Incognito rRNA and rDNA in databases and libraries.

I L Gonzalez and J E Sylvester

Genome Res. 1997 7: 65-70
Access the most recent version at doi:10.1101/gr.7.1.65

LETTER

Incognito rRNA and rDNA in Databases and Libraries

Iris L. Gonzalez[1,3] and James E. Sylvester[2]

Allegheny University of the Health Sciences, M.C.P. Hahnemann School of Medicine, Department of Pathology, Philadelphia, Pennsylvania 19102

Both ribosomal DNA (rDNA) and ribosomal RNA (rRNA) are over-represented in the starting material for genomic and cDNA libraries; thus, their sequences have the potential of repeatedly entering the various databases. When DNA (both transcribed and intergenic spacer regions) is used as query sequence, a great number of matches are found in the databases, particularly in the EST database, and to a lesser extent among genomic sequences and STSs, which are not identified as rDNA. We discuss the following explanations for the widespread occurrence of rDNA in cDNA and genomic DNA libraries: pseudogenes of rRNA in other genomic locations, mRNA-derived pseudogenes that reside in rDNA, cDNAs derived from rRNA [either by self-priming or by internal oligo(dT) priming], cDNAs derived from actual transcripts of the rDNA intergenic spacer, and genomic DNA contamination of RNA preparations. Because so many database entries contain unidentified rDNA, we recommend that all sequence submissions be checked (by the submitters) for the presence of structural RNAs in addition to repetitive sequences.

Various reports exist regarding contamination of genomic databases: One is that supposed gene sequences contain vector sequences (Lopez et al. 1992), another is that sequences supposedly derived from the nuclear genome actually contain some sequences from the mitochondrial genome (Wenger and Gassman 1995), and yet another is that libraries from one organism actually also contain DNAs derived from other species (Gersuk and Rose 1993; Kessin and Van Lookeren Campagne 1993; Dean and Allikmets 1995). There are other forms of possible physical or informational contamination, namely contamination of cDNA libraries with genomic DNA-derived clones, and the addition to databases of sequences that are partially or entirely ribosomal RNA (rRNA) or ribosomal spacers without identifying them as such in the sequence definitions. We became aware of this problem while conducting FASTA searches of the nucleotide databases using as query a distal sequence from an acrocentric chromosome p-arm, which, unbeknown to us, contained a 28S rRNA pseudogene fragment. To our surprise, we not only found the match with nuclear 28S rRNA but also matches to numerous known and unknown genes. The matches were all attributable to highly conserved 28S rRNA sequences, and the species from which they were derived ranged from animals and plants to unicellular eukaryotes. Below we present results from a more extended (but by no means complete) search using human nuclear ribosomal DNA (rDNA) sequences as query.

Present addresses: [1]DuPont Hospital for Children, Clinical Science, Wilmington, Delaware 19899; [2]The Nemours Children's Clinic, Jacksonville, Florida, 32207.
[3]Corresponding author.
E-MAIL gonzalez@lansol.net; FAX (302) 368-3331.

RESULTS AND DISCUSSION

FASTA searches of the database have revealed numerous sequences that match either the rRNAs or their gene spacers. The category with the most numerous matches is human STSs (sequence-tagged sites), which contain simple repeated motifs that are common to many genes and also are present in rDNA (e.g., CA, CTTT, GGC). These matches do not mean that the sequences are rDNA-derived; however, STSs are often primed from Alu elements (33 of which are present in the rDNA intergenic spacer), resulting in a fair collection of IGS clones. We will not consider them as contaminants, as the purpose of these sequences is to mark sites in the genome for mapping; they can be anonymous, and they do not purport to represent genes. However, based on their sequence, we can state that their map position is clearly acrocentric p-arm and that there is unnecessary redundant deposition of these sequences in the database. For example, there are 11 Alu-primed STSs from the IGS segment 36400–36520. But many sequences present in the EST (expressed sequence tag) database, presumably derived from poly(A)$^+$ RNA,

and also in the gene database, are either entirely or partially ribosomal sequences and are not identified as such. The lack of identification of these sequences leads us to classify them as "informational contaminants." We have divided these sequences into two categories: (1) those that match the transcribed region of rDNA, and (2) those that match the IGS of rDNA.

