BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS **224,** 265–270 (1996)
ARTICLE NO. 1018

Pseudogenes for the Human Uracil-DNA Glycosylase on Chromosomes 14 and 16

Henning Lund, Ingrid Eftedal, Terje Haug, and Hans E. Krokan[1]

UNIGEN Center for Molecular Biology, Norwegian University for Science and Technology, N-7005 Trondheim, Norway

Received May 22, 1996

Two clones containing nonfunctional pseudogenes for the human uracil-DNA glycosylase gene have been isolated. The sequences of the two clones that are homologous to the *UNG* cDNA span 670 and 580 bp, respectively. In the longest of these, a full length Sx type *Alu* sequence interrupts the homologous sequence. Chromosomal mapping locates the clones to chromosomes 16 and 14. Comparison of the pseudogene sequences to the cDNA sequence indicates that the pseudogenes diverged from the functional gene approximately 31 and 22 million years ago, which is before the point in evolution when great apes and hominides separated. © 1996 Academic Press, Inc.

Uracil in DNA can arise from misincorporation of dUMP during replication, or as a result of deamination of cytosine [1-3]. The latter of these processes will cause a transition mutation if not repaired before the next round of replication. Uracil-DNA glycosylase (UDG) removes uracil from DNA as the first step in the pathway for repair of such damages [4, 5]. The importance of UDG is reflected by the variety of organisms in which this activity have been detected [3, 6-9] and by the fact that all investigated human cells and tissues have measurable UDG activities, though at varying degrees [10, 11]. UDG has been purified from human placenta [12], and the corresponding cDNA has been cloned [13]. This cDNA has been proven to encode the major nuclear form of UDG, as well as the mitochondrial form of the enzyme [14, 15]. Recently cloning of the *UNG* upstream sequence and analysis of its organization was reported [16]. During the process of isolating genomic sequences for the uracil-DNA glycosylase gene, we identified two pseudogenes for this gene. Pseudogenes belong to one of two groups, they either result from DNA duplication and are linked to their functional productive counterpart, or they result from retrotranscription of RNA, usually mRNA. The latter group, termed processed pseudogenes, is characterized among other things by their lack of intervening sequences and presence of a 3 terminal poly A tail. They also in most cases are localized on another chromosome than the functional gene [17]. We present here the sequence, characterization and chromosomal localization of the two *UNG* pseudogenes.

1. MATERIALS AND METHODS

Isolation of clones and preparation of DNA. Clones were isolated from two different λ-libraries, an EMBL-3 library provided by Clontech and propagated in *E. coli* Y1090, and a λGEM-11 based custom made library from Novagen, propagated in the recD$^-$ *E. coli* strain ER1647. Screenings were performed as described by the respective manufacturers. The libraries were screened at high stringency using the 2 kb UNG15 cDNA as probe, and positive λ-clones were isolated according to Sambrook et. al. [18] For further characterization, restriction fragments were subcloned into plasmid pGEM-3zf (Promega). Plasmid DNA was prepared by the alkaline extraction procedure reported by Birnboim and Doly [18], or by using the QIAGEN midi-prep plasmid isolation kit.

Restriction mapping and southern blot hybridization. Restriction fragments of phage DNA were resolved through

[1] Corresponding author. Fax: + 47 73 59 87 05.

0006-291X/96 $18.00

1% agarose gels and transferred to GeneScreen Plus membranes (DuPont). Blots were hybridized as described by Sambrook et. al. [18], and exposed to Kodak XAR-5 x-ray film. Restriction digests and probe labeling were performed according to standard methods.

DNA sequencing. Nucleotide sequences were determined using Sanger's dideoxy chain termination method [20] from the T7/SP6 promoter primer sites of the pGEM3zf vector.

Chromosomal assignment. The pseudogenes were localized to chromosomes by hybridizing restriction fragments of the clones to a human/hamster hybrid panel provided by Oncor. The hybridizations were performed using standard methods as described by Sambrook et. al [18] and a PhosphorImager (Molecular Dynamics) was used to make images of the hybridized membrane.

