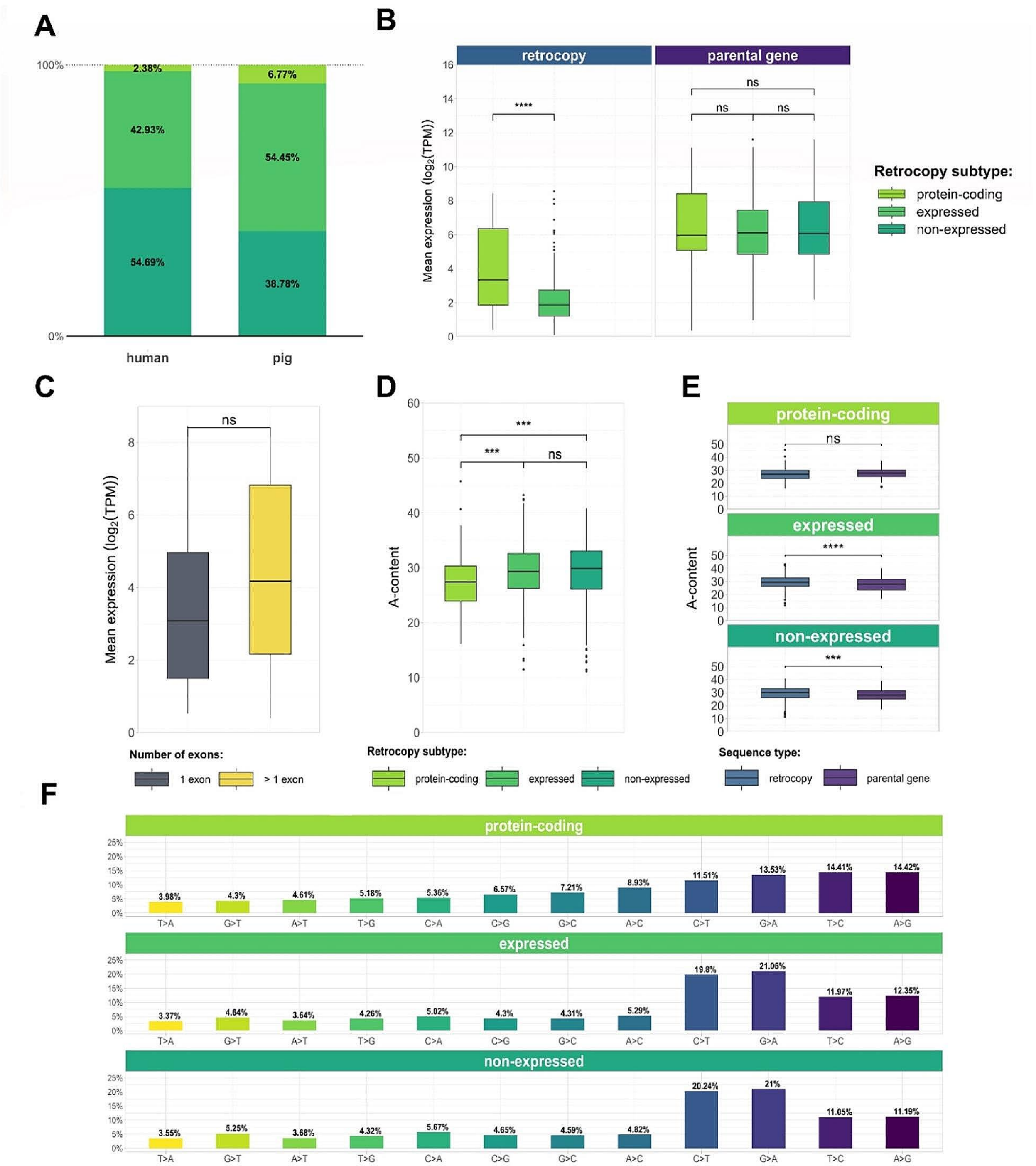

Fig. 7 Analyzes of pigs' retrocopies corresponding to previous calculations in humans. (**A**) The percentage of studied retrocopy subtypes in human and pig, (**B**) The expression of pigs' retrocopies and their progenitors, (**C**) Comparison of the mean expression level between single- and multiexon protein-coding retrocopies, (**D**) Comparison of the A-content in three groups of retrocopies and (**E**) between retrocopies and their progenitors, (**F**) Percentage of individual substitutions in all nucleotide changes. Values were transformed to $\log_2$ for visualization purposes. **** $p \leq 0.0001$, *** $p \leq 0.001$, ** $p \leq 0.01$, ns – not significant