Sequences That Match the Transcribed Region of rDNA

Sequences that match the transcribed region of rDNA can be further subdivided into those that are entirely rRNA and those that are composites of rRNA plus either a known or unknown sequence. Table 1 shows the frequency with which unidentified large subunit rRNA-containing clones appear in the EST database. The search was conducted with conserved region fragments covering the entire human 28S rRNA and reveals marked differences in the frequency of different fragments; these differences may be related to how well the bacterial host tolerates different clones.

rRNA Sequences

Sequences that are entirely rRNA include both the rRNAs and their external and internal transcribed spacers. Considering the abundance of rRNA transcripts, the most likely origin of these sequences in the EST database is actual reverse transcription of rRNA into cDNA, even though the RNAs were poly(A)$^+$ selected and the cDNAs were oligo(dT)-primed. Although it is usually difficult to see where or how the oligo(dT) could have primed, it is easy to envision that the secondary structure-rich rRNA could self-prime by hairpin formation. The alternative origin is from contaminating genomic DNA (0.5% of which is rDNA) that is digested with the restriction enzymes used at the cloning step. Table 2 contains several examples of such clones, originating from various libraries and from different species. One example is in perfect agreement with a special cloning strategy: cDNA L37707 was obtained in a search for transcripts containing trinucleotide repeats and was generated with oligonucleotide $[CAG]_8$ to prime the second strand; this sequence is 28S rRNA starting at a site with $[CAG]_2$.

Although our initial view was that all the examples found were nuisance database contamination, we also found an example that provides important information about rDNA transcription in vivo. D57487 is a human EST apparently derived from the unprocessed rRNA transcript, because it includes part of the 3′ external transcribed spacer, continues through the normal transcription termination signal into the intergenic spacer, and stops exactly at the second (T_3) of a series of spacer terminators. This demonstrates that readthrough does occur and that some of the fail-safe spacer termina-

Table 1. Frequency of rDNA in the EST Database

	rDNA identified		rDNA not identified		non-rDNA	
28S location[a]	human	nonhuman	human	nonhuman	human	nonhuman
7930–8330	21	5	12	30	118	14
8800–8950	0	1	0	0	121	78
9200–9600	24	17	16	35	57	47
9600–10000	17	21	10	64	56	30
10180–10550	12	9	7	39	80	46
10550–10850	6	6	2	45	105	37
11500–11890	47	9	21	90	24	2
11950–12250	18	1	46	45	63	26
12250–12600	41	8	51	62	38	0
12750–12970	16	36	16	85	18	29

FASTA searches conducted using human conserved region 28S rRNA fragments (with the exception of GC-rich 8800–8950) as query against the EST database. For each search, 200 alignments were inspected; increasing the number of 400 alignments yielded no additional rRNA fragments.
[a]28S location in U13369.

Table 2. Examples of Unidentified Fragments of Ribosomal Transcripts

Accession no.	Species	Sequence	rDNA location*	Percent similarity[a]	Length
D25770	human	5′ ETS	1447–1607	98.7	163
T04433	*Arabidopsis*	18D	1337–1734	98	400
M78253	human	18S	4976–5320	98.2	345
T38386	*Sacharomyces*	18S	75–634	94	562
D29194	human	18S + ITS1	5464–5710	93.3	256
F14864	pig	18S	3666–3803 (HU)	98.6	138
S29116	human	ITS1	5577–5748	96	175
Z36919	chicken	28S	12113–12202 (HU)	93.3	90
Z18874	human	28S	12120–12444	98.5	327
D57487	human	3′ ETS + IGS	13254–13505	93.3	257
L37707	human	28S	8854–8914	96.7	62

[a]rDNA location and percent similarity correspond to the species under consideration; when not available, the human sequence (U13369) was used and indicated (HU).

tors are as functional in vivo as they have been shown to be in vitro (Pfleiderer et al. 1990).

Apart from a few interesting cases, transcribed region clones clutter the EST database, to which large numbers of short (unidentified) fragments are submitted. To illustrate how prevalent the problem is, it is possible to reconstruct from the many fragments in the EST database 3032 bases of the 3374 nucleotide *Arabidopsis* 25S rRNA gene.

Composite rRNA Sequences

Sequences that are composites of rRNA and other sequences (see examples in Table 3) can arise in sev-

Table 3. Examples of Composite Clones Containing rRNA and Other Sequences

Accession no.	Species	Sequence	rDNA location[a]	Percent similarity[a]	Length
L25494	human	EST + 28S	12241–12359	89.5	134
T50883	human	P450IIb6 + 28S	12157–12227	100	70
R28706	potato	EST + 28S	3494–3738 (*Arabidopsis*}	94.2 (*Arabidopsis*)	244
X59474	mouse	Hox 2.9 + 28S	4261–4712	93.6	441
X54075	*Zea mays*	18 kD HSP + 28S	2961–3153 (rice)	97 (rice)	194
L36663	mouse	CArG box binding factor + 28S	3406–3742	97.4	383
M69055	rat	rIGFBP-6 + 28S	4629–4694	100	65
L26087	rat	*unc-18* homolog bound to syntaxin + 28S	4533–4685	98.7	154
T07450	human	EST + 5.8S	6658–6703	100	46
X78990	mouse	testin 2 + 28S	860–1213	98.6	354

