the morphological evolution of maize may well surpass that of any other plant species.

Acknowledgements

This work was supported by National Science Foundation grants (BSR-88-06889 and BSR-91-07175) and by Pioneer Hi-Bred, Int. of Johnston, Iowa.

References

1 Beadle, G.W. (1939) *J. Hered.* 30, 245–247

2 Beadle, G.W. (1972) *Field Mus. Natl Hist. Bull.* 43, 2–11

3 Doebley, J. (1990) *Econ. Bot.* 44 (3 Suppl.), 6–27

4 Mangelsdorf, P.C. and Reeves, R.G. (1939) *Texas Agric. Exp. Sta. Bull.* 574

5 Mangelsdorf, P.C. (1974) *Corn: its Origin, Evolution and Improvement*, Harvard University Press

6 Langham, D.G. (1940) *Genetics* 25, 88–107

7 Mangelsdorf, P.C. (1947) *Adv. Genet.* 1, 161–207

8 Rogers, J.S. (1950) *Genetics* 35, 541–558

9 Collins, G.N. and Kempton, J.H. (1920) *J. Agric. Res.* 19, 1–37

10 Doebley, J., Stec, A., Wendel, J. and Edwards, M. (1990) *Proc. Natl Acad. Sci. USA* 87, 9888–9892

11 Beadle, G.W. (1980) *Sci. Am.* 242, 112–119, 162

12 Mangelsdorf, P.C., MacNeish, R.S. and Galinat, W.C. (1967) in *The Prehistory of the Tehuacan Valley I* (Byers, D.S., ed.), pp. 178–200, University of Texas Press

13 Helentjaris, T. and Burr, B., eds (1989) *Development and Application of Molecular Markers to Problems in Plant Genetics*, Cold Spring Harbor Laboratory Press

14 Edwards, M.D., Stuber, C.W. and Wendel, J.F. (1987) *Genetics* 116, 113–125

15 Lander, E.S. and Botstein, D. (1989) *Genetics* 121, 185–199

16 Doebley, J. and Stec, A. (1991) *Genetics* 129, 285–295

17 Galinat, W.C. (1973) *Evolution* 27, 644–655

18 Mangelsdorf, P.C. (1952) in *Heterosis* (Gowen, J.W., ed.), pp. 175–198, State College Press, Ames, Iowa

19 Iltis, H.H. (1983) *Science* 222, 886–894

20 Galinat, W.C. (1969) *Mass. Agric. Exp. Sta. Bull.* 577, 1–19

21 Kempton, J.H. (1924) *J. Agric. Res.* 27, 537–596

J. Doebley is in the Department of Plant Biology, University of Minnesota, St Paul, MN 55108, USA.

Master genes in mammalian repetitive DNA amplification

PRESCOTT L. DEININGER, MARK A. BATZER, CLYDE A. HUTCHISON, III AND MARSHALL H. EDGELL

The analysis of species-specific subfamilies of both the LINE and SINE mammalian repetitive DNA families suggests that such subfamilies have arisen by amplification of an extremely small group of 'master' genes. In contrast to the master genes, the vast majority of both SINEs and LINEs appear to behave like pseudogenes in their inability to undergo extensive amplification.

A significant percentage of all mammalian genomes consists of interspersed repetitive DNA sequences. These are generally classified as SINEs (short interspersed elements) or LINEs (long interspersed elements). The SINEs range in size from 90 to 400 bp, while LINEs can be as large as 7000 bp. New copies of both types of element find their way into the genome via reverse transcription of an RNA intermediate, a process called retroposition or retrotransposition (reviewed in Ref. 1). The RNA intermediate involved in the retroposition of SINEs is transcribed by RNA polymerase III, while that for LINEs is thought to be produced by RNA polymerase II. Both SINEs and LINEs are present in the mammalian genome in copy numbers in excess of 100 000. Many of the major SINE and LINE families can be divided into subfamilies, which are defined in terms of common nucleotide variations at specific locations. All current evidence indicates that the subfamilies arose via amplification, rather than through any process acting to reduce sequence divergence among pre-existing members of the family within a species.

