

# Direct isolation of human transcribed sequences from yeast artificial chromosomes through the application of RNA fingerprinting

IVAN H. STILL, PAULINE VINCE, AND JOHN K. COWELL\*

Department of Neurosciences NC30, The Lerner Research Institute, Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195

Communicated by George R. Stark, Cleveland Clinic Foundation, Cleveland, OH, July 10, 1997 (received for review May 30, 1997)

**ABSTRACT** The identification of cDNA clones from genomic regions known to contain human genes is usually the rate-limiting factor in positional cloning strategies. We demonstrate here that human genes present on yeast artificial chromosomes (YACs) are transcribed in yeast host cells. We have used the arbitrarily primed RNA (RAP) fingerprinting method to identify human-specific, transcribed sequences from YACs located in the 13q12 chromosome region. By comparing the RAP fingerprints generated using defined, arbitrary primers from various fragmented YACs, megaYACs, and host yeast, we were able to identify and map 20 products transcribed from the human YAC inserts. This method, therefore, permits the simultaneous isolation and mapping of novel expressed sequences directly from whole YACs.

The generation of an extensive, integrated, physical and genetic map of the human genome (1) will provide a framework in which to refine the location of disease genes identified either through genetic linkage or through their association with a defined cytogenetic anomaly. It should, therefore, be possible to define the position of a disease gene relative to these maps, and therefore refine its location to a region typically spanning 0.5–1 Mb. However, the rate-limiting step in the identification of a particular disease gene is the isolation and mapping of candidate transcription units within the area of interest. In parallel to the physical and genetic mapping projects currently underway, a second mass screening project has been initiated to generate a transcription map of the human genome (2, 3). As a result of this, large numbers of novel cDNAs (expressed sequence tags, ESTs) are being deposited into the publicly accessible databases. Unfortunately, because of the considerable effort involved, their assignment to specific chromosomes has not kept pace. In addition, to be useful as candidate genes, these ESTs require a subregional localization along the chromosome. This is typically accomplished using specialized somatic cell hybrid mapping panels (4–6) or through the generation of radiation reduced hybrids (7). Due to the random and time-consuming nature of this work, it is difficult to apply this technology directly to the isolation of a particular disease gene known to be located within a particular yeast artificial chromosome (YAC).

Because YACs offer the advantage of maintaining the largest inserts, the large-scale physical maps currently under construction have used genomic DNA cloned in YAC vectors. A number of methods have been devised to isolate genes from these large regions of DNA. However, many of these methods require subcloning of the YACs into smaller vectors prior to analysis. The requirement of subcloning YACs for subsequent characterization through analysis of rare cutter sites (8), through the identification

of conserved sequences (9), or by exon trapping (10, 11) is time consuming and can have a variable degree of success. Transcriptional-based methods of isolating candidate genes, either by screening cDNA libraries directly with genomic probes (12, 13) or through hybrid selection (14, 15), are dependent upon the expression of the gene within the cDNA pool used.

The basic transcriptional machinery is thought to be conserved between higher and lower eukaryotes (16). Functional assays have shown that some components of the transcriptional apparatus can be interchanged between yeast and mammals. For instance, yeast TFIID can functionally interact with mammalian TATA-binding protein-associated factors and initiate transcription from polymerase II promoters *in vitro* (17, 18). The converse is also true (19), and in addition human TFIID is capable of complementing defects in TATA-dependent polymerase II transcription *in vivo* (20). The structure of TATA-dependent promoters differs between higher eukaryotes and yeast, particularly in regard to the spacing of the TATA-box and subsequent initiation of transcription (21). However, despite these differences, the yeast transcriptional machinery is able to recognize the TATA box in typical mammalian TATA-dependent promoters and initiate transcription *in vitro* (22, 23). Therefore, freed of transcriptional repressors that govern tissue-specific expression in mammalian systems, we would predict that human polymerase II promoters at least would be constitutively active in yeast. We demonstrate here that human genes are expressed from whole YACs in the yeast host cell, and also that novel human expressed sequences can be isolated directly from cultures of growing YACs through the application of arbitrarily primed RNA (RAP) fingerprinting (24). By comparing the RAP fingerprints of multiple overlapping YACs, or deletion constructs generated through YAC fragmentation (25), it is possible to simultaneously isolate and sublocalize these sequences within the YACs of interest.

## METHODS

**Isolation of Total RNA from Yeast.** Synthetic dextrose (SD) broth (5 ml) supplemented with 20  $\mu$ g/ml adenine, 20  $\mu$ g/ml tryptophan, 20  $\mu$ g/ml histidine, 20  $\mu$ g/ml isoleucine, and 20  $\mu$ g/ml uracil was inoculated with a single YAC colony and grown overnight at 30°C. Complete SD medium (15 ml), prewarmed to 30°C, was then added, and the culture grown for a further 4 h. Cells were harvested and total RNA was prepared according to the hot phenol method described by Wise (26). The integrity of the RNA was assessed by electrophoresis through 1% (wt/vol) agarose in 1 $\times$  TAE in the presence of 0.1% (wt/vol) SDS. To remove residual DNA contamination, the RNA was treated with RNase-free DNase

Abbreviations: YAC, yeast artificial chromosome; RAP, arbitrarily primed RNA; EST, expressed sequence tag; RT-PCR, reverse transcription-PCR.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. U82301–U82321). \*To whom reprint requests should be addressed at: Department of Neurosciences NC30, The Lerner Research Institute, Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1997 by The National Academy of Sciences 0027-8424/97/9410373-6\$2.00/0 PNAS is available online at <http://www.pnas.org>.

