SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

August 2016 Vol.59 No.8: 839–849
doi: 10.1007/s11427-016-5074-x

Evolutionary direction of processed pseudogenes

Guoqing Liu[1,2,4*], Xiangjun Cui[1,2], Hong Li[3] & Lu Cai[1,2]

[1] *School of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou 014010, China;*
[2] *Institute of Bioengineering and Technology, Inner Mongolia University of Science and Technology, Baotou 014010, China;*
[3] *School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China;*
[4] *Computational Systems Biology Lab, Department of Biochemistry and Molecular Biology, Institute of Bioinformatics, University of Georgia, Athens GA 30602, USA*

Received December 16, 2015; accepted March 8, 2016; published online June 16, 2016

While some pseudogenes have been reported to play important roles in gene regulation, little is known about the possible relationship between pseudogene functions and evolutionary process of pseudogenes, or about the forces responsible for the pseudogene evolution. In this study, we characterized human processed pseudogenes in terms of evolutionary dynamics. Our results show that pseudogenes tend to evolve toward: lower GC content, strong dinucleotide bias, reduced abundance of transcription factor binding motifs and short palindromes, and decreased ability to form nucleosomes. We explored possible evolutionary forces that shaped the evolution pattern of pseudogenes, and concluded that mutations in pseudogenes are likely determined, at least partially, by neighbor-dependent mutational bias and recombination-associated selection.

GC content, mutual information, transcription factor binding motifs, mutational bias, recombination

Citation: Liu, G., Cui, X., Li, H., and Cai, L. (2016). Evolutionary direction of processed pseudogenes. Sci China Life Sci 59, 839–849. doi: 10.1007/s11427-016-5074-x

INTRODUCTION

The functions of pseudogenes in gene regulation and genome evolution have been under debate (Balakirev and Ayala, 2003; Pink et al., 2011) since the first example of a pseudogene was identified for the 5S DNA of *Xenopus laevis* in 1977 (Jacq et al., 1977). In particular, many investigations of pseudogenes were triggered by the discovery of the regulatory role of mouse Makorin1-p1 pseudogene in the expression of its homologous Makorin1 gene (Hirotsune et al., 2003). Pseudogenes have been reported in various species of mammals, plants, insects, and bacteria (Lerat and Ochman, 2005; Harrison et al., 2001, 2003; Zhang et al., 2003; Thibaud-Nissen et al., 2009). Depending on the types of parent genes, pseudogenes are divided into two categories, pseudogenes generated from protein-coding genes and pseudogenes generated from RNA genes. Because of sequence degeneration, pseudogenes lose original functions of their parent genes. By convention, the loss of function for pseudogenes that arise from coding genes means they no longer have the ability to code for proteins. Pseudogenes that have remained highly similar to their parent genes can be detected throughout the genome; however, many pseudogenes are likely to have degenerated beyond the detection limit of sequence-similarity search algorithms. With respect to the mechanisms by which pseudogenes may arise, two major classes are defined: duplicated pseudogenes, and processed pseudogenes (or retro-pseudogenes) (Balakirev and Ayala, 2003). Processed pseudogenes are thought to be non-autonomous retrotransposons most likely produced with the assistance of a reverse transcriptase encoded by long interspersed elements (Esnault et al., 2000). Compared with duplicated pseudogenes, processed pseudogenes exhibit a much higher abundance in mammalian genomes

*Corresponding author (email: gqliu1010@163.com)

(Zhang et al., 2002). An additional class of pseudogenes called unitary pseudogenes has been defined (Zhang et al., 2010). Unitary pseudogenes are generated by mutations in functional genes and have no functional counterparts in a genome. Unlike processed pseudogenes, which account for a large proportion of the pseudogenes in the human genome (Zhang et al., 2003), unitary pseudogenes constitute only a small fraction (Zhang et al., 2010).

The identification of a non-trivial number of pseudogenes with explicit biological functions has challenged the popular belief that pseudogenes are non-functional and can simply be considered molecular fossils. For example, a nitric oxide synthase (NOS) pseudogene can regulate the paralogous protein-coding neuronal nNOS gene by producing antisense RNA that forms a duplex with the nNOS mRNA (Korneev et al., 1999); the mouse Makorin1-p1 pseudogene regulates the stability of the mRNA of its homologous Makorin1 gene (Hirotsune et al., 2003); a non-coding RNA transcribed from the MYLKP1 pseudogene can inhibit the expression of the transcript of its parent MYLK gene (Han et al., 2011); there is evidence of transcription for 10 out of 22 ABC transporter pseudogenes in the human genome and siRNA-mediated knockdown of one transcribed pseudogene demonstrated the role of the pseudogene in the regulation of its protein-coding counterpart (Piehler et al., 2008); some pseudogenes have been shown to generate endogenous small interfering RNAs and suppress gene expression via the RNA interference pathway in mouse oocytes (Tam et al., 2008; Watanabe et al., 2008), rice (Guo et al., 2009), and African Trypanosoma brucei (Wen et al., 2011). It was suggested that pseudogenes might serve as an alternative source of natural antisense transcripts that regulate the activity of the sense transcripts of their parent genes. Furthermore, if a pseudogene is transcribed, its transcripts can act as competitive endogenous RNAs (ceRNAs) that modulate the miRNA-mediated repression of other target genes by competing for miRNAs. For example, the tumor-suppressor gene PTEN was reported to be regulated by the abundance of its pseudogene (PTENP1) transcripts (Poliseno et al., 2010), though the regulatory effect of ceRNAs *in vivo* has been under debate (Tay et al., 2011; Denzler et al., 2014; Broderick and Zamore, 2014). This variety of functions already identified may suggest that pseudogenes have evolutionary roles, since the evolutionary fate of a pseudogene population is always related to its possible function(s). Over a long evolutionary time-scale, a young pseudogene with no function may acquire certain functions as it becomes older. If this kind of function acquisition can occur in a non-negligible pseudogene population (not just in rare cases), there must be some intrinsic and on-going evolutionary forces that drive the evolution and determine the fates of pseudogenes. In this respect, it was found that Alus, another kind of non-autonomous retro-transposons, evolve toward enhancers. Specifically, transcription factor binding motifs (TFBMs) and epigenetic marks such as nucleosome occupancy and histone methylation displayed increased enrichments in Alu sequences as Alus became older (Su et al., 2014). It is then natural to consider the possibility that pseudogenes evolve toward gene regulatory elements. This possibility can be investigated because several homologous pseudogenes originate from a single functional gene, particularly for high-transcriptional ribosomal genes and pseudogenes accumulate numerous mutations during evolution. A straightforward approach to investigate the direction of the pseudogene evolution is to evaluate the correlation between pseudogene sequence features and their evolutionary distances.

