Research

Molecular archeology of an SP100 splice variant revisited: dating the retrotranscription and *Alu* insertion events

Eric J Devor

Address: Molecular Genetics and Bioinformatics, Integrated DNA Technologies, 1710 Commercial Park, Coralville, Iowa 52241, USA.
E-mail: rdevor@idtdna.com

Published: 30 August 2001

Genome ***Biology*** 2001, **2(9)**:research0040.1–0040.6

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/9/research/0040

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 17 April 2001
Revised: 7 June 2001
Accepted: 24 July 2001

Abstract

Background: SP100 is a nuclear protein that displays a number of alternative splice variants. In Old World monkeys, apes and humans one of these variants is extended by a retroprocessed pseudogene, *HMG1L3*, whose antecedent gene is a member of the family of high-mobility-group proteins, HMG1. This is one of only a few documented cases of a retropseudogene being incorporated into another gene as a functional exon. In addition to the *HMG1L3* insertion, Old World monkey genomes also contain an *Alu* sequence within the last *SP100-HMG* intron. PCR amplification of the 3′ end of the *SP100* gene using genomic DNAs from human and New World and Old World monkey species, followed by direct sequencing of the amplicons has made dating the *HMG1L3* and *Alu* insertion events possible.

Results: PCR amplifications confirm that the *HMG1L3* retrotransposition into the *SP100* locus occurred after divergence of New World and Old World monkey lineages, some 35-40 million years ago. PCR amplification also shows that an upstream *Alu* sequence was inserted in the last *SP100-HMG* intron after divergence of the Old World monkey and ape lineages. Direct sequencing of the *Alu* in five Old World monkey species places the latter event at around 19 million years ago. Finally, ten single base mutations and one deletion in the *Alu* differentiate African from Asian Old World monkey species.

Conclusions: PCR and DNA sequence analysis of 'genetic fossils' such as retropseudogenes and *Alu* elements in primates give details as to the timing of such events and can reveal sequence features useful for other molecular phylogenetic applications.

Background

Retroprocessed pseudogenes, or retropseudogenes, are reverse transcripts of mature mRNAs retrotransposed to new locales within the genome [1]. Recently, these loci have received increasing attention [2]. Goncalves *et al.* [3] have shown that retropseudogenes are quite common in mammalian genomes; 23,000 to 33,000 are estimated to reside in the human genome. Studies of both point mutations [4] and indels (insertions/deletions) [5] in retropseudogenes have shown them to be excellent sources of background genetic information in a wide range of species. Thus, one of the emerging utilities of retropseudogenes is their role in providing markers for phylogenetic studies between species or between populations within species [6-10].

Among the retropseudogenes studied to date, the high-mobility-group (HMG) pseudogene *HMG1L3* is a member of a rare class in which all or part of the encoded protein is still expressed [11]. Seeler *et al.* [12] reported that the nuclear protein SP100 displays a number of alternative splice variants.

One of these, called SP100-HMG, is an 879 amino acid protein whose carboxy-terminal 170 residues bear a close similarity to the family of HMG proteins. Rogalla *et al*. [13] identified five retropseudogenes for which the antecedent gene is *HMG1*. Subsequently, Rogalla *et al*. [14] demonstrated that the carboxy-terminal extension of SP100-HMG is encoded by part of one of these *HMG-1* retropseudogenes. Denoted *HMG1L3*, this retrotranscribed copy was inserted at the 3′ end of the *SP100* gene and has become incorporated into the 3′ end of the *SP100* locus as an exon, resulting in the addition of a DNA-binding function to the SP100 protein.

