A large part of the coding region of the hominoid involucrin gene is of recent origin. This part of the gene, which we have called the modern segment, contains numerous repeats of a sequence of 10 codons, created by multiple duplications some of which consist of 3–12 repeats. We have sequenced two alleles of the involucrin gene in the owl monkey and found that the involucrin gene of this species also possesses a modern segment. By comparing the modern segment of the owl monkey with that of the hominoids, we find that only a part of this segment is shared by the two species. We call this part the early region because it must have originated in a common ancestor of the anthropoids. The rest of the hominoid modern segment does not correspond to any groups of repeats in the owl monkey and was therefore created after divergence of the two lineages. As in the hominoids, the latest additions to the modern segment of the owl monkey have been in its 5’ half, which possesses different duplication patterns in the two alleles. Lineage divergences within the anthropoids can be detected at different sites within the modern segment.

Introduction

The protein involucrin is synthesized during the programmed terminal differentiation of the keratinocytes of the epidermis and other stratified squamous epithelia (Rice and Green 1979). Involucrin is a substrate for epidermal transglutaminase and becomes incorporated into the cross-linked envelope of the corneocyte, thus contributing to the resistance of the epidermis.

In all mammals so far examined, the coding region of the involucrin gene contains a segment of contiguous repeats of a short sequence. In the human (Eckert and Green 1986), the gorilla (Teumer and Green 1989), and the orangutan (Djian and Green 1989), the segment of repeats is located at the same site within the coding region. This segment of repeats is absent from the involucrin gene of the lemur, the galago, and the pig, whose segments of repeats differ in sequence, in repeat length, and in location within the coding region (Tseng and Green 1988; M. Phillips, P. Djian and H. Green, unpublished experiments). We concluded that the segment of repeats found in the hominoids must have originated in the higher primates, and we therefore called it the modern segment of repeats, to distinguish it from the older segment of repeats present in lower animals. We have called the rest of the coding region, which flanks the segment of repeats, the ancestral segment, because it is shared, in large part, by all animals so far examined and has given rise at different times to the two segments of repeats.

It seemed reasonable that the involucrin gene of the owl monkey (Aotus trivir-
An esophageal biopsy from an owl monkey (*Aotus trivirgatus*) was provided by Dr. Norvall King of the New England Regional Primate Center to Dr. Robert Rice of the Harvard School of Public Health. Fibroblasts cultured from this biopsy by Dr. Rice were used as a source of DNA, prepared by Dr. Richard Eckert. A detailed description of the reagents and procedures has been given by Tseng and Green (1988).

**Restriction Mapping**

A restriction map was generated by cutting owl monkey genomic DNA with various single enzymes or pairs of enzymes. The digested DNA was then size-fractionated by electrophoresis through a 0.8% agarose gel, transferred to nylon membrane, and subjected to Southern analysis (Southern 1975) using as a probe the human involucrin-coding region (Tseng and Green 1988).

**Cloning**

The involucrin genes of owl monkey were cloned by the method of Nicholls et al. (1985). An *XbaI-HindIII* fragment of 4.2–4.4 kb containing the entire coding region was chosen for cloning. Two hundred micrograms genomic DNA were digested to completion with 500 units *XbaI* and 500 units *HindIII* in an overnight reaction, and the digested DNA was separated by electrophoresis through a preparative gel of 0.8% agarose. The parts of the gel containing DNA of 4.0–4.2 kb, 4.2–4.4 kb, and 4.4–4.6 kb were excised, and the DNAs electrophoresed from the gel slices were purified by diethylaminoethyl cellulose chromatography and ethanol precipitation. One-quarter of each fraction of DNA was examined for the presence of the involucrin gene by Southern analysis. The 4.0–4.2-kb fraction was found to be enriched for the small (S) allele, and the 4.2–4.4-kb fraction was enriched for the large (L) allele. One-tenth of
the DNA of each fraction was ligated with the vector pGEM-3 (Promega) digested with *Xba*I and *Hind*III. Transformants obtained by introducing the ligated DNA into competent *Escherichia coli* strain DH5α (Bethesda Research Laboratories) were divided into 10 pools (five for the L allele and five for the S allele), and DNA was prepared. Pools containing the involucrin gene were identified by Southern analysis of their DNA, by using the human involucrin probe. The corresponding transformants were identified by colony hybridization at low density. DNA from well-isolated positive colonies was subjected to restriction and Southern analysis to confirm the presence of the intact involucrin gene.

**Sequencing**

The plasmid containing the L allele [(pOM-I)X4.2H] was digested with *Bgl*II, which cleaves at the 5′ end of the coding region (fig. 1). The linearized DNA was then subjected to BAL31 exonuclease and digested with *Sma*I. The larger fragment was purified by electrophoresis through a 1% agarose gel and was recircularized. The nested deletions created by this procedure were sequenced from the T7 priming site by using reverse transcriptase of avian myeloblastosis virus.

The plasmid containing the S allele [(pOM-I)X4.0H] was digested by *Kpn*I and *Xba*I, and deletions were generated with exonuclease III (Henikoff 1984) by using a kit (Erase-a-base) made by Promega Biotec, Inc. Reaction conditions were those recommended by the manufacturer.