(GIBCO/BRL) according to manufacturer's instruction prior to reverse transcription.

**Reverse Transcription-PCR (RT-PCR).** Total RNA (500 ng) was reverse transcribed with random hexanucleotides (50 ng/ $\mu$ l) with Superscript (GIBCO/BRL) according to manufacturer's instructions. RT-PCR analysis of ESTs mapped to the 943E4 YAC was performed using 1  $\mu$ l of the RT reaction, 0.4  $\mu$ M primers, 2.0 mM MgCl<sub>2</sub>, 50 mM KCl, 20 mM Tris-HCl (pH 8.8 at 25°C), 250  $\mu$ M each dNTP, and 1 unit of *Taq* DNA polymerase. Samples were denatured at 94°C for 3 min followed by 40 cycles of denaturation at 94°C for 1 min, annealing at 50°C for 1 min, and extension at 72°C for 1 min. Reactions were carried out in an Ericomp (San Diego) thermocycler. Products were visualized on 3.5% agarose gels. RAP fingerprinting assays were carried out using 1  $\mu$ l of the RT reaction, 0.4  $\mu$ M arbitrary primer, 4.0 mM MgCl<sub>2</sub>, 50 mM KCl, 20 mM Tris-HCl (pH 8.8 at 25°C), 250  $\mu$ M each dNTP, and 1 unit of *Taq* DNA polymerase in the presence of 5  $\mu$ Ci [ $\alpha$ -<sup>32</sup>P]dCTP (1 Ci = 37 GBq). To promote annealing of the primer to sequences that match the six to eight bases at the 3' end of the primer, one low stringency PCR cycle of denaturation at 94°C for 5 min, annealing at 40°C for 5 min, and extension at 72°C for 5 min was performed, followed by 40 cycles of denaturation at 94°C for 1 min, annealing at 50°C for 1 min, and extension at 72°C for 1 min. Reactions were carried out in an Ericomp thermocycler.

**Visualization of RNA Fingerprint.** A total of 3  $\mu$ l of formamide stop mix (95% formamide/0.25% bromophenol blue/0.25% xylene cyanol) was added to 2  $\mu$ l of each RT-PCR. The sample was denatured at 85°C for 5 min and run on a 4% sequencing gel. The gel was dried under vacuum and exposed to Kodak XAR-5 film overnight at room temperature.

**Cloning and Analysis of Differential Display Products.** Products "differentially displayed" in RNA pools from YAC cultures, compared with the host yeast, were excised from the dried down polyacrylamide gel and recovered by the crush and soak method (27). Fragments were then reamplified with the same arbitrary primers. The localization of the differential display product was confirmed and refined by Southern blot hybridization to DNA prepared from fragmented YACs and the megaYACs 943E4 and 775F9. The PCR products were then cloned into pCR2.1 using the TA-cloning kit (Invitrogen) according to manufacturer's instructions and sequenced using an Applied Biosystems sequencer courtesy of the Cleveland Clinic Foundation sequencing core facility.

**cDNA Isolation and Characterization.** The fetal brain cDNA library (CLONTECH catalogue #HL3003a) was plated according to manufacturer's instructions and transferred to Hybond N+ (Amersham). A total of 300,000 cDNAs was screened with [ $\alpha$ -<sup>32</sup>P]dCTP-labeled YDD15 cDNA insert in Church buffer at 65°C for 18 h. Filters were washed in 2 $\times$  SSC for 30 min and exposed to Kodak XAR-5 autoradiograph film overnight at -80°C with intensification screens. cDNA inserts were subcloned into pBluescriptIISK(+) and sequenced as described above.

## RESULTS

**Expression of ESTs Mapped to the YAC 943E4.** To determine whether human genes are transcribed in yeast, we analyzed the expression of two known genes, ATP1AL1 and KIAA177, and five EST sequences, which have previously been localized to YAC 943E4 (2, 6, 28, 29). Total cellular RNA was isolated from logarithmically growing cultures of yeast containing this YAC using a modified hot phenol method (26). This RNA was treated with RNase-free DNase and then reverse transcribed into first strand cDNA using random hexanucleotides primers. Subsequent PCR amplification of the first strand cDNA with human-specific primers to each gene or EST demonstrated expression of all seven sequences in the YAC RNA pool (Fig. 1). Contamination of the RNA with YAC or yeast DNA was not evident as PCR amplification failed to produce amplification products without the initial inclusion of reverse transcriptase in the cDNA synthesis

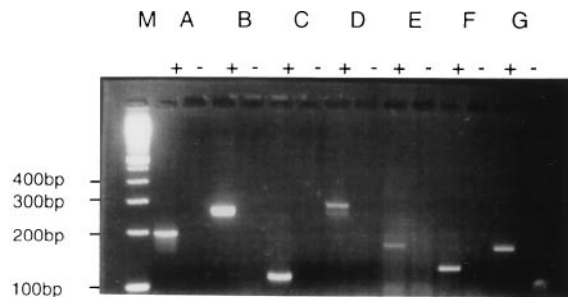


FIG. 1. Expression of ESTs and genes localized to YAC 943E4. Total RNA from yeast cells containing YAC 943E4 was reverse transcribed with random hexanucleotides (+), prior to PCR analysis with primers as previously described (30). A control reaction (-) following exactly the same procedure, but excluding reverse transcriptase during first strand synthesis, was performed simultaneously to determine whether samples were contaminated with DNA. ESTs and genes tested were as follows: lane A, KIAA177; lane B, ATP1AL1; lane C, D13S179E; lane D, D13S504E; lane E, D13S505E; lane F, D13S824E; and lane G, D13S182E. The size of the respective EST PCR products are as follows: KIAA177, 198 bp; ATP1AL1, 260 bp; D13S179E, 114 bp; D13S504E, 274 bp; D13S505E, 176 bp; D13S824E, 122 bp; and D13S182E, 176 bp. Marker (lane M): 100-bp ladder (GIBCO/BRL).

