ELSEVIER

Gene 259 (2000) 61–67

GENE
AN INTERNATIONAL JOURNAL ON GENES, GENOMES AND EVOLUTION

www.elsevier.com/locate/gene

Genomic scrap yard: how genomes utilize all that junk ☆

Wojciech Makałowski *

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received 31 May 2000; received in revised form 27 July 2000; accepted 25 August 2000
Received by T. Gojobori

Abstract

Interspersed repetitive sequences are major components of eukaryotic genomes. Repetitive elements comprise over 50% of the mammalian genome. Because the specific function of these elements remains to be defined and because of their unusual 'behavior' in the genome, they are often quoted as a selfish or junk DNA. Our view of the entire phenomenon of repetitive elements has to now be revised in light of data on their biology and evolution, especially in the light of what we know about the retroposons. I would like to argue that even if we cannot define the specific functions of these elements, we still can show that they are not useless pieces of the genomes. The repetitive elements interact with the whole genome and influence its evolution. Repetitive elements interact with the surrounding sequences and nearby genes. They may serve as recombination hot spots or acquire specific cellular functions such as RNA transcription control or even become part of protein coding regions. Finally, they provide very efficient mechanism for genomic shuffling. As such, repetitive elements should be called *genomic scrap yard* rather than *junk DNA*. Tables listing examples of recruited (exapted) transposable elements are available at http://www.ncbi.nlm.nih.gov/Makalowski/ScrapYard/.

Keywords: Exaptation; Genome shuffling; Genomic evolution; Junk DNA; Repetitive elements; Retrogenes

1. Introduction

Eukaryotic genomes are very complex and dynamic entities. Only a fraction of these genomes are occupied by protein coding exons, while the majority of non-exonic sequences consist of repetitive elements. For example, in mammalian genomes functional exons contribute to merely 2% of a genome, up to 50% of a genome is occupied by repetitive elements, while the remaining 48% is called unique DNA, most of which probably originated in mobile elements diverged over time beyond recognition (see Fig. 1). The complete sequences of human chromosome 21 and 22 revealed a surprisingly high number of pseudogenes (see discussion below).

In 1969 Masatoshi Nei first noticed the importance of non-exonic sequences and called them non-sense DNA (Nei, 1969). A few years later Suzumu Ono coined the term *junk DNA* to describe this phenomenon (Ohno, 1972). The abundance of the repetitive sequences has no immediate rational explanation; there are many very successful organisms with compact genomes, e.g. all prokaryotes, fugu among vertebrates, or *Arabidopsis thaliana* among flowering plants. Therefore many researchers view those elements as unnecessary ballast, burden for a genome, and compare them to parasites (Hickey, 1982), a selfish DNA exploiting eukaryotic genomes (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). With progress of the human genome project, our understanding of our genome increases, including the role and structure of non-coding sequences. At the same time, more and more biologists regard repetitive elements as a genomic treasure (Brosius, 1991; Nowak, 1994; Brosius, 1999). Several years ago (Makałowski, 1995) I introduced the concept of a scrap yard to describe the role of repetitive elements, particularly retrosequences, in genomic evolution. Recent years have witnessed accelerated progress in understanding of genome dynamics. It appears that different mobile elements play a significant role in this process. In this paper I review different contributions of repetitive elements to host genomic evolution.

Abbreviations: DAF, decay accelerating factor; LINE, long interspersed element; L1, LINE-1 element; LTR, long terminal repeat.

☆ Paper presented at the international symposium 'Evolution 2000: Biodiversity in Network of Bioinformation from the Standpoint of Devo-Evo View', Tokyo, March 5, 2000.

* Tel.: +1-301-435-5989. fax: +1-301-480-9241.
E-mail address: makalowski@ncbi.nlm.nih.gov (W. Makałowski)

