High polymorphism level of genomic sequences flanking insertion sites of human endogenous retroviral long terminal repeats

Irina Lavrentieva[a], Natalia E. Broude[b], Yuri Lebedev[a], Irving I. Gottesman[c], Sergei A. Lukyanov[a], Cassandra L. Smith[b], Eugene D. Sverdlov[a,*]

[a] *Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Miklukho-Maklaya 16/10, Moscow 117871, Russia*
[b] *Center for Advanced Biotechnology, Boston University, 36 Cummington Street, Boston, MA 02215, USA*
[c] *Department of Psychology, University of Virginia, Charlottesville, VA 22903, USA*

Received 2 December 1998

Abstract The polymorphism at the multitude of loci adjacent to human endogenous retrovirus long terminal repeats (LTRs) was analyzed by a technique for whole genome differential display based on the PCR suppression effect that provides selective amplification and display of genomic sequences flanking interspersed repeated elements. This strategy is simple, target-specific, requires a small amount of DNA and provides reproducible and highly informative data. The average frequency of polymorphism observed in the vicinity of the LTR insertion sites was found to be about 12%. The high incidence of polymorphism within the LTR flanks together with the frequent location of LTRs near genes makes the LTR loci a useful source of polymorphic markers for gene mapping.

Key words: Genomic differential display; Polymorphism; Suppression polymerase chain reaction; Monozygotic twins

1. Introduction

The analysis of the distribution of polymorphism throughout the human genome is important for the understanding of human evolution. Moreover, the detection of differences between the genomes of monozygotic (MZ) twins discordant for complex diseases could help to reveal the loci that contribute to the disorder. We searched for genomic differences between MZ twins discordant for schizophrenia, which is one of the complex disorders where the genetic component is essential [1]. To this end we used genomic differential display (GDD), a recently developed technique [2,3], which permits simultaneous comparative analysis of repeated sequences and their flanks in the multitude of genomic loci. Here we present data on a GDD analysis of subsets of loci containing inserted HERV long terminal repeats (LTRs). These LTRs are scattered throughout the whole human genome, occupy up to 1% of its length and are transmitted through the germline as stable Mendelian genes [4–6]. The HERVs and LTRs are thought to have been inserted into the germline 10–60 million years ago [7]. They might cause significant and evolutionarily important changes in expression patterns of neighboring genes due to a variety of transcription regulatory elements present in their structures, including promoters, enhancers, hormone-responsive elements, and polyadenylation signals. Earlier we found that the LTRs are frequently located close to human genes [8,9]. Being retrotransposons, HERVs could well be transposed to some blastomeres during the early stages after zygote splitting and MZ twin formation, thus generating genetically divergent 'identical' twins. We have shown a high occurrence of polymorphism in the loci surrounding the LTRs within the human genome. This finding, which is the subject of the current report, suggests that the LTR loci present a valuable source of polymorphic markers closely linked to genes and suitable for gene mapping.

2. Materials and methods

2.1. Samples

Non-phosphorylated oligonucleotides (Table 1) were purchased from Operon Technologies (USA). Fluorescent-labeled (Cy-5) primers were from Amitoff (USA). DNA from peripheral blood lymphocytes from anonymous donors and from cell lines was isolated by a standard phenol procedure. MZ twins discordant for schizophrenia were described in [10].

2.2. Preparation of adapter-ligated DNA

500 ng of human genomic DNA was digested in 50 μl with 20 U of *Hae*III or *Rsa*I restriction enzymes at 37°C for 90 min, and further incubated for 90 min after addition of 10 U of the fresh restriction enzyme. Digested DNA was phenol purified, precipitated with ethanol, dissolved in 20 μl of sterile water and ligated to an excess adapter (2 μM) at 16°C overnight. Ligation was carried out in a final volume of 30 μl containing 50 mM Tris-HCl, pH 7.6, 10 mM $MgCl_2$, 0.5 mM ATP, 10 mM dithiothreitol, 2 μM adapter (oligonucleotides 1 and 2, Table 1) and 5 U of T4 DNA ligase (Life Technologies, USA). The ligation was terminated by incubation at 75°C for 5 min. DNA was then separated from the excess primers by passing through QIAquick DNA Purification Kit (Qiagen, USA) and eluted with 50 μl of sterile water.