Rogalla *et al*. [14] performed a number of PCR amplifications using primer sequences from the 3′ end of the *SP100* locus. Different PCR primer combinations produced amplicons variously containing: the penultimate exon encoding a 14 amino acid joining region between SP100 and HMG1L3; the last *SP100* intron; and the entire *HMG1L3* pseudogene. Genomic DNA from human, chimpanzee, gorilla, gibbon and rhesus macaque was used in their study. Results suggest that the retro-transposition of *HMG1L3* into the *SP100* locus occurred at least 35 million years ago. In addition, a PCR amplicon produced from the rhesus macaque revealed the presence of an *Alu* sequence between the penultimate *SP100* exon and the *HMG1L3* insertion site that is not present in hominoid genomes. Here, I have used an expanded panel of New World and Old World monkey species to refine dating of both the *HMG1L3* retrotransposition and the *Alu* insertion events.

Results and discussion

Major features of the 3′ end of the *SP100* locus are shown in Figure 1. In addition to the spatial relationship among these features, the locations of PCR primers used in this study are indicated. Rogalla *et al*. [14] primers PICauf1 and a1PICdo amplify a 614 base pair (bp) amplicon in genomic DNA from human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), and gibbon (*Hylobates lar*) and a 900 bp amplicon in the rhesus macaque (*Macaca mulatta*). Here, this same primer pair is used against genomic DNA from *H. sapiens* and *M. mulatta* as well as additional Old World monkey species including the baboons *Papio anubis* and *Papio hamadryas*, the vervet monkey *Cercopithecus*

