Genomic Med. (2007) 1:9–17
DOI 10.1007/s11568-007-9002-9

REVIEW

Mechanism of *Alu* integration into the human genome

Jian-Min Chen · Claude Férec · David N. Cooper

Received: 31 January 2007 / Accepted: 6 March 2007 / Published online: 28 March 2007

Abstract LINE-1 or L1 has driven the generation of at least 10% of the human genome by mobilising *Alu* sequences. Although there is no doubt that *Alu* insertion is initiated by L1-dependent target site-primed reverse transcription, the mechanism by which the newly synthesised 3′ end of a given *Alu* cDNA attaches to the target genomic DNA is less well understood. Intrigued by observations made on 28 pathological simple *Alu* insertions, we have sought to ascertain whether microhomologies could have played a role in the integration of shorter *Alu* sequences into the human genome. A meta-analysis of the 1624 *Alu* insertion polymorphisms deposited in the Database of Retrotransposon Insertion Polymorphisms in Humans (dbRIP), when considered together with a re-evaluation of the mechanism underlying how the three previously annotated large deletion-associated short pathological *Alu* inserts were generated, enabled us to present a unifying model for *Alu* insertion into the human genome. Since *Alu* elements are comparatively short, L1 RT is usually able to complete nascent *Alu* cDNA strand synthesis leading to the generation of full-length *Alu* inserts. However, the synthesis of the nascent *Alu* cDNA strand may be terminated prematurely if its 3′ end anneals to the 3′ terminal of the top strand's 5′ overhang by means of microhomology-mediated mispairing, an event which would often lead to the formation of significantly truncated *Alu* inserts. Furthermore, the nascent *Alu* cDNA strand may be 'hijacked' to patch existing double strand breaks located in the top-strand's upstream regions, leading to the generation of large genomic deletions.

Keywords *Alu* insertion polymorphisms · Human genetic disease · Human genome evolution · L1 · LINE-1 · Retrotransposition

Abbreviations

DbRIP	Database of Retrotransposon Insertion Polymorphisms in humans
LINE-1 or L1	Long interspersed element-1
MMEJ	Microhomology-mediated end-joining
RT	Reverse transcriptase
TPRT	Target site-primed reverse transcription
TSDs	Target site duplications

Introduction

LINE-1 (long interspersed element-1) or L1-mediated retrotransposition has significantly impacted upon human genome evolution (for recent reviews, see Deininger et al.

Electronic supplementary material The online version of this article (doi:10.1007/s11568-007-9002-9) contains supplementary material, which is available to authorized users.

J.-M. Chen · C. Férec
INSERM, U613, 29220 Brest, France

J.-M. Chen (✉) · C. Férec
Etablissement Français du Sang-Bretagne, 46 rue Félix Le Dantec, 29220 Brest, France
e-mail: Jian-Min.Chen@univ-brest.fr

C. Férec
Faculté de Médecine de Brest et des Sciences de la Santé, Université de Bretagne Occidentale, 29238 Brest, France

C. Férec
Laboratoire de Génétique Moléculaire et d'Histocompatibilité, Centre Hospitalier Universitaire (CHU) de Brest, Hôpital Morvan, 29220 Brest, France

D. N. Cooper
Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK

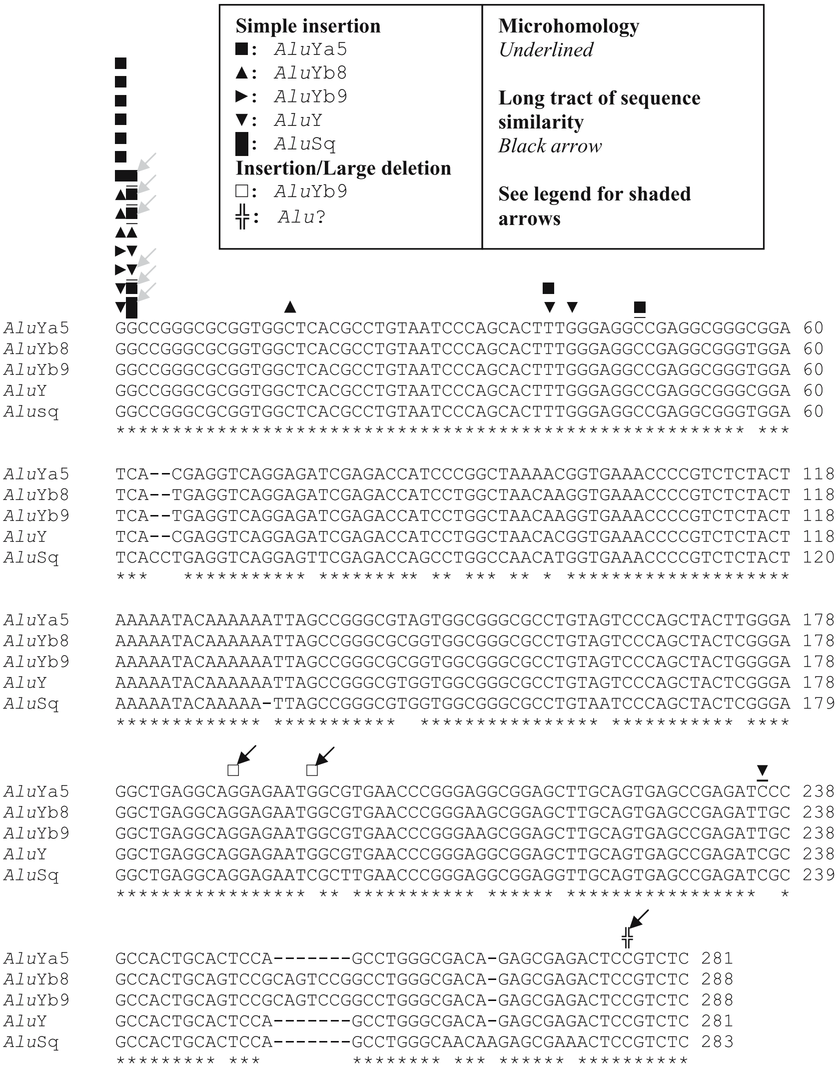
2003; Kazazian 2004; Han and Boeke 2005; Hedges and Batzer 2005) but has also given rise to human genetic disease (Chen et al. 2005, 2006). Intriguingly, L1 elements have driven the generation of some 10% of the human genome mass by mobilising *Alu* sequences (Lander et al. 2001; Batzer and Deininger 2002). Although there is no doubt that *Alu* insertion is initiated by L1 endonuclease and reverse transcriptase (RT)-dependent target site-primed reverse transcription (TPRT; Dewannieux et al. 2003; Hagan et al. 2003), the mechanism by which the newly synthesised 3′ end of a given *Alu* cDNA attaches to the target genomic DNA is less well understood. In this regard, the integration of full-length L1 elements has recently been proposed to occur *via* a template-jumping model whereas the integration of 5′-truncated L1 elements is thought to result predominantly from a microhomology-mediated end-joining (MMEJ) model (Zingler et al. 2005; Babushok et al. 2006). The integration of full-length *Alu* elements can also be explained, at least in principle, by the template-jumping model. However, unlike 5′-truncated L1 elements, 5′-truncated *Alu* elements appear by and large not to be integrated *via* the MMEJ model (Zingler et al. 2005).

