The human albumin-α-fetoprotein genomic domain contains 13 repetitive DNA elements randomly distributed throughout the symmetrical structures of these genes. These repeated sequences are located at different sites within the two genes. The human albumin gene contains five Alu elements within four of its 14 intervening sequences. Two of these repeats are located in intron 2, and the remaining three are located in introns 7, 8, and 11. The human α-fetoprotein gene contains three of these Alu elements, one in intron 4 and the remaining two in the 3'-untranslated region. In addition, the human α-fetoprotein gene contains a Kpn repeat and two classes of novel repeats that are absent from the human albumin gene. Six of the Alu elements within the two genes are bound by short direct repeats that harbor five base substitutions in 120 possible positions (60 bp times 2 termini). The absence of Alu repeats from analogous positions in rodents indicates that these repeats invaded the albumin-α-fetoprotein domain <85 Myr ago (the time of mammalian radiation). Furthermore, considering the conservation of terminal repeats flanking the Alu sequences of the albumin-α-fetoprotein domain (0.042 changes per site), we submit that the average time of Alu insertion into this gene family could have been as recently as 15–30 Myr ago.

Introduction

The Alu family of repetitive DNA elements is a prominent family of short sequences that are dispersed in mammalian genomes (Houck et al. 1979). The structure of the human Alu element consists of a head-to-tail dimer of two similar sequences ~130 bp each. Each of the monomers is terminated with a poly-A-rich tract, and the right-hand monomer also contains a 30-bp insert not present in the left-hand monomer. The complete dimer is therefore ~300 bp in length. The presence of Alu repeats in the human genome, at ~300,000 copies (Houck et al. 1979), has led to the suggestion that they may function in some vital housekeeping role such as DNA replication (Jelinek et al. 1980) or processing of pre-mRNA transcripts (Calabretta et al. 1981). In contrast, it has also been suggested that the Alu family has no function and represents selfish DNA (Doolittle and Sapienza 1980; Orgel and Crick 1980). In line with the latter view, Alu’s high copy number could have resulted from an opportunistic mode of replication, and it may not necessarily reflect a vital role of Alu repeats.

On the basis of structural similarity, the origin of Alu elements can be traced to the gene for 7SL RNA (Ullu and Tschudi 1984). The abundant cytoplasmic 7SL RNA...
functions in protein secretion as a component of the signal-recognition particle (Walter and Blobel 1982). This particle consists of six different polypeptides and one molecule of 7SL RNA, and it mediates the translocation of secretory proteins across the cytoplasmic reticulum. Although the 7SL RNA has a well-defined biological function, that of the related Alu repeat remains unknown. Furthermore, the Drosophila genome contains two copies of the 7SL RNA gene and no Alu repeats. This has led to the suggestion that the 7SL RNA gene is a progenitor of a processed pseudogene, the Alu element, that has recently spread to other locations in the human genome (Ullu and Tschudi 1984).

The purpose of the present study was to determine the DNA sequence and the precise location of the Alu repeats in the human albumin-AFP gene family in an effort to understand Alu's possible biological role and evolutionary origin. Five of these repeats were found within the albumin and three within the α-fetoprotein gene. All were localized within intronic or 3' flanking regions. The asymmetric location of Alu repeats within the human genes, coupled with their apparent absence from analogous rodent genes, suggests that they have "recently" invaded the human genes.

Material and Methods

Cloning

The entire human serum-albumin gene (Dugaiczyk et al. 1982; Hawkins and Dugaiczyk 1982) and α-fetoprotein gene (Minghetti et al. 1983) were previously isolated as sets of λ clones. Various restriction DNA fragments of these clones were subcloned into pBR322 and pUC8 using standard techniques (Maniatis et al. 1982). The subclones were subsequently used for DNA sequencing.

DNA Sequencing

The complete primary structures of the human serum-albumin and α-fetoprotein genes were determined using the DNA-sequencing procedure of Maxam and Gilbert (1980). The sequencing strategy and the structure of each gene are presented elsewhere (Minghetti et al. 1986; Gibbs et al. 1987).