mechanism against these genomic 'invaders'. However, this evolutionary 'war' is fought on both sides: host and parasite. Over time, transposable elements have therefore evolved their own defense mechanisms, making them at least partially resistant to the influence of HUSH. Human immunodeficiency viruses type 1 and 2 (HIV-1 and HIV-2), for example, use their viral auxiliary proteins to counteract HUSH restrictions. The viral proteins Vpx and Vpr antagonize SAMHD1, a factor that inhibits the reverse transcription. These molecules bridge the DCAF1 ubiquitin ligase substrate adaptor to SAMHD1 for subsequent ubiquitination and degradation [66, 67]. It appears that Vpx and Vpr counteract HUSH repression by a similar mechanism - an induction of its proteasomal degradation through the recruitment of DCAF1 [68, 69].

In the present study, we analyzed retrocopies in the context of HUSH complex repression. Our analyses of retrocopy expression levels confirmed previous findings that most of these molecules, including protein-coding retrogenes, have low expression levels [70]. In addition, we showed that their expression is significantly reduced compared to their cognate genes. Retrocopies contain very long exons resulting from the mechanism of their origin. The above support the studies of Seczynska et al. [27] and indicate that low expression of retroposed genes may be resulting from HUSH repression. However, our results demonstrate that the retrocopies have found a way to 'escape' the silencing of HUSH. This is mainly due to the evolutionary fate of retroposed genes. Initially, most retrocopies are deprived of regulatory elements and are considered to be 'dead on arrival'. To become transcriptionally active and thus targeted by HUSH, retrocopies need to acquire promoters. Published studies show that the vast majority of these retrocopies acquired a promoter *de novo* from a cryptic intergenic promoter (86%) [70]. Promoter acquisition is in many cases associated with the gain of a new 5' exon, and it has been shown that many transcriptionally active retrocopies gained 5' exons from upstream sequences. This implies the acquisition of introns, which are often very long. Exons can be acquired quite rapidly, and about 20% of young human retrogenes have non-parental 5' exons. According to Seczynska et al., introns, especially long ones, protect against HUSH repression [27]. Therefore, as more complex structures are obtained, retrocopies also gain at least some immunity to HUSH.

This complex has also been shown to target sequences with a substantial amount of adenine in the DNA sense strand [27], which is consistent with the context of retroelements evolution and DNA methylation. Deoxycytosine methylation occurs at the cytosine of the CpG dinucleotide, producing 5-methylcytosine (5mC), which mutates to thymine by spontaneous deamination [71]. As result there is observed CpG decay and the increase in TpG and CpA dinucleotide frequency. It is known that in primate genomes, for example, more than 40% of CpG islands are found within repetitive elements [72]. Accumulation of adenine has been observed as a result of methylation in *Alu* retroelements [73]. This supports the finding that HUSH defends the genome against DNA invasion and targets sequences with high adenine content. However, our results show that this may not be true for retroposed genes. Protein-coding retrogenes have, on average, lower expression than their progenitors, but do not differ in adenine content. We also found no correlation between the adenine content and the expression level. Therefore, other factors, such as the presence of long exons, seem to be more important. The expression of single exon retrocopies decreases with increasing exon length, independent of the adenine content. The presence of a long exon does not seem to have such a negative effect on expression when the retrocopy, whether protein-coding or non-coding, has acquired an intron.