RESULTS

GC content evolution

The decrease of GC content in pseudogenes was described many years ago; however, the underlying mechanism remains obscure. Zhang et al. compared the GC content of pseudogenes to that of their functional genes and that of their flanking regions; they found that the GC content of pseudogenes, in general, reduced toward that of their flanking regions (Zhang et al., 2002). A similar trend was detected for Hoppsigen pseudogenes in this study (Figure 1A). Moreover, by sorting the pseudogenes by evolutionary distance, we observed a decrease in GC content during evolution for both Gerstein and Hoppsigen pseudogene datasets

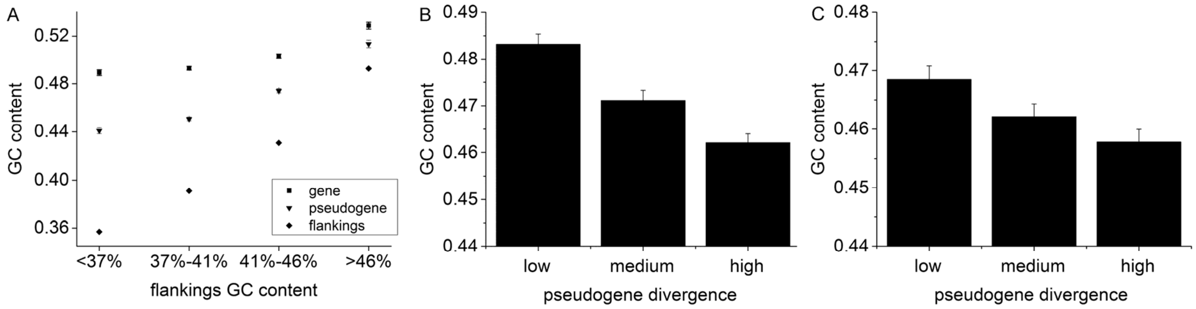

Figure 1 GC content of pseudogenes decreases during evolution. A, GC content of Hoppsigen pseudogenes decreases towards that of their flanking regions. B, Gerstein pseudogenes. C, Hoppsigen pseudogenes.

(Figure 1B and C).

A negative selection hypothesis was proposed to explain the decrease in GC content (Zhang et al., 2002). According to the hypothesis, the insertion of pseudogenes into the regions, in which the GC content remarkably differs from that of inserted pseudogenes can result in instability of the chromatin structure, and this kind of insertions are subjected to negative selection, which could drive the GC content of pseudogenes to decrease and thus resemble their flanking regions (Zhang et al., 2002, 2003). The more remarkable the difference in GC content between ancestral genes and pseudogene flanking regions (surrounding 50-kb regions) is, the more rapidly the GC content of pseudogenes decreases (Figure 1A). When the flanking regions of pseudogenes are redefined as regions of different sizes (15, 30 and 100 kb), we observed similar results, suggesting the independence of the GC decrease to the analyzed flanking size.

Dinucleotide bias evolution

To eliminate the fluctuation effect caused by short sequence length in the mutual information calculation, we concatenated the pseudogenes in each pseudogene divergence bin and calculated mutual information with a non-overlapping window of 5,000 bp. The calculated mutual information values are all above the corresponding fluctuation limit (FL=0.00,159 for 5,000-bp sequence). Our results show that mutual information for adjacent dinucleotides (k=0) is remarkably larger than that for non-adjacent dinucleotides (k=1,2…), which is consistent with the short-range dominance of base correlation in genomic sequences (Figure 2). More importantly, our results showed that mutual information for adjacent dinucleotides (k=0) tends to increase during evolution while mutual information for non-adjacent dinucleotides (k=1,2…) tends to decrease during evolution (Figure 2). Furthermore, mutual information values at k=2, 5, 8,...,20 are higher than those for other values of $k\neq 0$, which is a vestige of the 3-bp periodical occurrence of nucleotides in codons (Figure 2).

Evolution of transcription factor binding motifs

It has been shown that some non-coding elements in the human genome tend to evolve toward functional elements. For example, non-autonomous repeats, Alus, were thought to be a source of gene enhancers because transcription factor binding motifs are enriched in the Alu sequences and the enrichment increases with their evolutionary age. Apart from being a possible source of small RNA-generating elements, we tested the possibility that pseudogenes evolve toward gene regulatory elements by correlating the enrichment of possible TFBMs in pseudogenes with their evolutionary age. We identified possible TFBMs in pseudogenes using Match program (Kel et al., 2003) along with a library of positional weight matrices from the TRANSFAC 6.0 database (Matys et al., 2006). We used Match with the following options: “vertebrate matrix”, “high quality” and “minimize false positive rate”. After TFBMs were identified, pseudogenes were divided into three groups based on the identity between the pseudogenes and their ancestral genes. For each group, enrichment of TFBMs was measured by the number of TFBMs in the group divided by the total sequence length. There are several clues driving us to consider the possibility that TFBMs in pseudogenes enrich with evolutionary time. First, the GC content of pseudogenes decreases with evolutionary time and may approach to that of TFBMs. Second, the increase of nucleotide correlation in pseudogenes during evolution indicates enrichment of short motifs in pseudogenes. Our results, however, show that there is no enrichment of TFBMs in pseudogenes during evolution (Figure 3), suggesting that pseudogenes are not likely to evolve toward gene enhancers. Intriguingly, for Gerstein pseudogenes, we detected a decrease of TFBMs during evolution (Figure 3A).