Results

Alu Elements in the Human Albumin Gene

The complete nucleotide sequence of the human albumin gene was scanned for repetitive sequences using the Larson and Messing (1983) computer program. One partial and four complete Alu repeats were found (fig. 1). Two of the Alu repeats were found in a tail-to-tail arrangement, separated by only 100 bp, within intron 2 (fig. 1). Alu elements were not found 2 kb upstream or 7 kb downstream from the human albumin gene. Four of the Alu repeats are complete (i.e., possess complete left and right halves), whereas the fifth, located in intron 11, consists of only two-thirds of the complete Alu dimer. Alu 4, located in intron 8, has a deletion of 8 bp compared with the Alu consensus sequence. The identity between the Alu repeats in the albumin gene and a human Alu consensus sequence (Sawada et al. 1985) ranges from 71% to 90%.

Alu Repeats in the Human α-Fetoprotein Gene

The complete sequence of the human α-fetoprotein gene (Gibbs et al. 1987) reveals three Alu elements within 22 kb of chromosomal DNA (fig. 1). One element is found in intron 4; it is a complete, dimeric structure and is 92% identical with the
Invasion of Human Albumin-α-Fetoprotein

FIG. 1.—Schematic representation of the human serum-albumin and α-fetoprotein genes. Each gene is composed of 15 exons (boxes) separated by 14 introns (heavy lines). In both genes, exons 1 and 14 are partially untranslated and exon 15 (open boxes) is entirely untranslated. Numbers below exons and above introns indicate their sizes in base pairs. The positions of landmark sequences at the ends of each gene are also shown. Arrows indicate the position and orientation of the Alu repeats, which are expanded fivefold over the scale of the gene. The other arrows in the AFP gene represent the X, Xba, and Kpn repeats of this gene; these repeats are absent from the albumin gene.

human Alu consensus sequence. The remaining two are located 3' of the polyadenylation site. Of these, one is a partial Alu element representing the 5' half of the usual dimeric structure. The sequence of the remaining element is incomplete because it is at the end of the cloned DNA; it will, therefore, not be considered further here. The human α-fetoprotein gene contains other repeated elements, which are also shown in figure 1. These include a Kpn repeat, and two pairs of novel, low-copy-number repeats, which we have designated X repeats and Xba repeats. None are present in the albumin gene.

Alu Elements are Bound by Direct Repeats

The four complete Alu elements in the human albumin gene, 1-4 (fig. 2), are bound by short terminal repeats of 7 bp (TATTTAA), 8 bp (TGTTGGG), 8 bp (TAAGAAGA), and 15 bp (GACAGATCTATT), respectively. The four terminal repeats bear no sequence relationship to one another. Alu 5 is missing 100 bp of its 5' terminus, and thus no terminal repeats could be assigned. The close proximity of this repeat to the 5' splice junction of exon 12 suggests that the missing one-third could have interfered with the correct splicing and was therefore eliminated by some unknown mechanism.

One of the two Alu elements in the human AFP gene is flanked by a perfect 11-bp repeat (GGATGTTGTGG), the other by an imperfect 11-bp sequence (CTTTGTTCT).

Alu Repeats in the Rat Albumin Gene

The noncoding regions of the rat serum-albumin gene (Sargent 1981) were examined for the presence of Alu repeats. The reported sequence represents only 8 kb of the 15 kb comprising the entire gene. Intron 11, the only rat intron sequenced in
Invasion of Human Albumin-α-Fetoprotein

5

its entirety, corresponds to an Alu-containing intron in the human gene. This intron is clearly devoid of any sequence similarity to the rodent type 1 or type 2 (Page et al. 1981; Jelinek and Haynes 1983) Alu consensus. Rat intron 2 is interrupted by two gaps in the sequence data, one 100 bp and the other 150 bp. The sequenced portion of this intron has no similarity to the rodent Alu elements, and the small gaps in the sequence data could not accommodate two Alu repeats present in human intron 2. Rat introns 7 and 8 have gaps in the available data of 850 and 600 bp, respectively, and hence the presence or absence of Alu repeats within them could not be ascertained. It is clear that Alu repeats are absent from rat introns 2 and 11, corresponding to human Alu’s 1, 2, and 5, as well as from the remaining reported (Sargent 1981) rat intronic sequence.

