Kazazian *Mobile DNA* 2014, **5**:20
http://www.mobilednajournal.com/content/5/1/20

Mobile DNA

COMMENTARY Open Access

Processed pseudogene insertions in somatic cells

Haig H Kazazian Jr

Abstract

Processed pseudogenes are copies of messenger RNAs that have been reverse transcribed into DNA and inserted into the genome using the enzymatic activities of active L1 elements. Processed pseudogenes generally lack introns, end in a 3' poly A, and are flanked by target site duplications. Until recently, very few polymorphic processed pseudogenes had been discovered in mammalian genomes. Now several studies have found a number of polymorphic processed pseudogenes in humans. Moreover, processed pseudogenes can occur in somatic cells, including in various cancers and in early fetal development. One recent somatic insertion of a processed pseudogene has caused a Mendelian X-linked disease, chronic granulomatous disease.

Keywords: Processed pseudogenes, L1 retrotransposons, Polymorphism, Cancer, Chronic granulomatous disease

Background

Pseudogenes are sequences present in essentially all animal genomes that have many characteristics of genes, but are defective for production of protein. Of course, like most definitions that are 30 years old and based on incomplete information, this one has also been modified. We now know of many pseudogenes that are active in making proteins. Of the more than 14,000 pseudogenes in the human genome [1], at least 10% are no longer 'pseudogenes' and are active [1,2]. Many active 'pseudogenes' are gene duplicates that contain introns and are situated in close proximity to their active gene copies. These gene duplicates make up one class of pseudogenes. An interesting example of a duplicate pseudogene is the φζ gene in the α-globin gene cluster [3]. This pseudogene has only six nucleotide differences from its parent ζ (zeta) gene, and one of these differences leads to a nonsense codon. In eight populations studied, the nonsense codon is corrected by gene conversion in 15% to 50% of α-globin gene clusters. However, RNA emanating from the corrected φζ gene could not be detected [3].

Although there are many duplicate pseudogenes in the human genome, the majority of human pseudogenes, more than 7,800 [1], belong to the second class, and are called processed pseudogenes (PPs). The term processed pseudogene was first proposed in 1977 to describe a sequence of a 5S gene of *Xenopus laevis* [4]. PPs are found in the genomes of many animal species [2] and have the following characteristics: 1) their sequences are very similar to the transcribed portion of the parent gene; 2) they lack all or most introns, so they appear to be cDNA copies of processed mRNAs; 3) they have a poly A tail attached to the 3'-most transcribed nucleotide; and 4) they are flanked at their 5' and 3' ends by target site duplications (TSDs) of 5 to 20 nucleotides. The cDNA copies of mRNAs, the source of PPs, are inserted in far-flung regions of the genome [5]. At least 10% of PPs retain activity because when dispersed they have fortuitously landed close to an RNA polymerase II promoter [2]. We have known for ten years that the sequence characteristics of PPs are signs of mobilization by the endonuclease and reverse transcriptase activities of active LINE-1 (L1) elements [6,7]. In human cells, L1s have been shown to mobilize SINEs such as Alus [8,9], SVAs [10,11], and small nuclear (sn) RNAs [12], along with many mRNA transcripts. In mouse cells, L1s also mobilize B1 and B2 SINE elements [13]. More than 2,075 human genes are represented by at least one PP in the genome, while some genes, such as *GAPDH*, ribosomal proteins and actin β have 50 to 100 PPs [14]. Why 10% of human genes are represented by PPs, while the remaining 90% are not, is an important unanswered question.

A number of quite interesting PPs have been identified. In one example, the phosphoglycerate kinase gene, *pgk2*, is an active testis-expressed PP derived from the

Correspondence: kazazian@jhmi.edu
Institute for Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

X-linked *pgk1* gene [15]. Deficiency of pgk2 leads to severe reduction in male fertility [16]. Another example is the fgf4 (fibroblast growth factor 4) PP in a number of dog breeds. This activated fgf4 PP is responsible for a chondrodysplasia that leads to the short-legged phenotype of 19 dog breeds, including dachsund, basset hound and corgi [17]. A third example is the *CypA* pseudogene that has inserted into the *TRIM5* gene at least twice, once in the owl monkey [18] and another time in the macaque lineage [19,20]. The *TRIM-Cyp* fusion gene leads to HIV-1 resistance of the monkeys because the TRIM-Cyp fusion protein blocks entry of the virus into cells [18].