Evolution of palindrome abundance

Palindromes are important motifs that are involved in many cellular processes, such as DNA replication and gene regulation (Thukral et al., 1991; Hiratsu et al., 2000). To explore the change in palindrome abundance in pseudogenes during evolution, we compared palindrome abundances among the three pseudogene groups (see “MATERIALS AND METHODS” section) that differ in evolutionary distance. We identified short palindromes (<10 bp) in pseudogenes

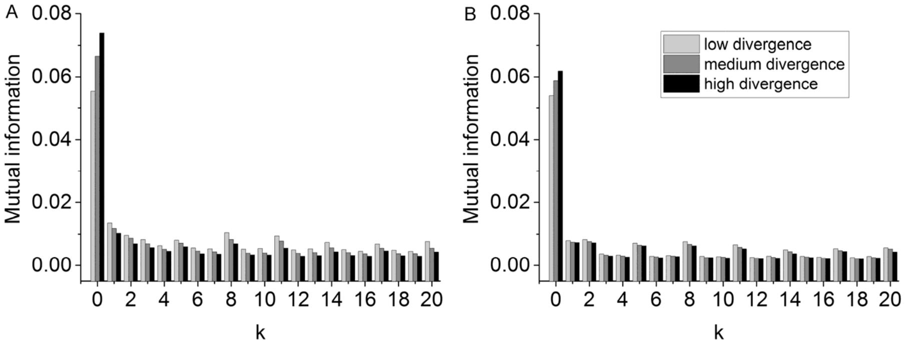

Figure 2 Change in mutual information of pseudogenes during evolution. A, Gerstein pseudogenes. B, Hoppsigen pseudogenes.

using the Spinnaker program (Lisnić et al., 2005). Palindromes longer than 10 bp are not analyzed because their occurrence in the pseudogenes is too small to be statistically significant. Palindrome abundance in random sequences has been shown to increase as sequence composition deviates from GC content of 50% (Liu et al., 2012). Our results, however, show that even though the GC content of pseudogenes deviates from 50% during evolution, short palindromes in pseudogenes do not show any increased enrichment during evolution (Figure 4). Instead, palindromes tend to be depleted from pseudogenes during evolution especially for Gerstein pseudogenes, suggesting that pseudogenes do not evolve toward possible functional sequences enriched with palindrome motifs. For Hoppsigen pseudogenes, although the old pseudogenes (high divergence group) exhibited a slight increase of palindromes of 8–10 bp when compared to the medium divergence group, there was no evidence of palindrome enrichment for the old pseudogenes as compared with their youngest stage (low divergence group).

Evolution of nucleosome positioning signals

There are several major nucleosome positioning signals encoded in genomic sequences. For example, high flexibility together with the 10-bp periodical occurrence of some dinucleotides in sequences can facilitate the nucleosome formation by allowing the DNA sequence to bend more easily around the histone octamer (Segal et al., 2006; Wang et al., 2012). We evaluate whether it is possible that mutations in pseudogenes are selected towards enhancing nucleosome positioning signals such as 10-bp periodicity and DNA flexibility to regulate the expression of its adjacent genes. The evaluation was performed by computing the average nucleosome occupancy for each pseudogene using our sequence-dependent deformation energy model, and compared the nucleosome occupancy between the three pseudogene bins with different divergence. Our results clearly indicate that the pseudogenes' ability to form nucleosomes decreases during evolution (Figure 5). Nucleosome occupancy estimated by Kaplan's model (Kaplan et al., 2009) also exhibited a similar trend (data not shown). Because GC content of a sequence has a strong positive correlation with its nucleosome-forming ability (Tillo and Hughes, 2009), it is very likely that that the reduced ability to form nucleosomes is caused by GC content decrease in pseudogenes. Moreover, in light of the finding that most pseudogenes do not tend to evolve towards functional elements such as TFBMs and palindromic motifs, it seems unnecessary for pseudogenes to maintain their high ability to form nucleosomes, which are important targets for various histone modifications and regulate gene expression.

High divergence of at-risk pseudogenes

In some cases, insertion of transposable elements, such as

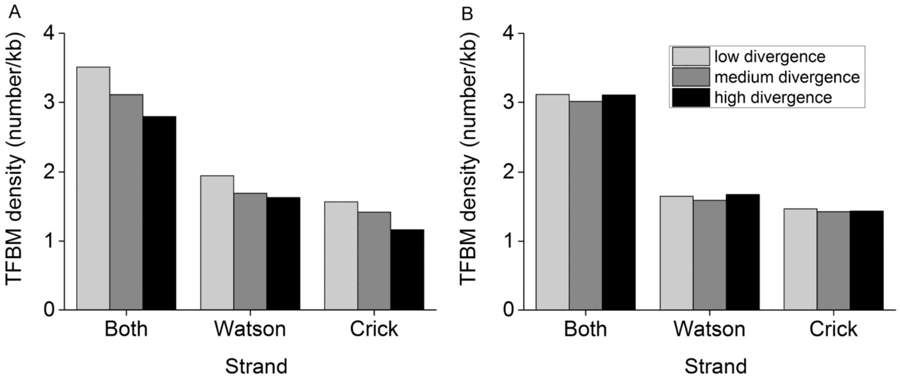

Figure 3 Change in the density of transcription factor binding motifs (TFBMs) during evolution. A, Gerstein pseudogenes. B, Hoppsigen pseudogenes.

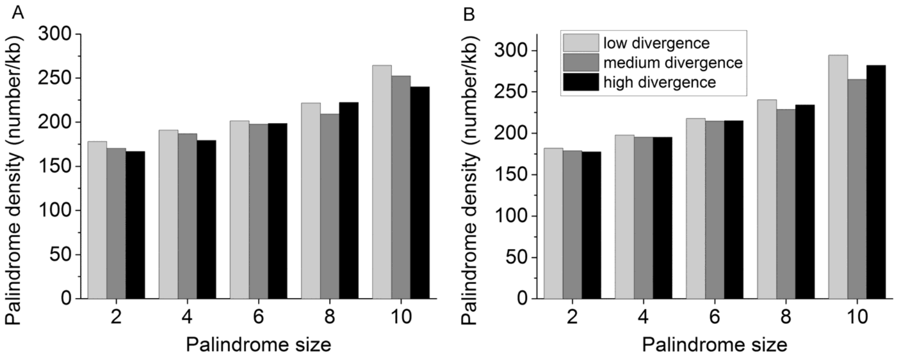

Figure 4 Change in pseudogene palindrome density during evolution. A, Gerstein pseudogenes. B, Hoppsigen pseudogenes.

