

Review

# Protein-Coding Genes' Retrocopies and Their Functions

Magdalena Regina Kubiak and Izabela Makalowska \*

Department of Integrative Genomics, Institute of Anthropology, Faculty of Biology, Adam Mickiewicz University in Poznan, 61-614 Poznan, Poland; magdalena.kubiak@amu.edu.pl

\* Correspondence: izabel@amu.edu.pl; Tel.: +48-61-8295835

Academic Editors: David J. Garfinkel and Katarzyna J. Purzycka

Received: 25 February 2017; Accepted: 11 April 2017; Published: 13 April 2017

**Abstract:** Transposable elements, often considered to be not important for survival, significantly contribute to the evolution of transcriptomes, promoters, and proteomes. Reverse transcriptase, encoded by some transposable elements, can be used in *trans* to produce a DNA copy of any RNA molecule in the cell. The retrotransposition of protein-coding genes requires the presence of reverse transcriptase, which could be delivered by either non-long terminal repeat (non-LTR) or LTR transposons. The majority of these copies are in a state of “relaxed” selection and remain “dormant” because they are lacking regulatory regions; however, many become functional. In the course of evolution, they may undergo subfunctionalization, neofunctionalization, or replace their progenitors. Functional retrocopies (retrogenes) can encode proteins, novel or similar to those encoded by their progenitors, can be used as alternative exons or create chimeric transcripts, and can also be involved in transcriptional interference and participate in the epigenetic regulation of parental gene expression. They can also act in *trans* as natural antisense transcripts, microRNA (miRNA) sponges, or a source of various small RNAs. Moreover, many retrocopies of protein-coding genes are linked to human diseases, especially various types of cancer.

**Keywords:** retrotransposon; retrotransposition; retrocopy; retrogene; gene duplication; genome evolution

---

## 1. Introduction

A large fraction of human and other eukaryotic genomes consist of sequences that originated, directly or indirectly, as a result of transposable elements (TE) activities. Most of these genomic elements are considered to be nonessential to survival. However, TEs have a significant influence on genome evolution. TEs are probably most commonly known as recombination hotspots; however, they also contribute to the evolution of promoters and proteomes. Considering the direct contribution to proteomes, two scenarios exist: the coding potential of a TE is “domesticated” to perform host cellular function or TE-derived sequences are exapted into a coding portion of existing genes to generate novel protein variants [1]. One of the most impressive examples of a domesticated TE is the recombination-activating protein RAG1 [2]. This protein was derived from the transposase gene of a *Transib* DNA transposon 500 million years ago [3]. Another great example is the domestication of the *gag* gene. As many as 85 *gag*-like genes might exist in the human genome [4]. The second direct contribution to proteomes is exaptation, i.e., cooption of different TE fragments to a new role. We call exapted those TE fragments that became a part of a coding sequence (CDS) but do not code for a protein domain attributed to their original function. In humans *Alu* sequences are major donors of exons, but exons acquired from other elements, such as LINEs, endogenous retroviruses, and DNA transposons, have also been reported [5]. Examples of the exaptation of an endogenous

retrovirus envelope (*env*) gene are the primate genes *Syncytin-1* and *Syncytin-2*, which might be involved in the formation of the placenta [6].

Reverse transcriptase (RT), encoded by some TEs, can be used in *trans* to produce a DNA copy of any RNA molecule in the cell. This copy, reintegrated into the genome, will most likely be “dead on arrival” because none of the regulatory elements can be copied in RNA-mediated gene duplication. Therefore, these sequences are often called retropseudogenes or processed pseudogenes. Although the majority of these retrocopies are in a state of “relaxed” selection and remain “dormant” because they are lacking regulatory regions, many become functional. The evolutionary path of these functional retrocopies, called retrogenes, is not uniform. In the course of evolution, they may undergo subfunctionalization and share their function with their parent [7], develop a brand new function (neofunctionalization) [7], or replace their progenitors [8].

Retrogenes were long considered to be unimportant copies, but are currently called “seeds of evolution” since they have made a significant contribution to molecular evolution [9]. It has been shown that retrogenes play an important role in the diversification of transcriptomes and proteomes and may be responsible for a wealth of species-specific features. Some of these differences are highly important in medical research and may be the reason why results from animal studies cannot be transferred to humans. For example, the functional mouse retrogene *Rps23r1* reduces Alzheimer’s beta-amyloid levels and tau phosphorylation [10]. This particular retrogene is rodent-specific and does not exist in the human genome. Another elegant example of the functional phenotypic effect of retroposition was demonstrated by *fgf4* retrogene studies. Insertion of this retrogene is responsible for chondrodysplasia in dogs. All breeds with short legs are carriers of the *fgf4* retrogene [11].

The discovery that retro sequences considered “junk DNA” may be functional and play a crucial role in shaping genome-specific features was one of the most surprising breakthroughs of human and other genome analyses. A large number of studies were recently performed to explore these unique sequences, yet our knowledge of retrogenes’ evolution is exceptionally limited. In this review paper, we present recent studies aiming to decipher the functions of transcriptionally active retrocopies of protein-coding genes.

## 2. Retrotransposons as a Source of Cellular Reverse Transcriptase

The possibility of the reverse flow of genetic information from RNA to DNA was initially proposed in research conducted on the chicken Rous sarcoma virus [12]. The suggestion that the viral RNA genome can be transcribed into a DNA sequence and integrated into the host genome, together with the subsequent discovery of adequate enzymes [13,14], received the Nobel Prize in 1975. At that time, various mobile genomic elements, such as *Ty* in yeast [15] and LINE1 in human [16], were found to encode a reverse transcriptase, which was quickly associated with their mobilization abilities. This abundant group of “jumping genes”, called retrotransposons, has been divided into two families characterized by the presence or absence of flanking long terminal repeats (LTRs) (Figure 1). The first group includes retroviral-like elements with LTRs, and the second consists mainly of long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) without LTRs. The LTR and non-LTR retrotransposons can be further subdivided into autonomous retrotransposons, which encode proteins required for mobilization, and nonautonomous retrotransposons, which utilize the retrotransposition machinery of the others.

The origin of retroelements and the evolutionary relation between them and retroviruses are not currently clear. The major hindrance in evolutionary analysis is the lack of unitary molecular characteristics across all retrosequences. Despite this limitation, several approaches have been used to portray the relations, for example, comparison of the reverse transcriptase [17] or ribonuclease H domains [18]. According to these analyses, non-LTR retrotransposons are the oldest group of retroelements and might be derived from various sequences, for example, having a common ancestor with RNA viruses [17] or originating from the prokaryotic group of II introns, also called retrointrons [18,19]. In the case of LTR retrotransposons, the evolutionary path is even more ambiguous. Doubtless, they are closely related with retroviruses, which is reflected in the common structural traits and notable similarities during the retrotransposition process. Autonomous LTR

retrotransposons contain *gag* and *pol* genes encoding structural proteins required for the formation of the virus-like particles and enzymes involved in reverse transcription and incorporation of new copies into the genome. The lack of the envelope (*env*) gene, which enables recognition and infection of the host cell, is considered a main difference between retrotransposons and retroviruses. Although *env*-like genes have been found in some retrotransposons, their function is not fully understood [20,21]. Due to these facts, it was suggested that retrotransposons evolved from retroviruses that lost their infectious properties as a result of mutational inactivation of the genes responsible for intercellular movement. Alternatively, LTR retrotransposons could have arisen from ancestral non-LTR retrotransposable elements, with acquisition of the *env* gene providing an opportunity for retrovirus evolution [17–19].

### LTR retrotransposon



(a)

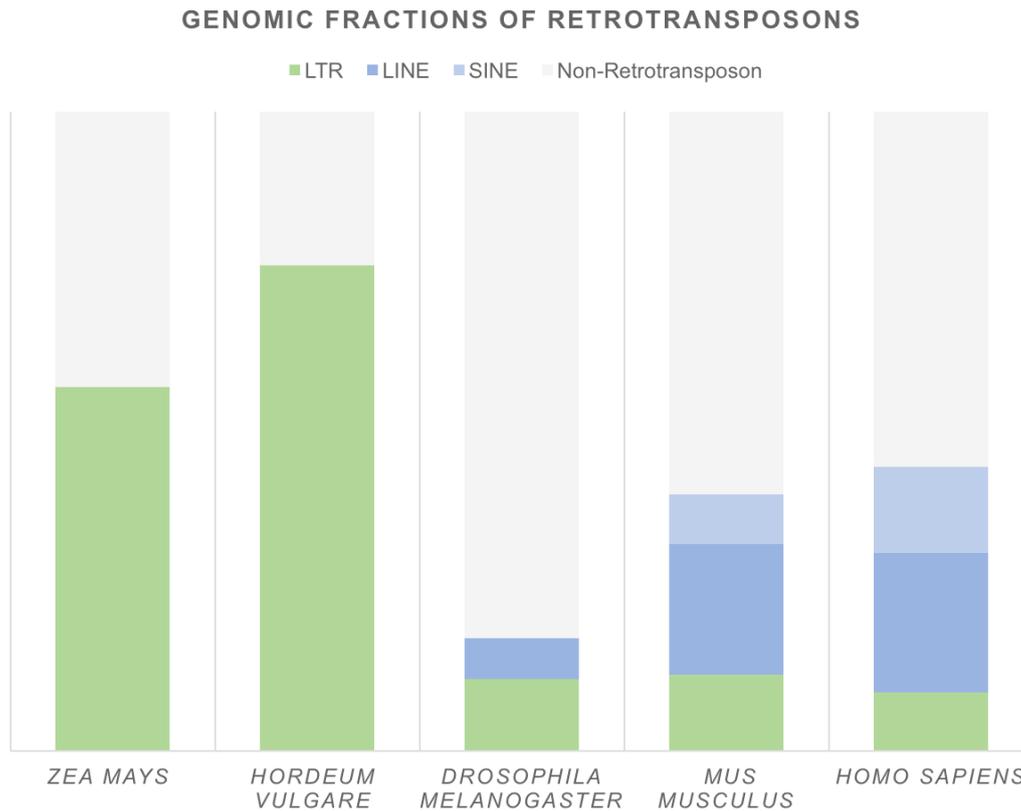
### non-LTR retrotransposon (LINE1)



(b)

**Figure 1.** Schematic representation of the structural organization of (a) long terminal repeat (LTR) and (b) non-LTR retrotransposons. A detailed description can be found in Sections 2.2 and 2.3.

Both LTR and non-LTR retrotransposons are abundant in eukaryotic genomes (Figure 2). In plants, the most prominent fraction is composed of LTR retrotransposons and is frequently associated with an enlarged genome size. A good example is *Zea mays*, in which insertions of retrotransposons have doubled the genome size during the past three million years of evolution [22]. LTR retrotransposons occupy at least 55% of the *Z. mays* and 76% of the *Hordeum vulgare* genomes [23–25] and compose up to 58% and 91% of the *Allium cepa* and *Asparagus officinalis* DNA contigs, respectively [26]. Non-LTR retrotransposons are less abundant and usually cover only a few percent of the whole genome. Nevertheless, analysis of 23 plant genomes indicated that structurally intact (and therefore potentially active) LINEs are present from model species, such as *Arabidopsis thaliana*, to very complex genomes, such as *Picea abies* [27]. In metazoa, compared to in plants, retrotransposons usually constitute a smaller genome fraction (Figure 2). A remarkable example is one of the model organisms, *Drosophila melanogaster*, in which retrotransposons compose approximately 17% of the genome. In mammals, non-LTR retrotransposons are the most common; they represent approximately 28% and 35% of the mouse and human genomes, respectively. In contrast, LTR elements compose only 12% and 9% of these genomes [28]. However, not all retrotransposon families are still active. For example, only three non-LTR retrotransposon families—long interspersed elements 1 (L1s), *Alu* elements, and SVAs (named after their composite parts: SINE-R, VNTR (variable number of tandem repeats), and an *Alu*-like sequence)—are actively mobilized in the human genome [29].



**Figure 2.** Genomic fractions of retrotransposons in selected genomes. Based on: *Zea mays* [23], *Hordeum vulgare* [25], *Drosophila melanogaster*, *Mus musculus* (genome version mm10), *Homo sapiens* (genome version hg38)—Repeat Masker online dataset [28].

### 2.1. Retrotransposition of Nonautonomous Retrotransposons and Gene Copies

While autonomous retrotransposons encode their own mobilization machinery, the generation of nonautonomous retrotransposons and retrotransposed gene copies has long remained unclear. In the early 1980s, the analysis of repetitive sequences originating from small nuclear RNA and 7SL RNA showed that some are flanked by short direct repeats [30]. Because analogous repeats were found in combination with endogenous retroviruses and other mobile elements, a similar mechanism involving the insertion of reverse-transcribed RNA into the genome was suggested.

The endogenous reverse transcriptase activity and its ability to retrotranspose cellular non-retrotransposon mRNA were tested in a minigene system in HeLa cells [31]. Copies of a reporter gene with all characteristics indicating RNA-dependent duplication, including lack of introns, the presence of a polyA tract, and short repeats flanking the sequence, were found. Furthermore, the experiment showed that retrotransposition of mRNA in mammalian cells is still an active process. Further experimentation with mouse and human cells transfected by vectors containing reverse transcriptase revealed functional differences between the enzymes from retroviruses and those from LINE retrotransposon [32]. These results demonstrated that, in contrast to retroviral reverse transcriptase, LINE RT could generate reverse transcripts from RNAs, which do not show any sequence specificity or similarity to the LINE themselves. Therefore, autonomous non-LTR retrotransposons might be the source of endogenous RT involved in the formation of other retrotransposed genetic elements. Experiments performed by the same group [33] confirmed that LINE1 can act in *trans* and give rise to new retroposed gene copies. However, LINE1 elements are more effective in *cis*, which could be a consequence of the close proximity of LINE1-derived mRNA and proteins during translation or the limited life-time of protein in the lack of LINE1 mRNA [34]. This conclusion also explains the abundance of LINE1 elements in the genome compared to retrotransposed genes, as it was demonstrated that no more than 0.05% of the retrotransposition events conducted by LINE1 are related to retrotransposed gene formation [34].

