MOLECULAR AND CELLULAR BIOLOGY, Sept. 1985, p. 2159–2171
0270-7306/85/092159-13$02.00/0
Copyright © 1985, American Society for Microbiology
Vol. 5, No. 9

Human U1 Small Nuclear RNA Genes: Extensive Conservation of Flanking Sequences Suggests Cycles of Gene Amplification and Transposition

LAUREL B. BERNSTEIN,[1]† TIM MANSER,[2]‡ AND ALAN M. WEINER[1]*

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06510,[1] and Department of Biology, Howard Hughes Medical Institute, University of Utah, Salt Lake City, Utah 84112[2]

Received 28 February 1985/Accepted 5 June 1985

The DNA immediately flanking the 164-base-pair U1 RNA coding region is highly conserved among the approximately 30 human U1 genes. The U1 multigene family also contains many U1 pseudogenes (designated class I) with striking although imperfect flanking homology to the true U1 genes. Using cosmid vectors, we now have cloned, characterized, and partially sequenced three 35-kilobase (kb) regions of the human genome spanning U1 homologies. Two clones contain one true U1 gene each, and the third bears two class I pseudogenes 9 kb apart in the opposite orientation. We show by genomic blotting and by direct DNA sequence determination that the conserved sequences surrounding U1 genes are much more extensive than previously estimated: nearly perfect sequence homology between many true U1 genes extends for at least 24 kb upstream and at least 20 kb downstream from the U1 coding region. In addition, the sequences of the two new pseudogenes provide evidence that class I U1 pseudogenes are more closely related to each other than to true genes. Finally, it is demonstrated elsewhere (Lindgren et al., Mol. Cell. Biol. 5:2190–2196, 1985) that both true U1 genes and class I U1 pseudogenes map to chromosome 1, but in separate clusters located far apart on opposite sides of the centromere. Taken together, these results suggest a model for the evolution of the U1 multigene family. We speculate that the contemporary family of true U1 genes was derived from a more ancient family of U1 genes (now class I U1 pseudogenes) by gene amplification and transposition. Gene amplification provides the simplest explanation for the clustering of both U1 genes and class I pseudogenes and for the conservation of at least 44 kb of DNA flanking the U1 coding region in a large fraction of the 30 true U1 genes.

Eucaryotes, and to a lesser extent procaryotes, often meet the demand for large amounts of a gene product by expanding single genes into multigene families. In most higher eucaryotes, abundant proteins such as the tubulins (8), actins (15), and histones (17) are encoded by multiple genes, as are almost all non-mRNA species (2, 25). A multigene family presents evolutionary problems not encountered for single-copy genes. How was the original gene duplicated or amplified? How is homogeneity maintained between individual members of the gene family? And, in certain cases (e.g., the 5S RNA genes of *Schizosaccharomyes pombe* [36]), how were the gene copies dispersed in the genome? We have been studying the multigene family for human U1 small nuclear RNA to gain insight into these questions.

U1 small nuclear RNA is an abundant, homogeneous, 164-nucleotide RNA species that participates in the splicing of pre-mRNA (26, 28, 42, 49, 65). The variety and abundance of human genomic sequences homologous to U1 RNA illustrate the complex structure and organization that can arise in a multigene family. Based on genomic blotting data (11), our minimum estimate for the number of homologies to U1 RNA in the human genome is 500 to 1,000 copies. However, all but about 30 of these copies (31) are unexpressed pseudogenes that have a wide variety of defects with respect to the true U1 genes (10, 11, 34, 40). DNA sequence analysis of six representative human U1 genes revealed dramatic homology among flanking sequences of U1 genes (35). Virtually perfect homology between individual U1 genes extends at least 2.6 kilobases (kb) upstream from the U1 coding region, and the first 100 nucleotides downstream also exhibit substantial although much less dramatic conservation of sequence. More recently, significant homology between U1 genes has been found to extend for at least 2.3 kb downstream as well (19). The true U1 genes are clustered, since the vast majority (if not all) of U1 genes are located in band 1p36 on the short arm of chromosome 1 (29, 30, 44). However, the average intergenic distance appears to exceed 15 kb (34, 35).

The structural features of cloned U1 pseudogenes suggest that both DNA- and RNA-mediated mechanisms are responsible for generating defective members of the extended U1 multigene family (11, 40). The DNA-mediated mechanisms (proposed for class I pseudogenes) work exclusively at the DNA level and thus preserve to some degree both the 5′ and 3′ flanking sequences of the true U1 genes. The RNA-mediated mechanisms (proposed for class II and class III pseudogenes) involve a reverse flow of genetic information from the small nuclear RNA back into genomic DNA, possibly through a cDNA intermediate (3, 11, 59, 61), and consequently, the flanking sequences of the true genes are absent from these pseudogenes.

In this paper and elsewhere (29), the relationship between the true U1 genes and the closely related class I U1 pseudogenes is addressed. Before this work, we did not know whether the class I pseudogenes were interspersed with the true genes or whether they were located separately from them. Thus, we could not determine whether class I pseudogenes were derived from the contemporary set of true U1

* Corresponding author.
† Present address: Institute of Molecular Biology, University of Oregon, Eugene, OR 97403.
‡ Present address: Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139.

genes or whether they were relics of a more ancient U1 gene family.

It is demonstrated elsewhere that human class I U1 pseudogenes are clustered in region 1q12-q22 within the long arm of chromosome 1 (29), whereas the true U1 genes are clustered in band 1p36 of the short arm on the opposite side of the centromere (30, 44). Here we present evidence that very good (and perhaps nearly perfect) DNA sequence homology between many individual U1 genes extends for at least 24 kb upstream and at least 20 kb downstream from the U1 coding region; the reduced level of homology previously observed for sequences immediately downstream from the U1 coding sequences (35) appears to represent a localized region of DNA sequence polymorphism. Thus, the intergenic distance between many true U1 genes exceeds 44 kb. We also present evidence that the distance between class I pseudogenes generally exceeds 35 kb (but can be as small as 9 kb) and that class I pseudogenes are more closely related to each other than to the true U1 genes. We argue that all of the data can best be accounted for by repeated cycles of gene amplification and recombination, in some cases accompanied by transposition. Furthermore, we suggest that the class I U1 pseudogenes are the aging ancestors of the contemporary functional U1 gene family.

MATERIALS AND METHODS

Construction and screening of a human cosmid library. A cosmid library was constructed essentially as described by Ish-Horowicz and Burke (20), except that the recipient pBR322-based cosmid vector was pHC79 (18) instead of pJB8 and the recipient *Escherichia coli* strain was the recA$^-$ strain 1046 instead of HB101. The inserted human genomic DNA fragments were generated by partial *Mbo*I digestion and size fractionation (20) of DNA from the blood of a 22-year-old male.

Two separate screenings of the cosmid library were performed. First, approximately 3×10^4 recombinants were screened by colony hybridization to nick-translated p5P2 (a subclone containing the 2.5-kb *Hin*dII-*Pvu*II fragment located 105 base pairs [bp] upstream of the U1 coding region of clone HSD2 [35]), as described previously (20). Positive colonies were screened with nick-translated pD2F (a U1 gene subclone containing the U1 coding region, 105 bp of 5′ flanking sequences, and 500 bp of 3′ flanking sequences [44]). Colonies hybridizing to both probes were purified; these yielded clones cosD1 and cosD21. In a second screen, approximately 5×10^4 recombinant colonies scoring as positive with the pD2F probe were hybridized with p5P2, and only those colonies that were positive with both probes were purified; these yielded the cosDA clone.

Restriction mapping and subcloning of recombinant cosmids. Cosmid DNAs were prepared and mapped by standard procedures (33). To quickly map the large inserts of these cosmids, restriction enzymes were chosen that cut each clone at only a few sites. Consequently, different sets of enzymes were used to map the three clones. (The possible significance of the differences between the restriction maps of cosD1 and cosD21 is discussed below.) We took advantage of several invariant restriction sites found within the 5′ flanking regions of U1 genes (*Bgl*II at −6, *Pvu*II at −100, and *Pst*I at −1500) to determine the precise locations and the orientations of the U1 coding regions of cosD1 and cosD21, with blots probed with p5P2.

The vector-insert junction subclone pD1Pv was created by complete digestion of cosD1 with *Pvu*II and by circularization of the resulting fragments by T4 DNA ligase treatment at high DNA dilution. Upon transfection into *E. coli*, the only viable species was pD1Pv, which retains parts of the pHC79 vector bearing the pBR322 origin of replication, the β-lactamase gene, and 3.8 kb of human DNA from the downstream junction of the cosD1 insert with vector DNA (Fig. 1). Similarly, the pD1Ec clone containing 400 bp of human DNA from the upstream junction of cosD1 DNA with vector sequences was constructed by complete *Eco*RI digestion of cosD1, followed by dilution, ligation, and transfection. Neither pD1Pv nor pD1Ec contains highly or moderately repetitive human DNA sequences, as judged by the absence of hybridization to nick-translated human placental DNA.

To isolate one of the two U1 pseudogenes of cosD8A, an 11-kb *Hin*dIII fragment containing the cosD8A-2 pseudogene was subcloned into the *Hin*dIII site of pBR322, creating plasmid pHin2 (Fig. 1). pHin2 was mapped by the type of blotting experiments described above for the cosmid clones. To orient the cosD8A-1 and cosD8A-2 pseudogenes, blots of cosD8A and pHin2 digests were hybridized at moderate stringency (42°C; 50% formamide, 5× SSC [1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate]) against the U1 gene 5′-flanking probe p5P2 (this probe has homology to both U1 gene and class I pseudogene 5′ flanking sequences). p5P2 hybridized both to the *Nco*I-*Sac*I fragment of pHin2 located next to the pseudogene cosD8A-2 and also to the short *Sal*I-*Mlu*I fragment of cosD8A adjacent to the cosD8A-1 pseudogene. Therefore, the cosD8A pseudogenes have been oriented in a head-to-head manner (Fig. 1).

DNA sequence analysis. For cosmid clones cosD1, cosD21, and cosD8A, DNA fragments spanning the U1 coding regions were subcloned into the M13 vectors mp8 and mp9 (39) and sequenced by the method of Sanger et al. (51). Klenow polymerase was a generous gift of C. Joyce, Yale University. To compare the sequences located 2 kb downstream from the lambda U1 clones HSD1 and HSD4, a 516-bp *Pst*I fragment mapping 2 kb downstream from the HSD1 U1 coding region was isolated; the map position of this fragment was identified by using the restriction map of HSD1 developed by Manser and Gesteland (see Fig. 1 of reference 35). The *Pst*I fragment of HSD1 hybridized strongly to a 700-bp *Bgl*II fragment of HSD4. Both of these fragments were subcloned into M13 mp9 and sequenced as described above. From the sequence data, as well as from our own mapping data (unpublished data) and that of Htun et al. (19), it is clear that the *Bgl*II fragment of HSD4 spans the entire *Pst*I fragment of HSD1. A typographical error in the original map of HSD4 (see Fig. 1 of reference 35) indicated the *Pvu*II site within the sequenced *Bgl*II fragment as a *Pst*I site.