0378-1119/00/$ - see front matter
PII: S0378-1119(00)00436-4

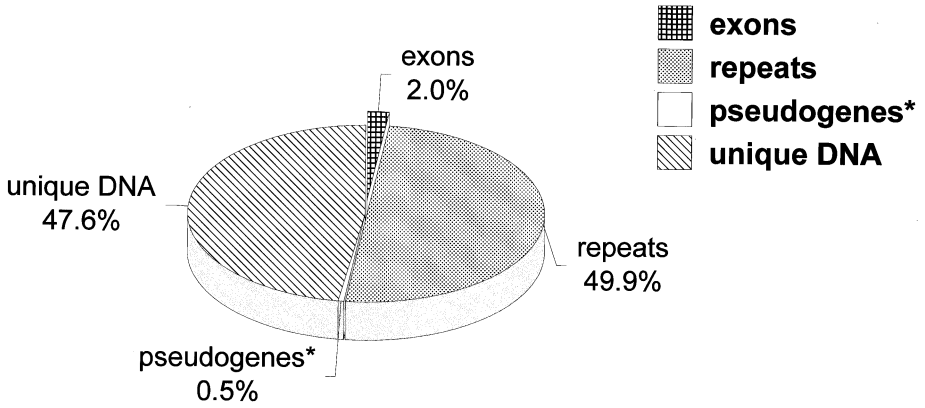

Fig. 1. Fractions of a human genome occupied by different sequence types. The pseudogenes fraction is an extrapolation from chromosome 21 data.

2. Repetitive elements and recombination events

Recombination is a very powerful factor of evolution that produces genetic variability by using already existing blocks of biological information. Computer simulations show that DNA sequences may evolve faster by homologous recombination than by point mutation (Levinson, 1994). The repetitive elements play an important role in the unequal homologous recombination events. Because of their sequence similarity, they enable pairing and exchange between unrelated fragments of chromatin, leading to deletion or duplication of a genomic fragment. Although most of the observed recombination events lead to pathological events (Kazazian, 1998; Deininger and Batzer, 1999), sometimes recombination has positive evolutionary effects, for example the human glycophorin gene family evolved through several duplication steps that involved recombination between Alu elements (see Fig. 2). Misaligned repetitive elements can promote unequal crossing-over events over long distances, for example recombination between two *Mariner* elements led to duplication of a 30 kb fragment of human chromosome 17 (Reiter et al., 1996).

3. Genomic motifs originated in retrosequences

One of the most direct influences of transposable elements on the host genome is their role in modulating of structure and expression of 'native' genes. This phe-

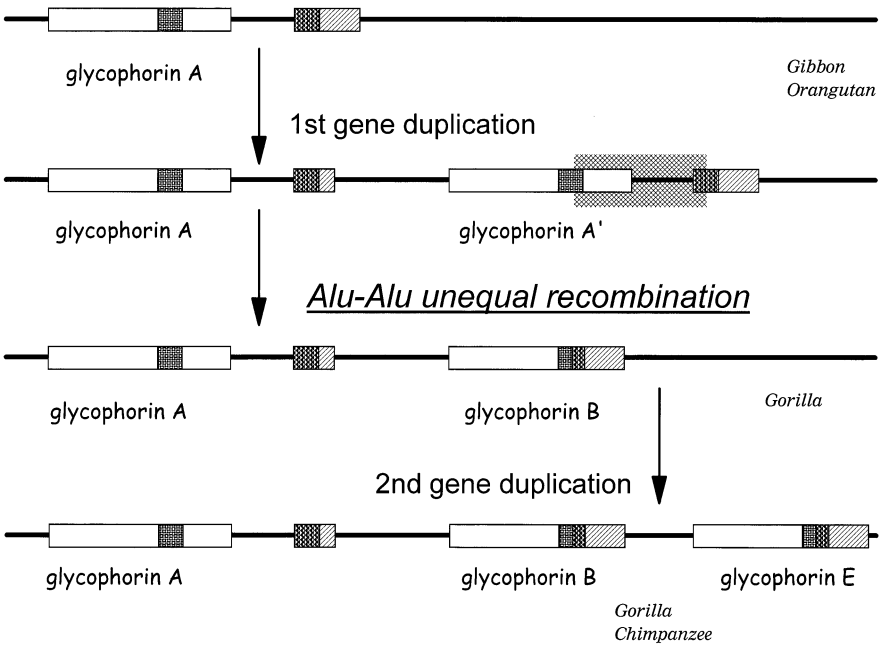

Fig. 2. Evolution of the primate glycophorin gene family. The primordial glycophorin gene was duplicated and one of the copies gave rise to the glycophorin B gene, through an unequal recombination mediated by Alu sequences. Another duplication led to the glycophorin E gene which is not completely fixed in the Gorilla species. The Alu elements are indicated by shaded boxes, the glycophorin genes by open boxes, and the genomic precursor sequence at the 3′ end of the glycophorin B and E genes is indicated by hatched boxes.

