Genomics 99 (2012) 340–346

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods

Alu insertion profiling: Array-based methods to detect *Alu* insertions in the human genome

Maurizio Cardelli *, Francesca Marchegiani, Mauro Provinciali

Advanced Technology Center for Aging Research, Scientific Technological Area, INRCA-IRCCS, Via Birarelli 8, 60121 Ancona, Italy

ARTICLE INFO

Article history:
Received 19 January 2012
Accepted 23 March 2012
Available online 1 April 2012

Keywords:
Alu
Retroelements
Transposons
Microarray
Tiling array

ABSTRACT

The analysis of the genetic variability associated to *Alu* sequences was hampered by the absence of genome-wide methodologies able to efficiently detect new polymorphisms/mutations among these repetitive elements. Here we describe two *Alu* insertion profiling (AIP) methods based on the hybridization of *Alu*-flanking genomic fragments on tiling microarrays. Protocols are designed to preferentially detect active *Alu* subfamilies. We tested AIP methods by analyzing chromosomes 1 and 6 in two genomic samples. In genomic regions covered by array-features, with a sensitivity of 2% (AIP1) −4% (AIP2) and 5% (AIP1) −8% (AIP2) for the old J and S *Alu* lineages respectively, we obtained a sensitivity of 67% (AIP1) −90% (AIP2) for the young Ya subfamily. Among the loci showing sample-to-sample differences, 5 (AIP1) −8 (AIP2) were associated to known *Alu* polymorphisms. Moreover, we were able to confirm by PCR and DNA sequencing 4 new intragenic *Alu* elements, polymorphic in 10 additional individuals.

1. Introduction

Alu sequences represent, with about one million copies, the most abundant retroelements in the human genome. These SINE (Short Interspersed Nuclear Elements) sequences, exclusively found in primate genomes, have been particularly active in the human lineage even after human–chimpanzee divergence, where they likely contributed to shape some of the human-specific characters such as brain size [1]. Until recently these genomic elements have been mainly considered just as molecular fossils, neutral residents of the human genome, part of the so-called "junk-" or, in the best case, "selfish-DNA". In recent years, however, the role of *Alu* elements in stress response and regulation of gene expression and proteome diversity has been supported by an increasing amount of evidences [2,3]; moreover, their role in RNA editing has been recently emphasized [4]. At the same time, these retroelements represent a powerful source of genetic diversity both at population and individual levels. In fact, it has been recognized that *Alu* sequences are an important source of genetic variability [5], with a possible effect on the phenotype even when inserted in introns, as in the well known cases of ACE and WNK1 genes [6,7]. Conversely, the risk that they represent as a potential source of mutation has been evidenced by many studies [8]. In particular, observations in disease cases and experiments on transposon model systems suggest a possible role of new, germline or somatic, *Alu* insertions in human diseases such as cancer [9,10]. The epigenomic alterations associated to *Alu* repeats in aging and cancer, and the fact that *Alu* and other retroelements are active and capable of inducing mutations not only in the germline but also in somatic cells, suggest that the role of *Alu* retroelements in these two conditions be studied [11,12]. Moreover, retroelement activation seems to be involved in inflammation, a mechanism with causal roles both in cancer and aging [13].

After the origin of the first *Alu* copies in the primate lineage, the *Alu* family has given rise to different subfamilies based upon diagnostic mutations shared by subfamily members, each of them originating from different *Alu* master genes active in different evolutionary periods [14]. The nomenclature *Alu* J, *Alu* S and *Alu* Y is used to indicate, respectively, old, intermediate and young *Alu* classes [14]. Each of these classes is composed of different lineages and subfamilies based on the presence of other diagnostic mutations shared by part of their members. It appears that in the human genome, while most *Alu* elements belong to one of the older subfamilies (class *Alu* J, class *Alu* S with the subfamilies Sx, Sq, Sp, Sc), the *Alu* members capable of producing new *Alu* insertions (at least in the germline) are restricted to a small subset of *Alu* Y members classified as part of the Ya (mainly the Ya5 subfamily) and *Alu* Yb (mainly the Yb8 subfamily) lineages and (more rarely) to other members of the *Alu* Y class [15].

Until recent years the study of *Alu*-associated genetic variability has been hampered by the lack of specific methodologies able to find, on a genomic scale and with good sensitivity, the "new" elements over a background of a million pre-existing copies; in fact, while some methodologies such as *Alu* PCR [16,17] and especially the allele-specific *Alu* PCR [18] were able to detect new *Alu* insertions without

* Corresponding author at: INRCA-IRCCS, Via Birarelli 8, 60121 Ancona, Italy. Fax: +39 071206791.

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2012.03.005

any a priori assumptions on their genomic localization, they were based on the electrophoretic separation of PCR products, limiting the number of the loci simultaneously analyzable and requiring time-consuming additional efforts for the characterization of the polymorphic bands. Consequently, polymorphisms or mutations due to new *Alu* insertions were mainly discovered through whole genome sequencing projects conducted on a small number of samples, or as occasional observations based on methods (such as locus-specific PCR) which likely underestimated their frequency [19]. Recently, however, specific methodologies to detect transposable elements (TE)-associated mutations and polymorphisms have been presented and, as highlighted in a recent review [20] they "promise to revolutionize our ability to analyze human genomes for TE-based variation important to studies of human variability and human disease". Among them, some of the most promising methods are those based on the hybridization of tiling arrays of a DNA probe enriched with transposon-flanking DNA fragments. Similar methods have been initially presented to detect polymorphisms associated to low-copy DNA transposons in the yeast genome [21,22] and have been defined as transposon insertion profiling (TIP)-chip [22]. Based on the same principle, we recently developed two *Alu* insertion profiling methods to detect *Alu*-associated genomic variability [23]. In the present paper, we describe the details of the two methods.