Subfamilies of SINEs and LINEs

Although comparison of a large array of their properties shows that these two repetitive families belong to different classes, they both share a common feature that seems unexpected for a family of selfish elements: the lineages of both families seem to be dominated by a very small number of 'master' elements. It is this shared feature of these two disparate families that is discussed below.

Alu subfamilies

A given SINE family is usually present in only a moderate number of related species[2]. In primates, the most abundant SINE is the Alu family. It has long been known that the Alu family members exhibit species-specific polymorphisms within various primate species[3]. More recently, a number of laboratories have identified a series of subfamilies of Alu elements within some primate species[4–9]. Each of these subfamilies has one or more diagnostic differences in their consensus sequence relative to the Alu consensus sequence. However, there are significant differences in the amount of sequence divergence seen in the different subfamilies[8,9], suggesting that they arose at a variety of evolutionary times. Furthermore, analysis of the diagnostic changes in the different subfamilies suggested that subfamilies could be placed in a sequential order[7]. Analysis of the times of insertions, by sequencing and polymerase chain reaction (PCR) amplification of orthologous primate loci, confirms that the youngest subfamily members have the most diagnostic changes[8], but the least sequence divergence overall. These data, along with those from the recent insertion events to be described below, suggest that as a new subfamily is established, the older subfamily stops amplifying.

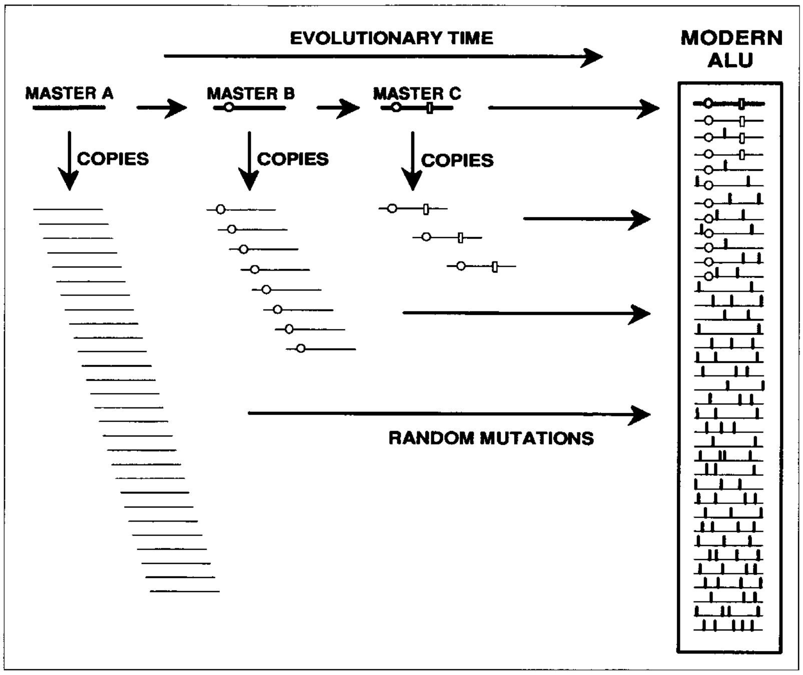

FIG **1**

The master gene model for formation of Alu subfamilies. In this model we propose that a master gene locus is responsible for the amplification of each subfamily of Alu sequences. The original master (master A) made many copies that were then subject to random mutation (vertical lines in the panel on the right). After a period of time, a mutation (○) occurred in the master A, resulting in master B, and all subsequently formed copies would contain this mutation (○), as well as accumulating random mutations. Alternatively, a new master Alu B could have been formed with those mutations and the previous master gradually silenced. These types of master changes would continue (a mutation ▯ results in master C) with each subfamily formation, resulting in the younger subfamilies containing fewer random mutations. The panel marked modern Alu represents the Alu sequences as they would look today if aligned. The coexistence of two master genes for any SINE or LINE will result in the simultaneous formation of different subfamilies.