2. RESULTS AND DISCUSSION

A total of 10^6 plaque from the EMBL-3 library were screened in order to isolate genomic *UNG* clones. 16 positive clones were isolated. Upon Southern analysis, these clones were found to be positive only for a probe from the most 3′ part of the cDNA-sequence (position 1923-1953 in the UNG15 cDNA). The clones were digested with several restriction endonucleases to obtain restriction maps and were subjected to Southern analysis with probes from different parts of the cDNA to identify regions homologous to the cDNA sequence. In one 14 kb clone the hybridizing sequence was found to be confined to a 2 kb *Sph*I fragment, which was subcloned into pGEM3zf and sequenced. This 2 kb clone was given the name UNGps2. The custom made recD$^-$ Novagen library was screened in an attempt to avoid problems with the recombinational events that was thought to be the cause of the underrepresentation of clones containing the 5′ part of *UNG*. Screening of a total of 10^6 plaque from this library yielded one positive 15 kb clone. *Pst*I restriction of this clone produced 5 fragments, of which only a 1.1 kb fragment was proven positive for hybridization with a cDNA probe upon Southern analysis. The 1.1 kb fragment, UNGps1, was sequenced after having been subcloned into pGEM3zf. Both sequences have been entered in the EMBL/GenBank/DDBJ databases with accession numbers X92985 (UNGps1) and X92986 (UNGps2).

Chromosomal Assignment

Clone UNGps1 was assigned to chromosome 16 by hybridizing a 618 bp *Taq*I fragment from this clone (position 422 to 1040) to the hybrid panel, as one single signal in the lane corresponding to this chromosome is present in the autoradiogram. (fig. 1a). A 561 bp *Sph*I/*BsaA*I fragment from UNGps2 (position 12 to 573) was employed as probe for localization of this clone. The resulting autoradiogram (fig 1b) shows one prominent signal in the chromosome 14 lane, with an additional weaker signal in the lane for chromosome 12. This signal corresponds to the signal that can be seen when hybridizing with the UNG cDNA (data not shown) and confirms the assignment of the *UNG* gene to this chromosome [21]. There is also a signal in the lane corresponding to chromosome 1. This lane, however, is known to be contaminated by chromosome 13 and 14. Hence, UNGps2 is concluded to be localized on chromosome 14.

Sequence Analysis

UNGps1: Over a span of 670 bases, corresponding to position 913 to 1583 in the UNG15 cDNA clone, this clone showed approximately 83% homology to the cDNA sequence. A full length *Alu* sequence, situated between position 58 and 380 in the UNGps1 clone and contributing an additional 322 bp (including two 15 bp direct repeats at the 5′ and 3′ end of the *Alu* sequence) to the length of this clone, interrupts the homologous sequence and is excluded from the homology calculation. The *Alu* sequence was classified using the Pythia server at pythia@ twinpeaks.bim.anl.gov, and was shown to belong to the class *Alu-Sx* [22]. The sequence of UNGps1 is aligned against the cDNA sequence in fig. 2A. In table 1 the single base differences between the cDNA sequence and the homologous part of the UNGps1 sequence are