Alu Repeats in the Mouse α-Fetoprotein Gene

The complete structure of the mouse α-fetoprotein gene is not known, but it contains at least one rodent Alu repeat within intron 1 (Young et al. 1982). In contrast, we find no Alu repeats in the first intron of the human AFP gene.

Discussion

Recent Time of Invasion of the Human Albumin and AFP Genes by Alu Repeats

There are several arguments in support of a recent invasion of this gene family. (1) The location of these repeats within each of the human genes is compatible with a model of albumin’s triplication from an ancestral one-domain precursor (fig. 3) only if the repeats were inserted into the gene after the triplication was completed. (2) Since the Alu repeats are absent from the rat albumin gene, their insertion into the human gene must postdate the time of mammalian radiation, estimated at ∼85 Myr ago (Romero-Herrera et al. 1973). A similar argument applies to the Alu’s of the mouse and human AFP gene; since the repeats are present in different locations in the same gene of the two species, their insertion must also postdate the time of mammalian radiation. (3) The time of the Alu insertions can also be estimated from the drift between the 5’ and 3’ terminal repeats. These repeats represent intronic DNA (Shapiro 1979; Jagadeeswaran et al. 1981) and hence must evolve at a high rate characteristic of introns, even if the Alu sequences were conserved because of a hypothetical function. The rate of substitution for introns has been estimated as being $3 \times 10^{-9}$/site/year (Li and Gojobori 1983), or 0.003/site/1 Myr. At this rate, 13.9 Myr are required for five substitutions to occur in 120 sites of intronic sequence. Since there are 120 sites (60 of 5’ and 60 of 3’) of terminal repeats in the six human Alu’s, with five (4.2%) differences

![Fig. 2.—Comparison of Alu sequences located in the human albumin and α-fetoprotein genes. The top string of nucleotides represents the human Alu consensus sequence. Below the consensus sequence, individual Alu sequences derived from the human albumin (Alb-1 to Alb-4) or human AFP (AFP-1 to AFP-2) genes are indicated. All Alu sequences are arranged in the 5’ to 3’ orientation for convenience. Only nucleotide differences between each Alu sequence and the consensus sequence are indicated. Dots represent nucleotides identical to the consensus sequence, whereas dashes represent gaps. Numbers above the consensus sequence indicate the locations of nucleotides. The bottom portion of this figure shows the comparison of direct repeats flanking the Alu elements within the human albumin and α-fetoprotein genes. The repeats from top to bottom are from Alu 1 through Alu 4 of the albumin gene followed by Alu 1 and Alu 2 of the AFP gene. For clarity, all Alu repeats are written in the 5’ to 3’ direction irrespective of their orientation within the two genes. The asterisks indicate nucleotide differences between the 5’ and 3’ direct repeats flanking a single Alu element.](https://academic.oup.com/mbe/article-abstract/4/1/1/1244505)
FIG. 3.—Panel A: A model depicting the evolution of the albumin and α-fetoprotein genes. This figure was taken from Sargent et al. (1981), except that the exons are designated differently. Exons are indicated by boxes, introns by solid lines, and Alu repeats for the albumin gene by asterisks. The location of Alu repeats within the present three-domain structures of albumin and AFP is incongruent with this model in that the repeats could not predate the duplication of the ancestral one-domain gene. The same is true for the other types of repeats in the AFP gene. Panel B: Symmetry between genetic domains within the albumin and AFP genes. Introns, exons, and repeats are designated as in panel A. The identical lettering of exons reflects the homology inferred from the nucleotide sequence similarity as well as the almost identical length of corresponding exons in each of the three domains. The threefold symmetry of these genes is further exemplified by the intron/exon junctions, which are placed at identical codon positions in each domain. These codon positions are indicated by the numbers above the exons. Asterisks indicate nonsymmetrical positions of Alu elements in the human albumin gene.

among them, this would indicate that on average the Alu elements were inserted into the human albumin-AFP genes ~14 Myr ago. The numbers are admittedly small and could be in error by a factor of 2, but similar values of 5% and 8% for the divergence of terminal repeats were obtained from a larger number of primate Alu sequences (Fukumaki et al. 1983; Sawada et al. 1985).

