al. 1994). Furthermore, cases of reverse mutation with contraction to normal size have been rarely observed in myotonic dystrophy (Brunner et al. 1993) and in fragile X syndrome (Antiñolo et al. 1996).

The GAA repeat size on FRDA alleles showed a negatively skewed distribution, with the majority of chromosomes having shorter allele length than the mode (Filla et al. 1996). The prevalence of contractions on expansions is consistent with this kind of distribution. The major factor determining contraction of FRDA alleles is paternal transmission. In our sample, all fathers transmit shortened or unvaried alleles to their children. These data were also confirmed by sperm DNA analysis. In all three cases, the amplification of the FRDA gene from sperm yielded a diffuse array of shorter products in comparison with blood. To test preferential amplification of the shortest alleles, we added increasing amounts of blood DNA to sperm DNA during amplification. We could detect the longest alleles at 20% of blood DNA concentration. We suggest that the majority of the gametes carry alleles shorter than blood, but we are unable to define their size distribution. Parental gender effect on repeat length variation has been described in other triplet diseases, occurring on male transmission in CAG repeat diseases and female transmission in fragile X syndrome and myotonic dystrophy. Parental gender effect is more evident in FRDA, apparently conditioning the direction of the variation. Also, GAA repeat size affected instability, since the tendency to expand was more pronounced in the maternal shortest alleles. Further analyses on larger samples are needed to clarify the role of allele size on its variability.

In this study, we found no association between extended haplotypes of the region and FRDA allele tendency to expand or to contract. In addition, the analysis of several instances of intergenerational transmission of the same allele confirms that the parental gender effect is more important than the putative effect of *cis*-acting elements.

In summary, our data suggest that (i) the FRDA GAA repeat is highly unstable during meioses, (ii) contractions outnumber expansions, (iii) both parental source and sequence length are important factors in variability of FRDA expanded alleles, and (iv) the tendency to contract or expand does not seem associated with particular haplotypes. The emerging picture of FRDA gene variability seems to be different from that proposed for other triplet diseases.

LUIGI PIANESE,[1,3,*] FRANCESCA CAVALCANTI,[1,3,*] GIUSEPPE DE MICHELE,[2] ALESSANDRO FILLA,[2] GIUSEPPE CAMPANELLA,[2] OLGA CALABRESE,[1,3] IMMA CASTALDO,[1,3] ANTONELLA MONTICELLI,[1] AND SERGIO COCOZZA[1]

Departments of [1]Molecular and Cellular Biology and Pathology and CEOS and of [2]Neurology, Federico II University, Naples; and [3]Neuromed, Pozzilli, Italy

Acknowledgments

This work was supported by Italian Telethon (grant 722 to S.C.) and by grants from Ministries of Health and of University.

References

Antiñolo G, Borrego S, Cabeza J C, Sánchez R, Sánchez J, Sánchez B (1996) Reverse mutation in fragile X syndrome. Am J Hum Genet 58:237–239

Bates G, Lehrach H (1994) Trinucleotide repeat expansions and human genetic disease. Bioessays 16:277–284

Brunner HG, Jansen G, Nillesen W, Nelen MR, de Die CE, Howeler CJ, van Oost BA, et al (1993) Brief report: reverse mutation in myotonic dystrophy. N Engl J Med 328:476–479

Campuzano V, Montermini L, Moltò MD, Pianese L, Cossèe M, Cavalcanti F, Monros E, et al (1996) Friedreich ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science 271:1423–1427

Filla A, De Michele G, Cavalcanti F, Pianese L, Monticelli A, Campanella G, Cocozza S (1996) The relationship between trinucleotide (GAA) repeat length and clinical features in Friedreich ataxia. Am J Hum Genet 59:554–560

Filla A, De Michele G, Marconi R, Bucci L, Carillo C, Castellano AE, Iorio L, et al (1992) Prevalence of hereditary ataxias and spastic paraplegias in Molise, a region of Italy. J Neurol 239:351–353

Pianese L, Cocozza S, Campanella G, Castaldo I, Cavalcanti F, De Michele G, Filla A, et al (1994) Linkage disequilibrium between FD1-D9S202 haplotypes and Friedreich's ataxia locus in a central-southern Italian population. J Med Genet 31:133–135

Telenius H, Almqvist E, Kremer B, Spence N, Squitieri F, Nichol K, Grandell U, et al (1995) Somatic mosaicism in sperm is associated with intergenerational $(CAG)_n$ changes in Huntington disease. Hum Mol Genet 4:189–195

Zhang L, Leeflang EP, Yu J, Arnheim N (1994) Studying human mutations by sperm typing: instability of CAG trinucleotide repeats in the human androgen receptor gene. Nat Genet 7:531–535

*These authors equally contributed to this work.