A computer analysis of the older Alu subfamily members has suggested that early in Alu family formation, several subfamilies may have been propagated during the same interval[9,10]. However, these analyses of the older and highly divergent Alu family members are difficult and relied to some extent on CpG dinucleotide positions that are known to mutate at extremely high rates in Alu family members. Thus, the exact nature of early Alu subfamilies is difficult to discern. The more recent subfamilies show a more distinctly sequential order of appearance of the diagnostic positions[8]. However, there is also evidence that, in a few cases, the most recent Alu family amplifications have involved closely related subfamilies that amplified in parallel[11,12]. It is unclear whether these amplifications represent evidence for multiple master genes active in the same time period[11,12], or represent allelic variation in a single master gene locus[8]. Nonetheless, it does seem that the vast majority of Alu subfamilies can be explained by a single series of sequential subfamily changes, each dependent on the previous subfamily.

L1 subfamilies

Although there are many different LINE families in the mammalian genome, the LINEs-1 or L1 family is several orders of magnitude more abundant than any other. Examination of L1 sequences by Southern analysis of DNA from primates[13] and murids[14] shows that the L1 family within each species or subspecies examined has species-specific diagnostic positions shared by a significant fraction of the total L1 population. In some species, such as human and mouse, more extensive analyses show that the L1 families in individual species can be divided into a small number of subfamilies[15–19]. In mice, four distinct subfamilies (V, F, C, A) are easily distinguishable by subfamily-specific sequences at the 5' end of the elements. Each of the four subfamilies seems to have attained its species-specific character at quite different times. A maximum parsimony model for the evolution of the more homologous L1 families in mice, on the basis of 3' sequences, suggests that the homology seen in these elements came from only two or three master elements[20]. Likewise, analysis of the 5' untranslated region of one of these major subfamilies (A) suggests that all of the substructure is derived from a very small number of master elements whose sequence was evolving over time[21]. A more extensive analysis based on 5' sequences from 30 L1 elements indicates that the F, C, and A subfamilies have all been derived from the same master at different times during evolution (N.B. Adey, pers. commun.).

Subfamilies in other repeated sequences

Much less information is available for other SINE families than for the Alu families. However, there are enough data on several SINE families to suggest that they show similar patterns of evolution. For example, there are at least three subfamilies of the rabbit C repeat, each having an apparently different age in the rabbit genome, and showing evidence of progressive changes in the subfamily consensus sequences[22]. Several repeated sequence families in rodents also show subfamily structures of apparently differing ages (reviewed in Ref. 23). Thus, it would appear that most of the mammalian retroposon and retrotransposon families share a similar, distinctive evolutionary pattern.

Subfamily formation

The master gene model

The relatively recent evolutionary origin of most of the SINE families and subfamilies, as well as studies of the evolution of orthologous repetitive elements in the genomes of different species, all point to the formation of the subfamily structure through an amplification process, rather than modification of existing family members. This can only be explained if the amplifications are occurring from a limited subset of the members of each family (Figs 1, 2B). A striking