2. Results

2.1. Alu insertion profiling method

To map the position of new *Alu* insertions we adopted a method similar to the TIP-chip approach [22], using genomic DNA fragmentation and vectorette (ligation-mediated) PCR to produce a DNA probe (corresponding to a wide set of small genomic regions flanking individual *Alu* elements) for the hybridization of a tiling array. The *Alu*-specific primer used in both AIP methods for the vectorette PCR step was designed according to the consensus sequence of the *Alu* Y class (in correspondence of diagnostic sites shared by the Ya lineage and most of the other lineages of the *Alu* Y class), in order to reduce as much as possible the signals generated by the "old" *Alu* families; the AIP2 method introduces, in addition, a primer extension step using a primer specific for some of the most active young subfamilies (Ya5, Ya8 and Ya4) in order to further increase the sensitivity of the method for the polymorphic and active Ya subfamilies. The high genomic density of *Alu* repeats suggested the use of high density arrays such as Affymetrix Tiling Arrays of the "2.0R" set, with an average resolution of 35 bp. Each Affymetrix Tiling Array contains features mapping a certain number of chromosomes, and the whole "2.0R" set, composed of seven arrays, has to be hybridized if a whole-genome scanning is needed. In the present work, aimed to optimize the method, we used the first array of the series, corresponding to chromosomes 1 and 6.

2.2. "One sample analysis": sensitivity of the methods for different Alu subfamilies

The results of "one sample" analysis are reported in Table 1 and Fig.1. In Table 1, we compared the different sensitivity of the two methods for different *Alu* subfamilies (we considered only the elements in fully analyzable genomic positions, i.e. those with 500 bp flanking regions composed for at least 60% of unique sequence). For both methods, using a given threshold, the sensitivity is much higher for young *Alu* subfamilies of Y class and especially for *Alu* Ya (and to a lesser extent for *Alu* Yb) than for older subfamilies of J and S classes. The AIP2 method shows a much higher sensitivity than AIP1 for *Alu* Ya elements, while its sensitivity for the other subfamilies seems to be only slightly higher. Given that both methods are based on the array hybridization of DNA fragments corresponding to the 5′ flanking region of each *Alu* element, *Alu* elements completely flanked by long regions of low complexity sequence cannot be detected, because the tiling arrays do not contain features corresponding to low complexity sequences (they would yield non-specific results). However, in Fig.1 we show that AIP methods (and AIP2 in particular) are quite robust in this regard, allowing the detection of *Alu* elements which have 100 bp or more of low complexity sequence at their 5′ flank (Fig.1 B) if a unique sequence (of at least 100 bp) is located not more distant than 500 bp.

Table 1
Repetitive elements detected by the two AIP methods in "one sample analysis".

Repetitive families	Number of considered repetitive elements[a] on Chr. 1 and Chr. 6	Detected by AIP1 [b]	Detected by AIP2 [b]
Alu Ya (Ya5, Ya8, Ya4, Ya1)	111	67%	90%
Alu Yb (Yb8, Yb9)	137	55%	57%
Other *Alu* Y lineages (Y, Yd, Yg, Yh)	6150	41%	48%
Alu J (Jo and Jb)	11,034	2%	4%
Alu S (Sx, Sg, Sc, Sp, Sq)	29,882	5%	8%

Percentage of non polymorphic *Alu* members of various families detected by "one sample analysis" with the two methods; the percentages refer to the average number of signals detected in the two samples of the chromosomes 1 and 6. An *Alu* element was considered "detected" if a significant signal was located at its 5′ flanking region, within 500 bp of its 5′ end.

[a] *Alu* elements considered for this table are those longer than 200 bp (to exclude short *Alu* fragments), fixed (not included among known *Alu* polymorphisms), and with at least 60% of unique (non repetitive) sequence in the 500 bp at their 5′ flank; members of *Alu* families and polymorphic *Alu* elements have been obtained by Repeat Masker track and RIPs track, on UCSC Genome Browser, version NCBI35/hg17.

[b] Calculated using a threshold of 3.4 on a log10 intensity scale.

2.3. "Two sample analysis": direct identification of sample-to-sample differences

The paired analysis of the signals obtained on chromosomes 1 and 6 allowed the identification of putative polymorphic loci showing differences between the two samples A and B. On the whole, 25 signals which were significantly different between the two samples were detected by AIP1 method, while 49 were obtained by AIP2 method (13 of such signals were detected with both methods). The workflow of the "two sample" analysis and the obtained results are summarized in Fig. 2. We observed that about half of the significant intervals were associated (located within 500 bp upstream) to known *Alu* insertions (14/25 with AIP1 and 22/49 with AIP2), about a third of which (5/14 and 8/22 for AIP1 and AIP2 respectively) already known as polymorphic, hence giving a confirmation of the specificity of (part) of the signals. Among the signals associated with known *Alu* elements, 13/14 (AIP1) and 16/22 (AIP2) were associated with elements belonging to young *Alu* subfamilies (in particular to Y, Ya, Yb lineages), the remaining being associated with *Alu* elements of old subfamilies (Sg, Sc, Sx and Jb). The 11 (AIP1) and 28 (AIP2) signals located far (more than 500 bp) from known *Alu* insertions putatively corresponded to new *Alu* insertions.