2.3. DNA amplification and labeling by PCR

Ligated DNA (3–5 ng) was amplified by PCR in a 25 μl reaction volume containing 1×PCR buffer for DisplayTaq, 2.5 mM $MgCl_2$, 250 μM each of dNTP, and 2.5 U of DisplayTaq DNA polymerase (Display Systems Biotech, USA). 0.2 μM of A1 primer and one of the T primers (Table 1) were used in the first round of PCR hot started by adding TaqStart Antibodies (Clontech, USA). Primers 6 and 7 were Cy-5 labeled. PCR mixtures were subjected to 20–25 amplification cycles (94°C for 15 s, 65°C for 20 s, and 72°C for 30 s) in the Omnigene Temperature Cycler (Hybaid, UK). PCR products obtained after the first round of PCR were 1000-fold diluted and amplified in the second PCR round. The conditions for PCR were the same as in the first round except that anchored A2 primers were used instead of A1 primer. PCR products were analyzed by electrophoresis in a 2% agarose gel and in a 6% PAGE using an ALF sequencing instrument (Pharmacia-Biotech, Sweden).

2.4. Display and analysis of LTR-containing sequences

PCR product (2 μl) was denatured for 3 min at 90°C in a stop solution (Pharmacia-Biotech, Sweden) containing 6 mg/ml of dextran blue and 0.1% SDS in deionized formamide, loaded onto a 6% denaturing PAGE and analyzed on an ALF express sequencing instrument. The result was visualized using the Fragment Manager software

*Corresponding author. Fax: (7) (095) 330-6538.
E-mail: eds@glasnet.ru

PII: S0014-5793(99)00004-6

provided with the instrument. A Cy5-labeled 50 bp ladder (Pharmacia-Biotech, Sweden) was used as a size marker.

2.5. Cloning and analysis of captured sequences

The PCR product obtained after the second PCR round was cloned using a TA cloning kit (InVitrogen, USA). Plasmid DNAs were isolated, and DNA sequences of randomly chosen clones were determined using a fmol DNA Cycle Sequencing System (Promega, USA) with an ALF automated sequencing instrument (Pharmacia-LKB, Sweden).

3. Results

3.1. Description of the technique

The procedure is based on a PCR suppression effect (PS effect) [11,12] (see Fig. 1). Briefly, it involves digestion of genomic DNA with a restriction enzyme (R) and tagging the resulting restriction fragments by ligation to the adapter. The adapter is a pair of complementary oligonucleotides (oligonucleotides 1 and 2, Table 1) of unequal length that form a double-stranded structure able to participate in ligation. Each DNA restriction fragment ligated to the adapters with the ends filled in by means of DNA polymerase repairing has inverted repeats at its termini (Fig. 1, line 2). Therefore, its single-stranded fragments contain self-complementary termini capable of forming intramolecular stem-loop structures (or pan-handle-like structures, Fig. 1, line 3 [12]). The ligated adapter is a 40 bases long oligonucleotide with high GC content to promote and strengthen the sticking of its self-comple-