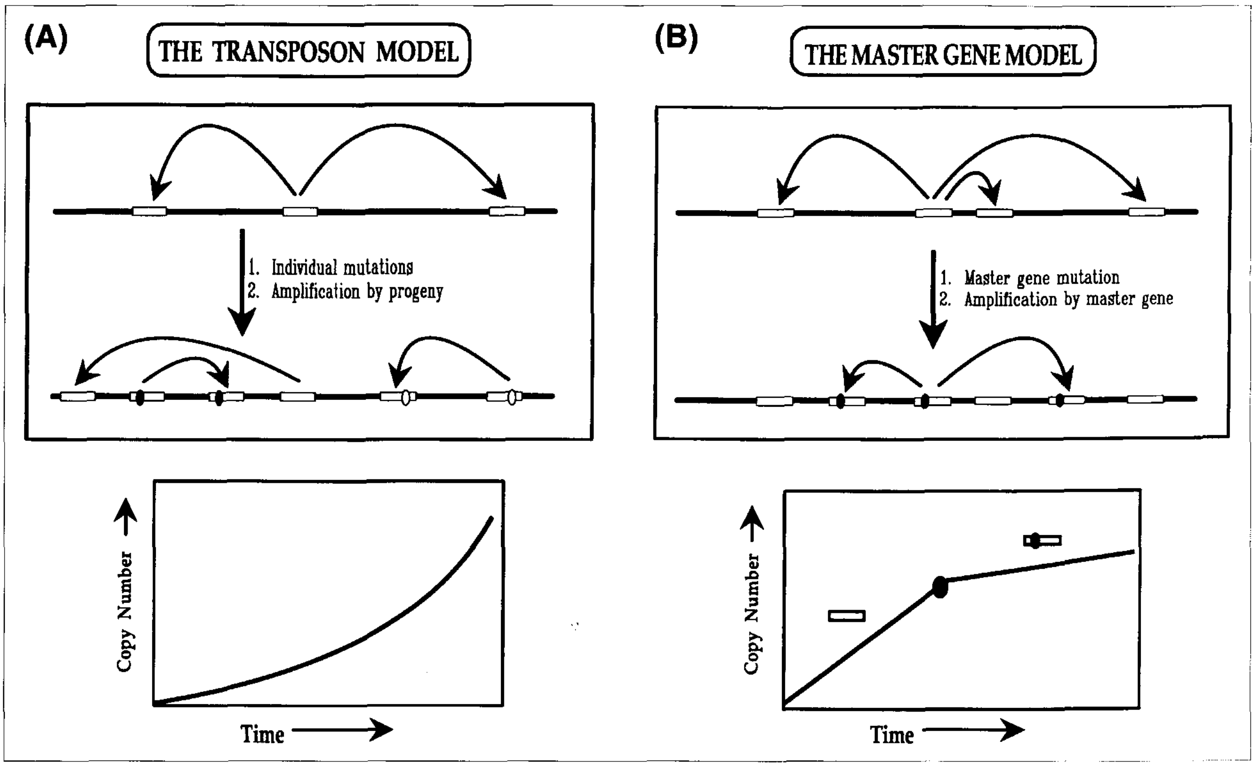

FIG 2

Different models for sequence amplification. (A) The transposon model. In this model each of the sequence copies would be capable of being amplified. As these elements each accumulate random mutations (○ and •), many independent and parallel subfamilies can be formed. If left unchecked, this process could lead to an ever-increasing amplification rate. (B) The master gene model. This model best fits the existing subfamily data. A single gene can make many copies, which are themselves pseudogenes. Thus, mutations in that master gene (•) appear as subfamilies, formed at different times in evolution (see also Fig. 1). In this model, the amplification rate would depend only on the master gene and therefore would be linear, with changes in rate occurring if mutations affect the master gene. In some repeated DNA families, it is possible that on rare occasions a copy could become a master gene, resulting in parallel evolution or sequential replacement of master genes.

example of Alu subfamilies involves the most recently inserted Alu sequences in humans: three of the most recently inserted Alu family members were found to be almost identical in sequence[24], whereas typical Alu family members show only about 70–75% sequence identity. This suggests that the three recent insertions were not progeny drawn at random from the very large Alu population, but were instead generated by a single Alu master gene. There are three additional examples of randomly isolated, recently inserted Alu elements, and these also show a very high degree of sequence identity with the previously identified recently inserted Alu elements[25–27]. One of these is an insertion into the *NF1* (neurofibromatosis) locus, which has inactivated the gene[27]. The insertion of this Alu element within the last generation represents direct evidence that current Alu insertions are occurring from this Alu master gene. Other data show that multiple closely related subfamilies have been formed in parallel in the human genome[11,12,28]. Although these data and some of the studies on the earliest Alu amplifications[9,10] suggest that there may have been a few independent Alu master genes, we believe it is also possible to explain the data by allelic variation at a single master gene locus[8].

There are also two excellent examples of recent L1 insertions, both of which were into the human factor VIII gene (see Ref. 29). Two different insertion events have been shown to have occurred in the last generation and they have closely related sequences in the regions common to both elements. Clearly, these two inserted elements were derived from either the same master L1 locus or a closely related subset of L1 elements; hence their sequence homology originated from this shared parent and not from the action of concerted evolution on pre-existing elements in the factor VIII gene.

Why are there so few master elements?

The master elements, as defined in this type of evolutionary analysis, must have some property (or properties) that separate them from progeny that are replicatively incompetent on an evolutionary timescale. The master elements must not only be able to create copies of themselves, but they must also continue to do so over a significant period of time in order to have any noticeable effect on subfamily structure. Thus, while it is possible that some of the progeny are not completely replicatively incompetent, they must be either much less effective than the master(s) or relatively quickly silenced. What features might contribute to replication competence and/or longevity of a master element?