Genomic blotting procedures. Placental and leukocyte DNA were prepared as described previously (23) or were received as a gift from Allan Wilson and Elizabeth Zimmer, University of California, Berkeley. Genomic blots were prepared and probed by the method of Southern (56) with modifications as described previously (60). In the haploid human genome the numbers of copies of sequences homologous to the U1 distant flanking probes pD1Pv and pD1Ec were estimated by densitometric comparison of genomic DNA band intensities with reconstruction lanes containing pBR322 DNA at concentrations representing 1 or 10 copies of a sequence per haploid genome (as calculated by assuming a haploid genome size of 3.2×10^9 bp). Copy number estimates were made by comparing the total amount of hybridization of the pD1Pv and pD1Ec probes to the pBR322 lanes (via vector sequences) with the total amount of hybrid-

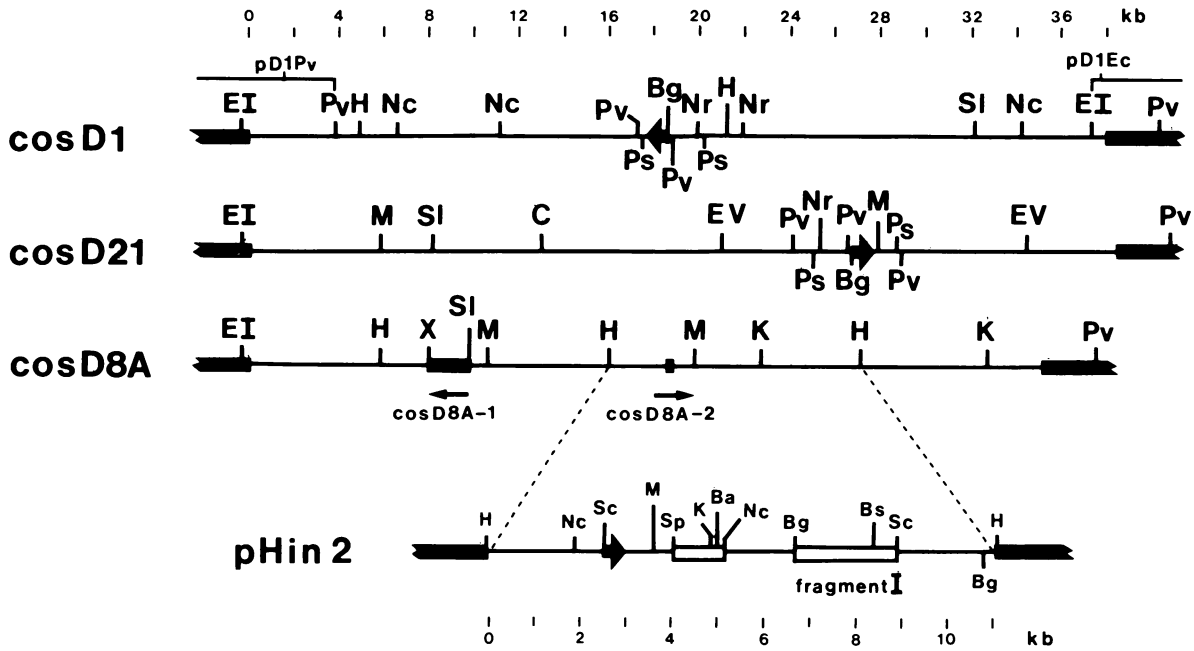


FIG. 1. Restriction maps of three regions of the human genome containing either true U1 genes or class I U1 pseudogenes. Exterior thick lines represent the ends of the vectors (pHC79 for cosD1, cosD21, and cosD8A; pBR322 for pHin2). Inserts of human DNA (thin lines) for cosD1 and cosD21 are marked with heavy arrows to show the locations and orientations of U1 RNA coding sequences. Dark boxes in cosD8A show the approximate locations of the two U1 coding regions, designated cosD8A-1 and cosD8A-2; their orientations are indicated below the boxes by thin arrows. The heavy arrow of pHin2 shows the precise location of the cosD8A-2 U1 sequence in this subclone of cosD8A. The significance of fragment I of pHin2 is discussed in the text. Restriction enzyme sites are denoted as follows: Ba, *Bam*HI; Bg, *Bgl*II; Bs, *Bst*EII; EI, *Eco*RI; EV, *Eco*RV; H, *Hin*dIII; K, *Kpn*I; M, *Mlu*I; Nc, *Nco*I; Nr, *Nru*I; Ps, *Pst*I; Pv, *Pvu*II; Sc, *Sac*I; Sl, *Sal*I; Sp, *Sph*I; X, *Xho*I. Partial brackets above the cosD1 map indicate the origins of subclones pD1Pv and pD1Ec, which were constructed by digestion and recircularization of cosD1. pD1Pv contains insert sequences from the left end of the cosD1 map and extends into vector sequences through the *Pvu*II site indicated at the far right end of the map. Likewise, pD1Ec includes insert sequences from the right end of the map and extends into vector sequences through the *Eco*RI site. The cosD1 and cosD21 maps show only a few selected sites near the U1 coding regions or the vector-insert junctions for the enzymes *Bgl*II, *Eco*RI, *Pst*I, and *Pvu*II. The four restriction maps are complete for the remainder of the enzymes shown, except that for cosD1 two small *Hin*dIII fragments (less than 2 kb) and two small *Nco*I fragments (less than 2 kb) have not been mapped.

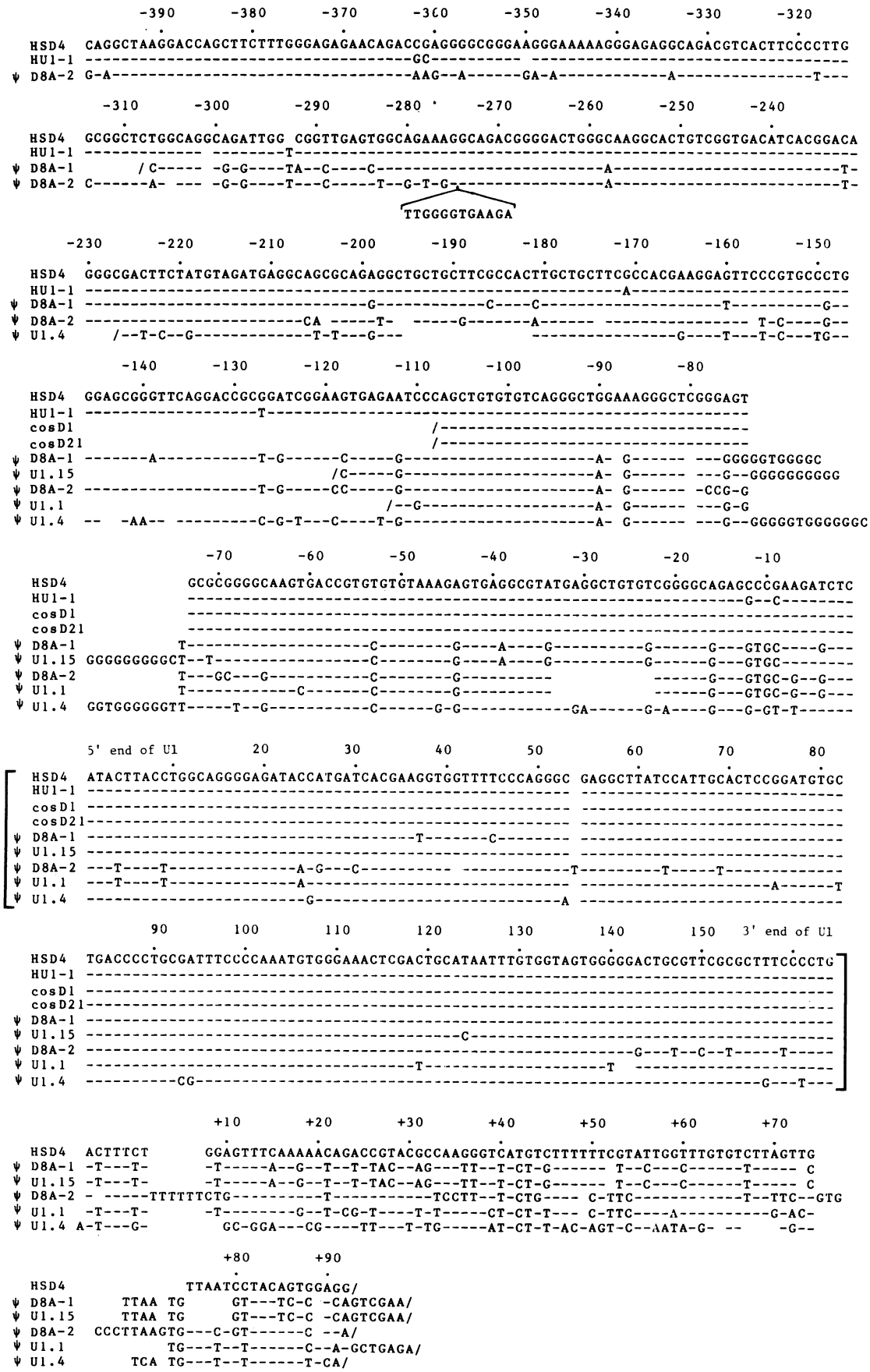
ization of the probes to genomic DNA (via insert sequences); the resulting figure of approximately 15 copies for either sequence was normalized to account both for the different ratios of vector and insert sequences in pD1Pv and pD1Ec and for the greater length of the insert sequence in pD1Pv as compared with pD1Ec.

RESULTS

Isolation and initial characterization of three cosmid clones containing U1 RNA homologies. With the intent of demonstrating physical linkage between two cloned U1 RNA genes, a human DNA library in cosmid vector pHC79 was constructed and screened for U1 genes. We used cosmid vectors because the 40-kb insert size afforded a better chance of finding two U1 loci on one recombinant cosmid than did the lambda phage vectors used previously (average insert size, 15 kb). A screen of approximately 80,000 colonies yielded 11 cosmids which hybridized strongly to both a U1 gene coding region probe (pD2F) and a U1 gene 5′ flanking sequence probe (p5P2). Hybridization with p5P2 indicated that these 11 cosmids contained either true U1 genes, whose immediate 5′ flanking sequences are virtually perfect matches for the probe sequence, or class I pseudogenes, whose 5′ flanking sequences share more than 80% homology with the true U1 genes.