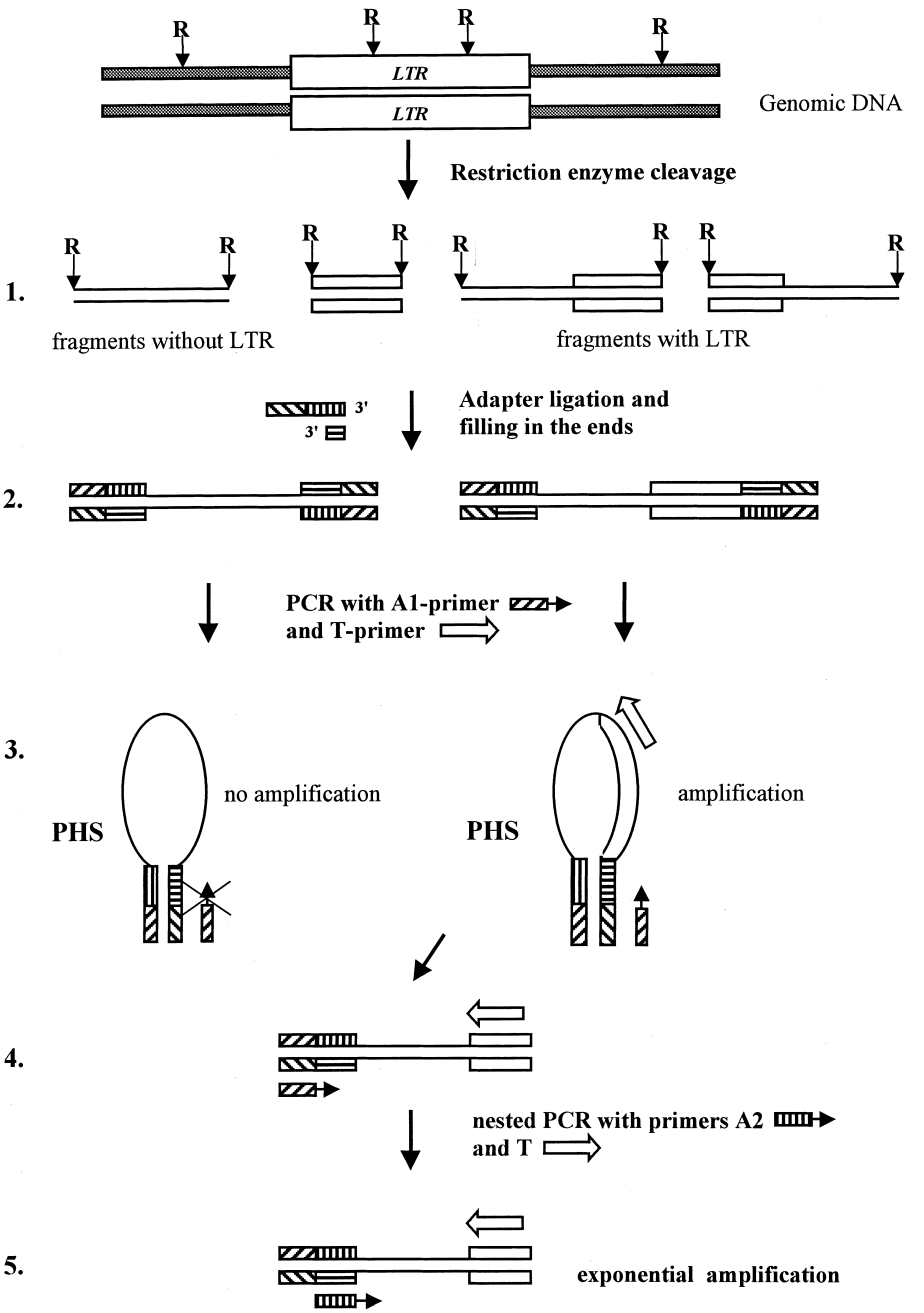

Fig. 1. A scheme of the PCR suppression technique used for selective amplification of human endogenous retroviral LTRs and their flanking sequences.

Table 1
Oligonucleotides used as adapters and primers (5′-3′)

1.	TGTAGCGTGAAGACGACAGAAAGGGCGTGGTGCGGAGGGCGGT	
2.	ACCGCCCTCCG	
3.	TGTAGCGTGAAGACGACAGAA	A1 primer
4.	AGGGCGTGGTGCGGAGGGCGGT	A2 core primer
5.	AGGGCGTGGTGCGGAGGGCGGTCC{N}$_x$	*Hae*III-A2 primer–(N)$_x$.
6.	AGGGCGTGGTGCGGAGGGCGGTCA{N}$_x$	*Rsa*I-A2 primer–(N)$_x$
7.	Cy5-TGGAGTCTGYTATGTCTWCYTCTTTCTAC	U3-LTR-out T primer
8.	Cy5-TTSTTTCTCTATACTTTGTCTCTGTGTCT	U5-LTR-out T primer

For oligonucleotides 5 and 6 $x = 2$, 3 or 4.