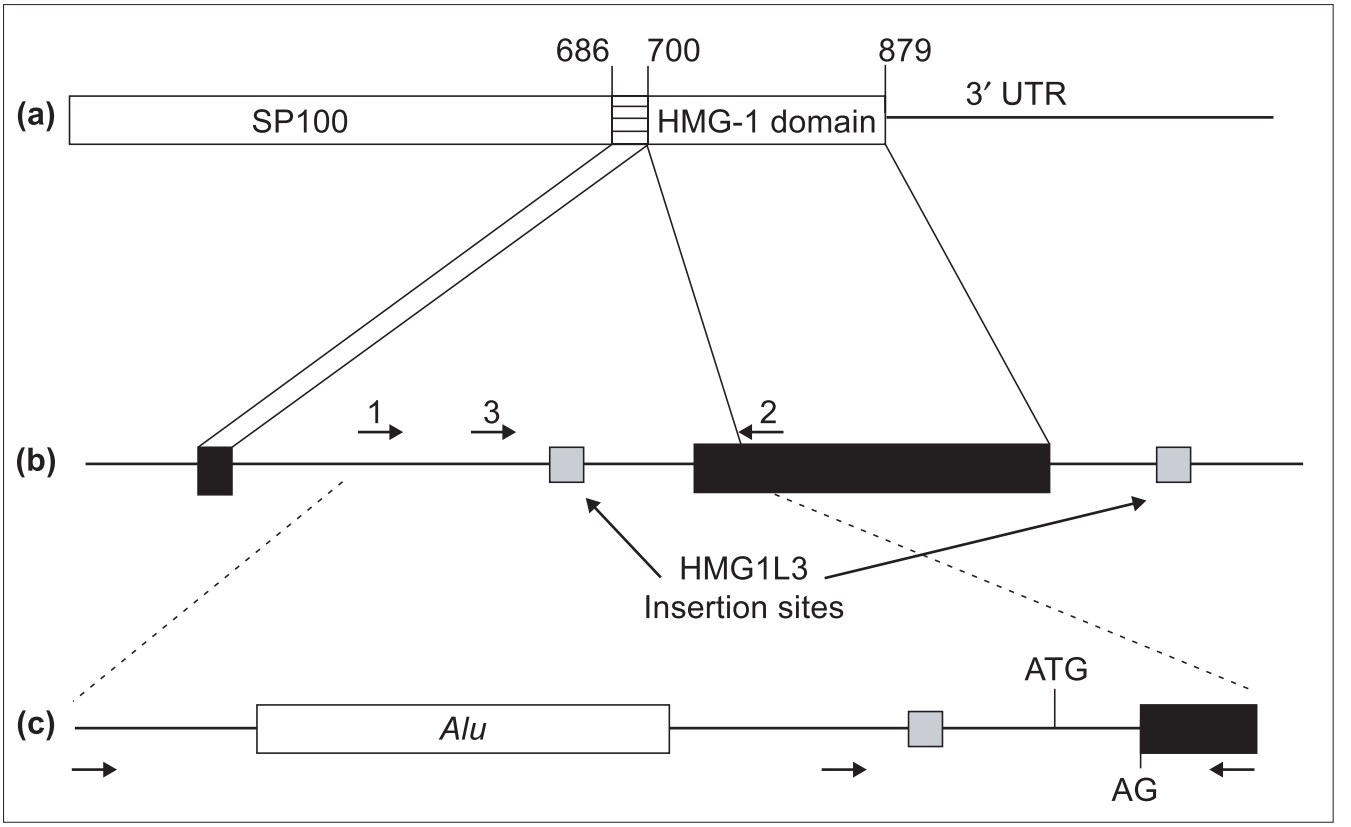

Figure 1
Schematic representation of the 3′ region of the *SP100* gene. **(a)** SP100-HMG splice variant showing the 14 amino acid joining region and the portion of the HMG1L3 retropseudogene that has been incorporated into the protein. **(b)** Genomic structure of the last two exons and the last intron of *SP100-HMG*. The *HMG1L3* insertion sites are indicated. Also shown are the locations of the PCR primer sequences PICauf1 (1), a1PICdo (2), and SP100-HMG3 (3). **(c)** Detail of the 900 base PICauf1/a1PICdo amplicon present in Old World monkeys showing the location of the *Alu* sequence.

aethiops, and the Asian macaque *Macaca assamensis*. In addition, genomic DNAs from three New World monkey species: spider monkey (*Ateles paniscus*), tamarin (*Leontopithecus saguinus*) and marmoset (*Callithrix jacchus*) are examined. Results of these PCR amplifications are shown in Figure 2; *H. sapiens* yields the expected 614 bp amplicon and all five Old World monkey species display the 900 bp amplicon. This indicates that the *Alu* sequence previously found in the rhesus macaque is present in a wide range of Old World monkey genomes. On the other hand, none of the three New World monkey species produced an amplicon with these primers, suggesting that neither the *HMG1L3* retropseudogene nor the *Alu* sequence is present in New World monkey genomes.

In support of the above suggestion, a third PCR primer, SP100-HMG3, was chosen from *SP100* genomic sequence upstream of the 5′ *HMG1L3* insertion site. Amplification with this primer and a1PICdo yields a 292 bp amplicon in human and Old World monkey samples but no product in the New World monkey samples (Figure 2). Together, these results demonstrate that New World monkey species do not have *HMG1L3*, but that it is probably present throughout the Old World monkeys as well as ape and human (Hominoidea) genomes. Clearly, the reverse transcription and retrotransposition of *HMG-1* that resulted in the creation of *HMG1L3* occurred after divergence of Old World primate species (Catarrhini) from New World primates (Platyrrhini), but prior to the divergence from the Catarrhini of the lineage leading to apes and humans. Estimates of the origin and subsequent phylogenetic radiation of the Anthropoidea offered by Kay *et al.* [15], places these events in late Eocene to middle Oligocene, or between 30 and 40 million years ago.

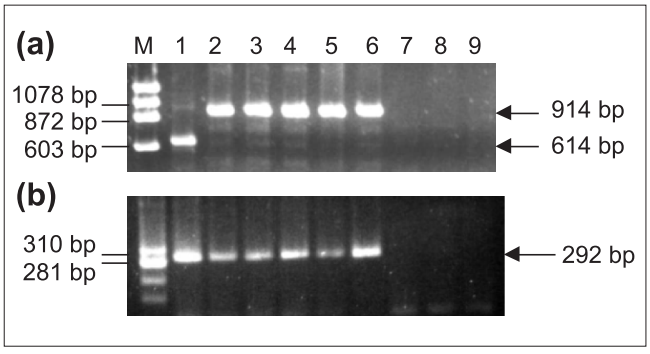

Figure 2
PCR amplicons from the 3′ region of the *SP100-HMG* gene. **(a)** Amplicons from the PICauf1/a1PICdo primer pair. **(b)** Amplicons from the SP100-HMG3/a1PICdo primer pair. The marker (M) is ϕX-174 *Hae*III. Genomic DNAs are: lane 1, human; lanes 2 and 3, baboon (*P. anubis* and *P. hamadryas*); lane 4, vervet monkey (*C. aethiops*); lanes 5 and 6, macaques (*M. mulatta* and *M. assamensis*); lanes 7-9, New World monkeys - spider monkey (*A. paniscus*), marmoset (*C. jacchus*) and tamarin (*L. rosalia*).

