to a similar extent as a CT base combination (submitted for publication).

References

Caceci, M. S., & Cacheris, W. P. (1984) *Byte 9* (5), 340–362.

Cantor, C. R., & Schimmel, P. R. (1980) in *Biophysical Chemistry*, W. Freeman and Co., San Francisco.

Chaconas, G., & van de Sande, J. H. (1980) *Methods Enzymol. 65*, 75–85.

Germann, M. W. (1989) Ph.D. Thesis, The University of Calgary.

Germann, M. W., Kalisch, B. W., & van de Sande, J. H. (1988) *Biochemistry 27*, 8302–8306.

Germann, M. W., Vogel, H. J., Pon, R. T., & van de Sande, J. H. (1989) *Biochemistry 28*, 6220–6228.

Haasnoot, C. A. G., Hilbers, C. W., van der Marel, G. A., van Boom, J. H., Singh, U. C., Pattabiraman, N., & Kollman, P. A. (1986) *J. Biomol. Struct. Dyn. 3*, 843–857.

Harvey, C. L., Gabriel, T. F., Wilt, E. M., & Richardson, C. C. (1971) *J. Biol. Chem. 246*, 4523–4530.

Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) in *Molecular Cloning, A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Orbons, L. P. M. (1987) Ph.D. Thesis, University of Leiden.

Pattabiraman, N. (1986) *Biopolymers 25*, 1603–1606.

Ramsing, N. B., & Jovin, T. M. (1988) *Nucleic Acids Res. 14*, 6659–6676.

Ramsing, N. B., Rippe, K., & Jovin, T. M. (1989) *Biochemistry 28*, 9528–9535.

Rippe, K., & Jovin, T. M. (1989) *Biochemistry 28*, 9542–9549.

Rippe, K., Ramsing, N. B., & Jovin, T. M. (1989) *Biochemistry 28*, 9536–9541.

Shchyolkina, A. K., Lysov, Y. P., Il'ichova, I. A., Chernyi, A. A., Golova, Y. B., Ckernov, B. K., Gottikh, B. P., & Florentiev, V. L. (1989) *FEBS Lett. 244*, 39–42.

Summers, M. F., Byrd, R. A., Gallo, K. A., Samsom, C. J., Zon, G., & Egan, W. (1985) *Nucleic Acids Res. 13*, 6375–6386.

Tchurikov, N. A., Chernov, B. K., Golova, Y. B., & Nechipurenko, Y. D. (1989) *FEBS Lett. 257*, 415–418.

van de Sande, J. H., Ramsing, N. B., Germann, M. W., Elhorst, W., Kalisch, B. W., Kitzing, E. V., Pon, R. T., Clegg, R. C., & Jovin, T. M. (1988) *Science 241*, 551–557.

Organization, Structure, and Polymorphisms of the Human Profilaggrin Gene‡

Song-Qing Gan,§ O. Wesley McBride,∥ William W. Idler,§ Nedialka Markova,§ and Peter M. Steinert*,§

Dermatology Branch and Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892

Received April 6, 1990; Revised Manuscript Received July 12, 1990

ABSTRACT: Profilaggrin is a major protein component of the keratohyalin granules of mammalian epidermis. It is initially expressed as a large polyprotein precursor and is subsequently proteolytically processed into individual functional filaggrin molecules. We have isolated genomic DNA and cDNA clones encoding the 5′- and 3′-ends of the human gene and mRNA. The data reveal the presence of likely "CAT" and "TATA" sequences, an intron in the 5′-untranslated region, and several potential regulatory sequences. While all repeats are of the same length (972 bp, 324 amino acids), sequences display considerable variation (10–15%) between repeats on the same clone and between different clones. Most variations are attributable to single-base changes, but many also involve changes in charge. Thus, human filaggrin consists of a heterogeneous population of molecules of different sizes, charges, and sequences. However, amino acid sequences encoding the amino and carboxyl termini are more conserved, as are the 5′ and 3′ DNA sequences flanking the coding portions of the gene. The presence of unique restriction enzyme sites in these conserved flanking sequences has enabled calculations on the size of the full-length gene and the numbers of repeats in it: depending on the source of genomic DNA, the gene contains 10, 11, or 12 filaggrin repeats that segregate in kindred families by normal Mendelian genetic mechanisms. This means that the human profilaggrin gene system is also polymorphic with respect to size due to simple allelic differences between different individuals. The amino- and carboxyl-terminal sequences of profilaggrin contain partial or truncated repeats with unusual un-filaggrin-like sequences on the termini. Such sequences are reminiscent of propeptides encountered in other structural protein systems. We suggest these sequences are required for the assembly of the accumulating protein into large keratohyalin granules among the keratin filaments in the granular cells and aid in later processing events.

Filaggrins represent an important class of intermediate filament-associated proteins (IFAPs) that function, at least in part, in the aggregation of keratin intermediate filaments into an organized "keratin pattern" during terminal stages of normal differentiation in mammalian epidermis (Dale et al., 1978, 1989; Steinert et al., 1981; Steinert & Roop, 1988). On the basis of data from both protein chemical studies (Harding & Scott, 1983; Resing et al., 1984, 1985) and more recent cloning experiments (Haydock & Dale, 1986; Rothnagel et al., 1987; Rothnagel & Steinert, 1990; McKinley-Grant et al., 1989), filaggrins are initially synthesized as large polyprotein precursors ("profilaggrins") consisting of many protein repeats

‡ The nucleic acid sequence in this paper has been submitted to GenBank under Accession Number J02929.

* To whom correspondence should be addressed at the Laboratory of Skin Biology, National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, Building 10, Room 12N238, Bethesda, MD 20892.

§ Dermatology Branch.

∥ Laboratory of Biochemistry.

arranged in tandem and accumulate in a nonfunctional phosphorylated form as F-keratohyalin granules late in epidermal differentiation (Fisher et al., 1987; Rothnagel et al., 1987; McKinley-Grant et al., 1989; Resing et al., 1989; Steven et al., 1989; Rothnagel & Steinert, 1990). Subsequently, this precursor is dephosphorylated and proteolytically cleaved by excision of a short peptide "linker" sequence to release functional filaggrin molecules (Resing et al., 1984, 1985, 1989; Haydock & Dale, 1986, 1990; Rothnagel et al., 1987; Rothnagel & Steinert, 1990; McKinley-Grant et al., 1989).