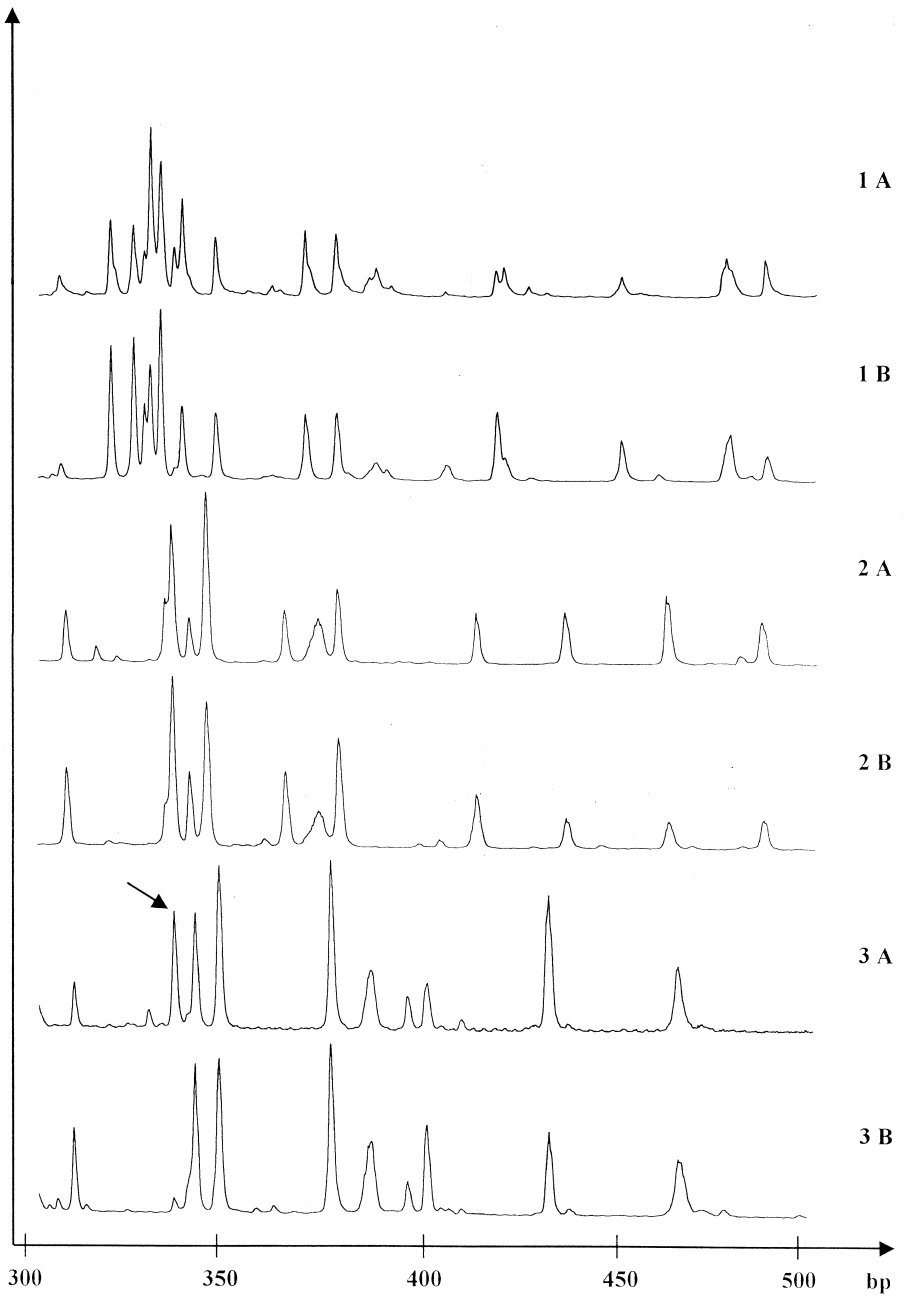

Fig. 2. LTR-containing genome subsets obtained from a pair of MZ twins (designated A and B) using different primer anchors. Genomic DNAs were digested with *Rsa*I, ligated to the adapter and amplified by a two-step PCR. The T primer U3-LTR-out (oligonucleotide 7, Table 1) was used at both PCR steps. The A primer used for the second PCR step was oligonucleotide 6 (Table 1) with anchors AT (lanes 1A and 1B), TG (lanes 2A and 2B), and TC (lanes 3A and 3B). The PCR products were fractionated by size on an ALF express sequencing instrument. Fluorescence intensity (*y*-axis) is plotted against fragment size in bp (*x*-axis). Each lane is independently auto-scaled. The arrow shows the band present in the DNA of co-twin A and absent from the DNA of co-twin B.

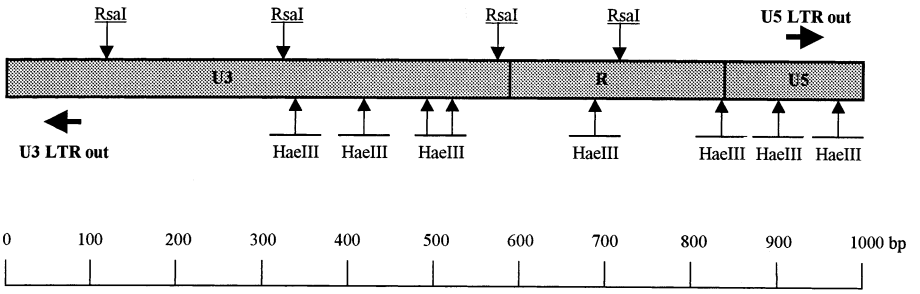

Fig. 3. The LTR consensus sequence. The U3, R and U5 regions as well as the recognition sites for restriction enzymes *Rsa*I and *Hae*III are indicated. Arrows indicate the positions of the T primers, U3-LTR-out and U5-LTR-out, used for the targeted amplification of the LTR flanking sequences. The bp scale is shown below.