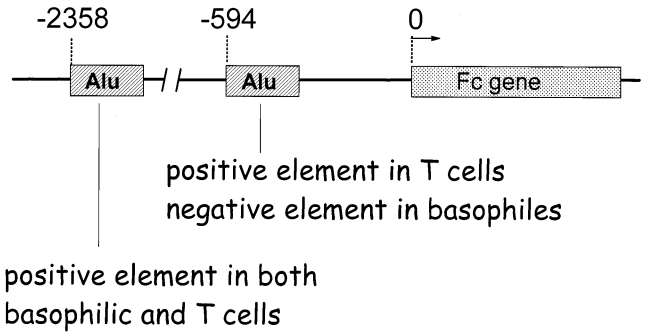

Fig. 3. Cell-specific regulation of the IgE receptor (FcεRI-γ) gene.

nomenon in recent years was a subject of several excellent reviews (Brosius, 1991, 1999). Here I will discuss only several examples. The most up-to-date list of such examples is maintained by Juergen Brosius from the University of Muenster, and is available on the Internet at http://www.crosswinds.net/%7Eexpath/references/addmat/add0101.htm. Here I will describe just a handful of examples of retroelements recruited by a host genome for a new role.

3.1. Transcriptional regulatory elements

After the discovery that long terminal repeats (integral parts of some retroelements) carry promoter and enhancer motifs, it became clear that integration of such elements in the proximity of a host gene must have an influence on this gene expression (Sverdlov, 1998). But not only LTRs can influence a gene expression. Also non-LTR retroposons can influence adjacent gene expression. For example, IgE receptor gene (FcεRI-γ) is cell-specific regulated by motifs that are part of Alu elements inserted upstream of FcεRI-γ gene (see Fig. 3). There are two Alu elements upstream of FcεRI-γ gene. The more distant serves as a positive element in both basophilic and T cell, while the other one (about 600 nt upstream of the transcription start) acts as a positive element in T cells but is a negative element in basophiles. Apparently, in the two types of cells different transcription factors are expressed that interact with '-600 Alu'.

3.2. Polyadenylation signals

Eukaryotic mRNA precursors are modified at their 3′ end. The newly synthesized RNA is cleaved 10–20 nt downstream of the A(A/U)UAAA sequence, which is called the polyadenylation signal. Since this signal can easily be created by a single point mutation within the poly-A tail of many retroelements, retroposition downstream to the coding region of a host gene could serve as a source of new polyadenylation signals. Indeed, several reports show poly-A signals originated in retroelements. Different retroposons in different organisms contribute poly-A signals. In Lagomorpha C repeat is a frequent contributor of poly-A signals, about 10% of the C repeats analyzed by Krane and Hardison contain an active signal (Krane and Hardison, 1990) and some of them create alternative transcripts (Boggaram et al., 1988). The insertion of an L1 element downstream of the open reading frame is responsible for the activation of a cryptic poly-A signal in the mice thymidylate synthase gene and for an unusual polyadenylation of the mRNA at the stop codon (Harendza and Johnson, 1990). Human THE-1 transposon and Alu elements are also known to contribute poly-A signals (Paulson et al., 1987; Makałowski, unpublished observation). An interesting case of evolution of 3′ UTR of the mice muscle γ-phosphorylase gene is schematically presented in Fig. 4.

