Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 8447–8451, November 1989
Evolution

Vectorial expansion of the involucrin gene and the relatedness of the hominoids

PHILIPPE DJIAN AND HOWARD GREEN

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

Contributed by Howard Green, July 17, 1989

ABSTRACT **In higher primates, the coding region of the gene for involucrin, an epidermal protein, is mostly composed of a recently generated (modern) segment of repeats of a sequence of 10 codons. While the rest of the coding region has evolved only by nucleotide substitutions, the modern segment has evolved by successive addition of repeats. This process has not taken place randomly; instead, the expansion of the modern segment has been progressive from 3′ to 5′ end, thus adding vectorially regions that have been defined as early, middle, and late. The relatedness of the human, chimpanzee, and gorilla may be analyzed with greatest sensitivity by comparing their middle regions. The chimpanzee involucrin gene is more closely related to that of the gorilla than to that of the human.**

Involucrin is a polypeptide substrate of keratinocyte transglutaminase (1, 2). During terminal differentiation of the cell, the enzyme is activated and involucrin becomes cross-linked to other proteins, forming an insoluble cell envelope.

The nucleotide sequence of the involucrin gene is known for the human (3), gorilla (4), orangutan (5), and owl monkey (6). The coding region of the gene contains a modern segment of repeats. This segment was generated within the higher primate lineage in stages that could be defined as early, middle, and late. The early region was completed in a common ancestor of all anthropoid primates, since it is present in species as widely separated as the human and the owl monkey (6). The middle region of the hominoids developed after divergence of their lineage from that of the owl monkey, since the involucrin gene of the latter does not contain a corresponding middle region. The late region must have been generated after the divergence of the hominoid species from each other, since it is unique in each hominoid so far examined (3–5).

We now report the nucleotide sequence of the involucrin gene of the chimpanzee (*Pan*)* and examine the relatedness of the segment of repeats in this gene to that of the human and gorilla. Relatedness is based on two criteria: similarity of repeat pattern and coincidence of marker nucleotides. Since the repeats of the modern segment are nearly identical to a consensus sequence, a non-consensus nucleotide marks a repeat and may therefore be used to establish that two repeats have a common origin.

MATERIALS AND METHODS

Vaginal biopsies were performed at the Yerkes Primate Center (Atlanta) on a pygmy chimpanzee (Laura) and a common chimpanzee (Garbo) and provided to R. H. Rice (Harvard School of Public Health), who prepared keratinocyte cultures, as described earlier (7, 8). A restriction map was derived from their DNA (5).

To obtain the pygmy chimpanzee gene, genomic DNA was digested with *Xba* I and *Hin*dIII, and the resulting fragments in the range of 4.5 kilobases (kb) were cloned (5). For the involucrin gene of the common chimpanzee, the same procedure was used except that genomic DNA was digested with *Bam*HI and fragments of 1.7 kb were cloned. The nucleotide sequences of the 2.5-kb *Xba* I–*Bam*HI fragment of the pygmy chimpanzee and the 1.66-kb *Bam*HI fragment of the common and the pygmy chimpanzee were determined (5).

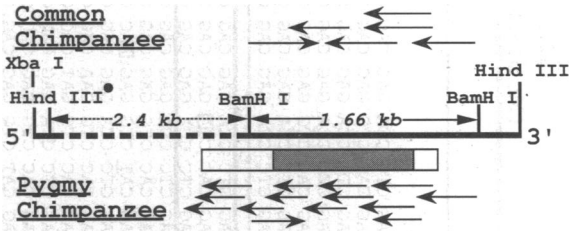

FIG. 1. Sequencing of the chimpanzee involucrin genes. The entire involucrin gene of the pygmy chimpanzee is included in the 4.5-kb *Xba* I–*Hin*dIII fragment. The dashed line indicates that the 2.4-kb *Hin*dIII–*Bam*HI fragment is not drawn to scale. The dotted *Hin*dIII site is found in the common chimpanzee but not in the pygmy chimpanzee. The coding region (boxed) contains 560 codons, slightly fewer than that of the human, in accord with the slightly smaller size of the protein as indicated by electrophoresis through polyacrylamide gels (9). Lying within the coding region is the segment of repeats (dotted). The arrows indicate the sequenced part of overlapping DNA clones made from deleted fragments.