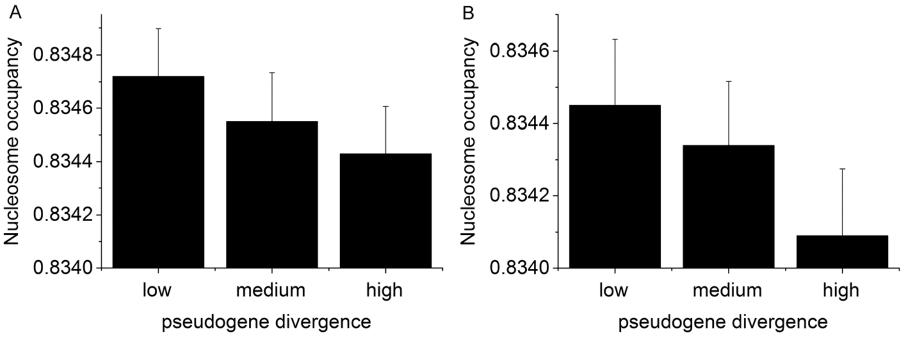

Figure 5 Change in the nucleosome-forming ability of pseudogenes during evolution.

long interspersed nuclear elements, short interspersed nuclear elements, and processed pseudogenes, in the genome can cause disease (Wei and Cao, 2016). We previously showed that homologous pseudogenes tend to cluster in low recombinant regions rather than in high recombinant regions, probably because insertion of homologous pseudogenes in high recombinant regions is negatively selected due to the risk of ectopic recombination between homologous pseudogenes (Liu et al., 2010). If closely located homologous pseudogenes indeed posed risk of instability to the genome, they should decay faster than those separated with a long genomic distance in order to reduce the risk of instability as a result of purifying (or negative) selection. To test this hypothesis, we compared the extent of divergence (relative to ancestors) between Hoppsigen at-risk and non-risk pseudogenes, defined according to their physical location and homology as described in our previous study (Liu et al., 2010). Specifically, the at-risk pseudogenes are homologous pseudogenes that are generated from the same gene and separated by an interval of less than 5 Mb on chromosomes. The remaining pseudogenes are called non-risk ER pseudogenes. We identified 598 at-risk pseudogenes and 4,273 non-risk pseudogenes in Hoppsigen dataset. As expected, we detected a higher divergence for at-risk pseudogenes when compared to non-risk pseudogenes (Figure 6, *t*-test: t=6.01, P=1×10^{-9}). Furthermore, we also found that closely distributed homologous pseudogenes exhibit higher divergence when compared to dispersedly located homologous pseudogenes (Figure 6, *t*-test: t=4.55, P=7.3×10^{-6}), suggesting the fast mutations in closely distributed homologous pseudogenes are likely to be driven by the risk-associated negative selection, which causes the at-risk pseudogenes highly diverge and thus avoid ectopic recombination. It is worth noting that closely located homologous segments may be subject to stronger gene conversion, homogenizing each copy. Therefore, the closely located homologous segments are expected to be more similar. However, in light of the high divergence of closely distributed homologous pseudogenes detected in this study, we speculate that this homogenizing effect by gene conversion may only happen between two similar genomic segments that are located close enough to each other along the genome.

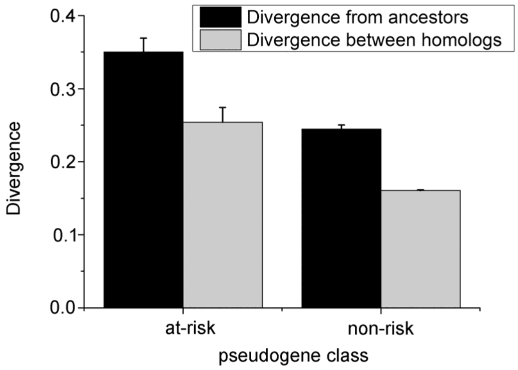

Figure 6 Comparison between at-risk and non-risk pseudogenes with respect to sequence divergence. Both divergence from ancestors and pair-wise divergence among homologs differ significantly between at-risk pseudogenes and non-risk pseudogenes.

We have to point out that our analysis is not quite rigorous because the extent of pseudogene divergence depends not only on selection or mutational bias, but also on the age of the pseudogenes. Nevertheless, we can reject the possibility that the observed difference in pseudogene divergence is caused by pseudogene age by simply assuming pseudogenes with different ages have the same distribution pattern across the genome, which seems rational as there is no evidence of distinct insertion bias between differently-aged pseudogenes. Accordingly, one may also expect that the at-risk pseudogenes should decay faster in regions of high recombination rates when compared to regions of low recombination rates to reduce the risk of ectopic recombination. However, after mapping the pseudogenes to deCODE sex-averaged recombination map (Kong et al., 2002) downloaded from UCSC (http://genome.ucsc.edu), we did not find such a significant difference in divergence of pseudogenes between the two recombination classes (*t*-test: t=0.12, P=0.23). It is possible that when the genomic distance between homologous pseudogenes is reduced to a certain threshold, they are equally subjected to strong negative selection, and become less sensitive to recombination variation. Regardless of recombination, it is clear that there is some risk-related effect in pseudogene mutation as described above, although the effect may influence only a small proportion of pseudogenes, such as closely located homologous pseudogenes.

DISCUSSION

In this study, we attempted to address some questions regarding the evolutionary direction of pseudogenes via characterization of pseudogenes with different ages as indicated by sequence divergence. In order to characterize pseudogene evolution, the following sequence descriptors were used: GC content, mutual information, nucleosome occupancy, abundance of TFBMs and abundance of palindromes. The first three descriptors indeed captured the main trends in pseudogene evolution, while no significant difference was found with respect to motif abundance. In other words, once they occur in the genome, processed pseudogenes gradually evolve toward low GC content, high mutual information, and reduced ability to form nucleosomes. The absence of functional motif enrichment in the pseudogenes suggests that the evolutionary pattern of pseudogenes characterized by the aforementioned descriptors is not driven by selection to increase functional elements. Another possible force driving pseudogene variation is the potential of pseudogenes to produce small RNAs by interacting with their homologous counterparts at the RNA level after transcription. Though such a phenomenon is theoretically possible, there is no evidence to support the hypothesis that specific evolutionary forces relevant to small RNA biogenesis are responsible for pseudogene mutational patterns. A straightforward approach to test this hypothesis would be to analyze the age-related enrichment of small RNAs in pseudogenes with transcriptional activity; this aspect will be the focus of future research efforts.

What is the driving force of the observed pseudogene evolution? In other words, why do pseudogene GC content and ability to form nucleosomes decrease with time, while mutual information increases remarkably? We believe that the reduced ability of pseudogenes to form nucleosomes is caused by the decline in GC content; therefore, we further discuss only the possible reasons for the decline in GC content and for the increase in mutual information. It was suggested that the increase in mutual information can be ascribed to neighbor-dependent mutations occurring in pseudogenes (Zhang and Gerstein, 2003). Neighbor-dependent mutations occur frequently in the human genome, particularly in non- or less functional regions such as pseudogenes, because of an intrinsic mutational bias that is generally thought to be related to DNA stability and other biological processes such as DNA replication and recombination/repair (Arndt and Hwa, 2005). Neighbor-dependent mutations can cause elevated dinucleotide bias, and hence affect mutual information. Notably, our analysis indicated that, in pseudogenes, the increase in mutual information is not caused by a minor proportion of repeats (data not shown). In addition, we point out that, although base correlation (here, the mutual information) in coding sequences is likely to become stronger under certain kinds of selection and become weaker due to random mutations (Luo et al., 1998), this is not the case for pseudogenes. If the majority of the mutations in pseudogenes are assumed to occur either randomly (i.e., without selection of any particular type of mutations that have selective advantage), or under some kind of selection at population level, as will be discussed later on, the above described intrinsic mutational bias of DNA can lead to increased bias in the dinucleotide composition of the pseudogenes.