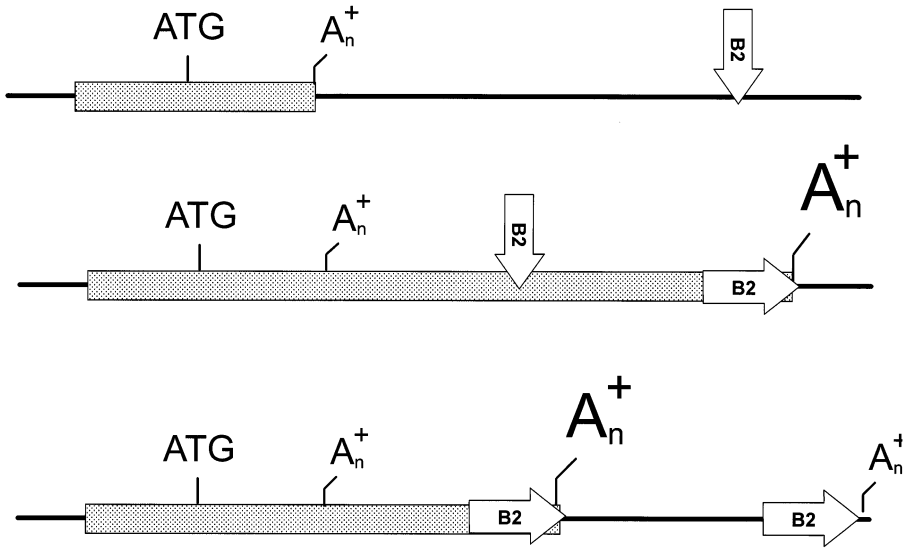

Fig. 4. The evolution of the mouse muscle γ-phosphorylase kinase 3′ UTR. The exons are represented by hatched boxes. Horizontal arrows represent B2 elements. Vertical arrows point to the insertion sites. Polyadenylation sites are represented by A+. The stop codons are marked by TGA triplets.

The examples presented above show the great potential of transposable elements to modify the 3′ end ofthe host mRNA. The current compilation of Brosius (http://www.crosswinds.net/%7Eexpath/references/addmat/add0101.htm) lists over 20 vertebrate genes with poly-A signal originated in retrosequences. Especially interesting are the cases of repeat insertions leading to alternative 3′ UTRs, although the physiological role of such alternative transcripts has yet to be determined.

3.3. Protein-coding sequences

The presence of a transposable element in the open reading frame of a host gene was first noticed in a disease phenotype. Single point mutation in an Alu element residing in the third intron of ornithine aminotransferase activated cryptic splicing sites, and consequently led to the introduction of a partial Alu element into an open reading frame (Mitchell et al., 1991). The in-frame STOP codon carried by an Alu cassette caused a truncated protein, and ornithine δ-aminotransferase deficiency was observed. This discovery led to the hypothesis that a similar mechanism is used for fast evolutionary changes in protein structure, leading to increased protein variability (Makałowski et al., 1994).

A recent survey of all vertebrate protein coding sequences (Makałowski, unpublished data) showed that mobile elements from all categories contribute to protein variability, but the primate Alu element seems to be predisposed for this role because of its abundance in the primate genome, the several cryptic splicing sites embedded into the element, and the 'Alu cassettes', uninterrupted by STOP codons, that can be created by those cryptic splicing sites (see Fig. 5) (Makałowski et al., 1994). Different mechanisms of mobile element insertion into the open reading frame of a host gene were discussed in detail previously (Makałowski et al., 1994; Makałowski, 1995). Here, I would like to briefly describe my favorite example of protein variability created by activation of cryptic splicing sites in intronic Alu. About 10% of the human decay accelerating factor (DAF) mRNA contains the Alu cassette (Caras et al., 1987). DAF is a cell membrane glycoprotein that binds the activated complement. The introduction of the Alu cassette into DAF mRNA creates a hydrophilic carboxy-terminal region in the peptide, which would inhibit migration of DAF into a cell membrane. Caras et al. observed that DAF translated from a wild-type message was membrane-bound, while the DAF peptide expressed from Alu-containing mRNA was not. They concluded that a fraction of the Alu-containing mRNA in a normal cell accounts for the soluble form of DAF (see Fig. 6).

Another interesting example of protein evolution influenced by SINE elements comes from Okada's group (Shimamura et al., 1998). They found an example of CHR-1 SINE that was inserted into the coding region of mRNA for the EP3 subtype of bovine prostaglandin E_2 receptor. Two out of four alternatively spliced EP3 messages include CHR-1 as a part of the coding sequence. Different carboxy-termini of this receptor, which are produced by alternative splicing, are responsible for modulation of the coupling with different G proteins that lead to the activation of different signaling pathways (Namba et al., 1993). In this case a transposable element is a source of protein domain that modifies