step (Fig. 1). These experiments indicated that human RNA transcripts are produced in the yeast host cell, and raised the possibility that transcribed human sequences could be isolated from YAC-containing yeast cells. Furthermore, because three of these ESTs, D13S179E, D13S182E, and D13S505E show highly restricted patterns of tissue expression (31), the analysis of YAC RNA may also provide a method of isolating transcripts that normally show a tissue-specific expression pattern, or that are only present at a low copy number in conventional cDNA libraries.

**RNA Fingerprinting of YACs.** Because the sequence requirements for correct polyadenylation of yeast mRNA differ significantly from mammalian systems (32), human polyadenylated mRNA cannot be isolated from yeast cells. In addition, because polyadenylation is required for mRNA stability and export from the nucleus (33), human transcripts from YACs are likely to be confined to the nucleus and unlikely to accumulate in yeast. Therefore, to identify novel transcribed sequences from YACs, we have modified the RNA fingerprinting by the RAP method described by Ralph *et al.* (24). First, total RNA was prepared from both exponentially growing YAC cultures and the yeast host strain AB1380. First strand cDNA synthesis was primed using random hexanucleotides to ensure a full representation of the length of any transcribed sequences. Second strand synthesis was then carried out using *Taq* DNA polymerase in the presence of a single, arbitrary primer at 40°C which promoted annealing of the oligonucleotide to sequences that match the six to eight bases at the 3' end of the primer. In the next step, this mixture of cDNA products was subjected to PCR amplification using the same arbitrary primer. To sample larger numbers of RNA species, and improve the recovery of YAC-specific products, the same first strand cDNA templates were amplified separately with different arbitrary primers, and the RNA fingerprint was resolved on 4% denaturing polyacrylamide gels. PCR products identified specifically in RNA pools from YAC cultures, compared with host yeast, were then excised and eluted from the gel using the crush and soak method (27). These cDNA products were then reamplified with the same primers at higher stringency. By hybridizing these PCR products to Southern blots containing DNA from the yeast host cells and yeast cells containing a variety of different overlapping YACs, it was possible to demonstrate the human origin of the transcribed sequences, and their approximate location within the region. Because no region of the RNA fingerprint gel is free from yeast sequences, each isolated gel fragment will inevitably contain some contaminating yeast cDNA. To purify

the human-derived cDNA products, the total amplification products from the gel slice were cloned into pCR2.1, and individual clones hybridized back to the YAC Southern blots to confirm their origin before sequencing.

**Nature of the Sequences Isolated by YAC Fingerprinting.** Because of our interest in cloning the genes responsible for the pathogenesis of an atypical myeloproliferative disorder associated with a specific t(8;13)(p11;q12) translocation (34), we have previously constructed a fine structure physical map encompassing the translocation breakpoint that is located in the chromosome 13q12.1 region (30). Using the unidirectional fragmentation method of Lewis *et al.* (25), a series of fragmented YACs was generated from the 943E4 YAC, which crosses the translocation breakpoint in 13q12.1. The physical relationship of the YACs used in the experiments described here is summarized in Fig. 2. In initial experiments, the RNA pools from the 943E4 YAC, the two fragmented YACs F10 and F18, and the overlapping CEPH megaYAC 775F9 were examined for the presence of YAC-derived transcripts. From the derived physical map, these YACs subdivide a 2.6-Mb region into five distinct sections (30).

Each RNA pool was amplified separately with a set of seven arbitrary primers. On average, each primer sampled a series of yeast RNA species ranging in size from 200 bp to 1.6 kb. However, comparison of the RNA fingerprints of each YAC with that of the host yeast strain AB1380, enabled the identification of several consistent YAC-specific products. An example of the YAC fingerprints obtained with three of the primers, ZF3, ZF6, and ZF14, is shown in Fig. 3. Some PCR products could only be identified in the fingerprint from one of the YACs and not in the other overlapping YACs. Subsequent analysis of these sequences indicated that they were either false positives (i.e., yeast sequences) or human sequences that also contained repetitive elements. When only those products that were evident in the YAC fingerprints from a minimum of two overlapping YACs were considered, between zero (for primer ZF-3) and 10 specific PCR products (ZF-14 primer) could be identified (Table 1). These PCR products ranged in size from 253 bp (YDD6) to 1075 bp (YDD20). Because no amplification products were produced

in control reactions where reverse transcriptase was omitted in the cDNA synthesis step, we could demonstrate that there was no contaminating YAC or yeast DNA in the RNA pool used for the differential display (data not shown).

By comparing RNA fingerprinting patterns for both fragmented YACs and the two overlapping YACs 775F9 and 943E4 simultaneously, we could identify human derived PCR products and map them to one of the five regions defined by the endpoints of these YACs. Each cDNA product was then isolated and hybridized to Southern blots containing DNA from the CEPH megaYACs and the series of fragmented YACs that we have previously generated from YAC 943E4 (30). These hybridization experiments suggested that, of 27 differential display products, 20 were of human origin and refined the localization of 17 of these transcribed sequences to specific subregions of the YAC contig (Fig. 2). Three of these sequences, YDD16, YDD17, and YDD18 could not be sublocalized because they contained repeat sequences. The other 7 products (25%) were shown to be derived from yeast cDNAs and were therefore not pursued further. Those cDNA products that hybridized to human-specific YAC fragments were cloned into the pCR2.1 vector using the TA-cloning kit (Invitrogen) and sequenced, as detailed above.