In order to understand the expression and function of this protein in detail and to explore its putative involvement in keratinizing disorders of the epidermis, we have recently isolated a cDNA clone encoding one full repeat of the human profilaggrin gene (McKinley-Grant et al., 1989). The full-length repeat was shown to be 324 amino acids (972 bp), which includes a linker of perhaps only 7 amino acids of sequence FLYQVST; that is, human profilaggrin consists of a tandem array of filaggrin molecules of about 317 amino acids separated by the linker sequence. The properties of such a deduced sequence are indistinguishable from those of isolated human filaggrin. By in situ hybridization, expression of the gene is tightly regulated at the transcriptional level in the granular layer. Although we showed the human gene is localized to chromosome position 1q21, no further information on gene structure and organization is available.

In this paper we have isolated and characterized both cDNA and genomic clones encoding the ends of the gene. This has enabled elucidation of the structure of the gene, the likely number of repeats, and the extent of the polymorphisms in it.

Materials and Methods

Molecular Biology Procedures. A human genomic library in EMBL-3, constructed from DNA isolated from a single placenta, kindly supplied by Dr. Frank Gonzales (National Cancer Institute, Bethesda, MD), was screened with the cDNA clone λHF10, which contains human filaggrin coding sequences (McKinley-Grant et al., 1989). Three clones were plaque purified, and their inserts were excised with *Sal*I. Filaggrin-positive fragments were subcloned into pGEM-3Z for preparation of DNA. Portions were further subcloned into M13 mp18 or mp19 vectors for sequencing with either Sequenase 2 (U.S. Biochemical Corp.) or TacTrac (Promega BioTec) according to the manufacturer's specifications and with synthetic oligonucleotides as primers. Portions of these clones or synthetic oligonucleotides derived from them, corresponding to 5′- or 3′-noncoding sequences, were used to reprobe the original λgt11 library (McKinley-Grant et al., 1989) to find cDNA clones also bearing 5′- or 3′-sequences. Similarly, λHF10 was used to screen this library to isolate longer cDNA clones containing multiple filaggrin repeats. The first-round signals of greatest intensity were sized by Southern blotting (Rothnagel et al., 1987; McKinley-Grant et al., 1989) and the longest were plaque purified. Table I summarizes the genomic DNA and cDNA clones used to generate sequence information.

DNA was obtained from a single placenta (Oncor Labs) or purified from 12 whole human foreskins (Maniatis et al., 1982).

Computer Analyses of Sequences. Protein sequence homologies, secondary structure prediction analyses, and nucleic acid sequence analyses were performed on the University of Wisconsin sequence analysis software packages compiled by the Wisconsin Genetics Computer Group (Devereux et al., 1984) and by use of the IBI Pustell sequence analysis software (version 2, International Biotechnologies Inc.).

Table I: Summary of Genomic DNA and cDNA Clones Used in This Work

clone name	sequence location	comments
gene clones		
gλHF5	5′-end of gene	see Figure 1
gλHF222	3′-end of gene	see Figure 2
cDNA clones		
λHF202	5′-end; bp 366–379 and 949–2076	see Figure 1
λHF604	5′-end; bp 1376–2971	see Figure 1
λHF373	5′-end; bp 1447–1989	see Figure 1
λHF223	3′-end; bp 2801–5732	see Figure 2
λHF10	unknown coding region; 1248 bp	McKinely-Grant et al. (1989)
λHF41	unknown coding region; 2832 bp	data not shown
λHF114	unknown coding region; 2325 bp	data not shown
λHF294	unknown coding region; 1800 bp	data not shown
λHF336	unknown coding region; 745 bp	data not shown

Results

Isolation of Clones for the 5′- and 3′-Ends of the Profilaggrin Gene. cDNA clone λHF10 established in a previous paper (McKinley-Grant et al., 1989) to encode a portion of the human profilaggrin mRNA was used as a probe to screen a human genomic library in EMBL-3. Three positive clones were identified and plaque purified to homogeneity. Their inserts were excised with *Sal*I (which does not cut within coding regions of the human gene), and the fragments which were filaggrin positive were as follows: gλHF5, 18 kbp; gλHF18, 4.5 kbp; gλHF222, 7.7 kbp. Each of these inserts was successfully subcloned into pGEM-3Z for further mapping analyses. By use of the restriction enzymes *Hgi*AI and *Xma*I (which cut each filaggrin repeat once to yield a repeat fragment of 0.972 kbp), clones gλHF18 and gλHF222 contained four full filaggrin repeats and clone gλHF5 contained two full filaggrin repeats, as well as bands of about 0.5, 2.5, and 16 kbp, respectively, that did not hybridize to the filaggrin probe. These sequences presumably represent flanking regions of the gene. The 0.5-kbp piece used as a probe cross-hybridized with the 2.5- but not the 16-kbp pieces, indicating that gλHF5 represented a different end of the gene from the others. We were unable to find any clones containing larger numbers of filaggrin repeats, presumably because *Bam*HI, used in constructing the genomic library, cuts each filaggrin repeat many times (McKinley-Grant et al., 1989). Subsequently, for DNA sequencing, the gλHF222 7.7-kbp piece was cut in half with *Sac*I. The gλHF5 clone was also cut with *Pvu*II to generate a 4.5-kbp piece carrying all of the filaggrin-positive sequences. In addition, the 0.972-kbp pieces obtained by *Xma*I digestion of both gλHF5 and gλHF222 were harvested. All of these fragments were subcloned into M13 vectors for sequencing. It became clear that gλHF5 encoded the 5′-end (Figure 1) and gλHF222 the 3′-end of the profilaggrin gene (Figure 2).

Isolation of cDNA Clones for the 5′- and 3′-Ends of the Profilaggrin mRNA. Synthetic oligomers 60 bp long corresponding to nucleotides 1605–1664 (see Figure 1) at the 5′-end of the gene and nucleotides 5461–5520 (see Figure 2) at the 3′-end of the gene were used as probes to rescreen a cDNA library in λgt11 prepared earlier (McKinley-Grant et al., 1989). Of about 1×10^6 pfu screened, only three clones positive for the 5′-end and one clone for the 3′-end were found. These numbers are far less than the total numbers of filaggrin clones in the library (about 2% of all plaques), suggesting that the ends of the mRNA have been substantially processed, as seems likely from Northern blots (McKinley-Grant et al., 1989). Clones λHF202 (1.145 kbp) and λHF373 (0.543 kbp) for the 5′-end and clone λHF223 (2.952 kbp) for the 3′-end were completely sequenced and are illustrated in Figures 1 and 2, respectively.