There is another class of PPs termed semi-processed pseudogenes, which retain some introns and are particularly prevalent in the mouse and rat. For example, in the mouse the preproinsulin II gene has two introns, while the preproinsulin I gene is a PP that retains one of the two introns [21]. However, until very recently the prevailing view has been that there is very little ongoing PP formation in mammals. Now we know that that view is wrong. There is significant PP formation in present day human beings.

Recent processed pseudogene insertions

About one year ago, a comprehensive paper on polymorphism among PPs in human beings appeared. Ewing *et al*. devised a bioinformatic pipeline to detect polymorphic PPs. Using discordant reads not present in reference genomes, they found 48 novel PP insertion sites among 939 low pass genomes from the 1,000 genomes project [22]. These PPs came from a wide variety of source genes, and were spread throughout the human chromosomes (Figure 1). All 48 of these polymorphic PPs were confirmed by locating the precise genomic insertion site. This group also studied the genome sequences of 85 human cancer-normal tissue pairs representing a variety of cancers. Among these cancers they found the first instances of somatic insertion of PPs; three PPs were predicted to occur in lung cancers

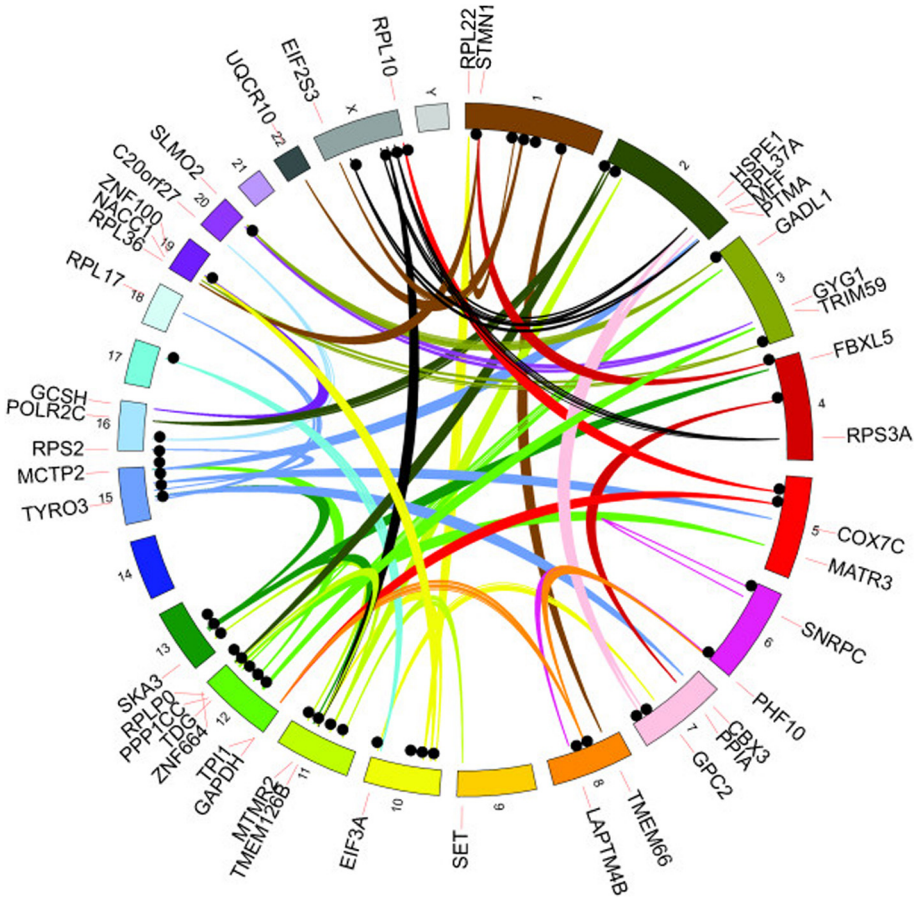

Figure 1 Locations of 48 non-reference gene processed pseudogene insertions sites in the human genome based on reads mapped to source genes. Discordant read mappings are represented by links colored based on chromosome of the source gene. Insertion sites are represented by black circles and the gene labels are based on the position of the source gene. Republished with permission from *Nature Communications*.

that were absent from paired normal tissue. The authors also estimated the rate of PP insertion in human beings at one insertion in every approximately 5,200 individuals/generation [22].