2.4. Characterization of 4 new intragenic Alu insertions on chromosomes 1 and 6

We checked and verified, by locus-specific PCR on A and B DNA samples and by DNA sequencing of the amplification products, part of the putative new *Alu* insertions detected by the "two sample" analysis, and in particular (for their possible functional significance) the 11 signals (3 detected by AIP1 and 10 by AIP2, including 2 signals detected by both methods) corresponding to intragenic loci. Among these 11 loci, we were able to confirm by locus-specific PCR/

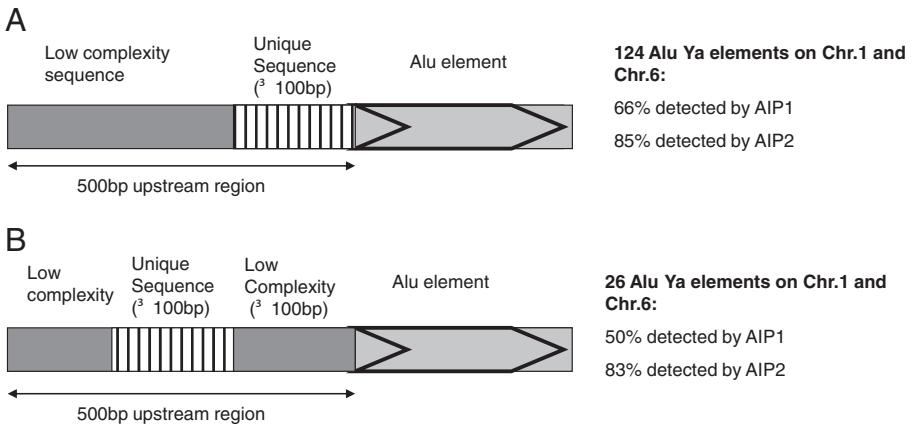

Fig. 1. The figure shows graphically two possible different compositions of the region located upstream to an *Alu* element, and the effect of the upstream sequence on the probability to detect the corresponding *Alu* element (evidenced in light gray). *Alu* Ya elements flanked at 5′ by at least 100 bp of unique sequence are represented in A, while *Alu* Ya elements flanked at 5′ by at least 100 bp of low complexity sequence and with 100 bp or more of unique sequence located within 500 bp are represented in B; the frequency with which the *Alu* Ya elements represented in A and B are detected by the two methods is reported. *Alu* elements considered for the reported data are the *Alu* Ya members longer than 200 bp (short *Alu* fragments are excluded) and fixed (known *Alu* polymorphisms are excluded); an *Alu* element was considered "detected" if a significant signal was located at its 5′ flanking region, within 500 bp of its 5′ end.

electrophoresis and DNA sequencing 4 polymorphic, previously unreported, *Alu* Y insertions: two of them (GenBank ID: **JN391997** and **JN391998**) on chromosome 1 and detected with both methods, one on chromosome 6 (GenBank ID: **JN391999**) detected only by AIP1, and one (GenBank ID: **JN392000**) on chromosome 6 detected only by AIP2. It is useful to note that, albeit in one case (GenBank ID: **JN391997**) the identification of the polymorphic locus was helped by the homozygous condition of both DNA samples (B homozygous for the "*Alu* insertion" vs. A homozygous for the "*Alu* absence"), the other 3 loci were identified as polymorphic despite the A and B DNA samples having one allele in common (the comparison was homozygous for "*Alu* absence" vs. heterozygous), demonstrating the efficacy of AIP methods for detecting mutations or polymorphisms in heterozygous condition.

The details of the new intragenic *Alu* insertions are reported in Table 2, and their sequence can be found in Supplementary File 2 and in GenBank (http://www.ncbi.nlm.nih.gov/genbank).

All the new *Alu* elements identified are inserted in a known gene (*MLK4*, *WDR64*, *NKAIN2*, *PARK2*), and are members of the *Alu* Y class; 3 of them, in particular, belong to the Ya lineage (2 Ya5 and one Ya4 members). After having analyzed, by locus-specific PCR, the genomic DNA samples of 10 additional individuals, the four new *Alu* insertions turned out to be common polymorphisms in this (Italian) population, and not individual mutations. However, given that none of these polymorphisms has been detected by the recent extensive search of new *Alu* polymorphisms accomplished by comparing eight human whole genome sequences from various ethnical groups [28], these polymorphisms are likely to be population-specific.

3. Discussion

We recently developed the *Alu* insertion profiling methods [23] based on the analysis of tiling microarrays. Although TIP-chip [22] and similar tiling-array-based methods [21] were initially applied to

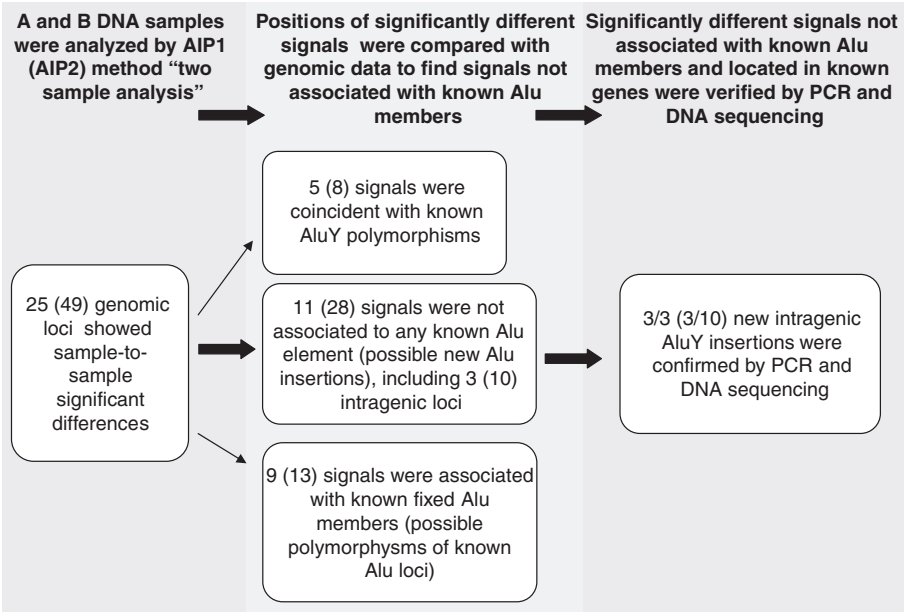

Fig. 2. Workflow followed for the analysis of the data obtained with AIP1 and AIP2 methods, and for the confirmation of part of the detected polymorphic loci.

Table 2
Characterization of four new Alu polymorphic loci.