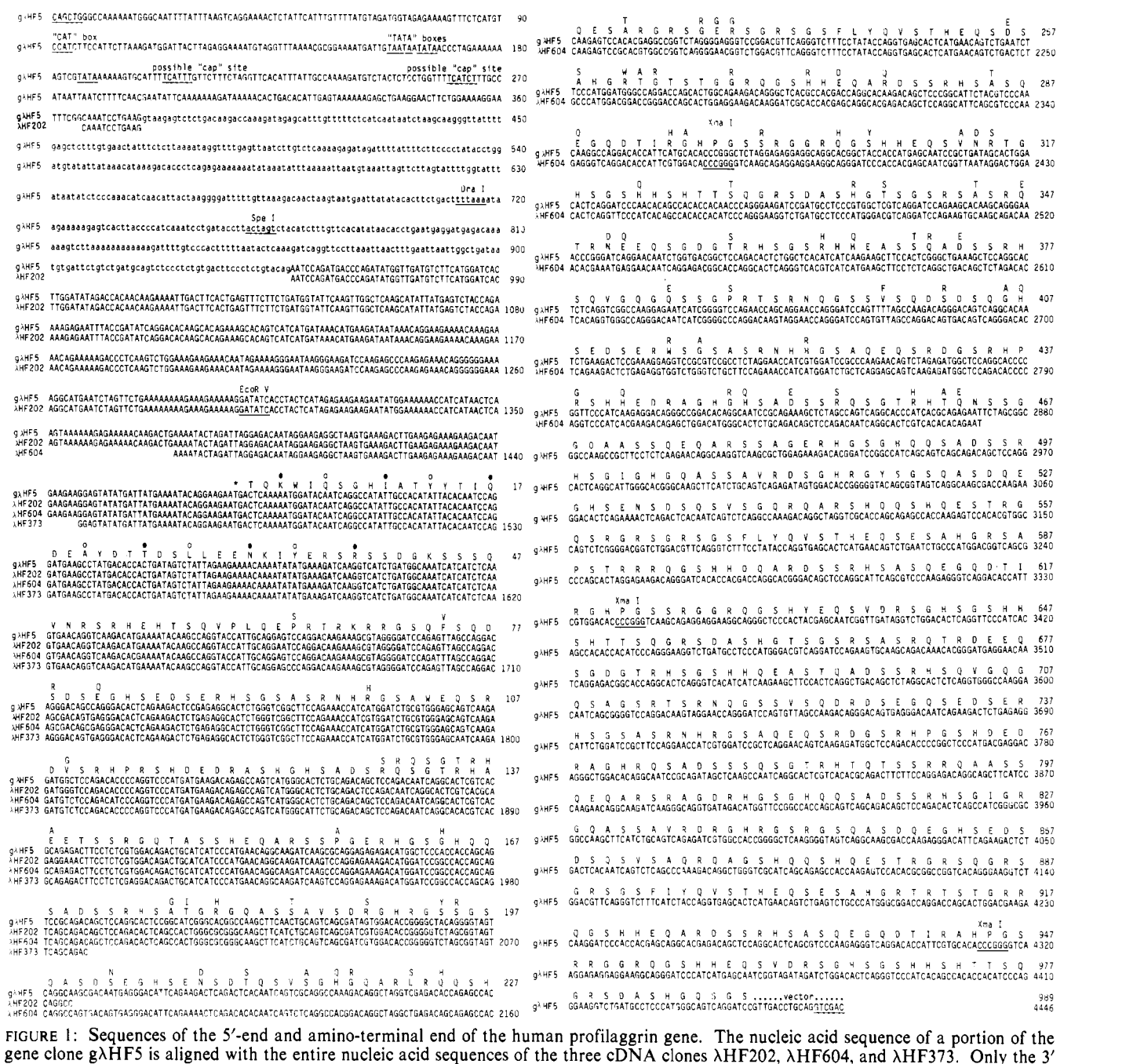

FIGURE 1: Sequences of the 5′-end and amino-terminal end of the human profilaggrin gene. The nucleic acid sequence of a portion of the gene clone gλHF5 is aligned with the entire nucleic acid sequences of the three cDNA clones λHF202, λHF604, and λHF373. Only the 3′ 4.456 kbp of gλHF5 from a *Pvu*II site (underlined at the beginning) to the cloning vector site of *Sal*I (underlined at the end) are shown. The two different *Xma*I fragments were ordered from the sequence of the intact gλHF5 sequence. The intron sequences in gλHF5 are presented in lower case. Putative regulatory sequences of the CAT and TATA boxes and the cap site are indicated. The three restriction enzyme sites *Dra*I, *Eco*RV, and *Spe*I used to size the entire gene and the *Xma*I sites are shown. The deduced amino acid sequences with variations are indicated with the single-letter code and are numbered from the initiation codon (*). The symbols (○) and (●) of the amino-terminal 40 residues mark likely a and d positions, respectively, that may form a coiled-coil α-helix.

Characterization of the 5′-End of the Profilaggrin Gene. Figure 1 shows the sequences of the genomic clone gλHF5 and cDNA clones λHF202, λHF373, and λHF604 that encode 5′ information of the gene. Several features are evident. First, all clones contain a unique in-frame ATG (at bp position 1477) that meets all of the criteria for a utilized initiating codon (Kozak, 1989). Their nucleotide sequences are identical prior to it and for the first 128 bp following it; in the next 1266 bp of overlapping sequences, there are 180 (14%) variations in nucleotide and 86 (20%) variations in amino acid sequence. The two different *Xma*I fragments identified by subcloning and sequencing in gλHF5 were ordered as shown in Figure 1. Second, comparisons of gλHF5 and λHF202 reveal the presence of an intron of 570 bp in the gene that splices the 5′-untranslated region (at bp 379), which meet the obligatory recognition sequence requirements for introns (Green, 1986). Primer extension experiments to define the likely "cap" site, using the 60-bp oligonucleotide from bp 1605–1664 described above and up to 100 μg of poly(A)-enriched epidermal RNA, were inconclusive due to the likely processed nature of the human filaggrin mRNA (McKinley-Grant et al., 1989). However, there are two possible cap sites at bp 203 and 261. These are preceded by potential "TATA" boxes at bp 163–185 and "CAT" boxes at bp 21 or 90–100 that fulfill the characteristics of functional genes. Finally, the available sequence data reveal several potential regulatory sequences such as the so-called epidermal-specific enhancer element and a retinoic acid responsive element (Blessing et al., 1987; Tseng & Green,

FIGURE 2: Sequences of the 3′-end and carboxyl-terminal end of the human profilaggrin gene. The nucleic acid sequence of a portion of gene clone gλHF222 is aligned with the entire nucleic acid sequence of cDNA clone λHF223. The unsequenced gap of 1420 bp in gλHF223 corresponds to the position of the two *Xma*I fragments that could not be ordered. The *Sal*I sites at the beginning (cloning vector site) and end (flanking gene sequences), likely polyadenylation signal sequences, and the middle *Sac*I site used in subcloning and sequencing are underlined. The three restriction enzyme sites *Dra*I, *Eco*RV, and *Spe*I used in sizing the gene and the *Xma*I sites are shown. The deduced amino acid sequences with variations are indicated with the single-letter code and are numbered from the 5′-end of gλHF222, through the unsequenced gap to the overlap region with λHF223 to the termination codon (*).