Because the majority of experiments were conducted on animals, mostly mammalian models, the knowledge of retrotransposition in plants is restricted. Moreover, LINE elements constitute only a small fraction of the plant genome; their retrotransposition and creation of new gene copies is limited, but possible [35]. Since the first plant retrocopy was reported in *Solanum tuberosum* in 1987 [36], many retrotransposed genes have been found in various plant species [37–40]. For instance, a recent analysis of *A. thaliana* identified 251 retrocopies, of which 216 were described as novel [40]. In the RetrogeneDB2 repository, 1821 retrocopies in 37 plant species were identified [41]. Interestingly, analysis of plant retrocopies showed that LTR retrotransposons might also be involved in retrocopy formation. One of the best documented cases is the *Bs1* retrotransposon, which participated in the creation of retrocopies of three different cellular genes in the *Z. mays* genome [42,43]. Moreover, several additional retrocopies showing the signatures of LTR-mediated retroposition were recently found in both invertebrates and vertebrates [44]. Tan and coworkers [44] showed that LINE1-linked retrotransposition is dominant in mammals, whereas in mosquito, zebrafish, and chicken, retrocopies are created by both LTR- and non-LTR-mediated mechanisms. All polymorphic retrocopies found in *Drosophila* were formed by LTR retroposition. This type of retrotransposition is also more common in plants. Recent analysis of retrocopies derived from circular RNA (circRNA) in the mouse genome has shown LTR sequences localized in the flanking regions, which may suggest LTR-mediated retrotransposition [45].

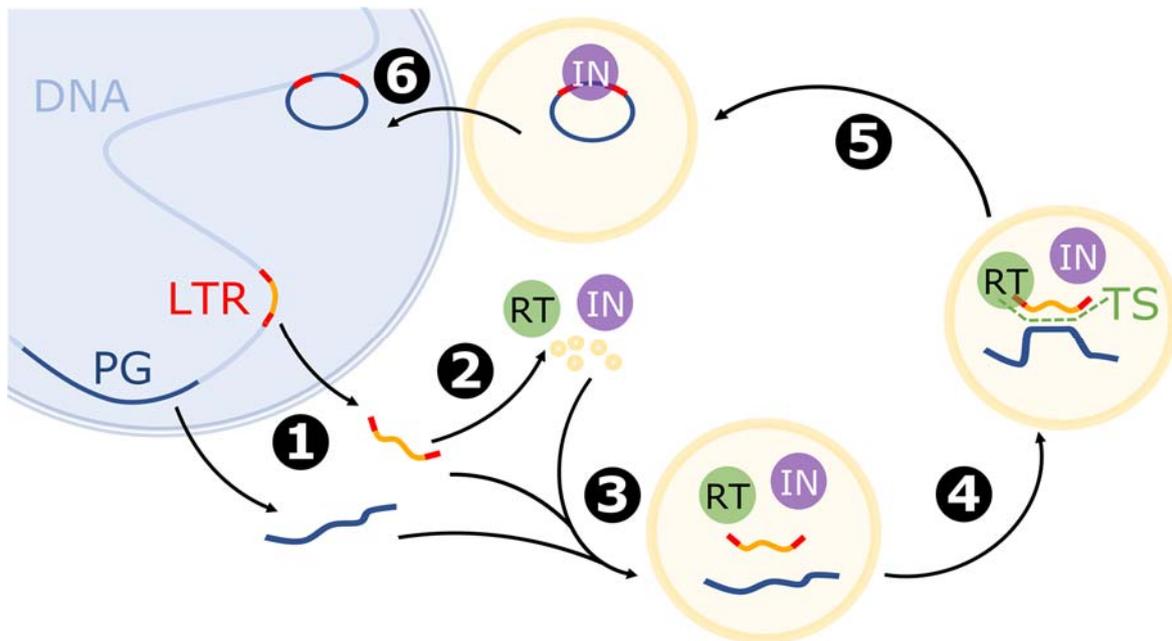
Although the majority of retrotransposons have lost the capacity for mobilization—for instance, no more than 100 L1 in the human genome encode functional retrotranspositional machinery [46]—they are found to play various roles. They can shape genomes by acting as recombination hotspots or participating in exon-shuffling. They may also be used as new regulatory elements and alter the chromatin structure, thus influencing neighboring gene expression [47]. However, due to the integration of retrotransposons in random sites and the high likelihood of deleterious effects of insertion [48], different mechanisms of retrotransposition regulation evolved. Multilevel pathways are present in cells, starting from epigenetic silencing of retrotransposons by genomic DNA methylation and histone modification [49] to post-transcriptional positive and negative regulation [50].

## 2.2. LTR Retrotransposon-Based Transposition

The autonomous LTR retrotransposons are a diverse group; however, some common traits exist. They usually span several kilobases, are flanked by long terminal repeats, and contain promoters for RNA polymerase II localized in LTRs; however, only one RNA Pol II transcribes the retroelement. They encode at least two genes, which can overlap, be separated by terminal codons, or be fused into a single open reading frame [21,51]. One of the genes, *pol*, encodes various enzymatic domains, including reverse transcriptase and integrase, while the second, *gag*, produces structural proteins involved in the formation of virus-like particles (VLPs) (Figure 1a).

The mechanism of retrotransposition is similar to that observed in retroviruses. The process begins with transcription of the LTR retrotransposon by RNA polymerase II, after which the newly synthesized RNA is transported to the cytoplasm [20,51]. Next, translation and formation of VLPs occurs. Reverse transcriptase, integrase, and RNA molecules are typically packed in VLPs, and by chance cellular mRNA may also be encapsulated [20,44]. Reverse transcription usually starts by annealing tRNA to a primer binding site near the 5' LTR. The microsimilarities between the LTR retrotransposon and cellular mRNA allow for template switching between these two molecules and therefore for incorporation of an mRNA sequence into the emerging cDNA [44]. Moreover, stretches of microsimilarity can appear in multiple places along the gene. Theoretically, template switching can occur several times, but frequently only a limited part of the parental mRNA is reverse transcribed. Furthermore, second strand synthesis occurs from a polypurine tract near the 3' LTR. Next, the LTR ends of the new cDNA are bound by integrase (IN) and the VLP localizes a nearby nucleus. Finally, the cDNA-IN complex is transported to the nuclei, and integrase cuts the cellular DNA and joins the released ends with LTRs from the retrotransposed copy [51]. The complete cDNA is finally integrated into the genome, generally at a random site (Figure 3). In contrast to

LINE1-mediated retroposition, new retrocopies acquire long tandem repeats and thus a promoter sequence, which enables further transcription and retrotransposition [44].



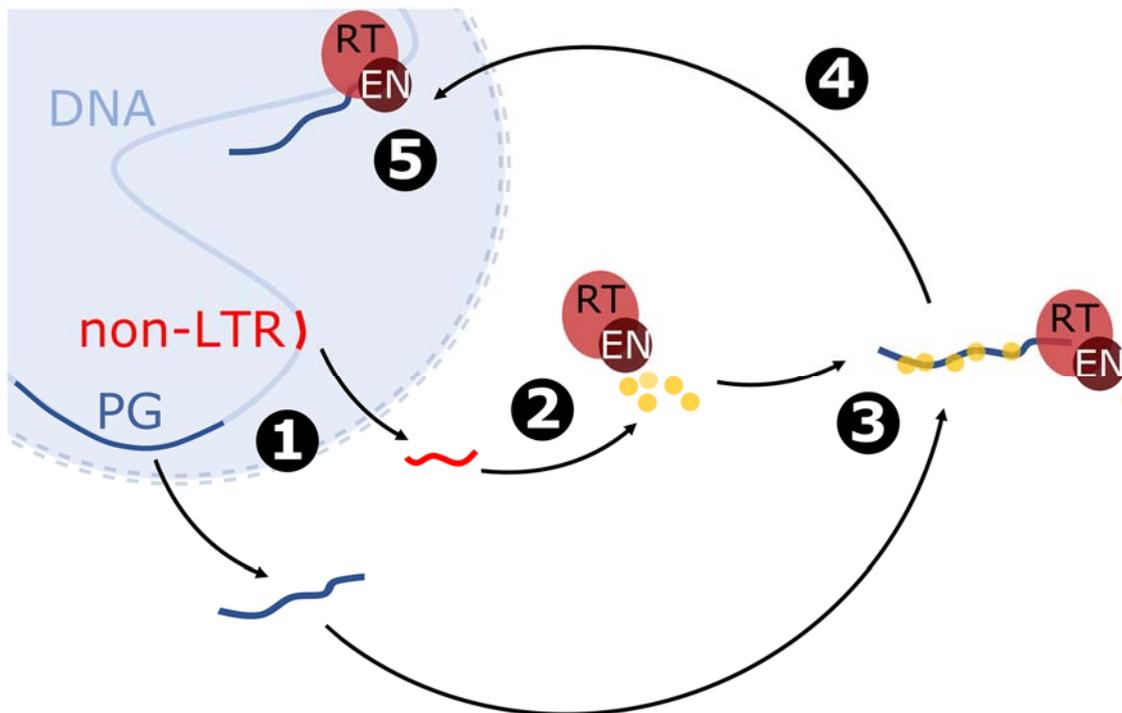
**Figure 3.** Model of mRNA retrotransposition mediated by the LTR retrotransposon (based on [20,44,51]). The blue sphere represents the nucleus. Stages: 1: Transcription of the LTR retrotransposon (LTR) and parental gene (PG). 2: Translation of LTR-encoded proteins, including reverse transcriptase (RT), integrase (IN), and proteins building virus-like particles (VLPs). 3: Formation of VLP with LTR-derived mRNA and parental gene mRNA. 4: Reverse transcription and template switch (TS). 5: Formation of the cDNA-IN complex. 6: Translocation and integration of chimeric cDNA into the genome.

### 2.3. Non-LTR Retrotransposition Mechanism

The most abundant non-LTR retrotransposon in mammals is LINE1 (L1); however, the set of the elements is quite variable between individuals [52]. For instance, a new insertion of L1 in humans occurs in between 1/95 and 1/270 of newborns [53]. Intact L1s are up to 6 kb in length and contain an internal promoter with sense and antisense activity localized in the 5' untranslated region [54]. We can distinguish two open reading frames: one encoding a smaller RNA-binding protein with chaperone activity [55] and the other encoding a larger protein with endonuclease and reverse transcriptase domains. Mutational analysis of these two open reading frames showed that both proteins are required for retrotransposition [56]. Additionally, a third, primate-specific open reading frame was recently found on the antisense strand [57], but the biological function of the transcript is not clear. The majority of L1 elements also contain a polyA tract in their 3' region (Figure 1b).

As experiments have shown, the LINE1 element can be transcribed by RNA polymerase II to mRNA, which is further used in translation and reverse transcription processes (Figure 4) [31,56]. The L1 mRNA is transported to the cytoplasm and translated, and the newly synthesized proteins interact with RNA in such a way that the one with enzymatic domains binds the 3' end, while the smaller ones with chaperone activity are attached along the entire RNA molecule [58]. The spatial distribution and proximity may influence the *cis* preference of L1 retrotransposition. However, small chaperone proteins encoded by L1 may also bind to other RNA molecules. For example, LINE1s are responsible for the *Alu* (from 7SL RNA) [59], snRNA [60] and mRNA retrocopies [33]. LINE1s can also mobilize primate-specific SVA retrotransposons [61,62], hY RNA [63] or even human endogenous retroviruses [64]. However, the observation that removal of the polyadenylation signal results in a loss of retrotransposition [65] indicates that the presence of a polyA tail may be one of the requirements for mobilizing RNA via a LINE1-derived mechanism.

The ribonucleoprotein particle formed by RNA and LINE1-encoded proteins needs to get close to the chromosome DNA, where the target-primed reverse transcription occurs. However, it is not clear how the import to the nucleus happens. It was proposed that retrotransposition occurs during cell division when the nuclear membrane is disrupted [66]. However, there is also evidence for retrotransposition in non-dividing cells [67,68] and, therefore, cell division may not be necessary for the ribonucleoprotein transport near genomic sequence. Nevertheless, the reverse transcriptase domain begins synthesis of a new DNA strand on an RNA template using the free 3'-OH resulting from endonuclease cleavage of the genomic sequence [69]. Alternatively, non-classic L1 insertion into pre-existing gaps in DNA can occur [70]. The subsequent steps have not been thoroughly studied, but it was proposed that a second nick is generated downstream, thereby enabling second DNA strand synthesis. The final step may also include creating target site duplication (TDS) of variable length. Additionally, in this process, the 5' site of the newly arisen copy is often truncated.



**Figure 4.** Model of mRNA retrotransposition mediated by non-LTR retrotransposon referring to LINE1. The blue sphere represents the nucleus; the dotted line shows the probability of nucleus membrane disruption. Description of the stages: 1: Transcription of the non-LTR retrotransposon (non-LTR) and parental gene (PG). 2: Translation of non-LTR-encoded proteins, including proteins with reverse transcriptase (RT) and endonuclease (EN) domains and RNA-binding proteins with chaperone activity. 3: Formation of ribonucleoprotein particle by binding of the proteins to the polyA tail of the parental gene mRNA. 4: Transport of ribonucleoprotein particle near the genome. 5: Reverse transcription and incorporation of the parental gene.

### 3. Number of Retrocopies across Genomes

The identification of retrocopies has been the subject of many studies and with the increasing number of sequenced genomes, as well as data from high-throughput experiments, new retrocopies are being discovered in various organisms. Examples of recent analyses include the comparative genomic study of the green algae retrogene repertoire [71] and inter-specific segregating retrocopies in cynomolgus and rhesus monkeys [72].

Basic annotations for retrotransposed gene copies, containing the localization and identified parental gene, are incorporated into several online databases. However, the most frequent description of these sequences is “processed pseudogene” or just “pseudogene”, which could be

misleading as it does not directly indicate origin by retrotransposition. The major difficulty in annotation is the lack of specific sequence motifs and the high rate of mutation accumulation in retrocopies. Even the most obvious feature, a lack of introns, does not have to be preserved, as retrogenes are known for intron incorporation [73]. Moreover, they may also acquire novel exons or be fused with another gene and used as an alternative exon [74]. On the other hand, intronless genes are not always an outcome of a retroposition event. For instance, single-exon histone-encoding genes are believed to have originated in prokaryota [75]. Other retrocopy traits, such as the polyA tails and insertion site repeats, are found only in evolutionarily young retrocopies. Regardless, the best approach to retrocopy identification so far is based on alignment of known protein-coding sequences to the genome.