Ewing *et al*. went on to study PP polymorphism among mice, finding 755 new polymorphic PPs with most PPs occurring in species and subspecies derived from wild mice. Among these, *Mus musculus castaneus*, *M.m. musculus*, and *M.m. spretus* had 213, 212 and 142 PPs in their genomes, respectively, that were not found in the inbred C57Bl6 genome. However, on average, each of the 12 inbred strains derived from C57Bl6 were genetically closer, but still differed from one another by 68 PPs on average. The much greater number of polymorphic PPs in mouse strains compared to individual human beings may be due to the much larger number of active L1s present in the mouse (approximately 3,000 versus approximately 100 in humans) [23,24]. Ewing *et al*. also studied the genome sequences of ten chimpanzees and found ten polymorphic PPs among these animals. This paper represented the first comprehensive look at the question of PP insertions in humans, mice and chimpanzees, and the first study of somatic insertion of PPs in cancer.

Two other papers demonstrating polymorphism of PPs in humans have now appeared. Using exon-exon junction spanning reads, Abyzov *et al*. found 147 novel putative processed pseudogenes among approximately 1,000 low–pass genome sequences [25]. Thirty-six of these 147 were confirmed as polymorphic in humans by detection of the genomic insertion point. Interestingly, the parental genes of non-reference PPs were significantly enriched among genes expressed at the M-to-G1 transition in the cell cycle. Schrider *et al*. also mapped processed pseudogenes among 17 individuals, mostly using exon-exon junction spanning reads from SOLID and 1,000 genomes data [26]. They found 21 PPs not present in the reference genome and presumably polymorphic; 17 of these 21 were confirmed by PCR (See [27] for a recent review of these papers).

Recently, Cooke *et al*. studied somatic PP insertion in cancer in greater detail [28]. They analyzed 660 cancer-normal pairs of sequenced samples at Wellcome Trust representing a variety of different cancers. In 17 or 2.5% of the cancers, they found 42 somatic PPs. The authors noted the presence of five PPs in non-small cell lung cancer among 27 cancers studied, similar to the Ewing *et al*. finding of somatic PPs in lung cancer. Additionally, they found two PPs in eleven colorectal cancer samples.

The PP insertions in cancer were thoroughly characterized and all had the molecular signatures of germ line L1 insertions. The majority had TSDs of 5 to 20 base pairs, 74% were 5' truncated (a percentage similar to that of human-specific L1s), 20% had inversions at their 5' ends due to 'twin priming' (again similar to the rate in germ line human L1 insertions) [29], and long poly A tracts. In a lung adenocarcinoma, one insertion was associated with an 8 kb deletion of the promoter and exon 1 of a tumor suppressor gene, *MGA1*. The deletion knocked out expression of that allele as determined by RNA-seq.

Among the PPs in cancer, most were derived from highly expressed transcripts, yet many were not. In addition, many PP insertions appeared to be early events in tumor formation, being present in an early lesion along with the tumor or in multiple sections of the same tumor. However, some PP insertions were shown to be later events in tumor progression because they were not detected in all sections of the same tumor.

A final paper nailed down the potential for PP formation during early development in humans. This paper by de Boer *et al*. described a case of the X-linked disorder, chronic granulomatous disease in a Dutch man [30]. This man, now a young adult, had suffered from multiple bouts of pulmonary aspergillosis as a child. On workup of his *CYBB* (cytochrome b-245, beta polypeptide) gene, the defective gene in the disorder and parenthetically the first human gene cloned by positional cloning [31], it was discovered that a PP insertion had knocked out the gene's activity.

There are three interesting aspects of this case. First, the insertion was a semi-processed pseudogene of the *TMF1* (TATA element modulatory factor) gene from chromosome 3 that had inserted into intron 1 of CYBB in reverse orientation. A PP had not been observed previously as a new insertion among 100 previous insertions (L1, Alu, SVA) in human Mendelian disease or cancer etiology [32]. Interestingly, *TMF1* is one of the about 10% of human genes that is represented by a single PP in the human reference genome sequence [14]. Second, the insertion was 3' truncated and contained exons 1 to 8 of *TMF1* along with intron 7 and much of intron 8. Transcription of *TMF1* had terminated after an alternative poly A signal, AGUAAA, in intron 8, and a 100 bp poly A tail was added to the transcript. After insertion of this semi-processed pseudogene in reverse orientation into intron 1 of *CYBB*, splicing had occurred into an excellent acceptor splice site and out of an excellent donor site in exon 2 of TMF1. The newly created 117 bp exon also contained a nonsense codon that caused the *CYBB* gene to be non-functional (Figure 2). Finally, the PP insertion had occurred during early embryonic development of the patient's mother. Roughly 10% to 20% of her lymphocytes contained the insertion as shown by qPCR.