Our proposal that Alu repeats are relatively young sequence elements in the albumin-AFP gene complex is supported by recent findings for the low-density lipoprotein (LDL)–receptor gene (Hobbs et al. 1985). The human gene was found to contain Alu elements at its 3' end, whereas neither the bovine nor baboon genes contain Alu repeats at this position. This indicates that the Alu elements were inserted at this location late in the evolution of primates.

Polyclonal Origin of the Alu Repeats

When individual Alu sequences in the human albumin-AFP gene complex are compared, the divergence among them (7%–34%) is considerably greater than that
among their terminal repeats (≈8%). These results could indicate that the Alu elements, inserted at the various sites, differed in sequence at the time of their insertion. In other words, contemporary Alu elements are conceivably of polyclonal origin.

Biological Role of Alu Repetitive DNA

It is unlikely that Alu repeats provide an essential biological function in light of the finding that they (1) appear to have arisen as a processed pseudogene from 7SL RNA (Ullu and Tschudi 1984), (2) are absent from the *Drosophila* genome, which nevertheless contains the 7SL gene, (3) are present in the human albumin gene but absent from homologous positions in the rat albumin gene, (4) are present as incomplete elements, as seen in the human albumin gene and in other human genes (Lee et al. 1984), and (5) are found in diverse locations, i.e., between genes (Bell et al. 1980; Shen et al. 1981) as well as in transcription units of genes (Nemer et al. 1984; Lehrman et al. 1985). Thus, there is little in common in the Alu theme to build a coherent hypothesis regarding their biological function. They reached the status of high copy number because they (1) can be transcribed from their own promoter by RNA polymerase III (Elder et al. 1981) and (2) conceivably can be reverse transcribed to undergo further rounds of reinsertion into the genome. They are then subject to evolutionary drift and may, in time, disappear into the genomic background.

Despite the apparent lack of a direct function, it has been argued that the preponderance of repetitive sequences may provide an overriding benefit by increasing genetic diversity that would drive nonhomologous recombination or inhibit gene conversion. Inhibition of gene conversion by Alu repeats has been suggested for the α1-α2 (Hess et al. 1983) and for the β-γ globin genes (Schimenti and Duncan 1984). This remains a possibility, but other effects of Alu repeats can be deleterious. A deletion of 5 kb at the 3' end of the LDL-receptor gene was shown to have occurred between two Alu repeats, resulting in a defective LDL receptor (Lehrman et al. 1985). Heterozygotes with this genotype are hypercholesterolemic, whereas homozygotes suffer heart attacks before the age of 20 years. In another example, a form of hereditary persistence of fetal hemoglobin was shown to be the result of a deletion of both the β- and δ-globin genes (Jagadeeswaran et al. 1982). This deletion also occurred at an Alu repeat.

Cataclysms of the Past?

The amplification and transposition of more than 10⁵ repetitive DNA elements into new genomic locations must have had a major impact on the structure and evolution of the genome. We submit that, at the time of occurrence, the spread of new DNA elements was, in effect, equivalent to a large increase in mutation rate. The pandemic spreading of such elements in the genome could have occurred during a relatively short period on the evolutionary time scale and could have seriously disrupted the reproductive equilibria of populations. At present, all seems quiet in our genome and we hear only periodic rumblings—e.g., in the form of deletions of our LDL receptors—of a more turbulent time in our evolution.

Acknowledgments

We thank Carol Boyd and Rita Zielinski for excellent technical help. We also thank Dr. Susumu Ohno for his critical evaluation of the manuscript and his very stimulating comments. This work was supported by NIH grant HD/GM 19281.
LITERATURE CITED


Masatoshi Nei, reviewing editor

Received April 14, 1986; revision received June 23, 1986.