Transcription from recent Alu and L1 subfamilies

Despite the presence of hundreds of thousands of copies, the majority of which have an RNA polymerase III promoter that is functional *in vitro*, there is very little transcription of Alu family members *in vivo*[30]. Transcripts have been identified for a truncated Alu element termed BC200 (Ref. 31), and there is an HS-1 subfamily-specific transcript[32]. The sequence of the BC200 transcript does not correspond to the Alu master gene, but the HS-1-specific transcript may represent transcription from the Alu master gene. There is also evidence for very limited transcription of other Alu elements, probably influenced by their chromosomal environment. However, it is clear that the vast majority of Alu sequences are transcriptionally silent in the cultured cells studied and there is preferential expression of the most recent subfamily.

It is interesting to note that the BC200 gene has generated at least two pseudogene copies (J. Martignetti and J. Brosius, pers. commun.). It is clear, however, that it has not generated many copies. Thus, despite its high level of expression in some tissues, it is much less adept at making copies than the master gene responsible for most Alu amplifications. The difference in amplification rates may be due to the lack of a required component such as reverse transcriptase, an alteration in self-priming ability, or may represent differences in the amount of transcription. Similar data suggest that in the rodent B2 family, only a small subset of family members are transcriptionally competent[33], and in the rodent ID family, a single gene contributes the majority of specific transcription[34].

There are several possible explanations for this limited transcription capability. The 7SL RNA gene, from which the Alu master gene was derived, requires approximately 37 bp of upstream sequence for its promoter to be active *in vitro*[35]. Thus, the master gene(s) may have appropriate upstream sequences that are not duplicated in the copies. Evolutionary evidence suggests that the Alu master gene is present in a methylation-free island, whereas the copies are extensively methylated[8]. Extensive genomic methylation has been confirmed for the most recently inserted Alu family members[36]. Either methylation or the high mutation rate associated with methylated DNA may also be partly responsible for the relative transcriptional silence of the Alu copies. It has been suggested that rather than a single major master locus being responsible for the sequential Alu subfamilies[8], there may have been a sequential replacement of the locus supplying the major Alu amplifications[7]. In the latter case, it has been suggested that it may be a very rare event for an Alu amplification to land in a methylation-free island, and those rare events may produce the next master gene as mutations or other factors gradually silence the previous master locus[36].

Although the data are less extensive, the situation is similar for the L1 family. Sequence analysis[37] of ten cDNA clones derived from unselected mRNA indicates that they all came from only one of the four mouse L1 subfamilies, A. That is, the predominant L1 mRNA is from the same subfamily that shows the most recent amplification. However, the mRNAs were clearly not from the master itself since they contained many mutations that would render the products nonfunctional.

Other factors in master gene activity

Although transcription is likely to be a major controlling influence on amplification rate, other factors may also contribute. One other possibility is the ability of an RNA to be reverse transcribed; perhaps there may be a factor that allows the transcript to be produced or compartmentalized wherever the reverse transcriptase activity is available. Alternatively, it has been proposed that SINEs may be highly efficient at reverse transcription because their 3' end may be able to fold over and self-prime the reverse transcription. Thus, the exact nature of sequences flanking the 3' end of an Alu insertion could also contribute to activity. Furthermore, some SINE transcripts have been demonstrated to be post-transcriptionally cleaved at their 3' ends[33]. Thus, the appropriate processing could be either a positive or negative influence on the ability of an RNA molecule to be reverse transcribed.

Full-length L1 elements as potential masters

Most L1 sequences are truncated with respect to the consensus sequence and therefore must be replicatively incompetent. However, there are several thousand full-length L1 sequences in the mammalian genome and hence some other features apart from length must determine which ones act as master genes. One could look for sequences that correspond to the functional element by using information from elements known to have been generated fairly recently. One way of doing this is to use phylogenetic analyses to derive 'ancestral' sequences that are presumed to represent the then active master gene[37]. In addition, a full-length element with completely intact open reading frames was isolated from the mouse[38]. However, neither of these elements has been shown to function as a master. A more direct approach is to use sequence identity between recent insertions to identify full-length elements that are candidate master copies. This has been done using sequence from the L1 elements recently inserted into the human factor VIII gene[29]. The full-length element that was identified carries short direct repeats and a poly(A) tail, suggesting that it is itself a retroposed sequence. It has complete open reading frames, one of which encodes a functional reverse transcriptase[39]. This element is currently the best candidate for an isolated master gene. Study of this functional L1 element is likely to provide a great deal of insight into the features that make a gene replicatively competent.