Recently, we have identified two pathological simple *Alu* insertions (termed #1 and #2, respectively) in the *CFTR* gene (manuscript submitted). Interestingly, #1 represents the shortest (starting position at 236) of the 28 currently known pathological simple *Alu* insertions (i.e. no loss of target gene sequence) that are informative with respect to the starting position of the *Alu* insert (Fig. 1). More interestingly, of the six 5′-truncated simple *Alu* insertions, #1 represents the only example of the occurrence of a 2 bp microhomology between the 3′ end of the top strand's 5′ overhang in the target sequence and the 3′ end of the nascent *Alu* cDNA (Supplementary Table S1). In addition, the second shortest pathological simple *Alu* insertion (starting position at 47) exhibited a one bp microhomology (Supplementary Table S1). In sharp contrast, none of the remaining four 5′-truncated simple *Alu* insertions (starting positions at 16, 39, 39, and 41, respectively) exhibited microhomology (Fig. 1; Supplementary Table S1). We were intrigued by this phenomenon and wondered whether microhomology could have played a role in the integration of shorter *Alu* sequences into the human genome. To test this idea, we performed a meta-analysis of the *Alu* insertion polymorphisms deposited in the Database of Retrotransposon Insertion Polymorphisms in Humans (dbRIP; http://falcon.roswellpark.org:9090/searchRIP.html; Wang et al. 2006). This analysis, when considered together with a re-evaluation of the mechanism underlying how the three previously annotated large deletion-associated short pathological *Alu* inserts (Chen et al. 2005) were generated, has enabled us to present a unifying model for *Alu* insertion in the human genome.

Identification of microhomology existing between the top strand's 5′ overhang and the sequence that lies 5′ to the truncation position in the *Alu* consensus sequence

The 1624 non-redundant *Alu* insertion polymorphisms deposited in dbRIP (as of December 6, 2006) were subjected to manual evaluation with respect to whether microhomology exists between the top strand's 5′ overhang and the sequence lying 5′ to the truncation position in the *Alu* consensus sequence, in line with previously established principles (e.g. Zingler et al. 2005; Babushok et al. 2006). Where a microhomology (the longest match where applicable) was identified, the top strand cleavage site was assigned as 3′ to the matched nucleotide(s) in the target sequence whilst the starting position of the 5′ truncated *Alu* insert was designated as the nucleotide 3′ to the matched base(s) in the *Alu* consensus sequence. Two examples—one involving a full-length *Alu* insert and the other involving a 5′ truncated *Alu* insert—are illustrated in Fig. 2. In many cases, this treatment yielded a modification of the originally defined end positions of the target site duplications (TSDs) and the start positions of the *Alu* inserts. Although detailed sequence information for each entry is given in Supplementary Tables S2–S6, several issues warrant further clarification here. First, that many of the entries can be alternatively annotated with respect to the microhomology question is due to the lack of a strict consensus sequence for top strand cleavage, although a weak preference for the sequence 5′-TYTN/R-3′ has recently been proposed (Gilbert et al. 2005). Second, a substantial proportion of the *Alu* insertion polymorphisms from dbRIP were excluded from further analysis; these included (i) entries overlapping with the pathological *Alu* insertional mutations listed in Supplementary Table S1, (ii) entries for which the repeat sequences and/or TSDs are unknown, (iii) full-length *Alu* insertions with additional nucleotides at their 5′ ends and (iv) various other entries that were uninformative with respect to the question of microhomology (Supplementary Table S6). Lastly, as is evident from inspection of Supplementary Tables S3 and S4, a significant proportion of the *Alu* insertions with starting positions at 2, 3 and 4 can be alternatively interpreted as full-length inserts; this issue will be addressed further at the end of the following section.

The sub-family of each selected *Alu* insert was checked/annotated using *RepeatMasker* (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker; as of December 6, 2006). Although in some cases, annotations were different from those previously reported in Chen et al. (2005, 2006) and dbRIP, this did not affect the conclusions of the study in any way. Consensus sequences of *Alu*Ya5, *Alu*Ya8, *Alu*Yb8, *Alu*Yb9, *Alu*Y, *Alu*Sq, *Alu*Yg6, *Alu*Yd8 and *Alu*Sp sub-families were taken from *Repbase* (http://www.girinst.org/repbase/update/browse.php; Jurka et al. 2005).