Genomic site[a]	Method used for the detection	Inserted *Alu* sequence	Frequency of the "*Alu* insertion" allele in 12 individuals[b]	Gene	Gene function
Chr1: 233,519,144–233,519,145	AIP1/AIP2	Genbank JN391997: *Alu* Ya5, homozygous insertion of 346 bp in sample B	8/24	*MLK4*, mixed-lineage kinase 4, exon 10, 3′ untranslated sequence	Mixed-lineage kinase 4, Ser/Thr protein kinase family, MAP kinase kinase kinase subfamily. Found mutated in colon cancer [24]
Chr1: 241,908,619–241,908,620	AIP1/AIP2	Genbank JN391998: *Alu* Ya5, heterozygous insertion of 322 bp in sample B	4/24	*WDR64*, WD repeat domain 64, intron 12	Locus associated with multiple sclerosis [25]
Chr6: 124,865,513–124,865,514;	AIP1	Genbank JN391999: *Alu* Ya4, heterozygous insertion of 314 bp in sample A	5/22	*NKAIN2*, Na+/K+ transporting ATPase interacting 2, intron 3	Transmembrane protein that interacts with the beta subunit of Na,K-ATPase (ATP1B1). A chromosomal translocation involving this gene is a cause of lymphoma. [26]
Chr6: 162,674,813–162,674,814	AIP2	Genbank JN392000: *Alu* Y, heterozygous insertion of 329 bp in sample B	4/24	*PARK2*, parkinson protein 2, E3 ubiquitin protein ligase (parkin), intron 3	Parkin protein; mutations cause autosomal recessive juvenile Parkinsonism [27]. The protein is a component of an E3 ubiquitin ligase complex.

[a] With respect to human genome assembly GRCh37/hg19.
[b] The frequency of the "*Alu* insertion" allele is calculated in 12 samples including A and B samples initially analyzed by AIP and confirmed by PCR, and samples from 10 additional individuals analyzed by PCR only.

map low copy number transposons in yeast, it was hypothesized that they could be modified and adapted to map high copy number retroelements in more complex genomes. However, the development of such methodologies was not immediate. In particular, the detection of new members belonging to the *Alu* family, characterized (with more than one million elements) by the highest copy number and genome density among human interspersed repetitive elements, needed specific requirements. While Huang and co-authors [29], who designed an efficient TIP-chip method to map LINE-1 sequences, reported without description the successful application of a specific TIP-chip methodology for *Alu* mapping, the present article contains the first described protocols for an array-based detection of *Alu* elements. To avoid the overlap of different *Alu*-associated signals we chose to produce the probes using short *Alu*-flanking DNA fragments (generated by using restriction enzymes with high cut frequency), and to conduct the hybridization on Affymetrix tiling arrays characterized by very high feature density. In the primer design, we considered that while the evolutionary old *Alu* subfamilies constitute the majority of *Alu* elements, almost all of the active and polymorphic *Alu* elements belong to one of the less numerous, 'young' subfamilies of the *Alu* Y lineage [5]; and that among them, the Ya lineage (including Ya5, Ya8, Ya4) includes most of the known disease-causing de novo *Alu* insertions [15,30]. Consequently, we aimed to obtain a good sensitivity and specificity for the subfamilies of the *Alu* Y lineage and in particular (with AIP2) for the *Alu* Ya lineage. The results show that both methods, but especially AIP2, provide a high probability to detect *Alu* Ya lineage members when they are flanked by a genomic region well represented by array features; importantly, the subfamily detected with the second best sensitivity is the other frequently polymorphic and active "young" *Alu* subfamily Yb [5]. Conversely, the fact that the members of old *Alu* families are detected only in a fraction of cases contributes to the reduction of the number of "useless" signals (noise).

By the "two sample analysis" we found, as expected, a series of genomic sites already known to be polymorphic for *Alu* insertions, and other loci possibly associated with new *Alu* insertions. The choice to check and to characterize the 11 new "candidate" insertions located within known genes was based on their possible functional significance. While each AIP method finally led to the detection of 3 confirmed new intragenic insertions, the use of both methods increased to 4 the total number of such insertions (two of them obtained by both methods). Importantly, the methods were revealed to be sufficiently sensitive to detect polymorphic sites even in case of homozygote vs. heterozygote comparison. The four newly discovered *Alu* insertions turned out to be common intragenic polymorphisms in known genes, and their clinical relevance should be verified, considering that intragenic *Alu* polymorphisms could have a functional role even when they are intronic [7]. In particular, the exonic *MLK4* polymorphism (GenBank ID: **JN391997**) can be of interest in the study of colorectal colon cancer-associated genetic variability [31]; while the *Alu* insertion GenBank ID: **JN392000**, located in the intron 3 of *PARK2* corresponding to a recombination hotspot prone to rearrangements involving *Alu* sequences [27], should be considered in the study of the genetic bases of juvenile Parkinson's disease [32].

Finally, we should be reminded that, albeit AIP1 and AIP2 are the first published protocols to detect *Alu* insertions based on microarray analysis, a method with the same aim but based on high-throughput sequencing has been recently presented by Witherspoon and co-authors [33]. As evidenced by a recent review [20], sequence-based and array-based methodologies have their own advantages and disadvantages when they are used to detect retroelement-associated genomic variability. In particular, sequence-based methods are capable of higher throughput than array-based methods when many samples are pooled and analyzed simultaneously to make efficient use of the next-generation sequencing platforms; on the other hand, they are likely

to have a higher per-sample cost than array-based methods when no more than a few samples are simultaneously analyzed. Moreover, array-based methods can easily allow the a priori restriction of the result analysis (or to use specifically designed custom arrays) to the genomic regions of interest such as specific chromosomes, exonic regions or class of genes, hence reducing the computational effort and/or the overall cost; in addition, the result analysis of tiling arrays can easily allow the selection of the best detection thresholds for the series of data, consequently favoring sensitivity or specificity even based on the aims of the study. Finally, it is necessary to remark that both approaches (array-based and sequence-based methods) have problems in the detection of *Alu* insertions flanked by repetitive elements, because genomic sites composed of repetitive elements are not represented on tiling arrays, and at the same time they yield unmappable results if sequenced. However, there are differences due to the different principles on which the methods are based: sequencing methods are based on the sequencing of very short *Alu*-flanking genomic fragments, and hence the obtained sequences are not useful to map an *Alu* element when the *Alu* element is *immediately* flanked by (even short) repetitive elements; on the other side, AIP methods are based on the (hybridization) analysis of some hundred base pairs in the *Alu* flanking region, and hence they can still detect with good sensitivity *Alu* sequences *immediately* flanked by repetitive sequences *if* a portion of repetitive sequence is present not far away (within 500 bp). Consequently, in studies in which the completeness of the results is requested, the use of array-based and sequence-based approaches can be the best choice, in order to minimize the number of undetectable *Alu* elements.

