Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 1283–1286, February 1989
Evolution

Divergent evolution of part of the involucrin gene in the hominoids: Unique intragenic duplications in the gorilla and human

JEFFREY TEUMER AND HOWARD GREEN

Department of Cellular and Molecular Physiology, Harvard Medical School, 25 Shattuck Street, Boston, MA 02115

Contributed by Howard Green, November 18, 1988

ABSTRACT **The gene for involucrin, an epidermal protein, has been remodeled in the higher primates. Most of the coding region of the human gene consists of a modern segment of repeats derived from a 10-codon sequence present in the ancestral segment of the gene. The modern segment can be divided into early, middle, and late regions. We report here the nucleotide sequence of three alleles of the gorilla involucrin gene. Each possesses a modern segment homologous to that of the human and consisting of 10-codon repeats. The early and middle regions are similar to the corresponding regions of the human allele and are nearly identical among the different gorilla alleles. The late region consists of recent duplications whose pattern is unique in each of the gorilla alleles and in the human allele. The early region is located in what is now the 3′ third of the modern segment, and the late, polymorphic region is located in what is now the 5′ third. Therefore, as the modern segment expanded during evolution, its 3′ end became stabilized, and continuing duplications became confined to its 5′ end. The expansion of the involucrin coding region, which began long before the separation of the gorilla and human, has continued in both species after their separation.**

Involucrin is a specialized protein of terminally differentiated keratinocytes. The involucrin gene is expressed as part of the program of terminal differentiation in all stratified squamous epithelia, including the epidermis (1–3); as a component of the cross-linked envelopes (1), involucrin must contribute to the resistance of the outer layers of the epithelium.

The coding region of the human involucrin gene lies within a single exon and consists of an ancestral segment and a modern segment (4, 5). The modern segment is composed of 39 repeats of a 10-codon sequence and accounts for two-thirds of the coding region. The modern segment must have evolved in the higher primates after they diverged from the prosimians because the lemur involucrin gene, which contains an ancestral segment similar to that of the human gene, has a segment of repeats different from the modern segment in both sequence and location (5).

The stages of evolution of the modern segment can be traced by comparing the repeat structure in the genes of different higher primates. We describe here the nucleotide sequences of three alleles of the gorilla involucrin gene* and compare them to the published sequence of the human gene. Part of the gorilla modern segment resembles that of the human, and part is different. The comparisons show how coding sequence was created by a process of duplication and deletion of repeats in a specific part of the coding region.

MATERIALS AND METHODS

DNA was obtained from two individuals of *Gorilla gorilla* (western lowland gorilla). Keratinocytes from gorilla K were derived from a vaginal biopsy obtained through Harold McClure (Yerkes Regional Primate Center, Emory University) and supplied to us by Robert Rice (Harvard School of Public Health). The keratinocytes were grown under the same conditions as human keratinocytes (6, 7). DNA from gorilla F was prepared from a fibroblast culture made available through the National Institutes of Health Aging Cell Repository (Camden, NJ) (repository no. AG5251). These fibroblasts were grown in the Dulbecco–Vogt modification of Eagle's medium supplemented with 10% fetal bovine serum. DNA was extracted by published procedures (8). Restriction maps of the gorilla involucrin genes were constructed by Southern blot hybridization using as probe plasmid pλI3-H6B, which contains the human involucrin gene (4), labeled with [^{32}P]dCTP by the oligonucleotide primer method (9). Analysis revealed the presence of two alleles in the genome of gorilla K (H. Tseng and H.G., unpublished data), consistent with the earlier finding that the keratinocytes of this gorilla expressed two involucrin proteins distinguishable by gel electrophoresis (10). Each allele was contained within a single *Hin*dIII fragment, one of about 5.8 kilobases (kb) and the other of about 6.1 kb. The genome of gorilla F was homozygous for a third allele, present in a single *Hin*dIII fragment of about 6.2 kb.