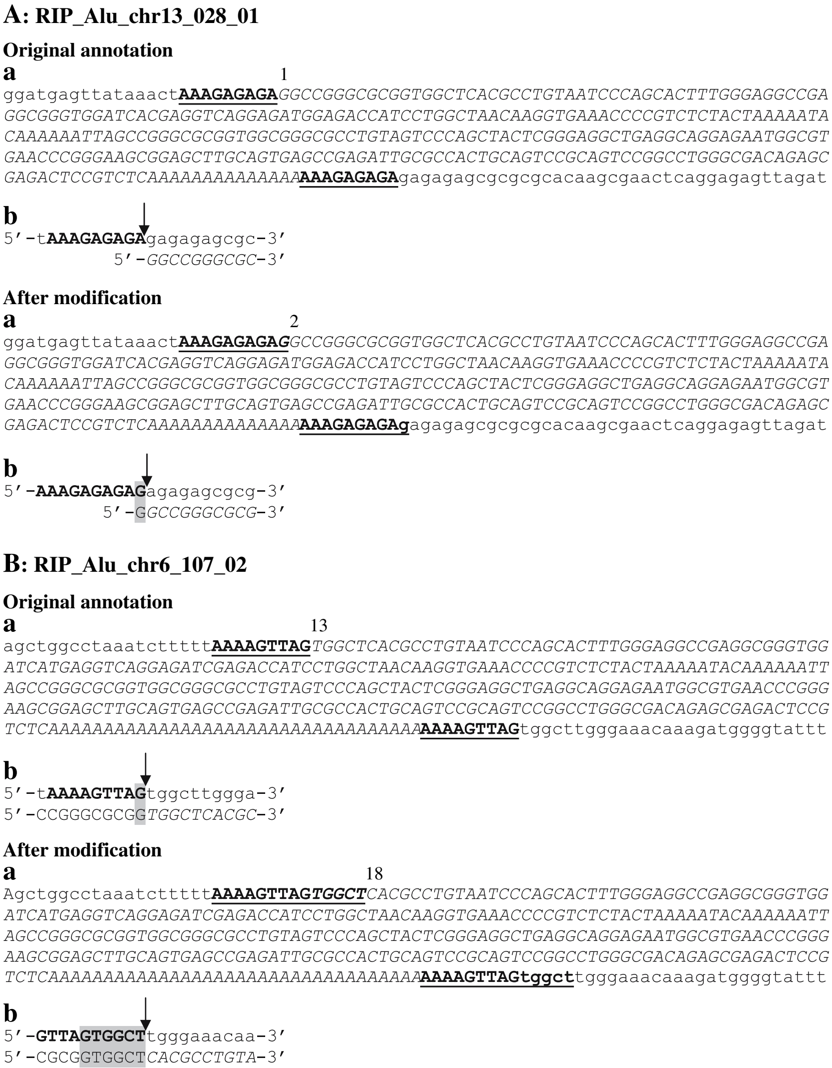
```
AluYa5    GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGA 60
AluYb8    GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGA 60
AluYb9    GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGA 60
AluY      GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGA 60
Alusq     GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGA 60
          ******************************************************** ***

AluYa5    TCA--CGAGGTCAGGAGATCGAGACCATCCCGGCTAAAACGGTGAAACCCCGTCTCTACT 118
AluYb8    TCA--TGAGGTCAGGAGATCGAGACCATCCTGGCTAACAAGGTGAAACCCCGTCTCTACT 118
AluYb9    TCA--TGAGGTCAGGAGATCGAGACCATCCTGGCTAACAAGGTGAAACCCCGTCTCTACT 118
AluY      TCA--CGAGGTCAGGAGATCGAGACCATCCTGGCTAACACGGTGAAACCCCGTCTCTACT 118
AluSq     TCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACATGGTGAAACCCCGTCTCTACT 120
          ***   *********** ********* ** *** ** * ********************

AluYa5    AAAAATACAAAAAATTAGCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGGGA 178
AluYb8    AAAAATACAAAAAATTAGCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGA 178
AluYb9    AAAAATACAAAAAATTAGCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTGGGGA 178
AluY      AAAAATACAAAAAATTAGCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGA 178
AluSq     AAAAATACAAAAA-TTAGCCGGGCGTGGTGGCGGGCGCCTGTAATCCCAGCTACTCGGGA 179
          ************* ***********  **************** *********** ****

AluYa5    GGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCCC 238
AluYb8    GGCTGAGGCAGGAGAATGGCGTGAACCCGGGAAGCGGAGCTTGCAGTGAGCCGAGATTGC 238
AluYb9    GGCTGAGGCAGGAGAATGGCGTGAACCCGGGAAGCGGAGCTTGCAGTGAGCCGAGATTGC 238
AluY      GGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCGC 238
AluSq     GGCTGAGGCAGGAGAATCGCTTGAACCCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGC 239
          ***************** ** *********** ****** *****************  *

AluYa5    GCCACTGCACTCCA-------GCCTGGGCGACA-GAGCGAGACTCCGTCTC 281
AluYb8    GCCACTGCAGTCCGCAGTCCGGCCTGGGCGACA-GAGCGAGACTCCGTCTC 288
AluYb9    GCCACTGCAGTCCGCAGTCCGGCCTGGGCGACA-GAGCGAGACTCCGTCTC 288
AluY      GCCACTGCACTCCA-------GCCTGGGCGACA-GAGCGAGACTCCGTCTC 281
AluSq     GCCACTGCACTCCA-------GCCTGGGCAACAAGAGCGAAACTCCGTCTC 283
          ********* ***        ******** *** ****** **********
```

Fig. 1 Alignment of the consensus sequences of five *Alu* sub-families. Dashes indicate gaps introduced so as to maximise alignment. Nucleotides identical between all sequences are indicated by asterisks. Pathological *Alu* insertions (including 28 simple ones and three associated with large genomic deletions) that are informative with respect to starting position in their respective *Alu* sub-family consensus sequences, are positioned accordingly in the aligned sequences. Note that the sub-family of the shortest *Alu* insert, which comprises CGTCTC plus A_{40} and is associated with the Δ1444 bp in the *SERPINC1* gene (Beauchamp et al. 2000; Chen et al. 2005), could not be assigned. Shaded arrows indicate either entries (underlined) that can be alternatively annotated as full-length *Alu* inserts or those that are not informative with respect to the 'microhomology' question (refer to Supplementary Table S1 for details). Note that (i) microhomology existing between the top strand's 5′ overhang and the sequence that lies 5′ to the truncation position in the *Alu* consensus sequence was identified in the same way as for the *Alu* insertion polymorphisms (see second section of the text) and (ii) only *Alu* inserts with starting position 6 or greater were regarded as 5′-truncated in accordance with Zingler et al. (2005)

Sequence alignments were performed with ClustalW (http://www.ebi.ac.uk/clustalw/#).

A trimodal length distribution of simple *Alu* inserts and the role of microhomology in generating shorter *Alu* inserts

Studies of recently inserted genomic L1 elements in the human genome (Myers et al. 2002; Pavlicek et al. 2002; Szak et al. 2002; Boissinot et al. 2004), pathological L1 direct insertions (Chen et al. 2005), and *de novo* L1 insertions in cultured human cells (Gilbert et al. 2002; 2005) as well as in a transgenic mouse model (Babushok et al. 2006) have consistently shown that simple L1 inserts display a bimodal length distribution with a large peak of short (<2 kb) and a smaller peak of longer (~6 kb) integrations. Although the exact mechanism underlying this bimodal distribution remains controversial (e.g. Farley et al. 2004; Gilbert et al. 2005), the generation of the abundant short L1 inserts would appear to be facilitated by the presence of microhomologies frequently found between



A: RIP_Alu_chr13_028_01

Original annotation

a 1

ggatgagttataaact**AAAGAGAGA***GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGA*
GGCGGGTGGATCACGAGGTCAGGAGATGGAGACCATCCTGGCTAACAAGGTGAAACCCCGTCTCTACTAAAAATA
CAAAAAATTAGCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGT
GAACCCGGGAAGCGGAGCTTGCAGTGAGCCGAGATTGCGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACAGAGC
*GAGACTCCGTCTCAAAAAAAAAAAAAA***AAAGAGAGA**gagagagcgcgcacaagcgaactcaggagagttagat

b

5′-t**AAAGAGAGA**gagagagcgc-3′
5′-*GGCCGGGCGC*-3′

After modification

a 2

ggatgagttataaact**AAAGAGAGA*G****GCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGA*
GGCGGGTGGATCACGAGGTCAGGAGATGGAGACCATCCTGGCTAACAAGGTGAAACCCCGTCTCTACTAAAAATA
CAAAAAATTAGCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGT
GAACCCGGGAAGCGGAGCTTGCAGTGAGCCGAGATTGCGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACAGAGC
*GAGACTCCGTCTCAAAAAAAAAAAAAA***AAAGAGAGAg**agagagcgcgcacaagcgaactcaggagagttagat

b

5′-**AAAGAGAGAG**agagagcgcg-3′
5′-G*GCCGGGCGCG*-3′

B: RIP_Alu_chr6_107_02

Original annotation

a 13

agctggcctaaatcttttt**AAAAGTTAG***TGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGG*
ATCATGAGGTCAGGAGATCGAGACCATCCTGGCTAACAAGGTGAAACCCCGTCTCTACTAAAAATACAAAAAATT
AGCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGG
AAGCGGAGCTTGCAGTGAGCCGAGATTGCGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACAGAGCGAGACTCCG
*TCTCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA***AAAAGTTAG**tggcttgggaaacaaagatggggtattt

b

5′-t**AAAAGTTAG**tggcttggga-3′
5′-CCGGGCGCGG*TGGCTCACGC*-3′

After modification

a 18

Agctggcctaaatcttttt**AAAAGTTAG*TGGCT****CACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGG*
ATCATGAGGTCAGGAGATCGAGACCATCCTGGCTAACAAGGTGAAACCCCGTCTCTACTAAAAATACAAAAAATT
AGCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGG
AAGCGGAGCTTGCAGTGAGCCGAGATTGCGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACAGAGCGAGACTCCG
*TCTCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA***AAAAGTTAGtggct**tgggaaacaaagatggggtattt