With one of the paradoxes which are not rare in the history of science, the most abundant human genomic sequences only marginally benefited, up to now, from the striking scientific progress of the "genomic era". In fact, despite the decades-old discoveries of their nature of active, mutagenic and potentially deleterious retroelements and their impressive role in genome evolution, a rather fundamental question has yet to be answered: how big is the impact of the genome-resident *Alu* elements in the germline and somatic genomic variation, in human health and disease? In particular, the study of the role of *Alu* retranspositional activity in genetic diseases, somatic mosaicism and cancer can be still considered in its early days. Most of such (apparently surprising) delay has been likely due to the scarcity of efficient analytical methodologies applicable in this field of study. We hope that the *Alu* profiling methods presented here, together with the recent developments in next-generation sequencing, will facilitate the work of the increasing number of scientists that are, or will be, engaged in the difficult task of highlighting the "dark side" of the genome.

4. Materials and methods

4.1. Genomic DNA extraction

Blood was obtained by venipuncture from 12 healthy unrelated subjects from central Italy after an informed consent was obtained; genomic DNA was extracted from buffy coat using QIAamp DNA Blood Mini Kit (Qiagen) and its quantity and quality verified using gel electrophoresis and nanodrop spectrophotometric measurement. Samples were diluted at 72 ng/μl. The two DNA samples used for the AIP analysis are hereafter indicated as A and B.

4.2. Alu insertion profiling methodology

Two AIP strategies have been tested, both based on the use of Affymetrix tiling microarrays and designed to preferentially target genomic sites associated to some of the young (active) *Alu* subfamilies. The two strategies differ in the method used to obtain the probes (labeled *Alu*-flanking DNA fragments) for microarray hybridization. In the first method (AIP1) the specificity is based on the use of a primer ("Ya5R") which has perfect complementarity with the consensus sequence of most of the *Alu* subfamilies of the Y class, including Ya subfamilies (Ya5, Ya8, Ya1, Ya4), but not Yb subfamilies (Yb8 and Yb9); consensus sequences of older *Alu* families are not perfectly complementary to this primer and, in particular, the most abundant old *Alu* families do not offer a good annealing due to an insertion of two bases positioned near the 3′ end of the annealing site (the former is true, in particular, for Sx, Jo, Jb families, while Sg and Sc consensus sequences differ only in one base from this primer). In AIP2 the Ya5R primer is also used, but we added a previous primer extension step which uses a different primer ("Ya rev", complementary to Ya5, Ya8, Ya4) to further enrich the probe with DNA sequences flanking the *Alu* Ya subfamily members. Fig. 3 illustrates the steps of the two methods.

4.2.1. Enzymatic digestions

Each aliquot of genomic DNA was digested in parallel with three restriction endonucleases. The enzymes were chosen to give a frequency of cut on the human genome of 1/250 to 1/500 bp; the production of short fragments for the probe preparation is necessary due to the high density of *Alu* sequences on the genome: *Alu* density is on average 10.8% [34], corresponding (considering a mean *Alu* length of 300 bp) to mean inter-*Alu* genomic sequences of about 2500 bp; but some genomic regions have an *Alu* density higher than 40% [35], corresponding to mean inter-*Alu* spacers as short as 450 bp. In detail, 2 μg of genomic DNA were digested in distinct reactions with SfcI (25 U of enzyme, reaction for 12 h at 25 °C followed by 1 h at 37°) AluI (25 U of enzyme, reaction for 16 h at 37°) and DdeI (25 U of enzyme, reaction for 16 h at 37 °C) enzymes in reaction volumes of 50 μl, followed by enzyme inactivation by holding for 20 min at 65 °C. All enzymes were provided by New England Biolabs, and the reactions were conducted using the specific buffers (and bovine serum albumin solution for SfcI) indicated by the producer.

4.2.2. Ligation of vectorette linkers

The DNA fragments generated by restriction endonucleases were ligated, in separate reactions, to vectorette linker oligonucleotides [36]. Each linker was composed of a constant reverse strand (Linker_R_JB9408: CTC TCC CTT CTC GAA TCG TAA CCG TTC GTA CGA GAA TCG CTG TCC TCT CCT TC) pre-annealed with a forward strand which is different for each reaction and designed to match the ends generated by each enzyme (Linker_F_SfcI: TRY AGA AGG AGA GGA CGC TGT CTG TCG AAG GTA AGG AAC GGA CGA GAG AAG GGA GAG; Linker_F_AluI: GAA GGA GAG GAC GCT GTC TGT CGA AGG TAA GGA ACG GAC GAG AGA AGG GAG AG; Linker_F_DdeI: TNA GAA GGA GAG GAC GCT GTC TGT CGA AGG TAA GGA ACG GAC GAG AGA AGG GAG AG).

Each ligation reaction was conducted in 25 μl reactions using 0.8 μg (20 μl) of digested DNA sample, 100 pmol of preannealed vectorette double strand linker, 400 U of T4 DNA ligase (New England Biolabs), 0.5 μl of T4 DNA Ligase reaction buffer (New England Biolabs), 1 mM final ATP concentration; ligation was conducted for 12 h at 16 °C, 2 h at 25 °C, 20 min at 65 °C. Ligation products were purified using MinElute Reaction Cleanup columns (Qiagen) to eliminate unlinked linker oligonucleotides and directly used for the step 3 (AIP 1 method), or treated (AIP2) with an additional enrichment step (2b).