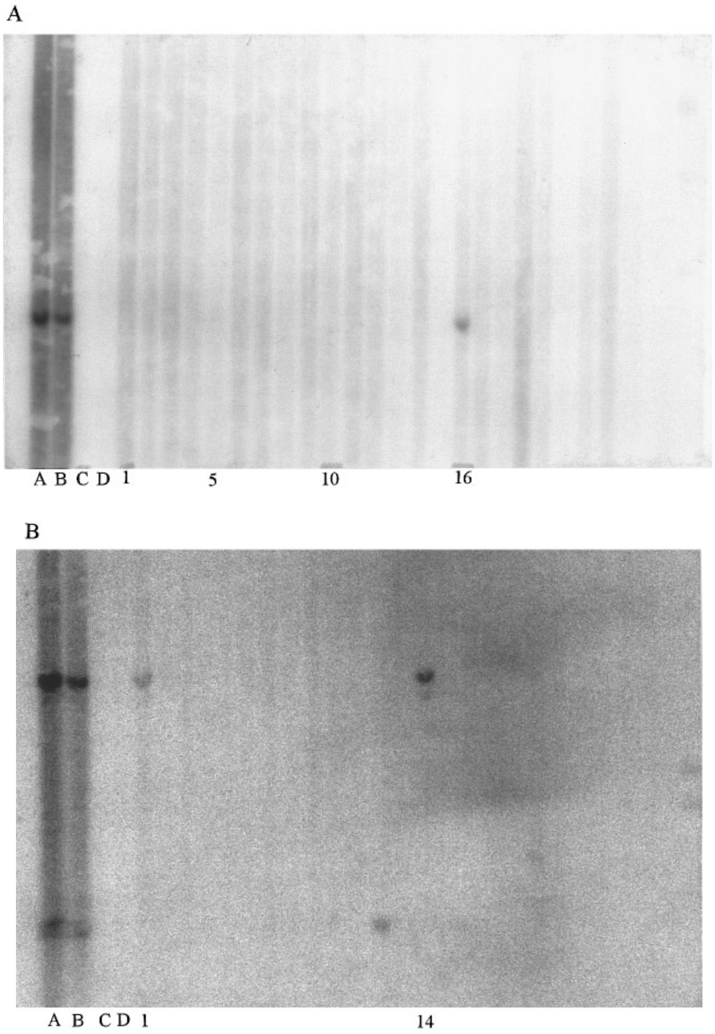

FIG. 1. PhosphorImager scans of the hybrid panel after hybridizing with the UNGps1 (A) and UNGps2 (B) probes. Lane A, Male human genomic DNA; lane B, Female human genomic DNA; lanes C and D, Mouse and hamster genomic DNA; the subsequent numbering of lanes correspond to the human chromosome number.

summarized, along with the expected values for non functional pseudogenes as presented in [23]. The region of homology corresponds to a segment near the 5′ end of the 1.2 kb 3-terminal exon of the *UNG*-gene (Terje Haug and Frank Skorpen, unpublished results). It does not, however, span any exon/intron splice sites. Neither does this sequence reach as far as to the 3′-end of the cDNA sequence. Based on the degree of homology between this clone and the *UNG* cDNA, and on the fact that the homologous sequence of UNGps1 is interrupted by an *Alu* sequence, it can be concluded that this is a nonfunctional pseudogene for the human uracil-DNA glycosylase.

UNGps2: This clone showed 90% homology to the 580 3′ basepairs of the *UNG* cDNA sequence when a 17 base pair insert was left out. A sequence similar to the polyadenylation signal from *UNG* cDNA, differing in a single-base insert, is found in the UNGps2 sequence. This sequence is followed 15 bases further downstream by a sequence of approximately 30 bases with a high