Address for correspondence and reprints: Dr. Giuseppe De Michele, Clinica Neurologica, Università "Federico II," Via S. Pansini 5, 80131 Napoli Italy. E-mail: cocozza@cds.unina.it

0002-9297/97/6002-0027$02.00

Am. J. Hum. Genet. 60:463–466, 1997

Contamination of Sequence Databases with Adaptor Sequences

To the Editor:

Because of the exponential increase in the amount of DNA sequences being added to the public databases on a daily basis, it has become imperative to identify sources of contamination rapidly. Previously, contaminations of sequence databases have been reported, to alert the scientific community to the problem (Anderson 1993; Binns 1993; Gersuk and Rose 1993; Kessin and Van Lookeren Campagne 1993; Dean and Allikmets 1995). These contaminations can be divided into two categories. The first category comprises host sequences that have been difficult for submitters to manage or control. Examples include anomalous sequences derived from *Escherichia coli,* which are inserted into the chromosomes (and plasmids) of the bacterial hosts (Binns 1993). Insertion sequences are highly mobile and are capable of transposing themselves into plasmids during cloning manipulation. Another example of the first category is the infection with yeast genomic DNA (Anderson 1993; Gersuk and Rose 1993) or with bacterial DNA (Kessin and Van Lookeren Campagne 1993; Dean and Allikmets 1995) of some commercially available cDNA libraries from Clontech. The second category of database contamination is due to the inadvertent inclusion of nonhost sequences. This category includes incorporation of cloning-vector sequences and multicloning sites in the database submission. M13-derived artifacts have been common, since M13-based vectors have been widely used for subcloning DNA fragments (Lamperti et al. 1992; Lopez et al. 1992; Reynolds 1994). Recognizing this problem, the National Center for Biotechnology Information (NCBI) started to screen, in April 1994, all sequences directly submitted to GenBank, against a set of vector data retrieved from GenBank by use of key-word searches, such as "vector." In this report, we present evidence for another sequence artifact that is widespread but that, to our knowledge, has not yet been reported.

Recently, we examined the 5′ UTR of rat orphan receptor TR4, identified novel sequences, and determined the genomic organization of the 5′-UTR portion of the gene (Yoshikawa et al. 1996). During our analysis of the genomic DNA, we could not find the sequence that corresponded to the 14 nucleotides (5′-GAATTCGGCACGAG-3′) contained in the 5′ end of a previously reported rat TR4 cDNA (Chang et al. 1994). When we screened the GenBank database with this 14-nucleotide sequence, using the FASTA similarity search, we were surprised to find that hundreds of cDNA sequences from both eukaryotes and prokaryotes begin with this oligonucleotide fragment (table 1). Subsequently, we realized that this oligonucleotide was identical to the *Eco*RI adaptor sequence used in the ZAP libraries marketed by Stratagene. An adaptor is a short, double-stranded oligonucleotide containing a restriction-enzyme (RE) recognition site(s). It usually is ligated to both ends of the cDNA to permit incorporation of the cDNA into a cloning site of a vector. An adaptor usually carries one blunt end (which can be ligated to cDNA) and one cohesive terminus (which can be ligated to a compatible terminus in the vector). This experience led us to investigate the extent of the problem.

In this study, we took the sequences of *Eco*RI site-containing adaptors published in catalogs of various companies and surveyed them for matched cDNA sequences in GenBank release 94.0 (April 1996) and in European Molecular Biology Library (EMBL) release 46.0 (March 1996), using the FASTA program in the GCG package, with the parameter of word size fixed at four (interval of six). We displayed the top 1,000 hits and assorted them according to the following groups, which are shown in table 1: 5′ end versus 3′ end, year of submission, company selling the adaptor, and complete matches versus partial deletion of RE-site matches.