1988), but additional sequencing and other functional assays will be necessary to identify all such sequences and those likely to exist further upstream.

The deduced amino acid sequences from the initiating codon reveal a conserved aliphatic–polar sequence for the first 40 amino acids of no sequence homology to the filaggrin repeating sequences. Residues 41–71 reveal about 50% sequence homology, while residues beyond 71 are highly homologous (>85%). The first FLYQVST sequence, which represents the linker region that is cleaved to release individual functional filaggrin molecules (McKinley-Grant et al., 1989), occurs at residue 245; that is, the first portion of the gene encodes a truncated filaggrin repeat with an unusual amino-terminal end.

Analysis of the likely secondary structure reveals that the first 40 residues have an α-helical conformation. In searching both GenBank and NBRF sequence data banks, we found the sequences I-A-T-Y (residues 10–13) and L-L-E (residues 27–29) occur elsewhere only in the coiled-coil sequences of several IF proteins, including human keratin 1 (Johnson et al., 1985; Steinert et al., 1985), which is coexpressed in this tissue with profilaggrin. Apart from this, the amino-terminal 40 residues share little significant homology with any IF or other coiled-coil α-helical protein. However, this sequence possesses a weak heptad pattern of the form $(a\text{-}b\text{-}c\text{-}d\text{-}e\text{-}f\text{-}g)_n$, suggesting that it may form a coiled coil. In most established coiled-coil proteins, at least 70% of the a and d positions are occupied by residues with hydrophobic side chains (Conway & Parry, 1988). In this case, 55% (6 of 11) of the a and d residues meet this requirement.

Residues 41–71, which have been less conserved, are likely

```
Human 765  C G H S S D I S - K Q L G F S Q S Q R Y Y Y Y E
             :   :  . .  :  :  :  ::    : : : :
Mouse 315  - G Y E S I F T A K H L D F N Q S H S Y Y Y Y -
```

FIGURE 3: Homology of the carboxyl-terminal sequences of human and mouse (Rothnagel et al., 1987; Rothnagel & Steinert, 1990) profilaggrins: (:) identity; (·) homologous residues; (−) deletion.

to possess a folded structure due to the presence of several turns.

Characterization of the 3′-End of the Profilaggrin Gene. Figure 2 shows the sequences of the genomic clone gλHF222 and the cDNA clone λHF223 that encode 3′ information of the gene. Both clones possess the termination codon (at bp 5193) and the entire A-T-rich 3′-noncoding region. Like the mouse profilaggrin gene (Rothnagel et al., 1987; Rothnagel & Steinert, 1990), there are no introns in the coding end. The nucleic acid sequences are identical beyond the last FLYQVST sequence (at bp 4138), suggesting that this part of the gene has been conserved. Prior to this, in the last complete filaggrin repeat (bp 3165–4137), there are 62 (8%) variations in nucleotide and 30 variations (11%) in amino acid sequence in the overlap region. The *Xma*I fragments of gλHF222 were subcloned into M13 and sequenced, and four different repeat sequences were found. Two repeats, corresponding to the first and fourth of the intact clone gλHF222, were recognized and are shown in Figure 2.

The deduced amino acid sequence of Figure 2 shows that in the last repeat the sequence deviates completely from filaggrin-like sequences after amino acid residue 1594. While the carboxyl-terminal 137 residues have no homology with typical filaggrin repeat sequences, the last 23 residues share 59% homology with the carboxyl-terminal end of mouse filaggrin, including a striking -Y-Y-Y-Y terminal sequence (Figure 3; Rothnagel et al., 1987; Rothnagel & Steinert, 1990). Thus, like mouse profilaggrin, the human gene posssses a truncated and modified repeat at its carboxyl-terminal end. The carboxyl-terminal 137 residues are highly charged (24 basic, 13 acidic) and hydrophobic (23%). Secondary structural predictions suggest little or no organized structure, having frequent turns.

Sequence Polymorphisms of the Human Profilaggrin Gene System. The data of Figures 1 and 2 have revealed considerable sequence variation between adjacent repeats on the same genomic clones. This was particularly evident in a sequencing reaction using gλHF222 and a synthetic oligonucleotide primer corresponding to the linker region (bp 4138–4155 of Figure 2) which hybridizes to gλHF222 in five locations. A sequencing gel covering approximately 90–220 bp from the linker region (Figure 4) reveals that 12% of the base positions are heterogeneous. In order to understand these sequence polymorphisms in more detail, additional cDNA clones were obtained from the λgt11 library. The longest clones were isolated and were λHF114 (2.325 kbp) and λHF41 (2.832 kbp) and a third that was λHF223 (see Figure 2). (Several other long clones were concateners of *Eco*RI fragments which had randomly ligated together during preparation of the library, but provided more filaggrin sequence data.) Together with these new clones and the cDNA and genomic DNA clones described above and previously (McKinley-Grant et al., 1989), we are able to compile a data base of sequence information on human filaggrin, including sequences from multiple individual persons and some with multiple adjacent repeats. Figure 5 shows a "consensus" sequence map for human filaggrin sequences with variations. Of 26 partial or complete filaggrin repeats, we found that all repeats are precisely 972 bp (324 amino acids) long, except those repeats located at the 5′- and 3′-ends of the

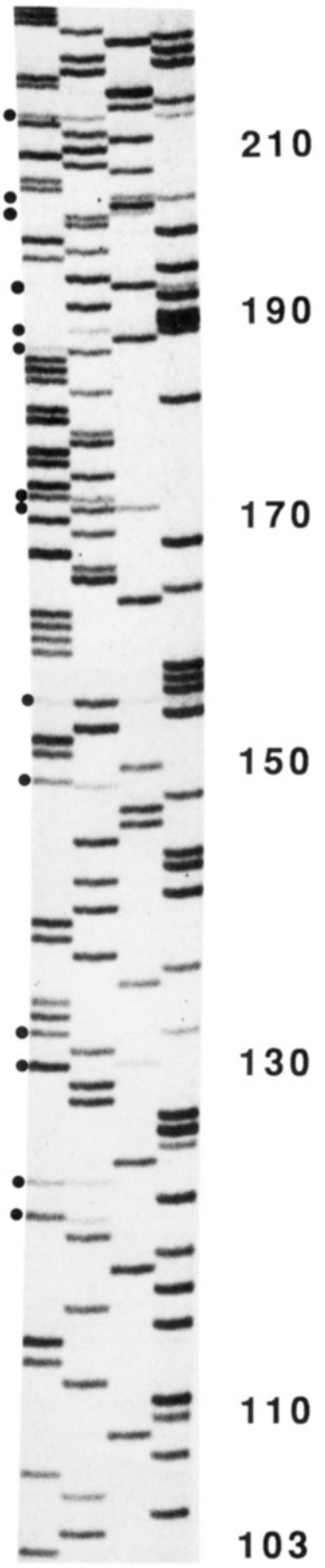

FIGURE 4: A synthetic oligonucleotide corresponding to a linker region (bp 4138–4155 of Figure 2) which hybridizes to gλHF222 in five locations was used as a primer. The sequencing reaction shows multiple bands in several base positions (spots), indicating that the sequences following the linker are variable. The numbers refer to the base pairs from the linker. The lanes are, from left to right, G, A, T, and C.