Results illustrated in Figure 2 also show that the 300 bp *Alu* sequence found in the region between the penultimate exon of *SP100* and the *HMG1L3* insertion site in the genome of *Macaca mulatta* is present in the genomes of other Old World monkey species from Asia, the Indian subcontinent and Africa. Previous results [14] clearly show that the *Alu* is not present in any hominoid genome. Again, relying on the anthropoid phylogeny of Kay *et al.* [15], insertion of the *Alu* would have to have occurred after the divergence of the hominoids, or not more than 25 million years ago. An alternative view is that the *Alu* sequence insertion in *SP100* occurred prior to the divergence of the hominoids, perhaps even at the same time as the *HMG1L3* insertion, but that it was lost in the line leading to Hominoidea after divergence. However, the latter possibility is unlikely, for the following reasons: individual *Alu* sequences arise via unique insertion events; they are inserted in a sequence-independent manner into breaks in genomic DNA; and those breaks are subsequently repaired with the *Alu* embedded at the break point [16]. Once inserted, *Alu* sequences remain stable features of the host genome [17]. Although *Alu* sequences have been lost from host genomes, their excision is never as clean as their insertion. Either only part of the *Alu* sequence is lost or a loss of flanking genomic DNA occurs along with loss of the *Alu* sequence [18-19].

To determine which of the two scenarios is applicable to the SP100-HMG *Alu*, PICauf1/a1PICdo amplicons from human, baboon, vervet monkey and three macaque genomes were cloned and sequenced (GenBank Accession numbers AF 377332, AF 377333, AF 377334, AF 377335, AF 377336 and AF378670). Consensus amplicon sequences from the five Old World monkey species and from three unrelated humans are presented in Figure 3. Comparison of the Old World monkey consensus sequence with the human consensus sequence shows that loss of the *Alu* among the Hominoidea subsequent to divergence from the Catarrhini would have required a perfect reversal of the insertion. In fact, the only sequence deletion is seen among the Old World monkey amplicons. This 22 base deletion (position 140-162) is near the position 186 *Alu* insertion site. If the sequence of this deletion is the same or nearly the same as that retained in the human consensus, it can form a hairpin with flanking poly(T)s and may, thus, have been lost during the repair process that occurred as part of the *Alu* insertion event.

An alignment of the *Alu* sequences from the five Old World Monkey species is presented in Figure 4. Two features of these sequences suggest a late, that is, post-divergence, origin of the insertion. First, all five sequences are consistent with a Class IV *Alu* based on the classification of Britten *et al.* [20] and, more specifically, with the *AluY* group from the nomenclature of Batzer *et al.* [21], both of which are regarded as late origin 'master' *Alu* sequences. Second, disregarding both diagnostic sites and CpG dimers, there are few sequence variations among the five species. With the