A

```
aggaagcggcaccatgtactacagacGGCTCATCCCTCCCCTTTG
                                        TG

TCAGTGTATAGAGGGTTCTTTGGATGTAGACACTTTTC
  T CA        Δ                     (322 bp Alu)

AAAGACCAATGAGCTGCTGCAGAAGTCTGGCAAGAAGCCCATTGA
T  A            TCT        C A       T   C

CTGGAAGGAGCTGTGATCATCAG CTGAGG
                  C   AC  A   ggtggtctttctcca

GGTGGCCTTTGAGAAGCTG  CTGTTAACGTATTTGCCagttacg
     TT        A CTGC     G A         caggttt

aagttccact   GAAAATTTTCCTATTAATTCTTAAGTACTCTG
atgaaggcggaat        C  C

CATAAGGGGGAAAAGCTTCCAGAAAGCAGCCATGAACCAG
        A C<   Δ     >     TT    A  G   ccctt

        GCTGTCCAGGAATGGCAGCTGTATCCAACCACAAACA
gaacccag A  C   A           T      GA  G C

ACAAAGGCTACCCTTTGACCAAATGTCTTTCTCTGCAACATGGCT
G     A A  T     G                  TT

TCGGCCTAAAATATGC         AGAAGACAGATGAGGTCAAA
                agcttctgc

TACTCAGTTGGCTCTCTTTATCTCCCTTGCCTTTAT GGTGAAACA
      C C       C C C A     A       T   T   TG

GGGGAGATGTGCACCTTTCAGGCACAGCCCTAGTTTGGCGCCTGC
TT     G          TT  T     TG   C    T  A

TGCTCCTTGGTTTTGCCT    GGTTAGACTTTCAGTGAC AGAT
  T               ggtg               A TA  G

GTTGGGGTGTTTTTGCTTAGAAAGGtccccttgtctcagccttgc
         A       G A     gatctcctgggcccaaccca

aggGCAGGCATGCCAGTCTCTGCCAGTTCCACTGCCCCCTTGATC
ΔΔΔ        T     TG A     A AT  C TT        A

TTTGAAGGAGTCCTCAGGCCCCTCGC       AGCATAAGGATG
   C     TC        TT   cttgaagga    GG  A

TTTTGCAACTTTCCAGAA TCTGGCCCAGAAATTAGGGCTCAATTT
    TTGCG    T    C                A  A  G GC

CCTgattgt
 T
```

B

```
tcagccttgcagggcagGCATGCCAGTCTCTGCCAGTTCCACTGCCCC
                                    G   G   G

CTTGATCTTTGAAGGAGTCCTCAGGCCCCTCGCAGCATAAGGATGTTT
Δ                          T    A          C

TGCAACTTTCCAGAATCTGGCCCAGAAATTAGGGCTCAATTTCCTGAT
     G                                         C

TGTAGTAGAGGTTAAGA                 TTGCTGTGAGCTT
           CA    tacttaaggaatagccc         A C

TATCAGATAAGAGACCGAGAGAAGTAAGCTGGGTCTTGTTATTCCTTG
       C G T Δ   G       G                  T

GGTGTTGGTGGAATAAGCAGTGGAATTTGAACAAGG AAGAGGAGAAA
                      A            AA

AGGGAATTTTGTCTTTAT   GGGGTGGGGTGATTTT CTCCTAGGGT
         CA     TATGG   TG           G

TATGTCCAGTTGGGGTTTTTAAGGCAGCACAGACTGCCAAGTACTGTT
   C  A                            A    A G   ΔΔ

TTTTTTAACCGACTGAAATCACTTTGGGATATTTTTTCCTGCAACACT
      T T A                  G  A    T G

GGAAAGTTTTAGTTTTTTAAGAAGTACTCATGCAGATATATATATATA
      A  AGΔΔ       A             T  G

TATTTTTCCCAGTCCTTTTTTTAAGAGACGGTCTTTATTGGG TCTGCA
 ΔΔΔΔ G  AT C Δ              AA           G

CCTCCATCCTTGATCTTGTTAGCAATGCTGTTTTTGCTGTTAGTCGGG
                      TG                   C A

TTAGAGTTGGCTCTACGCGAGGTTTGTTAATAAAAGTTTGTTAAAAGT
           C     T G      G ___G___ A

Tcaaaaaaaaaaaaaaaaacccg
 ttgaaaagttttgaaaataataataacaatag
```

FIG. 2. Alignment of (A) UNGps1 and (B) UNGps2 to the UNG15 cDNA sequence. The upper line shows the complete cDNA sequence, the lower line indicates where and how the pseudogene sequences differ from the cDNA sequence. Homologous regions are written in capital letters, a deletion in the pseudogene sequence compared to the cDNA sequence is indicated by a Δ; an insertion in the pseudogene is indicated by a gap in the cDNA sequence. 〈Δ〉 indicates that the region between the brackets is deleted. The place of insertion of the 322 bp *Alu* sequence in UNGps1 is indicated.