mRNA (Figures 1 and 2). In the full-length and partial repeats on gλHF5 and gλHF222 clones from the same individual, 100 of 324 (31%) of the residue positions are variable, of which 14 (4%) vary more than twice (Figure 5, capitals). Of these variations, all but four can be accounted for by sin-

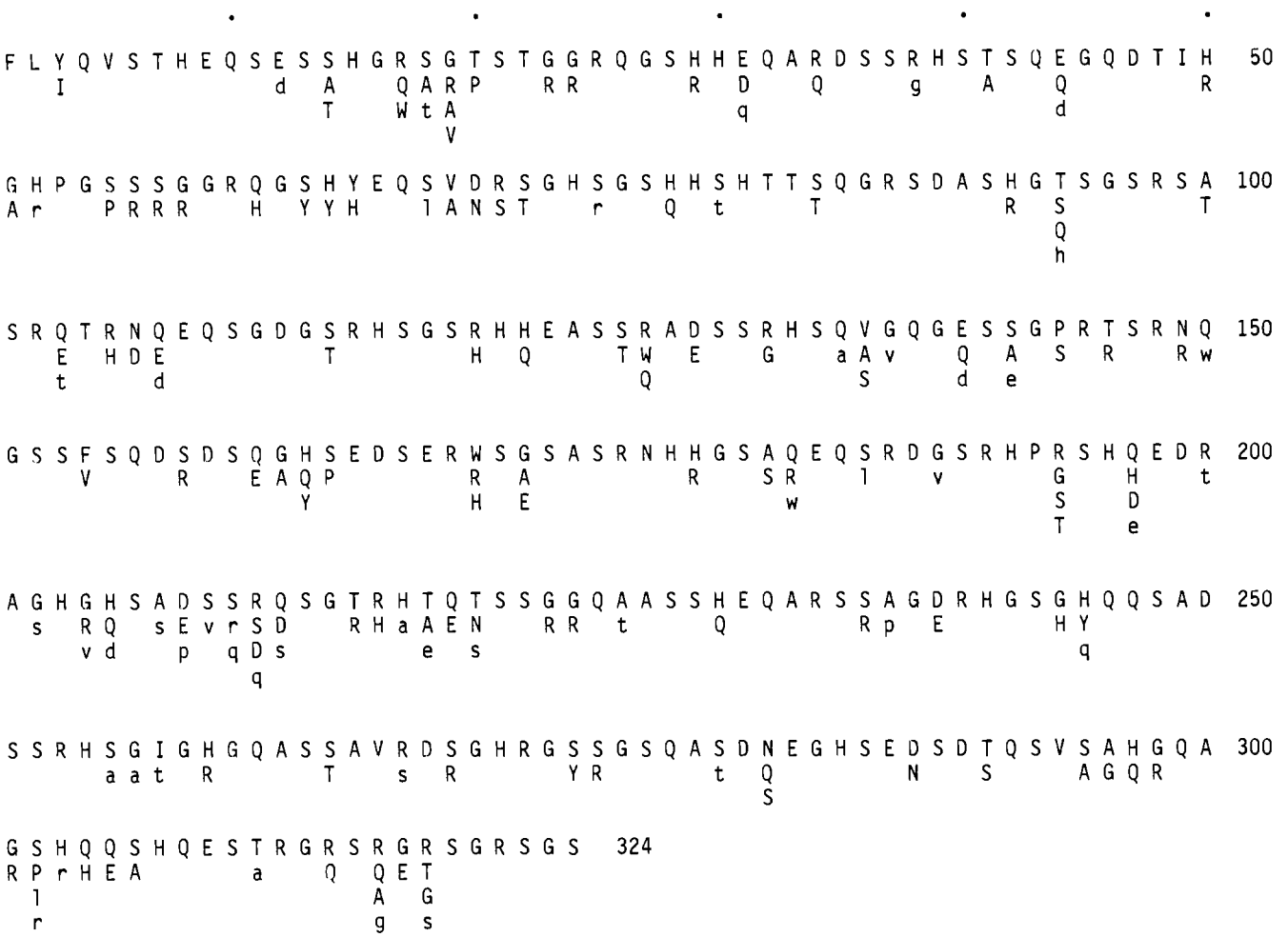

FIGURE 5: A consensus amino acid sequence map of a human filaggrin repeat. A comparison of the full-length and partial repeats of clones gλHF5 and gλHF222 derived from a single individual (placenta) is shown in capital letters. Additional variations encountered in a total of 20 other partial or complete filaggrin repeats of nine other cDNA clones (and thus probably from different individuals) are shown in lower-case letters. Whereas most of the former represent conservative amino acid substitutions arising from single-base changes in the codons utilized, many of the latter involve more complicated mutations and involve nonconservative amino acid substitutions.

gle-base changes in the codons utilized. When all available sequence data are considered (Figure 5), 126 of 324 (39%) of the residue positions are variable, with several (a total of 32, 9%) more than twice. Most of the additional 26 variations would have required multiple mutations in the codons utilized. The longest conserved region occurs in the vicinity of the linker (residues 319–13). This corresponds to the region in the DNA sequence where several restriction enzymes such as *Hgi*AI cut each repeat, generating the superstoichiometric repeat on Southern blots (McKinley-Grant et al., 1989).

A comparison of the full-length and partial repeats from the two genomic clones (Figure 5, capitals) reveals that 60% of amino acid sequence variations are conservative; only 7% involve exchanges between hydrophilic and hydrophobic residues, but 33% involve changes in charge. While the molecular masses of these repeats vary little (34 ± 0.2 kDa), their p*I* vary more widely (8.3 ± 1.1). Human filaggrin has been shown to consist of multiple isoelectric variants, attributed to incomplete dephosphorylation or desimidation (Harding & Scott, 1983), but these data clearly show that another major reason for charge heterogeneity is sequence polymorphism.