**Sequence Analysis of Human-Specific Sequences.** To determine the identity of the human-specific cDNA products, each clone was sequenced and then compared against the GenBank nucleotide database using the BLAST server at the National Center for Biotechnology Information.

**Identity with Known Genes or ESTs.** BLAST searches with two clones revealed overlaps with entries in dbEST. Clone YDD19, showed 92.7% identity with a human endothelial cell line clone (Gb:AA187540) over a 229-bp overlap and 84% homology to a cDNA clone derived from a mouse placental library (Gb:AA023961). Conceptual translation of these clones and YDD19 indicates that YDD19 contains the final 68 amino acids of a partial ORF of 122 amino acids. Sequence analysis also established a match to the 5' end of a 1578-bp IMAGE clone (Gb:R18527), which is part of an unmapped cluster of five overlapping cDNAs isolated from infant brain,

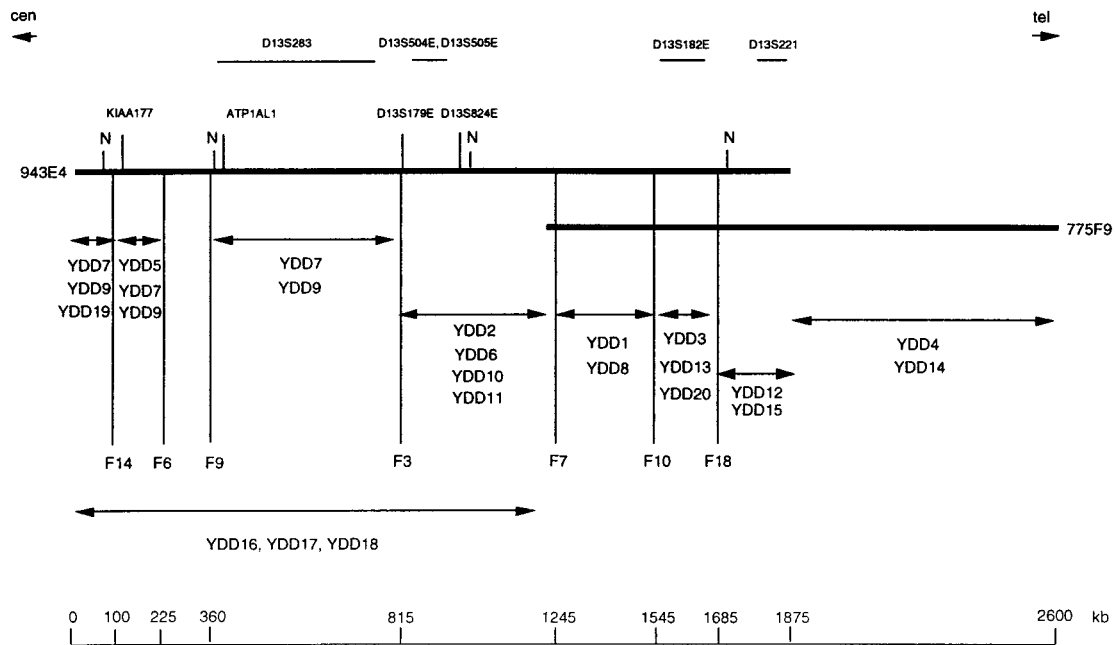


FIG. 2. Transcriptional map for the 2.6-Mb region of 13q12 defined by YACs 943E4 and 775F9. The locations of known ESTs, genes, microsatellite markers and *NotI* (N) restriction sites mapped to this region are shown above the 943E4 YAC (30). The vertical lines depict the endpoints of the fragmented YACs (F) used in this study as described by Still *et al.* (30). The distance (in kb) of these endpoints from the centromeric end of the 943E4 YAC is shown on the scale at the bottom of the figure. The locations of transcribed sequences derived from the RAP fingerprinting are shown relative to the framework of fragmented YAC endpoints, assigned ESTs, and genes. The assignment of each human cDNA product is indicated by the arrowed horizontal bars. The three YDD products (YDD16–18), which contain Alu repeats, could only be given a broad location based upon the RAP fingerprinting analysis.