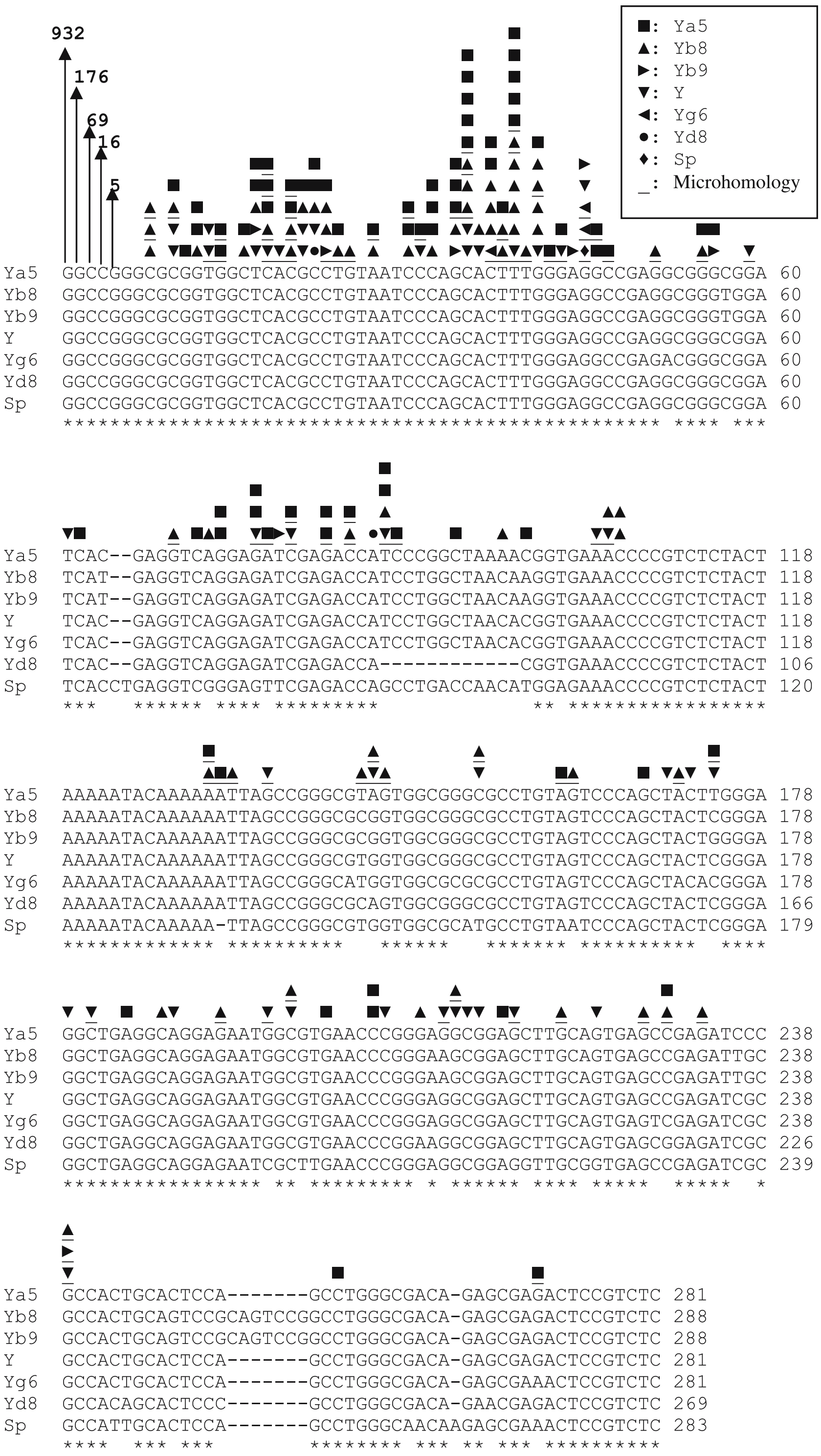
b

5′-**GTTAGTGGCT**tgggaaacaa-3′
5′-CGCGGTGGCT*CACGCCTGTA*-3′

Fig. 2 Two examples of how the starting positions of *Alu* inserts were modified, taking into account the question of 'microhomology'. Both examples (**A** and **B**) were taken from dbRIP, the *Database of Retrotransposon Insertion Polymorphisms in Humans* (http://falcon.roswellpark.org:9090/searchRIP.html). (**a**) Target site duplications (TSDs) are highlighted in bold and underlined; *Alu* sequence plus the poly(A) tail are italicised; the starting position of the *Alu* insert is indicated by an Arabic numeral. (**b**) *Top sequence:* ±10 bp flanking the top strand cleavage site (indicated by an arrow) deduced from **a**; *lower sequence*: whilst italicised sequence on the right side corresponds to the ten 5′-most nucleotides of the *Alu* insert illustrated in **a**, sequence not italicised on the left side was taken from the *Alu* insert's respective consensus sequence at corresponding positions where applicable. Microhomology is shaded wherever applicable. Note that in **A**, re-assigning the first G of the originally annotated full-length *Alu* insert into the upstream TSD resulted in the generation of a one base-microhomology between the top strand's 5′ overhang and the now 5′-truncated (1 bp) *Alu* insert. In **B**, re-assigning the 5′-most TGGCT of a 5′-truncated *Alu* insert into the upstream TSD resulted in the generation of more extensive microhomology

the top strand's 5′ overhang in the target genomic sequence and the 3′ end of the nascent L1 RT-transcribed cDNA strand (Zingler et al. 2005; Babushok et al. 2006).