To tentatively classify the U1 homologies on the cosmids as true genes or pseudogenes, we took advantage of the fact that restriction-fragment-length polymorphisms can be used to divide the U1 genes into a small number of groups. For instance, the great majority of U1 genes are found on *Pst*I fragments of either 3.9 or 2.65 kb (35). The 3.9-kb *Pst*I U1 group and the 2.6-kb *Pst*I U1 group appear as multiple copy bands above a background of fainter bands on blots of human genomic DNA hybridized to a U1 coding region probe. All of the previously cloned U1 genes are representatives of one or the other *Pst*I group (35; unpublished data). However, restriction-fragment-length polymorphisms cannot be used to divide the 30 U1 genes into unique groups, since some enzymes define three or more major groups. For example, *Pvu*II divides the U1 genes into four major groups, but two U1 genes belonging to different *Pst*I groups may belong to the same *Pvu*II group (35). Thus, a recombinant cosmid is very likely to represent a true U1 gene if digestion with several different restriction enzymes invariably produces a U1-containing fragment that corresponds to one of the major genomic U1 fragments characteristic of each enzyme.

Of the 11 cosmids, only 2 (cosD1 and cosD21) carried the U1 homology on restriction fragments whose lengths were always characteristic of true U1 gene groups (data not shown). A third cosmid (cosD8A) produced, for each restriction enzyme used, two fragments hybridizing to the U1 probe. The sizes of both of these fragments seldom corresponded to major U1 gene groups, suggesting that both U1 homologies were class I pseudogenes (data not shown).


-390 -380 -370 -360 -350 -340 -330 -320
HSD4 CAGGCTAAGGACCAGCTTCTTTGGGAGAGAACAGACCGAGGGGCGGGAAGGGAAAAAGGGAGAGGCAGACGTCACTTCCCCTTG
HU1-1 GC
ψ D8A-2 G-A AAG--A GA-A A T
-310 -300 -290 -280 -270 -260 -250 -240
HSD4 GCGGCTCTGGCAGGCAGATTGG CGGTTGAGTGGCAGAAAGGCAGACGGGGACTGGGCAAGGCACTGTCGGTGACATCACGGACA
HU1-1 T
ψ D8A-1 /C----- -G-G----TA--C----C A T-
ψ D8A-2 C------A- --- -G-G----T---C-----T--G-T-G- A T-
TTGGGGTGAAGA
-230 -220 -210 -200 -190 -180 -170 -160 -150
HSD4 GGGCGACTTCTATGTAGATGAGGCAGCGCAGAGGCTGCTGCTTCGCCACTTGCTGCTTCGCCACGAAGGAGTTCCCGTGCCCTG
HU1-1 A
ψ D8A-1 G C C T G--
ψ D8A-2 CA -----T- -----G A T-C----G--
ψ U1.4 /--T-C--G T-T---G- G----T---T-C---TG--
-140 -130 -120 -110 -100 -90 -80
HSD4 GGAGCGGGTTCAGGACCGCGGATCGGAAGTGAGAATCCCAGCTGTGTGTCAGGGCTGGAAAGGGCTCGGGAGT
HU1-1 T
cosD1 /
cosD21 /
ψ D8A-1 -------A----------T-G------C-----G----------A- G------ ---GGGGGTGGGGC
ψ U1.15 /C-----G----------A- G------ ---G--GGGGGGGGGG
ψ D8A-2 ------------------T-G-----CC-----G----------A- G------ -CCG-G
ψ U1.1 /--G----------A- G------ ---G-G
ψ U1.4 -- -AA-- ---------C-G-T---C----T-G----------A- G------ ---G--GGGGGTGGGGGGC
-70 -60 -50 -40 -30 -20 -10
HSD4 GCGCGGGGCAAGTGACCGTGTGTGTAAAGAGTGAGGCGTATGAGGCTGTGTCGGGGCAGAGCCCGAAGATCTC
HU1-1 G--C--------
cosD1
cosD21
ψ D8A-1 T----------C--------G----A----G----------G------G---GTGC----G---
ψ U1.15 GGGGGGGGGCT--T----------C--------G----A----G----------G------G---GTGC--------
ψ D8A-2 T---GC---G----------C--------G---------- ------G---GTGC-G--G---
ψ U1.1 T----------C--------G---------- ------G---GTGC-G--G---
ψ U1.4 GGTGGGGGGTT-----T--G----------C------G-G----------GA------G-A----G---G-GT-T------
5' end of U1 20 30 40 50 60 70 80
HSD4 ATACTTACCTGGCAGGGGAGATACCATGATCACGAAGGTGGTTTTCCCAGGGC GAGGCTTATCCATTGCACTCCGGATGTGC
HU1-1
cosD1
cosD21
ψ D8A-1 T-------C--------
ψ U1.15
ψ D8A-2 ---T----T----------A-G---C---------- ----------T---------T-----T----------
ψ U1.1 ---T----T----------A---------- ----------A------T
ψ U1.4 ----------G----------A
90 100 110 120 130 140 150 3' end of U1
HSD4 TGACCCCTGCGATTTCCCCAAATGTGGGAAACTCGACTGCATAATTTGTGGTAGTGGGGGACTGCGTTCGCGCTTTCCCCTG
HU1-1
cosD1
cosD21
ψ D8A-1
ψ U1.15 C
ψ D8A-2 G---T--C--T-----T-----
ψ U1.1 T T
ψ U1.4 ----------CG G---T---
+10 +20 +30 +40 +50 +60 +70
HSD4 ACTTTCT GGAGTTTCAAAAACAGACCGTACGCCAAGGGTCATGTCTTTTTTCGTATTGGTTTGTGTCTTAGTTG
ψ D8A-1 -T---T- -T-----A--G--T--T-TAC--AG---TT--T-CT-G------ T--C---C------T----- C
ψ U1.15 -T---T- -T-----A--G--T--T-TAC--AG---TT--T-CT-G------ T--C---C------T----- C
ψ D8A-2 - -----TTTTTTCTG----------T----------TCCTT--T-CTG---- C-TTC----------T--TTC--GTG
ψ U1.1 -T---T- -T--------G--T-CG-T----T-T-----CT-CT-T--- C-TTC----A----------G-AC-
ψ U1.4 A-T---G- GC-GGA---CG----TT---T-TG-----AT-CT-T-AC-AGT-C--AATA-G- -- -G--
+80 +90
HSD4 TTAATCCTACAGTGGAGG/
ψ D8A-1 TTAA TG GT---TC-C -CAGTCGAA/
ψ U1.15 TTAA TG GT---TC-C -CAGTCGAA/
ψ D8A-2 CCCTTAAGTG---C-GT------C --A/
ψ U1.1 TG---T--T------C--A-GCTGAGA/
ψ U1.4 TCA TG---T--T------T-CA/

These three clones (cosD1, cosD21, and cosD8A) were subjected to further analysis, and the restriction maps are shown in Fig. 1.

The two U1 pseudogenes of cosD8A were found to be oriented divergently by mapping an 11-kb *Hin*dIII fragment of cosD8A subloned into pBR322 (plasmid pHin2). The maps of the cosD1 and cosD21 genes can be aligned with each other by using the *Nru*I sites which appear 1.4 kb upstream of the U1 coding sequence. However, for most other infrequently cutting enzymes, we did not find any obvious correspondence between the maps of cosD1 and cosD21 at distances greater than a few kb from the U1 coding region. Thus, these two U1 gene loci share a core of homology extending for a few kb on either side of the U1 sequence but have different restriction maps at larger distances from the U1 coding region. Moreover, these genes belong to different *Pst*I and *Pvu*II groups (data not shown). We conclude that cosD1 and cosD21 belong to different subfamilies of U1 genes. We will demonstrate below that at least one of these two loci, cosD1, is representative of a subfamily of U1 genes that is repeated many times in the human genome.

DNA sequence analysis of the U1 genes and pseudogenes. To confirm that the single U1 loci on cosmids cosD1 and cosD21 are true U1 genes and that the two U1 loci on cosD8A are class I pseudogenes, we sequenced the U1 coding regions and immediate flanking DNA. These sequences are compared (Fig. 2 and 3) to those of two previously characterized U1 genes, HSD4 and HU1-1, both of which are known to be transcriptionally active in vivo as well as in vitro (32a, 43, 54). Several class I pseudogenes whose sequences we reported previously (U1.1, U1.4, and U1.15) also are included in Fig. 2; all pseudogenes are indicated by a ψ.

As predicted by the genomic blotting data, the U1 loci of cosD1 and cosD21 appear to be true genes, by several criteria. First, the U1 RNA sequence is perfectly conserved in each. Second, the 5′ flanking regions of the U1 loci in both cosD1 and cosD21 are virtually identical to the perfectly conserved 5′ flanking regions of the true genes: the 5′ flanking sequences of cosD1, cosD21, HSD4, and HU1-1 diverge from each other by less than 1% (Fig. 2). Third, as in true U1 genes, the 3′ flanking sequences of cosD1 and cosD21 are well conserved at least 20 bp beyond the U1 coding region and thereafter exhibit occasional mismatches and insertions or deletions with repect to one another (Fig. 3). Most of the divergent nucleotides in the 3′ flanking sequences of cosD1 and cosD21 are common to a subset of genes and appear to divide the true U1 genes into groups; however, these groups do not seem to be correlated with the groups of U1 genes defined on a larger scale by restriction site mapping.

The genomic blotting data also predicted that cosD8A would contain two class I U1 pseudogenes, and this was confirmed by DNA sequence analysis (Fig. 2). The U1

```
                +10        +20        +30        +40        +50
HSD2   ACTTTCTGGAGTTTCAAAAACAGACCGTACGCCAAGGGTCATGTCTTTTTTCGT
HSD4   ----------------------------------------------------------
HSD5   ----------------------------------------------------------
HSD7   ----------------------------------G-T-----------------------
cosD1  ----------------------------------------A----A--------T--
HSD1   ------------------------T------T----------------C-TT-
HSD6   ------------------------T------T----C----------C--T--
HU1-1  --------------------GT----T------T---------A-----------
HSD3   --------------------GT----T-/
cosD21 --------------------GT----T------T---------A--------TT-

             +60        +70        +80        +90        +100
HSD2   ATT  GGTTTGTGTCTTAGTTG                TTAATCCTACAGTGGAG/
HSD4   ---  ----------------                ----------------/
HSD5   ---  ---G/
HSD7   --   ----/
cosD1  C--  ----------------                ---------------CA/
HSD1   G--TTT---------T--G----GCG CTTAAATG---------------T-/
HSD6   ---  --A---------G-------T----GG-------------------/
HU1-1  T--  -----------G-------------------------------------/
cosD21 ---  ---------T---------- ----------------------T-/
```

FIG. 3. Comparison of the 3′ flanking sequences of nine true U1 genes. The first 100 3′ flanking nucleotides of true U1 gene HSD2 are compared with the sequences of nine other true U1 genes. Sequence data is presented as in Fig. 2. Sequences for clones HSD1 through HSD7 are taken from Manser and Gesteland (35). The sequence of HU1-1 is from Murphy et al. (43).

coding regions of both the cosD8A-1 and cosD8A-2 loci exhibit point mutations and single base insertions or deletions but retain ~95% overall homology with U1 RNA. However, in comparison with true U1 genes, the flanking sequences average only ~85% homology upstream and ~65% homology downstream from the U1 coding region. This suggests that the class I U1 pseudogenes were at one time subjected to a selection pressure which was greater within the U1 coding region than within the flanking sequences. Curiously, sequences about 5 kb downstream from the cosD8A-2 coding sequence (Fig. 1, fragment I of pHin2) cross-hybridize to a probe that is specific for the 5′ flanking sequences of a true U1 gene (p5P2, see above). Fragment I could be (i) the orphaned remnant of a third U1 pseudogene that was deleted from the cosD8A locus, (ii) part of the 5′ flanking sequences of cosD8A-1 that were separated from cosD8A-1 by insertion of cosD8A-2, or (iii) part of the 5′ flanking sequences of cosD8A-2 that were orphaned by inversion of a DNA segment containing cosD8A-2. A more detailed DNA sequence analysis of the cosD8A locus would be required to distinguish among these possibilities.