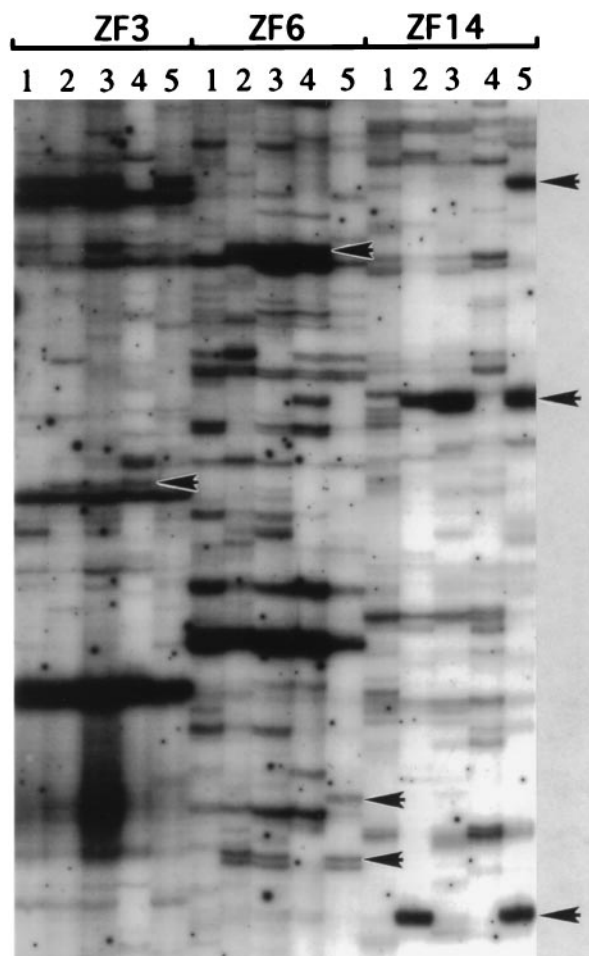


FIG. 3. RNA fingerprints of total RNA from YACs. RNA fingerprints were obtained from RNA pools derived from AB1380 host cells (lane 1), YAC 943E4 (lane 2), YAC F18 (lane 3), YAC F10 (lane 4), and YAC 775F9 (lane 5), using primers ZF3, ZF6, and ZF14 as described in the text. Arrows indicate five products consistently differentially amplified in RNA samples from overlapping YACs compared with the yeast host strain. Variable degrees of contamination with yeast mitochondrial RNA may account for some inconsistencies between background yeast bands. Note that the product amplified in three samples with the ZF3 primer was later determined to be a yeast artefact.

and breast (3NbHBst) cDNA libraries (31). The sequence provided by the YDD19 clone, therefore, joins two UNIGENE clusters which constitute 2.3 kb of a conserved gene which maps to a defined 100-kb region of 13q12.1 (Fig. 2).

Product YDD10 overlaps 81 bp of the IMAGE cDNA clone 62186 (Gb:T41117), which has homology to the BTF3 general transcription factor (31). Analysis of the sequence currently available for this cDNA clone indicated that the binding site for the arbitrary primer showed a match of 12/16 bp, with an exact match of 6 nt at the 3' end of the primer. Recently, this clone was mapped via radiation hybrid analysis to the region between D13S221 and D13S289 (35). However, using RNA fingerprinting, our results suggest that this gene resides within a 400-kb region just centromeric to D13S221 (Fig. 2).

**Human Endogenous Retrovirus-Like Sequence (HuERS).** YDD1 was identified in the RAP fingerprints from YACs F10, F18, 943E4, and 775F9, thereby mapping this clone to the region between the centromeric end of 775F9 and the F10 endpoint. Subsequent Southern blot analysis indicated that this product was transcribed from a single locus, located in the region between the endpoints of the F7 and F10 fragmented YACs (Fig. 2). Sequence analysis of YDD1 demonstrated numerous homologies with human cDNAs in the dbEST database. Further analysis revealed

that this was due to an 85.2% identity over a 398-bp overlap with an endogenous retrovirus-like sequence that is present in the 3' untranslated region of a number of different genes (36). It has been estimated that between 10 and 40 copies of this retroviral element are present per haploid genome (36). The RNA fingerprint of these YACs, therefore, indicates that an actively transcribed locus containing this retroviral element is located within this region of 13q12.

**Long Terminal Repeat (LTR) of Endogenous Retroviral Element ERV-9.** Two other clones, YDD7 and YDD9, showed significant homology to the LTR of the ERV-9 family of retroviral elements. We were able to identify and map at least three loci that hybridized to the YDD7 differential display product (Fig. 2), although it is unclear whether all of these loci are transcriptionally active. This class of retroviral element was originally isolated in cDNAs expressed in teratocarcinoma cells (37) and, in the case of the highly conserved zinc finger gene, ZNF-80, the presence of a single 5' ERV-9 LTR directs transcription of this gene in cells of the T cell and myeloid lineages (38). We have previously shown that YAC 943E4 crosses the t(8;13)(p11;q12) translocation breakpoint in two archival tumor samples from patients with an atypical myeloproliferative disorder associated with T cell leukemia and eosinophilia (34). Hence, the characterization of the transcriptionally active ERV-9-related sequences located on 943E4 may be useful candidates for the gene rearranged in this disorder.

**Sequences with no Homology to Known Genes or ESTs.** The remaining 12 sequences have no significant homology with entries in GenBank. Of these, 6 cDNA fragments contained partial ORFs ranging in size between 73 amino acids (YDD13) and 139 amino acids (YDD20) long (Table 1). None of these ORFs revealed significant homology in the SwissProt database, indicating that these sequences are derived from previously uncloned genes. The remaining cDNA fragments that contained no ORF, or an ORF of less than 70 amino acids, were presumed to be derived from noncoding regions or intronic sequences.

To isolate fuller length cDNA clones from these differential display products, we carried out a preliminary PCR-based screen of five cDNA libraries (bone marrow, fetal brain, fetal kidney, fetal retina, and adult hippocampus) using primers developed from the sequence of each differential display product. This analysis indicated that YDD4 was expressed in the bone marrow and fetal retina, whereas YDD15 was only identified in the fetal brain (data not shown). The remaining novel expressed sequences were not identified in the cDNA libraries tested, and therefore may represent transcripts expressed in different tissues to those examined. From this analysis, we decided to screen a fetal brain library using the YDD15 differential display product. Two cDNA clones were identified, suggesting that this transcript may be specifically expressed at very low levels (0.0007%), in the fetal brain. Sequence analysis of these clones indicated that both cDNAs contained the entire 739-bp sequence corresponding to the original differential display product. The 2.1-kb cDNA sequence generated from these clones failed to identify any homologous sequences in the GenBank database. In addition, no significant ORF was identified, suggesting that this clone represents part of the 3'-untranslated region of a novel gene mapping to the telomeric region of 943E4.