mentary ends. PCR of the DNA fragments with such termini will be suppressed if the PCR primer used for amplification is targeted at the 5′ ends of the ligated adapter (A1 primer). The suppression of PCR is a result of a competition for the same target sequence between the A1 primer and the end of the fragment (Fig. 1). The stem-loop structures formed after each PCR denaturation step prevent primer binding and thus suppress PCR at the initiation step (Fig. 1, line 3, left part). Even if the DNA polymerase occasionally overcomes the block imposed by the stems, the newly synthesized DNA template will again form the stem-loop structure, making PCR inefficient. However, the outcome of PCR will be different if, in addition to the A1 primer, another primer (T primer), targeted to the single-stranded part of the stem-loop structure of the fragment is used for amplification (Fig. 1, line 3, right part). The T primer can interact with its target sequence within the fragment and can be used by DNA polymerase for initiation of the DNA synthesis. The newly synthesized PCR product has two different termini and can not form stem-loop structures (Fig. 1, line 4). This PCR product

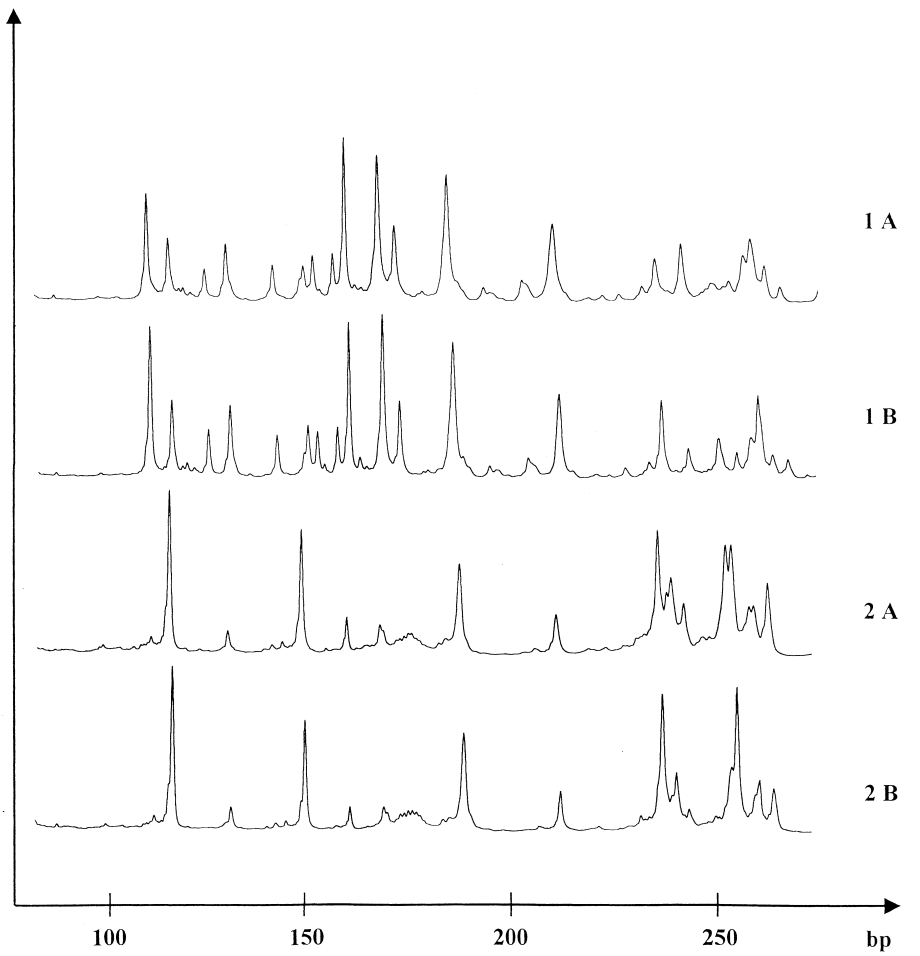

Fig. 4. Complexity reduction of the LTR-containing genome subsets obtained by the GDD technique. DNAs of the MZ twins (the same pair as in Fig. 2) was used for GDD. The T primer was oligonucleotide 7, the A primers were oligonucleotides 6 with anchors TT (lanes 1A and 1 B) and TTC (lanes 2A and 2 B) (Table 1). For other symbols see Fig. 2.

