An Alu element retroposition in two families with Huntington disease defines a new active Alu subfamily

Gordon B. Hutchinson, Susan E. Andrew, Helen McDonald, Y. Paul Goldberg, Rona Graham, Johanna M. Rommens and Michael R. Hayden

Department of Medical Genetics, University of British Columbia, Vancouver, BC V6T 1Z4, and Department of Genetics, The Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada

Received May 10, 1993; Accepted June 8, 1993

ABSTRACT

Alu repetitive elements represent the most common short interspersed elements (SINEs) found in primates, with an estimated 500,000 members in the haploid human genome. Considerable evidence has accumulated that these elements have dispersed in the genome by active transcription followed by retroposition, and that this process is ongoing. Sequence variation between the individual elements has led to the hierarchical classification of Alu repeats into families and subfamilies. Young subfamilies that are still being actively transposed are of considerable interest, and the identification of one such subfamily (designated 'PV') has led to the hypothesis that the most recent retroposition events are due to a single master Alu source gene. In the course of our search for the gene causing Huntington disease, we have detected an Alu retroposition event in two families. Sequence analysis demonstrates that this Alu element is not a member of the PV subfamily, but is similar to 5 other Alu elements in the GenBank database. Together, these Alu elements, all of which contain a 7 base-pair internal duplication, define a distinct subfamily, designated as the Sb2 subfamily, providing evidence for a second actively retroposing Alu source gene. These data provide support for multiple source genes for Alu retroposition in the human genome.

INTRODUCTION

There are approximately 500,000 Alu repeat elements in the haploid human genome, comprising in total 5% of human DNA. These repeats are about 300 nucleotides in length, have a dimeric structure, and appear to have arisen as a duplication of a small untranslated RNA similar to the 7SL RNA signal recognition particle (1). Comparative studies of primates demonstrate that while many Alu repeats were present in their current locations prior to primate speciation, others have inserted more recently (2). There is convincing evidence that Alu elements have spread by insertion throughout the genome through the mechanism of retroposition. Evidence for this derives from the recognition that the first half of the Alu sequence contains a polymerase III promoter, and that Alu insertions are flanked by direct repeats of the host sequence, consistent with insertion into staggered nicks in the host DNA sequence (1). Moreover, the terminal 3' ends have tracts of adenosine, consistent with a history of reverse transcription through a polyadenylated mRNA intermediate (1).

There are specific nucleotide variations in individual Alu element that allow their classification into distinct subfamilies. At least five separate nomenclatures have been described (3,4,5,6,7). The age of each family has been estimated through the extent of divergence from the Alu element consensus. This includes the intrafamily variability due to random mutation following retroposition, the CpG dinucleotide content and the purity of elements' poly-A tracts. The most recently retroposed Alu elements have inserted following speciation, and therefore are not found in the orthologous location in the genomes of other primates. Furthermore, there may be polymorphism within a species, so that a particular Alu element may be present in some individuals and absent in others.

All of the recently-inserted Alu elements that have been so far described are members of one family, variously termed Conserved, Class IV, A, CS or Sb (3,4,5,6,7). A member of the Sb class has been found in the Gorilla β-globin locus, but not in the orthologous location in the chimpanzee and humans (8), while another member is present near the human α-fetoprotein locus but is absent at the equivalent location in the gorilla (9). Some members of the Sb family have been further sub-classified as the 'PV' (10) subfamily. This family was previously thought to be specific to humans, and has also been described as the 'human specific' (HS) (11) subfamily. However, a member of this family has recently been identified in DNA from other primates (12), and therefore we prefer the PV classification.

Some members of the PV subfamily are polymorphic in the human genome, including an Alu element within the tissue plasminogen activator (TPA) locus (13) and at the Mlvi-2 locus (a putative oncogene) (14). Other PV members have been found to be transcriptionally active (15), a necessary pre-requisite for retroposition. Evidence of ongoing insertion of PV family members into the human genome derives from the finding of...
disease-related de novo mutations caused by PV Alu reposition in the Factor IX (FIX) gene (16) and the neurofibromatosis type 1 (NF1) gene (17).

The current predominance of the PV Alu subfamily in recent reposition events has led to two different theories to explain their origins (figure 1). The 'master gene model' proposes that a master gene locus is responsible for the amplification of all subfamilies of Alu sequences. The copies that this master gene creates are not themselves active in reposition. With this model, different subfamilies are created by mutations of the master source gene, and not by the creation of new master genes at other locations (fig. 1A) (18). As a consequence each new Alu element will have sequence characteristics of all earlier versions of the master source gene, with the addition of nucleotide changes unique to its specific subfamily. As such, all Alu subfamilies will be serially related. An alternative explanation, the 'transposon model', postulates that multiple Alu elements are, or were, potential sources for ongoing reposition, but that some have been more successful than others (19). With this model one would expect to identify Alu subfamilies that share some features of a common ancestor, but have other unique sequence changes that are inconsistent with a single master as origin for each subfamily (fig 1B).