As with genome evolutionary patterns in general, the decrease in GC content of pseudogenes can also be addressed in terms of mutational bias and possible selection, which were addressed in our study. A negative selection hypothesis was proposed for the evolution of the base composition of Alus towards the composition of their flanking regions (Pavlicek et al., 2001). In the negative selection theory, Alus are stable in regions with similar base composition, but negatively selected if they differ from their flanking regions in terms of base composition. A similar hypothesis was previously invoked to explain the decline in GC content of pseudogenes (Zhang et al., 2002, 2003). Our present results show that there is another kind of negative selection relevant to genome stability that may partially account for the decrease in the GC content of pseudogenes. Namely, the results suggest that closely located homologous pseudogenes with a high risk of inducing genome instability via recombination (i.e., at-risk pseudogenes) are subject to negative selection and decay faster than non-risk pseudogenes.

Processed pseudogenes resemble Alu transposons in many aspects, such as biogenesis, rapid decay of poly(A) tracts over time, high GC content of young elements, etc. However, our data indicate that their evolutionary fates differ significantly. Pseudogenes degenerate toward their flanking regions under negative selection, whereas Alus are prone to evolve toward enhancers under positive selection, even though aforementioned negative selection may still affect Alu evolution. In some aspects, the epigenetic marks in pseudogenes resemble thouse in Alus, while in other aspects they do not. For example, methylation occurs more in young Alus. Specifically, CpG sites abundant in Alu regions are subject to extensive methylation, and then the methylated CpGs readily mutate into TpG, as Alus get older. Considering the high substitution rate from CpG to TpG in pseudogenes (Zhang and Gerstein, 2003), it is very likely that the methylation dynamics in pseudogenes resemble those in Alus. However, given that the distance between the methylated sites in pseudogenes and nearby genes is, on average, larger than that in Alus, the methylation pattern in pseudogenes may have much less impact on the expression of nearby genes. At the level of chromatin structure, older Alus are rich in histone modification marks (Su et al., 2014), which can regulate the expression of nearby genes. In contrast, as our data indicate that pseudogenes are less occupied by nucleosomes over time, and hence histone modification marks in pseudogenes may become scarce over time. Unlike pseudogenes, Alu elements, especially the older ones, contain many TFBMs (Polak and Domany, 2006; Su et al., 2014). We think the difference can be ascribed largely to two factors: population (or amplification capability) and sequence composition. There are approximately 1.1 million copies of Alu in the human genome, which represents a higher abundance than those of both pseudogenes (approximately 20,000) and genes that can generate processed pseudogenes through retro-transposition. Although a large number of Alus might be inactive or less active in terms of transcription, it seems evident that Alus still have higher capacity of amplification through retro-transposition than pseudogenes due to their tremendous population. It is preferable for the selection to rely on easy-to-use repeats like Alu to expand TFBMs, because the Alu elements are not only densely distributed in the genome, but may also be a better source of binding motifs than pseudogenes. A similar previous prediction that some repeats harbor better binding motifs than random promoter templates proved to be true (Bourque et al., 2008). In other words, for evolution, Alus provide a way to create numerous similar TFBMs in a short time through their insertion in the proximal upstream regions of genes and the inclusion of a few subsequent point mutations, if needed, to optimize the binding sites. In contrast, an individual processed pseudogene cannot reach a very high copy number, and thus pseudogenes are less likely to generally serve as small regulatory elements. Furthermore, the insertion of pseudogenes, most of which are longer than Alus, in the vicinity of genes might be subject to stronger negative selection than in the case of Alus (Liu et al., 2010). There might be another mutually exclusive interpretation for the difference in evolutionary fate between Alus and pseudogenes. That is, if pseudogenes are rich in the signatures of binding motifs at their origins, and Alus lack such signatures at their appearance, random mutations may result in a lower and a higher density of binding sites, respectively, in old pseudogenes and old Alus than in their younger counterparts. However, this possibility seems unlikely because the processed pseudogenes stem from gene mRNA transcripts, which may have much less propensity to be bound by transcription factors, although a small fraction (1.8%) of transcription factor binding sites were shown to be located in coding exons (Stergachis et al., 2013). Then, is it possible that post-transcriptional regulatory properties differ between pseudogenes and Alu elements, leading to their distinct evolutionary trajectories? Previous studies demonstrated the presence of a large number of endogenous siRNAs and piRNAs in mouse oocytes and embryonic stem cells, most of which are derived from retrotransposons, bidirectional transcription, and antisense transcripts (Babiarz et al., 2008; Watanabe et al., 2008). It can be hypothesized that transcribed pseudogenes having high sequence similarity with their parent genes can serve as a more suitable pool of small regulatory RNAs or ceRNAs that can regulate the expression of their protein-coding counterparts than some other kinds of sources like Alus or long non-coding RNAs. Although in some cases, transcribed pseudogenes are known to be involved in post-transcriptional regulatory networks (Tam et al., 2008; Guo et al., 2009; Poliseno et al., 2010), a thorough comparative study is required to evaluate their relative potential of being a source of small RNA and ceRNA, and of global regulatory effect at population level. In addition, possible roles of non-coding RNAs, which can be derived from pseudogenes, in epigenetic regulation (e.g. histone modification and chromatin remodeling) still remain an open question (Chen and Xue, 2016).

The sequencing of the human genome revealed that the human genome contains fewer protein-coding genes than *Arabidopsis thaliana* and much fewer protein-coding genes than originally expected. Many biologists believed that, to achieve the phenotypic complexity of humans, the human genome must have complex gene regulatory systems. With the discovery of miRNA, numerous other non-coding RNAs, as well as pervasive alternative splicing, the above idea has been deeply reinforced. Upon publication of the results of the ENCODE project (The ENCODE Project Consortium, 2012), the authors had assigned functions to 80% of the human genome, which suggested that most of the human genome sequences are functional. Consequently, some biologists claimed, with great excitement, that the concept of junk DNA is dead (Pennisi, 2012; Ecker et al., 2012). This view was later criticized by evolutionary biologists (Eddy, 2012, 2013; Niu and Jiang 2013; Graur et al., 2013; Doolittle, 2013). What the ENCODE Consortium had found was simply that most of the human genome is active. “Active” does not equal “functional”. In this context, we need to carefully examine each of the major groups of so-called functional non-coding sequences. de Souza et al. examined the claimed functionality of transposable elements and found that most are still inconclusive (de Souza et al., 2013). Similar attempts have been carried out for transcription factor binding sites (Paris et al., 2013), alternatively splicing (Wang et al., 2014), and non-coding RNAs (Palazzo and Lee, 2015). In this study, we showed evidence that most of the pseudogenes are unlikely to be functional. These results contribute to our understanding of the genome.