The two alleles were sequenced in the opposite direction by first subcloning the *Hinc*II-*Bam*HI fragments from (pOM-I)X4.2H and (pOM-I)X4.0H (fig. 1) into pGEM-3 cut with the same enzymes. This reverses the orientation of the coding sequence relative to the T7 primer site in the vector. The deletions were then made by treatment with *Exo*III after digestion with *Kpn*I and *Bam*HI.

Sequencing was done by the chain termination method (Sanger et al. 1977) by using a Sequenase kit (United States Biochemicals) and the conditions suggested by the manufacturer. Each nucleotide was sequenced at least once in each direction.

**Results**

Two Involucrin Alleles in the Owl Monkey

Restriction analysis of genomic DNA from a single owl monkey resulted in the map shown in figure 1. At the involucrin locus there were two alleles with a size difference of 150–200 bp. The two alleles were cloned, and the coding regions were sequenced. The keratinocytes of the owl monkey synthesized two involucrin molecules of different molecular weight (Parenteau et al. 1987), and, as expected, the L and S alleles differed in the number of repeats in the modern segment.

The Coding Region Flanking the Modern Segment of Repeats

The coding region of both alleles has the same basic structure as that of the human gene (fig. 2). It is divided at the same point into two parts by the segment of repeats. That part located 5′ of the segment of repeats is of equal length in human and owl monkey, but each species has an extra codon in a different position. That part located 3′ of the segment of repeats begins and terminates similarly, but the owl monkey alleles lack two codons present in the human. Comparing the combined 5′ and 3′ parts of the L allele with those of the human gives a mismatch frequency of 9.0%. The 5′ parts of the two alleles of the owl monkey differ in five nucleotides; the 3′ parts are identical (overall divergence 0.8%).
The Modern Segment of Repeats: A and B Repeats

This segment consists of two types of repeats designated A and B (figs. 2, 3, table 1). The consensus sequences of the A and B repeats are identical for the last seven codons, but the first three codons (AAG, CAC, and CTG) of the A repeat encode...
Table 1
Codon Frequency at Each Position in A and B Repeats (L/S allele)

<table>
<thead>
<tr>
<th>CODON POSITION</th>
<th>A repeat (L)</th>
<th>B repeat (S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AAG</td>
<td>GAG</td>
</tr>
<tr>
<td>2</td>
<td>CAT</td>
<td>CTC</td>
</tr>
<tr>
<td>3</td>
<td>CAG</td>
<td>CCA</td>
</tr>
<tr>
<td>4</td>
<td>GCC</td>
<td>CCA</td>
</tr>
<tr>
<td>5</td>
<td>CAG</td>
<td>GTA</td>
</tr>
<tr>
<td>6</td>
<td>CCA</td>
<td>GAG</td>
</tr>
<tr>
<td>7</td>
<td>GCA</td>
<td>GTA</td>
</tr>
<tr>
<td>8</td>
<td>CCA</td>
<td>GGT</td>
</tr>
<tr>
<td>9</td>
<td>CCA</td>
<td>CTA</td>
</tr>
<tr>
<td>10</td>
<td>CCA</td>
<td>CTG</td>
</tr>
</tbody>
</table>

NOTE.—The underscored sequences are consensus sequences.

Because of two deletions, position 5 of B repeats contains fewer codons.

KHL, and the first three codons (GAG, CTC, and CCA) of the B repeat encode ELP. In the L allele, the B repeats are somewhat less divergent from their nucleotide consensus (8.7%) than are the A repeats from their consensus (12.9%); the difference is greater at the protein level, for the B repeats deviate from their consensus at 16% of amino acids, whereas the A repeats deviate at 30%. The figures are similar for the S allele. GAG at positions 1 and 4, CAG at positions 5, 6, and 9, and GGG at position 8 have been nearly uniformly preserved in B repeats but have been frequently substituted in A repeats (table 1). An exception is at position 7, where the codon GAG has been more conserved in the A repeats than in the B repeats.