[a]rDNA location and percent similarity correspond to the species under consideration; when not available, another species is indicated in parentheses. (Human rDNA) U13369; (*Arabidopsis* 28S-containing clone.) X52320; (mouse 28S) X00525; (rice 28S) M11585; (rat 28S) V01270.

eral ways: transcripts of genes that contain a rRNA pseudogene (usually only a fragment); chimeric composites of rRNA and mRNA reverse transcripts that are joined in the cloning procedure; and chimeric composites of cDNA and genomic DNA fragments also generated during cloning. It is difficult to distinguish among these, but some obvious chimeras have the stated cloning site separating a rRNA fragment from another sequence fragment (unfortunately, many sequences in the database do not include cloning information). For example, EST L25494 contains an unknown sequence separated by an *Eco*RI site from a 28S rRNA fragment. Another obvious chimera is T50883, a human clone containing 70 bases of 28S rRNA adjacent to a small fragment of P450IIB6 coding sequence (a whole P450 is also in the database). There are several sequences of identified genes that contain rRNA fragments in either the upstream (5′) or downstream (3′) untranslated regions, and thus could be examples of sequences containing rRNA pseudogenes (they also could be chimeric artifacts). These include X59474 (mouse *Hox 2.9* with 441 bases from the 3′ end of 28S rRNA), X54075 (maize 18-kD heat shock protein with 194 bases of 28S rRNA), L36663 (mouse CArG box binding factor with 28S rRNA).

One puzzling and potentially interesting example is X78990, mouse testin2. This sequence includes 354 bases of 28S rRNA at the 5′ end of its coding sequence, of which 200 bases are designated as "CpG island" and 105 bases are part of the deduced testin2 open reading frame (ORF). The match to 28S rRNA was not recognized by Divecha and Charleston (1995). It certainly would be exciting to find that 28S rRNA can perform a dual role as part of the ribosome and also be translated! There is a recent report of a short functional peptide coded within the large subunit rRNA of *Escherichia coli* (Tenson et al. 1996); however, the conservative interpretation is that the mouse testin2 clone is a composite clone.

It has been suggested by an anonymous reviewer that rRNA fragments within messages could be a natural occurrence and could have biological significance. We agree that this is a concept worthy of exploration, which would open a whole new avenue of experimental work.

Sequences That Match the rDNA IGS

The rRNA IGSs are generally believed to be nontranscribed, except for short transcripts originating from spacer promoters not far upstream of the main promoter (e.g., see Cassidy et al. 1987). Three possible scenarios can be envisioned to explain the origin of numerous unidentified ESTs matching the IGSs: First, the IGS, or part of it, is transcribed and real IGS cDNAs are being cloned. This is an exciting possibility, especially when ESTs originating from different libraries cluster in one location. A very high percentage of sequence similarity to the published rDNA sequence (U13369) should be found while allowing for differences due to a 3% spacer sequence variability (Gonzalez and Sylvester 1995) and sequencing errors. How to recognize a real IGS transcript from its sequence alone is a problem. If it were a translated transcript, one would look for ORFs; however, a regulatory transcript is not translated and would require experimental detection on Northern blots. The second possibility is that the IGS contains pseudogenes and that we are finding matches to the parent genes that are in the libraries. In this case, the sequence similarity between the IGSs and the ESTs would be lower because of the accumulation of mutations in the pseudogene. This can be tested experimentally by probing suitably digested genomic DNA with the fragment in question: If there is a pseudogene, one would find a strong signal arising from the 400 copies of pseudogene in rDNA and weak signals arising from the genomic locations of the parent gene exons. The third possibility is that the cDNA libraries are contaminated with rDNA-derived clones. Unfortunately, this is the most likely explanation for most or all of the matches between the rDNA IGSs and ESTs, because many ESTs begin at *Xho*I sites (*Xho*I being the enzyme used for cloning) and at $[A]_n$ or $[T]_n$ tracts in rDNA where $[dT]_n$ could have primed.

Here we will discuss single clones and groups of overlapping clones that can be ascribed to DNA contamination of cDNA libraries (see Table 4 for examples discussed below). Redundant clones are often derived from different libraries, indicating that DNA contamination is a widespread problem.