Each involucrin-containing *Hin*dIII fragment was cloned by using a modification (5) of the technique described by Nicholls *et al.* (11). *Hin*dIII-digested DNAs were separated on preparative agarose gels. Slices were cut from the gels, and the DNA was electroeluted from each slice and concentrated. A portion of the DNA from each slice was analyzed by Southern blot hybridization for the presence of involucrin sequences. DNA from slices containing involucrin DNA was ligated to the *Hin*dIII-digested and dephosphorylated plasmid pGEM-3 (Promega Biotec). The ligation mixture was used to transform competent HB101 bacteria. Bacterial colonies were screened for the presence of involucrin DNA by colony-lift hybridization (8) using pλI3-H6B as a probe. After four rounds of screening and purification, pure colonies containing the involucrin sequence were obtained. Plasmids containing the small and medium alleles, isolated from the gorilla K DNA library, were designated pIGorH5.8 and pIGorH6.1, respectively. In pIGorH5.8, nested deletions were generated in the 3′ direction from an *Xba* I site (located 5′ to the promoter sequences) and from an *Xho* I site (located in the intron), and the sequence of the coding strand was obtained. In pIGorH6.1, nested deletions were generated in the 5′ direction from an *Nco* I site located 3′ to the involucrin gene, and the noncoding strand of the gene was sequenced. The large allele, isolated from the gorilla fibroblast library, was designated pIGorH6.2. A 2.3-kb *Bgl* II–*Bam*HI fragment, which contained the entire coding region of the large allele, was subcloned into the *Bam*HI site of pGEM-3. This subclone, pIGorH6.2(Bgl2.3Bam), was used to generate nested deletions in the 5′ direction starting at a *Bam*HI site located just 3′ to the coding region. All deletions were

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*The gorilla sequences reported in this paper are being deposited in the EMBL/GenBank data base (accession no. J04499).