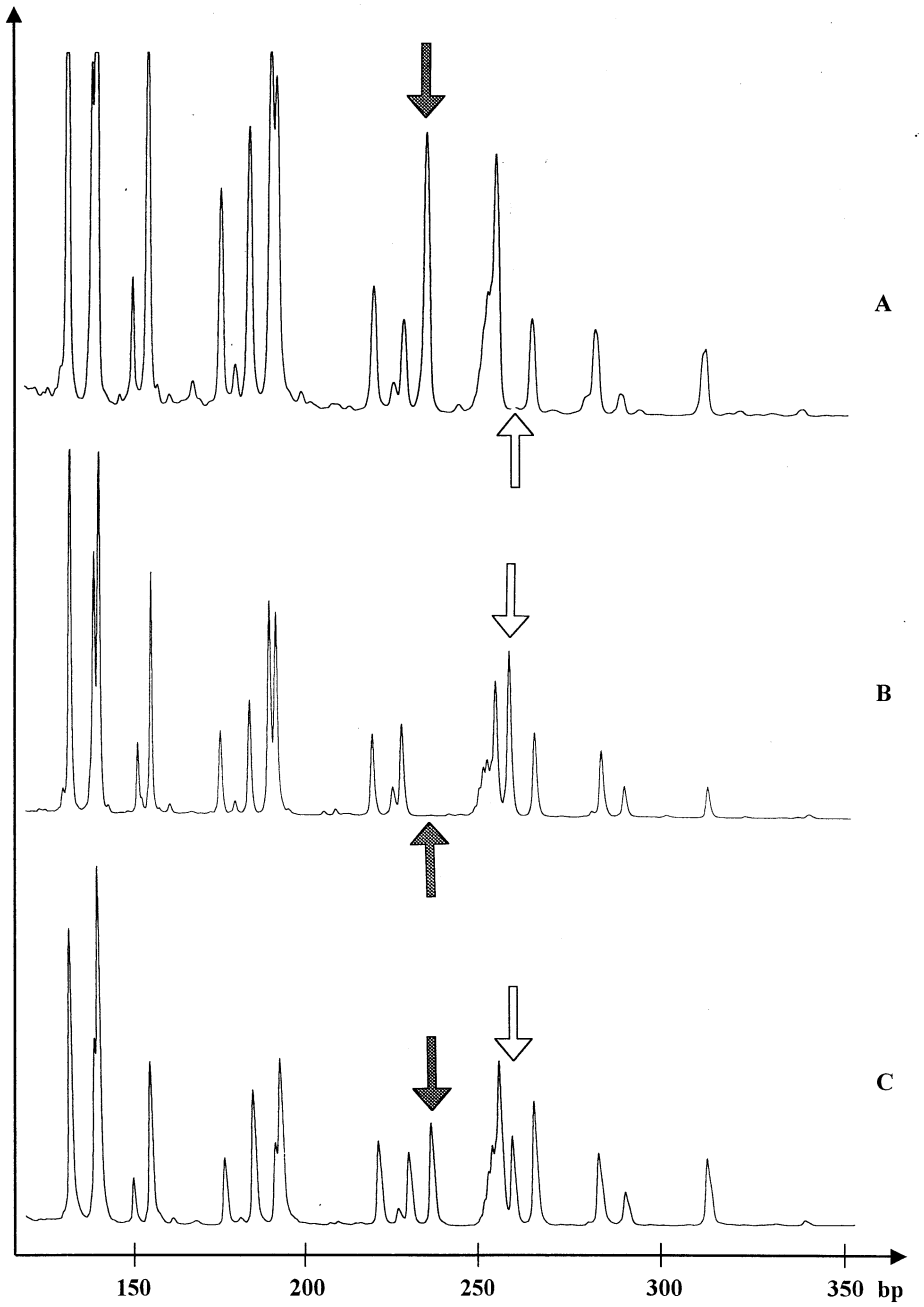

Fig. 5. The LTR-containing genome subsets amplified from DNAs of three unrelated individuals. Each individual was a co-twin of an MZ twins pair. The T primer was oligonucleotide 7 and the A primer was oligonucleotide 6 with the AT anchor (Table 1). The PCR products were analyzed as shown in Fig. 3. Different polymorphic fragments are marked by gray arrows and open ones.

can be efficiently amplified by PCR using the A1 and T primers. To increase the specificity of the amplification, nested PCR with primers A2 and T is used (Fig. 1, line 5). This approach ensures the efficient amplification of only those fragments that contain the targeted sequence. The PS technique was successfully applied for selective amplification of unique genomic DNA or cDNA targeted sequences [11,13]. Recently, we extended the field of application of the PS and demonstrated its usefulness in selective amplification of the multitude of the sequences flanking CAG-repeated elements in the genome [2].

3.2. Selective amplification of LTRs and their flanks

LTR sequences were amplified directly from the genomic DNA by two-step PCR (Fig. 2). In this experiment genomic DNAs from one MZ twin pair were digested with restriction enzyme *Rsa*I, and the restriction fragments were ligated to the adapter (oligonucleotides 1 and 2, Table 1). DNAs were separated from excess adapter and amplified by a two-step PCR. In the first PCR step, the A1 primer (oligonucleotide 3, Table 1) corresponded to the 5′-outermost part of the ligated adapter, and the T primer was targeted at the U3 region of the LTR sequence (oligonucleotide 7, Table 1) (Fig. 3). In other experi-