RESULTS

Restriction Map of the Chimpanzee Involucrin Gene. Over the 25 kb containing the involucrin gene and its flanking DNA, the restriction map of the chimpanzee is similar to that

```
                                        ATG TCC CAG
    CAA CAC ACA CTG CCA GTG ACC CTC TCC CCT
    GCC CTC AGT CAG GAG CTC CTC AAG ACT GTT
    CCT CCT CCA GTC AAT ACC CAG CAG GAG CAA
    ATG AAA CAG CCA ACT CCA CTG CCT CCC CCA
    TGC CAG AAG ATG CCT GTC GAG CTC CCA GTG
    GAG GTC CCA TCA AAG CAA GAG GAA AAG CAC
    ATG ACT GCT GTA AAG GGA CTG CCT GAG CAA
5'  GAA TGT GAG CAA CAG CAG CAG GAG CCA CAG
    GAG CAG GAG CTG CAG CAA CAG CAC TGG GAA
    CAG CAT GAG GAA TAT CAG AAA GCA GAA AAC
    CCA GAG CAG CAG CTT AAG CAG GAG AAA GCA
    CAA AGG GAT CCG CAG CTA AAC AAA CAG CTG
    GAA GAA GAG AAG AAG CTC TTA GAC CAG CAA
    CTG GAT CAA GAG CTA GTC AAG AGA GAT GAG
    CAA CTG GGA ATG AAG AAA GAG CAA CTG TTG
                Segment of Repeats
    GAG CAG CCT GTG TTT GCC CCA GCT CCA GGC
    CAG GTC CAA GAC ATT CAA TCA GCC CTG CCC
3'  ACA AAG GGA GAA GTA TTG CTT CCC CTA GAG
    CAC CAG CAG CAG AAG CAG GAG GTG CAG TGG
    CCA CCC AAA CAT AAA TAA
```

FIG. 2. Coding region flanking the segment of repeats in the chimpanzee. Overlined nucleotides differ from those of either human or gorilla; underlined nucleotides differ from both.

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 sequence reported in this paper for *Pan paniseus* has been deposited in the GenBank data base (accession no. M26514).