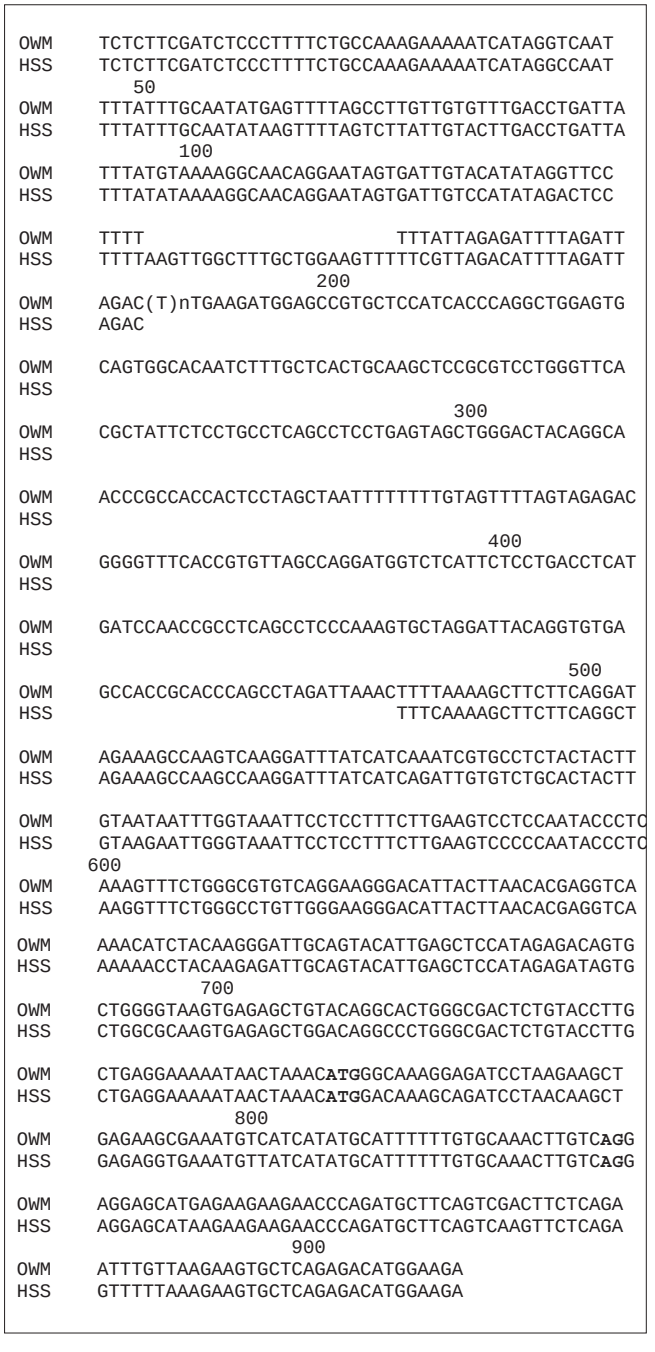

Figure 3
Alignment of consensus PICauf1/a1PICdo amplicon sequences from five Old World Monkey species (OWM) with the consensus PICauf1/a1PICdo amplicon sequence from three unrelated humans (HSS). The *HMG-1* start codon at position 761 and the end of the last *SP100-HMG* intron at position 831 are indicated in bold type.

exception of four mutations found only in one or another of the five species, the variations that are in evidence fall into two types. One type, composed of fourteen single base changes and one deletion, is shared among all five species and the other type, composed of ten single base changes and one deletion, is common to either the African species *P. anubis* and *C. aethiops* or the Asian macaque species but not both. The shared variants could be a feature of the ancestral *Alu*, but those that are segregated clearly arose after insertion and after the divergence of the macaques from the rest of the catarrhines some 8 to 10 million years ago [22,23].

On the basis of these results, the most parsimonious scenario involves insertion of the *Alu* into the 3′ region of the catarrhine *SP100* gene and loss of the 22 base upstream sequence after hominid-catarrhine divergence between 20 and 25 million years ago. The most recent point at which these events might have occurred is 10 million years ago, the time at which the Cercopithecidae, represented by *C. aethiops*, and the Papionidae, represented by baboons and macaques, diverged [22,23]. This gives a window of 10-15 million years for the *Alu* insertion. Should members of the Colobinae, such as *Colobus*, *Presbytis* or *Nasalis*, have the *Alu*, the upper limit would be pushed back to 16-18 million years ago and restrict the insertion window to only 5-10 million years [24]. Taking an estimate of 5×10^{-9} nucleotide substitutions per site per year for pseudogenes [25], mutations in the *Alu* sequences shown here suggest a date on the order of 19 million years ago for the insertion event. This is consistent with both the molecular and paleontologic data.