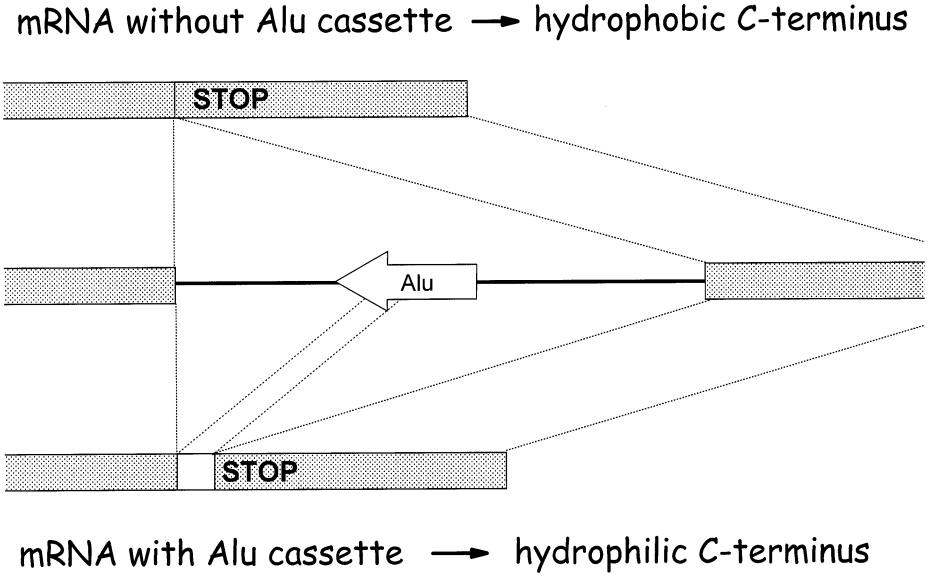

Fig. 5. Potential splicing sites in the Alu consensus sequence. Top vertical bars indicate the potential splicing sites in an Alu in the sense orientation (here, I define the orientation of an Alu element as 'sense' if the polyadenyl tail is downstream with respect to the direction of transcription of the host gene, and 'antisense' if it is in the opposite orientation). The bottom vertical bars indicate splicing sites in an Alu in the antisense orientation.

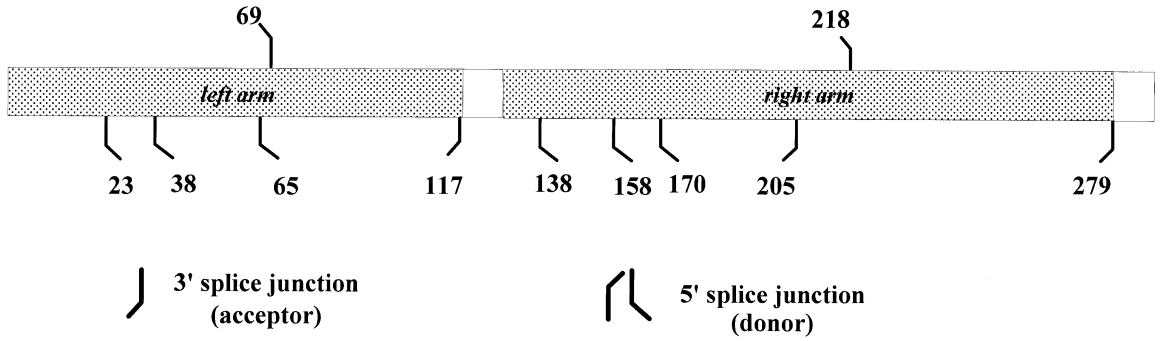

Fig. 6. Alternative splicing in DAF mRNA. Only 3′ fragments of the gene and messages are shown. Open boxes in the genomic sequence represent the exons and a solid line represents the intron. The Alu is represented by a shaded arrow. In mRNA schemes boxes represent ORFs and a solid line 3′ UTR. A shaded square represents the fragment of ORF that originated within the Alu sequence.

the specificity of the receptor and increases its physiological flexibility.

As mentioned above, exhaustive scanning of vertebrate protein coding regions showed that all types of mobile elements contribute to protein variability. As of May 2000, the list of genes with mobile elements contributing to protein coding sequences consists of over 200 records. This list is available as an electronic appendix to this article at: http:/www.ncbi.nlm.nih.gov/Makalowski/ScrapYard.