In table 1, the positive hits are tabulated from 1991 to 1996. Since all the hits appeared after the individual adaptors were put on the market by company representatives (see table 1), these figures seem to contain few fortuitous matches. The total number of cDNAs that have adaptor sequences in the 5′ end is 553, and that in the 3′ end is 175, for a total of 728 matches. The number of cDNAs in which an adaptor sequence is located somewhat internally—that is, the sequence is not an immediate extension of the extreme 5′ end or 3′ end but is located 2–100 nucleotides from an end—amounted to 144 (20%) of 728 total matches. In addition to 100% matches between adaptor and cDNA sequences, there was a sizable number in which the match began with part or all of the *Eco*RI site (GAATTC) deleted, as indicated in table 1. We believe that the contamination at the 3′ end of the cDNA originated from randomly primed cDNA libraries. (If an oligo(dT)-primed cDNA is used, the 3′ sequence has a poly(A) stretch upstream of an adaptor sequence, thus increasing the likelihood that the adaptor would not be included in the submission.) Among the different adaptor sequences that we examined, the contamination with Stratagene's ZAP library adaptor was most prominent. This may reflect the popularity of this library series but also could be the result of adaptor sequences in some libraries being unavailable for examination. It is also notable that the contamination with an adaptor sequence in the 5′ end of cDNA is increasing year by year (table 1).

We maintain that this adaptor contamination is an important artifact in the gene sequences in GenBank release 94.0 (April 1996) (Benson et al. 1996), for several reasons. (1) In this study, we have examined only some *Eco*RI site–containing adaptor sequences in cDNAs. If we had examined sequences originating from genomic clones, other adaptors, linkers, and anchor sequences used in RACE (rapid amplification of cDNA ends) PCR, the contamination likely would have turned out to be more widespread. (2) The sequence contamination primarily occurs at a crucial site—the 5′ end of the cDNA. The 5′ ends of cDNAs are especially important in the identification of

Table 1

cDNAs in the GenBank/EMBL Databases with End Sequences That Are Homologous to *Eco*RI Adaptors

Year Submitted and Type of Match	Stratagene: (G)AATTCGGCACGAG[a] (ZAP Library Adaptor[b])		Stratagene: (G)AATTCGCGGCCGC[a] (*Eco*RI*Not*I Adaptor[b])		Promega: (G)AATTCCGTTGCTGTCG[a] (*Eco*RI Adaptor[c])		GIBCO-BRL: (G)AATTCGCGGCCGCGTCGAC[a] (*Eco*RI/*Not*I Adaptor[d])		Clontech: (G)AATTCGCGGCCGCGTCGAC[a] (*Eco*RI Adaptor[e])	
	5′ end	3′ end	5′ end	3′ end	5′ end	3′ end	5′ end	3′ end	5′ end	3′ end
1991:										
Complete[f]	8	6	1	1	0	0	0	0	0	0
Minus *Eco*RI site[g]	1	0	0	0	0	0	1	0	0	0
1992:										
Complete[f]	25	5	7	4	3	2	0	0	0	0
Minus *Eco*RI site[g]	2	1	1	1	1	0	0	0	0	0
1993:										
Complete[f]	42	15	24	13	3	2	0	0	0	0
Minus *Eco*RI site[g]	11	1	2	1	2	2	0	0	0	0
1994:										
Complete[f]	46	11	13	14	8	2	0	0	1	0
Minus *Eco*RI site[g]	10	2	4	3	6	2	1	0	1	0
1995:										
Complete[f]	111	22	39	29	10	3	1	1	1	1
Minus *Eco*RI site[g]	38	2	8	2	5	7	4	0	5	0
1996 (to April):										
Complete[f]	50	5	21	11	2	0	5	1	3	1
Minus *Eco*RI site[g]	18	2	3	0	1	0	2	0	2	0
Total	362	72	123	79	41	20	14	2	13	2

[a] Sequence is 5′ to 3′ direction. Both forward and reverse sequences of the adaptors were matched against 5′ and 3′ ends of cDNAs, respectively. The "G" in parentheses on the 5′ end of the adaptors are not included in the manufacturers' sequences but were added in the search, because cDNA sequences that are identical to these sequences contain a "G" in the 5′ end.