As shown in Fig. 3, a trimodal length distribution of the 1402 informative *Alu* insertion polymorphisms is apparent: a major peak of full-length or almost full-length inserts (starting positions at 1–5; termed Group I for ease of discussion) with a frequency of ~85% (1198/1402), a smaller peak of 115 inserts initiating from positions 8–47 (frequency, ~8%; termed Group II), and the remaining inserts beginning from after position 51 to the end (termed Group III). The major peak was not unanticipated since (i) a full-length *Alu* insert is <290 bp and (ii) the L1 RT is believed to be of high processivity, by analogy with the property of *Bombyx mori* R2Bm RT (Bibillo and Eickbush 2002; Gilbert et al. 2005). Here it is worth noting that the observed frequency of Group I inserts is consistent with the finding that genome-wide ~90% of *Alu* insertions are full-length [with full-length being defined as those elements initiating within the first five nucleotides of the consensus



```
Ya5  GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGA 60
Yb8  GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGA 60
Yb9  GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGA 60
Y    GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGA 60
Yg6  GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGACGGGCGGA 60
Yd8  GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGA 60
Sp   GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGA 60
     *************************************************** **** ***

Ya5  TCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAACGGTGAAACCCCGTCTCTACT 118
Yb8  TCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACAAGGTGAAACCCCGTCTCTACT 118
Yb9  TCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACAAGGTGAAACCCCGTCTCTACT 118
Y    TCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACACGGTGAAACCCCGTCTCTACT 118
Yg6  TCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACACGGTGAAACCCCGTCTCTACT 118
Yd8  TCAC--GAGGTCAGGAGATCGAGACCA-------------CGGTGAAACCCCGTCTCTACT 106
Sp   TCACCTGAGGTCGGGAGTTCGAGACCAGCCTGACCAACATGGAGAAACCCCGTCTCTACT 120
     ***   ****** **** *********               ** *****************

Ya5  AAAAATACAAAAAAATTAGCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGGGA 178
Yb8  AAAAATACAAAAAAATTAGCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGA 178
Yb9  AAAAATACAAAAAAATTAGCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTGGGGA 178
Y    AAAAATACAAAAAAATTAGCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGA 178
Yg6  AAAAATACAAAAAAATTAGCCGGGCATGGTGGCGCGCGCCTGTAGTCCCAGCTACACGGGA 178
Yd8  AAAAATACAAAAAAATTAGCCGGGCGCAGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGA 166
Sp   AAAAATACAAAAA-TTAGCCGGGCGTGGTGGCGCATGCCTGTAATCCCAGCTACTCGGGA 179
     ************* **********   ******   ******* **********  ****

Ya5  GGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCCC 238
Yb8  GGCTGAGGCAGGAGAATGGCGTGAACCCGGGAAGCGGAGCTTGCAGTGAGCCGAGATTGC 238
Yb9  GGCTGAGGCAGGAGAATGGCGTGAACCCGGGAAGCGGAGCTTGCAGTGAGCCGAGATTGC 238
Y    GGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCGC 238
Yg6  GGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGTCGAGATCGC 238
Yd8  GGCTGAGGCAGGAGAATGGCGTGAACCCGGAAGGCGGAGCTTGCAGTGAGCGGAGATCGC 226
Sp   GGCTGAGGCAGGAGAATCGCTTGAACCCGGGAGGCGGAGGTTGCGGTGAGCCGAGATCGC 239
     ***************** ** ********* * ****** **** *****  *****  *

Ya5  GCCACTGCACTCCA-------GCCTGGGCGACA-GAGCGAGACTCCGTCTC 281
Yb8  GCCACTGCAGTCCGCAGTCCGGCCTGGGCGACA-GAGCGAGACTCCGTCTC 288
Yb9  GCCACTGCAGTCCGCAGTCCGGCCTGGGCGACA-GAGCGAGACTCCGTCTC 288
Y    GCCACTGCACTCCA-------GCCTGGGCGACA-GAGCGAGACTCCGTCTC 281
Yg6  GCCACTGCACTCCA-------GCCTGGGCGACA-GAGCGAAACTCCGTCTC 281
Yd8  GCCACAGCACTCCC-------GCCTGGGCGACA-GAACGAGACTCCGTCTC 269
Sp   GCCATTGCACTCCA-------GCCTGGGCAACAAGAGCGAAACTCCGTCTC 283
     ****  *** ***        ******** *** ** *** **********
```

Fig. 3 Global survey of *Alu* insertion polymorphisms selected from dbRIP (Wang et al. 2006). The Figure is presented essentially in the same manner as Fig. 1. However, for full-length or near full-length entries (i.e. starting positions at 1–5), only the total number is provided, respectively. See Supplementary Tables S2–S5 for details of all entries

sequence; Zingler et al. (2005)]. Thus, by contrast with the situation pertaining with L1 elements, for most *Alu* sequences the process of cDNA synthesis would have a high probability of completion before being counteracted by the host repair machinery.

The smaller peak constituting Group II is however intriguing. On the one hand, all 115 truncations occurred within a relatively short region of 40 bases that is well-conserved between different *Alu* sub-families (Fig. 3). On the other hand, microhomology was only evident in 34.8%

Table 1 Correlation between the Presence of Microhomology (1–7 bp) and the length of the 5′ truncation of *Alu* insertion polymorphisms[a]

Starting positions	Number of entries manifesting microhomology (A)	Total number of entries (B)	% (A/B)
8–47	40	115	34.8
	23 (1 bp)		20.0
	17 (≥2 bp)		14.8
51–106	15	38	39.5
	10 (1 bp)		26.3
	5 (≥2 bp)		13.2
131–288	29	51	56.8
	17 (1 bp)		33.3
	12 (≥2 bp)		23.5

[a] Data from Fig. 3

of the 115 entries (Fig. 3; Table 1). With respect to the mechanism underlying the generation of these Group II *Alu* insertions, we currently envisage two possible models, one operating at the level of transcription (i.e. from DNA to RNA), the other at the level of reverse transcription (i.e. from the RNA to the nascent cDNA strand). Both models are predicated upon the assumption that the behaviour of L1 RT is similar to that of *Bombyx mori* R2 RT, which readily jumps from the 5′ terminal end of the R2 RNA but very inefficiently from internal positions (Bibillo and Eickbush 2004). The first of these models proposes that the truncations arise through the use of alternative transcriptional start sites, in the context of the internal RNA polymerase III promoter [see Fig. 1 in Murphy and Baralle (1983) and Fig. 1 in Shankar et al. (2004) for the RNA polymerase III promoter structure and location within the *Alu* element itself]. This proposition is based upon two observations. First, the Group II inserts are located entirely within the A- and B-box consensus sequences of the polymerase III promoter (Murphy and Baralle 1983; Shankar et al. 2004); this strongly implies the involvement of alternative transcription sites in the generation of these 5′ truncated *Alu* inserts. Second, the use of alternative transcription start sites is not infrequent in genes that are transcribed by RNA polymerase II, although this has not been empirically demonstrated for RNA polymerase III transcripts. Formation of Group II inserts would proceed in the same way as for full-length inserts: upon reaching the 5′ end of the truncated *Alu* RNA, the L1 RT would jump from the RNA template to the 3′ end of the top strand's 5′ overhang [see Fig. 3A in Zingler et al. (2005) and Fig. 5D, 2 in Babushok et al. (2006)]. The alternative model proposes that the truncations result from the degradation of *Alu* RNA by cellular RNase H (Ostertag and Kazazian 2001a; Zingler et al. 2005), the clustering of truncation sites being due to the occurrence of a specific secondary structure that prevents further RNA degradation by binding to *trans*-stabilising factors. Under this model, the formation of these truncated insertions would be identical to that envisaged under the first model, given that L1 RT can process to the 5′ end of a 5′ degraded *Alu* RNA.