Sequences 2 kb downstream from the ends of true U1 genes are almost as highly conserved as the 5′ flanking sequences. Manser and Gesteland (35) originally observed that the 5′ flanking sequences of true U1 genes are almost perfectly conserved as far as 2.6 kb upstream, whereas the first 100 bp of 3′ flanking sequences appeared to be extremely polymorphic (31, 35; Fig. 3). However, Htun et al. (19) subsequently demonstrated that seven true U1 genes have very similar restriction maps for at least 2.3 kb downstream. To determine whether downstream sequences diverge more

FIG. 2. Comparison of the sequences of four true U1 genes and five class I U1 pseudogenes. The top line shows the sequence of the true U1 gene HSD4 (35) supplemented with sequence data from our own laboratory (M. Mangin, V. Hoffarth, and A. M. Weiner, unpublished data); the sequences of three other true U1 genes and five class I U1 pseudogenes (marked by ψ) are compared with it. Dashes indicate matching bases, blank spaces indicate the absence of a corresponding base at that position, and a slash denotes the limits of sequence data. Nucleotide positions are numbered as follows. 5′ flanking sequences are given negative numbers, beginning with −1 for the base preceding the RNA cap site. U1 coding sequences are numbered 1 through 164, begining with the first base encoding U1 RNA. 3′ flanking sequences are given positive numbers, starting with +1 for the first base downstream of the U1 coding region. The 3′ flanking sequences of true genes other than HSD4 are not shown here; instead they are shown in Fig. 3. The two pseudogenes of the cosD8A clone are abbreviated here as D8A-1 and D8A-2 (for cosD8A-1 and cosD8A-2, respectively). The sequence of HU1-1 is from Murphy et al. (43) and Lund and Dahlberg (31). The sequences of U1.1, U1.4, and U1.15 are from Denison and Weiner (11). The sequence of cosD8A-2 very closely resembles the sequence of clone U1-8, reported by Monstein et al. (40, 41).

FIG. 4. Comparison of sequences 2 kb downstream from the U1 coding regions of two true U1 genes. The top line shows the sequence of a *Pst*I restriction fragment located 2 kb downstream from the U1 coding region of the true U1 gene HSD1. This sequence is compared to the sequence of part of a *Bgl*II fragment located 2 kb downstream from the true U1 gene HSD4. Sequence data are presented as in Fig. 2. *Pst*I and *Pvu*II sites are overlined for HSD1 and underlined for HSD4.

quickly than upstream sequences and whether the restriction map differences downstream are due to minor sequence heterogeneity or to gross sequence divergence, we compared sequences located 2 kb downstream from the ends of the U1 coding regions in two previously characterized U1 genes (HSD1 and HSD4 [35]). Surprisingly, these distant downstream sequences (Fig. 4) are much better conserved than the sequences immediately following the U1 genes (cf. Fig. 4 with 3). Over the entire 516-bp region (Fig. 5), these two loci diverge by only about 4% as a result of 19 scattered single-base mismatches (approximately 1 every 20 bp) and two small insertions or deletions. The same two loci diverge by at least 10% immediately 3′ to the U1 coding regions (Fig. 3).

We note that a single base insertion in HSD4 relative to HSD1 explains why a *Pst*I restriction site is present in HSD1 but absent from the corresponding position in HSD4. Anticipating our analysis of U1 gene organization below, we emphasize that as little as 4% sequence divergence can introduce significant restriction site polymorphism into two related sequences while still preserving many of the restriction sites (for instance, both HSD1 and HSD4 share the *Pvu*II site indicated in Fig. 5). In this light, the apparent differences in the maps of distinct U1 gene loci (35; corrected as in reference 19) do not contradict the evidence presented here that the flanking sequences of many U1 genes are extensively conserved. However, as mentioned above, the differences found between the restriction maps of the two U1 gene clones cosD1 and cosD21 at large distances from the U1 coding region may indicate the existence of divergent subfamilies of U1 genes. Therefore, we cannot rule out the possibility that some other U1 genes may diverge from the cosD1 subfamily (which probably includes HSD1 and HSD4, since they also belong to the 2.6-kb *Pst*I group) at distances further upstream or downstream from the coding region. Nevertheless, the homologous regions found 2 kb downstream from the coding regions of the two U1 genes HSD1 and HSD4 are almost as highly conserved as the virtually identical 5′ flanking regions of these genes. We will argue later that conservation of such extensive flanking sequences could result from coamplification of the U1 genes together with a large tract of flanking DNA.

Flanking sequence homology extends over 24 kb upstream and 20 kb downstream from the true U1 genes. The selected sequence comparisons described above (and in references 19 and 35) prove that many of the ~30 true U1 genes in the human genome share a high degree of flanking sequence homology at distances as far away as 2.6 kb upstream and 2.3 kb downstream from the U1 coding region. However, previous studies with U1 gene probes against whole-genome Southern blots (11, 35) suggested that the homology between many U1 gene loci may extend well beyond a few kb. To determine how far these homologous regions actually extend, we performed genomic blotting experiments based on the following rationale. If all U1 genes differ completely from one another at some distance from the U1 coding region, then a distant flanking fragment probe from one U1 locus should hybridize to a genomic blot at single-copy intensity. At the other extreme, if all U1 genes have identical flanking sequences at a distance from the coding region, the probe should hybridize to a single band on a genomic blot representing 30 copies of the probe sequence (i.e., the repetition frequency of the U1 genes). In an intermediate case, in which U1 genes are slightly divergent at this distance, we

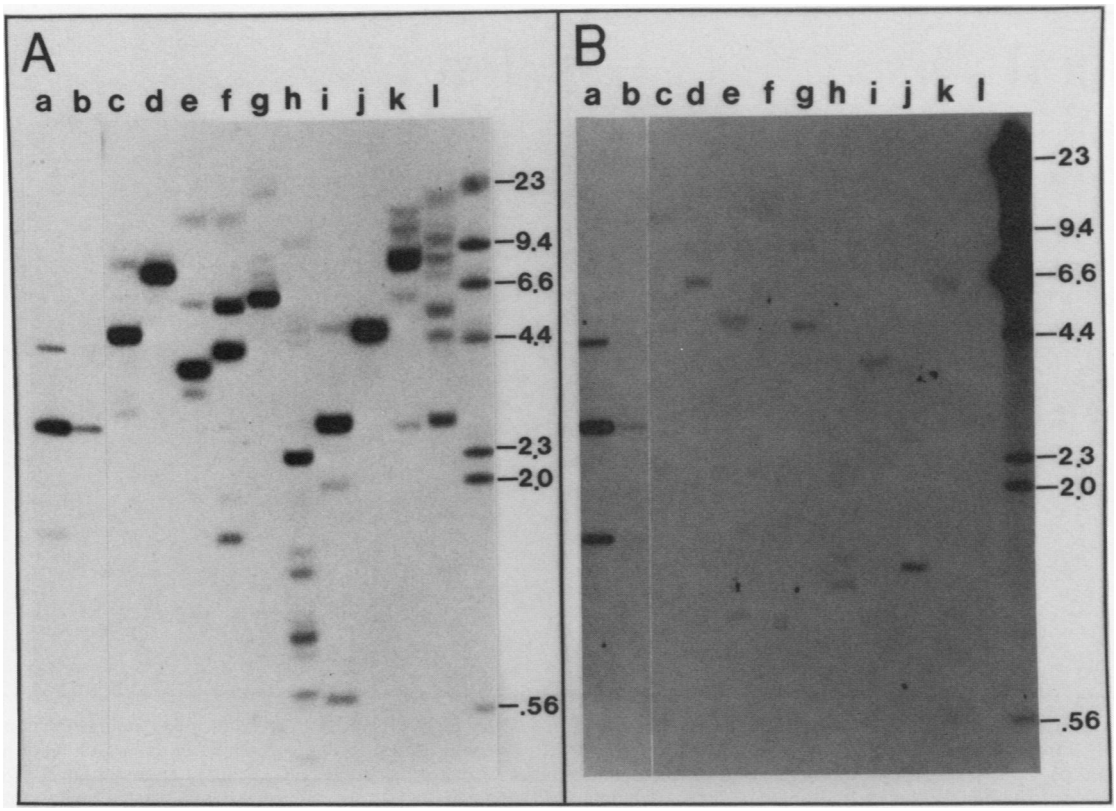


FIG. 5. Estimation of the copy number of sequences at least 20 kb to either side of the true U1 gene cosD1. Portions (10 μg) of the same sample of human placental DNA were digested by restriction enzymes and used to prepare two identical genomic blots, which were hybridized either to nick-translated pD1Pv (A) or to pD1Ec (B), as described in the text. Hybridizations and washes were performed at 42°C. The pD1Ec probe gave a weaker signal merely because it is much smaller than the pD1Pv probe (0.4 versus 3.8 kb). Enzymes used to digest placental DNA were *Bgl*II (lanes c), *Bam*HI (lanes d), *Eco*RI (lanes e), *Hin*dIII (lanes f), *Kpn*I (lanes g), *Msp*I (lanes h), *Pst*I (lanes i), *Pvu*II (lanes j), *Xba*I (lanes k), and *Xmn*I (lanes l). Lanes a and b contain 135 and 13.5 pg, respectively, of pBR322 DNA partially digested with *Acc*I; these lanes represent 10 and 1 copy equivalents of a 2.7-kb sequence (A), or of a 3.8-kb sequence (B), per haploid human genome.

would detect a few bands on the genomic blot indicative of restriction site polymorphism.

We chose the cosD1 cosmid to serve as a representative true U1 gene for this analysis. Since this clone contains more than 19 kb of flanking DNA to either side of the U1 coding region, we could identify and isolate restriction fragments that are relatively distant from the U1 sequence. The cosD1 subclones pD1Pv (insert size, 3.6 kb) and pD1Ec (insert size, 0.4 kb) contain fragments of DNA at the junctions of the cosD1 insert with the vector (Fig. 1 and above) and represent sequences lying 19 kb upstream (pD1Ec) or 19 kb downstream (pD1Pv) from the U1 coding region. Since the U1 genes may belong to subfamilies which differ from each other at large distances from the U1 coding region, we will use the term cosD1 family of genomic sequences to describe those genomic sequences which hybridize to either the pD1Pv or the pD1Ec probe.