Protein-coding retrogenes had significantly lower amounts of adenine than the other two categories of retrocopies. It has previously been shown that the exonic sequences contain more CpG than intergenic [52] and intronic DNA [74], making them more susceptible to mutation. It has also been shown that CpG-containing codons are subject to greater purifying selection than less mutable sites at identical codon positions [75, 76]. In addition, high GC-content promote nuclear export of mRNAs, especially in intron-poor mRNAs, and is important in distinguishing functional RNAs from junk transcripts [77]. These GC-rich regions likely recruit protein factors such as the THO complex, SR proteins and RBM33, which recruit nuclear transport receptors [78]. The above highlights the differences between protein-coding retrogenes and the remaining two categories of retrocopies – expressed but non-coding and non-expressed. Protein-coding retrogenes probably gained promoters soon after retrotransposition before losing coding potential due to mutations, which immediately put them under selective pressure and preserved CpG-containing codons. As a result, the adenine content, and therefore also CG-content, of the protein-coding retrogenes does not differ from that of the parental genes. In contrast, in retrocopies that were not transcriptionally active for long periods of time or did not acquire promoters at all, both nucleotides at the CpG site were free to undergo neutral nucleotide substitution. In the absence of negative selection, TpG and CpA dinucleotides accumulated as a consequence of cytosine methylation and following mutations to thymine. Our results corroborate those of Subramanian and Kumar, who demonstrated the over-time decay of CpG in pseudogenes [52]. Non-coding retrocopies inherited a high CpG content from their protein-coding parents, and since they no longer code for proteins, these highly mutable sites have escaped selective pressure. Thus, even in a relatively short time, they could accumulate enough adenine to differ from their parents. The results of the studies of Seczynska et al. [27] suggest that sequences with high adenine content are more susceptible to HUSH silencing. However, our study showed no correlation between adenine content and the expression level in any group of retrocopies. On the other hand, there is a significant decrease in the CG-content, and since high CG-content has been found to correlate with the nuclear transport of intron-poor genes [77], this may be a more important factor responsible for the low level of expression of retrocopies than A-content. In addition, we cannot exclude other factors, such as promoter architecture. For example, it has been shown that the promoters of retrocopies have depleted CpG islands and are bound to fewer transcription factors than the original genes [79].

Conclusions

In summary, the results of our study show that the presence of long exons has a negative effect on the level of retrocopy expression. We have also shown that intron gain provides some protection against possible HUSH repression and makes the expression level less dependent on the transcript length. The above may suggest that retrocopies are under some control of the HUSH complex. However, we cannot exclude other factors, such as GC-content and/or promoter architecture.

Abbreviations

5mC	5-Methylcytosine
DCAF1	DDB1 And CUL4 Associated Factor 1
H3K9me3	Histone H3 lysine 9 trimethylation
HIV-1 and HIV-2	Human immunodeficiency viruses type 1 and 2
HUSH	Human silencing hub
KRAB-ZNF	Zinc finger protein containing the Krüppel associated box (KRAB)
LINE1	Long interspersed nuclear elements 1
MORC2	MORC Family CW-Type Zinc Finger 2
MPP8	M-phase phosphoprotein 8
PPHLN1	Periphilin
SAMHD1	SAM domain HD domain-containing protein 1
SETDB1	SET Domain Bifurcated Histone Lysine Methyltransferase 1
TASOR	Transgene activation suppressor
TPM	Transcripts per million
Vpr	Viral protein R
Vpx	Viral protein X

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13062-024-00507-9.

Supplementary Material 1

Author contributions

Conceptualization, I.M., Methodology, I.M., J.K.-M., and J.C.-B.; Software, J.K.-M., and M.R.K.; Investigation, I.M., J.K.-M., and J.C.-B.; Data Curation, I.M., J.K.-M., and J.C.-B.; Writing - Original Draft Preparation, J.K.-M. and J.C.-B.; Writing - Review & Editing, I.M., and J.K.-M.; Visualization, J.K.-M.; Supervision, I.M. All authors have read and agreed to the published version of the manuscript.

Data availability

The data presented in this study are openly available at SRA NCBI, ENA EBI, or ENCODE databases, and its analysis and results presented in this work do not violate any copyrights.

