Copyright © 1995 by the Genetics Society of America

Letter to the Editor

Poly(dC) Segments and Cloning Artifacts in Databases

SEQUENCE databases are useful tools for the scientific community. When new sequence data are obtained, an important part of the analysis relies on the comparison of this new sequence with all the sequences contained in the databases, either to find phylogenetic relationships or to identify functionally significant regions in it. For this reason, great care should be taken to avoid any database "contamination" with artifactual sequences since they could influence on the reliability of the data obtained. Some database contaminations have already been reported (*e.g.*, see LOPEZ *et al.* 1992; ANDERSON 1993). Here we report what we consider a new kind of contamination that could implicate ~500 sequences. These sequences contain poly(dC) segments that seem to have been wrongly assigned to them, and some of these sequences could be derived from cloning artifacts.

During the isolation of mesoderm specific cDNAs, we screened a Drosophila cDNA library constructed in λgt10 as described in HUYNH *et al.* (1985). We purified several clones and, as a first step in the analysis, we sequenced both ends of all of them. Our sequences were compared to all known DNA sequences by running the WORDSEARCH program from the University of Wisconsin GCG software package, and we found a sequence entry (accession no. X61180) showing a very abnormal pattern of similarity (see Figure 1): the first 227 nucleotides of that DNA were identical to our sequence, and the rest did not show any significant similarity to it. Furthermore, we noted that the junction between the homologous and the nonhomologous segment in X61180 was the sequence CCCCCCCCCC. These remarkable features prompted us to check further what could be the origin of this paradoxical pattern.

Sequence entry X61180 corresponds to a cDNA isolated during the cloning and analysis of the *l(1)ogre* locus (WATANABE and KANKEL 1990) and is derived from a library prepared by POOLE *et al.* (1985). For the preparation of that library, it is described that, after the synthesis of the first strand of the cDNA, a poly(dG) tail was created by terminal transferase tailing. Then, an oligo(dC) was used to prime the synthesis of the second strand. This library (as perhaps many other libraries) contains some clones that include DNA fragments that have been artificially ligated. There is at least one reported case of such cloning artifacts that were found during the analysis of other Drosophila cDNA derived from that library [see MATERIALS AND METHODS of SCHEUWLY *et al.* (1986)]. In this case, the cloning artifacts were detected because several cDNAs from the same gene were analyzed. Given these precedents, we considered that the sequence described by WATANABE and KANKEL (1990) could include a cloning artifact. To check this possibility we used the primers indicated in Figure 1 for two independent polymerase chain reac-

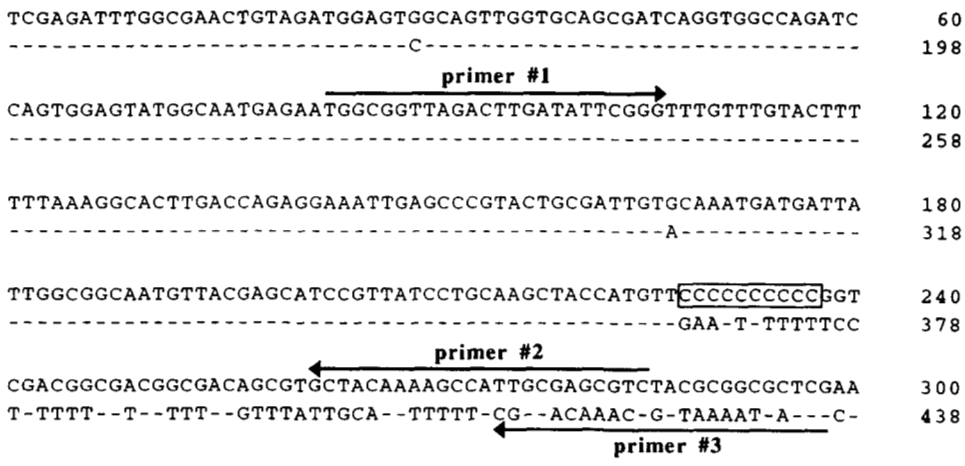

FIGURE 1.—Alignment of the published Drosophila *l(1)ogre* cDNA (top) and the sequence of one of our cDNAs (bottom). The coordinates of the *ogre* sequence are as described (WATANABE and KANKEL 1990). The boxed poly(dC) tract separates the homologous and nonhomologous regions between both sequences. Arrows represent the locations of the primers used for PCR amplification on *D. melanogaster* embryonic cDNA following standard procedures. PCR with primers #1 and #3 yields a product of the expected size (210 bp); no amplification is obtained when using primers #1 and #2, as expected if this sequence is not present in the cDNA preparation, confirming that it is the result of an artifactual ligation.

tion (PCR) amplifications. Our results confirm that the first 227 nucleotides of the sequence X61180 were artificially ligated to the cDNA of the Drosophila *l(1)ogre* locus.

After this finding, we considered that there might be other sequence entries in the databases with the same kind of artifacts, since the oligo(dC) approach has been used for the construction of several other libraries of different origins. These artifacts can be particularly important in the case of cDNA sequences because they are the basis to infer protein structures and to design further experiments. To test this, we screened the complete GenBank and EMBL databases for sequences containing poly(dC) segments longer than 10 nucleotides. This search shows that ~2000 sequences contain such segments. Of course, not all the sequences containing this segment do necessarily contain a cloning artifact. In many cases we have found cDNA sequences containing poly(dC) segments at the 5′ ends, usually following an *Eco*RI site. This strongly suggests that these segments are derived from the construction of the corresponding library and have been wrongly assigned to the sequence of the transcript. But cloning artifacts are by far more difficult to detect because we can not exclude the possibility that poly(dC) segments do exist in the genome. Thus, it would be very useful to check the validity of cDNA sequences containing internal poly(dC) [or poly(dG)] segments and derived from cDNA libraries made by the oligo(dC) priming method. In such cases, we propose that several cDNA clones should be characterized or that RT-PCR or genomic PCR should be performed, to deduce the primary structure of the mRNA. Cloning artifacts would be easily identified in this way.

NURIA PARICIO, JAVIER TEROL,
RUBÉN D. ARTERO and MANUEL PÉREZ-ALONSO
Department of Genetics
University of Valencia
46100 Burjasot, Spain

LITERATURE CITED

ANDERSON, C., 1993 Genome databases worry about yeast (and other) infections. Nature **259:** 1685.

HUYNH, T. V., R. A. YOUNG and R. W. DAVIS, 1985 Constructing and screening cDNA libraries in λgt10 and λgt11, pp. 49–78 in *DNA Cloning: A Practical Approach,* edited by D. M. GLOVER. IRL Press, Oxford.

LOPEZ, R., T. KRISTENSEN and H. PRYDZ, 1992 Database contamination. Nature **355:** 211.

POOLE, S. J., C. M. KAUVAR, B. DREES and T. KORNBERG, 1985 The *engrailed* locus of *Drosophila:* structural analysis of an embryonic transcript. Cell **40:** 37–43.

SCHNEUWLY, S., A. KUROIWA, P. BAUMGARTNER and W. J. GEHRING, 1986 Structural organization and sequence of the homeotic gene *Antennapedia* of *Drosophila melanogaster.* EMBO J. **5:** 733–739.

WATANABE, T., and D. R. KANKEL, 1990 Molecular cloning and analysis of *l(1)ogre,* a locus of *Drosophila melanogaster* with prominent effects on the postembryonic development of the central nervous system. Genetics **126:** 1033–1044.