A-content, which is reminiscent of a polyA tail. The UNGps2- and cDNA sequences are aligned in fig. 2B. The point mutations of UNGps2 compared to the cDNA sequence are summarized in table 2. It can be concluded that UNGps2 is also a pseudogene for *UNG.*

Since these clones do not span any intron/exon splice sites it is not possible to check them for lack of intervening sequences, which is one of the major hallmarks of processed pseudogenes. Their location on different chromosomes than the *UNG* gene and the possible poly A tail of UNGps2, however, indicates that these sequences both should be considered as processed pseudogenes.

The base substitution patterns given in tables 1 and 2 are interesting compared to the expected values given in the same tables. For UNGps1 the base substitution pattern is not very different from what can be expected in nonfunctional pseudogenes. UNGps2, however, shows a pattern

TABLE 1
Base Substitutions in UNGps1[a]

Base in functional gene	Base in UNGps1				Total
	A	T	C	G	
A		3.9 (4.7 ± 1.3)	3.9 (5.0 ± 0.8)	9.8 (9.4 ± 1.3)	17.6 (19.1)
T	5.9 (4.4 ± 1.1)		14.7 (8.2 ± 1.3)	5.9 (3.3 ± 1.2)	26.5 (15.9)
C	6.7 (6.5 ± 1.1)	20.6 (21.0 ± 2.1)		4.0 (4.2 ± 0.5)	31.3 (31.7)
G	12.7 (20.7 ± 0.2)	8.8 (7.2 ± 1.1)	2.0 (5.3 ± 1.1)		23.5 (33.2)
Total	25.3 (31.6)	33.3 (32.9)	20.6 (18.5)	19.7 (16.9)	

[a] The figures show the percentage of the total number of base substitutions in each category. The figures in parentheses are the expected values for base substitutions in pseudogenes as presented in [23].

that deviates significantly from the expected values. As can be seen in the lower row of table 2, what is expected is a drift towards an increased A/T content, but this has not happened in the case of UNGps2. This is thought to be explained by the fact that this pseudogene fragment corresponds to the 3′ part of *UNG* which already has an A/T content of about 60%. This approximately equals the average A/T content in non-coding regions of the genome and is very close to the predicted equilibrium level in pseudogenes [24]. Therefore, there has been no net drift towards an increase of this content after the pseudogene was established.

Age

Using an average base substitution rate of 4.18 × 10^{-9} substitutions per site per year [25], a calculation based on the single base substitutions only indicates that pseudogene UNGps1 is approximately 31 million years old. The same calculation gives an age of approximately

TABLE 2
Base Substitutions in UNGps2[a]

Base in functional gene	Base in UNGps2				Total
	A	T	C	G	
A		7.8 (4.7 ± 1.3)	0 (5.0 ± 0.8)	17.6 (9.4 ± 1.3)	25.4 (19.1)
T	5.9 (4.4 ± 1.1)		15.7 (8.2 ± 1.3)	9.8 (3.3 ± 1.2)	31.4 (15.9)
C	2.0 (6.5 ± 1.1)	11.8 (21.0 ± 2.1)		3.9 (4.2 ± 0.5)	17.7 (31.7)
G	19.6 (20.7 ± 0.2)	3.9 (7.2 ± 1.1)	2.0 (5.3 ± 1.0)		25.5 (33.2)
Total	27.5 (31.6)	23.5 (32.9)	17.7 (18.5)	31.3 (16.9)	

[a] Presentation is as in Table 1.

22 million years for the UNGps2 pseudogene. Thus, both these pseudogenes were established before the point in evolution at which great apes and hominids diverged [26]. According to Britten [27], the process of insertion of *Alu-Sx* sequences ceased about 30 mill. years ago. Therefore, the *Alu* sequence in UNGps1 must have been inserted a short time after the pseudogene itself was established.