3.4. Retrogenes

All eukaryotic genomes are populated by a number of pseudogenes. For example, the gene catalogue of the whole human chromosome 21 consists of 225 records, among which 59 are pseudogenes. While some of pseudogenes resemble a structure of an active gene (i.e. the same number of exons and introns), others resemble rather mature (spliced) mRNA. The latter are often flanked by short direct repeats, a hallmark of retroposition. Inactivation (a molecular death) is not always a fate of a retrogene. If integration occurs downstream of an active promoter, the retrogene may escape extinction by utilizing a nearby resident promoter. This seemed to be the mechanism by which a *Drosophila jingwei* (*jgw*) gene was born (Long and Langley, 1993). The *jgw* gene only exists in two sibling species, *Drosophila yakuba* and *Drosophila teissieri*, that diverged only 2.5 million years ago. The gene consists of four exons, three relatively short (less than 100 nt) and one very long. Long and coworkers (Long and Langley, 1993) showed that the last, long exon is a retroposed sequence originated in alcohol dehydrogenase gene (see Fig. 7). Three 5′ exons were recruited after retroposition from another unrelated gene. The donor, *yellow-emperor* (*ymp*) gene, is widely distributed among different *Drosophila* species (Long et al., 1999). Interestingly, both *jgw* and *ymp* genes are transcriptionally active, though their specific function is still not clear. The story of the *jingwei* gene suggests that many eukaryotic genes with unusually long exons may be created in this way. Indeed, the current work of Long's group in different *Drosophila* species shows that it might be quite a common mechanism of creating new genes. It also suggests a mechanism of exon shuffling, a very important and powerful way of creating protein variability.

Retrogenes can also be rescued by other means. If the integration of a retrogene is followed by retroposition of other elements, especially those carrying a regulatory signal, such a gene may escape genomic oblivion. The evidence that this mechanism is used comes from the evolution of the mammalian α-globin locus. The Θ-globin sequence in the α-globin cluster is transcribed in the catherini including baboons, orangutans, and humans, but not in the prosimians and rabbit, where it is present as a pseudogene. It appears that the Θ-globin sequence was rescued by an Alu element which inserted

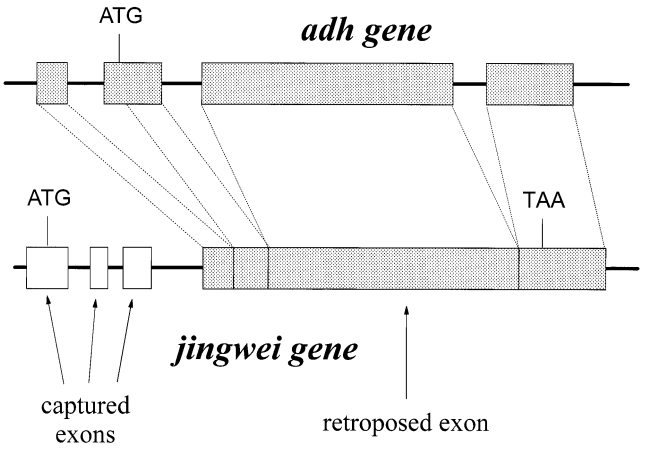

Fig. 7. Origin of the *Drosophila jingwei* gene. The structures of the alcohol dehydrogenase and jingwei genes are shown. Boxes represent the exons. The shaded boxes of the adh gene represent parts of the gene which were retroposed and gave rise to the *jingwei* exon 4. Shaded boxes in the *jingwei* gene represent ORF and open boxes untranslated regions of the jingwei mRNA. Captured exons were donated by duplicated *yellow-emperor* gene.

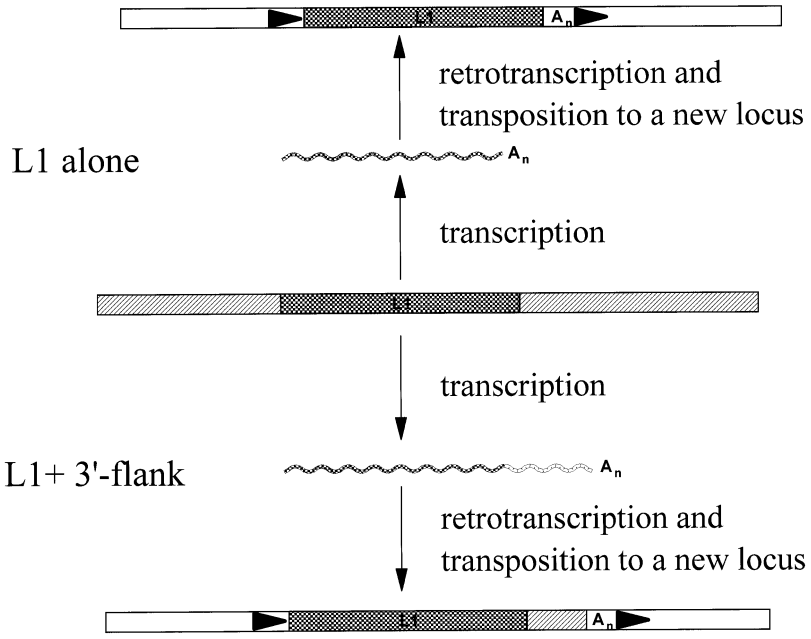

Fig. 8. L1 retroposition without (top) and with (bottom) 3′ transduction. In some cases (bottom part of the figure), an additional genomic sequence is incorporated into the L1 message and consequently it can be moved into a new genomic loci. L1 is represented by check-board boxes, A_n represents the polyadenylation tail, and black triangles represent target site duplications.