The Ensembl genome browser is one of the major resources of publicly available genomic information. In addition to datasets analyzed in an automated way, manually curated data from the HAVANA project are included [76]. Annotation of retrocopies, called “processed pseudogenes”, is based on an imperfect alignment between the genome and protein sequences, which enables multi-exon and single-exon gene models to be obtained. These single-exon annotations are interpreted as intronless, and therefore possibly retrotransposed, gene copies [77]. A similar method was applied in the PseudoPipe pipeline to predict all types of pseudogenes in the eukaryotic genomes stored on the Pseudogene.org server [78]. A slightly different approach was applied in UCSC Genome Browser, which besides Ensembl is the largest collection of genomic annotations. Here, the method for retrocopy identification was based on the RetroFinder pipeline, in which mRNA sequences were aligned to the genome [74,79]. Similar to the previous approaches, multi-exon and single-exon hits were obtained. The intronless candidates for retrocopies were then passed through multistage feature-based selection. For instance, a number of ancestral genes and putative retrocopy exons, as well as the presence and position of the polyA tail, were considered. PseudoPipe and RetroFinder, together with the HAVANA annotations, were used to produce a high-confidence dataset of pseudogenes, including retrocopies, in the human and mouse reference genomes in the GENCODE project [80].

More conservative approaches for retrotransposed gene copy annotation were applied in retroduplication-dedicated databases. The main purpose of utilizing more restrictive prediction criteria is the minimization of false positive results. On the other hand, they also provide expanded information enriched by potential function, interspecies conservation, and expression studies. There are three retrocopy-specific databases available. The first, HOPPSIGEN, is focused on human and mouse genomes [81]. A wider range of retrocopies, identified in six primate genomes, is available in RCPedia [82]. In addition, RetrogeneDB2 stores retrocopy information for 62 animal and 37 plant species [41]. Moreover, this database includes expression validation based not only on RNA-seq experiments but also on expressed sequence tag (EST), transcription start sites (TSS), and chromatin immunoprecipitation sequencing (ChIP-Seq) data. Another distinguishing feature of RetrogeneDB2 is the inclusion of data from retrocopy number variation studies that show retrocopies’ indel frequencies across human populations [83].

The numbers of predicted retroposed pseudogenes across databases are summarized in Table 1. Although all the described methods rely on the alignment of known multi-exon coding genes (in the form of nucleotide or amino acid sequences) to the genome, the number of retrocopies differs because of the distinct filtering strategies, as well as the applied tools and parameters. For instance, the HOPPSIGEN dataset was obtained using the BLAST alignment tool [84], while in RetrogeneDB2, database retrocopies were identified on the basis of sequence alignments generated by LAST [85] (Table 2). Additional information about putative retrogenes can also be acquired from intronless gene databases. For instance, the IGD database [86] and SinEx DB [87] contain sets of single-exon coding genes for human and 10 mammalian genomes, respectively.

**Table 1.** Comparison of retrocopy sets available in the described databases. In addition, the number of retrogenes annotated in the *Homo sapiens*, *Pan troglodytes*, *Macaca mulatta*, and *Mus musculus* genomes is shown.

Database	Plants	Animals	Number of Retrocopies			
			<i>Homo sapiens</i>	<i>Pan troglodytes</i>	<i>Macaca mulatta</i>	<i>Mus musculus</i>
Non-specific databases						
Ensembl <sup>1</sup>	+	+	10,815	69	182	6999
UCSC <sup>2</sup>	+	+	13,742	–	–	18,456
GENCODE <sup>3</sup>	–	+	9074	–	–	6151
Pseudogene.org <sup>4</sup>	+	+	8739	7505	–	9809
Retrogene-dedicated databases						
HOPPSIGEN	–	+	5206	–	–	3428
RCPedia	–	+	7831	7733	7544	–
RetrogeneDB2	+	+	4611	3285	2377	4148

<sup>1</sup> Ensembl genome browser release 86 dataset filtered by “processed pseudogene” in BioMart; <sup>2</sup> UCSC RetroGenes v9 (*H. sapiens*) and v6 (*M. musculus*) dataset statistics; <sup>3</sup> GENCODE human (Release 25, GRCh38.p7), Mouse (Release M11, GRCm38.p4), “processed pseudogenes” with “gene” and “level 1” statuses were chosen from the comprehensive gene annotation GTF file; <sup>4</sup> class “processed” was selected for all organisms: chimp build 50 (CHIMP2), mouse build 84 (GRCm38), human build 83 (NCBI38).

**Table 2.** Approaches used in retrocopy-specific databases.

Database	Sequence Aligned to the Genome	Tool
HOPPSIGEN	gene coding sequence	TBLASTX
RCPedia	entire transcript	BLAT
RetrogeneDB2	protein	LAST

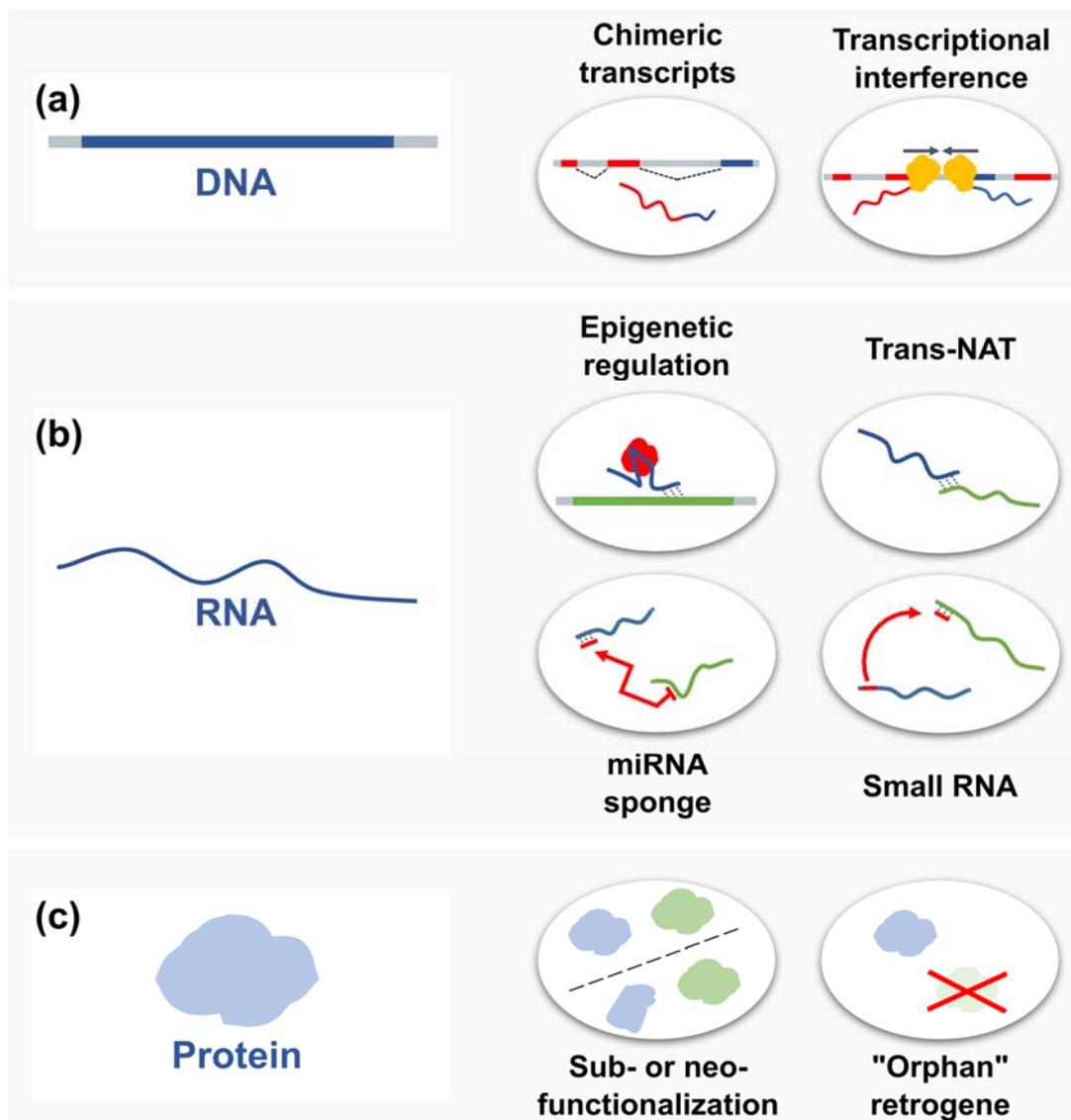
#### 4. Molecular Functions of Genes Retrocopies

In the retroposition process, the parental regulatory elements are usually not inherited, and the new copy often slowly decays and is silenced by the accumulation of degenerative mutations in a process called pseudogenization or nonfunctionalization. However, a large number of transcripts originating from retrocopies were found in cells, which suggests the acquisition of active promoters. Retrogenes may use regulatory machinery of nearby genes or utilize distant CpG-rich sequences [88] and occasionally parts of their own sequence [89] to promote transcription. For instance, when the parental gene has multiple transcription start sites and the one located upstream of the promoter region is used, the retrotransposed transcript may contain a prominent part of the core promoter. A good example of such a case is the *PABP3* retrogene. Analysis of the sequence similarity between the abovementioned retrocopy and the parental gene showed high conservation of the 5' upstream region, suggesting that the retrogene arose from a gene transcript containing a fragment of the promoter [89].

After gene duplication, the genome contains two similar genes, which may potentially have the same function. This initial functional redundancy, depending on the deleterious or beneficial impact on the organism, may be eliminated or preserved during evolution. Signatures of purifying selection, like intact open reading frames or lower rates of non-synonymous to synonymous mutation, are frequently used as evidence to support the putative functionality of the new copy [7]. However, sequence conservation is not direct evidence of functionality. Experiments focused on the molecular characteristics of retrogene products are usually necessary to confirm expression and to assess the function of the analyzed molecule.

The development of high-throughput sequencing methods has enabled wide characterization of genomes, transcriptomes, and even epigenomes of various organisms, as well as particular organs

and tissues. Sequencing experiments provide information about gene expression, methylation patterns, and DNA–protein interactions, which may be used to create a complex description of retrogene functionality and regulation. However, the analysis of short reads produced by next-generation sequencers is challenging because of the difficulty in assigning them to one of two or more highly similar sequences, such as the parental gene and its retrocopies. Nevertheless, new technologies that produce long reads from a single molecule have improved, and retrocopy expression analyses should eventually become less problematic. While analytical problems may always exist, our current knowledge about retrocopies and their function is quite extensive, and many different examples are well documented (Figure 5).



**Figure 5.** Selected functions of retrogenes. (a) On the DNA level, for example, retrogenes may be used as alternative exons and create chimeric transcripts or be involved in transcriptional interference; (b) as RNAs, they can participate in epigenetic regulation of parental gene expression and act as trans-natural antisense transcripts, microRNA (miRNA) sponges or a source of various small RNAs; (c) retrogenes can also encode proteins, which might retain the parental gene function (subfunctionalization), evolve a new function (neofunctionalization), or even functionally replace the parental gene (“orphan” retrogenes).

#### 4.1. Protein-Coding Retrogenes

Retention of highly similar expressed sequences is often disadvantageous; therefore, conservation of the same gene function in its retrocopy is rare. Zhang [90] suggested that duplicates could be possible only in cases of highly demanded genes, such as rRNAs and histones. Although sharing the same function appears to be a natural consequence of gene duplication, retrogenes are often regulated in a different way than their ancestor genes because of distinct regulatory mechanisms. The main consequence of differences in regulatory machinery is spatio-temporal division of expression. In the most popular duplication-degeneration-complementation (DDC) model, the parental gene and retrocopy subdivide the ancestral function [91]. This mode of retrocopy evolution is called subfunctionalization.

One interesting but complex example of retrogene evolution is illustrated by a pair of retrocopies in *A. thaliana* [92]. The parental gene *CYP98A3* encodes the meta-hydroxylase engaged in the plant phenolic pathway and lignin biosynthesis. Two retrocopies of this gene, *CYP98A8* and *CYP98A9*, encode similar enzymes, which specialize in 3'- and 5'-hydroxylation of derivatives of spermidine localized in the pollen coat and wall. Analysis of the evolutionary history of the *CYP98* family shows that in *Brassicaceae*, the parental gene *CYP98A3* was retrotransposed and the *CYP98A8/9* ancestor retrogene went through tandem duplication [92]. Through further evolution, the loss of one copy was observed in some lineages; therefore, the remaining copy, *CYP98A8*, preserved the 3'- and 5'-hydroxylase activity. However, in the *A. thaliana* lineage, two copies were conserved, and subdivision of function occurred. *CYP98A9* acts as the 3'-hydroxylase, while *CYP98A8* acts as the 5'-hydroxylase. Moreover, the authors suggest that *CYP98A9* retroprotein may have developed an additional function and play a role in flavonoid metabolism [92].

An example of subfunctionalization in humans is the cell cycle gene *CDC14B* and its retrocopy, *CDC14Bretro* [93]. Retrotransposition of the *CDC14Bpar* transcript occurred approximately 18–25 million years ago, and the ancestral function was probably conserved until the separation of African and Asian apes. However, in the African apes' ancestor genome, several mutations in the 5' end of retrogene sequence were fixed, and subcellular localization of the encoded protein was shifted from the microtubule to endoplasmic reticulum [93]. The authors suggest that the relocation of the retrogene protein was due to a change in substrate and/or interaction partners and was related to the novel function development and specific testis/brain expression. Thus, this example may not be subfunctionalization but rather acquisition of a novel function that was previously not reported for the parental gene, i.e., neofunctionalization [94]. Another example where it is difficult to differentiate subfunctionalization from neofunctionalization is the *RAB6C* retrogene, which was retrotransposed approximately 21–25 million years ago in primates from the *RAB6A* gene [95]. In contrast to the parental gene, it is expressed in the centrosome, whereas the ancestor protein is found in the Golgi apparatus. This subcellular shift is probably a result of C-terminal extension impeding interactions with the Golgi. A *RAB6C* depletion experiment resulted in tetraploidization and duplication of the centrosome. Therefore, the retrogene-encoded protein developed a new function and is perhaps responsible for controlling cell cycle progression [95].