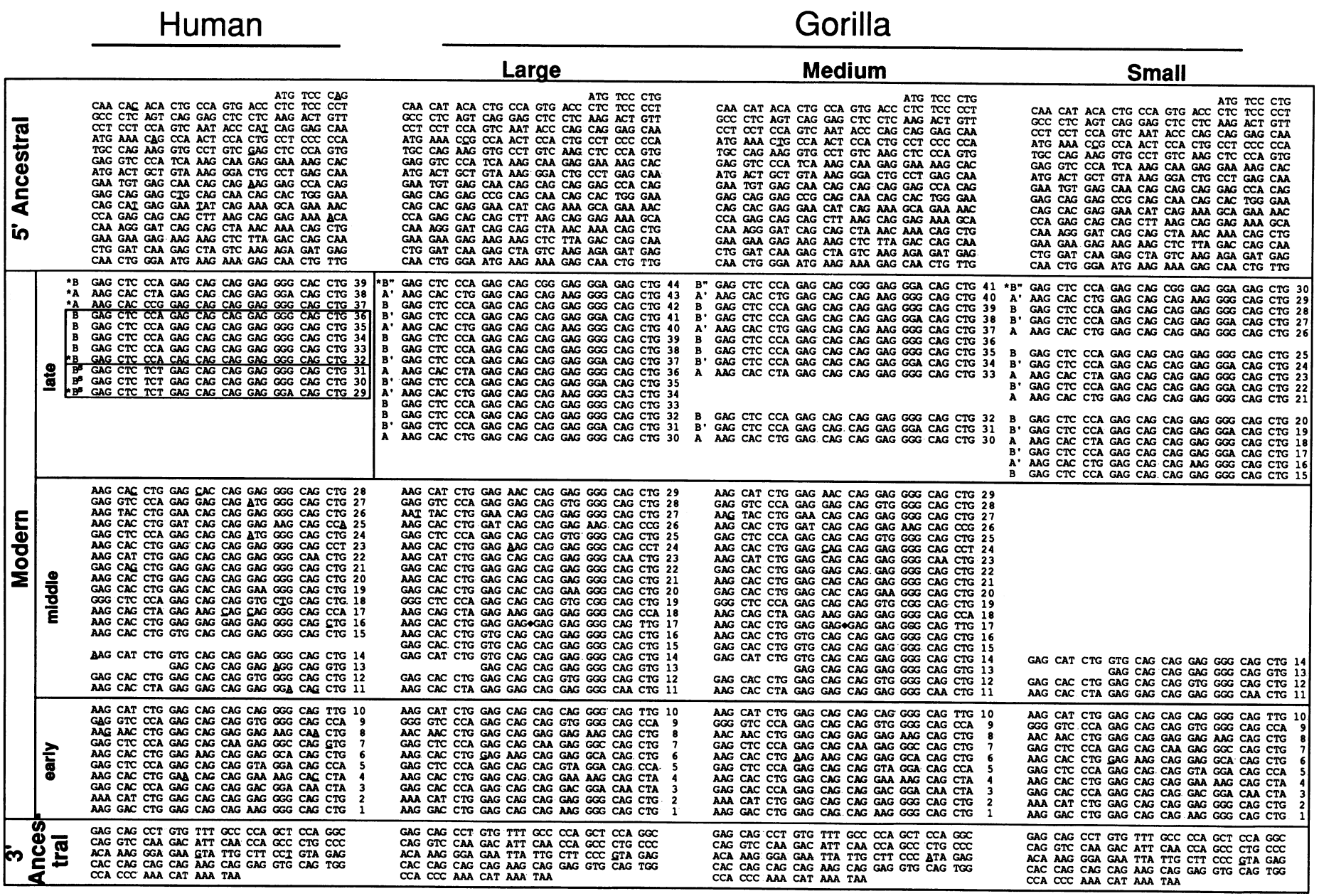

FIG. 1. (*Legend appears at the bottom of the opposite page.*)

Table 1. Divergence of ancestral and modern segments

	% base mismatches (mismatches per total bases compared)	
	Human:gorilla	Gorilla:gorilla
Ancestral	2.1 (37/1791)	0.2 (4/1791)
Early modern	2.1 (19/900)	0.2 (2/900)
Middle modern	2.9 (34/1173)	0.2 (2/783)

In each comparison, base mismatches were totalled for all three possible pairings and divided by the total number of bases in the three pairings. Deletions and insertions were not counted as mismatches.

generated by the exonuclease III method (12) with the Erase-a-Base system (Promega Biotec). All sequencing was performed by the dideoxy chain-termination method with the Klenow enzyme, reverse transcriptase (Promega Biotec), or Sequenase (United States Biochemical).

RESULTS

As in the human involucrin gene (4), the coding region for each of the gorilla alleles lies within the second exon. A single intron separates this exon from a 43-base-pair first exon (not shown). This overall structure of the gorilla gene is identical to that of the human gene.

The Ancestral Segment and the Modern Segment of Repeats. Shown in Fig. 1 are the nucleotide sequences of the gorilla involucrin coding regions for the small allele (495 amino acids), the medium allele (606 amino acids), and the large allele (636 amino acids). These are compared with the single known human allele (585 amino acids), whose sequence has been described (4). The ancestral segment is divided into two parts by the modern segment; in all four alleles there are 153 codons in the 5′ part of the ancestral segment and 45 codons in the 3′ part. Thus, the modern segment is in exactly the same position with respect to the ancestral segment in all four alleles.

The modern segment consists of 39 repeats in the human and 44, 41, and 30 repeats in the large, medium, and small alleles, respectively, of the gorilla. We have divided the modern segment into early, middle, and late regions. Comparison of the genes of the two species shows close correspondence of the repeats of the early and middle regions. In the medium and large gorilla alleles, the two regions contain a total of 29 repeats, 1 more than the 28 repeats of the human. This results from the presence in the gorilla alleles of an extra repeat not present in the human (repeat 15 in the medium and large alleles). There is also an extra glutamic acid codon (GAG) in repeat 17 of the medium and large gorilla alleles but not in the corresponding position in repeat 16 of the human. In the small gorilla allele, a block of 15 repeats, presumably including the extra repeat and the repeat containing the extra GAG codon, appears to have been deleted. All three gorilla alleles, like the human allele, contain a three-codon deletion at the same position in repeat 13 in the middle region. The nucleotide divergence between the alleles for the different sections of the coding region is given in Table 1.