To conclude, we investigated the evolutionary characteristics of human processed pseudogenes; we found several evolutionary patterns of pseudogenes, and explored the possible driving forces for pseudogene evolution. Our results indicate that pseudogenes evolve toward low GC content, strong dinucleotide bias, and reduced ability to form nucleosomes, which can be explained by negative selection associated with chromatin structure. In addition, our data suggest that pseudogene population is unlikely to evolve toward functional elements, which contrasts with the trends previously found for Alu transposons.

MATERIALS AND METHODS

Pseudogene sequences

We used two datasets of human processed pseudogenes. One is composed of ribosomal protein (RP) pseudogenes, which are generated from evolutionary conserved RP genes. We used this dataset because RP pseudogene family is the largest pseudogene family (a set of pseudogenes assigned to a gene family) in the human genome (over 2000 pseudogenes generated from 78 ribosomal protein genes), which makes it more statistically helpful for studying pseudogene evolutionary characteristics than small-sized pseudogene families. We downloaded the primary DNA sequences and annotations (hg16-based) for 2,536 RP pseudogenes (Zhang et al., 2002) from the pseudogene database (http://pseudogene.org). After excluding duplicated pseudogenes and pseudogenic fragments, we obtained 1,931 processed pseudogenes generated from 78 ribosomal protein genes. This set of pseudogenes, identified by Gerstein lab, is denoted as Gerstein pseudogenes in this study.

The other set of processed pseudogenes was retrieved from the Hoppsigen database (Khelifi et al., 2005). The Hoppsigen dataset was shown to be a high-quality genome-scale pseudogene dataset (Khelifi et al., 2005). Gerstein and Hoppsigen pseudogenes were identified by similarity search against the human genome using protein sequences and coding sequences of genes, respectively, as queries. Multiple alignment results of Hoppsigen processed pseudogenes (7,105) were obtained by WWW-Query search (http://pbil.univ-lyon1.fr/search/quick_test.php). To analyze recombination rates (hg18-based) for Hoppsigen pseudogenes, we successfully mapped 5,055 Hoppsigen processed pseudogenes on the human genome (hg18 version) using Blat (Kent, 2002).

A feature (e.g. mutual information) of a pseudogene is subjected to two factors: the corresponding feature of the gene that generated the pseudogene and the divergence of the pseudogene. We did not focus on the influence of ancestral genes on pseudogene features, but rather on how features of pseudogenes change with their evolutionary distance (divergence), but. Therefore, we prepared each of the pseudogene datasets in a specific manner. We did not simply divide pseudogenes into three bins according to evolutionary divergence, as it is commonly done. Instead, each pseudogene family that contained at least three pseudogenes was divided into three equal-sized bins according to evolutionary divergence, which constituted three pseudogene groups: low divergence, medium divergence, and high divergence. The sample size of each pseudogene group for Gerstein and Hoppsigen pseudogenes was 651 and 905, respectively. Pseudogene families with less than 3 pseudogenes were not included in our analysis. Thus, we expected to avoid possible compositional bias caused by influence of ancestral genes on corresponding pseudogenes, and to assure that results based on this dataset can reflect evolutionary characteristics of pseudogenes.

Divergence of pseudogenes

We used the divergence of Gerstein pseudogenes, which is available at http://pseudogene.org. Based on the downloaded multiple alignment results for each Hoppsigen pseudogene family (including pseudogenes and corresponding protein-coding gene), we computed their divergence from corresponding consensus sequences by using MEGA3 (Kumar et al., 2004) with Kimura's two-parameter model (Kimura, 1980). It is well accepted that a consensus sequence obtained from multiple-sequence alignment can represent, although roughly, the ancestor of the aligned homologous sequences.

Mutual information

Mutual information (Kullback, 1959) in genomic sequences is defined as

$$D_2 = \sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_i p_j} \quad (1)$$

where p_i and p_j (i, j=A, G, C, T) are, respectively, the probability of occurrence of nucleotide i and j, and p_{ij} is the probability of single-base-step occurrence of dinucleotide ij in a sequence. Mutual information has also been described as base correlation (Luo et al., 1998).

Generalized mutual information for two nucleotides with a distance of k bp (k=0, 1, 2,…) can be defined as

$$D_{2+k} = \sum_{i,j} p_{i(k)j} \log_2 \frac{p_{i(k)j}}{p_i p_j} \quad (2)$$

where D_2 quantifies the overall deviation of observed probabilities of 16 dinucleotides from those expected by chance. A high value of mutual information indicates a high dinucleotide bias and implicates some biological sense of genomic sequences that is related to selection or mutational bias. Double-stranded versions of mutual information can be computed from the sequence concatenated with its inverted complementary sequence. Individualized treatment for the sense and anti-sense strands of transcriptional sequences seems important for substitution pattern analysis in light of non-complementarity of substitutions in some genomes (Singh et al., 2005). However, in most cases, there is no significant difference between single-stranded and double-stranded sequence descriptors. Furthermore, no evidence of non-complementarity patterns of substitution for human pseudogenes has been reported though a proportion of pseudogenes can be transcribed. Therefore, we assessed dinucleotide bias from both strands of pseudogene sequences.

Nucleosome occupancy estimates

In vivo nucleosome positions along the genome are determined jointly by intrinsic sequence preference, and many other non-sequence factors, such as chromatin remodelers, transcription factor binding, etc. In our study, we focused on the sequence-directed nucleosome formation ability of pseudogenes, and therefore we used two sequence-based models to predict nucleosome occupancy. One is based on deformation energy of DNA molecules (Liu et al., 2015), and the other is a bioinformatics method proposed by Kaplan et al. 2009. The deformation energy-based model is outlined below.