## DISCUSSION

The rate-limiting step of any positional cloning strategy is the ability to identify transcripts from a region of interest. The construction of extensive physical maps covering large parts of the human genome has concentrated around DNA cloned in YACs. Whereas these vectors provide the vehicle for cloning large fragments of DNA, the subsequent mapping of transcription units is hampered by their very size. A relatively rapid method of identifying transcribed sequences from these genomic clones would therefore greatly accelerate the identification of disease genes. Our motivation for pursuing an RNA

Table 1. Summary of the characteristics of transcribed sequences isolated from YAC RNA fingerprints

cDNA fragment	Arbitrary primer	Size, bp	Sequence Homology	Comments
YDD1	ZF1	465	emb:x08776	Retroviral element: HuERS-P1-1
YDD2	ZF6	278	Unique	Untranslated region
YDD3	ZF6	344	Unique	85 amino acid ORF
YDD4	ZF6	358	Unique	Untranslated region
YDD5	ZF6	408	Unique	Untranslated region
YDD6	ZF10	253	Unique	75 amino acid ORF
YDD7	ZF10	336		ERV9 LTR
YDD8	ZF11	444	Unique	Untranslated region
YDD9	ZF11	443		ERV9 LTR
YDD10	ZF14	255	gb:t41117	74 amino acid ORF
YDD11	ZF14	274	Unique	91 amino acid ORF
YDD12	ZF14	334	Unique	Untranslated region
YDD13	ZF14	463	Unique	73 amino acid ORF
YDD14	ZF14	552	Unique	Untranslated region
YDD15	ZF14	739	Unique	Untranslated region
YDD16	ZF11	434	Alu repeat	Untranslated region
YDD17	ZF11	440	Alu + Mer37 repeat	Untranslated region
YDD18	ZF11	879	Alu repeat	Untranslated region
YDD19	ZF11	1,005	gb:aa187540	Also homology to murine cDNA
YDD20	TKS1	1,075	Unique	144 amino acid ORF

The primer (ZF6–ZF14 or TKS1) with which each cDNA product was identified is listed next to the clone name. Sequences of the primers ZF6–ZF14 are from Ralph *et al.* (24); sequence of TKS1 primer is 5'-gacttggcctggcaccg-3'. Only open reading frames (ORFs) of >70 amino acids were classified as significant, although smaller ORFs generated by the presence of termination codons close to the ends of the clones could not be discounted as legitimate coding sequence. This was noted for YDD19, which contains the final 68 amino acids of a 122-amino acid ORF identified in the homologous overlapping cDNA (see text). Sequence homologies are noted in the homology and comments section.

fingerprinting approach from cultures of YAC-containing yeast cells was, therefore, to develop a rapid means of isolating gene probes from defined regions of the genome, which was potentially independent of tissue specific expression and transcript copy number in mammalian cells.

We have shown that human RNA transcripts from documented genes and ESTs are present in yeast cells that contain YACs with human inserts. Based upon this finding, we subsequently adapted the RAP fingerprinting method of Ralph *et al.* (24) and applied this technique to the detection of novel human transcribed sequences from YACs covering a 2.6-Mb region of chromosome 13q12. At present, 40% of the clones isolated match either known genes or ESTs in the databases, or contain large ORFs. Another two clones (10%) matched sequences from the 5' ends of retroviral elements that are known to be expressed in some human cells. The other 50% of the sequences isolated by differential display were demonstrably not derived from yeast RNA and so must originate from the human component of the YAC. It was possible to confirm that this was the case for two sequences because, using PCR, cDNAs corresponding to these clones were present in several different libraries. Using one of these clones, we were able to isolate a large (2.1 kb) cDNA that contained all of the sequence which we had isolated by differential display. It has been estimated that only 3–5% of the human genome is transcribed (39). Hence, if these clones originated from nonspecific transcription events or through random cloning, we would expect that only 1 of the 20 clones identified would resemble a gene sequence. However, at least 60% of the clones that we have isolated appear to be derived from human genes. Although this report concentrates on the detailed analysis of two large, overlapping YACs mapping to 13q12, we have recently obtained similar results using YACs from different regions of the genome. Thus, we have described a method of isolating transcribed sequences, which can be applied to different regions of the genome cloned in YACs. However, even though we have clearly demonstrated that this novel method can be used to identify such sequences, we have not addressed important questions concerning the mechanisms behind the production of human transcripts in yeast cells.

The first step in the initiation of transcription from polymerase II promoters requires the binding of the TFIID protein to the TATA element situated upstream of the transcription start site. Previous *in vitro* analysis has demonstrated that yeast TFIID can recognize, bind, and commence the assembly of the transcriptional complex at the TATA box in mammalian TATA-promoter constructs (22, 23). We have shown that transcription of all the genes currently localized to a 1.9-Mb YAC mapping to 13q12 occurs in yeast cells. This demonstrates that the human genes present on YACs are constitutively transcribed in yeast. Importantly, transcripts that are either tissue-specific, temporally expressed, or expressed at low levels in normal tissues may also be identified in the total RNA pool derived from cultures of growing YACs. The selection of the transcription site for these transcripts, however, is determined by the yeast transcriptional machinery, specifically TFIIB and the RNA polymerase itself (40). This results in the transcription start site shifting to a point downstream of the typical mammalian initiation site resulting in the generation of shorter transcripts (17, 23, 40). Therefore, mammalian transcripts derived from transcribed loci on YACs in yeast are unlikely to be intact. This in itself is not a drawback to our approach of finding human transcribed sequences from YACs, because this only removes 30 bp from the 5' end of the transcribed sequence.