A retrogene-encoded protein may participate in an ancestral or new metabolic pathway, as shown above, but the process of retrocopy translation itself can also have an impact on the parental gene. An interesting example of this phenomenon is the connexin 43 (*Cx43*) gene and its retrocopy. Both encode proteins; however, retrogene expression is limited to breast cancer [96,97]. The *Cx43* retrogene has an intact open reading frame and encodes a protein of the same size as the parental gene, yet the ancestral function is not fully retained. While the parental *Cx43* gene is involved in cell growth control and intercellular communication, the retrogene seems to be not engaged in the second one [96]. Interestingly, further research indicated that the translational machinery preferentially binds to the retrogene, causing a shift in the parental gene mRNA from a polyribosome to monoribosome fraction, resulting in decreased expression of *Cx43* [97]. Silencing of the retrogene resulted in increased *Cx43* RNA and protein levels, supporting the regulatory role of the retrogene [97].

Another study of tumor-suppressor gene *TP53* suggested that gene duplicates, which arose via retrotransposition, play a role in the reduction of cancer risk in elephants [98]. Two of 19 retrogenes of the *TP53* gene, *TP53RTG12* and *TP53RTG19*, can be translated and enhance the DNA-damage response. In comparison to the parental *TP53* protein, there are three significant differences: truncation in the DNA binding domain, lack of a nuclear localization signal, and lack of an oligomerization domain. However, the interaction motif, which enables binding to a negative regulator, is conserved. Therefore, the researchers suggested two putative models of function for this protein. The retrogene-encoded proteins may bind to and block a *TP53* negative regulator or may directly bind to the parental gene protein and prevent its ubiquitination [98].

A retrogene may not only share a function with the parental gene but may become a functional replacement after pseudogenization or deletion of its progenitor. The first so-called “orphan” retrogenes were discovered as a result of a comparative analysis of worm, chicken, and human genes. All 25 such cases identified in the human genome represent known and well-studied genes that were not previously recognized as retrocopies. Moreover, seven of them are associated with various human diseases, including diabetes, attention-deficit/hyperactivity disorder, congestive heart failure, and Huntington’s disease. One of them, linked to hereditary spastic paraplegia, is the *CHMP1B* retrogene encoding chromatin-modifying protein 1B. The parental gene was pseudogenized in the ancestor of Old World and New World monkeys, but it is still active in rodents [8]. Another study identified a partially overlapping set of 10 “orphan” retrogenes [99]. Retroduplication and loss of parental genes was also found as a major mechanism involved in genome evolution and the generation of intronless genes in tunicates [100].

Another scenario of retrocopy function is dosage compensation when the level of the parental gene product is, for whatever reason, insufficient. For example, two testis-specific retrogenes, *RPL10L* and *RPL39L*, may compensate for their parental genes, which are inactivated during spermatogenesis [101]. A similar mechanism was proposed for the *HNRNP G-T* retrogene, which may functionally replace the parental protein in the course of meiosis [102].

Evidence of retrogene translation was also found during high-throughput data analysis. Because protein levels correlate with the levels of mRNA associated with polyribosomes, Mascarenhas and colleagues [103] analyzed the polyribosome loading of all RNA classes. An RNA sequencing experiment performed on cytosolic extracts and polyribosomal fractions showed that 18 pseudogenes exhibit significant polyribosome enrichment, which may suggest protein-coding potential [103]. Sixteen of these pseudogenes were found to be retrocopies.

Direct evidence of retrogene translation may be obtained via mass spectrometry. One interesting example of proteomic data utilization was an attempt to improve and further refine mouse genome annotations [104]. Analysis of 10.5 million tandem mass spectra enabled confirmation of the translation of known genes as well as identification of new protein-coding genes. Unique peptide hits were reported for nine retrocopies. One of them, retrotransposed from peptidylprolyl isomerase A (*PPIA*) gene, has two protein-coding variants that differ in the 5’ region. Intriguingly, none of these translated retrogenes have orthologs in the human genome [104]. Tandem mass spectra were also used in the proteomic profiling of 30 human adult and fetal tissues and primary hematopoietic cells [105]. In these studies, translation of more than 17,000 known protein-coding genes and 808 novel coding regions, including 140 pseudogenes, was confirmed. Although the authors did not discriminate between pseudogene subtypes, retrogenes can be identified in this dataset. For instance, nine peptide sequences were matched uniquely with the fibrillar-like 1 (*FBLL1*) retrogene. Some of the identified pseudogenes, like MAGE family member B6 pseudogene 1 (*MAGEB6P1*), had common for retrocopies testes-specific expression, while others, like voltage-dependent anion channel 1 pseudogene 7 (*VDAC1P7*), had broad expression patterns [105].

#### 4.2. Consequences of Retrogene Insertion for the Host and nearby Genes

Across the 84,483 retrogenes annotated in 62 animal species deposited in RetrogeneDB, approximately 20% (18,468) are inserted into the intron of another gene. In the case of the human genome, this proportion is even larger, and so-called nested retrocopies constitute 44% of the total.

As mentioned previously, a retrocopy may use a promoter of the host gene to become transcriptionally active. Retrocopies are also frequently found in gene-rich and actively transcribed chromatin regions [88]. Depending on the position of the retrocopy, different destinies for the emerging transcripts are proposed. In the “hitchhike” scenario, a retrocopy inserted close to the 5′ end of the host gene uses a 5′ untranslated region (UTR) for its own transcription without a disruption of the host gene functions [7,88]. However, the production of resulting chimeric transcripts may theoretically reduce the normal host gene transcription level. Alternatively, a retrocopy insertion near the 3′ end of the host gene may produce a new exon for a new splice isoform [88].

The first reported chimeric transcript generated by retrocopy insertion into a previously existing gene was *jingwei* observed in *Drosophila* [106]. A decade later, a fusion gene was found in a vertebrate during a study focused on resistance to human immunodeficiency virus type 1 (HIV-1) [107]. Interestingly, retroposition of cyclophilin A (*PPIA*) into the *TRIM5* gene, which occurred after the divergence of New World and Old World monkeys, resulted in the origin of a novel protein that was probably able to attach ubiquitin to HIV-1 virion proteins [107]. However, chimeric transcript creation is not a widespread phenomenon across the human and mouse genomes. Baertsch and colleagues [74] analyzed 726 highly expressed retrocopies and identified only 34 cases as potential gene fusions. In another study, new chimeric transcripts of 13 human and 14 mouse retrocopies, together with the upstream exons of the host gene, were identified [99]. As an example, the authors present the mouse *Taf9* retrogene, which uses the first two exons of the *Ak6* gene to become active.

The retrogene or part of it may also be used as an alternatively spliced exon of the host gene or a new 3′ exon of a nearby gene. The *BRCA1* gene, for example, has an internal retrogene-derived exon, which generates a 22 amino acid cassette. In *SCP2*, *HLA-F*, and *KIAA0415*, alternatively spliced 3′ end exons arising from antisense retrocopy insertions were found [74]. Chimeric transcripts were also found in an analysis focused on human-specific retrocopies harboring 5′ CpG islands [108]. One is composed of exon 8 of the *RNF13* gene and *TMEM183A-r* retrogene and another is composed of the *HSF2BP* gene and *H2BFS* retrogene. In both cases the retrogene is located antisense to the parental gene orientation; therefore, these chimeric transcripts may potentially form RNA–RNA duplexes with their progenitors and participate in their regulation. The third chimeric transcript identified in these studies, created between the *VRK2* gene and *EIF3P3* retrogene, was expressed only in malignant prostate cancer cell lines [108]. Retrocopies expressed as a part of chimeric transcripts were also observed in plants. For instance, analysis of the retrogene repertoire in the genome of rice showed that more than one-third of the identified retrocopies recruit additional coding exons from nearby genes [38]. Retrogene-derived exons are found in many other plants’ proteins; for example, a polygalacturonase-inhibiting protein encoded by a chimeric gene acting against *Aspergillus niger* polygalacturonase [38,109].

As already mentioned, the origination of chimeric transcripts may influence the expression level of a host gene. However, this is not the only way in which retrocopies regulate the expression of other genes. A group of *cis*-regulatory transcription-related functions was proposed for neighboring genes [110]. First, two genes localized in the same genomic locus and transcribed from independent promoters may directly impede each other’s transcriptional processes [110]. This suppressive influence, called transcriptional interference, is a result of interactions between the dominant and sensitive promoters of overlapping genes. According to the promoter orientation and arrangement, several mechanisms of this process were proposed, including promoter competition, sitting duck interference, occlusion by another transient promoter occupation, collision of elongation complexes, and roadblocks precluding transcription [110]. Although transcriptional interference appears to be less frequent in higher eukaryotes, examples of this process have been reported; for instance, the results of an analysis focused on mouse and human genes that overlap in antisense orientation were consistent with the transcriptional collision model [111]. Transcriptional interference was also analyzed in the context of intronic retroelements and single-exon nested genes [112]. Using a minigene system and different deletion constructs, the *KTI12* retrogene located in the intron of the *TXNDC12* gene was analyzed. This particular case was selected based on the identification of three different ESTs, suggesting forced exonization of a portion of the *TXNDC12* intron upstream of the

retrogene. The experimental results confirmed that expression of the *KTI12* retrocopy imposes utilization of cryptic acceptor splice sites and premature termination of *TXNDC12* gene transcription.

Host genes may also be affected by intronic retrogene methylation. Epigenetic regulation of expression is one of the mechanisms proposed for retrotransposon activity suppression; thus, a link between retrotransposition and methylation is strongly suggested. Recent analysis of retrocopy-associated CpG islands showed that 68% of them are methylated, which in comparison to the whole human genome is a significant proportion [113]. DNA methylation of cytosine bases at the CpG dinucleotides is a basic modification that occurs during genomic imprinting. Few examples of imprinted retrogenes have been reported. For instance, a systematic screen of known genes in mouse led to the identification of 11 imprinted retrogenes, of which three (*Mcts2*, *Dnajb3*, and *Oxct2a*) were nested in an intron of another gene. Interestingly, imprinting of these retrogenes is conserved in humans [114]. Detailed studies of the transcriptionally active *Mcts2* retrogene, inserted into the fourth intron of the *H13* gene, revealed that the retrogene's promoter is silenced by methylation in the female germline [115]. Surprisingly, the choice of polyA signal by the host gene depends on this epigenetic promoter modification. It was shown that expression of the retrogene from the paternal allele forces utilization of the upstream polyA site, resulting in a truncated transcript of the host gene [115]. Similar observations were reported for the *Nap1 l5* retrogene and *Herc3* host gene [116] and other retrogenes (for review, see: [117]).

#### 4.3. Retrocopy Impact on Parental DNA

The main implication for the coexistence of highly similar sequences in a genome is the possibility of direct exchange of DNA fragments in a homologous recombination event. Retrogenes, like other genomic duplicates, can participate in gene conversion. However, this process was shown to be less frequent for genes localized in distant genomic regions [118]; therefore, it is less likely to occur between retrocopy-parental gene pairs, as they are in most cases localized on different chromosomes [119]. The retrotransposition process is also proposed as a possible mechanism in the RNA-mediated intron loss observed in Eukaryota [120–122]. In this model, precise deletion of the parental gene intron occurs as a result of recombination between genomic DNA and spliced, reverse-transcribed mRNA. Intron loss analysis conducted on 684 groups of orthologous genes from seven eukaryotic species supported the proposed mechanism [123]. Another study suggested that higher reverse transcriptase activity is connected to higher frequencies of intron loss and larger numbers of retrocopies. A correlation between these events was observed in mammals [124].

Retrogenes may have an impact on parental genes by contributing to epigenetic regulation of their expression. Nuclear antisense transcripts can act as a scaffold for chromatin remodeling complexes and therefore guide them to genomic loci on the basis of sequence complementarity [125]. It was also demonstrated that retrogenes may participate in this mechanism. A well-known example is the tumor-suppressor gene *PTEN*, whose expression is regulated at multiple levels by the *PTENpg1* retrogene [126,127]. *PTENpg1* has three different transcripts, of which two are antisense to the parental gene. One isoform, called *PTENpg1 asRNA alpha*, recruits the DNA methyl transferase 3A to the parental promoter [127]. As a result, repression of parental gene expression occurs by the addition of three methyl groups to histone H3 Lys27.

#### 4.4. Retrogene Regulatory Functions on RNA Level

##### 4.4.1. Trans-Natural Antisense Transcripts

The vast majority of transcribed retrogenes do not have conserved open reading frames and therefore have roles other than protein-coding functionality. Retrogenes may be expressed independently from both DNA strands, which increases the range of possible interactions on the RNA level. For instance, bioinformatic analysis of ESTs showed that retrogenes constitute 15% of the 87 pseudogenes that were found to be expressed from an antisense strand [128]. Moreover, recent studies focused on long non-coding RNAs that overlap retrocopies across the human genome identified three retrocopy-derived antisense RNAs (asRNAs). These retrocopies are potentially

capable of forming RNA:RNA duplexes with their parental genes [129]. One of these long non-coding RNAs (lncRNAs) derived from the *HNRNPA1P7* retrogene may contribute to heterogeneous nuclear ribonucleoprotein A1 (*hnRNPA1*) pre-mRNA processing. The bioinformatic analysis strongly suggested that asRNA might mask the 5' splice site in the sixth intron of parental gene transcript and therefore enable the expression of isoform with a longer sixth exon. Another identified lncRNA, antisense to a ribosomal protein L23a (*RPL23A*) retrocopy, is potentially able to mask microRNA target sites in seven splice forms of the parental gene, thus controlling their stability [129]. An interesting example was found in the snail *Lymnaea stagnalis*. A retrocopy of a gene encoding nitric oxide synthase (NOS) includes a region of antisense homology to its progenitor's transcript. The antisense region of the pseudogene transcript forms an RNA:RNA duplex with the NOS-encoding mRNA and prevents its translation [130]. Antisense and sense transcripts can also interact. *PTENpg1*, described earlier in the context of epigenetic modification, has a second antisense transcript, *PTENpg1 asRNA beta*, which stabilizes the expression of the sense transcript [127].