In the course of our search for the gene causing Huntington disease, we have discovered an Alu reposition event in the Huntington disease region (4p16.3) that segregates with the disease in two families (20). Here we present evidence that this Alu element is a member of the Sb subfamily, but does not have the nucleotide changes consistent with it being a PV Alu. Sequence characteristics and comparison with other sequences in the GenBank database suggest that it has been transcribed from a source gene not serially related to the PV class, thus favouring the transposon model of Alu element reposition.
MATERIALS AND METHODS

Identification of transcribed sequences

A direct cDNA selection strategy, termed 'Gene Tracking', was used to identify transcripts from the minimal region containing the Huntington disease (HD) gene. Complementary DNA clones were physically mapped and classified into transcription units by cross-hybridization to each other and to RNA from various tissue types. A full description of the technique has been published elsewhere (21,22).

Sequencing

Sequencing was performed using fluorescent dye terminators and an ABI 373 automated sequencer. Sequence data were edited and assembled using the TED and XDAP software available in the Staden package (23).

Database searches

The Flat DB E-mail Network Server (E-mail address: flatnetserv@smlab.eg.gunma-u.ac.jp) implementing the FASTA algorithm (24), and the BLAST (25) e-mail server (E-mail address: blast@ncbi.nhn.nih.gov) were used for searches of the nucleic acid databases.

RESULTS AND DISCUSSION

Identification of an Alu insertion in the Huntington disease region

The cDNA clones from the Huntington region were used to search for disease-associated rearrangements. Using one of these clones (GT 48), we detected an insert of approximately 330 base pairs in two of 250 patients (20). This rearrangement segregated with Huntington disease in both of the families of these patients. Further cloning and sequencing identified this insertion as an Alu element located within an intron of the α-2-adducin gene between markers D4S95 and D4S182. The insertion occurred at exactly the same nucleotide position in both patients, was flanked by identical 9 base pair direct repeats and was not seen in 1000 control chromosomes. We identified 14 of 687 (2%) control chromosomes that had the same core haplotype extending about a megabase from D4S95 to D4S98. None of these control chromosomes had this rearrangement. Further history revealed that the ancestors of these two patients' families had lived in two Scottish towns 50 km. apart, strongly suggesting that this insertion represents a single retroposition event.

Alignment of the HD-associated Alu element with known active Alu elements

Previous reports of PV Alu insertions have been described in the neurofibromatosis (17) and factor IX (16) genes. There is also a report of an Alu insertion that inactivated the cholinesterase (CHE) gene (26) The HD Alu was aligned with these known Alu insertions, as well as the consensus sequences for PV and Sb Alu elements and the general Alu consensus (Fig. 2). It is evident from the alignment that, while the HD and CHE Alu retroposons are members of the Sb subfamily, they lack all 5 nucleotide substitutions diagnostic of the PV class (27). Moreover, they are closely related to each other, sharing several nucleotide substitutions as compared to the Sb and PV groups. Of particular interest is a 7 base-pair insertion following nucleotide 250. This insertion appears to have been created by a C to G transversion at position 249 followed by a 7 base-pair internal tandem duplication.

We conducted FASTA and BLAST searches of the GenBank database using a 20 base-pair segment containing the 7 base-pair duplication (GCCTGGCTMCATGCTGAAACGCTG) and found the highest-scoring matches to be the CHE insertion and four other Alu sequences. These four Alu elements were located in the 3' UTR of the human low-density lipoprotein receptor (LDLR) gene, near the 3' end of human lecithin-cholesterol acyltransferase (LCAT) gene, near the 5' end of the biliary glycoprotein (BGP) gene, and in a large cosmids contig from chromosome 4 near D4S113, HDAC (accession numbers L00352, X04981, X67277 and M63480, respectively). These four Alu insertions, together with the retropositions in the CHE gene and in the two patients with HD share the substitution at position 249 followed by the 7 base-pair duplication of nucleotides 246 to 252, as well as 6 other nucleotide changes (positions 57, 100, 145, 212, 237 and 253) that clearly distinguish them from other Sb subfamily members (figure 2). It is extremely unlikely that these multiple correlated changes could have occurred independently following retroposition, and strongly favours the hypothesis that these changes were present in the source gene from which these sequences were transcribed. These sequence changes are the diagnostic features of a new Alu subfamily, designated Sb2 (28). The Sb2 Alu family shares 8 of the 9 diagnostic characteristics of the Sb subfamily, but none of the diagnostic features of the PV subfamily, indicating that elements of the Sb2 Alu subfamily were derived from the Sb group as copies of a source gene distinct from the source gene for the PV class of Alu elements. The Sb2 and PV groups do not therefore appear to be serially related (figure 1).
by tandem duplications. The PYTHIA e-mail server (7), however, identifies the adjacent Alu elements as members of other subfamilies (Sx and Sp), implying that the three adjacent Alu elements are the result of separate retroposition events.