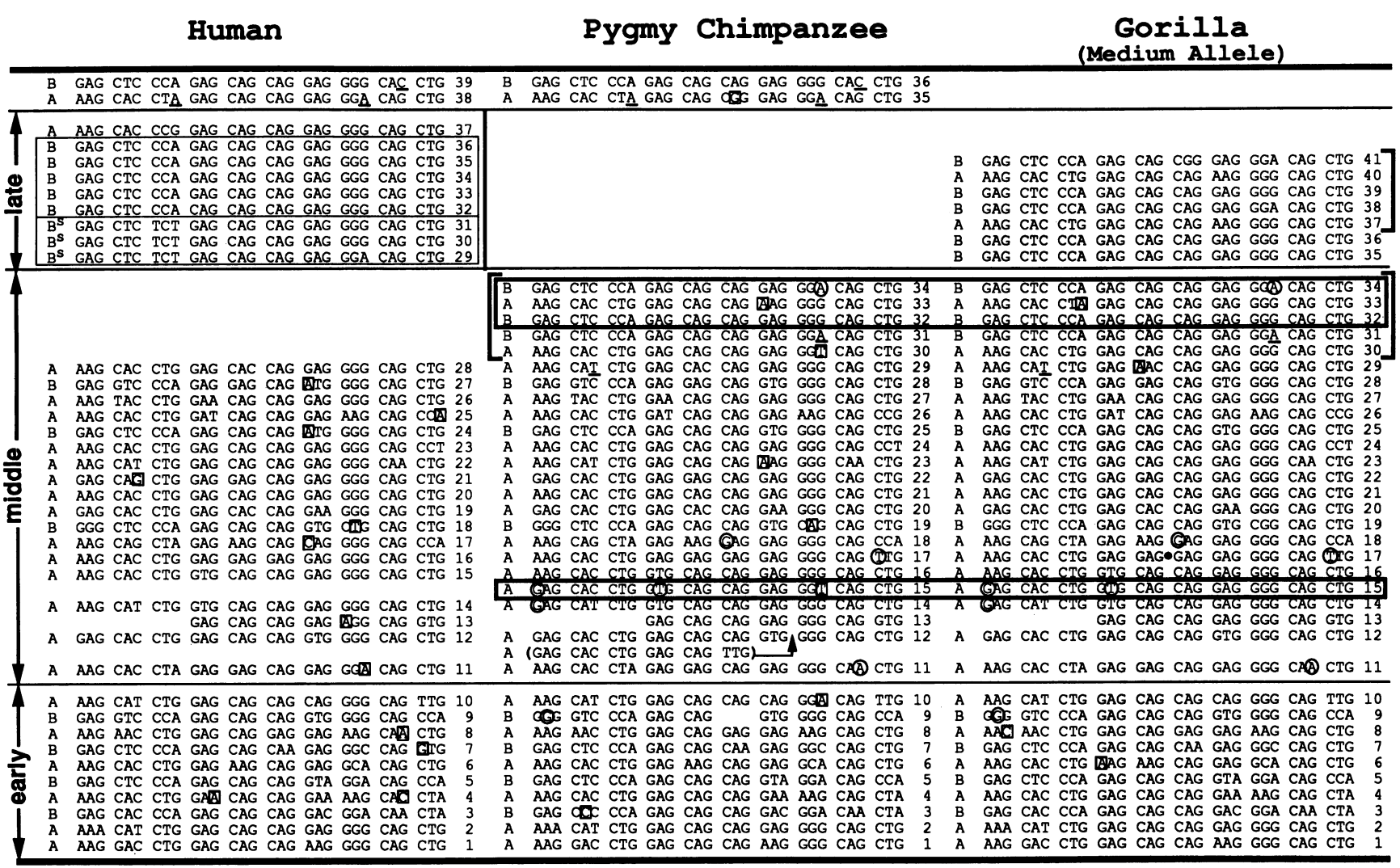

FIG. 3. Segment of repeats. This segment is divided into early, middle, and late regions. Repeats are numbered in the 3′ to 5′ direction and are designated as A or B. Repeat 12 of chimpanzee contains a fragment of an A repeat inserted between codons 7 and 8 (bracketed and the place of insertion indicated by an arrow). The solid dot in repeat 17 of the gorilla indicates the insertion of an extra GAG codon (4). Uniquely shared repeats 15 and 32–34 of chimpanzee and gorilla are framed. Repeats 30–34 constitute a BABBA block (bracketed). These repeats of the gorilla, previously assigned to the late region (4), are now moved to the middle region because of their correspondence with the chimpanzee repeats. Repeats 37–41 of the gorilla late region (bracketed) also form a BABBA block. The two quasi-invariant blocks of repeats that compose most of the late region of the human are framed; these cannot be aligned with the late region of the gorilla, as indicated by the vertical bar. B^s indicates a modified B repeat encoding a serine (TCT) in the third position and found only in the human late region. All marker nucleotides shared only by the chimpanzee and gorilla are circled. All marker nucleotides shared by two of the three species but also by the orangutan (5) and/or *M. fascicularis* are underlined. All marker nucleotides present in only one species are boxed.

of the human and gorilla. However, both the common chimpanzee (*Pan troglodytes*) and the pygmy chimpanzee (*Pan paniscus*) possess a *Bam*HI site 344 nucleotides 3′ of the translation initiation codon, and the common chimpanzee alone possesses a *Hin*dIII site located 5′ of the transcription start site (Fig. 1).