Nucleosome formation ability is negatively correlated with deformation energy of a DNA segment. The bending energy of a *L*-bp sequence was calculated by

$$\begin{aligned} E_\mathrm{b} &= \sum_{-(L-1)/2}^{(L-1)/2} E_\mathrm{b}(i) \\ &= \sum_{-(L-1)/2}^{(L-1)/2} \left[\frac{F_\mathrm{b}^2}{2k_\rho(i)} \cos^2 \Omega_i + \frac{F_\mathrm{b}^2}{2k_\tau(i)} \sin^2 \Omega_i \right] \end{aligned} \tag{3}$$

where

$$F_\mathrm{b} = \frac{\alpha - \sum_i \rho_0(i) \cos \Omega_i - \sum_i \tau_0(i) \sin \Omega_i}{\sum_i \frac{\cos^2 \Omega_i}{k_\rho(i)} + \sum_i \frac{\sin^2 \Omega_i}{k_\tau(i)}} \tag{4}$$

In the above two equations, $\rho(i)$ and $\tau(i)$ are, respectively, the actual roll and tilt angles of the dinucleotide at step *i*; $\rho_0(i)$ and $\tau_0(i)$, which are dependent on the dinucleotide at step *i*, are, respectively, the roll and tilt angles without torque; and $k_\rho(i)$ and $k_\tau(i)$ are the dinucleotide-dependent force constants; Ω_i is the cumulative helical twist at step *i*, counted from the dyad point. A constant helical twist of 34.8°, which is the average twist for the 1kx5 X-ray crystal structure of nucleosome-bound DNA, was used for all dinucleotide steps (Richmond and Davey, 2003). For a 147-bp nucleosomal sequence, only the central 129-bp contribute to its bending (Richmond and Davey, 2003). Hence, *L* was set to 129. The empirical parameters of the model including force constants (k_ρ and k_τ) and equilibrium structural parameters (ρ_0 and τ_0) for 10 dinucleotides (complementary dinucleotides are considered to be the same) were taken from Morozov et al. (Morozov et al., 2009).

Nucleosome occupancy at the *j*th base-pair site was estimated by

$$O_j = \sum_{i=j-(l-1)/2}^{j+(l-1)/2} e^{-\beta E_j} \Big/ l \tag{5}$$

where E_j is the deformation energy of the segment defined from positions j–64 to j+64, l=51, and inverse temperature β is assumed to be 1 for simplicity.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *We thank Ms. Zhaoxia Zhao for her help in data preparation. We are also grateful to anonymous reviewers and editors for their help in improving the manuscript. This work was supported by the National Natural Science Foundation (61102162, 61271448, 31260274, 31260219), the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (NJYT-14-B10), and the Inner Mongolia Science & Technology Plan (20140401).*

Arndt, P.F., and Hwa, T. (2005). Identification and measurement of neighbor-dependent nucleotide substitution processes. Bioinformatics 21, 2322–2328.

Babiarz, J.E., Ruby, J.G., Wang, Y., Bartel, D.P., and Blelloch, R. (2008). Mouse ES cells express endogenous shRNAs, siRNAs, and other microprocessor-independent, Dicer-dependent small RNAs. Genes Dev 22, 2773–2785.

Balakirev, E.S., and Ayala, F.J. (2003). Pseudogenes: are they "junk" or functional DNA? Annu Rev Genet 37, 123–151.

Bourque, G., Leong, B., Vega, V.B., Chen, X., Lee, Y.L., Srinivasan, K.G., Chew, J.L., Ruan, Y., Wei, C.L., Ng, H.H., and Liu, E.T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res 18, 1752–1762.

Broderick, J.A., and Zamore, P.D. (2014). Competitive endogenous RNAs cannot alter microRNA function *in vivo*. Mol Cell 54, 711–713.

Chen, J., and Xue, Y. (2016). Emerging roles of non-coding RNAs in epigenetic regulation. Sci China Life Sci 59, 227–235.

Denzler, R., Agarwal, V., Stefano, J., Bartel, D.P., and Stoffel, M. (2014). Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. Mol Cell 54, 766–776.

de Souza, F.S., Franchini, L.F., and Rubinstein, M. (2013). Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? Mol Biol Evol 30, 1239–1251.

Doolittle, W.F. (2013). Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci USA 110, 5294–5300.

Ecker, J.R., Bickmore, W.A., Barroso, I., Pritchard, J.K., Gilad, Y., and Segal, E. (2012). Genomics: ENCODE explained. Nature 489, 52–55.

Eddy, S.R. (2013). The ENCODE project: missteps overshadowing a success. Curr Biol 23, R259–R261.

Eddy, S.R. (2012). The C-value paradox, junk DNA and ENCODE. Curr Biol 22, R898–R899.

Esnault, C., Maestre, J., and Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. Nat Genet 24, 363–367.

Graur, D., Zheng, Y., Price, N., Azevedo, R.B.R., Zufall, R.A., and Elhaik, E. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol 5, 578–590.

Guo, X., Zhang, Z., Gerstein, M.B., and Zheng, D. (2009). Small RNAs originated from pseudogenes: *cis*- or *trans*-acting? PLoS Comput Biol 5, e1000449.

Han, Y.J., Ma, S.F., Yourek, G., Park, Y.D., and Garcia, J.G. (2011). A transcribed pseudogene of MYLK promotes cell proliferation. FASEB J 25, 2305–2312.

Harrison, P.M., Echols, N., and Gerstein, M.B. (2001). Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. Nucleic Acids Res 29, 818–830.

Harrison, P.M., Milburn, D., Zhang, Z., Bertone, P., and Gerstein, M. (2003). Identification of pseudogenes in the *Drosophila melanogaster* genome. Nucleic Acids Res 31, 1033–1037.

Hiratsu, K., Mochizuki, S., and Kinashi, H. (2000). Cloning and analysis of the replication origin and the telomeres of the large linear plasmid pSLA2-L in *Streptomyces rochei*. Mol Gen Genet 263, 1015–1021.

Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S., Yagami, K., Wynshaw-Boris, A., and Yoshiki, A. (2003). An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature 423, 91–96.

Jacq, C., Miller, J.R., and Brownlee, G.G. (1977). A pseudogene structure in 5S DNA of *Xenopus laevis*. Cell 12, 109–120.

Khelifi, A., Duret, L., and Mouchiroud, D. (2005). HOPPSIGEN: a database of human and mouse processed pseudogenes. Nucleic Acids Res 33, D59–D66.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., and Segal, E. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458, 362–366.

Kel, A.E., Goessling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., and Wingender, E. (2003). Match (TM): a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31, 3576–3579.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res 12, 656–664.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16, 111–120.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S.T., Frigge, M.L., Thorgeirsson, T.E., Gulcher, J.R., and Stefansson, K. (2002). A high resolution recombination map of the human genome. Nat Genet 31, 241–247.