The Late Region. All repeats of the modern segment can be divided into two types, A and B (H. Tseng and H.G., unpublished data); the consensus type A repeat encodes Lys-His-Leu as the first three amino acids, while the consensus type B repeat encodes Glu-Leu-Pro. The last seven codons of both repeat types have the same consensus. Each repeat in the late region belongs to one of these classes but can be designated more precisely as described in Fig. 2.

Human	Gorilla Large	Gorilla Medium	Gorilla Small
*B	*B"	B"	*B"
*A	A'	A'	A'
*A	B	B	B
B	B'	B'	B'
B	A'	A'	A
B	B	B	
B	B	B	B
*B	B'	B'	B'
B^S	A	A	A
B^S	B'		B'
*B^S	A'		A
	B		
	B	B	B
	B'	B'	B'
	A	A	A
			B'
			A'
			B

```
A   (KHL) AAG CAC CTG/A GAG CAG CAG GAG GGG CAG CTG
A'  (KHL) AAG CAC CTG GAG CAG CAG AAG GGG CAG CTG
B   (ELP) GAG CTC CCA GAG CAG CAG GAG GGG CAG CTG
B'  (ELP) GAG CTC CCA GAG CAG CAG GAG GGA CAG CTG
B"  (ELP) GAG CTC CCA GAG CAG CGG GAG GGA CAG CTG
B^S (ELS) GAG CTC TCT GAG CAG CAG GAG GGG CAG CTG
```

FIG. 2. Late region of the modern segment. Amino acids encoded by the first three codons of each type of repeat are given in parentheses. The nucleotide change represented by each prime is underlined. The asterisks denote a repeat containing a single variable nucleotide change other than that indicated by the prime. In the human allele, a B^s indicates a B repeat with a serine codon in position 3.

The late region of the human contains two invariant blocks (4). One consists of five repeats of type B; the other consists of three repeats of a subtype designated as type B^s, which encodes Glu-Leu-Ser as the first three amino acids. The B^s repeat probably arose through modification of the third codon of a B repeat, leading to the substitution of a serine codon for a proline codon. Type B and B^s repeats constitute most of the late region of the human gene.

None of the gorilla alleles contains a late region resembling that of the human. None contains a block of invariant repeats nor any B^s repeats. Instead, the late regions of the gorilla alleles are built of duplicated blocks of mixed A and B repeats, and the final pattern is different in each allele. Nevertheless, features common to all gorilla late regions can be distinguished by a proper alignment of the repeats, as shown in Fig. 2.

A plausible scheme for the generation of the three late regions from a single block of six repeats is shown in Fig. 3. The late region of the large allele may be generated by two duplications, one of three repeats, and the other of six repeats. The late region of the medium allele may be derived from that of the large allele by deletion of a block of three repeats. The late region of the small allele may be derived from that of the large allele by one nucleotide substitution,

FIG. 1 (*on opposite page*). Coding regions of the three gorilla alleles and the single human allele. The coding regions consist of an ancestral segment divided into two parts by a modern segment containing 30–44 repeats of a 10-codon sequence. The modern segment, whose repeats are numbered from the 3′ end, consists of early, middle, and late regions; ♦ indicates the position of an extra GAG codon in repeat 17 of the medium and large alleles. These repeats are the only ones containing a sequence of consecutive GAG codons (five in the two gorilla alleles, and four in the corresponding human repeat 16). Repeats 15 in the large and medium alleles are identical and have no counterpart in the human allele. The two blocks of nearly invariant B and B^s repeats, which comprise most of the late region of the human allele, are boxed. The late region is different for each gorilla allele but always contains irregularly alternating A and B repeats (see also Fig. 2). Underlined bases in the human gene indicate mismatches between the human and the gorilla alleles. Underlined bases in the gorilla alleles indicate mismatches among different gorilla alleles.