Both clone 75639 of the Stratagene ovary library (represented by 3′ and 5′ sequences T58431/T58463) and clone 79581 of the Stratagene lung library (represented by 3′ and 5′ sequences T62861/T62711) appear to be derived from oligo(dT) priming at T-rich segments of the IGS preceding these sequences (at 31936–31955 and 41032–41056 of U13369, respectively).

The 3′ end of clone 118145 of the Stratagene lung library (3′ and 5′ sequences T92433/T91479) is formed by a naturally occurring *Xho*I restriction site in rDNA. *Xho*I was one of the enzymes used in cloning. The sequences are 98.4–100% identical to the rDNA sequence. Unless most of the IGS is nonspe-

Table 4. Examples of cDNA Clones Matching the Human rDNA IGS

Accession no.	Library	Clone no.	End	rDNA location	Percent similarity	Length
T58431	Strat. ovary	75639	3'	31952–32274	98.7	323
T58463	Strat. ovary	75639	5'	31984–32403	98.3	421
T62861	Strat. lung	79581	3'	41054–41348	99	294
T62711	Strat. lung	79581	5'	41057–41343	95.6	296
T92433	Strat. lung	118145	3'	19142–19509	98.4	372
T91479	Strat. lung	118145	5'	19077–19184	100	109
T94179	Strat. lung	119180	5'	18203–18486	100	248
T94095	Strat. lung	119180	3'	18498–18890	98.3	372
D56572	Clontech human aorta	6572	3'	37767–38150	96.8	382
D57294	Clontech human aorta	6572	3'	37818–38150	99.4	332
T92332	Strat. lung	118182	3'	35636–35989	98.8	260
T92381	Strat. lung	118182	5'	35724–35920	98	196
T98250	Soares fetal liver	122006	3'	22906–23018	77.9	113
T93159	Strat. lung	118656	5'	17230–17472	97.9	244

Numbering as in U13369.

cifically transcribed, it is not likely that this sequence is a cDNA, as it contains mostly pyrimidine-rich simple sequence and three Alu element fragments.

A group of five Stratagene lung library clones, [117431, 81173, 118500, 119180, 79017 (T89912, T70021, T92647, T94179/T94095, T61953)] represents 1100 bp of overlapping sequence between nucleotide 17893 and 18944 in the IGS. Two of the sequences are of the 3′ ends; one of these ends is at a natural *Xho*I site in rDNA, and the other is nearby. Again, *Xho*I was the 3′ cloning site, and the similarity to the rDNA sequence ranges from 96.5% to 100%, suggesting that the clones are either rDNA derived or that the synthesized cDNA was cut at the internal *Xho*I site. Although this region contains some sequence of average composition, and computer analysis reveals several ORFs longer than 50 amino acids in various reading frames, one end consists of an entire Alu repetitive element, making it less likely to be a cDNA from a mature message.

Clone 6572 from the Clontech human aorta library (represented by sequences D56572/57294) is 99.4% similar to rDNA. It could be derived from oligo(dT) priming on rDNA, at an A-rich IGS segment following the sequence 38151–38358, or from oligo(dT) priming on a transcript containing this A-rich segment.

A group of 12 overlapping clones [represented by sequences T02858, T92332/T92381, T51823, T58159/T63096 (a chimeric sequence), T93312/T93989, T92440, T62217, H68123, N23637, N22926, T57471, T61528] matches a region that is immediately adjacent to and briefly overlaps a known pseudogene in the IGS (pseudo-CDC27HS; Tugendreich et al. 1993). But this group of clones does not represent CDC27HS cDNA and does not extend pseudo-CDC27HS: The parent CDC27HS gene has a sequence similarity of 91% to rDNA between nucleotide 33673 and 35699, whereas this group of clones has similarities of 95.7–99.1% to rDNA between nucleotide 35619 and 36798, indicating that it is probably derived from rDNA. Moreover, this region also contains a retroposed SINE (Alu 8) that is known to have entered the primate genome before the entry/fixation of pseudo-CDC27HS (Gonzalez et al. 1993). Although it is not clear how all of these clones originated, three of them begin at sites where rDNA has naturally occurring *Xho*I sites and could easily be derived from contaminating genomic DNA. This group of clones includes representatives from four different cDNA libraries.

Another probable instance of genomic DNA contamination of the cDNA library is sequence T93159 (at nucleotide 17230–17472), which has 98% sequence similarity to IGS. It begins within an Alu element, ends in a simple sequence pyrimidine-

rich region, and is an unlikely prospect for a real transcript as the Alu element belongs to an old family and does not have an intact box B sequence required for transcription.