Coding Region of the Chimpanzee Gene. The parts of the coding region located 5′ and 3′ of the segment of repeats are shown in Fig. 2. As in the other hominoids, the two parts contain a total of 198 codons and are separated at the same point by the segment of repeats. There were no nucleotide differences between the common and pygmy chimpanzees in the protein-encoding regions flanking the segment of repeats that have been sequenced in both animals.

The Modern Segment of Repeats: Early and Middle Regions. The segment of repeats of the pygmy chimpanzee is compared in Fig. 3 with the previously published segments of repeats of the human (3) and of the medium allele of the gorilla (4). Repeats are classified as A or B. The first three codons in the A repeat have the consensus sequence AAG CAC CTG and in the B repeat the consensus sequence GAG CTC CCA; the last seven codons are the same in A and B repeats (4, 6). In the hominoids, the nucleotide sequence of the repeats is 91–93% identical to either the A or B consensus.

The early regions of the genes in the two chimpanzees differ at three nucleotide positions, and these are the only differences between the modern segments of the two genes. The common chimpanzee possesses A instead of C in the second position of the second codon of repeat 3, C instead of T in the second position of the third codon of repeat 4, and A instead of C in the second position of the third codon of repeat 9.

The early region of the pygmy chimpanzee has been aligned, repeat for repeat, with the early regions of the human and gorilla (Fig. 3). There has been a deletion of a single CAG codon in repeat 9 of the chimpanzee. Since this codon is present in the early regions of all the other hominoids, including the orangutan (5), and in the owl monkey as well (6), the deletion must have occurred in the chimpanzee lineage after its separation from the other hominoid lineages. A coincident marker (non-consensus) nucleotide is present in the first codon of repeat 9 of the chimpanzee and gorilla genes.

In the middle region, most repeats of the chimpanzee can be matched with corresponding repeats in the human and gorilla. However, there is more extensive repeat-sharing between the chimpanzee and the gorilla. Repeat 15 of these species, a probable duplicate of repeat 14, is not present in either the human or the orangutan (Fig. 4). The gibbon also lacks this repeat but possesses repeats corresponding to those immediately above and below (unpublished data). Similarly, repeats 32–34 of the chimpanzee and gorilla are not present in the other hominoids. We conclude that repeats 15 and 32–34 were created in a common proximate ancestor of the gorilla and chimpanzee. Repeats 30–34 of chimpanzee and gorilla have the distinctive pattern BABBA.

Further evidence that the chimpanzee gene is closer to that of the gorilla than to that of the human is provided by coincidences of marker nucleotides (Fig. 3). There is no marker nucleotide shared by the human and chimpanzee alone. But in addition to one in repeat 9 of the early region, there are seven coincident marker nucleotides shared by the middle regions of chimpanzee and gorilla alone (in repeats 11, 14, 15, 17, 18, and 34). There are two coincident marker nucleotides in repeats 29 and 31 of the chimpanzee and gorilla, but they are not specific to the African apes since they are also present in the orangutan.

The orangutan lacks repeats 16, 21, 24, and 32–34 of the African apes but contains an extra repeat circled in Fig. 4. Repeats corresponding to 30 and 31 of the African apes are present in the orangutan (and the gibbon) but must have been deleted in the human.