To date, somatic retrotransposition in Mendelian disease has been rarely found. Among the 100 cases mentioned above, there is only a somatic insertion into the adenomatous polyposis coli (APC) tumor suppressor

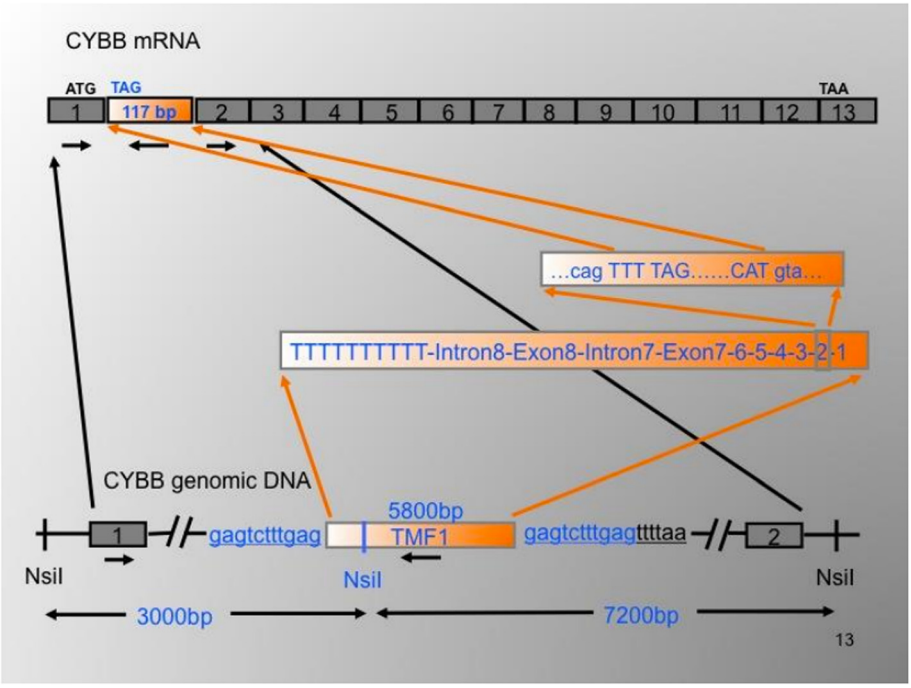

Figure 2 Orientation of the TMF1 insertion in intron 1 of the *CyBB* gene (below), leading to an extra exon between exons 1 and 2 in the CYBB mRNA (above). Republished with permission from *Human Mutation* published by Wiley.

gene in a colorectal cancer case [33] and somatic and germ line mosaicism in the mother of a patient with the X-linked disease, choroideremia [34]. Thus, after more than 20 years since the discovery of the first retrotransposition events due to L1 and Alu elements [35,36], we finally have definitive evidence of retrotransposition of processed pseudogenes in human somatic cells (cancer and early development).

These papers beg the question, why do PP insertions not occur more frequently? Another recent paper has provided evidence that the RNAs associated with the L1 ORF1 protein in the L1 ribonucleoprotein particle (L1 RNP) contain a preponderance of those mRNAs that form PPs [37]. These mRNAs also have a much greater capacity for reverse transcription by L1 ORF2 protein than mRNAs that do not form PPs [37,38]. Now that we know that PP formation can occur in somatic cells, it is logical that those mRNAs that are both located in L1 RNPs and capable of reverse transcription have the inside track in PP formation. Messenger RNAs that lack what it takes to associate with the L1 RNP and be reverse transcribed, perhaps due to deficient cellular concentration or their sequence characteristics, are unable to form PPs. However, the story is not quite so simple since the majority of mRNAs that have formed PPs in the human genome do not appear to be associated with the L1 RNP. Thus, the demonstration of somatic PP insertions leads to a new as yet unanswered question: What are the important factors that increase the likelihood that a particular mRNA will become a processed pseudogene?

Conclusions

Although perhaps unexpected, the evidence is overwhelming that PPs continue to insert in the germ line and in somatic cells of human beings.

Abbreviations

PP: processed pseudogene; L1: LINE1-long interspersed element; RNP: ribonucleoprotein particle..

Competing interests

The author declares that he has no competing interests.

Authors' contributions

HHK conceived and wrote the manuscript.