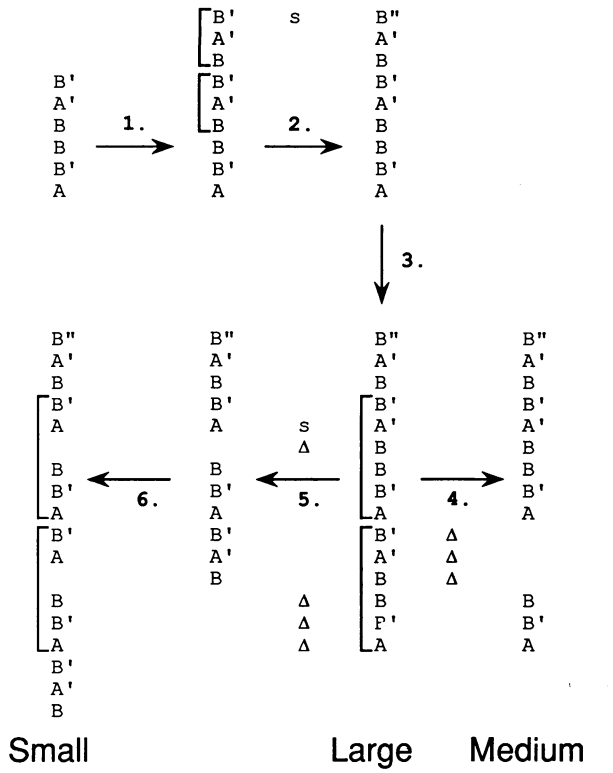

FIG. 3. Scheme for the generation of the late region of the three gorilla alleles. Steps: 1, duplication of a block of three repeats; 2, single nucleotide substitution; 3, duplication of a block of six repeats, yielding the large allele; 4, deletion of a block of three repeats, yielding the middle allele; 5, single nucleotide substitution and deletion of one repeat and of a block of three repeats; 6, duplication of a block of five repeats, yielding the small allele. s, Nucleotide substitution; Δ, Deletion of a repeat; brackets, duplicated blocks of repeats.

deletion of a single repeat and of a block of three repeats, and duplication of a block of five repeats.

In the early and middle regions of all alleles of gorilla and human, the most common repeat is of type A, whereas in the late regions, the most common repeat is of type B. Therefore, the duplications and deletions that generated the late regions have altered their ratio of B/A repeats. In the gorilla, this ratio changes from 0.3 in the early and middle region to 1.9 in the late region. That this occurred in the gorilla alleles is not surprising if, as proposed, their late regions were generated mainly from a duplicated block consisting of 4B repeats and 2A repeats. However, the human, whose pattern of duplication in the late region is different from that of the gorilla, has also formed a late region enriched in B repeats (B/A = 4.5). Thus, both species exhibit a tendency toward change of repeat composition in the late region.

DISCUSSION

In an ancestor of all higher primates, the early region of the modern segment was created by successive, intraexonic duplications. The early region probably became stabilized early in the history of the higher primates, since it is present in the owl monkey (H. Tseng and H.G., unpublished data). The middle region, which is not present in the owl monkey, was created or stabilized in some later ancestor of human and gorilla and has undergone only small changes after their separation. The late regions of human and gorilla, where the major differences lie, formed by means of different patterns of duplications and deletions after separation of their lineages.