[b] First available in 1990.

[c] First available in 1988.

[d] First available in 1991.

[e] First available in 1994.

[f] Number of sequences that are identical to the adaptor sequence.

[g] Number of sequences that are identical to the adaptor sequence minus a complete *Eco*RI site (GAATTC).

transcription-initiation sites and in the study of promoter regulation and gene organization. (3) When the rapid increase of sequence submission is considered, adaptor contamination likewise will increase rapidly until addressed. (4) Since very high sequencing accuracy is desired, even small sources of contamination can have large effects within the very low allowable-sequencing-error rate. (5) In terms of the actual number of sequences shown to be contaminated, this study has demonstrated a much larger number than have been seen in previous studies—for example, 45 sequences in the largest previous study (Lamperti et al. 1992) versus the 553 sequences that we found at 5′ ends.

One way to prevent these contaminations is to add a query about any adaptor sequence in the sequence-submission format of BankIt, which is the Internet submission program developed by NCBI. As shown in this report, the contaminations of adaptors comprise several categories and multiple sequences from multiple companies. We realize that, compared with vector sequences, adaptor contamination may be more difficult for the database center to eliminate automatically at the time of submission, because of possible random matches with the relatively short adaptor sequences. This problem of internal random matches could be addressed by filtering matches only in the first few nucleotides on each of the ends. In the absence of an automatic filtering mechanism as described above, inclusion of an adaptor sequence in a submission can be decreased only by the vigilance of each individual submitter.

We believe that alerting the scientific community with regard to the preponderance of adaptor sequences in the sequence databases could save time and effort. It is extremely important that artifactual sequences in the databases are recognized quickly, particularly since the Human Genome Project has spawned a tremendous growth in data submissions, arising from the flurry in sequencing of entire cDNA libraries, whole chromosomes, and whole genomes.

Takeo Yoshikawa, Alan R. Sanders,
and Sevilla D. Detera-Wadleigh

Unit on Gene Mapping and Expression
Clinical Neurogenetics Branch
National Institute of Mental Health
Bethesda

References

Anderson C (1993) Genome database worry about yeast (and other) infections. Science 259:1685

Benson D, Boguski M, Lipman DJ, Ostell J (1996) GenBank. Nucleic Acids Res 24:1–5

Binns M (1993) Contamination of DNA database sequence entries with *Escherichia coli* insertion sequences. Nucleic Acids Res 21:779

Chang C, da Silva SL, Ideta R, Lee Y, Yeh S, Burbach JPH (1994) Human and rat TR4 orphan receptors specify a subclass of the steroid receptor superfamily. Proc Natl Acad Sci USA 91:6040–6044

Dean M, Allikmets R (1995) Contamination of cDNA libraries and expressed-sequence-tags databases. Am J Hum Genet 57:1254–1255

Gersuk VH, Rose TM (1993) Database contamination. Science 260:605

Kessin RH, Van Lookeren Campagne MM (1993) Database contamination. Science 260:605

Lamperti ED, Kittelberger JM, Smith TF, Villa-Komaroff L (1992) Corruption of genomic databases with anomalous sequence. Nucleic Acids Res 20:2741–2747

Lopez R, Kristensen T, Prydz H (1992) Database contamination. Nature 355:211

Reynolds TL (1994) Vector DNA artifacts in the nucleotide sequence database. Biotechniques 16:1124–1125

Yoshikawa T, Makino S, Gao X-M, Xing G-Q, Chuang D-M, Detera-Wadleigh SD (1996) Splice variants of rat TR4 orphan receptor: differential expression of novel sequences in the 5′-untranslated region and C-terminal domain. Endocrinology 137:1562–1571

Address for correspondence and reprints: Dr. Takeo Yoshikawa, Unit on Gene Mapping and Expression, Clinical Neurogenetics Branch, National Institute of Mental Health, National Institutes of Health, Building 10, Room 3N218, Bethesda, MD 20892-1274. E-mail: squid@helix.nih.gov

0002-9297/97/6002-0028$02.00