Implications of the master gene model

There appear to be very few replicatively competent SINEs or LINEs in the mammalian genome. Because all the subfamily studies are based on evolutionary analyses, they can only detect master genes that have been active at a relatively high level over a significant length of time. It is possible that a much larger fraction of both the SINEs and the LINEs are capable of very low levels of amplification, or of amplification that is silenced rapidly on an evolutionary timescale. The number of L1 masters seems some-what larger than that for the Alu family but is still very small. It seems clear that the vast majority of LINE progeny are replicatively incompetent – full-length elements as well as truncated ones. Thus, although they have generated many pseudogenes, the SINEs and LINEs are remarkably inefficient at producing replicatively competent copies. Thus, the amplification rate of such a DNA family (Fig. 2A) would be likely to follow a very different pattern from that where most of the progeny were replicatively competent (Fig. 2B).

If a significant fraction of repeated sequence copies were transpositionally competent, the amplification rate would be expected to grow exponentially as more and more progeny were produced (unless other environmental variables were greatly to affect retroposition frequency). With the master gene model, amplification would be controlled by the state of the master genes, and thus the amplification rate would be very sensitive to changes in these genes. Such changes might result in significant periods of decreased amplification (as seems to have occurred over evolutionary time with the Alu family[8]), and also periods of very rapid amplification.

Many transposable elements actually impose limitations on their own amplification rate and hence the amplification model in Fig. 2A may be too simplistic. However, while the large coding capacity of LINEs suggests they may be able to regulate their own copy number, it is hard to imagine that SINEs have similar powers unless the transcripts themselves have regulatory capacity. It is surprising that such disparate types of high copy number repetitive families should both have the intuitively unexpected property of being generated by a very small number of elements. At this point it is not known what shared feature gives rise to this property.

One of the most important implications of the master gene model has to do with function. If there are only a few master genes how is it that they have managed to persist in the genome over such long periods of time? Traditionally one concludes that sequences present in low copy number (replicatively incompetent elements) must provide a useful function to avoid mutational inactivation and clearance from the genome. If those master genes provide a function, are the large numbers of pseudogenes a direct or indirect consequence of that function? Alternatively, one could explore whether or not there is a way for the large reservoir of inactive elements to provide a new mechanism of sequence retention in the genome.

Acknowledgements

This research was supported by USPHS Grants RO1 HG00340 (to P.L.D.), AI08998 and GM21313 (to M.H.E. and C.A.H.) and a grant from the Cancer Crusaders (to M.A.B.).