Declarations

Ethics approval and consent to participate

The data was obtained from openly available databases gathering RNA-seq data - SRA NCBI, ENA EBI, or ENCODE databases, and downloading does not require additional consent or ethics approvals.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 17 April 2024 / Accepted: 29 July 2024
Published online: 02 August 2024

References

1. Nei M. Gene duplication and nucleotide substitution in evolution. Nature. 1969;221:40–2. https://doi.org/10.1038/221040a0
2. Ohno S. Evolution by Gene Duplication. Berlin: Springer; 1970.
3. Chen S, et al. New genes as drivers of phenotypic evolution. Nat Rev Genet. 2013;14:645–60. https://doi.org/10.1038/nrg3521
4. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010;20:1313–26. https://doi.org/10.1101/gr.101386.109
5. de Koning AP, et al. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 2011;7:e1002384. https://doi.org/10.1371/journal.pgen.1002384
6. Esnault C, et al. Human LINE retrotransposons generate processed pseudogenes. Nat Genet. 2000;24:363–7. https://doi.org/10.1038/74184
7. Wei W, et al. Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol. 2001;21:1429–39. https://doi.org/10.1128/MCB.21.4.1429-1439.2001
8. Kaessmann H, et al. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet. 2009;10:19–31. https://doi.org/10.1038/nrg2487
9. Long M. Evolution of novel genes. Curr Opin Genet Dev. 2001;11:673–80. https://doi.org/10.1016/s0959-437x(00)00252-5
10. Zhang Z, et al. Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet. 2004;20:62–7. https://doi.org/10.1016/j.tig.2003.12.005
11. Troskie RL, et al. Processed pseudogenes: a substrate for evolutionary innovation: retrotransposition contributes to genome evolution by propagating pseudogene sequences with rich regulatory potential throughout the genome. BioEssays. 2021;43:e2100186. https://doi.org/10.1002/bies.202100186
12. Fablet M, et al. Evolutionary origin and functions of retrogene introns. Mol Biol Evol. 2009;26:2147–56. https://doi.org/10.1093/molbev/msp125
13. Okamura K, Nakai K. Retrotransposition as a source of new promoters. Mol Biol Evol. 2008;25:1231–8. https://doi.org/10.1093/molbev/msn071
14. Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000;154:459–73. https://doi.org/10.1093/genetics/154.1.459
15. Force A, et al. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999;151:1531–45. https://doi.org/10.1093/genetics/151.4.1531
16. Kubiak MR, et al. Complex analysis of retroposed genes' contribution to human genome, proteome and transcriptome. Genes (Basel). 2020;11. https://doi.org/10.3390/genes11050542
17. Poliseno L, et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010;465:1033–8. https://doi.org/10.1038/nature09144