Acknowledgements

The author thanks John Goodier, Adam Ewing, Szilvia Solyom and Tara Doucet for critical comments on the manuscript. The author is supported by an RO1 grant from NIH and a P50 grant from the NIH.

Received: 30 April 2014 Accepted: 20 May 2014
Published: 2 July 2014

References

1. Zhang Z, Harrison PM, Liu Y, Gerstein M: **Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome.** *Genome Res* 2003, **13:**2541–2558.
2. Kabza M, Ciomborowska J, Makałowska I: **RetrogeneDB- a database of animal retrogenes.** *Mol Biol Evol*. in press.

3. Hill AV, Nicholls RD, Thein SL, Higgs DH: **Recombination within the human embryonic ζ-globin locus: a common ζ – ζ chromosome produced by gene conversion of the φζ gene.** *Cell* 1985, **42:**809–819.
4. Jacq C, Miller JR, Brownlee GG: **A pseudogene structure in 5S DNA of Xenopus laevis.** *Cell* 1977, **12:**109–120.
5. Vanin EF: **Processed pseudogenes: characteristics and evolution.** *Annu Rev Genet* 1985, **19:**253–272.
6. Esnault C, Maestre J, Heidmann T: **Human LINE retrotransposons generate processed pseudogenes.** *Nat Genet* 2000, **24:**363–367.
7. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV: **Human L1 retrotransposition: cis preference versus trans-complementation.** *Mol Cell Biol* 2001, **21:**1429–1439.
8. Kajikawa M, Okada N: **LINEs mobilize SINEs in the eel through a shared 3′ sequence.** *Cell* 2002, **111:**433–444.
9. Dewannieux M, Esnault C, Heidmann T: **LINE-mediated retrotransposition of marked Alu sequences.** *Nat Genet* 2003, **35:**41–48.
10. Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH Jr: **Retrotransposition of marked SVA elements by human L1s in cultured cells.** *Hum Mol Genet* 2011, **20:**3386–3400.
11. Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, Löwer J, Strätling WH, Löwer R, Schumann GG: **The non-autonomous retrotransnposon SVA is trans-mobilized by the human LINE-1 protein machinery.** *Nucleic Acids Res* 2012, **40:**1666–1683.
12. Gilbert N, Lutz S, Morrish TA, Moran JV: **Multiple fates of L1 retrotransposition intermediates in cultured human cells.** *Mol Cell Biol* 2005, **25:**7780–7795.
13. Dewannieux M, Heidmann T: **L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells.** *J Mol Biol* 2005, **349:**241–247.
14. Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, Reymond A, Hubbard TJ, Harrow J, Gerstein MB: **The GENCODE pseudogene resource.** *Genome Biol* 2012, **13:**R51.
15. McCarrey JR, Thomas K: **Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene.** *Nature* 1987, **326:**501–505.
16. Danshina PV, Geyer CB, Dai Q, Goulding EH, Willis WD, Kitto GB, McCarrey JR, Eddy EM, O'Brien DA: **Phosphoglycerate kinase 2 (PGK2) is essential for sperm function and male fertility in mice.** *Biol Reprod* 2010, **82:**136–145.
17. Parker HG, VonHoldt BM, Quignon P, Margulies EH, Shao S, Mosher DS, Spady TC, Elkahloun A, Cargill M, Jones PG, Maslen CL, Acland GM, Sutter NB, Kuroki K, Bustamante CD, Wayne RK, Ostrander EA: **An expressed *fgf4* retrogene is associated with breed-defining chondrodysplasia in domestic dogs.** *Science* 2009, **325:**995–998.
18. Sayah DM, Sokolskaja E, Berthoux L, Luban J: **Cyclophilin A retrotransposition into *TRIM5* explains owl monkey resistance to HIV-1.** *Nature* 2004, **430:**569–573.
19. Virgen CA, Kratovac Z, Bieniasz PD, Hatziioannou T: **Independent genesis of chimeric TRIM5-cyclophilin proteins in two primate species.** *Proc Natl Acad Sci U S A* 2008, **105:**3563–3568.
20. Wilson SJ, Webb BL, Ylinen LM, Verschoor E, Heeney JL, Towers GJ: **Independent evolution of an antiviral TRIMCyp in rhesus macaques.** *Proc Natl Acad Sci U S A* 2008, **105:**3557–3562.