Identical Southern blots containing human placental DNA digested with a variety of restriction enzymes and probed with pD1Pv or with pD1Ec (Fig. 5) show patterns of hybridization which in general reflect very good homology of sequences at a distance from U1 genes. For most of the enzymes used in the blots shown in Fig. 5, both probes hybridized to one very intense band and to as many as four faint bands, with a total copy number per lane of approximately 15 per haploid genome, as determined by densitometry (see above). Some lanes of these genomic blots contain a single intense band and no minor bands, such as the *Pvu*II lane for pD1Pv (Fig. 5A, lane j) and the *Xba*I lane for pD1Ec (Fig. 5B, lane k).

Since the single *Pvu*II band hybridizing with pD1Pv (Fig. 5A, lane j) migrates at 4.5 kb, the cosD1 map (Fig. 1) can now be extended by nearly 1 kb past the leftmost end of the insert, to the next *Pvu*II site. Likewise, since the major *Eco*RI band hybridizing with pD1Ec (Fig. 5B, lane e) is 5 kb long, the rightmost insert end of the cosD1 map can be extended by about 5 kb, to the next *Eco*RI site. Therefore, the cosD1 family of genomic sequences (which are identified here by hybridization to the distant flanking region subclones pD1Pv and pD1Ec) must be nearly identical over a distance of at least 24 kb upstream and at least 20 kb downstream from the U1 coding region. Since a second U1 gene is not found within this region, we can estimate an intergenic distance of more than 44 kb for U1 genes of the cosD1 subfamily.

We noted earlier, based on the marked differences in the maps of cosD1 and cosD21 (Fig. 1), that U1 genes may be divided into several subfamilies. If the distant flanking sequences of another subfamily of U1 genes (for instance, a cosD21 subfamily) were unrelated to those of the cosD1 subfamily, only the genes of the cosD1 subfamily would contribute to the observed hybridization signal. The copy numer of 15 found for distant flanking sequences of cosD1 (Fig. 5) suggests that U1 genes of the cosD1 subfamily might account for roughly half of the ~30 true U1 genes in the human genome (31). However, given the technical limita-

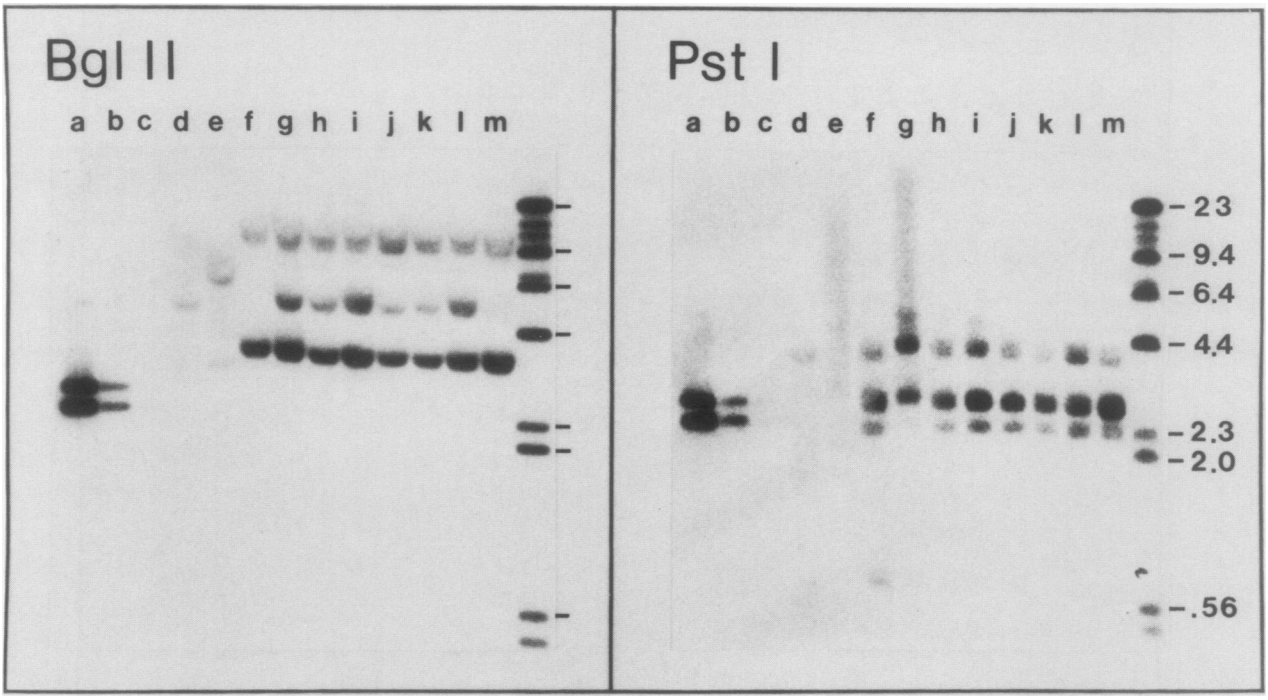


FIG. 6. Variation between individuals in the distribution of U1 genes among *Bgl*II and *Pst*I genomic blot bands. Portions (10 μg) of leukocyte DNA from various primates were digested with *Bgl*II or with *Pst*I and used to prepare genomic blots, which were hybridized to nick-translated p5P2, as described in the text. Hybridizations and washes were performed at 50°C. Primate genomic DNAs were from the following sources: chimpanzee (lanes d); orangutan (lanes e); humans of Oriental (lanes f) and Caucasian (lanes g to m) background. The DNA in lane g of the *Pst*I blot was incompletely digested. Lanes a, b, and c contain p5P2 DNA digested with *Hin*dII and *Pvu*II, at amounts equivalent to 15, 1.5, and 0.15 copies, respectively, of the p5P2 insert per haploid human genome.

tions of the gene quantitation experiment, we cannot rule out an error of a factor of 2 in copy number determination. Thus, the cosD1 subfamily of U1 genes could conceivably represent most of the ~30 true U1 genes in the human genome. We do not yet know whether cosD21 defines another subfamily of U1 genes or whether cosD21 is merely a single-copy locus. If the distant flanking sequences of the cosD21 locus are imperfectly homologous to those of the cosD1 subfamily, the cosD21-like genes may be represented in the fainter bands on the blots of Fig. 5. Alternatively, these faint bands could represent class I U1 pseudogenes.

Individual variation in the distribution of the true U1 genes among groups defined by restriction-fragment-length polymorphisms. To assess individual variation in the U1 multigene family, we obtained genomic DNA prepared from the blood of eight humans and two lower primates. Southern blots prepared from *Bgl*II or *Pst*I digests of these DNAs were probed with the U1 gene 5′ flanking probe p5P2. Since all known human U1 genes have a *Bgl*II site at −6, the intense 3.5-kb band of the *Bgl*II blot (Fig. 6, lanes f to m) indicates that most of the U1 genes of each human individual examined have a *Bgl*II site at −3.5 kb. Absence of the −3.5-kb *Bgl*II site apparently results in the appearance of bands of 5 or 9.5 kb, with the 9.5-kb band presumably indicating the absence from some U1 gene loci of two upstream *Bgl*II sites (at −3.5 and −5.0 kb). In agreement with previous work by others (31), we found that the relative intensities of these three bands vary from one individual to another, but the sizes of the bands remain constant, indicating an underlying similarity in the upstream flanking sequences of U1 genes in different individuals.

The *Pst*I blot (Fig. 6) depicts a similar situation for the downstream flanking sequences of U1 genes. Since the p5P2 probe overlaps the invariant *Pst*I site at −1.5 kb in human U1 genes, it hybridizes to fragments downstream from this site (i.e., containing the U1 coding region) as well as to those upstream. The two U1-containing *Pst*I fragments migrate at either 3.9 or 2.6 kb, probably determined by the presence or absence of a *Pst*I site about 1 kb downstream from the U1 coding sequence (35), and the upstream *Pst*I fragment is always 2.3 kb in length (Fig. 6, lanes f to m; the DNA sample in lane g is underdigested). When the *Pst*I patterns of different individuals were compared, the relative intensities of the 3.9- and 2.6-kb bands were found to vary; however, in no individual were new bands found which would indicate additional polymorphism.

The invariant sizes of the fragments hybridizing with p5P2 in different human genomic DNAs digested with either *Bgl*II or *Pst*I lend further support to our contention that human U1 genes have highly homologous flanking sequences both upstream and downstream from the U1 coding region. In addition, we note that the upstream flanks of the U1 genes in lower primates (chimpanzee and orangutan; lanes d and e, respectively, of Fig. 6) apparently retain sufficient similarity to the upstream flanks of human U1 genes to hybridize to the p5P2 probe at high stringency, but insufficient homology to preserve the human *Bgl*II and *Pst*I maps. No 5′ flanking homology to human U1 genes has been found in mouse, frog, chicken, or fruitfly genomes (31).

DISCUSSION

From studies in this paper and elsewhere (29), a clearer picture of the genomic organization of human U1 genes and class I U1 pseudogenes has emerged. By probing genomic

blots with 5′ and 3′ sequences far from the U1 coding region itself, we demonstrated that sequences flanking many of the ~30 true U1 genes are well conserved for at least 24 kb upstream and 20 kb downstream from the coding region. By restriction mapping and DNA sequence analysis of the cosD8A locus, we presented additional evidence that the class I U1 pseudogenes are more closely related to each other than to the true U1 genes (also see reference 11). It is shown elsewhere that true U1 genes appear to reside exclusively at the subtelomeric band 1p36 on the short arm of human chromosome 1 (confirming earlier work in references 30 and 44), whereas the class I U1 pseudogenes appear to reside exclusively in 1q12-q22 near the heterochromatin on the opposite side of the centromere. Admittedly, the details of genomic evolution are difficult to reconstruct by examination of a contemporary genome; however, we believe that the simplest interpretation of all of our data is that the human U1 genes are organized as a large, somewhat polymorphic tandem array and that the contemporary array was derived from a more ancient tandem array of U1 genes (now class I U1 pseudogenes) by gene amplification and transposition.