ments the T primer targeted at the U5 region of the LTR sequence was used (not shown). The sequences of the T primers correspond to the LTR HERV-K consensus sequence [9]. PCR was directed outwards of the LTR as shown in Fig. 3. The PCR product generated in the first PCR step was reamplified with the same T primer and *Rsa*-A2 primers (oligonucleotide 6, Table 1) corresponding to the innermost part of the ligated adapter. The *Rsa*-A2 primers contained three components: (a) a 22 nucleotides long 3′-terminal portion of the ligated adapter (A2 core sequence, oligonucleotide 4, Table 1); (b) the CA dinucleotide corresponding to the ends of the *Rsa*I fragments, and (c) the 3′-terminal di-, tri- or tetranucleotides (anchors), designated $(N)_x$, that anchored the A primer to the complementary genomic DNA sequence adjacent to the *Rsa*I recognition site. The anchors ensure the amplification of only those restriction fragments that contained terminal sequences complementary to the anchors, considerably reducing the complexity of the amplicon mixture. The T primers (Table 1) were fluorescently labeled, and the labeled PCR products were analyzed in a high resolution PAGE using an ALF express sequencing instrument (Pharmacia, Sweden).

In the experiments depicted in Fig. 2, *Rsa*I-A2 primers with one of the three different dinucleotide anchors (AT, TG, or TC) each were used for amplification. Theoretically, a dinucleotide anchor must reduce the complexity of the amplicons 16-fold.

Two features of the patterns obtained are worth mentioning. First, the patterns are different for different anchors as follows from the comparison of lines 1, 2 and 3 (Fig. 2), which is most probably a result of amplification of different subsets of fragments directed by the anchor used. Second, the patterns were almost identical for DNAs of monozygotic twins irrespective of the anchors used, as judged from the pairwise comparison of lanes 1A and 1B, 2A and 2B, and 3A and 3B. This close similarity of the patterns demonstrates the high reproducibility of the technique, supporting our earlier data obtained with the primers designed for CAG triplet repeat amplification [2].

Fig. 4 shows a significant reduction in complexity when a TTC trinucleotide anchor is used instead of a TT dinucleotide. In a random sequence a 64-fold decrease in complexity is found with a trinucleotide anchor as compared to the 16-fold decrease for a dinucleotide anchor. These results show that the specificity of amplification is controlled by the anchor structure, and the complexity of the displayed set can be modulated by the length and sequence of the anchor. The patterns were also characteristically dependent on the restriction endonuclease used for the digestion of genomic DNA, they were different for *Rsa*I and *Hae*III (not shown).

The PCR products obtained with the LTR-specific T primer (U3-LTR-out, Table 1) were cloned, and randomly chosen clones were sequenced. All sequences minus primers were compared to DNA sequences deposited in GenBank using BLAST. Eleven out of 12 clones possessed high homology to known HERV-K LTR sequences, confirming that the subset of LTR-containing fragments was specifically amplified.

3.3. Polymorphisms in the LTR loci

Fig. 5 displays the patterns of 150–350 bp long LTR-containing fragments amplified from DNAs of three unrelated individuals. Each of the individuals was a co-twin of an MZ twin pair. The patterns for co-twins appeared to be identical (not shown). Although the major amplification pattern was common to DNAs from all three individuals, several bands were different and specific for each DNA. These differences reflect polymorphisms, the frequencies of which can be calculated from the number of positions where the differences were observed. In total, using A primers with 16 different dinucleotide anchors, we displayed 700 peaks. Differences between patterns were detected in 90 positions. That gives a polymorphism frequency in the fragments originated from the LTR loci as high as $90/700 = 0.12$.

Several primer combinations revealed genomic differences between MZ twins. For example, Fig. 2 (3A) shows that a 350 bp long fragment is present in the DNA from twin A and is absent from the DNA of co-twin B. Several of these differences are now under further investigation.

4. Discussion

We applied the targeted genomic differential display technique to search for differences between MZ twins in the multitude of loci containing LTRs of human endogenous retroviruses. Unique features of these repetitive sequences dictated the choice of LTRs. On the one hand, they contain a set of transcription regulatory elements such as promoters and enhancers that can influence transcription of the neighboring genes. On the other hand, LTRs as retro elements can be retrotransposed during and after twinning, thus causing germ and somatic mutations reflected as genomic differences and potentially capable of modulating the expression of particular genes. These differences in turn might manifest themselves phenotypically as discordance. Therefore, LTRs seem to be promising in searches for differences between closely related genomes.