just upstream of this pseudogene (Kim et al., 1989). The CCAAT motif carried by the Alu, along with a TATA sequence existing downstream of the retroposition, created a fully functional promoter and enabled transcription of the Θ-globin sequence. This example presents not only a beneficial potential existing in transposable elements, but also a mechanism of bringing back to life a pseudogene.

4. Genome shuffling

Genomes are dynamic entities, shaped by different evolutionary forces. In the previous section we discussed how new genes can be assembled from chunks of existing ones. Retroposition of mRNA sequences has some limits though, it enables shuffling of coding (and UTR) sequences only. As discussed above, some transposable elements can be a source of regulatory elements which can be moved around a genome. Recent studies of human L1 (LINE-1) element suggest a new mechanism of genome shuffling. L1, as a retrotransposon, replicates within mammalian genome using reverse transcriptase, which copies the retrotransposon RNA into DNA. L1 usually moves only its own sequence from one genomic location to another (see Fig. 8). But during studies of de novo pathological insertions of human L1 elements, it has been noticed that in some cases additional (non-L1) sequences were incorporated downstream of the L1 element (Fig. 8) (Miki et al., 1992; Holmes et al., 1994; McNaughton et al., 1997). These observations were followed up and confirmed by both in vivo and in silico studies (Moran et al., 1999; Pickeral et al., 2000). Both studies showed that the co-mobilization (or 3′ transduction) process is very efficient and can move up to several kilobases of non-L1 DNA to a new genomic location. Pickeral et al. estimated that about 1% of the human genome could be shuffled by L1-driven transduction (Pickeral et al., 2000).

5. Conclusions

The examples presented above indicate that transposable elements are not useless DNA. They interact with the surrounding genomic environment and increase host evolvability by serving as:

1. recombination hot spots
2. a source of ‘ready-to-use’ motifs
   (a) transcriptional regulatory elements
   (b) polyadenylation signals
   (c) protein coding sequences
3. a mechanism for genomic shuffling.

All these examples are probably just the tip of the iceberg, many more are waiting to be discovered, and even more will never be discovered because their original donors (tranposable elements) mutated beyond recognition. The genomes are dynamic entities. New functional elements appear and old ones become extinct. The information reviewed above suggests that transposable elements are the major evolutionary force in shaping eukaryotic genomes. As such, they should not be viewed

as genomic parasites, but rather as genomic symbionts that create a *genomic scrap yard*, the source of 'junk' that natural selection utilizes in its evolutionary experiments.

Acknowledgements

I would like to thank Juergen Brosius for many fruitful discussions on the subject. Tables listing examples of recruited (exapted) transposable elements are available at: http://www.ncbi.nlm.nih.gov/Makalowski/ScrapYard/ or http://www.crosswinds.net/%7Eexpath/references/addmat/add0101.htm.

References

Boggaram, V., Qing, K., Mendelson, C.R., 1988. The major apoprotein of rabbit pulmonary surfactant. Elucidation of primary sequence and cyclic AMP and developmental regulation. J. Biol. Chem. 263, 2939–2947.

Brosius, J., 1991. Retroposons — seeds of evolution. Science 251, 753

Brosius, J., 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238, 115–134.

Caras, I.W., Davitz, M.A., Rhee, L., Weddell, G., Martin Jr., D.W., Nussenzweig, V., 1987. Cloning of decay-accelerating factor suggests novel use of splicing to generate two proteins. Nature 325, 545–549.

Deininger, P.L., Batzer, M.A., 1999. Alu repeats and human disease. Mol. Genet. Metab. 67, 183–193.

Doolittle, W.F., Sapienza, C., 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284, 601–603.

Harendza, C.J., Johnson, L.F., 1990. Polyadenylylation signal of the mouse thymidylate synthase gene was created by insertion of an L1 repetitive element downstream of the open reading frame. Proc. Natl. Acad. Sci. USA 87, 2531–2535.