Yeast TFIID has a high affinity for the TATA box, but also retains an ability to bind nonspecifically to DNA (41). *In vitro* kinetic studies indicate that TFIID bound nonspecifically, subsequently slides to higher affinity functional TATA elements followed by stabilization by associated factors (41). Hence, although there are controls to ensure transcription from in-context initiation start sites, it is also possible that promiscuous initiation from sites that resemble TATA elements could also occur on rare occasions. Therefore, if this were to occur within regions of the human YAC insert, some of the noncoding sequences identified by YAC RNA fingerprinting may actually originate from regions of the human YAC insert not normally transcribed in humans. We are currently subjecting these clones to a more rigorous expression analysis by RT-PCR to determine whether or not they are derived from legitimate transcripts.

In our previous analysis of this region using ESTs described in the database, we were able to map seven genes and ESTs to defined regions within the 943E4 YAC (30). By comparing the RAP fingerprints of different overlapping YACs, we have now isolated a further 20 expressed sequences and directly mapped these to discrete subregions. Furthermore, because we have only used a subset of the arbitrary primers described by Ralph *et al.* (24), and each resulted in a very different RAP profile, different oligonucleotide primers could sample additional transcripts or different fragments of the same transcripts already identified. Eleven clones mapped to regions known to contain EST sequences (30); however, none of the clones showed homology to the sequences currently available from these ESTs. It has been proposed that the pericentromeric long arm of chromosome 13 is particularly gene rich (39); however, it is unclear how many genes might be present in the 2.6-Mb region under consideration. If the genes are spaced at average intervals of 50 kb, then more than 40 genes could reside within this area. If this is the case, then it is perhaps not surprising that we did not identify the few genes that have already been placed in the region. Also, because ESTs represent relatively small fragments of cDNAs, it is also possible that some of these products represent different parts of a transcription unit also defined by a presently unmapped EST.

Several methods have been developed to isolate novel genes from large stretches of DNA (8–10). Initially approaches require the subcloning of these fragments into smaller vectors for subsequent analysis. The technique that we have described here overcomes the need for subcloning of YACs into smaller vectors by sampling sequences transcribed from whole YACs. More recently, the technique of hybrid capture has been used to isolate cDNAs directly from YACs (14), although, this technique is more successfully employed with cosmids, or bacterial artificial chromosomes (15). This method, although efficient, still requires that the gene of interest is transcribed in the cDNA pool used. The demonstration that rare tissue-restricted cDNAs are transcribed from YACs suggests that genes subject to strict temporal or tissue-specific regulation may be isolated from these YAC RNA pools. In addition, a major problem with hybrid capture and direct cDNA screening with genomic probes is the possibility that pseudogenes and repetitive sequences present on the YAC will hybridize to non-YAC-derived cDNAs in the cDNA library used, potentially generating a series of false positive cDNA clones. YAC RNA fingerprinting, however, only samples loci on the YAC that are transcriptionally active in yeast. Those cDNA clones isolated by RAP fingerprinting, which also contain repetitive elements, such as Alu or human endogenous retroviral-like sequence (HuERS), can therefore only have arisen by transcription through a repetitive element resident in the transcription unit itself. Furthermore, comparison of RNA fingerprints from overlapping or fragmented YACs not only provides internal controls against completely random initiation of transcription, but also permits the direct sublocalization of a differential display product, and therefore, the respective gene to a specific subregion of the YAC. This directed approach overcomes the random nature of mapping cDNAs using somatic cell hybrids and radiation reduced cell hybrids, by isolating gene probes directly from defined regions of the genome. We are now continuing to determine the specific tissue expression patterns of the transcribed sequences isolated in this study using RT-PCR, but the demonstration that longer cDNA clones can be isolated using the fragments obtained from RNA fingerprints offers a rapid means of isolating the corresponding genes.

We thank Sandya Rani, Ph.D., and Richard Ransohoff, M.D., for providing the ZF primers, and for their helpful discussions on RNA fingerprinting.