The technique used takes advantage of PCR suppression, which allows highly specific amplification of the targeted sequences directly from genomic DNA. By changing the T primer design, the technique can be adapted to any specific target in repetitive sequences of the human genome [2,3]. Another advantage of the technique is the possibility of controlled reduction in the complexity of displayed fragments by using primers with appropriate anchors. All 16 possible dinucleotide anchors allow one to display almost the complete variety of LTR-containing fragments of the human genome and to identify the anchors that reveal differences between co-twins. As the next step of the analysis one can use trinucleotide anchors to simplify the pattern and make the difference more clear-cut. When needed, the complexity can be even further reduced by additional extension of the anchor. Finally, the difference can be isolated and cloned.

The results of this study reveal a high level of polymorphism in the loci of the human genome comprising LTRs. The polymorphism might arise due to several reasons: (a) dimorphisms in the sites of the restriction enzymes used for cleaving genomic DNA, (b) dimorphisms in the T primer binding sites, (c) variations in the fragment lengths due to insertions or deletions, and (d) rare LTR insertion dimorphism. At the moment it is hardly possible to assess the relative contributions of these three factors to the high frequency of polymorphism observed. However, this is probably related to the fact that LTRs are frequently located in variable parts of the genome. In particular, it was demonstrated that retro

elements, including LTRs, are often located within introns or intergenic regions, where they are less harmful for genome functions [4,5]. At the same time they are preferentially inserted into transcriptionally active DNA regions [4,5] and hence can be frequently found in the vicinity of genes.

The LTR-containing fragments were specifically amplified from the human genomic DNA: 90% of randomly chosen clones contained targeted LTR sequences. Using different dinucleotide anchors, we were able to display 20–60 LTR-linked fragments for each DNA. The human genome contains about 10^4 LTR sequences [4–6], which, taking into account the 16-fold reduction in complexity, would suffice to display about 500 fragments. Clearly, the real number should be less because not all possible LTRs match the T primer to provide efficient amplification and, also, too long and too short amplicons fall outside the range suitable for the technique used.

As a preliminary result, the observed polymorphism between MZ twins should be mentioned (Fig. 2). It suggests that the application of the technique to a comparison of discordant MZ twins would be helpful in identifying the genes involved in complex genetic diseases.

Acknowledgements: This work supported by NATO Grant HTECH.LG940570, by the Russian National Human Genome Program and partially supported by Grants NIMH-41176, AIBS2154 DOA and HL55001 NIH.

References

[1] Petronis, A. and Kennedy, J.L. (1995) Am. J. Psychiatry 152, 164–172.
[2] Broude, N., Chandra, A. and Smith, C.L. (1997) Proc. Natl. Acad. Sci. USA 94, 4548–4553.
[3] Broude, N., Storm, N., Malpel, S., Graber, J.H., Lukyanov, S., Sverdlov, E. and Smith, C.L. (1998) Genet. Anal. Biomol. Eng. (in press).
[4] Leib-Mosch, C., Haltmeier, M., Werner, T., Geigl, E.-M., Brack-Werner, R., Francke, U., Erfle, V. and Hehlmann, R. (1993) Genomics 18, 261–269.
[5] Lower, R., Lower, J. and Kurth, R. (1996) Proc. Natl. Acad. Sci. USA 93, 5177–5184.
[6] Patience, C., Wilkinson, D. and Weiss, R. (1997) Trends Genet. 13, 116–120.
[7] Steinhuber, S., Brack, M., Hunsmann, G., Schwelberger, H., Dierich, M.P. and Vogetseder, W. (1995) Hum. Genet. 96, 188–192.
[8] Vinogradova, T., Volik, S., Lebedev, Y., Shevchenko, Y., Lavrentyeva, I., Khil, P., Grzeschik, K.-H., Ashworth, L.K. and Sverdlov, E. (1997) Gene 199, 255–264.
[9] Lavrentieva, I., Khil, P., Vinogradova, T., Akhmedov, A., Lapuk, A., Shahova, O., Lebedev, Y., Monastyrskaya, G. and Sverdlov, E. (1997) Hum. Genet. 102, 107–116.
[10] Torrey, E.F., Bowler, A.E., Taylor E.H. and Gottesman, I.I. (Eds.) (1994) Schizophrenia and Manic-depressive Disorder, Basic Books, New York.
[11] Siebert, P.D., Chenchik, A., Kellogg, D.E., Lukyanov, K.A. and Lukyanov, S.A. (1995) Nucleic Acids Res. 23, 1087–1088.
[12] Lukyanov, S.A., Gurskaya, N.G., Lukyanov, K.A., Tarabykin, V.S. and Sverdlov, E.D. (1994) Bioorgan. Khim. (Moscow) 20, 701–704.
[13] Lukyanov, K.A., Matz, M.V., Bogdanova, E.A., Gurskaya, N.G. and Lukyanov, S.A. (1996) Nucleic Acids Res. 24, 2194–2195.