18. Ciomborowska J, et al. Orphan retrogenes in the human genome. Mol Biol Evol. 2013;30:384–96. https://doi.org/10.1093/molbev/mss235
19. Krasnov AN, et al. A retrocopy of a gene can functionally displace the source gene in evolution. Nucleic Acids Res. 2005;33:6654–61. https://doi.org/10.1093/nar/gki969
20. Kabza M, et al. Inter-population differences in retrogene loss and expression in humans. PLoS Genet. 2015;11:e1005579. https://doi.org/10.1371/journal.pgen.1005579
21. Parker HG, et al. An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science. 2009;325:995–8. https://doi.org/10.1126/science.1173275
22. Abegglen LM, et al. Potential mechanisms for cancer resistance in elephants and comparative cellular response to DNA damage in humans. JAMA. 2015;314:1850–60. https://doi.org/10.1001/jama.2015.13134
23. Vinckenbosch N, et al. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA. 2006;103:3220–5. https://doi.org/10.1073/pnas.0511307103
24. Allshire RC, Madhani HD. Ten principles of heterochromatin formation and function. Nat Rev Mol Cell Biol. 2018;19:229–44. https://doi.org/10.1038/nrm.2017.119
25. Aravin AA, Hannon GJ. Small RNA silencing pathways in germ and stem cells. Cold Spring Harb Symp Quant Biol. 2008;73:283–90. https://doi.org/10.1101/sqb.2008.73.058
26. Ecco G, et al. KRAB zinc finger proteins. Development. 2017;144:2719–29. https://doi.org/10.1242/dev.132605
27. Seczynska M, et al. Genome surveillance by HUSH-mediated silencing of intronless mobile elements. Nature. 2022;601:440–5. https://doi.org/10.1038/s41586-021-04228-1
28. Seczynska M, Lehner PJ. The sound of silence: mechanisms and implications of HUSH complex function. Trends Genet. 2023;39:251–67. https://doi.org/10.1016/j.tig.2022.12.005
29. Tchasovnikarova IA, et al. GENE SILENCING. Epigenetic silencing by the HUSH complex mediates position-effect variegation in human cells. Science. 2015;348:1481–5. https://doi.org/10.1126/science.aaa7227
30. Liu N, et al. Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. Nature. 2018;553:228–32. https://doi.org/10.1038/nature25179
31. Robbez-Masson L, et al. The HUSH complex cooperates with TRIM28 to repress young retrotransposons and new genes. Genome Res. 2018;28:836–45. https://doi.org/10.1101/gr.228171.117
32. Rosikiewicz W, et al. RetrogeneDB-a database of plant and animal retrocopies. Database (Oxford). 2017;2017. https://doi.org/10.1093/database/bax038
33. Cunningham F, et al. Ensembl 2022. Nucleic Acids Res. 2022;50:D988–95. https://doi.org/10.1093/nar/gkab1049
34. Szczesniak MW, et al. Towards a deeper annotation of human lncRNAs. Biochim Biophys Acta Gene Regul Mech. 2020;1863:194385. https://doi.org/10.1016/j.bbagrm.2019.05.003
35. Kodama Y, et al. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res. 2012;40:D54–56. https://doi.org/10.1093/nar/gkr854
36. Gibson R, et al. Biocuration of functional annotation at the European nucleotide archive. Nucleic Acids Res. 2016;44:D58–66. https://doi.org/10.1093/nar/gkv1311
37. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74. https://doi.org/10.1038/nature11247
38. O'Leary NA, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–745. https://doi.org/10.1093/nar/gkv1189
39. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. https://doi.org/10.1038/nmeth.1923
40. Wang L, et al. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012;28:2184–5. https://doi.org/10.1093/bioinformatics/bts356
41. Patro R, et al. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9. https://doi.org/10.1038/nmeth.4197
42. Shao Y, et al. GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. Genome Res. 2019;29:682–96. https://doi.org/10.1101/gr.238733.118
43. Altschul SF, et al. Basic local alignment search tool. J Mol Biol. 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2
44. Shen W, et al. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE. 2016;11:e0163962. https://doi.org/10.1371/journal.pone.0163962
45. Team, RC. R: a language and environment for statistical computing. R Foundation for Statistical Computing; 2021.
46. Constantin A, P.I. Ggsignif: R package for displaying significance brackets for 'ggplot2'. PsyArxiv. 2021. https://doi.org/10.31234/osf.io/7awm6. https://psyarxiv.com/7awm6
47. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-; 2016.
48. van den Brand T. (2022). ggh4x: Hacks for 'ggplot2'. R package version 0.2.3.
49. Min SH. (2023). smplot2: smplot2 - a package for statistical data visualization. R package. version 0.1.0.
50. Wickham H, Jennifer Bryan MA, McGowan WCL, François R, Grolemund G, Hayes A, Hester LHJ, Kuhn M, Pedersen T, Miller E. Stephan Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, Hiroaki Yutani. Welcome to the tidyverse. J Open Source Softw. 2019;4:1686.
51. Balasubramanian S, et al. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. Genome Biol. 2009;10:R2. https://doi.org/10.1186/gb-2009-10-1-r2
52. Subramanian S, Kumar S. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res. 2003;13:838–44. https://doi.org/10.1101/gr.1152803
53. Zhang Z, Gerstein M. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res. 2003;31:5338–48. https://doi.org/10.1093/nar/gkg745
54. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 1977;267:275–6. https://doi.org/10.1038/267275a0
55. Li WH, et al. Pseudogenes as a paradigm of neutral evolution. Nature. 1981;292:237–9. https://doi.org/10.1038/292237a0
56. Marques AC, et al. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 2005;3:e357. https://doi.org/10.1371/journal.pbio.0030357
57. Feng Q, et al. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell. 1996;87:905–16. https://doi.org/10.1016/s0092-8674(00)81997-2
58. Brosius J. Retroposons–seeds of evolution. Science. 1991;251:753. https://doi.org/10.1126/science.1990437
59. Szcześniak MW, et al. Primate and rodent specific intron gains and the origin of retrogenes with splice variants. Mol Biol Evol. 2011;28:33–7. https://doi.org/10.1093/molbev/msq260
60. Burki F, Kaessmann H. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. Nat Genet. 2004;36:1061–3. https://doi.org/10.1038/ng1431
61. Chen S, et al. Frequent recent origination of brain genes shaped the evolution of foraging behavior in drosophila. Cell Rep. 2012;1:118–32. https://doi.org/10.1016/j.celrep.2011.12.010
62. Rosso L, et al. Mitochondrial targeting adaptation of the hominoid-specific glutamate dehydrogenase driven by positive darwinian selection. PLoS Genet. 2008;4:e1000150. https://doi.org/10.1371/journal.pgen.1000150
63. Bryzghalov O, et al. Retroposition as a source of antisense long non-coding RNAs with possible regulatory functions. Acta Biochim Pol. 2016;63:825–33. https://doi.org/10.18388/abp.2016_1354
64. Marques AC, et al. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. Genome Biol. 2008;9:R54. https://doi.org/10.1186/gb-2008-9-3-r54
65. Tchasovnikarova IA, et al. Hyperactivation of HUSH complex function by Charcot-Marie-tooth disease mutation in MORC2. Nat Genet. 2017;49:1035–44. https://doi.org/10.1038/ng.3878
66. Hrecka K, et al. Vpx relieves inhibition of HIV-1 infection of macrophages mediated by the SAMHD1 protein. Nature. 2011;474:658–61. https://doi.org/10.1038/nature10195
67. Laguette N, et al. SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. Nature. 2011;474:654–7. https://doi.org/10.1038/nature10117
68. Chougui G, et al. HIV-2/SIV viral protein X counteracts HUSH repressor complex. Nat Microbiol. 2018;3:891–7. https://doi.org/10.1038/s41564-018-0179-6
69. Yurkovetskiy L, et al. Primate immunodeficiency virus proteins Vpx and Vpr counteract transcriptional repression of proviruses by the HUSH complex. Nat Microbiol. 2018;3:1354–61. https://doi.org/10.1038/s41564-018-0256-x
70. Carelli FN, et al. The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Res. 2016;26:301–14. https://doi.org/10.1101/gr.198473.115