- Chumakov, I. M., Rigault, P., Le Gall, I., Bellanne-Chantelot, C., Billault, A., *et al.* (1995) *Nature (London)* **377** (Suppl.), 175–297.
- Adams, M. D., Kerlavage, A. R., Fleischmann, R. D., Fuldner, R. A., Bult, C. J., *et al.* (1995) *Nature (London)* **377** (Suppl.), 3–174.
- Auffray, C., Behari, G., Bois, F., Bouchier, C., Da Silva, C., Devignes, M. D., Duprat, S., Houlgatte, R., Jumeau, M. N., Lamy, B., Lorenzo, F., Mithcell, H., Mariage-Samson, R., Piétu, G., Pouliot, Y., Sabastiani-Kabaktchis, C. & Tessier, A. (1995) *C. R. Acad. Sci. Paris* **318**, 263–272.
- Hawthorn, L. A. & Cowell, J. K. (1996) *Cytogenet. Cell Genet.* **72**, 72–77.
- Roberts, T., Auffray, C. & Cowell, J. K. (1996) *Genomics* **36**, 337–340.
- Still, I. H., Roberts, T., Bia, B., Hawthorn, L. A., Auffray, C. & Cowell, J. (1996) *Genomics* **33**, 159–166.
- Gyapay, G., Schmitt, K., Fizames, C., Jones, H., Vega-Czarny, N., Spillett, D., Muselet, D., Prud'Homme, J. F., Dib, C., Auffray, C., Morissette, J., Weissenbach, J. & Goodfellow, P. N. (1996) *Hum. Mol. Genet.* **5**, 339–346.
- Bickmore, W. & Bird, A. P. (1992) *Methods Enzymol.* **216**, 224–244.
- Monaco, A. P., Neve, R. L., Colletti-Feener, C., Bertelson, C. J., Kurnit, D. M. & Kunkel, L. M. (1986) *Nature (London)* **323**, 646–650.
- Buckler, A. J., Chang, D. D., Graw, S. L., Brook, D. J., Haber, D. A., Sharp, P. A. & Houseman, D. E. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 4005–4009.
- Church, D. M., Stotler, C. J., Rutter, J. L., Murrell, J. R., Trofatter, J. A. & Buckler, A. J. (1994) *Nat. Genet.* **6**, 98–105.
- Elvin, P., Slynn, G., Black, D., Graham, A., Butler, R., Riley, J., Anand, R. & Markham, A. F. (1990) *Nucleic Acids Res.* **18**, 3913–3917.
- Wallace, M. R., Marchuk, D. A., Anderson, L. B., Letcher, R., Odeh, H. M., Saulino, A. M., Fountain, J. W., Brereton, A., Nicholson, J., Mitchell, A. L., Brownstein, B. H. & Collins, F. S. (1990) *Science* **249**, 181–186.
- Tagle, D. A., Swaroop, M., Lovett, M. & Collins, F. S. (1993) *Nature (London)* **361**, 751–753.
- Yamakawa, K., Mitchell, S., Hubert, R., Chen, X.-N., Colbern, S., Huo, Y.-K., Gadomski, C., Kim, U.-J. & Kornberg, J. R. (1995) *Hum. Mol. Gen.* **4**, 709–716.
- Struhl, K. (1995) *Annu. Rev. Genet.* **29**, 651–674.
- Buratowski, S., Hahn, S., Sharp, P. A. & Guarente, L. (1988) *Nature (London)* **334**, 37–42.
- Keaveney, M., Beckenstam, A., Feigenbutz, M., Vreind, G. & Stunnenberg, H. G. (1993) *Nature (London)* **362**, 562–566.
- Flanagan, P. M., Kelleher, R. J. I., Feaver, W. J., Lue, N. F., LaPointe, J. W. & Kornberg, R. D. (1990) *J. Biol. Chem.* **265**, 11105–11107.
- Cormack, B. P., Strubin, M., Stargell, L. A. & Struhl, K. (1994) *Genes Dev.* **8**, 1335–1343.
- Struhl, K. (1989) *Annu. Rev. Biochem.* **58**, 1051–1077.
- Lue, N. F., Flanagan, P. M., Sugimoto, K. & Kornberg, R. D. (1989) *Science* **246**, 661–664.
- Prentice, H. L. & Kingston, R. E. (1992) *Nucleic Acids Res.* **20**, 3383–3390.
- Ralph, D., McClelland, M. & Welsh, J. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 10710–10714.
- Lewis, B. C., Shah, N. P., Braun, B. S. & Denny, C. T. (1992) *GATA* **9**, 86–90.
- Wise, J. A. (1991) *Methods Enzymol.* **194**, 295–415.
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainville, NY).
- Berry, R., Stevens, T. J., Walter, N. A., Wilcox, A. S., Rubano, T., Hopkins, J. A., Goold, R., Soares, M. B. & Sikela, J. M. (1995) *Nat. Genet.* **10**, 415–423.
- Guilford, P., Ben Arab, S., Blanchard, S., Levilliers, J., Weissenbach, J., Belkahlia, A. & Petit, C. (1994) *Nat. Genet.* **6**, 24–28.
- Still, I. H., Roberts, T. & Cowell, J. K. (1997) *Ann. Hum. Genet.* **61**, 15–24.
- Boguski, M. S. & Schuler, G. D. (1995) *Nat. Genet.* **10**, 369–371.
- Proudfoot, N. (1991) *Cell* **74**, 671–674.
- Jackson, R. J. & Standart, N. (1990) *Cell* **62**, 15–24.
- Kempski, H., MacDonald, D., Michalski, A. J., Roberts, T., Goldman, J. M., Cross, C. P. & Cowell, J. K. (1995) *Genes Chromosomes Cancer* **12**, 283–287.
- Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., *et al.* (1996) *Science* **274**, 540–546.
- Harada, F., Tsukada, N. & Kato, N. (1987) *Nucleic Acids Res.* **15**, 9153–9162.
- La Mantia, G., Maglione, D., Pengue, G., Di Cristofano, A., Simeone, A., Lanfrancone, L. & Lania, L. (1991) *Nucleic Acids Res.* **19**, 1513–1520.
- Di Cristofano, A., Strazullo, M., Longo, L. & La Mantia, G. (1995) *Nucleic Acids Res.* **23**, 2823–2830.
- Saccone, S., de Sario, A., Valle, G. D. & Bernardi, G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 4913–4917.
- Li, Y., Flanagan, P. M., Tschochner, H. & Kornberg, R. D. (1994) *Science* **263**, 805–807.
- Coleman, R. A. & Pugh, B. F. (1995) *J. Biol. Chem.* **270**, 13850–13859.
- Nehls, M., Pfeifer, D. & Boehm, T. (1994) *Oncogene* **9**, 2169–2175.
- Virgilio, L., Narducci, M. G., Isobe, M., Billips, L. G., Cooper, M. D., Croce, C. M. & Russo, G. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 12530–12534.