Materials and methods

Genomic DNA samples

Genomic DNA samples from New World and Old World monkey species were obtained through the generosity of a number of investigators. Human genomic DNAs were extracted from whole blood samples collected by the author under informed consent.

PCR amplification and amplicon sequencing

PCR primers were synthesized at Integrated DNA Technologies using standard phosphoramidite chemistry. Sequences PICauf1, 5′-TCTCTTCGATCTCCCTTTTCTG-3′ and a1PICdo 5′-TCTTCCATGTCTCTGAGCACTTCT-3′ were previously published [14]. PCR conditions used for these primers are 94°C for 5 min, followed by 35 cycles of 94°C for 30 sec; 53°C for 30 sec; 72°C for 45 sec with a final extension of 72°C for 7 min. These amplifications are optimal at 1 mM $MgCl_2$ concentration. Other primers used in this study: SP100-HMG3, 5′-CAAGGGACATTACTTAAC-ACGAGG-3′; SP100-HMG4, 5′-GGATGGACTTGATCTCTTGACC-3′; and SP100-HMG5, 3′-AGTCATGACATAGTGTGCCTGG-3′, were selected from *SP100-HMG* sequences deposited in GenBank (Accession numbers AF076675 and AF146342). Amplifications using SP100-HMG3 and a1PICdo were carried out under the same conditions as above with an annealing temperature of 55°C at 1.5 mM $MgCl_2$ and those involving SP100-HMG4 and SP100-HMG5 at an annealing temperature of 54°C at 1.5 mM $MgCl_2$. Amplicons were resolved on 1.4% agarose gels.