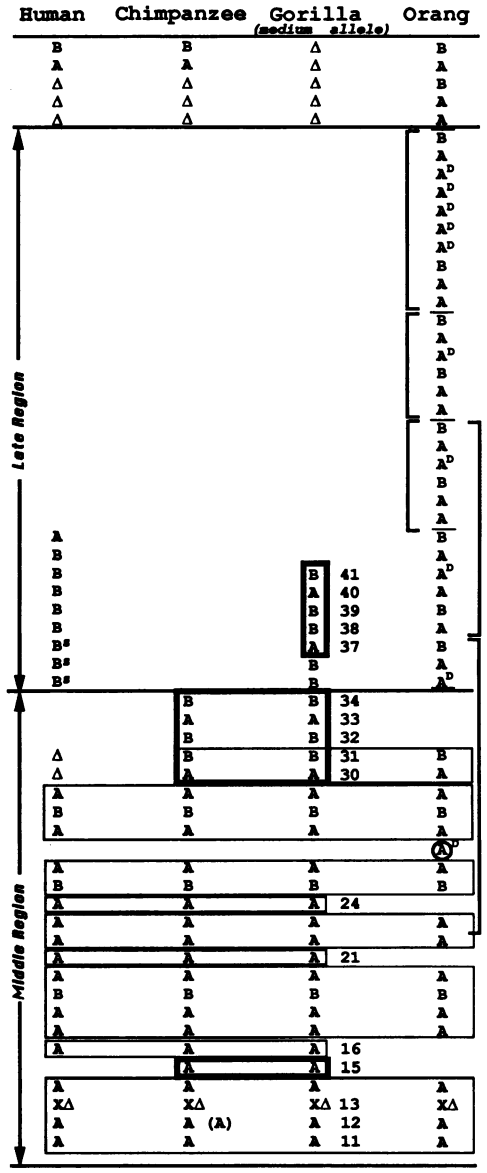

FIG. 4. Repeat patterns in the middle and late regions. Each 10-codon repeat is designated as A or B, except for repeat 13, which lacks the first 3 codons in all species and is designated XΔ. (A) is the partial A repeat containing 6 codons inserted into repeat 12 of the chimpanzee. Repeats present in more than one species are framed. The middle regions of the chimpanzee and gorilla and the late region of the gorilla contain the pattern BABBA (also framed). Brackets show block duplications in the orangutan. A^D indicates an A repeat in which the fourth codon has been changed from GAG to GAT, encoding aspartic acid. The only repeat unique to the orangutan middle region is circled.

Late Region. The late region is different in each hominoid. The late region of the human is mostly composed of two blocks of quasi-invariant repeats, one consisting of five B repeats, the other of three modified B repeats (B^s) (4).

The chimpanzee gene is unusual in that it contains no late region at all. The late region of the gorilla does not resemble that of the human. Instead, it consists of mixed A and B repeats and includes the pattern BABBA, which is present at the 5′ end of the middle region in both gorilla and chimpanzee.

In the medium allele of the gorilla [one of three polymorphic alleles of that species sequenced (4)], BABBA occurs in

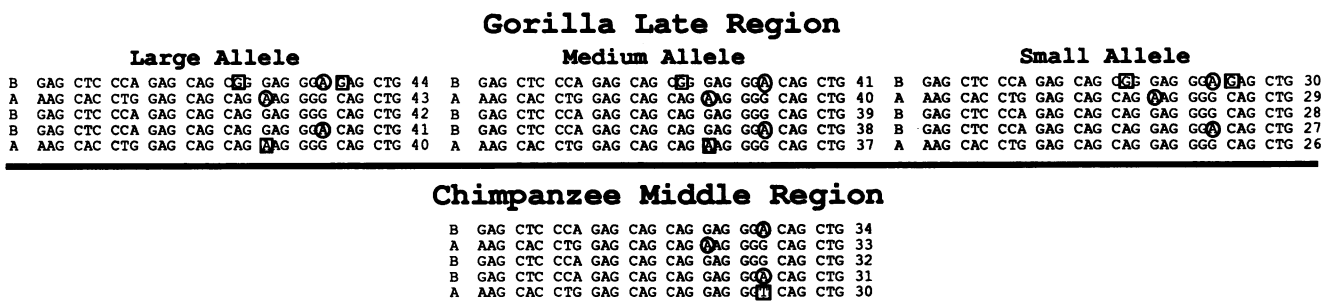

FIG. 5. BABBA blocks of gorilla and chimpanzee. One BABBA block of the late region of every gorilla allele is shown. For each, the repeat numbers are indicated on the right. Three marker nucleotides present at the same positions in the BABBA block of the chimpanzee are circled. Marker nucleotides not shared by chimpanzee and gorilla are boxed. The nucleotide sequence of the gorilla BABBA blocks is from Teumer and Green (4).