Are U1 genes organized in a tandem array? Our observations indicate that U1 genes belonging to the cosD1 subfamily share at least 44 kb of flanking DNA sequence homology. All true U1 genes were previously known to share a core of sequence homology extending about 2 kb on either side of the U1 coding region (19, 35). In addition, all true U1 genes are clustered in a single chromosomal band (44). From this we conclude that human U1 genes must be organized in one of two ways: either (i) the 44 kb of flanking homology surrounding many true U1 genes extends for more than 24 kb in the upstream direction and more than 20 kb in the downstream direction until reaching the adjacent U1 genes in a large and somewhat polymorphic tandem array, or (ii) the flanking homology ultimately degenerates on either side of each gene so that individual U1 genes exist as large islands of conserved sequence clustered in a sea of unique DNA. We argue below that the generation of a tandem array of U1 genes with extensive flanking homology can be readily accounted for by the experimentally documented characteristics of spontaneous gene amplification in mammalian cells. In addition, we show that the known characteristics of gene amplification provide a simple and straightforward explanation for the chromosomal mapping data for both true U1 genes and class I U1 pseudogenes (29). In contrast, the generation of a clustered but nontandem array of U1 genes would require a mechanism for localized duplication and transposition of a large unit of DNA exceeding 44 kb, an unprecedented possibility which we regard as implausible.

Tandem arrays of small nuclear RNA genes have been documented previously in humans as well as in other organisms. Researchers in our laboratory (60) and others (64) have demonstrated that the genes encoding human U2 RNA are organized as a tandem array of 10 to 20 essentially identical 6-kb repeat units, each containing a single U2 gene. The human U2 tandem array maps to a single site on chromosome 17 (28a). The vast majority of U1 genes in *Xenopus laevis* are arranged in a tandem array of 1.8-kb repeat units (32, 66), and the *Xenopus* U2 genes lie within a separate tandem array with an 830-bp repeat unit (38). The genes encoding the U1 and U2 RNAs of sea urchins also are arranged in separate tandem arrays with repeat units of 1.4 and 1.1 kb, respectively (6, 7). Although the U1 and U2 genes of mice (37, 45) and rats (58, 62) and the U1 genes of chickens (12) are not organized in a simple tandem array, at least some of the genes in each of these multigene families are clustered and share homologous flanking sequences. Thus, these genes might be part of an extremely polymorphic tandem array that has been severely scrambled by genetic exchange, as would appear to be the case for the human class I U1 pseudogenes on cosmid cosD8A (see below).

The repeat units of a tandem array are often nearly identical, as is the case for the U1 and U2 genes of sea urchins (7) and the U2 genes of *X. laevis* (38) and humans (60, 64). However, polymorphism between individual repeat units in a tandem array is also common and has been documented for the histone (17) and rRNA gene repeats (reviewed in reference 9) of *Drosophila*; for the 5S RNA genes (25), the rRNA genes (4), and both the major and minor types of U1 gene tandem repeats (32, 66) of *X. laevis*; and for the human rRNA genes (2, 27). Many tandemly repeated satellite sequences such as the primate alphoid families, and even simple sequence satellities, show polymorphism between repeat units (21, 53). Some polymorphisms leave the basic DNA sequence organization of the individual repeat units intact; these include variable numbers of an internally repetitious spacer sequence (4, 25) or mutations such as single base changes and small insertions or deletions (2, 27, 32). Other polymorphisms alter the basic organization of the original repeat unit; these include larger insertions or deletions (e.g., the *Drosophila* histone gene repeat [17]) or homologous but unequal recombination between sequences that are present more than once within each individual repeat unit (47, 48). In fact, polymorphisms consisting of variable numbers of a repetitious spacer sequence may arise from homologous but unequal recombination between fortuitous internal repeats within the basic repeat unit (55).

The human U1 genes HSD1 and HSD4 have slightly different restriction maps (19, 35); however, the presence of an additional *Pst*I site 2 kb downstream from the U1 sequence in HSD1 results from the insertion or deletion of only a single base in one gene relative to the other (Fig. 4). Thus, the restriction site polymorphisms found within the several kb immediately flanking the human U1 genes reflect a very minor level of DNA sequence heterogeneity between individual U1 repeat units and is entirely consistent with the existence of a tandem array of human U1 genes.

A model for the evolution of the U1 multigene family. How might a large, polymorphic tandem array of U1 genes have arisen? The idea that gene amplification could serve as a means both to establish (5) and to maintain (13) a homogeneous gene family was proposed many years ago and has been refined ever since (see reference 63 for a recent review); we present only the basic arguments here. In a multigene family, natural selection cannot act forcefully upon any individual gene since the preponderance of functional genes will partially protect the organism from the effects of almost all mutations (both favorable and unfavorable) except for those few that are dominant. Thus, the 30 individual U1 genes that constitute the U1 multigene family might have been expected to diverge from each other, but in fact U1 RNA (like most structural RNA species) is found to be homogeneous (10). One way to regenerate a uniform array of genes from an array that has diverged is through repeated cycles of DNA sequence amplification. Amplification of a well-preserved gene from the divergent array could effectively homogenize the gene family by overwhelming the old array with a new set of identical, fully functional genes. Freed from selection pressure, the old array might then degenerate further or be deleted.

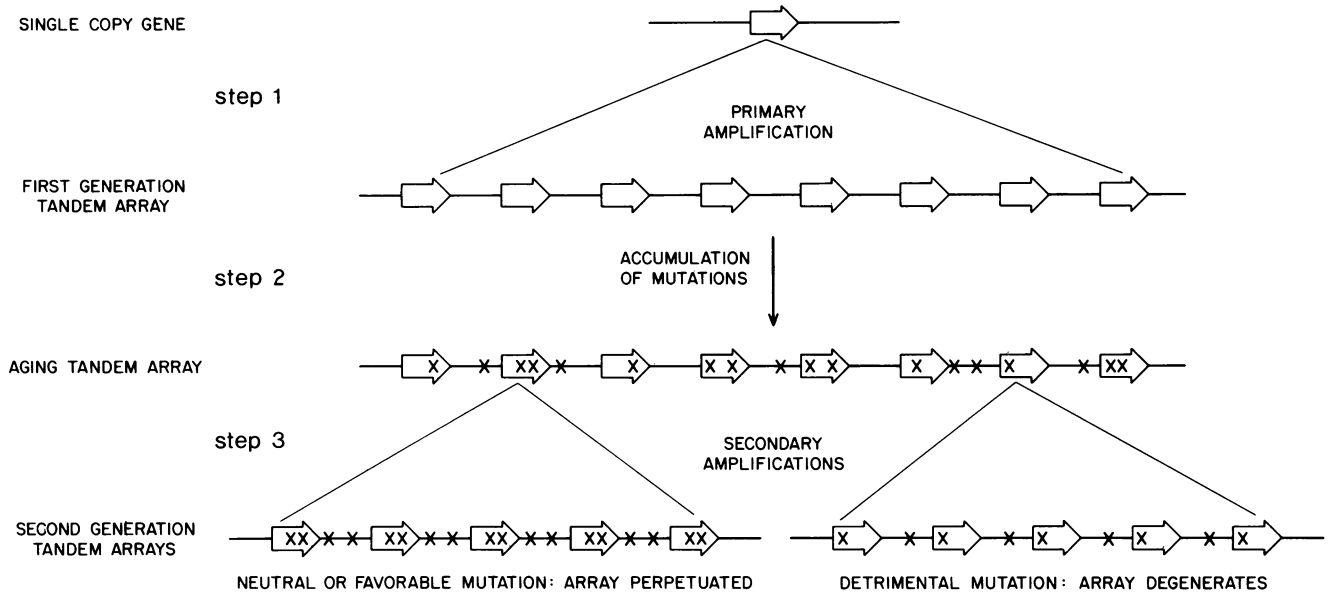


FIG. 7. Model for the evolution and maintenance of tandemly repeated gene families by multiple cycles of gene amplification. A primary amplification event (step 1) results in the tandem duplication of a single-copy gene (thick arrow). As the first-generation tandem array begins to age (step 2), individual genes accumulate favorable, unfavorable, and neutral mutations (represented by X's). In step 3, secondary amplification of a gene containing a favorable mutation confers a selective advantage, thereby fixing the array in the genome, whereas secondary amplification of a detrimentally mutated gene results in an array which might degenerate or be deleted.

Some of what is now understood about the process of gene amplification is directly relevant to the model for the evolution of the human U1 family of genes and class I pseudogenes (presented in detail below). Johnston et al. (22) provided experimental evidence that spontaneous DNA sequence amplifications occur at a remarkably high frequency, approaching 10^{-3} events per cell per generation per locus in cultured somatic cells. Moreover, Roberts and Axel (47) and Roberts et al. (48) showed that DNA sequence amplification in somatic cells occurs quickly and can produce as many as 50 copies of an amplified gene in just one generation. Since the contemporary human U1 multigene family consists of about 30 members, a single gene amplification event could in principle have enabled a new U1 array to be genetically dominant in the presence of an older degenerating array of U1 genes. In addition, a variety of studies have shown that the unit of amplified DNA in somatic cells can be as small as 40 kb and as large as 2,000 kb (48). Apparently, very large tracts of flanking DNA can be coamplified together with the selected gene without harming the cell; this helps to rationalize the conservation of at least 44 kb of DNA flanking the human U1 genes of the cosD1 subfamily. In fact, by taking into account the maximum resolution achievable by in situ chromosomal mapping (approximately 5,000 to 10,000 kb [50]), and the estimate of 30 U1 genes per human genome (31), the maximum length of the repeat unit for human U1 genes is about 300 kb.

Although gene amplification has been best studied in cultured somatic cells, where it is the most common mechanism by which cells acquire resistance to treatment with various cytotoxic drugs, gene amplification is not peculiar to immortal cell lines. The dihydrofolate reductase gene, for example, is amplified in tumor cells that survive chemotherapy with methotrexate (52), and the c-*myc* oncogene is amplified both in primary cultures and in cells taken directly from a patient with the HL-60 promyelocytic leukemia (46), and also in malignant neuroendocrine tumor cells from a human colon carcinoma (1). Finally, because the model for the evolution of the human U1 gene family requires that DNA sequence amplifications be heritable, we emphasize that there is no reason a priori why gene amplification in germ line cells of the living organism (i.e., mitotically dividing oogonia and spermatogonia, not primary oocytes and spermatocytes undergoing meiosis) should differ significantly from that observed experimentally for somatic cells.

In light of the known characteristics of DNA sequence amplification described above, we now present a detailed model for the evolution of the human U1 family of genes and class I pseudogenes (Fig. 7). Step 1 depicts the amplification of a primordial U1 gene to create a first-generation tandem array of identical U1 genes at 1q12-q22, currently the home of the class I U1 pseudogenes. As the array of functional U1 genes undergoes genetic drift (step 2), each gene acquires unique mutations. Some of the genes begin to degenerate into nonfunctional class I pseudogenes, thus accounting for the contemporary location of these pseudogenes at 1q12-q22.