Hickey, D.A., 1982. Selfish DNA: a sexually-transmitted nuclear parasite. Genetics 101, 519–531.

Holmes, S.E., Dombroski, B.A., Krebs, C.M., Boehm, C.D., Kazazian Jr., H.H., 1994. A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. Nat. Genet. 7, 143–148.

Kazazian Jr., H.H., 1998. Mobile elements and disease. Curr. Opin. Genet. Dev. 8, 343–350.

Kim, J.H., Yu, C.Y., Bailey, A., Hardison, R., Shen, C.K., 1989. Unique sequence organization and erythroid cell-specific nuclear factor-binding of mammalian theta 1 globin promoters. Nucleic Acids Res. 17, 5687–5700.

Krane, D.E., Hardison, R.C., 1990. Short interspersed repeats in rabbit DNA can provide functional polyadenylation signals. Mol. Biol. Evol. 7, 1–8.

Levinson, G., 1994. Crossovers generate random recombinants under Darwinian selection. In: Maes, R. (Ed.), Artificial Life IV. MIT Press, Cambridge, MA, pp. 90–101.

Long, M., Langley, C.H., 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Science 260, 91–95.

Long, M., Wang, W., Zhang, J., 1999. Origin of new genes and source for N-terminal domain of the chimerical gene, *jingwei*, in *Drosophila*. Gene 238, 135–141.

Makałowski, W., 1995. SINEs as a genomic scrap yard: an essay on genomic evolution. In: Maraia, R.J. (Ed.), The Impact of Short Interspersed Elements (SINEs) on the Hpst Genome. R.G. Landes, Austin, TX, pp. 81–104.

Makałowski, W., Mitchell, G.A., Labuda, D., 1994. Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet. 10, 188–193.

McNaughton, J.C., Hughes, G., Jones, W.A., Stockwell, P.A., Klamut, H.J., Petersen, G.B., 1997. The evolution of an intron: analysis of a long, deletion-prone intron in the human dystrophin gene. Genomics 40, 294–304.

Miki, Y., Nishisho, I., Horii, A., Miyoshi, Y., Utsunomiya, J., Kinzler, K.W., Vogelstein, B., Nakamura, Y., 1992. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Cancer Res. 52, 643–645.

Mitchell, G.A., Labuda, D., Fontaine, G., Saudubray, J.M., Bonnefont, J.P., Lyonnet, S., Brody, L.C., Steel, G., Obie, C., Valle, D., 1991. Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation. Proc. Natl. Acad. Sci. USA 88, 815–819.

Moran, J.V., DeBerardinis, R.J., Kazazian Jr., H.H., 1999. Exon shuffling by L1 retrotransposition. Science 283, 1530–1534.

Namba, T., Sugimoto, Y., Negishi, M., Irie, A., Ushikubi, F., Kakizuka, A., Ito, S., Ichikawa, A., Narumiya, S., 1993. Alternative splicing of C-terminal tail of prostaglandin E receptor subtype EP3 determines G-protein specificity. Nature 365, 166–170.

Nei, M., 1969. Gene duplication and nucleotide substitution in evolution. Nature 221, 40–42.

Nowak, R., 1994. Mining treasures from 'junk DNA'. Science 263, 608–610.

Ohno, S., 1972. So much 'junk' DNA in our genome. In: Smith, H.H. (Ed.), Brookhaven Symposia in Biology No. 23. Gordon and Breach, New York, pp. 366–370.

Orgel, L.E., Crick, F.H., 1980. Selfish DNA: the ultimate parasite. Nature 284, 604–607.

Paulson, K.E., Matera, A.G., Deka, N., Schmid, C.W., 1987. Transcription of a human transposon-like sequence is usually directed by other promoters. Nucleic Acids Res. 15, 5199–5215.

Pickeral, O.K., Makałowski, W., Boguski, M.S., Boeke, J.D., 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome Res. 10, 411–415.

Reiter, L.T., Murakami, T., Koeuth, T., Pentao, L., Muzny, D.M., Gibbs, R.A., Lupski, J.R., 1996. A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. Nat. Genet. 12, 288–297.

Shimamura, M., Nikaido, M., Ohshima, K., Okada, N., 1998. A SINE that acquired a role in signal transduction during evolution. Mol. Biol. Evol. 15, 923–925.

Sverdlov, E.D., 1998. Perpetually mobile footprints of ancient infections in human genome. FEBS Lett. 428, 1–6.