#### 4.4.2. MicroRNA Sponges

Continuing with the example of *PTENpg1*, another gene regulation mechanism should be mentioned. The sense transcript of *PTENpg1* shares many microRNA binding sites with the parental gene and consequently exhibits “sponging” activity, manifested by releasing miRNA-mediated repression of *PTEN* [126]. *PTEN* was identified as a tumor-suppressor gene that negatively regulates the phosphatidylinositol 3-kinase signaling pathway involved in cell proliferation control [131]. Downregulation of *PTEN*, which occurs when the retrogene transcript is absent, enhances proliferation and cell growth and decreases sensitivity to cell death, which promotes tumorigenesis [126,131]. The *PTEN-PTENpg1* endogenous competition for shared microRNA molecules has an oncosuppressive effect on different human cancers, including prostate and colon cancers [126], melanoma [132], renal cell carcinoma [133], and hepatocellular carcinoma [134]. Since the identification of *PTENpg1* as a microRNA sponge and functional antisense RNA and its role in tumorigenesis, many retrogene-derived non-coding RNAs have been analyzed in this context [135–138].

One interesting example is the high mobility group A1 (*HMG A1*) gene and its two retrogenes, *HMG A1P6* and *HMG A1P7* [139]. The parental gene encodes proteins involved in chromatin architecture organization and therefore gene expression regulation. Whereas in adult normal cells *HMG A1* proteins are expressed at a very low level, they are overexpressed in cancers. Analysis of the retrogene–parental gene expression pattern in thyroid and ovarian carcinomas showed a positive correlation and suggested gene co-regulation [139]. Both investigated retrogenes conserved the miRNA target site of the parental gene. The abilities of miRNA–retrogene interactions were experimentally evaluated by transfection of miRNAs into human breast adenocarcinoma cells. A significant reduction in the *HMG A1*, *HMG A1P6*, and *HMG A1P7* mRNA levels was observed, which strongly supports the retrogene miRNA “sponging” activity. These retrogenes may also regulate the expression of other genes, including cancer-related ones, such as *HMG A2*, *VEGF*, and *EZH2* [139]. Another research group found that the *HMG A1P7* retrogene is also involved in the regulation of *H19* non-coding gene and *IGF2* gene expression in human breast cancer [140]. In contrast to the oncosuppressive function of *PTENpg1*, *HMG A1P6* and *HMG A1P7* show oncogenic activity and contribute to cancer progression.

Retrogenes may also compete with parental genes for other molecules. For instance, the *Cx43* retrogene transcript shows higher affinity to translational machinery than the parental gene mRNA and therefore decreases the parental protein level [97].

#### 4.4.3. Small RNA

Retrogenes have also been proposed as a source of several classes of small RNAs. For example, they may provide the sequence for novel miRNA genes. Primate-specific *miR-492* may be expressed from both *KRT19* gene and retrogene loci [141,142]. Interestingly, *hs-miR-492* was proposed to play a role in the progression of hepatoblastoma [142] and was found to be a proto-oncogenic miRNA,

acting as a cell proliferation promoter in breast cancer [143]. Several other miRNAs stored in miRBase [144], the main source of miRNA annotations, lie within retrogenes. For instance, *hsa-miR-622*, which acts as a suppressor of tumorigenesis in cancers, including hepatocellular carcinoma [145], overlaps with the genomic locus of the *KRT18P27* retrocopy. Other examples include *hsa-miR-7161*, located in the *TATDN2P2* retrocopy, and *hsa-miR-4788*, overlapping *HMGB3P13*. However, experimental analysis must be conducted to verify the actual miRNA coding potential.

Piwi-interacting RNA (piRNA) involved in the formation of silencing complexes in animal germ lines can also be derived from retrocopies. Total small RNA profiling of the marmoset testis showed abundant expression of piRNAs [146]. Across various clusters, four antisense-oriented piRNAs localized within retrocopies were found, and the authors suggested that the piRNAs originating from these clusters might regulate the parental gene expression by cleaving mRNAs. Moreover, retrogene-derived piRNAs appear to be species-specific because clusters found in the marmoset were absent in the mouse [146]. Recent bioinformatic analysis focused on the small RNA of human sperm showed that piRNA clusters also contain non-coding genes, one of which overlaps with the *NPAP1P6* retrocopy [147]. Interestingly, recent evolutionary analysis showed that the parental gene of the *NPAP1P6* retrocopy, *NPAP1*, was created via duplication of a retrotransposed ancestral paralog derived from the vertebrate nucleoporin gene *POM121* [148]. *NPAP1* is a primate-specific imprinted gene that encodes a nuclear pore-associated protein associated with Prader–Willi syndrome. As the authors emphasized, this syndrome is linked with testis dysfunction, which supports the possible relation between many sperm piRNAs and the analyzed retrocopy [147]. Several additional piRNAs from retrocopies, including *IMPDH1P5*, *TMX2P1*, and *RP11-545A16.3*, which may potentially interact with the protein-coding genes due to high sequence complementarity, were identified [147]. Six retrogene-derived piRNAs that potentially regulate parental genes were found in late mouse spermatocytes [149]. For one of them, experimental evidence of post-transcriptional regulation of *Stambp* gene was provided. By generating two mouse strains with gene-trap insertions upstream of the retrocopy, the authors demonstrated relationships between the piRNA precursor, piRNA, and the parental gene levels, which clearly suggests a role of the retrocopy in *Stambp* regulation [149]. PiRNA clusters overlapping retrocopies were also found in other recent studies (e.g., [150,151]).

Antisense transcripts from retrocopies can pair with other transcripts, including those of the parental gene. This double-stranded RNA may be processed into endogenous small interfering RNA (endo-siRNA). Analysis of the small RNA profiles from wild-type and seven RNA-silencing mutants of *A. thaliana* showed overrepresentation of siRNAs in the transposons, retroelements, and pseudogenes, which may suggest that these sequences are regulated by siRNA-generating systems [152]. A year later, two scientific reports published in parallel showed that endo-siRNAs annotated in the retrocopy locus may regulate their parental genes' expression in mouse oocytes [153,154]. For instance, 77 small RNAs were mapped to an expressed retrocopy (*Gm15681*) of the protein phosphatase 4 regulatory subunit 1 gene (*Ppp4r1*). Moreover, almost all were oriented antisense to the parental gene, which strongly suggested that the siRNAs originated from the retrocopy-parental gene double stranded RNA (dsRNA) region [154]. In another analysis focused on developing rice grains [155], among 145 pseudogenes identified as good candidates for generating antisense small RNAs, 16.6% were retrocopies. However, their *cis*-activity was questioned due to a low rate of identified gene–retrogene complementary regions from which siRNA can be produced [155]. A cluster of siRNAs derived from pseudogenes was also identified in African *Typanosoma brucei* [156]. More detailed studies were conducted for endo-siRNAs derived from human pseudogenes in hepatocellular carcinoma. A well-documented example is the human endo-siRNA from the retrogene of the protein phosphatase 1K (*PPM1K*) gene. The retrocopy can fold into a hairpin structure due to inverted repeats and can be processed in at least two endo-siRNAs, one of which downregulates the parental gene and NIMA-related Kinase 8 (*NEK8*) gene, inhibiting cell proliferation in hepatocellular carcinoma. Thus, the retrogene could be considered a tumor-suppressor gene [157].

Retrocopy-derived small RNAs were also found in the human transcriptome during characterization of the non-coding RNA repertoire. A higher density of small RNA was observed in

retrocopies than in duplicated pseudogenes or coding genes. Interestingly, transcription-dependent H3K9me3 enrichment was observed in some cases, suggesting that pseudogene-derived small RNAs, including retrocopy-derived RNAs, may play a role in modulating the epigenetic suppression of those pseudogenes, as well as neighboring gene expression [158].

## 5. Retrogenes in Diseases

Many retrocopies of protein-coding genes are linked to human diseases, especially various types of cancer (for review, see: [159–162]). It was recently suggested that expression of evolutionarily young non-coding genes in tumors might be considered a new biological phenomenon [163]. A good example of a cancer-related retrogene is the *RHOB* gene, a tumor suppressor of the Rho GTPases family, which arose via retroposition in the early stage of vertebrate evolution [164]. Another retrogene, *UTP14c*, was linked to ovarian cancer predisposition [165]. An analysis of 293 samples representing 13 cancer and normal tissue types revealed 218 pseudogenes expressed only in cancer samples. Out of them, 178 were observed in multiple cancers and 40 were identified in a single cancer type only [166].

As many reports have shown, retrocopies can be used as diagnostic biomarkers, such as the *INTS6P1* retrogene, for which a low expression level in plasma is linked with hepatocellular carcinoma [167], or as prognostic markers, such as tumor-suppressive *PTENpg1* [126,132–134] and the oncogenic *HMGA1P6* and *HMGA1P7* [139,140] retrogenes mentioned in previous sections. Another retrocopy of the *HMGA1* gene is also linked with disease; its overexpression has been found in human type 2 diabetes. Further analysis indicated that this retrocopy post-transcriptionally regulates parental gene expression by competing for critical RNA stability factor and, as a result, suppresses the expression of the insulin receptor gene. Therefore, this retrocopy contributes to insulin resistance [168].

Due to the reactivation of retrotransposons in somatic cells during cancer development, the formation of new retrocopies was observed [169,170]. Somatic acquired retrocopies are present in lung and colorectal cancers. Moreover, insertions occur not only in intragenic regions but also in other genes, which may have implications for their expression. For instance, a *KRT6A* retrocopy replaced the 3' UTR of the *MLL* gene transcript and a *PTPN12* retrocopy caused deletion of the promoter and first exon of the *MGA* gene [170].

Retrogenes may also play active roles in the regulation of signaling pathways involved in inflammation [171]. A mouse retrocopy of ribosomal protein S15A gene, called *Lethe* (*Rps15a-ps4*), is induced by inflammatory cytokines TNF $\alpha$  and IL-1 $\beta$ . Moreover, it can bind and block RelA homodimers, which are required for NF- $\kappa$ B activation; therefore, *Lethe* plays the role of a negative inflammatory response regulator. Interestingly, the retrogene is expressed in an age-dependent manner [171]. Mutation in another retrogene, *TACSTD2* (tumor-associated calcium signal transducer 2) causes gelatinous drop-like corneal dystrophy, leading to blindness [172]. An insertion of a retrocopy, similar to insertion of *L1* or *Alu* elements, may disrupt a gene structure. An example of such event is insertion of a retrocopy of *TMF1* gene into the *CYBB* gene on the X chromosome. This insertion induced aberrant *CYBB* mRNA splicing and introduced a premature stop codon that resulted in chronic granulomatous disease [173].

Retrocopies were also incorporated into analyses conducted in the context of neurodegenerative disorders, including Alzheimer's, Huntington's, and Parkinson's diseases [174], as well as muscular dystrophy [175].

As described above, the chimeric transcript resulting from cyclophilin A (*PPIA*) retrogene insertion into the *TRIM5* host gene in the owl monkey is involved in HIV-1 resistance [107]. Retrocopies may also be associated with the host response during pathogen infection in humans, as significant expression changes were observed after HIV-1 and human type 2 adenovirus infection [176,177]. For instance, expression pattern analysis after HIV-1 infection of human T-cells showed that the most upregulated pseudogene group was a group of retrocopies. Additionally, retrocopies accounted for eight out of 13 cases of underexpressed pseudogenes [176]. These results

suggest that both tandemly duplicated and retroposed pseudogenes may be involved in host–pathogen interaction pathways.

## 6. Retroposition and Genetic Variation

Retroposition gives rise to considerable genetic variation between individuals. Recent developments in sequencing technology allow researchers to move beyond the analysis of individual genomes from model organisms to the study of retrocopies within a population. The 1000 Genomes Project [178] could be mentioned as an example of a large-scale sequencing project that enables the exploration of differences in copy-number variation within human populations. Recent studies [66,169,179], focused on the retrocopy repertoire in human populations, uncovered a total of 208 polymorphic retrocopies [180] called retroduplication variations (RDVs). Moreover, in two of them [66,169], RDV polymorphisms were used as genomic markers for the reconstruction of human population history. In another study, concentrated on retrocopies deletions, 214 indels that affected 190 retrocopies were identified. Out of them, 68 were found to be ancestral (i.e., their orthologs were found in at least one another Hominidae species) and the polymorphism of these retrocopies clearly resulted from a deletion. This study also showed a variation in the retrocopies' expression level [83].

## 7. Conclusions

Retrocopies were long considered non-functional pseudogenes or even “junk DNA”; however, current studies show that they contribute significantly to molecular evolution. Retrocopies have been found as factors shaping differences between species, individuals, or even tissues and cell types; therefore, they are considered a source of genetic polymorphism. They are also important players in complex cellular pathways, including immune response and tumorigenesis. Moreover, a large and increasing range of putative retrocopy functions make them an interesting subject of molecular and medical studies. The numerous studies performed to date have enriched our comprehension of the course and dynamics of retrocopy–gene interactions at the DNA, RNA, and protein levels. Nevertheless, many questions remain unsolved, and further analyses are necessary to accurately describe the plant and animal retrocopy repertoire, evolution, and functions.

**Acknowledgments:** This work was supported by the National Science Centre [grant No. 2013/11/B/NZ2/02598] and the KNOW Poznan RNA Centre [grant No. 01/KNOW2/2014].

**Author Contributions:** M.R.K. drafted the manuscript. Both authors edited, corrected, and approved the manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Makalowski, W.; Kischka, T.; Makałowska, I. Contribution of transposable elements to human proteins. In *Encyclopedia of Life Sciences*; John Wiley & Sons, Ltd.: Chichester, UK, 2017.
2. Agrawal, A.; Eastman, Q.M.; Schatz, D.G. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* **1998**, *394*, 744–751.
3. Kapitonov, V.V.; Jurka, J. Harbinger transposons and an ancient HARBI1 gene derived from a transposase. *DNA Cell Biol.* **2004**, *23*, 311–324.
4. Campillos, M.; Doerks, T.; Shah, P.K.; Bork, P. Computational characterization of multiple Gag-like human proteins. *Trends Genet.* **2006**, *22*, 585–589.
5. Gotea, V.; Makałowski, W. Do transposable elements really contribute to proteomes? *Trends Genet.* **2006**, *22*, 260–267.
6. Mi, S.; Lee, X.; Li, X.; Veldman, G.M.; Finnerty, H.; Racie, L.; LaVallie, E.; Tang, X.Y.; Edouard, P.; Howes, S.; et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **2000**, *403*, 785–789.
7. Kaessmann, H.; Vinckenbosch, N.; Long, M. RNA-based gene duplication: Mechanistic and evolutionary insights. *Nat. Rev. Genet.* **2009**, *10*, 19–31.