In step 3, spontaneous secondary amplification events, sometimes accompanied by translocation, create new tandem arrays of U1 genes and class I pseudogenes. One of the secondary arrays of U1 genes is transposed to band 1p36 on the opposite side of the centromere; this array confers a selective advantage on the organism and is preserved as the contemporary family of U1 genes. Secondary arrays of defective U1 genes which confer a selective disadvantage would degenerate or be deleted; arrays of selectively neutral sequences might be retained as class I pseudogenes. Secondary amplifications of the kind we invoked are known to occur at even higher frequency than that of the primary amplification event (22). In addition, although DNA sequence amplification can produce an array of genes at the site of the resident chromosomal gene copy (52), the amplified DNA is often transposed to a completely new chromosomal site (57), perhaps using extrachromosomal amplified

DNA or "double minute" chromosomes as intermediates (1, 52).

Clustering and recombination of class I U1 pseudogenes. Class I U1 pseudogenes are far more divergent from each other than are the true U1 genes and therefore appear to be more ancient than the true U1 genes. In another paper (29), these class I U1 pseudogenes are mapped to chromosome 1, bands q12-q22, and we propose that class I U1 pseudogenes are the decaying ancestors of an older tandem array of human U1 genes, one member of which was amplified and transposed to 1p36 in the course of evolution.

Comparison of the sequences flanking U1 genes and class I pseudogenes (Fig. 2) supports the idea that separate arrays of genes and pseudogenes were created by distinct events. In both the 5′ and 3′ flanking regions, all pseudogenes appear more closely related to each other than to the true U1 genes. Some differences between true U1 genes and the class I pseudogenes are common to all the pseudogenes that have been sequenced (Fig. 2); these differences are mostly single base changes (e.g., the G at position −16) and small deletions (e.g., at positions −80 and −88). Thus, the class I U1 pseudogenes appear to represent the descendants of a primordial U1 gene that was significantly different from the gene that was amplified, at a later time in evolution, to create the modern functional U1 gene family. Other differences between true U1 genes and class I pseudogenes are found in only a subset of the pseudogenes (e.g., the insertion of G residues at position −75 and the deletion of 11 bp at position −23). In particular, pseudogenes U1.15 and cosD8A-1 are closely related, as are pseudogenes U1.1 and cosD8A-2. The existence of polymorphisms that are shared by different class I U1 pseudogenes indicates that these pseudogenes have exchanged sequence information by gene conversion or have been subject to more than one round of gene amplification.

According to the gene amplification model (Fig. 7), class I pseudogenes in the aging U1 tandem array should reside on repeat units which are equal to or larger than the size of the U1 gene repeat units in the second-generation tandem array (unless the primordial gene was located at one end of the original tandem array). Otherwise, more than one primordial U1 gene would be found within the contemporary repeat unit. In fact, with the exception of the cosD8A locus, we observe that class I U1 pseudogenes are usually separated from each other by a distance exceeding 35 kb. Although we have not analyzed the genomic copy number of sequences distantly flanking a class I pseudogene, eight class I U1 pseudogenes isolated from the human genome in cosmid vectors contain a single U1 sequence within approximately 35 kb of genomic DNA (T. Manser, unpublished data), and the three class I U1 pseudogenes (U1.1, U1.4, and U1.15) characterized previously from a human lambda vector library each contain a single U1 sequence within the 15-kb insert (11). In contrast to these other characterized class I U1 pseudogenes, only the cosD8A clone (Fig. 1) contains two pseudogenes 9 kb apart, as well as a region of U1 5′ flanking sequence homology (fragment I) located just 5 kb downstream from one of the two pseudogenes. Thus, the cosD8A locus does not appear to represent a majority of class I U1 pseudogenes; moreover, the divergent orientations of the two cosD8A pseudogenes (Fig. 1) and the presence of an orphaned 5′ flanking sequence suggest that the cosD8A locus has been subject to multiple recombination events.

Recombination between the true U1 genes. The extraordianry conservation of flanking sequences in human U1 genes of the cosD1 subfamily (over 44 kb) and human U2 genes (an essentially perfect 6-kb repeat unit) strongly suggests to us that these particular tandem arrays, once established by gene amplification, have been perpetuated by recombination events that can homogenize the array in situ without significant DNA rearrangement. This view is directly supported by the observation that U1 genes can be divided into groups which share certain DNA sequence polymorphisms (Fig. 3; 35). Such groups must reflect partial homogenization of the true U1 genes by gene conversion, by homologous but unequal recombination, or by reamplification of individual repeat units. U1 gene homogenization by recombination is indirectly supported by the observation that the class I U1 pseudogenes (which presumably represent an aging array of U1 genes) can also be divided into groups based on shared polymorphisms in both the coding and flanking regions (Fig. 2; see above). Finally, the human rRNA genes provide evidence for extensive recombination between the repeat units of another, unrelated tandem array in the human genome (2, 27).

The human U2 genes are found in an essentially homogeneous tandem array, unlike the extended family of 30 U1 genes and perhaps hundreds of class I pseudogenes; moreover, only a few class I U2 pseudogenes exist (16, 60, 64). Why is the human U2 family so much cleaner than the human U1 family? One possible explanation is that the U2 genes are more efficiently homogenized in situ than are the U1 genes and therefore may not require repeated cycles of gene amplification to maintain homogeneity. For example, both unequal sister chromatid exchange and gene conversion require the alignment of nonequivalent genes within the tandem array, and the frequency of such events might be increased by the much shorter length of the U2 repeat unit (6 kb) in comparison to the U1 repeat (more than 44 kb). Alternatively, the U2 repeat unit might contain a recombinogenic element that is functionally similar to the chi octanucleotide sequence of *E. coli* (21, 55a) or to the HOT1 locus found within each repeat unit of the tandemly arrayed rDNA in *Saccharomyces cerevisiae* (24). Perhaps the absence of such a recombinogenic element in the vicinity of the cellular dihydrofolate reductase gene (or the presence of too many such elements) could account for the recent observation of Federspiel et al. (14) that sequences flanking the amplified dihydrofolate reductase gene in several methotrexate-resistant mouse cell lines are subject to extensive and ongoing DNA sequence rearrangement.

Because it is difficult to deduce the mechanisms of genomic evolution by studying the biological record, we are currently using the techniques of somatic cell genetics to create cell lines with artificial arrays of tandemly repeated U1 and U2 genes linked to selectable markers. This may enable us to apply our theories to a direct experimental test.

ACKNOWLEDGMENTS

We thank Catherine Joyce for generous supplies of Klenow polymerase, Elizabeth Zimmer and Allan Wilson for gifts of human and simian genomic DNA, Kimberly Mowry for preparing samples of human genomic DNA, and Marguerite Mangin and Vernita Hoffarth for providing sequence data for HSD4. We are grateful to Vernita Hoffarth for expert technical assistance with M13 subcloning and DNA sequencing and to R. F. Gesteland for providing laboratory facilities and advice on the initial construction and isolation of the U1 cosmid clones. We thank Manuel Ares, Norman Arnheim, Robin Chadwick, Uta Francke, and Valerie Lindgren for helpful suggestions in the course of the work.

This work was supported by Public Health Service grants

GM-31073 and GM-31335 from the National Institutes of Health and by grant PCM83-15602 from the National Science Foundation.

LITERATURE CITED

1. **Alitalo, K., M. Schwab, D. C. Lin, H. E. Varmus, and J. M. Bishop.** 1983. Homogeneously staining chromosomal regions contain amplified copies of an abundantly expressed cellular oncogene (c-*myc*) in malignant neuroendocrine cells from a human colon carcinoma. Proc. Natl. Acad. Sci. U.S.A. **80:** 1707–1711.
2. **Arnheim, N.** 1983. Concerted evolution of multigene families, p. 38–61. *In* M. Nei and R. K. Koehn (ed.), Evolution of genes and proteins. Sinauer Associates, Inc., Sunderland, Mass.
3. **Bernstein, L. B., S. M. Mount, and A. M. Weiner.** 1983. Pseudogenes for human samll nuclear RNA U3 appear to arise by integration of self-primed reverse transcripts of the RNA into new chromosomal sites. Cell **32:**461–472.
4. **Boseley, P. T., T. Moss, M. Mächler, R. Portmann, and M. Birnstiel.** 1979. Sequence organization of the spacer DNA in a ribosomal gene unit of *Xenopus laevis*. Cell **17:**19–31.
5. **Britten, R. J., and D. E. Kohne.** 1968. Repeated sequences in DNA. Science **161:**529–540.
6. **Brown, D. T., G. F. Morris, N. Chodchoy, C. Sprecher, and W. F. Marzluff.** 1985. Structure of the sea urchin U1 RNA repeat. Nucleic Acids Res. **13:**537–556.
7. **Card, C. O., G. F. Morris, D. T. Brown, and W. F. Marzluff.** 1982. Sea urchin small nuclear RNA genes are organized in distinct tandemly repeating units. Nucleic Acids Res. **10:** 7677–7688.
8. **Cleveland, D. W.** 1983. The tubulins: from DNA to RNA to protein and back again. Cell **34:**330–332.
9. **de Cicco, D. V., and D. M. Glover.** 1983. Amplification of rDNA and type I sequences in *Drosophila* males deficient in rDNA. Cell **32:**1217–1225.
10. **Denison, R. A., S. W. Van Arsdell, L. B. Bernstein, and A. M. Weiner.** 1981. Abundant pseudogenes for small nuclear RNAs are dispersed in the human genome. Proc. Natl. Acad. Sci. U.S.A. **78:**810–814.
11. **Denison, R. A., and A. M. Weiner.** 1982. Human U1 RNA pseudogenes may be generated by both DNA- and RNA-mediated mechanisms. Mol. Cell. Biol. **2:**815–828.
12. **Earley, J. M., III, K. A. Roebuck, and W. E. Stumph.** 1984. Three linked chicken U1 RNA genes have limited flanking DNA sequences homologies that reveal potential regulatory signals. Nucleic Acids Res. **12:**7411–7421.
13. **Edelman, G. M., and J. A. Gally.** 1970. Arrangement and evolution of eukaryotic genes, p. 962–972. *In* F. O. Schmitt (ed.), Neurosciences: second study program. Rockefeller University Press, New York.
14. **Federspiel, N. A., S. M. Beverley, J. W. Schilling, and R. T. Schimke.** 1984. Novel DNA rearrangements are associated with dihydrofolate reductase gene amplification. J. Biol. Chem. **259:**9127–9140.
15. **Fyrberg, E. H., J. W. Mahaffey, B. J. Bond, and N. Davidson.** 1983. Transcripts of the six *Drosophila* actin genes accumulate in a stage- and tissue-specific manner. Cell **33:**115–123.
16. **Hammarström, K., G. Westin, C. Bark, J. Zabielski, and U. Pettersson.** 1984. Genes and pseudogenes for human U2 RNA: implications for the mechanism of pseudogene formation. J. Mol. Biol. **179:**157–169.
17. **Hentschel, C. C., and M. L. Birnstiel.** 1981. The organization and expression of histone gene families. Cell **25:**301–313.
18. **Hohn, B., and J. Collins.** 1980. A small cosmid for efficient cloning of large DNA fragments. Gene **11:**291–298.
19. **Htun, H., E. Lund, and J. E. Dahlberg.** 1984. Human U1 RNA genes contain an unusally sensitive nuclease S1 cleavage site within the conserved 3′ flanking region. Proc. Natl. Acad. Sci. U.S.A. **81:**7288–7292.
20. **Ish-Horowicz, D., and J. F. Burke.** 1981. Rapid and efficient cosmid cloning. Nucleic Acids Res. **9:**2989–2998.
21. **Jeffreys, A. J., V. Wilson, and S. L. Thein.** 1985. Hypervariable 'minisatellite' regions in human DNA. Nature (London) **314:**67–73.
22. **Johnston, R. N., S. M. Beverley, and R. T. Schimke.** 1983. Rapid spontaneous dihydrofolate reductase gene amplification shown by fluorescence-activated cell sorting. Proc. Natl. Acad. Sci. U.S.A. **80:**3711–3715.
23. **Kan, Y. W., A. M. Dozy, R. Trecartin, and D. Todd.** 1977. Identification of a nondeletion defect in α-thalassemia. N. Engl. J. Med. **297:**1081–1084.
24. **Keil, R. L., and G. S. Roeder.** 1984. *Cis*-acting, recombination-stimulating activity in a fragment of the ribosomal DNA of *S. cerevisiae*. Cell **39:**377–386.
25. **Korn, L. J., and D. F. Bogenhagen.** 1982. Organization and transcription of *Xenopus* 5S ribosomal RNA genes, p. 1–27. *In* H. Busch and L. Rothblum (ed.), The cell nucleus, vol. 12, rDNA part C. Academic Press, Inc., New York.
26. **Krämer, A., W. Keller, B. Appel, and R. Lührmann.** 1984. The 5′ terminus of the RNA moiety of U1 small nuclear ribonucleoprotein particles is required for the splicing of messenger RNA precursors. Cell **38:**299–307.
27. **Krystal, M., P. D'Eustachio, F. H. Ruddle, and N. Arnheim.** 1981. Human nucleolus organizers on nonhomologous chromosomes can share the same ribosomal gene variants. Proc. Natl. Acad. Sci. U.S.A. **78:**5744–5748.
28. **Lerner, M. R., J. A. Boyle, S. M. Mount, S. L. Wolin, and J. A. Steitz.** 1980. Are snRNPs involved in splicing? Nature (London) **283:**220–224.