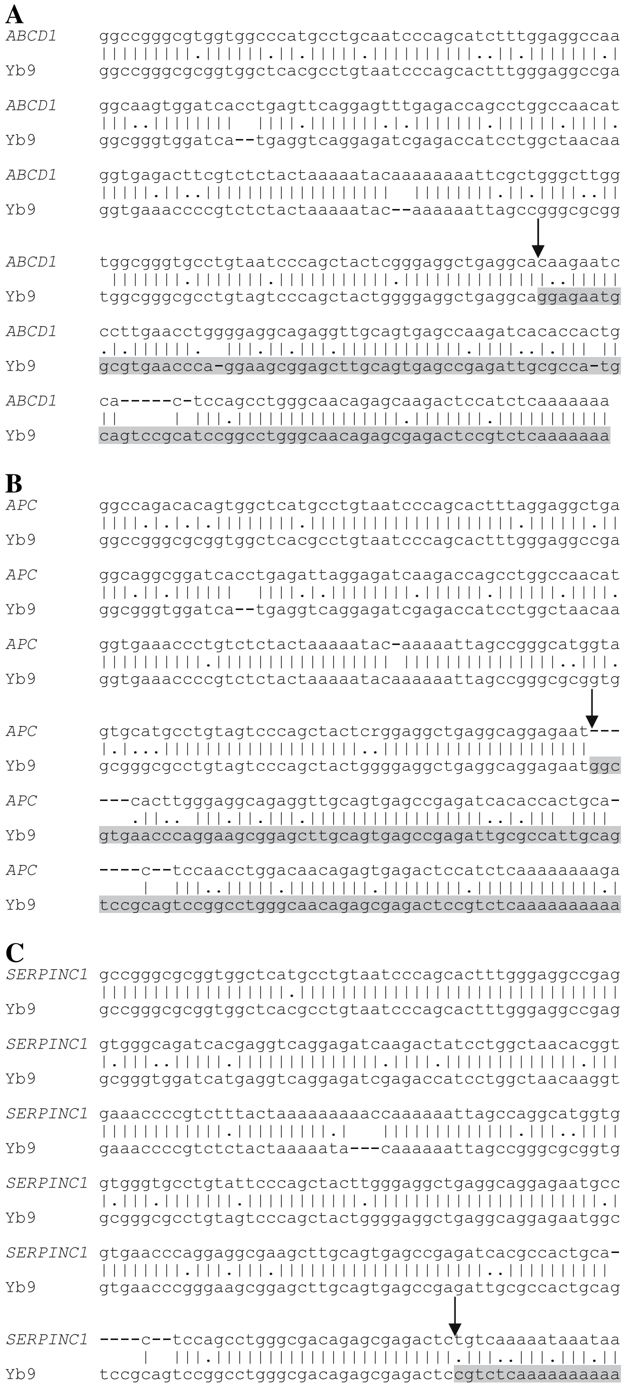
As mentioned above, only 34.8% of the Group II *Alu* inserts were found to exhibit microhomology. By contrast, microhomology was found in some 50% (44/89) of the Group III *Alu* inserts. As a matter of fact, in the context of the 5′ truncated *Alu* insertion polymorphisms (i.e. starting positions, 8–271), there exists a positive correlation between the presence of microhomology and the length of the 5′ truncation (Table 1), thereby suggesting an important role of the MMEJ mechanism in generating shorter *Alu* inserts. Under this model, the generation of most of the shorter *Alu* inserts could have been promoted by the inadvertent annealing of the microhomology present between the 3′ end of the nascent *Alu* cDNA strand and the 3′ end of the top strand's 5′ overhang. This would then be followed by the premature termination of nascent cDNA strand synthesis with concomitant initiation of second *Alu* cDNA strand synthesis by either a second L1 RT or a host DNA repair enzyme. In addition, we should point out that our finding differs from the recent genome-wide analysis that has concluded that 5′ truncated *Alu* elements exhibit no (or only a weak) tendency to exhibit microhomology (Zingler et al. 2005). The discrepancy may be due to one or more of the following reasons. Firstly, Zingler et al. (2005) did not address the microhomology issue in relation to the different lengths of 5′ truncation. Secondly, these authors used only computer-generated data with respect to the analysis of the 5′ truncated *Alu* insertions. In other words, they did not analyse the relevant data manually. As shown in Supplementary Tables S3–S6, our manual evaluation led to the re-annotation of a significant fraction of the dbRIP entries.

Finally, as in the case of the pathological *Alu* insertional mutations (Supplementary Table S1), most of the near full-length *Alu* insertion polymorphisms (i.e. starting positions at 2–5) can be alternatively interpreted as *bona fide*

Table 2 Near Full-Length *Alu* insertion polymorphisms (i.e. starting positions 2–5 in accordance with their respective consensus sequences) that can be alternatively interpreted as full-length insertions[a]

Starting position	Number of entries that can be alternatively interpreted as full-length insertions	Total number of entries
2	145	176
3	60	69
4	15	16
5	0	5

[a] See Supplementary Tables S3 and S4 for detailed information



```
A
ABCD1     ggccgggcgtggtggcccatgcctgcaatcccagcatctttggaggccaa
          |||||||||.||||||.||.|||||.||||||||||..||.|||||||.|
Yb9       ggccgggcgcggtggctcacgcctgtaatcccagcactttgggaggccga

ABCD1     ggcaagtggatcacctgagttcaggagtttgagaccagcctggccaacat
          |||..||||||||  ||||.|||||||.|.|||||||.||||||.||||.
Yb9       ggcgggtggatca--tgaggtcaggagatcgagaccatcctggctaacaa

ABCD1     ggtgagacttcgtctctactaaaaatacaaaaaaaattcgctgggcttgg
          |||||.||..||||||||||||||||||  ||||||||.||.||||..||
Yb9       ggtgaaaccccgtctctactaaaaatac--aaaaaattagccgggcgcgg

ABCD1     tggcgggtgcctgtaatcccagctactcgggaggctgaggcacaagaatc
           |||||||.|||||||.|||||||||||.||||||||||||||..|||||
Yb9       tggcgggcgcctgtagtcccagctactggggaggctgaggcaggagaatg

ABCD1     ccttgaacctggggaggcagaggttgcagtgagccaagatcacaccactg
          .|.||||||.  |||.||.|||.||||||||||||.||||..|.||| ||
Yb9       gcgtgaaccca-ggaagcggagcttgcagtgagccgagattgcgcca-tg

ABCD1     ca-----c-tccagcctgggcaacagagcaagactccatctcaaaaaaa
          ||     | |||.||||||||||||||||.|||||||.|||||||||||
Yb9       cagtccgcatccggcctgggcaacagagcgagactccgtctcaaaaaaa

B
APC       ggccagacacagtggctcatgcctgtaatcccagcactttaggaggctga
          ||||.|.|.|.||||||||.||||||||||||||||||||.||||||.||
Yb9       ggccgggcgcggtggctcacgcctgtaatcccagcactttgggaggccga

APC       ggcaggcggatcacctgagattaggagatcaagaccagcctggccaacat
          |||.||.||||||  ||||.|.||||||||.||||||.||||||.||||.
Yb9       ggcgggtggatca--tgaggtcaggagatcgagaccatcctggctaacaa

APC       ggtgaaaccctgtctctactaaaaatac-aaaaattagccgggcatggta
          ||||||||||.|||||||||||||||||| |||||||||||||||..|||.
Yb9       ggtgaaaccccgtctctactaaaaatacaaaaaattagccgggcgcggtg

APC       gtgcatgcctgtagtcccagctactcrggaggctgaggcaggagaat---
          |.|...|||||||||||||||||||..||||||||||||||||||||
Yb9       gcgggcgcctgtagtcccagctactggggaggctgaggcaggagaatggc

APC       ---cacttgggaggcagaggttgcagtgagccgagatcacaccactgca-
             .||..  |||.||.|||.||||||||||||||||..|.|||  ||||
Yb9       gtgaacccaggaagcggagcttgcagtgagccgagattgcgccattgcag

APC       ----c--tccaacctggacaacagagtgagactccatctcaaaaaaaaga
              |  |||..|||||.||||||||.||||||||.||||||||||||.|
Yb9       tccgcagtccggcctgggcaacagagcgagactccgtctcaaaaaaaaaa

C
SERPINC1  gccgggcgcggtggctcatgcctgtaatcccagcactttgggaggccgag
          ||||||||||||||||||.|||||||||||||||||||||||||||||||
Yb9       gccgggcgcggtggctcacgcctgtaatcccagcactttgggaggccgag

SERPINC1  gtgggcagatcacgaggtcaggagatcaagactatcctggctaacacggt
          |.|||..|||||.||||||||||||||.||||.|||||||||||||.|||
Yb9       gcgggtggatcatgaggtcaggagatcgagaccatcctggctaacaaggt

SERPINC1  gaaaccccgtctttactaaaaaaaaaccaaaaaattagccaggcatggtg
          ||||||||||||.|||||||||.|   |||||||||||||.|||..||||
Yb9       gaaaccccgtctctactaaaaata---caaaaaattagccgggcgcggtg

SERPINC1  gtgggtgcctgtattcccagctacttgggaggctgaggcaggagaatgcc
          |.|||.|||||||.|||||||||||.||||||||||||||||||||||.|
Yb9       gcgggcgcctgtagtcccagctactggggaggctgaggcaggagaatggc

SERPINC1  gtgaacccaggaggcgaagcttgcagtgagccgagatcacgccactgca-
          ||||||||.|||.|||.||||||||||||||||||||..||||||||||
Yb9       gtgaacccgggaagcggagcttgcagtgagccgagattgcgccactgcag

SERPINC1  ----c--tccagcctgggcgacagagcgagactctgtcaaaataaataa
              |  |||.|||||||||||||||||||||||.|||..|||.|||.||
Yb9       tccgcagtccggcctgggcgacagagcgagactccgtctcaaaaaaaaaa
```