21. Soares MB, Schon E, Henderson A, Karathanasis SK, Cate R, Zeitlin S, Chirgwin J, Efstratiadis A: **RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon.** *Mol Cell Biol* 1985, **5:**2090–2103.
22. Ewing AD, Ballinger TJ, Earl D, Broad Institute Genome Sequencing and Analysis Program and Platform, Harris CC, Ding L, Wilson RK, Haussler D: **Retrotransposition of gene transcripts leads to structural variation in mammalian genomes.** *Genome Biol* 2013, **14:**R22.
23. Goodier JL, Ostertag EM, Du K, Kazazian HH Jr: **Characterization of a novel active L1 retrotransposon subfamily in the mouse.** *Genome Res* 2001, **11:**1677–1685.
24. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH Jr: **Hot L1s account for the bulk of retrotransposition in the human population.** *Proc Natl Acad Sci U S A* 2003, **100:**5280–5285.
25. Abyzov A, Iskow R, Gokcumen O, Radke DW, 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.
26. Schrider DR, Navarro FC, Galante PA, Parmigiani RB, Camargo AA, Hahn MW, de Souza SJ: **Gene copy-number polymorphism caused by retrotransposition in humans.** *PLoS Genet* 2013, **9:**e1003242.
27. Richardson SR, Salvador-Palomeque C, Faulkner GJ: **Diversity through duplication: whole-genome sequencing reveals novel gene retrocopies in the human population.** *Bioessays* 2014, **36:**475–481.
28. Cooke SL, Shlien A, Marshall J, Pipinikas CP, Martincorena I, Tubio JM, Li Y, Menzies A, Mudie L, Ramakrishna M, Yates L, Davies H, Bolli N, Bignell GR, Tarpey PS, Behjati S, Nik-Zainal S, Papaemmanuil E, Teixeira VH, Raine K, O'Meara S, Dodoran MS, Teague JW, Butler AP, Iacobuzio-Donahue C, Santarius T, Grundy RG, Malkin D, Greaves M, Munshi N, *et al*: **Processed pseudogenes acquired somatically during cancer development.** *Nat Commun* 2014, **5:**3644.
29. Ostertag EM, Kazazian HH Jr: **Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition.** *Genome Res* 2001, **11:**2059–2065.
30. de Boer M, van Leeuwen K, Geissler J, Weemaes CM, van den Berg TK, Kuijpers TW, Warris A, Roos D: **Primary immunodeficiency caused by an exonized retroposed gene copy inserted in the *CYBB* gene.** *Hum Mutat* 2014, **35:**486–496.
31. Royer-Pokora B, Kunkel LM, Monaco AP, Goff SC, Newburger PE, Baehner RL, Cole FS, Curnutte JT, Orkin SH: **Cloning the gene for an inherited human disorder–chronic granulomatous disease–on the basis of its chromosomal location.** *Nature* 1986, **322:**32–38.
32. Hancks DC, Kazazian HH Jr: **Active human retrotransposons: variation and disease.** *Curr Opin Genet Dev* 2012, **22:**191–203.
33. Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, Vogelstein B, Nakamura Y: **Disruption of the *APC* gene by a retrotransposal insertion of L1 sequence in a colon cancer.** *Cancer Res* 1992, **52:**643–645.
34. van den Hurk JA, Meij IC, Seleme MC, Kano H, Nikopoulos K, Hoefsloot LH, Sistermans EA, de Wijs IJ, Mukhopadhyay A, Plomp AS, de Jong PT, Kazazian HH, Cremers FP: **L1 retrotransposition can occur early in human embryonic development.** *Hum Mol Genet* 2007, **16:**1587–1592.
35. Kazazian HH Jr, Wong C, Youssoufian H, Scott AF, Phillips D, Antonarakis SE: **A novel mechanism of mutation in man: haemophilia A due to de novo insertion of L1 sequences.** *Nature* 1988, **332:**164–166.
36. Wallace MR, Andersen LB, Saulino AM, Gregory PE, Glover TW, Collins FS: **A de novo Alu insertion results in neurofibromatosis type 1.** *Nature* 1991, **353:**864–866.
37. Mandal PK, Ewing AD, Hancks DC, Kazazian HH Jr: **Enrichment of processed pseudogene transcripts in L1-ribonucleoprotein particles.** *Hum Mol Genet* 2013, **22:**3730–3748.
38. Kulpa DA, Moran JV: **Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles.** *Nat Struct Mol Biol* 2006, **13:**655–660.

doi:10.1186/1759-8753-5-20
Cite this article as: Kazazian: Processed pseudogene insertions in somatic cells. *Mobile DNA* 2014 **5**:20.