28a. **Lindgren, V., M. Ares, Jr., A. M. Weiner, and U. Francke.** 1985. Human genes for U2 small nuclear RNA map to a major adenovirus 12 modification site on chromosome 17. Nature (London) **314:**115–116.

29. **Lindgren, V., L. B. Bernstein, A. M. Weiner, and U. Francke.** 1985. Human U1 small nuclear RNA pseudogenes do not map to the site of the U1 genes in 1p36 but are clustered in 1q12-q22. Mol. Cell. Biol. **5:**2172–2180.
30. **Lund, E., C. Bostock, M. Robertson, S. Christie, J. L. Mitchen, and J. E. Dahlberg.** 1983. U1 small nuclear RNA genes are located on human chromosome 1 and are expressed in mouse-human hybrid cells. Mol. Cell. Biol. **3:**2211–2220.
31. **Lund, E., and J. E. Dahlberg.** 1984. True genes for human U1 small nuclear RNA: copy number, polymorphism, and methylation. J. Biol. Chem. **259:**2013–2021.
32. **Lund, E., J. E. Dahlberg, and D. J. Forbes.** 1984. The two embryonic U1 small nuclear RNAs of *Xenopus laevis* are encoded by a major family of tandemly repeated genes. Mol. Cell. Biol. **4:**2580–2586.

32a. **Mangin, M., M. Ares, Jr., and A. M. Weiner.** 1985. U1 small nuclear RNA genes are subject to dosage compensation in mouse cells. Science **229:**272–275.

33. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
34. **Manser, T., and R. F. Gesteland.** 1981. Characterization of small nuclear RNA U1 gene candidates and pseudogenes from the human genome. J. Mol. Appl. Genet. **1:**117–125.
35. **Manser, T., and R. F. Gesteland.** 1982. Human U1 loci: genes for human U1 RNA have dramatically similar genomic environments. Cell **29:**257–264.
36. **Mao, J., B. Appel, J. Schaack, S. Sharp, H. Yamada, and D. Söll.** 1982. The 5S RNA genes of *Schizosaccharomyces pombe*. Nucleic Acids Res. **10:**487–500.
37. **Marzluff, W. F., D. T. Brown, S. Lobo, and S.-S. Wang.** 1983. Isolation and characterization of two linked mouse U1b small nuclear RNA genes. Nucleic Acids Res. **11:**6255–6270.
38. **Mattaj, I. W., and R. Zeller.** 1984. *Xenopus laevis* U2 snRNA genes: tandemly repeated transcription units sharing 5′ and 3′ flanking homology with other RNA polymerase II transcribed genes. EMBO J. **2:**1883–1891.
39. **Messing, J., and J. Vieira.** 1982. A new pair of M13 vectors for selecting either DNA strand of double-digested restriction fragments. Gene **19:**269–276.
40. **Monstein, H.-J., K. Hammarström, G. Westin, J. Zabielski, L. Philipson, and U. Pettersson.** 1983. Loci for human U1 RNA:

structural and evolutionary implications. J. Mol. Biol. **167:** 245–257.
41. **Monstein, H.-J., G. Westin, L. Philipson, and U. Pettersson.** 1982. A candidate gene for human U1 RNA. EMBO J. **1:**133–137.
42. **Mount, S. M., I. Pettersson, M. Hinterberger, A. Karmas, and J. A. Steitz.** 1983. The U1 small nuclear RNA-protein complex selectively binds a 5′ splice site *in vitro*. Cell **33:**509–518.
43. **Murphy, J. T., R. R. Burgess, J. E. Dahlberg, and E. Lund.** 1982. Transcription of a gene for human U1 small nuclear RNA. Cell **29:**265–274.
44. **Naylor, S. L., B. U. Zabel, T. Manser, R. Gesteland, and A. Y. Sakaguchi.** 1984. Localization of human U1 small nuclear RNA genes to band p36.3 of chromosome 1 by *in situ* hybridization. Somatic Cell Mol. Genet. **10:**307–313.
45. **Nojima, H., and R. D. Kornberg.** 1983. Genes and pseudogenes for mouse U1 and U2 small nuclear RNAs. J. Biol. Chem. **258:**8151–8155.
46. **Nowell, P., J. Finan, R. Dalla Favera, R. C. Gallo, A. ar-Rushdi, H. Romanczuk, J. R. Selden, B. S. Emanuel, G. Rovera, and C. M. Croce.** 1983. Association of amplified oncogene c-*myc* with abnormally banded chromosome 8 in a human leukaemia cell line. Nature (London) **306:**494–497.
47. **Roberts, J. M., and R. Axel.** 1982. Gene amplification and gene correction in somatic cells. Cell **29:**109–119.
48. **Roberts, J. M., L. B. Buck, and R. Axel.** 1983. A structure for amplified DNA. Cell **33:**53–63.
49. **Rogers, J., and R. Wall.** 1980. A mechanism for RNA splicing. Proc. Natl. Acad. Sci. U.S.A. **77:**1877–1879.
50. **Ruddle, F. H.** 1981. A new era in mammalian gene mapping: somatic cell genetics and recombinant DNA methodologies. Nature (London) **294:**115–120.
51. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. **74:**5463–5467.
52. **Schimke, R. T.** 1984. Gene amplification in cultured cells. Cell **37:**705–713.
53. **Singer, M.** 1982. Highly repeated sequences in mammalian genomes. Int. Rev. Cytol. **76:**67–112.
54. **Skuzeski, J. M., E. Lund, J. T. Murphy, T. H. Steinberg, R. R. Burgess, and J. E. Dahlberg.** 1984. Synthesis of human U1 RNA. II. Identification of two regions of the promoter essential for transcription initiation at position +1. J. Biol. Chem. **259:**8345–8352.
55. **Smith, G. P.** 1976. Evolution of repeated DNA sequences by unequal crossover. Repetitiousness as the natural state of DNA whose sequence is not controlled by selective forces. Science **191:**528–535.
55a. **Smith, G. R.** 1983. Chi hotspots of generalized recombination. Cell **34:**709–710.
56. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98:**503–517.
57. **Stallings, R. L., A. C. Munk, J. L. Longmire, C. E. Hildebrand, and B. D. Crawford.** 1984. Assignment of genes encoding metallothioneins I and II to Chinese hamster chromosome 3: evidence for the role of chromosome rearrangement in gene amplification. Mol. Cell. Biol. **4:**2932–2936.
58. **Tani, T., N. Watanabe-Nagasu, N. Okada, and Y. Ohshima.** 1983. Molecular cloning and characterization of a gene for rat U2 small nuclear RNA. J. Mol. Biol. **168:**579–594.
59. **Van Arsdell, S. W., R. A. Denison, L. B. Bernstein, A. M. Weiner, T. Manser, and R. F. Gesteland.** 1981. Direct repeats flank three small nuclear RNA pseudogenes in the human genome. Cell **26:**11–17.
60. **Van Arsdell, S. W., and A. M. Weiner.** 1984. Human genes for U2 small nuclear RNA are tandemly repeated. Mol. Cell. Biol. **4:**492–499.
61. **Van Arsdell, S. W., and A. M. Weiner.** 1984. Pseudogenes for human U2 small nuclear RNA do not have a fixed site of 3′ truncation. Nucleic Acids Res. **12:**1463–1471.
62. **Watanabe-Nagasu, N., Y. Itoh, T. Tani, K. Okano, N. Koga, N. Okada, and Y. Ohshima.** 1983. Structural analysis of gene loci for rat U1 small nuclear RNA. Nucleic Acids Res. **11:** 1791–1801.
63. **Weiner, A. M., and R. A. Denison.** 1983. Either gene amplification or gene conversion may maintain the homogeneity of the multigene family encoding human U1 small nuclear RNA. Cold Spring Harbor Symp. Quant. Biol. **47:**1141–1149.
64. **Westin, G., J. Zabielski, K. Hammarström, H.-J. Monstein, C. Bark, and U. Pettersson.** 1984. Clustered genes for human U2 RNA. Proc. Natl. Acad. Sci. U.S.A. **81:**3811–3815.
65. **Yang, V. W., M. R. Lerner, J. A. Steitz, and S. J. Flint.** 1981. A small nuclear ribonucleoprotein is required for splicing of adenoviral early RNA sequences. Proc. Natl. Acad. Sci. U.S.A. **78:**1371–1375.
66. **Zeller, R., M.-T. Carri, I. W. Mattaj, and E. M. DeRobertis.** 1984. *Xenopus laevis* U1 snRNA genes: characterisation of transcriptionally active genes reveals major and minor repeated gene families. EMBO J. **3:**1075–1081.