8. Ciomborowska, J.; Rosikiewicz, W.; Szklarczyk, D.; Makałowski, W.; Makałowska, I. “Orphan” retrogenes in the human genome. *Mol. Biol. Evol.* **2013**, *30*, 384–396.
9. Brosius, J. Retroposons—Seeds of evolution. *Science* **1991**, *251*, 753.
10. Zhang, Y.; Liu, S.; Zhang, X.; Li, W.-B.; Chen, Y.; Huang, X.; Sun, L.; Luo, W.; Netzer, W.J.; Threadgill, R.; et al. A functional mouse retroposed gene *Rps23r1* reduces Alzheimer’s beta-amyloid levels and Tau phosphorylation. *Neuron* **2009**, *64*, 328–340.
11. Parker, H.G.; VonHoldt, B.M.; Quignon, P.; Margulies, E.H.; Shao, S.; Mosher, D.S.; Spady, T.C.; Elkahlon, A.; Cargill, M.; Jones, P.G.; et al. An expressed *Fgf4* retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* **2009**, *325*, 995–998.
12. Temin, H.M. The nature of the provirus of Rous sarcoma. *Natl. Cancer Inst. Monogr.* **1964**, *17*, 557–570.
13. Baltimore, D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* **1970**, *226*, 1209–1211.
14. Temin, H.M.; Mizutani, S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* **1970**, *226*, 1211–1213.
15. Garfinkel, D.J.; Boeke, J.D.; Fink, G.R. Ty element transposition: Reverse transcriptase and virus-like particles. *Cell* **1985**, *42*, 507–517.
16. Mathias, S.L.; Scott, A.F.; Kazazian, H.H.; Boeke, J.D.; Gabriel, A. Reverse transcriptase encoded by a human transposable element. *Science* **1991**, *254*, 1808–1810.
17. Xiong, Y.; Eickbush, T.H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **1990**, *9*, 3353–3362.
18. Malik, H.S.; Eickbush, T.H. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res.* **2001**, *11*, 1187–1197.
19. Boeke, J.D. The unusual phylogenetic distribution of retrotransposons: A hypothesis. *Genome Res.* **2003**, *13*, 1975–1983.
20. Havecker, E.R.; Gao, X.; Voytas, D.F. The diversity of LTR retrotransposons. *Genome Biol.* **2004**, *5*, 225.
21. Eickbush, T.H.; Jamburuthugoda, V.K. The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res.* **2008**, *134*, 221–234.
22. SanMiguel, P.; Gaut, B.S.; Tikhonov, A.; Nakajima, Y.; Bennetzen, J.L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **1998**, *20*, 43–45.
23. Meyers, B.C.; Tingey, S.V.; Morgante, M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* **2001**, *11*, 1660–1676.
24. Messing, J.; Bharti, A.K.; Karlowski, W.M.; Gundlach, H.; Kim, H.R.; Yu, Y.; Wei, F.; Fuks, G.; Soderlund, C.A.; Mayer, K.F.X.; et al. Sequence composition and genome organization of maize. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 14349–14354.
25. International Barley Genome Sequencing Consortium; Mayer, K.F.X.; Waugh, R.; Brown, J.W.S.; Schulman, A.; Langridge, P.; Platzer, M.; Fincher, G.B.; Muehlbauer, G.J.; Sato, K.; et al. A physical, genetic and functional sequence assembly of the barley genome. *Nature* **2012**, *491*, 711–716.
26. Vitte, C.; Estep, M.C.; Leebens-Mack, J.; Bennetzen, J.L. Young, intact and nested retrotransposons are abundant in the onion and asparagus genomes. *Ann. Bot.* **2013**, *112*, 881–889.
27. Heitkam, T.; Holtgräwe, D.; Dohm, J.C.; Minoche, A.E.; Himmelbauer, H.; Weisshaar, B.; Schmidt, T. Profiling of extensively diversified plant LINES reveals distinct plant-specific subclades. *Plant J.* **2014**, *79*, 385–397.
28. RepeatMasker Home Page. Available online: <http://www.repeatmasker.org/> (accessed on 16 January 2017).
29. Konkel, M.K.; Batzer, M.A. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Semin. Cancer Biol.* **2010**, *20*, 211–221.
30. Van Arsdell, S.W.; Denison, R.A.; Bernstein, L.B.; Weiner, A.M.; Manser, T.; Gesteland, R.F. Direct repeats flank three small nuclear RNA pseudogenes in the human genome. *Cell* **1981**, *26*, 11–17.
31. Maestre, J.; Tchénio, T.; Dhellin, O.; Heidmann, T. mRNA retroposition in human cells: Processed pseudogene formation. *EMBO J.* **1995**, *14*, 6333–6338.
32. Dhellin, O.; Maestre, J.; Heidmann, T. Functional differences between the human LINE retrotransposon and retroviral reverse transcriptases for in vivo mRNA reverse transcription. *EMBO J.* **1997**, *16*, 6590–6602.
33. Esnault, C.; Maestre, J.; Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* **2000**, *24*, 363–367.

34. Wei, W.; Gilbert, N.; Ooi, S.L.; Lawler, J.F.; Ostertag, E.M.; Kazazian, H.H.; Boeke, J.D.; Moran, J.V. Human L1 retrotransposition: *Cis* preference versus *trans* complementation. *Mol. Cell. Biol.* **2001**, *21*, 1429–1439.
35. Zhu, Z.; Tan, S.; Zhang, Y.; Zhang, Y.E. LINE-1-like retrotransposons contribute to RNA-based gene duplication in dicots. *Sci. Rep.* **2016**, *6*, 24755.
36. Drouin, G.; Dover, G.A. A plant processed pseudogene. *Nature* **1987**, *328*, 557–558.
37. Benovoy, D.; Drouin, G. Processed pseudogenes, processed genes, and spontaneous mutations in the *Arabidopsis* genome. *J. Mol. Evol.* **2006**, *62*, 511–522.
38. Wang, W.; Zheng, H.; Fan, C.; Li, J.; Shi, J.; Cai, Z.; Zhang, G.; Liu, D.; Zhang, J.; Vang, S.; et al. High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* **2006**, *18*, 1791–1802.
39. Sakai, H.; Mizuno, H.; Kawahara, Y.; Wakimoto, H.; Ikawa, H.; Kawahigashi, H.; Kanamori, H.; Matsumoto, T.; Itoh, T.; Gaut, B.S. Retrogenes in rice (*Oryza sativa* L. ssp. japonica) exhibit correlated expression with their source genes. *Genome Biol. Evol.* **2011**, *3*, 1357–1368.
40. Abdelsamad, A.; Pecinka, A. Pollen-specific activation of *Arabidopsis* retrogenes is associated with global transcriptional reprogramming. *Plant Cell* **2014**, *26*, 3299–3313.
41. Rosikiewicz, W.; Kabza, M.; Kosiński, J.; Ciomborowska, J.; Kubiak, M.R.; Makołowska, I. RetrogeneDB—A database of plant and animal retrocopies. *Database (Oxford)*, under review.
42. Jin, Y.K.; Bennetzen, J.L. Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the Bs1 retroelement of maize. *Plant Cell* **1994**, *6*, 1177–1186.
43. Elrouby, N.; Bureau, T.E. A novel hybrid open reading frame formed by multiple cellular gene transductions by a plant long terminal repeat retroelement. *J. Biol. Chem.* **2001**, *276*, 41963–41968.
44. Tan, S.; Cardoso-Moreira, M.; Shi, W.; Zhang, D.; Huang, J.; Mao, Y.; Jia, H.; Zhang, Y.; Chen, C.; Shao, Y.; et al. LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. *Genome Res.* **2016**, *26*, 1663–1675.
45. Dong, R.; Zhang, X.-O.; Zhang, Y.; Ma, X.-K.; Chen, L.-L.; Yang, L. CircRNA-derived pseudogenes. *Cell Res.* **2016**, *26*, 747–750.
46. Brouha, B.; Schustak, J.; Badge, R.M.; Lutz-Prigge, S.; Farley, A.H.; Moran, J.V.; Kazazian, H.H. Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 5280–5285.
47. Elbarbary, R.A.; Lucas, B.A.; Maquat, L.E. Retrotransposons as regulators of gene expression. *Science* **2016**, *351*, aac7247.
48. Goodier, J.L.; Kazazian, H.H. Retrotransposons revisited: The restraint and rehabilitation of parasites. *Cell* **2008**, *135*, 23–35.
49. Crichton, J.H.; Dunican, D.S.; Maclennan, M.; Meehan, R.R.; Adams, I.R. Defending the genome from the enemy within: Mechanisms of retrotransposon suppression in the mouse germline. *Cell. Mol. Life Sci.* **2014**, *71*, 1581–1605.
50. Pizarro, J.G.; Cristofari, G. Post-Transcriptional Control of LINE-1 Retrotransposition by Cellular Host Factors in Somatic Cells. *Front. Cell Dev. Biol.* **2016**, *4*, 14.
51. Schulman, A.H. Retrotransposon replication in plants. *Curr. Opin. Virol.* **2013**, *3*, 604–614.
52. Beck, C.R.; Collier, P.; Macfarlane, C.; Malig, M.; Kidd, J.M.; Eichler, E.E.; Badge, R.M.; Moran, J.V. LINE-1 retrotransposition activity in human genomes. *Cell* **2010**, *141*, 1159–1170.
53. Ewing, A.D.; Kazazian, H.H. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res.* **2010**, *20*, 1262–1270.
54. Speek, M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol. Cell. Biol.* **2001**, *21*, 1973–1985.
55. Martin, S.L.; Bushman, F.D. Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol. Cell. Biol.* **2001**, *21*, 467–475.
56. Moran, J.V.; Holmes, S.E.; Naas, T.P.; DeBerardinis, R.J.; Boeke, J.D.; Kazazian, H.H. High frequency retrotransposition in cultured mammalian cells. *Cell* **1996**, *87*, 917–927.
57. Denli, A.M.; Narvaiza, I.; Kerman, B.E.; Pena, M.; Benner, C.; Marchetto, M.C.N.; Diedrich, J.K.; Aslanian, A.; Ma, J.; Moresco, J.J.; et al. Primate-specific ORF0 contributes to retrotransposon-mediated diversity. *Cell* **2015**, *163*, 583–593.
58. Khazina, E.; Truffault, V.; Büttner, R.; Schmidt, S.; Coles, M.; Weichenrieder, O. Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. *Nat. Struct. Mol. Biol.* **2011**, *18*, 1006–1014.

59. Dewannieux, M.; Esnault, C.; Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* **2003**, *35*, 41–48.
60. Doucet, A.J.; Droc, G.; Siol, O.; Audoux, J.; Gilbert, N. U6 snRNA Pseudogenes: Markers of Retrotransposition Dynamics in Mammals. *Mol. Biol. Evol.* **2015**, *32*, 1815–1832.
61. Hancks, D.C.; Goodier, J.L.; Mandal, P.K.; Cheung, L.E.; Kazazian, H.H. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum. Mol. Genet.* **2011**, *20*, 3386–3400.
62. Raiz, J.; Damert, A.; Chira, S.; Held, U.; Klawitter, S.; Hamdorf, M.; Löwer, J.; Strätling, W.H.; Löwer, R.; Schumann, G.G. The non-autonomous retrotransposon SVA is *trans*-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res.* **2012**, *40*, 1666–1683.
63. Perreault, J.; Noël, J.-F.; Brière, F.; Cousineau, B.; Lucier, J.-F.; Perreault, J.-P.; Boire, G. Retropseudogenes derived from the human Ro/SS-A autoantigen-associated hY RNAs. *Nucleic Acids Res.* **2005**, *33*, 2032–2041.
64. Pavlíček, A.; Paces, J.; Elleder, D.; Hejnar, J. Processed pseudogenes of human endogenous retroviruses generated by LINES: Their integration, stability, and distribution. *Genome Res.* **2002**, *12*, 391–399.
65. Doucet, A.J.; Wilusz, J.E.; Miyoshi, T.; Liu, Y.; Moran, J.V. A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition. *Mol. Cell* **2015**, *60*, 728–741.
66. Abyzov, A.; Iskow, R.; Gokcumen, O.; Radke, D.W.; Balasubramanian, S.; Pei, B.; Habegger, L.; 1000 Genomes Project Consortium; Lee, C.; Gerstein, M. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome Res.* **2013**, *23*, 2042–2052.
67. Kubo, S.; del Selemé, M.C.; Soifer, H.S.; Perez, J.L.G.; Moran, J.V.; Kazazian, H.H.; Kasahara, N. L1 retrotransposition in nondividing and primary human somatic cells. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8036–8041.
68. Macia, A.; Widmann, T.J.; Heras, S.R.; Ayllon, V.; Sanchez, L.; Benkaddour-Boumzaouad, M.; Muñoz-Lopez, M.; Rubio, A.; Amador-Cubero, S.; Blanco-Jimenez, E.; et al. Engineered LINE-1 retrotransposition in nondividing human neurons. *Genome Res.* **2017**, *27*, 335–348.
69. Cost, G.J.; Feng, Q.; Jacquier, A.; Boeke, J.D. Human L1 element target-primed reverse transcription in vitro. *EMBO J.* **2002**, *21*, 5899–5910.
70. Sen, S.K.; Huang, C.T.; Han, K.; Batzer, M.A. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Res.* **2007**, *35*, 3741–3751.
71. Jąkałski, M.; Takeshita, K.; Deblieck, M.; Koyanagi, K.O.; Mąkałowska, I.; Watanabe, H.; Mąkałowski, W. Comparative genomic analysis of retrogene repertoire in two green algae *Volvox carteri* and *Chlamydomonas reinhardtii*. *Biol. Direct* **2016**, *11*, 35.
72. Zhang, X.; Zhang, Q.; Su, B. Emergence and evolution of inter-specific segregating retrocopies in cynomolgus monkey (*Macaca fascicularis*) and rhesus macaque (*Macaca mulatta*). *Sci. Rep.* **2016**, *6*, 32598.
73. Szcześniak, M.W.; Ciomborowska, J.; Nowak, W.; Rogozin, I.B.; Mąkałowska, I. Primate and rodent specific intron gains and the origin of retrogenes with splice variants. *Mol. Biol. Evol.* **2011**, *28*, 33–37.
74. Baertsch, R.; Diekhans, M.; Kent, W.J.; Haussler, D.; Brosius, J. Retrocopy contributions to the evolution of the human genome. *BMC Genom.* **2008**, *9*, 466.
75. Slesarev, A.I.; Belova, G.I.; Kozyavkin, S.A.; Lake, J.A. Evidence for an early prokaryotic origin of histones H2A and H4 prior to the emergence of eukaryotes. *Nucleic Acids Res.* **1998**, *26*, 427–430.
76. Ashurst, J.L.; Chen, C.-K.; Gilbert, J.G.R.; Jekosch, K.; Keenan, S.; Meidl, P.; Searle, S.M.; Stalker, J.; Storey, R.; Trevanion, S.; et al. The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res.* **2005**, *33*, D459–D465.
77. Aken, B.L.; Ayling, S.; Barrell, D.; Clarke, L.; Curwen, V.; Fairley, S.; Fernandez Banet, J.; Billis, K.; García Girón, C.; Hourlier, T.; et al. The Ensembl gene annotation system. *Database* **2016**, *2016*, baw093.
78. Zhang, Z.; Carriero, N.; Zheng, D.; Karro, J.; Harrison, P.M.; Gerstein, M. PseudoPipe: An automated pseudogene identification pipeline. *Bioinformatics* **2006**, *22*, 1437–1439.
79. Speir, M.L.; Zweig, A.S.; Rosenbloom, K.R.; Raney, B.J.; Paten, B.; Nejad, P.; Lee, B.T.; Learned, K.; Karolchik, D.; Hinrichs, A.S.; et al. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.* **2016**, *44*, D717–D725.
80. Pei, B.; Sisu, C.; Frankish, A.; Howald, C.; Habegger, L.; Mu, X.J.; Harte, R.; Balasubramanian, S.; Tanzer, A.; Diekhans, M.; et al. The GENCODE pseudogene resource. *Genome Biol.* **2012**, *13*, R51.
81. Khelifi, A.; Adel, K.; Duret, L.; Laurent, D.; Mouchiroud, D.; Dominique, M. HOPPSIGEN: A database of human and mouse processed pseudogenes. *Nucleic Acids Res.* **2005**, *33*, D59–D66.