repeats 37–41. A BABBA block also appears in the late regions of both small and large alleles of the gorilla. The BABBA blocks in the late regions of the three gorilla alleles (Fig. 5) are clearly related to the BABBA block of the middle region.† This repeat pattern occurs nowhere else in the coding region of the gorilla or the chimpanzee. It occurs nowhere in the entire coding region of the human or orangutan, nor does it occur in the involucrin gene of the gibbon or the Old World monkey *Macaca fascicularis* (unpublished data).

Further evidence of the relatedness of the BABBA blocks is provided by coincidences of marker nucleotides. The BABBA blocks of the gorilla late region share coincident marker nucleotides not only with the BABBA block of the gorilla middle region but also with the BABBA block of the chimpanzee middle region (Fig. 5). The BABBA repeat pattern must have been generated in a lineage common only to the gorilla and the chimpanzee, after this lineage had separated from that of the human. After the divergence of the chimpanzee lineage from that of the gorilla, the latter alone added more duplications containing the BABBA block to form its late region.

Chimpanzee repeats 35 and 36 and human repeats 38 and 39 share three coincident marker nucleotides that are absent from the last two repeats of the gorilla alleles (Fig. 3) (4). These two repeats are part of a group of five repeats present in the orangutan (Fig. 4 and ref. 5) and in *M. fascicularis* (unpublished data) and therefore do not belong to the species-specific late region. Three of the five repeats have been deleted in the human and chimpanzee, and all five have been deleted in the gorilla.

Nucleotide Divergence Between the Coding Regions of the Three Species. In contrast with the informative nature of the repeat patterns, study of overall nucleotide divergence between the coding regions of the three species add little useful information (Table 1). The chimpanzee sequence is overall closer to that of the gorilla than to that of the human. The strength of the conclusion is limited by the small number of nucleotides compared, but the result is in accord with the much stronger evidence from the repeat patterns.

†In the large allele of the gorilla (4), repeats 30–35 might constitute a middle-region BABBA block in which an internal B repeat has been duplicated to produce BABBBA. This block shares three marker nucleotides with the chimpanzee BABBA block. In the small allele, the boundary between middle and late regions has been altered by a large deletion of the 5′ end of the middle region (4). Repeats 18–22 of this allele constitute a second BABBA block of the late region with only two of the three shared marker nucleotides. Repeats 15–17 of this allele, which share two marker nucleotides with the chimpanzee BABBA block, might be the remains of a partially deleted BABBA block of the middle region.

DISCUSSION

Relatedness of the Three Species by Different Criteria. There has been a long-continuing controversy as to whether the chimpanzee is more closely related to the human or to the gorilla. From their similarities in gross morphology and physiology, the chimpanzee and the gorilla were thought to be most closely related (10–14). Comparisons of the thermal stabilities of interspecific DNA hybrids (15), the sequence of $\Psi\eta$-globin pseudogenes (16), the sequence of the intergenic DNA of the β-globin gene cluster (17), and the mobility of cell proteins in two-dimensional electrophoresis (18) have favored a closer relatedness of the human and chimpanzee, whereas studies of mitochondrial DNA (19–21) have favored a closer relatedness of chimpanzee and gorilla. From comparisons of chromosome structure, the human and the chimpanzee (22, 23), the chimpanzee and the gorilla (24), and the human and the gorilla (25) have all been proposed as the most closely related. Comparisons of blood groups have not been informative (26). Critical examination of the evidence from these and other studies of genes and proteins has emphasized the difficulties in reaching conclusions as to the comparative relatedness of the three species (19, 20, 27–31).