The Early Region

At the extreme 3' end of the modern segment, both alleles of the owl monkey possess 10 repeats that correspond closely to the first 10 repeats of the human modern segment (fig. 2). We call this part of the modern segment the early region. The resemblance between early regions of human and owl monkey is made clear by the identical pattern of alternation of A and B repeats and by coincidence of the same nonconsensus nucleotide (marker nucleotide) in identical positions. Some marker nucleotides appear twice at corresponding positions in both genes; for example, T occurs twice in the second position of codon 7 (repeats 5 and 9), AA occurs twice in the first and second positions of codon 8 (repeats 4 and 8), and CA occurs twice in the second and third positions of codon 10 (repeats 5 and 9). Thus, in some common ancestor of human and owl monkey, four consecutive repeats in the early region were tandemly duplicated (blocks I and II of fig. 2), thus placing marker nucleotides at two corresponding positions in the two blocks of each species. One codon was later deleted from block I of the owl monkey but not from the corresponding position of block II.
FIG. 3.—Comparison of duplication patterns in the L and S alleles. The sequences of the two alleles are aligned according to their repeat patterns. The framed repeat blocks occurring in one or both alleles are numbered I–VIII. An additional duplication of two repeats present in both alleles is enclosed by a dashed frame. Coincident marker nucleotides in a pair of duplicated blocks are boxed, except for blocks I and II. Coincident marker nucleotides in repeat blocks of early regions of both the L allele and the human allele are shown in fig. 2.
Other marker nucleotides appear in only one of the duplicate blocks I or II of both species and must therefore have been generated after duplication of the four repeats. For example, G is the first nucleotide of the sixth codon only once in the 10 repeats of the early region, and it is present in repeat 8 of both species. Similarly, A is the third nucleotide of the sixth codon only once, and it is present in repeat 7 of both species. The same applies to C in the second position of the eighth codon in repeat 6. Thus, some divergence had occurred between the two blocks before separation of the human and owl monkey lineages.

The early region of the S allele is identical with that of the L allele (fig. 3), except for two mismatches of the third nucleotide, one in repeat 5 and the other in repeat 7.

Middle and Late Regions

These two regions were defined earlier by comparison of the human and gorilla genes (Teumer and Green 1989). In the owl monkey, there is no group of repeats that can be matched to the middle or late regions of the human gene (fig. 2). The pattern of A and B repeats and the positions of marker nucleotides are different in the two species, and there is no three-codon deletion in repeat 13 of the owl monkey, as there is in the hominoids. By comparing the owl monkey sequence with that of other new-world monkeys, it should be possible to determine whether they share a common middle region. In the absence of such information, repeats 11–35 of the owl monkey L allele and repeats 11–30 of the S allele cannot be separated into middle and late regions.

It should be noted that the two owl monkey alleles are very similar up to repeat 16 (fig. 3) but that 5' of this point the coding region has been extended by duplications of different blocks of repeats in the two alleles (fig. 3); for example, block III (consisting of repeats 12–16) was duplicated to give block IV in the L allele but not in the S allele. (As a possible alternative, repeat 16 of the S allele could be matched with repeat 21 of the L allele, suggesting a deletion of five repeats spanning blocks III and IV.) Block V was duplicated to produce block VI in the S allele but not in the L allele. Block VII was duplicated to produce block VIII in the L allele but not in the S allele. Duplicate blocks share the same repeat pattern and have corresponding marker nucleotides.

Two additional possible duplications are not numbered in figure 3, because they are less certain. Repeats 4 and 5 of the early region have apparently been duplicated as a block in both alleles, to give repeats 17 and 18 in the S allele and repeats 22 and 23 in the L allele. Repeat 29 in the L allele has most likely been duplicated to give repeat 30, but the corresponding repeat 27 of the S allele has not been duplicated.

Some indication of the relative timing of the duplications of these blocks of repeats can be obtained from the amount of nucleotide divergence between the members of each pair (fig. 3). In both alleles, blocks I and II have a mismatch frequency of 12.7%, whereas the other duplicated blocks (III–VIII) have a mismatch frequency of <2%. This is consistent with the general pattern of expansion of the modern segment; that is, the oldest part is the 3' end, and the newest part is at or close to the 5' end.

Discussion

All involucrins so far known possess a segment of short repeats of unknown significance; perhaps they confer properties appropriate to a substrate of transglutaminase (Moore et al. 1987) or facilitate the interaction of a protein with cell membrane receptors (Kochan et al. 1986; Klickstein et al. 1988). It seems likely that regular
tandem short repeats within apolipoprotein A-I are important for its lipid-binding properties (Fitch 1977).

The modern segment of repeats present in the involucrin gene of the higher primates is not orthologous to the segment of repeats present in the gene of the pig, lemur, or galago. The modern segment of repeats was built by duplications of two basic units (A and B repeats) in various combinations. Both types of repeat were present at an early stage in the generation of the modern segment.

It is now clear that this evolution of the modern segment began in a common ancestor of all higher primates, since not only the three hominoid species examined but also the owl monkey all share a closely related early region. In all four species, the early region consists of 10 repeats containing the same repeat pattern, including a duplicated block of four repeats, and corresponding marker nucleotides. The early region of the owl monkey is distinctive only in having a deletion of a single codon in repeat 5. It may be postulated that all anthropoids will contain the early region.

As the middle region common to the human, gorilla, and orangutan does not correspond to any part of the modern segment in the owl monkey, this middle region probably developed in the common hominoid lineage after the latter diverged from the new-world monkeys. Whether a common middle region also formed in the lineage leading to the new-world monkeys remains to be determined. However, it is clear that, after completion of the early region, its repeat structure became stable and that the site of repeat addition in the owl monkey and in the hominoids shifted in the 5′ direction. This implies that the duplication process was not spatially random.

Sequence Availability

These sequences have been deposited in GenBank under accession numbers M25313 and M25314.

Acknowledgment

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

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received January 24, 1989; final revision received May 16, 1989

Accepted May 16, 1989