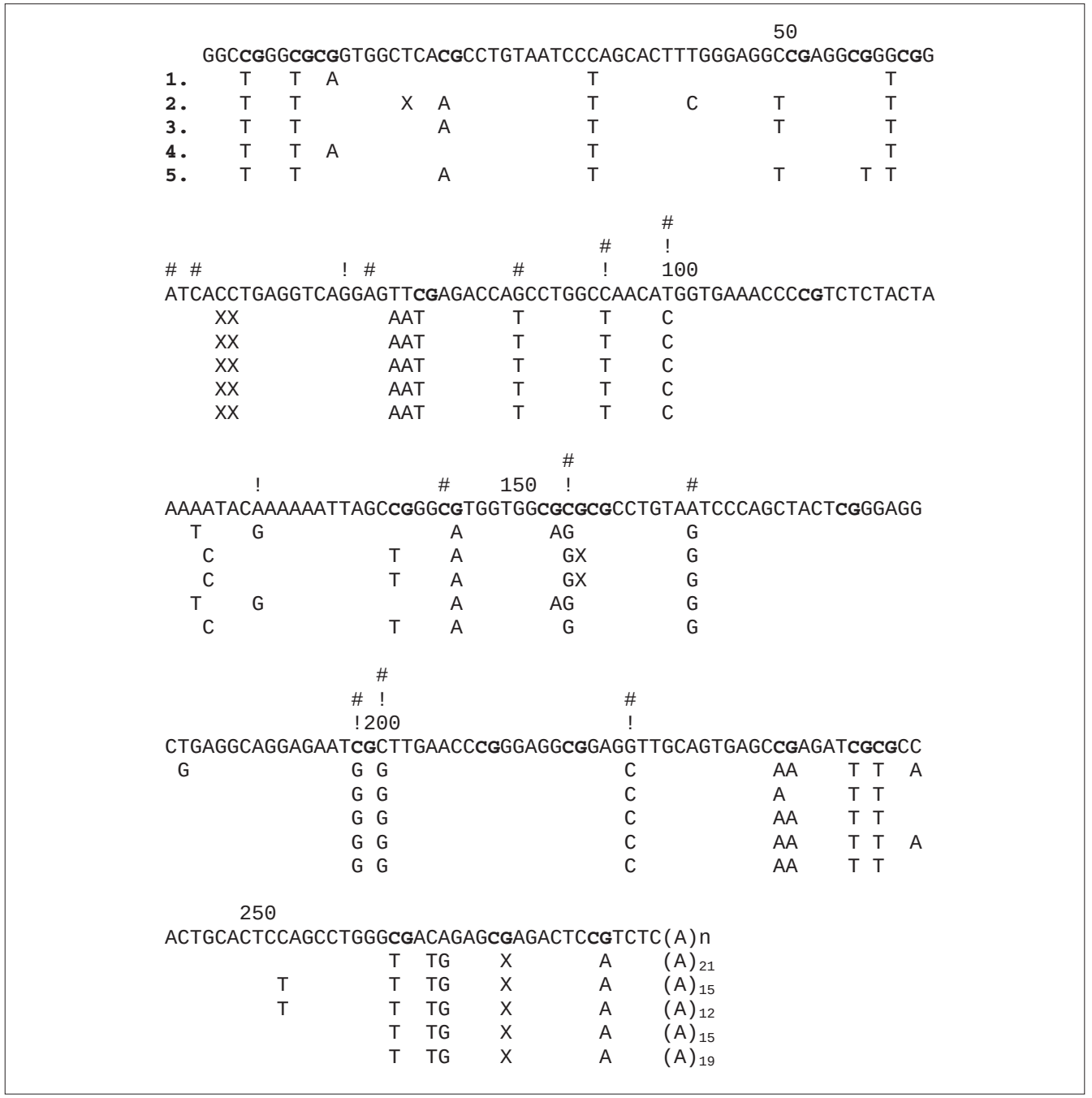

Figure 4
Comparison of the SP100-HMG *Alu* sequences among the five Old World monkey species in this study. The top line is the *Alu* consensus sequence reported by Batzer *et al.* [21]. The five species represented are: 1, vervet monkey (*C. aethiops*); 2, assamese macaque (*M. assamensis*); 3, rhesus macaque (*M. mulatta*); 4, olive baboon (*P. anubis*); and 5, pigtail macaque (*Macaca nemestrina*). Nucleotides diagnostic of a Class IV *Alu* from the scheme of Britten *et al.* [20] are denoted with an exclamation mark (!), nucleotides diagnostic of an *AluY* from the nomenclature of Batzer *et al.* [21] are denoted with a hash (#), deletions are represented by X, and CpG dimers are indicated in bold type.

PCR amplicons selected for sequencing were cloned into the TOPO-TA PCR cloning vector (Invitrogen, Carlsbad, USA). Sequencing was performed in both directions on an Applied Biosystems Model 310 Automated Fluorescence Sequencer.

Acknowledgements
I thank Moses Schanfield, Edward Max, Boris Lapin and the Southwest Foundation for Biomedical Research for their generosity in providing genomic DNA samples. Amplicon sequencing was carried out by Susanna Rezikyan at IDT.