4.2.3. Only for the AIP2

Purified ligated DNA samples were subjected to a primer extension step aimed to enrich the DNA fragments which flank *Alu* Y elements. Mix containing 1× PCR buffer, 1.5 mM Mg2+, 50 μM dATP, dTTP, dGTP, 40 μM dCTP, 10 μM dCTP biotin, 10 pmol of the primer Ya_rev (ACC GTT TTA GCC GGG A), template DNA 500 ng, Taq polymerase 1 U, H_2O to 25 μl. Denaturation 95 °C, 12 min (Taq added

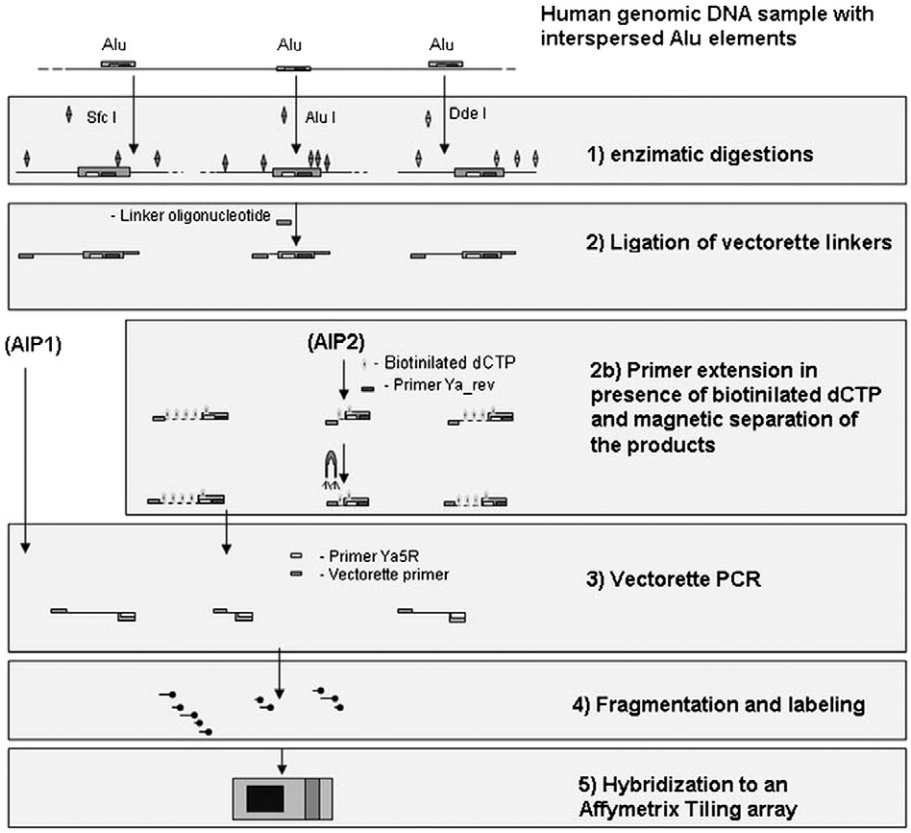

Fig. 3. *Alu* profiling methods. 1) enzymatic digestions in 3 separate reactions with SfcI, AluI and DdeI restriction enzymes; rhombi represent restriction sites; 2) Ligation of oligo linkers to the digested DNA fragments; 2b) Primer extension in presence of biotinylated dCTP using a primer complementary to *Alu* Ya elements and magnetic separation of primer extension products (only for AIP2 method); 3) Vectorette PCR with a primer complementary to part of the *Alu* Y consensus sequence and a primer complementary to the linker oligo; 4) After having pooled the 3 kinds of PCR products obtained for each genomic DNA sample (one for each of the 3 initial DNA restriction reactions and the subsequent steps), the DNA is fragmented and labeled to produce the hybridization probe; 5) The probe is hybridized to an Affymetrix 2.0R A tiling array.

in the last 2 min of denaturation); annealing 55 °C 20 min; extension 72 °C, 20 min. After subtraction of unincorporated biotin-dCTP through MinElute Reaction Cleanup Kit (Qiagen), primer extension products have been purified by magnetic separation using M-280 streptavidin Dynabeads.

4.2.4. Vectorette PCR

Reactions were conducted in 50 µl, using EuroTaq reaction buffer (Euroclone) 1×, vectorette primer (CTC TCC CTT CTC GAA TCG TAA) 20 pmol, primer Ya5R (TCT CGA TCT CCT GAC CTC GT) 20 pmol, dCTP 0.2 mM, dATP 0.2 mM, dGTP 0.2 mM, dTTP 0.16 mM, dUPT 0.04 mM, Mg^{2+} 2.5 mM, DNA sample from step 2 (AIP1) or from step 2b (AIP2) 160 ng, EuroTaq (Euroclone) 5 U. PCR cycles: 5 min 94 °C, (30 s 94 °C, 30°s 61 °C, 1 min 72 °C) 42 cycles, 7 min 72 °C. For each sample 3 TIP-chip PCR reactions were conducted, one for each digestion/ligation. The enrichment was checked using RT-PCR for a specific locus containing an *Alu* Ya5 element compared to two single copy genes.

4.2.5. Fragmentation and labeling

After each TIP-chip PCR, products were purified through MinElute Reaction Cleanup Kit (Qiagen). 3 µg of the three reactions for each genomic DNA sample were pooled and treated for fragmentation and labeling using the Gene Chip WT Double Stranded DNA Terminal Labeling kit (Affymetrix) following manufacturer's instructions.

4.2.6. Hybridization and scanning

Labeled samples were hybridized to Tiling arrays 2.0RA (Affymetrix), using Affymetrix 640 hybridization oven, Affymetrix 450 fluidic station, hybridization buffers contained in the Gene chip Hybridization, Wash and Stain Kit (Affymetrix), and following manufacturer's instructions. Arrays were scanned using an Affymetrix Gene Chip scanner 3000.