The Sb2 Alu family member in the LDLR gene is situated in the 3' untranslated region and, while not known to cause mutation directly by its retroposition, has nevertheless been noted to recombine with another Alu element in the last intron of that gene, thereby deleting the last exon, which codes for the cytoplasmic domain of the protein. The resulting defective receptor is unable to internalize LDL, leading to familial hypercholesterolemia (30).

How old is the Sb2 subfamily?

Table 1 shows a matrix summarizing the number of changes noted between several Alu insertions in the PV and Sb2 subfamilies when compared with each other, their family consensus and the general and Sb family consensus. Insertions, deletions, duplications and substitutions are each counted as one change. When estimating the relative age of sequences, the CpG positions were considered as uninformative, since transitions from CpG to TpG and from CpG to CpA can occur from deamination following methylation, and this process occurs more frequently than other substitutions. Table 1 shows the number of CpG changes in parentheses. There are 7 informative changes between the Sb2 and Sb subfamilies, whereas there are only 4 differences between the PV consensus and the Sb. This suggests that the Sb2 subfamily diverged earlier than the PV Alu elements. The average age of individual PV Alu element insertions has previously been estimated at 2.7 million years, by assuming a divergence from the PV consensus following retroposition at a rate of 0.15% per million years (11). The 6 Sb2 subfamily members identified here have a total of 11 non-CpG differences from the Sb consensus over 300 nucleotides (0.61%). Using the same procedure, the calculated average age of Sb2 Alu element insertions is 4.1 million years. This figure should be interpreted with caution given the small number of Sb2 Alu elements available for analysis. The LDLR Alu element has also accumulated 4 CpG mutations, and is noted to have 3 further substitutions in its poly-A tract and 2 in its direct repeat which suggests that it is the oldest member of the 4 Sb2 subfamily members so far identified.

It is relevant to consider whether the Sb2 subfamily is still active. The presence of Sb2 members in two isolated families (in the HD patients) and our failure to identify the same insertion in 1000 control chromosomes or in 248 other HD patients argues that the insertion on the HD chromosome took place relatively recently. Further evidence for this being a recent event is the absence of the rearrangement in all 14 control chromosomes with the same core haplotype.

Northern analysis of HeLa cell RNA has previously shown transcription of PV subfamily members (10). The evidence thus suggests that both the Sb2 subfamily and the PV subfamily are transcriptionally active.

New Alu retroposition events and models of retroposition

We have identified a recent Alu retroposition event in the 4p16.3 region, approximately 200 kilobases from the recently identified HD defect (32). Sequence characteristics of this Alu insertion establish it as a member of the Sb subfamily, but with nucleotide changes that make it distinct from the PV subgroup of the Sb family. This provides evidence for the existence of at least two separate Alu source genes for the human population that have diverged from the Sb consensus.

Delineation of the subfamily structure of Alu elements provides insight into the mechanisms of Alu element propagation in the primate genome. Although this repetitive DNA exists in a high copy number, Alu retroposition is a rare event. A study of 14 Alu sequences revealed no differences in their location in human, chimpanzee and gorilla (33,34), and most members of the PV subfamily that have that have been studied in search of insertion polymorphism have been found to be fixed in the human genome (35). Since evidence points to the PV subfamily being the youngest so far identified, having originated only shortly before human divergence from the great apes, it follows that only a small fraction (perhaps as few as 500 to 2000) (10,36) of the 500,000 Alu elements in the haploid human genome have inserted since the divergence of human, chimpanzee and gorilla, about 5-10 million years ago (37).

The sequence analysis of an Alu element retroposed in two families with HD has lead to the identification of a new active Alusubfamily, 'Sb2', with a unique internal duplication and other sequence changes which readily distinguish it from the PV Alu subfamily. This finding is inconsistent with the 'master gene' model of Alu retroposition, since the Sb2 subfamily is not serially related to the PV subfamily. This finding favours the multiple source gene, or 'transposon' model of Alu retroposition which has been previously advanced using other data (7,12,27,28). The source Alu elements for the Sb2 and PV subfamily may both be currently active in generating new Alu retroposition events.
ACKNOWLEDGEMENTS

This work is supported by the MRC (Canada), the Canadian Genetic Diseases Network and the Huntington Society of Canada. M.R.H. is an established investigator for the B.C. Children's Hospital.

NOTE ADDED

While this manuscript was undergoing review, we became aware of a paper in press (28) which describes the same Alu subfamily noted here. After contacting the author, we have therefore modified the original name we had chosen for this subfamily to match that used in the other reference (Sb2) to avoid confusion.

REFERENCES