◀ **Fig. 4** Pairwise alignment of the top strand sequences (from 5′ to 3′) overlapping the presumed upstream breakpoints of the *ABCD1* (Kutsche et al. 2002), *APC* (Su et al. 2000) and *SERPINC1* (Beauchamp et al. 2000) genes and their respective *Alu* inserts. Dashes indicate gaps introduced in order to maximise alignment. Identical nucleotides are identified by vertical bars. The putative upstream breakpoints are denoted by vertical arrows. *Alu* sequences contained within the inserts are shaded. Unshaded *Alu* sequences are derived from the consensus *Alu* Yb9 sequence at corresponding positions. For the sake of simplicity, the sub-family of the precursor sequence that generated the shortest *Alu* insert associated with the 1444 bp deletion in the *SERPINC1* gene (Beauchamp et al. 2000) was also arbitrarily designated Yb9 (this does not affect the conclusions drawn owing to the high sequence identity manifested by the members of the *Alu* sub-families; see Fig. 1)

full-length insertions (Table 2). Assuming that L1 RT is of high processivity and given that a full-length *Alu* element is < 290 bp, we believe that most, if not all, of the above entries that can be alternatively interpreted are genuinely full-length insertions. Consequently, we propose that *Alu* insertions should be regarded as full-length whenever possible. Finally, it should be noted that all *Alu* insertions with starting positions beyond five, analysed in this study, cannot be alternatively interpreted to be full-length.

Large deletion-associated short *Alu* inserts appear to be integrated through qualitatively different mechanisms

It is no longer in dispute that L1-mediated retrotransposition generates large genomic deletions, as evidenced by complementary observations made in the context of *in vitro* studies (Gilbert et al. 2002, 2005; Symer et al. 2002), identification of disease-causing mutations (Chen et al. 2005; Mine et al. 2007) and genome-wide analysis (Callinan et al 2005; Han et al. 2005). As we already pointed out in our previous meta-analytical study (Chen et al. 2005), the regions spanning the upstream deletion breakpoints in the target *ABCD1*, *APC* and *SERPINC1* genes were annotated as *Alu* sequences by *RepeatMasker* and hence share significant similarity with the *Alu* inserts of interest (Fig. 4). *Alu* retrotransposition-mediated deletions have also been identified in the human genome in an evolutionary context (Callinan et al. 2005), but it is unclear whether these lesions share the same sequence features as noted in the three above-mentioned pathological mutations.

The generation of the three disease-causing large genomic deletions associated with *Alu* insertions can in principle be accounted for by the model illustrated in Fig. 6B from Gilbert et al. (2002): each event was putatively initiated by L1 endonuclease cleavage on the bottom strand but, unlike the typical process of TPRT leading to the generation of a simple insertional event, the L1 RT-transcribed *Alu* cDNA strand appears to have invaded a double strand break located far upstream of the bottom strand nick/break (Chen et al. 2005). This model can be further refined in the light of new developments in the field. Thus, in a genome-wide analysis of both human and chimpanzee data sets, Han et al. (2005) observed a significant positive correlation between the size of the L1 direct insertion and the size of the associated deletions. Han et al. (2005) surmised that the longer the newly synthesised L1 cDNA strand was, the higher would be the probability of forming sufficient complementarity between the end of the L1 cDNA and the region flanking the 5′ end of the L1 insertion in the ancestral sequence. This is indeed a plausible explanation for the generation of large genomic deletions created upon L1 insertion. This model cannot however be readily extrapolated to cases of large genomic deletions caused by insertions of *Alu* elements, simply because the *Alu* inserts in the three disease-causing events are significantly 5′ truncated (see Fig. 1). This notwithstanding, the model of Han et al. (2005) stimulated us to propose a refined model for the generation of large genomic deletions caused by *Alu* insertions: the significant sequence similarity existing between the regions spanning the top strand's upstream deletion breakpoints and the newly synthesised *Alu* cDNA strands in all three cases (Fig. 4) suggests that the longer the stretch of complementarity, the higher the likelihood of a newly synthesised *Alu* cDNA strand annealing to a double strand break-containing far-upstream region. In this refined model, the position of the *Alu* truncation would be specified by the position of the double strand break in the top strand whereas the synthesis of the *Alu* cDNA strand might not necessarily need to be completed in order to obtain sufficient complementarity for strand annealing/invasion.