4.3. Tiling microarray analysis

The efficacy of the two used AIP methods was tested by analyzing genomic DNA samples from A and B DNA samples, using the "Affymetrix Tiling Array 2.0R A" for chromosomes 1 and 6. Affymetrix Tiling Analysis Software (TAS) was used for the analysis.

Two kind of analysis were performed: a single sample analysis in which results for A and B samples were analyzed separately, and a "two sample" analysis in which the A and B array signals were directly compared to find "different" signals. For the single sample analysis, after the analysis of signal intensity an interval analysis was performed to detect stretch of features (of at least 70 bp) above the detection threshold. Thresholds of 3.1 or 3.4 for signal intensity were used. Similarly, the "two sample analysis" was conducted in two steps consisting, respectively, of a two sided probe analysis in which for each feature of the array a p-value was obtained to test the hypothesis of different signal intensities between sample A and sample B, and an interval analysis aimed to identify a series of consecutive features (of at least 70 bp) yielding significantly different signals. A signal threshold of 32 for the p-value (on a −10log10 scale) was used. The obtained "bed" files containing the detected intervals on chromosome 1 and 6 were then analyzed using the Galaxy software [37] to analyze the position of significant intervals with respect to annotated genomic features (known genes, known *Alu* elements, known *Alu* polymorphisms). An example of AIP results (signal intensity for A and B samples and intervals of significantly different signals) visualized by Integrated Genome Browser [38] is

shown in Supplementary Fig. 3. The loci detected by "two sample analysis", corresponding to significant intervals located far (more than 500 bp) from known *Alu* insertions, and located inside known genes, were verified by locus specific-PCR and DNA sequencing. Oligonucleotide primers used for locus-specific PCR and sequencing are reported in Supplementary File 1. Primers and sequencing service were provided by MWG (Eurofins MWG Gmbh, Ebersberg).

Supplementary data to this article can be found online at doi:10.1016/j.ygeno.2012.03.005.

Acknowledgments

This work was supported by a project grant by the Fondazione per la Ricerca sul Cancro Fernanda e Gaudenzio Renzi, Largo Sarnano 12, Ancona, Italy.