We believe that the resolution inherent in earlier molecular comparisons of the hominoids was limited by lack of a gene that changed sufficiently during the short evolutionary history of these species. The involucrin gene provides the necessary resolution and its repeat structure is consistent only with a grouping of chimpanzee and gorilla. The similarities between these two species are summarized in Table 2.

Expansion of the Involucrin Gene as a Hitherto Undescribed Form of Evolution. In the hominoids, most protein-encoding genes differ mainly by a small number of nucleotide substitutions. The involucrin gene is of interest for the study of evolution in the higher primates because its coding region has

Table 1. Nucleotide divergence of the parts of the involucrin coding regions shared by chimpanzee, gorilla [medium allele], and human

Species	Nucleotide mismatches/total nucleotides	%
Human/pygmy chimpanzee	29/1422	2.0
Human/gorilla	32/1422	2.3
Pygmy chimpanzee/gorilla	19/1422	1.3
Common chimpanzee/ pygmy chimpanzee	3/1078	0.3

The entire coding region of the pygmy chimpanzee was sequenced; however, for the parts of the coding region flanking the segment of repeats, comparison with the common chimpanzee was incomplete because the sequence of 344 nucleotides extending upstream of the *Bam*HI site in this species was not determined.

Table 2. Points of closer similarity between chimpanzee and gorilla

Repeat pattern (middle and late regions)	
1. Repeats 15 and 32-34 present in chimpanzee and gorilla, but lacking in human and orang. No repeat shared by human and chimpanzee alone.	
2. Gorilla late region lacking homogeneous block of B repeats or any B^S repeats as in human; instead possesses alternating A and B repeats containing repeat pattern BABBA sharing coincident marker nucleotides with BABBA block of chimpanzee middle region.	
Marker nucleotides shared and not shared by corresponding repeats (early and middle regions)	
present in human alone	12
present in chimpanzee alone (omitting 3 in the insertion in repeat 12)	7
present in gorilla alone	4
present in orang alone	11
shared by human and chimpanzee alone	0
shared by human and gorilla alone	0
shared by human and orang alone	0
shared by gorilla and orang alone	0
shared by chimpanzee and orang alone	0
shared by chimpanzee and gorilla alone	8

been rapidly expanded by addition of repeats, in a process that has continued through multiple events of speciation. For this reason, the involucrin gene of higher primates should be considered a nascent gene, in contrast with mature mammalian genes whose evolution has been studied earlier. During each stage of the development of the modern segment, new repeats have been added mainly to a particular location and the process of repeat addition has been vectorial.

This order makes it possible to distinguish between conservative and innovative periods in the evolution of different regions of the modern segment. All 10 repeats of the early region must have been generated in a common anthropoid ancestor (6); the early region then became conservative in that no repeats were added or deleted during the subsequent evolution of man and the great apes, while at the same time the entire middle region was generated by duplications. This is shown in Fig. 4, where the repeats of all the great apes and the human are aligned. In the middle regions, 15 repeats are common to all, 3 are common only to the African apes and man, 4 are common only to the chimpanzee and the gorilla, none is common only to human and chimpanzee, and 1 is present only in the orangutan.

The late region, in contrast to the middle region, was generated after separation of the hominoid lineages from each other. This region varies greatly in size between different species, the chimpanzee being the only hominoid lacking it altogether. The polymorphism of the late region in the gorilla (4), owl monkey (6), and Old World monkeys (9) suggests that in many species this region is still in an innovative phase.

The modern segment of the involucrin gene should permit analysis of lineage branching throughout the evolution of higher primates. The rapid expansion of the involucrin gene and the vectorial nature of that expansion have made this gene specially important for the evaluation of relatedness from DNA sequence.