Size of the Full-Length Human Profilaggrin Gene. The data of Figures 1 and 2 reveal the presence of *Dra*I, *Eco*RV, and *Spe*I restriction enzyme sites that occur only in the conserved 5′- and 3′-ends of the gene which permit calculations of the size of the full-length gene. These calculations assume that the distances between restriction enzyme sites and the first FLYQVST linker sequence (bp 2212 of Figure 1) and between the last FLYQVST linker sequence (bp 4138 of Figure 2) and the restriction enzyme sites have been conserved, as our sequence data indicate. Surprisingly, the sample of genomic DNA used, from a different source than used previously (McKinley-Grant et al., 1989), yielded two bands of equal intensity with each enzyme of sizes 13.2 and 12.2, 13.0 and 12.1, and 15.3 and 14.2 kbp, respectively (Figure 6A). This means that this DNA sample contains profilaggrin genes having 10 and 11 full filaggrin repeats, in addition to the partial and modified repeats at the 5′- and 3′-ends. The genomic DNA utilized previously (McKinley-Grant et al., 1989) yielded a single band that corresponds to a gene with 12 full filaggrin repeats. These observations were explored further with DNA from another 12 individuals which was cut with *Dra*I (Figure 6b). All samples contain either one band or two bands of equal intensity of three size classes about 1 kbp apart and correspond to genes containing 11 only, 12 only, 10 and 11, 11 and 12, or 10 and 12 repeats.

These apparent allelic forms of the human profilaggrin gene were further examined with DNA derived from many individuals in several three-generation kindreds (CEPH cell lines; White et al., 1990) with no known involved keratinizing disorders of the skin and of several different racial and ethnic groups. When cut with *Eco*RV, DNA from two kindred families (Figure 7), as well as 24 other families (data not shown), revealed only one or two bands in all cases, corre-

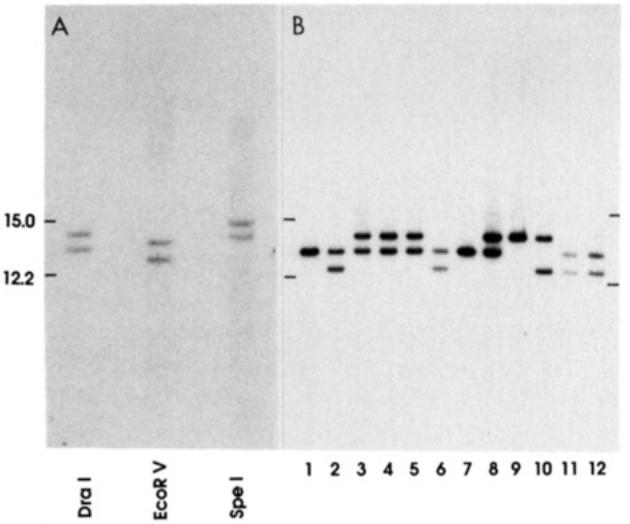

FIGURE 6: Size of the human profilaggrin gene. (A) Genomic DNA from a single source was cut with the three enzymes *Dra*I, *Spe*I, and *Eco*RV, electrophoresed for 3 days to maximize resolution, processed by Southern blotting, and probed with a coding probe [λHF10; McKinley-Grant et al. (1989)]. The sizes of the two bands in each case were measured with respect to high molecular weight markers (Bethesda Research Labs). With the *Dra*I data for example, the number of repeats was calculated as follows: the distance from the proximal *Dra*I site at the 5′-end to the first FLYQVST linker is 1.590 kbp (Figure 1); the distance from the last FLYQVST linker to the proximal *Dra*I site at the 3′-end is 1.169 kbp (Figure 2); the sizes of the *Dra*I fragments shown here are 13.4 and 12.4 kbp; the size of each filaggrin repeating unit is 0.972 kbp; the number of repeats is therefore (13.4 − 1.59 − 1.169)/0.972 = 10.9 and (12.4 − 1.59 − 1.169)/0.972 = 9.9. The numbers for *Eco*RV and *Spe*I calculated the same way are 11.0 and 10.1 and 11.1 and 9.9, respectively. (B) DNA from 12 individuals was digested with *Dra*I and processed as above. The three levels of bands correspond to 10, 11, or 12 repeats.

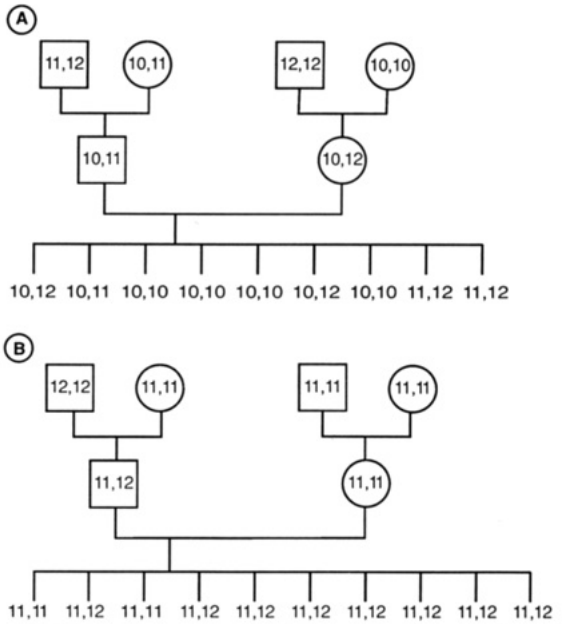

FIGURE 7: Mendelian segregation of human profilaggrin size alleles. DNA from transformed lymphocytes of the several members of two three-generation kindred families (CEPH cell lines; White et al., 1990) was cut with *Eco*RV and characterized on Southern blots as in Figure 6. In all cases, one or two bands corresponding to 10, 11, or 12 repeats were obtained, which were segregated between the various family members as shown.

sponding in size to 10, 11, or 12 filaggrin repeats. The distributions of the repeat numbers in the various family members (Figure 7) indicate normal Mendelian inheritance. Thus, based on the analyses of the DNA of more than 300 different

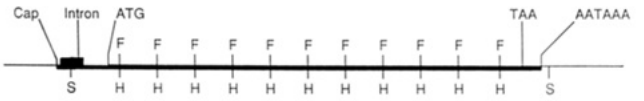

FIGURE 8: Structure of the human profilaggrin gene containing 10 filaggrin repeats. S = *Spe*I, which cuts only in the conserved flanking regions; H = *Hgi*AI, which cuts in the conserved linker region; F = phenylalanine of the first consensus residue of each repeat. The positions of the cap site, intron, initiating codon, termination codon, and polyadenylation signal sequences are shown.

individuals (26 of 40 CEPH families and 14 other individuals; a total of 44 shown in this paper), it is clear that the range in size of the normal profilaggrin gene within the human population is limited.