One further point warrants further discussion. It is possible that the top strand's upstream double strand break may be attributable to the activity of L1 endonuclease (Gasior et al. 2006). Were this to be the case, this could predict an active role for L1-mediated retrotransposition in creating large genomic deletions. It should however be emphasised that the L1 endonuclease used to generate the top strand's upstream double strand break may not necessarily be the same as that used to create the bottom strand's first nick (Mine et al. 2007), by analogy to the proposition that two different L1 RT molecules may be used for twin-priming, leading to L1 inversion (Ostertag and Kazazian 2001b). It is equally possible that the top strand's upstream double strand break was created independently of L1 endonuclease. Were this to be the case, ''a fascinating scenario would present itself: the organism could have 'hijacked' the L1 machinery to repair an existing double strand break through a mechanism akin to single strand annealing.'' (Chen et al. 2005). In this particular context, L1 integration may represent a 'host/parasite battleground' as it has been termed by Gilbert et al. (2005), in which L1 integration finds itself in a 'race' to complete cDNA synthesis before being 'hijacked' to patch an upstream double strand break.

A unified model for *Alu* insertion into the human genome

Based upon the above observations, we propose a unified model for *Alu* insertion in the human genome. Since *Alu* elements are comparatively short, L1 RT is usually able to complete nascent *Alu* cDNA strand synthesis before jumping to the 3′ end of the top strand's 5′ overhang, resulting in the generation of either full-length (i.e. Group I events) or 5′ truncated (i.e. Group II events) *Alu* inserts. Alternatively, the synthesis of the nascent *Alu* cDNA strand may be terminated prematurely if its 3′ end anneals to the 3′ terminal of the top strand's 5′ overhang by means of microhomology-mediated mispairing, an event which would often lead to the formation of significantly truncated (Group III) *Alu* inserts. Furthermore, the nascent *Alu* cDNA strand may be 'hijacked' to patch existing double strand breaks located in the top-strand's upstream regions (which should usually comprise *Alu*-rich sequences), leading to the generation of large genomic deletions. Clearly, the unified model proposed here is likely to be subjected to further modification/revision by new studies as they emerge.

Acknowledgement This work was supported by the INSERM (Institut National de la Santé et de la Recherche Médicale), France.

References

Babushok DV, Ostertag EM, Courtney CE, Choi JM, Kazazian HH Jr (2006) L1 integration in a transgenic mouse model. Genome Res 16:240–250

Batzer MA, Deininger PL (2002) *Alu* repeats and human genomic diversity. Nat Rev Genet 3:370–379

Beauchamp NJ, Makris M, Preston FE, Peake IR, Daly ME (2000) Major structural defects in the antithrombin gene in four families with type I antithrombin deficiency–partial/complete deletions and rearrangement of the antithrombin gene. Thromb Haemost 83:715–721

Bibillo A, Eickbush TH (2002) High processivity of the reverse transcriptase from a non-long terminal repeat retrotransposon. J Biol Chem 277:34836–34845

Bibillo A, Eickbush TH (2004) End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. J Biol Chem 279:14945–14953

Boissinot S, Entezam A, Young L, Munson PJ, Furano AV (2004) The insertional history of an active family of L1 retrotransposons in humans. Genome Res 14:1221–1231

Callinan PA, Wang J, Herke SW, Garber RK, Liang P, Batzer MA (2005) *Alu* retrotransposition-mediated deletion. J Mol Biol 348:791–800

Chen JM, Stenson PD, Cooper DN, Férec C (2005) A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. Hum Genet 117:411–427

Chen JM, Férec C, Cooper DN (2006) LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease: mutation detection bias and multiple mechanisms of target gene disruption. J Biomed Biotechnol 2006:56182

Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr (2003) Mobile elements and mammalian genome evolution. Curr Opin Genet Dev 13:651–658

Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked *Alu* sequences. Nat Genet 35:41–48

Farley AH, Luning Prak ET, Kazazian HH Jr (2004) More active human L1 retrotransposons produce longer insertions. Nucleic Acids Res 32:502–510

Gasior SL, Wakeman TP, Xu B, Deininger PL (2006) The human LINE-1 retrotransposon creates DNA double-strand breaks. J Mol Biol 357:1383–1393

Gilbert N, Lutz-Prigge S, Moran JV (2002) Genomic deletions created upon LINE-1 retrotransposition. Cell 110:315–325

Gilbert N, Lutz S, Morrish TA, Moran JV (2005) Multiple fates of L1 retrotransposition intermediates in cultured human cells. Mol Cell Biol 25:7780–7795

Hagan CR, Sheffield RF, Rudin CM (2003) Human *Alu* element retrotransposition induced by genotoxic stress. Nat Genet 35:219–220

Han JS, Boeke JD (2005) LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression? Bioessays 27:775–784

Han K, Sen SK, Wang J, Callinan PA, Lee J, Cordaux R, Liang P, Batzer MA (2005) Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. Nucleic Acids Res 33:4040–4052

Hedges DJ, Batzer MA (2005) From the margins of the genome: mobile elements shape primate evolution. Bioessays 27:785–794

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Kazazian HH Jr (2004) Mobile elements: drivers of genome evolution. Science 303:1626–1632

Kutsche K, Ressler B, Katzera HG, Orth U, Gillessen-Kaesbach G, Morlot S, Schwinger E, Gal A (2002) Characterization of breakpoint sequences of five rearrangements in *L1CAM* and *ABCD1* (*ALD*) genes. Hum Mutat 19:526–535

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowki J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, International human genome sequencing consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Mine M, Chen JM, Brivet M, Desguerre I, Marchant D, de Lonlay P, Bernard A, Férec C, Abitbol M, Ricquier D, Marsac C (2007) A large genomic deletion in the *PDHX* gene caused by the retrotranspositional insertion of a full-length LINE-1 element. Hum Mutat 28:137–142

Murphy MH, Baralle FE (1983) Directed semisynthetic point mutational analysis of an RNA polymerase III promoter. Nucleic Acids Res 11:7695–7700

Myers JS, Vincent BJ, Udall H, Watkins WS, Morrish TA, Kilroy GE, Swergold GD, Henke J, Henke L, Moran JV, Jorde LB, Batzer MA (2002) A comprehensive analysis of recently integrated human Ta L1 elements. Am J Hum Genet 71:312–326

Ostertag EM, Kazazian HH Jr (2001a) Biology of mammalian L1 retrotransposons. Annu Rev Genet 35:501–538

Ostertag EM, Kazazian HH Jr (2001b) Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res 11:2059–2065

Pavlicek A, Paces J, Zika R, Hejnar J (2002) Length distribution of long interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection. Gene 300:189–194

Shankar R, Grover D, Brahmachari SK, Mukerji M (2004) Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile *Alu* elements. BMC Evol Biol 4:37

Su LK, Steinbach G, Sawyer JC, Hindi M, Ward PA, Lynch PM (2000) Genomic rearrangements of the *APC* tumor-suppressor gene in familial adenomatous polyposis. Hum Genet 106: 101–107

Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, Boeke JD (2002) Human L1 retrotransposition is associated with genetic instability *in vivo*. Cell 110:327–338

Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD (2002) Molecular archeology of L1 insertions in the human genome. Genome Biol 3(10):research0052

Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P (2006) dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum Mutat 27:323–329

Zingler N, Willhoeft U, Brose HP, Schoder V, Jahns T, Hanschmann KM, Morrish TA, Lower J, Schumann GG (2005) Analysis of 5′ junctions of human LINE-1 and *Alu* retrotransposons suggests an alternative model for 5′-end attachment requiring microhomology-mediated end-joining. Genome Res 15:780–789