82. Navarro, F.C.P.; Galante, P.A.F. RCPedia: A database of retrocopied genes. *Bioinformatics* **2013**, *29*, 1235–1237.
83. Kabza, M.; Kubiak, M.R.; Danek, A.; Rosikiewicz, W.; Deorowicz, S.; Polański, A.; Makałowska, I. Inter-population Differences in Retrogene Loss and Expression in Humans. *PLoS Genet.* **2015**, *11*, e1005579.
84. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.
85. Kielbasa, S.M.; Wan, R.; Sato, K.; Horton, P.; Frith, M.C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **2011**, *21*, 487–493.
86. Louhichi, A.; Fourati, A.; Rebaï, A. IGD: A resource for intronless genes in the human genome. *Gene* **2011**, *488*, 35–40.
87. Jorquera, R.; Ortiz, R.; Ossandon, F.; Cárdenas, J.P.; Sepúlveda, R.; González, C.; Holmes, D.S. SinEx DB: A database for single exon coding sequences in mammalian genomes. *Database* **2016**, *2016*, baw095.
88. Vinckenbosch, N.; Dupanloup, I.; Kaessmann, H. Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 3220–3225.
89. Okamura, K.; Nakai, K. Retrotransposition as a source of new promoters. *Mol. Biol. Evol.* **2008**, *25*, 1231–1238.
90. Zhang, J. Evolution by gene duplication: An update. *Trends Ecol. Evol.* **2003**, *18*, 292–298.
91. Force, A.; Lynch, M.; Pickett, F.B.; Amores, A.; Yan, Y.L.; Postlethwait, J. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **1999**, *151*, 1531–1545.
92. Liu, Z.; Tavares, R.; Forsythe, E.S.; André, F.; Lugan, R.; Jonasson, G.; Boutet-Mercey, S.; Tohge, T.; Beilstein, M.A.; Werck-Reichhart, D.; et al. Evolutionary interplay between sister cytochrome P450 genes shapes plasticity in plant metabolism. *Nat. Commun.* **2016**, *7*, 13026.
93. Rosso, L.; Marques, A.C.; Weier, M.; Lambert, N.; Lambot, M.-A.; Vanderhaeghen, P.; Kaessmann, H. Birth and Rapid Subcellular Adaptation of a Hominoid-Specific CDC14 Protein. *PLoS Biol.* **2008**, *6*, e140.
94. Rastogi, S.; Liberles, D.A. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol. Biol.* **2005**, *5*, 28.
95. Young, J.; Ménétrey, J.; Goud, B. RAB6C is a retrogene that encodes a centrosomal protein involved in cell cycle progression. *J. Mol. Biol.* **2010**, *397*, 69–88.
96. Kandouz, M.; Bier, A.; Carystinos, G.D.; Alaoui-Jamali, M.A.; Batist, G. Connexin43 pseudogene is expressed in tumor cells and inhibits growth. *Oncogene* **2004**, *23*, 4763–4770.
97. Bier, A.; Oviedo-Landaverde, I.; Zhao, J.; Mamane, Y.; Kandouz, M.; Batist, G. Connexin43 pseudogene in breast cancer cells offers a novel therapeutic target. *Mol. Cancer Ther.* **2009**, *8*, 786–793.
98. Sulak, M.; Fong, L.; Mika, K.; Chigurupati, S.; Yon, L.; Mongan, N.P.; Emes, R.D.; Lynch, V.J. TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. *Elife* **2016**, *5*, e11994.
99. Carelli, F.N.; Hayakawa, T.; Go, Y.; Imai, H.; Warnefors, M.; Kaessmann, H. The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Res.* **2016**, *26*, 301–314.
100. Kim, D.S.; Wang, Y.; Oh, H.J.; Choi, D.; Lee, K.; Hahn, Y. Retroduplication and loss of parental genes is a mechanism for the generation of intronless genes in *Ciona intestinalis* and *Ciona savignyi*. *Dev. Genes Evol.* **2014**, *224*, 255–260.
101. Uechi, T.; Maeda, N.; Tanaka, T.; Kenmochi, N. Functional second genes generated by retrotransposition of the X-linked ribosomal protein genes. *Nucleic Acids Res.* **2002**, *30*, 5369–5375.
102. Elliott, D.J.; Venables, J.P.; Newton, C.S.; Lawson, D.; Boyle, S.; Eperon, I.C.; Cooke, H.J. An evolutionarily conserved germ cell-specific hnRNP is encoded by a retrotransposed gene. *Hum. Mol. Genet.* **2000**, *9*, 2117–2124.
103. Mascarenhas, R.; Pietrzak, M.; Smith, R.M.; Webb, A.; Wang, D.; Papp, A.C.; Pinsonneault, J.K.; Seweryn, M.; Rempala, G.; Sadee, W. Allele-Selective Transcriptome Recruitment to Polysomes Primed for Translation: Protein-Coding and Noncoding RNAs, and RNA Isoforms. *PLoS ONE* **2015**, *10*, e0136798.
104. Brosch, M.; Saunders, G.I.; Frankish, A.; Collins, M.O.; Yu, L.; Wright, J.; Verstraten, R.; Adams, D.J.; Harrow, J.; Choudhary, J.S.; et al. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and “resurrected” pseudogenes in the mouse genome. *Genome Res.* **2011**, *21*, 756–767.
105. Kim, M.-S.; Pinto, S.M.; Getnet, D.; Nirujogi, R.S.; Manda, S.S.; Chaerkady, R.; Madugundu, A.K.; Kelkar, D.S.; Isserlin, R.; Jain, S.; et al. A draft map of the human proteome. *Nature* **2014**, *509*, 575–581.

106. Long, M.; Langley, C.H. Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. *Science* **1993**, *260*, 91–95.
107. Sayah, D.M.; Sokolskaja, E.; Berthou, L.; Luban, J. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* **2004**, *430*, 569–573.
108. Mori, S.; Hayashi, M.; Inagaki, S.; Oshima, T.; Tateishi, K.; Fujii, H.; Suzuki, S. Identification of Multiple Forms of RNA Transcripts Associated with Human-Specific Retrotransposed Gene Copies. *Genome Biol. Evol.* **2016**, *8*, 2288–2296.
109. Jang, S.; Lee, B.; Kim, C.; Kim, S.-J.; Yim, J.; Han, J.-J.; Lee, S.; Kim, S.-R.; An, G. The OsFOR1 gene encodes a polygalacturonase-inhibiting protein (PGIP) that regulates floral organ number in rice. *Plant Mol. Biol.* **2003**, *53*, 357–369.
110. Shearwin, K.E.; Callen, B.P.; Egan, J.B. Transcriptional interference—A crash course. *Trends Genet.* **2005**, *21*, 339–345.
111. Osato, N.; Suzuki, Y.; Ikeo, K.; Gojobori, T. Transcriptional interferences in *cis* natural antisense transcripts of humans and mice. *Genetics* **2007**, *176*, 1299–1306.
112. Kaer, K.; Branovets, J.; Hallikma, A.; Nigumann, P.; Speek, M. Intronic L1 retrotransposons and nested genes cause transcriptional interference by inducing intron retention, exonization and cryptic polyadenylation. *PLoS ONE* **2011**, *6*, e26099.
113. Grothaus, K.; Kanber, D.; Gellhaus, A.; Mikat, B.; Kolarova, J.; Siebert, R.; Wieczorek, D.; Horsthemke, B. Genome-wide methylation analysis of retrocopy-associated CpG islands and their genomic environment. *Epigenetics* **2016**, *11*, 216–226.
114. Wood, A.J.; Roberts, R.G.; Monk, D.; Moore, G.E.; Schulz, R.; Oakey, R.J. A screen for retrotransposed imprinted genes reveals an association between X chromosome homology and maternal germ-line methylation. *PLoS Genet.* **2007**, *3*, e20.
115. Wood, A.J.; Schulz, R.; Woodfine, K.; Koltowska, K.; Beechey, C.V.; Peters, J.; Bourc'his, D.; Oakey, R.J. Regulation of alternative polyadenylation by genomic imprinting. *Genes Dev.* **2008**, *22*, 1141–1146.
116. Cowley, M.; Wood, A.J.; Böhm, S.; Schulz, R.; Oakey, R.J. Epigenetic control of alternative mRNA processing at the imprinted Herc3/Nap115 locus. *Nucleic Acids Res.* **2012**, *40*, 8917–8926.
117. McCole, R.B.; Oakey, R.J. Unwitting hosts fall victim to imprinting. *Epigenetics* **2008**, *3*, 258–260.
118. Ezawa, K.; Oota, S.; Saitou, N. Genome-wide search of gene conversions in duplicated genes of mouse and rat. *Mol. Biol. Evol.* **2006**, *23*, 927–940.
119. Pan, D.; Zhang, L. Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: A novel strategy to estimate gene duplication rates. *Genome Biol.* **2007**, *8*, R158.
120. Nishioka, Y.; Leder, A.; Leder, P. Unusual alpha-globin-like gene that has cleanly lost both globin intervening sequences. *Proc. Natl. Acad. Sci. USA* **1980**, *77*, 2806–2809.
121. Derr, L.K. The involvement of cellular recombination and repair genes in RNA-mediated recombination in *Saccharomyces cerevisiae*. *Genetics* **1998**, *148*, 937–945.
122. Mourier, T.; Jeffares, D.C. Eukaryotic intron loss. *Science* **2003**, *300*, 1393.
123. Roy, S.W.; Gilbert, W. The pattern of intron loss. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 713–718.
124. Zhu, T.; Niu, D.-K. Frequency of intron loss correlates with processed pseudogene abundance: A novel strategy to test the reverse transcriptase model of intron loss. *BMC Biol.* **2013**, *11*, 23.
125. Rinn, J.L.; Kertesz, M.; Wang, J.K.; Squazzo, S.L.; Xu, X.; Bruggmann, S.A.; Goodnough, L.H.; Helms, J.A.; Farnham, P.J.; Segal, E.; et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **2007**, *129*, 1311–1323.
126. Poliseno, L.; Salmena, L.; Zhang, J.; Carver, B.; Haveman, W.J.; Pandolfi, P.P. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **2010**, *465*, 1033–1038.
127. Johnsson, P.; Ackley, A.; Vidarsdottir, L.; Lui, W.-O.; Corcoran, M.; Grandér, D.; Morris, K.V. A pseudogene long noncoding RNA network regulates PTEN transcription and translation in human cells. *Nat. Struct. Mol. Biol.* **2013**, *20*, 440–446.
128. Muro, E.M.; Andrade-Navarro, M.A. Pseudogenes as an alternative source of natural antisense transcripts. *BMC Evol. Biol.* **2010**, *10*, 338.
129. Bryzghalov, O.; Szcześniak, M.W.; Makołowska, I. Retroposition as a source of antisense long non-coding RNAs with possible regulatory functions. *Acta Biochim. Pol.* **2016**, *63*, 825–833.
130. Korneev, S.A.; Park, J.H.; O'Shea, M. Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J. Neurosci.* **1999**, *19*, 7711–7720.