Tandemly repeated short sequences are present in noncoding regions of animal genomes (13) and may be highly polymorphic, the frequency of gain or loss of repeats being as high as 5% per generation (14). The coding regions of a number of genes also contain short tandem repeats (15–18), and some such genes are highly polymorphic (19). In the yeast RNA polymerase II gene, some repeats are required for proper function of the protein (20). Repeats in a protein can promote interactions with other proteins (21). Perhaps, in the case of involucrin, repeats are important for the cross-linking process. This suggestion is supported by the recent finding that the guinea pig vesicular clotting protein (SVP-1), which is known to become highly cross-linked by an extracellular transglutaminase (22), is also composed of multiple short repeats (23). Although the repeats are not homologous to those of involucrin, they do, like those of involucrin, contain conserved lysine and glutamine residues, and the number of repeats correlates with the cross-link frequency.

The involucrin gene is unusual in that so much of its coding region was created within the primate lineage and in that a significant part was created within quite recent times. It is noteworthy that in both gorilla and human, the most recently created sequence is mainly confined to the same region. The evolutionary process by which this directed expansion occurred may still be active, a suggestion supported by the fact that the two gorillas examined contained three alleles, each polymorphic in the repeat pattern in the late region. The expansion of the involucrin molecule through the addition of repeats has likely conferred distinctive properties on the primate epidermis.

This work was aided by postdoctoral fellowship PF-2924 from the American Cancer Society (to J.T.), a grant from the National Cancer Institute (to H.G.), and a gift from Johnson and Johnson.

1. Rice, R. & Green, H. (1979) *Cell* **18**, 681–694.
2. Banks-Schlegel, S. & Green, H. (1981) *J. Cell Biol.* **90**, 732–737.
3. Watt, F. M., Jordan, P. W. & O'Neill, C. O. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 5576–5580.
4. Eckert, R. L. & Green, H. (1986) *Cell* **46**, 583–589.
5. Tseng, H. & Green, H. (1988) *Cell* **54**, 491–496.
6. Rheinwald, J. R. & Green, H. (1977) *Nature (London)* **265**, 421–424.
7. Simon, M. & Green, H. (1985) *Cell* **40**, 677–683.
8. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY).
9. Feinberg, A. P. & Vogelstein, B. (1983) *Anal. Biochem.* **132**, 6–13.
10. Parenteau, N. L., Eckert, R. L. & Rice, R. H. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 7571–7575.
11. Nicholls, R. D., Hill, A. V. S., Clegg, J. B. & Higgs, D. R. (1985) *Nucleic Acids Res.* **13**, 7569–7578.
12. Henikoff, S. (1984) *Gene* **28**, 351–359.
13. Southern, E. M. (1975) *J. Mol. Biol.* **94**, 51–69.
14. Jeffreys, A. J., Royle, N. J., Wilson, V. & Wong, Z. (1988) *Nature (London)* **332**, 278–281.
15. Argos, P., Pedersen, K., Marks, M. D. & Larkins, B. A. (1982) *J. Biol. Chem.* **257**, 9984–9990.
16. Koide, T., Foster, D., Yoshitake, S. & Davie, E. W. (1986) *Biochemistry* **25**, 2220–2225.
17. Galinski, M. R., Arnot, D. E., Cochrane, A. H., Barnwell, J. W., Nussenzweig, R. S. & Enea, V. (1987) *Cell* **48**, 311–319.
18. Weber, J. L. (1988) *Exp. Parasitol.* **66**, 143–170.
19. McLean, J. W., Tomlinson, J. E., Kuang, W.-J., Eaton, D. L., Chen, E. Y., Fless, G. M., Scanu, A. M. & Lawn, R. M. (1987) *Nature (London)* **330**, 132–137.
20. Nonet, M., Sweetser, D. & Young, R. A. (1987) *Cell* **50**, 909–915.
21. Kochan, J., Perkins, J. & Ravetch, J. V. (1986) *Cell* **44**, 689–696.
22. Notides, A. C. & Williams-Ashman, H. G. (1967) *Proc. Natl. Acad. Sci. USA* **58**, 1991–1995.
23. Moore, J. T., Hagstrom, J., McKormick, D. J., Harvey, S., Madden, B., Holicky, E., Stanford, D. R. & Wieben, E. D. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 6712–6714.