References

[1] R.J. Britten, Transposable element insertions have strongly affected human evolution, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 19945–19948.
[2] W.M. Chu, R. Ballard, B.W. Carpick, B.R. Williams, C.W. Schmid, Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR, Mol. Cell. Biol. 18 (1998) 58–68.
[3] S.L. Ponicsan, J.F. Kugel, J.A. Goodrich, Genomic gems: SINE RNAs regulate mRNA production, Curr. Opin. Genet. Dev. 20 (2010) 149–155.
[4] R.D. Walters, J.F. Kugel, J.A. Goodrich, InvAluable junk: the cellular impact and function of Alu and B2 RNAs, IUBMB Life 61 (2009) 831–837.
[5] M.A. Batzer, P.L. Deininger, Alu repeats and human genomic diversity, Nat. Rev. Genet. 3 (2002) 370–379.
[6] H.K. Hamdi, R. Castellon, A genetic variant of ACE increases cell survival: a new paradigm for biology and disease, Biochem. Biophys. Res. Commun. 318 (2004) 187–191.
[7] M. Putku, K. Kepp, E. Org, S. Sõber, D. Comas, M. Viigimaa, G. Veldre, P. Juhanson, P. Hallast, N. Tõnisson, HYPertension in ESTonia (HYPEST), S. Shaw-Hawkins, M.J. Caulfield, BRItish Genetics of HyperTension (BRIGHT), E. Khusnutdinova, V. Kožich, P.B. Munroe, M. Laan, Novel polymorphic AluYb8 insertion in the WNK1 gene is associated with blood pressure variation in Europeans, Hum. Mutat. 32 (2011) 806–814.
[8] P.A. Callinan, M.A. Batzer, Retrotransposable elements and human disease, Genome Dyn. 1 (2006) 104–115.
[9] M. Dewannieux, C. Esnault, T. Heidmann, LINE-mediated retrotransposition of marked Alu sequences, Nat. Genet. 35 (2003) 41–48.
[10] M.K. Konkel, M.A. Batzer, A mobile threat to genome stability: the impact of non-LTR retrotransposons upon the human genome, Semin. Cancer Biol. 20 (2010) 211–221.
[11] V.P. Belancio, A.M. Roy-Engel, R.R. Pochampally, P. Deininger, Somatic expression of LINE-1 elements in human tissues, Nucleic Acids Res. 38 (2010) 3909–3922.
[12] H. Xie, M. Wang, F. Bonaldo Mde, C. Smith, V. Rajaram, S. Goldman, T. Tomita, M.B. Soares, High-throughput sequence-based epigenomic analysis of Alu repeats in human cerebellum, Nucleic Acids Res. 37 (2009) 4331–4340.
[13] M. Provinciali, A. Barucca, M. Cardelli, F. Marchegiani, E. Pierpaoli, Inflammation, aging, and cancer vaccines, Biogerontology 11 (2010) 615–626.
[14] M.A. Batzer, P.L. Deininger, U. Hellmann-Blumberg, J. Jurka, D. Labuda, C.M. Rubin, C.W. Schmid, E. Zietkiewicz, E. Zuckerkandl, Standardized nomenclature for Alu repeats, J. Mol. Evol. 42 (1996) 3–6.
[15] P. Deininger, Alu elements: know the SINEs, Genome Biol. 12 (2011) 236.
[16] D. Sinnett, J.M. Deragon, L.R. Simard, D. Labuda, Alumorphs human DNA polymorphisms detected by polymerase chain reaction using Alu-specific primers, Genomics 7 (1990) 331–334.
[17] M. Cardelli, Alu PCR, Methods Mol. Biol. 687 (2011) 221–229.
[18] A.M. Roy, M.L. Carroll, D.H. Kass, S.V. Nguyen, A.H. Salem, M.A. Batzer, P.L. Deininger, Recently integrated human Alu repeats: finding needles in the haystack, Genetica 107 (1999) 149–161.
[19] J.M. Chen, C. Férec, D.N. Cooper, LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease: mutation detection bias and multiple mechanisms of target gene disruption, J. Biomed. Biotechnol. 2006 (2006) 56182.
[20] D.A. Ray, M.A. Batzer, Reading TE leaves: new approaches to the identification of transposable element insertions, Genome Res. 21 (2011) 813–820.
[21] A. Gabriel, J. Dapprich, M. Kunkel, D. Gresham, S.C. Pratt, M.J. Dunham, Global mapping of transposon location, PLoS Genet. 2 (2006) e212.
[22] S.J. Wheelan, L.Z. Scheifele, F. Martínez-Murillo, R.A. Irizarry, J.D. Boeke, Transposon insertion site profiling chip (TIP-chip), Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 17632–17637.
[23] M. Cardelli, F. Marchegiani, C. Franceschi, F. Lattanzio, M. Provinciali, Alu insertion site profiling in the human genome, N. Biotechnol. 2 (2010) s38.
[24] A. Bardelli, D.W. Parsons, N. Silliman, J. Ptak, S. Szabo, S. Saha, S. Markowitz, J.K. Willson, G. Parmigiani, K.W. Kinzler, B. Vogelstein, V.E. Velculescu, Mutational analysis of the tyrosine kinome in colorectal cancers, Science 300 (2003) 949.
[25] J.L. McCauley, R.L. Zuvich, Y. Bradford, S.J. Kenealy, N. Schnetz-Boutaud, S.G. Gregory, S.L. Hauser, J.R. Oksenberg, D.P. Mortlock, M.A. Pericak-Vance, J.L. Haines, Follow-up examination of linkage and association to chromosome 1q43 in multiple sclerosis, Genes Immun. 10 (2009) 624–630.
[26] H. Tagawa, I. Miura, R. Suzuki, H. Suzuki, Y. Hosokawa, M. Seto, Molecular cytogenetic analysis of the breakpoint region at 6q21-22 in T-cell lymphoma/leukemia cell lines, Genes Chromosomes Cancer 34 (2002) 175–185.
[27] J. Mitsui, Y. Takahashi, J.H. GotoTomiyama, S. Ishikawa, H. Yoshino, N. Minami, D.I. Smith, S. Lesage, H. Aburatani, I. Nishino, A. Brice, N. Hattori, S. Tsuji, Mechanisms of genomic instabilities underlying two common fragile-site-associated loci, PARK2 and DMD, in germ cell and cancer cell lines, Am. J. Hum. Genet. 87 (2010) 75–89.
[28] F. Hormozdiari, C. Alkan, M. Ventura, I. Hajirasouliha, M. Malig, F. Hach, D. Yorukoglu, P. Dao, M. Bakhshi, S.C. Sahinalp, E.E. Eichler, Alu repeat discovery and characterization within human genomes, Genome Res. 21 (2011) 840–849.
[29] C.R. Huang, A.M. Schneider, Y. Lu, T. Niranjan, P. Shen, M.A. Robinson, J.P. Steranka, D. Valle, C.I. Civin, T. Wang, S.J. Wheelan, H. Ji, J.D. Boeke, K.H. Burns, Mobile interspersed repeats are major structural variants in the human genome, Cell 141 (2010) 1171–1182.
[30] P.L. Deininger, M.A. Batzer, Alu repeats and human disease, Mol. Genet. Metab. 67 (1999) 183–193.
[31] R.X. Shao, N. Kato, L.J. Lin, R. Muroyama, M. Moriyama, T. Ikenoue, H. Watabe, M. Otsuka, B. Guleng, M. Ohta, Y. Tanaka, S. Kondo, N. Dharel, J.H. Chang, H. Yoshida, T. Kawabe, M. Omata, Absence of tyrosine kinase mutations in Japanese colorectal cancer patients, Oncogene 26 (2007) 2133–2135.
[32] D.M. Kay, C.F. Stevens, T.H. Hamza, J.S. Montimurro, C.P. Zabetian, S.A. Factor, A. Samii, A. Griffith, J.W. Roberts, E.S. Molho, D.S. Higgins, S. Gancher, L. Moses, S. Zareparsi, P. Poorkaj, T. Bird, J. Nutt, G.D. Schellenberg, H. Payami, A comprehensive analysis of deletions, multiplications, and copy number variations in PARK2, Neurology 75 (2010) 1189–1194.
[33] D.J. Witherspoon, J. Xing, Y. Zhang, W.S. Watkins, M.A. Batzer, L.B. Jorde, Mobile element scanning (ME-Scan) by targeted high-throughput sequencing, BMC Genomics 11 (2010) 410.
[34] D. Grover, M. Mukerji, P. Bhatnagar, K. Kannan, S.K. Brahmachari, Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition, Bioinformatics 20 (2004) 813–817.
[35] M. Cardelli, F. Marchegiani, L. Cavallone, F. Olivieri, S. Giovagnetti, E. Mugianesi, R. Moresi, R. Lisa, C. Franceschi, A polymorphism of the YTHDF2 gene (1p35) located in an Alu-rich genomic domain is associated with human longevity, J. Gerontol. A Biol. Sci. Med. Sci. 61 (2006) 547–556.
[36] J. Riley, R. Butler, D. Ogilvie, R. Finniear, D. Jenner, S. Powell, R. Anand, J.C. Smith, A.F. Markham, A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones, Nucleic Acids Res. 18 (1990) 2887–2890.
[37] D. Blankenberg, G.V. Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, J. Taylor, Galaxy: a web-based genome analysis tool for experimentalists, Curr. Protoc. Mol. Biol. 89 (2010) 19.10.1–19.10.21.
[38] J.W. Nicol, G.A. Helt, S.G. Blanchard, A. Raja, A.E. Loraine, The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets, Bioinformatics 25 (2009) 2730–2731.