131. Stambolic, V.; Suzuki, A.; de la Pompa, J.L.; Brothers, G.M.; Mirtsos, C.; Sasaki, T.; Ruland, J.; Penninger, J.M.; Siderovski, D.P.; Mak, T.W. Negative regulation of PKB/Akt-dependent cell survival by the tumor suppressor PTEN. *Cell* **1998**, *95*, 29–39.
132. Polisenio, L.; Haimovic, A.; Christos, P.J.; Vega, Y.; Saenz de Miera, E.C.; Shapiro, R.; Pavlick, A.; Berman, R.S.; Darvishian, F.; Osman, I. Deletion of PTENP1 pseudogene in human melanoma. *J. Investig. Dermatol.* **2011**, *131*, 2497–2500.
133. Yu, G.; Yao, W.; Gumireddy, K.; Li, A.; Wang, J.; Xiao, W.; Chen, K.; Xiao, H.; Li, H.; Tang, K.; et al. Pseudogene PTENP1 functions as a competing endogenous RNA to suppress clear-cell renal cell carcinoma progression. *Mol. Cancer Ther.* **2014**, *13*, 3086–3097.
134. Chen, C.-L.; Tseng, Y.-W.; Wu, J.-C.; Chen, G.-Y.; Lin, K.-C.; Hwang, S.-M.; Hu, Y.-C. Suppression of hepatocellular carcinoma by baculovirus-mediated expression of long non-coding RNA PTENP1 and microRNA regulation. *Biomaterials* **2015**, *44*, 71–81.
135. Wang, L.; Guo, Z.-Y.; Zhang, R.; Xin, B.; Chen, R.; Zhao, J.; Wang, T.; Wen, W.-H.; Jia, L.-T.; Yao, L.-B.; et al. Pseudogene OCT4-pg4 functions as a natural micro RNA sponge to regulate OCT4 expression by competing for miR-145 in hepatocellular carcinoma. *Carcinogenesis* **2013**, *34*, 1773–1781.
136. Peng, H.; Ishida, M.; Li, L.; Saito, A.; Kamiya, A.; Hamilton, J.P.; Fu, R.; Olaru, A.V.; An, F.; Popescu, I.; et al. Pseudogene INTS6P1 regulates its cognate gene INTS6 through competitive binding of miR-17-5p in hepatocellular carcinoma. *Oncotarget* **2015**, *6*, 5666–5677.
137. Ye, X.; Fan, F.; Bhattacharya, R.; Bellister, S.; Boulbes, D.R.; Wang, R.; Xia, L.; Ivan, C.; Zheng, X.; Calin, G.A.; et al. VEGFR-1 Pseudogene Expression and Regulatory Function in Human Colorectal Cancer Cells. *Mol. Cancer Res.* **2015**, *13*, 1274–1282.
138. Di Sanzo, M.; Aversa, I.; Santamaria, G.; Gagliardi, M.; Panebianco, M.; Biamonte, F.; Zolea, F.; Faniello, M.C.; Cuda, G.; Costanzo, F. FTH1P3, a Novel H-Ferritin Pseudogene Transcriptionally Active, Is Ubiquitously Expressed and Regulated during Cell Differentiation. *PLoS ONE* **2016**, *11*, e0151359.
139. Esposito, F.; De Martino, M.; Petti, M.G.; Forzati, F.; Tornincasa, M.; Federico, A.; Arra, C.; Pierantoni, G.M.; Fusco, A. HMGA1 pseudogenes as candidate proto-oncogenic competitive endogenous RNAs. *Oncotarget* **2014**, *5*, 8341–8354.
140. De Martino, M.; Forzati, F.; Marfella, M.; Pellecchia, S.; Arra, C.; Terracciano, L.; Fusco, A.; Esposito, F. HMGA1P7-pseudogene regulates H19 and Igf2 expression by a competitive endogenous RNA mechanism. *Sci. Rep.* **2016**, *6*, 37622.
141. Devor, E.J. Primate microRNAs miR-220 and miR-492 lie within processed pseudogenes. *J. Hered.* **2006**, *97*, 186–190.
142. Von Frowein, J.; Pagel, P.; Kappler, R.; von Schweinitz, D.; Roscher, A.; Schmid, I. MicroRNA-492 is processed from the keratin 19 gene and up-regulated in metastatic hepatoblastoma. *Hepatology* **2011**, *53*, 833–842.
143. Shen, F.; Cai, W.-S.; Feng, Z.; Li, J.-L.; Chen, J.-W.; Cao, J.; Xu, B. MiR-492 contributes to cell proliferation and cell cycle of human breast cancer cells by suppressing SOX7 expression. *Tumour Biol.* **2015**, *36*, 1913–1921.
144. Griffiths-Jones, S.; Grocock, R.J.; van Dongen, S.; Bateman, A.; Enright, A.J. miRBase: MicroRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **2006**, *34*, D140–D144.
145. Song, W.-H.; Feng, X.-J.; Gong, S.-J.; Chen, J.-M.; Wang, S.-M.; Xing, D.-J.; Zhu, M.-H.; Zhang, S.-H.; Xu, A.-M. MicroRNA-622 acts as a tumor suppressor in hepatocellular carcinoma. *Cancer Biol. Ther.* **2015**, *16*, 1754–1763.
146. Hirano, T.; Iwasaki, Y.W.; Lin, Z.Y.-C.; Imamura, M.; Seki, N.M.; Sasaki, E.; Saito, K.; Okano, H.; Siomi, M.C.; Siomi, H. Small RNA profiling and characterization of piRNA clusters in the adult testes of the common marmoset, a model primate. *RNA* **2014**, *20*, 1223–1237.
147. Pantano, L.; Jodar, M.; Bak, M.; Ballescà, J.L.; Tommerup, N.; Oliva, R.; Vavouri, T. The small RNA content of human sperm reveals pseudogene-derived piRNAs complementary to protein-coding genes. *RNA* **2015**, *21*, 1085–1095.
148. Neumann, L.C.; Feiner, N.; Meyer, A.; Buiting, K.; Horsthemke, B. The imprinted *NPAP1* gene in the Prader-Willi syndrome region belongs to a *POM121*-related family of retrogenes. *Genome Biol. Evol.* **2014**, *6*, 344–351.
149. Watanabe, T.; Cheng, E.; Zhong, M.; Lin, H. Retrotransposons and pseudogenes regulate mRNAs and lncRNAs via the piRNA pathway in the germline. *Genome Res.* **2015**, *25*, 368–380.

150. Gebert, D.; Ketting, R.F.; Zischler, H.; Rosenkranz, D. piRNAs from Pig Testis Provide Evidence for a Conserved Role of the Piwi Pathway in Post-Transcriptional Gene Regulation in Mammals. *PLoS ONE* **2015**, *10*, e0124860.
151. Milligan, M.J.; Harvey, E.; Yu, A.; Morgan, A.L.; Smith, D.L.; Zhang, E.; Berengut, J.; Sivananthan, J.; Subramaniam, R.; Skoric, A.; et al. Global Intersection of Long Non-Coding RNAs with Processed and Unprocessed Pseudogenes in the Human Genome. *Front. Genet.* **2016**, *7*, 26.
152. Kasschau, K.D.; Fahlgren, N.; Chapman, E.J.; Sullivan, C.M.; Cumbie, J.S.; Givan, S.A.; Carrington, J.C. Genome-wide profiling and analysis of *Arabidopsis* siRNAs. *PLoS Biol.* **2007**, *5*, e57.
153. Tam, O.H.; Aravin, A.A.; Stein, P.; Girard, A.; Murchison, E.P.; Cheloufi, S.; Hodges, E.; Anger, M.; Sachidanandam, R.; Schultz, R.M.; et al. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **2008**, *453*, 534–538.
154. Watanabe, T.; Totoki, Y.; Toyoda, A.; Kaneda, M.; Kuramochi-Miyagawa, S.; Obata, Y.; Chiba, H.; Kohara, Y.; Kono, T.; Nakano, T.; et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **2008**, *453*, 539–543.
155. Guo, X.; Zhang, Z.; Gerstein, M.B.; Zheng, D. Small RNAs originated from pseudogenes: *cis*- or *trans*-acting? *PLoS Comput. Biol.* **2009**, *5*, e1000449.
156. Wen, Y.-Z.; Zheng, L.-L.; Liao, J.-Y.; Wang, M.-H.; Wei, Y.; Guo, X.-M.; Qu, L.-H.; Ayala, F.J.; Lun, Z.-R. Pseudogene-derived small interference RNAs regulate gene expression in African *Trypanosoma brucei*. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 8345–8350.
157. Chan, W.-L.; Yuo, C.-Y.; Yang, W.-K.; Hung, S.-Y.; Chang, Y.-S.; Chiu, C.-C.; Yeh, K.-T.; Huang, H.-D.; Chang, J.-G. Transcribed pseudogene  $\psi$ PPM1K generates endogenous siRNA to suppress oncogenic cell growth in hepatocellular carcinoma. *Nucleic Acids Res.* **2013**, *41*, 3734–3747.
158. Guo, X.; Lin, M.; Rockowitz, S.; Lachman, H.M.; Zheng, D. Characterization of Human Pseudogene-Derived Non-Coding RNAs for Functional Potential. *PLoS ONE* **2014**, *9*, e93972.
159. Fatima, R.; Akhade, V.S.; Pal, D.; Rao, S.M. Long noncoding RNAs in development and cancer: Potential biomarkers and therapeutic targets. *Mol. Cell. Ther.* **2015**, *3*, doi:10.1186/s40591-015-0042-6.
160. Poliseno, L.; Marranci, A.; Pandolfi, P.P. Pseudogenes in Human Cancer. *Front. Med.* **2015**, *2*, 68.
161. Zhang, X.; Zhang, J.; Ping, X.; Wang, Q.-L.; Lu, X. Pseudogene transcripts: Participants in tumorigenicity and promising therapeutic targets. *Leuk. Res.* **2016**, *42*, 105–106.
162. Shi, X.; Nie, F.; Wang, Z.; Sun, M. Pseudogene-expressed RNAs: A new frontier in cancers. *Tumour Biol.* **2016**, *37*, 1471–1478.
163. Kozlov, A.P. Expression of evolutionarily novel genes in tumors. *Infect. Agents Cancer* **2016**, *11*, doi:10.1186/s13027-016-0077-6.
164. Prendergast, G.C. Actin' up: RhoB in cancer and apoptosis. *Nat. Rev. Cancer* **2001**, *1*, 162–168.
165. Rohozinski, J.; Edwards, C.L.; Anderson, M.L. Does expression of the retrogene UTP14c in the ovary predispose women to ovarian cancer? *Med. Hypotheses* **2012**, *78*, 446–449.
166. Kalyana-Sundaram, S.; Kumar-Sinha, C.; Shankar, S.; Robinson, D.R.; Wu, Y.-M.; Cao, X.; Asangani, I.A.; Kothari, V.; Prensner, J.R.; Lonigro, R.J.; et al. Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell* **2012**, *149*, 1622–1634.
167. Lui, K.Y.; Peng, H.-R.; Lin, J.-R.; Qiu, C.-H.; Chen, H.-A.; Fu, R.-D.; Cai, C.-J.; Lu, M.-Q. Pseudogene integrator complex subunit 6 pseudogene 1 (INTS6P1) as a novel plasma-based biomarker for hepatocellular carcinoma screening. *Tumour Biol.* **2016**, *37*, 1253–1260.
168. Chiefari, E.; Iiritano, S.; Paonessa, F.; Le Pera, I.; Arcidiacono, B.; Filocamo, M.; Foti, D.; Liebhaber, S.A.; Brunetti, A. Pseudogene-mediated posttranscriptional silencing of HMGA1 can result in insulin resistance and type 2 diabetes. *Nat. Commun.* **2010**, *1*, 40.
169. Ewing, A.D.; Ballinger, T.J.; Earl, D.; Broad Institute Genome Sequencing and Analysis Program and Platform; Harris, C.C.; Ding, L.; Wilson, R.K.; Haussler, D. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol.* **2013**, *14*, R22.
170. Cooke, S.L.; Shlien, A.; Marshall, J.; Pipinikas, C.P.; Martincorena, I.; Tubio, J.M.C.; Li, Y.; Menzies, A.; Mudie, L.; Ramakrishna, M.; et al. Processed pseudogenes acquired somatically during cancer development. *Nat. Commun.* **2014**, *5*, 3644.
171. Rapicavoli, N.A.; Qu, K.; Zhang, J.; Mikhail, M.; Laberge, R.-M.; Chang, H.Y. A mammalian pseudogene lncRNA at the interface of inflammation and anti-inflammatory therapeutics. *eLife* **2013**, *2*, e00762.

172. Tsujikawa, M.; Kurahashi, H.; Tanaka, T.; Nishida, K.; Shimomura, Y.; Tano, Y.; Nakamura, Y. Identification of the gene responsible for gelatinous drop-like corneal dystrophy. *Nat. Genet.* **1999**, *21*, 420–423.
173. De Boer, M.; van Leeuwen, K.; Geissler, J.; Weemaes, C.M.; van den Berg, T.K.; Kuijpers, T.W.; Warris, A.; Roos, D. Primary immunodeficiency caused by an exonized retroposed gene copy inserted in the CYBB gene. *Hum. Mutat.* **2014**, *35*, 486–496.
174. Costa, V.; Esposito, R.; Aprile, M.; Ciccodicola, A. Non-coding RNA and pseudogenes in neurodegenerative diseases: “The (un)Usual Suspects”. *Front. Genet.* **2012**, *3*, 231.
175. Feng, Q.; Snider, L.; Jagannathan, S.; Tawil, R.; van der Maarel, S.M.; Tapscott, S.J.; Bradley, R.K. A feedback loop between nonsense-mediated decay and the retrogene DUX4 in facioscapulohumeral muscular dystrophy. *eLife* **2015**, *4*, doi:10.7554/eLife.04996.
176. Gupta, A.; Brown, C.T.; Zheng, Y.-H.; Adami, C. Differentially-Expressed Pseudogenes in HIV-1 Infection. *Viruses.* **2015**, *7*, 5191–5205.
177. Zhao, H.; Chen, M.; Lind, S.B.; Pettersson, U. Distinct temporal changes in host cell lncRNA expression during the course of an adenovirus infection. *Virology* **2016**, *492*, 242–250.
178. 1000 Genomes Project Consortium; Abecasis, G.R.; Auton, A.; Brooks, L.D.; DePristo, M.A.; Durbin, R.M.; Handsaker, R.E.; Kang, H.M.; Marth, G.T.; McVean, G.A. An integrated map of genetic variation from 1092 human genomes. *Nature* **2012**, *491*, 56–65.
179. Schrider, D.R.; Navarro, F.C.P.; Galante, P.A.F.; Parmigiani, R.B.; Camargo, A.A.; Hahn, M.W.; de Souza, S.J. Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS Genet.* **2013**, *9*, e1003242.
180. Richardson, S.R.; Salvador-Palomeque, C.; Faulkner, G.J. Diversity through duplication: Whole-genome sequencing reveals novel gene retrocopies in the human population. *Bioessays* **2014**, *36*, 475–481.



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).