Korneev, S.A., Park, J.H., and O'Shea, M. (1999). Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. J Neurosci19, 7711–7720.

Kullback, S. (1959). Information Theory and Statistics. New York: Wiley.

Kumar, S., Tamura, K., and Nei, M. (2004). MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 5, 150–163.

Lerat, E., and Ochman, H. (2005). Recognizing the pseudogenes in bacterial genomes. Nucleic Acids Res 33, 3125–3132.

Lisnić, B., Svetec, I.K., Sarić, H., Nikolić, I., and Zgaga, Z. (2005). Palindrome content of the yeast *Saccharomyces cerevisiae* genome. Curr Genet 47, 289–297.

Liu, G., Liu, J., and Zhang, B. (2012). Compositional bias is a major determinant of the distribution pattern and abundance of palindromes in *Drosophila melanogaster*. J Mol Evol 72, 130–140.

Liu, G., Li, H., and Cai, L. (2010). Processed pseudogenes are located preferentially in regions of low recombination rates in the human genome. J Evol Biol 23, 1107–1115.

Liu, G., Feng, F., Zhao, X., and Cai, Lu. (2015). Nucleosome organization around pseudogenes in the human genome. BioMed Res Int 2015, 821596.

Luo, L., Lee, W., Jia, L., Ji, F., and Tsai, L. (1998). Statistical correlation of nucleotides in a DNA sequence. Phys Rev E 58, 861–871.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., and Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 34, D108–D110.

Morozov, A.V., Fortney, K., Gaykalova, D.A., Studitsky, V.M., Widom, J., and Siggia, E.D. (2009). Using DNA mechanics to predict *in vitro* nucleosome positions and formation energies. Nucleic Acids Res 37, 4707–4722.

Niu, D.K., and Jiang, L. (2013). Can ENCODE tell us how much junk DNA we carry in our genome? Biochem Biophys Res Commun 430, 1340–1343.

Palazzo, A.F., and Lee, E.S. (2015). Non-coding RNA: what is functional and what is junk? Front Genet 6, 2.

Paris, M., Kaplan, T., Li, X.Y., Villalta, J.E., Lott, S.E., and Eisen, M.B. (2013). Extensive divergence of transcription factor binding in *Drosophila* embryos with highly conserved gene expression. PLoS Genet 9, e1003748.

Pavlicek, A., Jabbari, K., Paces, J., Paces, V., Hejnar, J.V., and Bernardi, G. (2001). Similar integration but different stability of Alus and LINEs in the human genome. Gene 276, 39–45.

Pennisi, E. (2012). ENCODE project writes eulogy for junk DNA. Science 337, 1159–1161.

Piehler, A.P., Hellum, M., Wenzel, J.J., Kaminski, E., Haug, K.B., Kierulf, P., and Kaminski, W.E. (2008). The human ABC transporter pseudogene family: evidence for transcription and gene-pseudogene interference. BMC Genomics 9, 165.

Pink, R.C., Wicks, K., Caley, D.P., Punch, E.K., Jacobs, L., and Carter, D.R.F. (2011). Pseudogenes: pseudo-functional or key regulators in health and disease. RNA 17, 792–798.

Polak, P., Domany, E. (2006). Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics 7, 133.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., and Pandolfi, P.P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465, 1033–1038.

Richmond, T.J., and Davey, C.A. (2003). The structure of DNA in the nucleosome core. Nature 423, 145–150.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature 442, 772–778.

Singh, N.D., Arndt, P.F., and Petrov, D.A. (2005). Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. Genetics 169, 709–722.

Stergachis, A.B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., Raubitschek, A., Ziegler, S., LeProust, E.M., Akey, J.M., and Stamatoyannopoulos, J.A. (2013). Exonic transcription factor binding directs codon choice and affects protein evolution. Science 342, 1367–1372.

Su, M., Han, D., Boyd-Kirkup, J., Yu, X., Han, J.D.J. (2014). Evolution of Alu elements toward enhancers. Cell Rep 7, 376–385.

Tam, O.H., Aravin, A.A., Stein, P., Girard, A., Murchison, E.P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R.M., and Hannon, G.J. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature 453, 534–538.

Tay, Y., Kats, L., Salmena, L., Weiss, D., Tan, S.M., Ala, U., Karreth, F., Poliseno, L., Provero, P., Di Cunto, F., Lieberman, J., Rigoutsos, I., and Pandolfi, P.P. (2011). Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. Cell 147, 344–357.

The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Thibaud-Nissen, F., Ouyang, S., and Buell, C.R. (2009). Identification and characterization of pseudogenes in the rice gene complement. BMC Genomics 10, 317.

Thukral, S.K., Eisen, A., and Young, E.T. (1991). Two monomers of yeast transcription factor ADR1 bind a palindromic sequence symmetrically to activate ADH2 expression. Mol Cell Biol 11, 1566–1577.

Tillo, D., and Hughes, T.R. (2009). G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics 10, 442.

Wang, J.Y., Wang, J., and Liu, G. (2012). Calculation of nucleosomal DNA deformation energy: its implication for nucleosome positioning. Chromosome Res 20, 889–902.

Wang, M., Zhang, P., Shu, Y., Yuan, F., Zhang, Y., Zhou, Y., Jiang, M., Zhu, Y., Hu, L., Kong, X., and Zhang, Z. (2014). Alternative splicing at GYNNGY 5′ splice sites: more noise, less regulation. Nucleic Acids Res 42, 13969–13980.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., Surani, M.A., Sakaki, Y., and Sasaki, H. (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature 453, 539–543.

Wei, L., and Cao, X. (2016). The effect of transposable elements on phenotypic variation: insights from plants to humans. Sci China Life Sci 59, 24–37

Wen, Y.Z., Zheng, L.L., Liao, J.Y., Wang, M.H., Wei, Y., Guo, X.M., Qu, L.H., Ayala, F.J., and Lun, Z.R. (2011). Pseudogene-derived small interference RNAs regulate gene expression in African *Trypanosoma brucei*. Proc Natl Acad Sci USA 108, 8345–8350.

Zhang, Z.D., Frankish, A., Hunt, T., Harrow, J., and Gerstein, M.B. (2010). Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol 11, R26.

Zhang, Z., and Gerstein, M. (2003). Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acid Research 31, 5338–5348.

Zhang, Z., Harrison, P.M., Liu, Y., and Gerstein, M. (2003). Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res13, 2541–2558.

Zhang, Z., Harrison, P., and Gerstein, M. (2002). Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res 12, 1466–1482.