A group of 18 sequences, representing six libraries, is clearly not derived from the IGS (e.g., T98250) and may represent a retroposon in rDNA. They are all very similar to each other and match IGS 22714–23023 [a repeated segment dubbed "butterfly" (Gonzalez and Sylvester 1995)] at 70–75% similarity. These clones could represent cDNAs from parent genes that match the pseudogene present in the rDNA. The sequence divergence is high (25–30%), indicating entry and fixation long ago. An alternative and more likely explanation is that they could belong to a repetitive and possibly retroposed group; this is supported by the fact that butterfly is repeated in another location of rDNA and flanked by direct repeats.

The above examples and discussion have focused on two problems: contamination of information and contamination of libraries. The first is easy to solve and should be the responsibility of sequence submitters. Because many sequences in the database include in their annotations the presence of repetitive elements, such as Alu elements, LINEs, and MER sequences, it is also desirable to check for the presence of structural RNAs. Their transcribed sequences are highly conserved, so the search should be easy; moreover, a complete intergenic spacer is available for the human. Also, at present, the same sequence is submitted repeatedly just because there are multiple clones representing it in a library; there should be a way to bin sequences that are redundant or are variants of each other. The second problem, that of library construction, is technical in nature and is the responsibility of the various library builders (Mistry et al. 1993).

METHODS

FASTA searches (GCG package, Genetics Computer Group, Madison, WI) were performed using as query 200- to 400-base fragments of human rDNA; the best 200 matches were reviewed in EST searches, and the best 400 were reviewed in gene database searches. The latter database contains hundreds of known rRNA genes, which led us to concentrate on the EST database. Only the conserved regions of the 28S rRNA coding region were used, because the variable regions have numerous and nonsignificant matches because of their skewed GC-rich base composition. The same approach was used for the transcribed spacers, from which microsatellite type segments had to be omitted. For the intergenic spacer (IGS), fragments free of repetitive Alu and LINE elements were used; sequences containing microsatellites were also omitted. Human rDNA sequence numbers indicated in the text are from the complete rDNA repeat U13369.

ACKNOWLEDGMENTS

We thank Paul Keenan (Allegheny University) for computer assistance. We express appreciation to anonymous reviewers for their thoughtful suggestions. This work was supported by National Institutes of Health grant GM R01 41625.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Cassidy, B.G., H.-F. Yang-Yen, and L.I. Rothblum. 1987. Additional RNA polymerase I initiation site within the nontranscribed spacer region of the rat rRNA gene. *Mol. Cell. Biol.* **7:** 2388–2396.

Dean, M. and R. Allikmets. 1995. Contamination of cDNA libraries and Expressed-Sequence-Tags databases. *Am. J. Hum. Genet.* **57:** 1255–1256.

Divecha, N. and B. Charleston. 1995. Cloning and characterization of two new cDNAs encoding murine triple LIM domains. *Gene* **156:** 283–286.

Gersuk, V.H. and T.M. Rose. 1993. Database contamination: Letters. *Science* **260:** 605.

Gonzalez, I.L. and J.E. Sylvester. 1995. Complete sequence of the 43-kb human ribosomal DNA repeat: Analysis of the intergenic spacer. *Genomics* **27:** 320–328.

Gonzalez, I.L., S. Tugendreich, P. Hieter, and J.E. Sylvester. 1993. Fixation times of retroposons in the ribosomal DNA spacer of human and other primates. *Genomics* **18:** 29–36.

Kessin, R.H. and M.M. Van Lookeren Campagne. 1993. Database contamination: Letters. *Science* **260:** 605.

Lopez, R., T. Kristensen, and H. Prydz. 1992. Database contamination. *Nature* **355:** 211.

Mistry, A., R. Greenlee, and K. Fong. 1993. Database contamination: Letters. *Science* **260:** 605.

Pfleiderer, C., A. Smid, I. Bartsch, and I. Grummt. 1990. An undecamer DNA sequence directs termination of human ribosomal gene transcription. *Nucleic Acids Res.* **18:** 4727–4736.

Tenson, T., A. DeBlasio, and A. Mankin. 1996. A functional peptide encoded in the *Escherichia coli* 23S rRNA. *Proc. Natl. Acad. Sci.* **93:** 5641–5646.

Tugendreich, S., M.S. Boguski, M.S. Seldin, and P. Hieter. 1993. Linking yeast genetics to mammalian genomes: identification and mapping of the human homologue of cdc27 via the EST database. *Proc. Nat. Acad. Sci.* **90:** 10031–10035.

Wenger, R.H. and M. Gassmann. 1995. Mitochondria contaminate databases. *Trends Genet.* **11:** 167–168.

Received August 20, 1996; accepted in revised form December 3, 1996.