71. Bestor TH. The DNA methyltransferases of mammals. Hum Mol Genet. 2000;9:2395–402. https://doi.org/10.1093/hmg/9.16.2395
72. Lander ES, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921. https://doi.org/10.1038/35057062
73. Xing J, et al. Alu element mutation spectra: molecular clocks and the effect of DNA methylation. J Mol Biol. 2004;344:675–82. https://doi.org/10.1016/j.jmb.2004.09.058
74. Palazzo AF, et al. mRNA nuclear export: how mRNA identity features distinguish functional RNAs from junk transcripts. RNA Biol. 2024;21:1–12. https://doi.org/10.1080/15476286.2023.2293339
75. Schmidt S, et al. Hypermutable non-synonymous sites are under stronger negative selection. PLoS Genet. 2008;4:e1000281. https://doi.org/10.1371/journal.pgen.1000281
76. Ying H, Huttley G. Exploiting CpG hypermutability to identify phenotypically significant variation within human protein-coding genes. Genome Biol Evol. 2011;3:938–49. https://doi.org/10.1093/gbe/evr021
77. Palazzo AF, Kang YM. GC-content biases in protein-coding genes act as an mRNA identity feature for nuclear export. BioEssays. 2021;43:e2000197. https://doi.org/10.1002/bies.202000197
78. Huang Y, et al. SR splicing factors serve as adapter proteins for TAP-dependent mRNA export. Mol Cell. 2003;11:837–43. https://doi.org/10.1016/s1097-2765(03)00089-3
79. Fraimovitch E, Hagai T. Promoter evolution of mammalian gene duplicates. BMC Biol. 2023;21:80. https://doi.org/10.1186/s12915-023-01590-6

Publisher's Note