Discussion

Data reported in this paper on the isolation of portions of the human profilaggrin gene, as well as data from both protein chemical (Harding & Scott, 1983; Resing et al., 1984, 1985, 1989; McKinley-Grant et al., 1989) and other recent cloning experiments (Haydock & Dale, 1986; Rothnagel et al., 1987; McKinley-Grant et al., 1989; Rothnagel & Steinert, 1990), have now firmly established that filaggrins are expressed from huge genes of relatively simple structure. The genes consist of several tandemly arranged polynucleotide repeats of 972 bp [human; this paper and McKinley-Grant et al. (1989)], 750 bp (mouse; Rothnagel & Steinert, 1990), and 1272 bp (rat; Haydock & Dale, 1986) and are devoid of introns in coding regions. Thus, the genes encode large polyprotein precursors consisting of numerous tandem filaggrin repeats.

This paper provides details of the 5′- and 3′-ends and thus of the structure of the entire human profilaggrin gene (Figure 8). Even though the complete gene has not been isolated, by taking advantage of certain restriction enzyme sites that occur only in the conserved flanking regions, we are able to calculate the number of repeats in it. Whereas a sample of DNA obtained previously from one individual yielded only one band when cut with the enzymes *Dra*I and *Eco*RV (McKinley-Grant et al., 1989), we now find in another single DNA sample that these enzymes and *Spe*I generate two bands (Figure 6A). Further, analysis of DNA from an additional 12 foreskins and DNA from many members of 26 CEPH kindred families of several different racial and ethnic origin reveals one or two bands with *Dra*I (Figure 6) or *Eco*RV (Figure 7). The precise 1-kbp difference in size of the bands in different individuals (Figures 6 and 7), the conservation of the flanking sequences of the gene, and the multiplicity of sites for these three restriction enzymes (Figures 1 and 2) make it improbable that these 1-kbp variations in size can be due to mutations in all three restriction enzyme sites simultaneously. The most likely explanation of these results is that the profilaggrin genes in different individuals can contain 10, 11, or 12 full filaggrin repeats; that is, the human profilaggrin gene system is polymorphic with respect to the numbers of repeats. These data may mean that there are multiple genes containing varying numbers of repeats within any one individual, although if this were the case, such multiple genes must be tightly linked to the 1q21 region (McKinley-Grant et al., 1989). It is more likely, however, that there is only one gene per haploid genome, but the two copies of the gene in any one individual can contain variable numbers of repeats due to simple allelic differences. This notion is further supported by the finding (Figure 7) that the different-sized bands corresponding to different numbers of repeats segregate in kindred families by normal Mendelian processes. These three allelic variants may have arisen by unequal meiotic recombinations earlier in evolution and have

since been conserved. Even though our data represent more than 300 individuals of several racial and ethnic groups, we cannot exclude the possibility of additional allelic size variants in a wider population survey.

One important conclusion of this finding is that it would appear that the formation of a normal terminally differentiated human epidermis is not critically dependent on the precise amount of functional filaggrin produced from the precursor gene; that is, to date we see a variation of 20% (10–12 repeats). The rationale for such variability is not yet clear.

In our initial report on the human filaggrin system (McKinley-Grant et al., 1989), we recognized the probability of sequence variation between neighboring repeats. In this paper we have compared the sequences of 11 different clones including clones containing multiple adjacent repeats and find (Figure 5) that all repeats are precisely 972 bp (324 amino acid residues) long but display a bewildering array of sequence variations. Such variations mean that the human filaggrin system is doubly polymorphic: in addition to variable numbers of repeats in profilaggrin, functional filaggrin also consists of a heterogeneous population of molecules of similar size but of considerable charge and sequence heterogeneity. There is as much variation between neighboring repeats on the same clone from the same individual as between repeats on different clones from different individuals (Figures 4 and 5). So far, we have found 39% of the 324 amino acid positions per repeat are variable (Figure 5). Our data base contains information from nine clones from two different foreskin cDNA libraries and two genomic clones, all obtained from individuals of similar ethnic origin. Accordingly, we expect more variations will appear when a larger portion of the human population is sampled. Nevertheless, most of the identified variations represent conservative changes. Few if any changes involve the appearance of a different type of amino acid that might be expected to significantly change the structural properties of the filaggrin molecules. Thus, although our data base is limited in size, it seems likely that generally only conservative substitutions are tolerated and that such changes are not randomly distributed in the normal human profilaggrin gene. Furthermore, our data show islands of tight sequence conservation, which explains why some restriction enzymes cleave DNA regularly. The most notable region is in the vicinity of the linker (residues 319–13, Figure 5), as might be expected since this is recognized by a common set of proteolytic processing enzyme(s). In contrast, mouse filaggrin repeat sequences seem to have been highly conserved, yet the linker region is somewhat variable (Rothnagel & Steinert, 1990). Future work will be directed toward an understanding of the structural and functional significance of these sequence variations.

We demonstrate here that the human profilaggrin gene contains an intron in the 5′-untranslated region. Interestingly, other genes expressed in mammalian epidermis such as involucrin (Eckert & Green, 1988) and loricrin (D. Hohl and P. Steinert, unpublished results) and epidermal derivatives such as trichohyalin (Rothnagel & Rogers, 1986; Fietz et al., 1990) also possess simple gene structures. None of these genes contain introns in coding portions, and all possess a single intron in their 5′-untranslated regions. Each of these genes encode proteins having peptide or polypeptide repeats that display considerable sequence variations yet retain certain prominent structural motifs. The lack of introns within or between the repeats probably reflects the simple evolutionary processes of amplification and/or duplication involved in their formation. The reason for their sequence variations is not clear at this time. However, the fact that each is expressed in a moribund tissue and ultimately functions in a dead cell to afford a barrier against the environment reminds us of the earlier view (Fraser et al., 1972) that such variations, providing they retain certain essential structural motif(s), are tolerated because they retain no further effect on the life of the organism. A further point of interest for future consideration is that most if not all of these proteins probably function in some way as intermediate filament-associated proteins, by interacting directly or indirectly with the keratin IFs of the various cell types (Steinert & Roop, 1988).