This investigation was aided by a grant from the National Cancer Institute.

1. Rice, R. H. & Green, H. (1979) *Cell* **18,** 681–694.
2. Simon, M. & Green, H. (1988) *J. Biol. Chem.* **263,** 18093–18098.
3. Eckert, R. L. & Green, H. (1986) *Cell* **46,** 583–589.
4. Teumer, J. & Green, H. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 1283–1286.
5. Djian, P. & Green, H. (1989) *Mol. Biol. Evol.* **6,** 469–477.
6. Tseng, H. & Green, H. (1989) *Mol. Biol. Evol.* **6,** 461–468.
7. Rheinwald, J. & Green, H. (1975) *Cell* **6,** 331–344.
8. Simon, M. & Green, H. (1985) *Cell* **40,** 677–683.
9. Parenteau, N. L., Eckert, R. L., & Rice, R. H. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 7571–7575.
10. Tuttle, R. H. (1969) *Science* **166,** 953–961.
11. Martin, L. (1985) *Nature (London)* **314,** 260–263.
12. Ciochon, R. L. (1985) in *Primate Evolution and Human Origins,* eds. Ciochon, R. L. & Fleagle, J. C. (Benjamin/Cummings, Menlo Park, CA), pp. 345–362.
13. Gantt, D. G. (1986) in *Comparative Primate Biology,* eds. Swindler, D. R. & Erwin, J. (Liss, New York), Vol. 1, pp. 453–475.
14. Andrews, P. (1987) in *Molecules and Morphology in Evolution: Conflict or Compromise?* ed. Patterson, C. (Cambridge Univ. Press, Cambridge), pp. 23–53.
15. Sibley, C. G. & Ahlquist, J. E. (1987) *J. Mol. Evol.* **26,** 99–121.
16. Miyamoto, M. M., Slightom, J. L. & Goodman, M. (1987) *Science* **238,** 369–373.
17. Maeda, N., Wu, C.-I., Bliska, J. & Reneke, J. (1988) *Mol. Biol. Evol.* **5,** 1–20.
18. Goldman, D., Giri, P. R. & O'Brien, S. J. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 3307–3311.
19. Ferris, S. D., Wilson, A. C. & Brown, W. M. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 2432–2436.
20. Hixson, J. E. & Brown, W. M. (1986) *Mol. Biol. Evol.* **3,** 1–18.
21. Templeton, A. R. (1983) *Evolution* **37,** 221–244.
22. Dutrillaux, B., Rethore, M. O. & Lejeune, J. (1975) *Ann. Genet.* **18,** 153–161.
23. Yunis, J. J. & Prakash, O. (1982) *Science* **215,** 1525–1530.
24. Bianchi, N. O., Bianchi, M. S., Cleaver, J. E. & Wolff, S. (1985) *J. Mol. Evol.* **22,** 323–333.
25. Miller, D. A. (1977) *Science* **198,** 1116–1124.
26. Socha, W. W. (1986) in *Comparative Primate Biology,* eds. Swindler, D. R. & Erwin, J. (Liss, New York), Vol. 1, pp. 299–333.
27. Mai, L. L. (1983) in *New Interpretations of Ape and Human Ancestry,* eds. Ciochon, R. L. & Corruccini, R. S. (Plenum, New York), pp. 87–114.
28. Nei, M., Stephens, J. C. & Saitou, N. (1985) *Mol. Biol. Evol.* **2,** 66–85.
29. Ruvolo, M. & Smith, T. F. (1986) *Mol. Biol. Evol.* **3,** 285–289.
30. Holmquist, R., Miyamoto, M. M. & Goodman, M. (1988) *Mol. Biol. Evol.* **5,** 201–216.
31. Marks, J., Schmid, C. W. & Sarich, V. M. (1988) *J. Hum. Evol.* **17,** 769–786.