ACKNOWLEDGMENTS

This work was supported by the Norwegian Research Council. We greatly appreciate the contribution from Dr. Berit Johansen and Eva Støkken in the initial phase of this work.

REFERENCES

1. Lindahl, T., and Nybergh, B. (1974) *Biochemistry* **13,** 3405–3410.
2. Tye, B. K., Nyman, P. O., Lehman, I. R., Hochhauser, S., and Weiss, B. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 154–157.
3. Wist, E., Unhjem, O and Krokan, H. (1978) *Biochem. Biophys. Acta.* **520,** 253–270.
4. Franklin, W. A., and Lindahl, T. (1988) *EMBO J.* **7,** 3617–3622.
5. Dianov, G., Price, A., and Lindahl, T. (1992) *Mol. Cell. Biol.* **12,** 1605–1612.
6. Sekiguchi, M., Hayakawa, H., Makino, F., Tanaka, K., and Odaka, Y. (1976) *Biochem. Biophys. Res. Commun.* **73,** 293–299.
7. Caradona, S. J., and Cheng, Y. C. (1980) *J. Biol. Chem.* **255,** 2293–2300.
8. Talpaert-Borlé, M., Campagnari, F., and Creissen, D. M. (1982) *J. Biol. Chem.* **257,** 1208–1214.
9. Arenaz, P., and Sirover, M. A. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 5822–5826.
10. Krokan, H. E., Haugen, Aa., Myrnes, B., and Guddal, P. H. (1983) *Carcinogenesis* **4,** 1559–1564.
11. Myrnes, B., Giercksky, K.-E., and Krokan, H. (1983) *Carcinogenesis* **4,** 1565–1658.
12. Wittwer, C. U., Bauw, G., and Krokan, H. E. (1989) *Biochemistry* **28,** 780–784.
13. Olsen, L. C., Aasland, R., Wittwer, C. U., Krokan, H. E., and Helland, D. E. (1989) *EMBO J.* **8,** 3121–3125.
14. Slupphaug, G., Olsen, L. C., Helland, D., Aasland, R., and Krokan, H. E. (1991) *Nucleic Acids Res.* **19,** 5131–5137.
15. Slupphaug, G., Markussen, F.-H., Olsen, L. C., Aasland, R., Aarsæther, N., Bakke, O., Krokan, H. E., and Helland, D. E. (1993) *Nucleic Acids Res.* **21,** 2579–2584.
16. Haug, T., Skorpen, F., Lund, H., and Krokan, H. E. (1994) *FEBS Letters,* **353,** 180–184.
17. Vanin, E. F. (1985) *Annu. Rev. Genet.* **19,** 253–272.
18. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) A Laboratory Manual, 2nd ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
19. Birnboim, H. C., and Doly, J. (1979) *Nucleic Acids Res.* **7,** 1513–1523.
20. Sanger, F., Nicklen, S., and Coulson, A. A. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
21. Aasland, R., Olsen, L. C., Spurr, N., Krokan, H. E., and Helland, D. E. (1990) *Genomics* **7,** 139–141.
22. Jurka, J., and Milosavljevic, A. (1991) *J. Mol. Evol.* **32,** 105–121.
23. Li, W.-H., and Graur, D. (1991) Fundamentals of Molecular Evolution, Sunderland, Sinauer Assoc.
24. Bulmer, M. (1986) *Mol. Biol. Evol.* **3,** 322–329.
25. Li, W.-H., Luo, C.-C., and Wu, C.-I. (1985) *in* Moleculary Evolution Genetics (Macintyre, R. J., Ed.), pp. 1–94, Plenum Press, New York.
26. Kimura, M. (1983) *in* The Neutral Theory of Molecular Evolution, pp. 65–76, Cambridge Univ. Press, Cambridge.
27. Britten, R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 6148–6150.