Examination of the sequences at the amino- and carboxyl-terminal ends of human profilaggrin reveals the presence of modified repeats that either start with unusual sequences before merging into or end with unusual sequences in the midst of the "consensus" filaggrin repeats. Their structural and chemical properties are strikingly different from those of the filaggrin repeat sequences: (1) the amino-terminal sequence is α-helical, is likely to form or participate in the formation of a coiled coil, is strongly acidic, and is notably enriched in aromatic amino acids; (ii) the carboxyl-terminal sequence is strongly basic and also hydrophobic; (iii) whereas the filaggrin repeat sequences contain an average of 22 potential phosphorylation sites per repeat (Resing et al., 1985, 1989; Steinert, 1988), the amino- and carboxyl-terminal ends contain 0 and 2 such sites, respectively; (iv) the sequences appear to have been highly conserved (Figures 1 and 2). Although searches in data bases with the carboxyl-terminal sequences have revealed no similarities to other proteins (except with the carboxyl-terminal end of mouse filaggrin; Figure 3), the amino-terminal sequence reveals modest homologies with certain keratin IF chains because of a potential to form a coiled-coil α-helical structure. Therefore, these sequences may serve an important role in the function of profilaggrin, distinct from its content of several filaggrin repeats. The hydrophobic nature of the carboxyl-terminal sequences may aid in the proteolytic processing, but other functions, if any, will have to await further experiments. With respect to the amino-terminal sequences, we note similar α-helical sequences are present on other structural proteins, including procollagens (Bornstein & Traub, 1980). Thus by analogy with the procollagens, we suggest the following two possibilities for the function of the amino-terminal sequences on human profilaggrin. They may aid in the accumulation of the profilaggrin in the epidermis by interaction with coiled-coil sequences on the adjacent keratin IF, so as to in effect anchor the accumulating deposit of protein. Alternatively, this could be accomplished when two (or more) adjacent profilaggrin molecules associate by interaction of their coiled-coil sequences to form a macroscopic aggregate of protein. A third or concurrent function related to their hydrophobic nature may be to aid in proteolytic processing, as proposed for the carboxyl-terminal and linker regions.

Several authors have hitherto referred to the initial translation product of this gene system as profilaggrin (Resing et al., 1984, 1985, 1989; Dale et al., 1989; Haydock & Dale, 1986). The use of this term now seems fully justified in view of the data described in this paper which clearly demonstrate the presence of propeptide sequences at the termini.

In summary, we have characterized cDNA and genomic DNA clones encoding the ends of the human profilaggrin gene, which provide novel information on the extraordinary polymorphisms of this gene system and which will now permit more detailed studies on its expression and function in normal and abnormal epidermal differentiation.

ACKNOWLEDGMENTS

We thank Drs. Sherri Bale, John DiGiovanna, Bernhard Korge, and David Parry for numerous helpful discussions during the course of this work.

REFERENCES

Blessing, M., Zentgraf, H., & Jorcano, J. L. (1987) *EMBO J. 6*, 567–575.

Bornstein, P., & Sage, H. (1980) *Annu. Rev. Biochem. 49*, 957–1003.

Conway, J. F., & Parry, D. A. D. (1988) *Int. J. Biol. Macromol. 10*, 79–98.

Dale, B. A., Holbrook, K. A., & Steinert, P. M. (1978) *Nature (London) 276*, 729–731.

Dale, B. A., Resing, K. A., Haydock, P. V., Fleckman, P., Fisher, C., & Holbrook, K. A. (1989) in *The Biology of Wool and Hair* (Rogers, G. ., Reis, P. J., Ward, K. A., & Marshall, R. C., Eds.) pp 97–115, Chapman and Hall, London.

Devereux, J., Haeberli, P., & Smithies, O. (1984) *Nucleic Acids Res. 12*, 387–395.

Eckert, R. L., & Green, H. (1988) *Cell 46*, 583–589.

Fietz, M. J., Presland, R. B., & Rogers, G. E. (1990) *J. Cell Biol. 110*, 427–436.

Fisher, C., Haydock, P. V., & Dale, B. A. (1987) *J. Invest. Dermatol. 88*, 661–664.

Fraser, R. D. B., MacRae, T. P., & Rogers, G. E. (1972) *The Keratins, Their Composition, Structure and Biosynthesis*, Charles C. Thomas, Springfield, IL.

Green, M. R. (1986) *Annu. Rev. Genet. 20*, 671–708.

Harding, C. R., & Scott, I. R. (1983) *J. Mol. Biol. 170*, 651–673.

Haydock, P. V., & Dale, B. A. (1986) *J. Biol. Chem. 261*, 12520–12525.

Haydock, P. V., & Dale, B. A. (1990) *DNA Cell Biol. 9*, 251–261.

Johnson, L. D., Idler, W. W., Zhou, X.-M., Roop, D. R., & Steinert, P. M. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 1896–1900.

Kozak, M. (1989) *J. Cell Biol. 108*, 229–234.

Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) in *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

McKinley-Grant, L. J., Idler, W. W., Bernstein, I. A., Parry, D. A. D., Cannizzaro, L., Croce, C. M., Huebner, K., Lessin, S. L., & Steinert, P. M., (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 4848–4852.

Resing, K. A., Walsh, K. A., & Dale, B. A. (1984) *J. Cell Biol. 99*, 1372–1378.

Resing, K. A., Dale, B. A., & Walsh, K. A. (1985) *Biochemistry 24*, 4167–4176.

Resing, K. A., Walsh, K. A., Haugen-Scofield, J., & Dale, B. A. (1989) *J. Biol. Chem. 264*, 1837–1845.

Rogers, G. E., Kuczek, E. S., McKinnon, P. J., Presland, R. B., & Fietz, M. J. (1989) in *Biology of Wool and Hair* (Rogers, G. E., Reis, P. J., Ward, K. A., & Marshall, R. C., Eds.) pp 69–85, Chapman and Hall, London.

Rothnagel, J. A., & Rogers, G. E. (1986) *J. Cell Biol. 102*, 1419–1429.

Rothnagel, J. A., & Steinert, P. M. (1990) *J. Biol. Chem. 265*, 1862–1865.

Rothnagel, J. A., Mehrel, T., Idler, W. W., Roop, D. R., & Steinert, P. M. (1987) *J. Biol. Chem. 262*, 15643–15648.

Steinert, P. M. (1988) *J. Biol. Chem. 263*, 13333–13339.

Steinert, P. M., & Roop, D. R. (1988) *Annu. Rev. Biochem. 57*, 503–625.

Steinert, P. M., Cantieri, J. S., Teller, D. C., Lonsdale-Eccles, J. D., & Dale, B. A. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 4097–4101.

Steinert, P. M., Parry, D. A. D., Idler, W. W., Johnson, L. D., Steven, A. C., & Roop, D. R. (1985) *J. Biol. Chem. 260*, 6142–7149.

Steven, A. C., Bisher, M. E., Roop, D. R., & Steinert, P. M. (1989) *J. Cell Biol. 109*, 258.

Tseng, H., & Green, H. (1988) *Cell 54*, 491–496.

White, R. L., Lalouel, J.-M., Nakamura, Y., Daus-Keller, H., Green, P., Bowden, D. W., Mathew, C. G. P., Easton, D. F., Robson, E. B., Morton, N. E., Gusella, J. F., Haines, J. L., Retief, A. E., Kidd, K. K., Murray, J. C., Lathrop, G. M., & Cann, H. M. (1990) *Genomics 6*, 393–412.