References

1 Weiner, A.M., Deininger, P.L. and Efstratiadis, A. (1986) *Annu. Rev. Biochem.* 55, 631–661

2 Deininger, P.L. and Daniels, G.R. (1986) *Trends Genet.* 2, 76–80

3 Daniels, G.R. *et al.* (1983) *Nucleic Acids Res.* 11, 7579–7593

4 Slagel, V. *et al.* (1987) *Mol. Biol. Evol.* 4, 19–29

5 Willard, C., Nguyen, H.T. and Schmid, C.W. (1987) *J. Mol. Evol.* 26, 180–186

6 Jurka, J. and Smith, T. (1988) *Proc. Natl Acad. Sci. USA* 85, 4775–4778

7 Britten, R.J., Baron, W.F., Stout, D.B. and Davidson, E.H. (1988) *Proc. Natl Acad. Sci. USA* 85, 4770–4774

8 Shen, M.R., Batzer, M.A. and Deininger, P.L. (1991) *J. Mol. Evol.* 33, 311–320

9 Quentin, Y. (1988) *J. Mol. Evol.* 27, 194–202

10 Jurka, J. and Milosavljevic, A. (1991) *J. Mol. Evol.* 33, 105–121

11 Matera, A.G., Hellmann, U., Hintz, M. and Schmid, C.W. (1990) *Nucleic Acids Res.* 18, 6019–6023

12 Leeflang, E.P. *et al. J. Mol. Evol.* (in press)

13 Sakaki, Y., Kurata, Y., Miyake, T. and Saigi, K. (1983) *Gene* 24, 179–190

14 Jubier-Maurin, V. *et al.* (1985) *J. Mol. Biol.* 184, 547–564

15 Loeb, D.D. *et al.* (1986) *Mol. Cell. Biol.* 6, 168–182

16 Padgett, R.W., Hutchison, C.A., III and Edgell, M.H. (1988) *Nucleic Acids Res.* 16, 739–749

17 Jurka, J. (1989) *J. Mol. Evol.* 29, 496–503

18 Adey, N.B., Schickman, S.A., Hutchison, C.A., III and Edgell, M.H. (1991) *J. Mol. Biol.* 221, 367–373

19 Jubier-Maurin, V. *et al. Mol. Biol. Evol.* (in press)

20 Hardies, S.C. *et al.* (1986) *Mol. Biol. Evol.* 3, 109–125

21 Schickman, S.A., Adey, N.B., Edgell, M.H. and Hutchison, C.A., III *Mol. Biol. Evol.* (in press)

22 Krane, D.E., Clark, A.G., Cheng, J-F. and Hardison, R.C. (1991) *Mol. Biol. Evol.* 8, 1–30

23 Deininger, P.L. (1989) in *Mobile DNA* (Berg, D.E. and Howe, M.M., eds), pp. 619–636, American Society for Microbiology

24 Deininger, P.L. and Slagel, V.K. (1988) *Mol. Cell. Biol.* 8, 4566–4569

25 Ryan, S.C. and Dugaiczyk, A. (1989) *Proc. Natl Acad. Sci. USA* 86, 9360–9364

26 Stoppa-Lyonnet, D., Carter, P.E., Meo, T. and Tosi, M. (1990) *Proc. Natl Acad. Sci. USA* 87, 1551–1555

27 Wallace, M.R. *et al.* (1991) *Nature* 353, 864–866

28 Batzer, M.A. *et al.* (1990) *Nucleic Acids Res.* 18, 6793–6798

29 Dombroski, B.A. *et al.* (1991) *Science* 254, 1805–1808

30 Paulson, K.E. and Schmid, C.W. (1986) *Nucleic Acids Res.* 14, 6145–6158

31 Watson, J.B. and Sutcliffe, J.G. (1987) *Mol. Cell. Biol.* 7, 3324–3327

32 Matera, A.G., Hellmann, U. and Schmid, C.W. (1990) *Mol. Cell. Biol.* 10, 5424–5432

33 Maraia, R.J. (1991) *Nucleic Acids Res.* 19, 5695–5702

34 DeChiara, T.M. and Brosius, J. (1987) *Proc. Natl Acad. Sci. USA* 84, 2624–2628

35 Ullu, E. and Weiner, A.M. (1985) *Nature* 318, 371–374

36 Schmid, C.W. (1991) *Nucleic Acids Res.* 19, 5613–5617

37 Schickman, S.A., Severynse, D.M., Edgell, M.H. and Hutchison, C.A., III (1992) *J. Mol. Biol.* 224, 559–574

38 Sheehee, W.R. *et al.* (1987) *J. Mol. Biol.* 196, 757–767

39 Mathias, S.L. *et al.* (1991) *Science* 254, 1808–1810

P.L. Deininger is in the Department of Biochemistry and Molecular Biology, Louisiana State University Medical Center, New Orleans, LA 70112, USA, and in the Laboratory of Molecular Genetics, Alton Ochsner Medical Foundation, New Orleans, LA 70121, USA; M.A. Batzer is in the Department of Biochemistry and Molecular Biology, Louisiana State University Medical Center, New Orleans, LA 70112, USA; C.A. Hutchison, III and M.H. Edgell are in the Department of Microbiology, University of North Carolina, Chapel Hill, NC 27599, USA; present adress for M.A.B.: Human Genome Center, Biomedical Sciences Division, L-452, Lawrence Livermore National Laboratory, PO Box 5507, Livermore, CA 94550, USA.