References

1. Vanin EF: **Processed pseudogenes: characteristics and evolution.** *Annu Rev Genet* 1985, **19:**253-272.
2. Mighell AJ, Smith NR, Robinson PA, Markham AF: **Vertebrate pseudogenes.** *FEBS Lett* 2000, **468:**109-114.
3. Goncalves I, Duret L, Mouchiroud D: **Nature and structure of human genes that generate retropseudogenes.** *Genome Res* 2000, **10:**672-678.
4. Gojobori T, Li W-H, Graur D: **Patterns of nucleotide substitution in pseudogenes and functional genes.** *J Mol Evol* 1982, **18:**360-369.
5. Ophir R, Graur D: **Patterns and rates of indel evolution in processed pseudogenes from humans and murids.** *Gene* 1997, **205:**191-202.
6. Casane D, Boissinot S, Chang BH, Shimmin LC, Li W: **Mutation pattern variation among regions of the primate genome.** *J Mol Evol* 1997, **45:**216-226.
7. Devor EJ: **Use of molecular beacons to verify that the serine hydroxy-methyltransferase pseudogene SHMT-ps1 is unique to the order Primates.** *Genome Biol* 2001, **2**(2):research0006.1-0006.5
8. Devor EJ, Dill-Devor RM, Magee HJ, Waziri R: **Serine hydroxymethyltransferase pseudogene SHMT-ps1: A unique genetic marker of the order Primates.** *J Exp Zool* 1998, **282:**150-156.
9. Pompei F, Ciminelli BM, Modiano G: **Two ethnic-specific polymorphisms in the human beta pseudogene of hemoglobin.** *Hum Biol* 1998, **70:**659-666.
10. Boyson JE, Iwanaga KK, Urvater JA, Hughes AL, Golos TG, Watkins, DI: **Evolution of a new nonclassical MHC class I locus in two Old World primate species.** *Immunogenetics* 1999, **49:**86-98.
11. Brosius J: **Genomes were forged by massive bombardments with retroelements and retrosequences.** *Genetica* 1999, **107:**209-238.
12. Seeler JS, Marchio A, Sitterlin D, Transy C, Dejean A: **Interaction of SP100 with HP1 proteins: A link between the promyelocytic leukemia-associated nuclear bodies and the chromatin compartment.** *Proc Natl Acad Sci USA* 1998, **95:**7316-7321.
13. Rogalla P, Borda Z, Meyer-Bolte K, Tran KH, Hauke S, Nimzyk R, Bullerdiek J: **Mapping and molecular characterization of five *HMG1*-related DNA sequences.** *Cytogenet Cell Genet* 1998, **83:**124-129.
14. Rogalla P, Kazmierczak B, Flohr AM, Hauke S, Bullerdiek J: **Back to the roots of a new exon - the molecular archaeology of a SP100 splice variant.** *Genomics* 2000, **63:**117-122.
15. Kay RF, Ross C, Williams BA: **Anthropoid origins.** *Science* 1997, **275:**797-804.
16. Hamdi H, Nishio H, Zielinski R, Dugaiczyk A: **Origin and phylogenetic distribution of *Alu* DNA repeats: Irreversible events in the evolution of primates.** *J Mol Biol* 1999, **289:**861-871.
17. Minghetti PP, Dugaiczyk A: **The emergence of new DNA repeats and the divergence of primates.** *Proc Natl Acad Sci USA* 1993, **90:**1872-1876.
18. Miura O, Sagahara Y, Nakamura Y, Hirosawa S, Aoki N: **Restriction fragment length polymorphism caused by a deletion involving *Alu* sequences within the human α_2-plasmin inhibitor gene.** *Biochemistry* 1989, **28:**4934-4938.
19. Edwards MC, Gibbs RA: **A human dimorphism resulting from loss of an *Alu*.** *Genomics* 1992, **14:**590-597.
20. Britten RJ, Baron WF, Stout DB, Davidson EH: **Sources and evolution of human *Alu* repeated sequences.** *Proc Natl Acad Sci USA* 1988, **85:**4770-4774.
21. Batzer MA, Deininger PL, Hellmann-Blumberg U, Jurka J, Labauda D, Rubin CM, Schmid CW, Zietkiewicz E, Zuckerkandl E: **Standardized nomenclature for Alu repeats.** *J Mol Evol* 1996, **42:**3-6.
22. Disotell TR, Honeycutt RL, Ruvolo M: **Mitochondrial DNA phylogeny of the Old-World monkey tribe Papionini.** *Mol Biol Evol* 1992, **9:**1-13.
23. Morales JC, Melnick DJ: **Phylogenetic relationships of the macaques (Cercopithecidae: *Macaca*), as revealed by high resolution restrcition site mapping of mitochondrial ribsomal genes.** *J Hum Evol* 1998, **34:**1-23.
24. Szalay FS, Delson E: *Evolutionary History of the Primates.* New York: Academic Press; 1979.
25. Miyata T, Yasunaga T: **Rapidly evolving mouse alpha-globin-related pseudogene and its evolutionary history.** *Proc Natl Acad Sci USA* 1981, **78:**450-453.