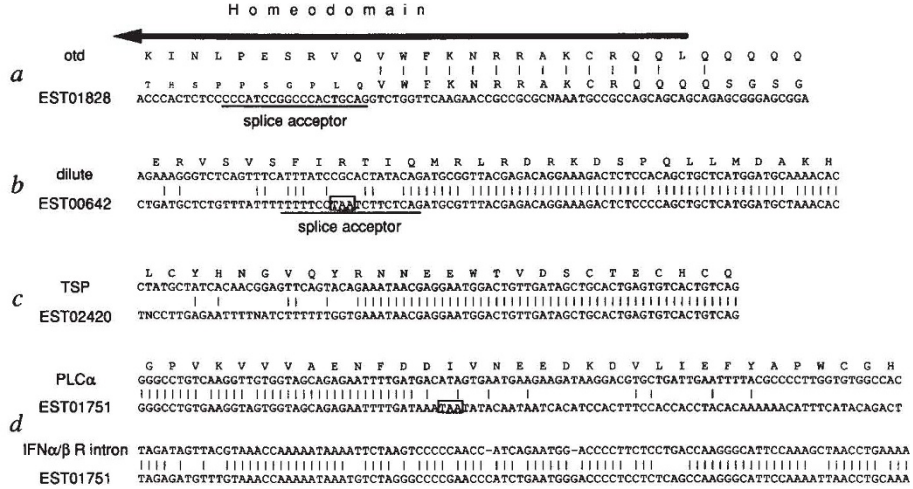
Introns in sequence tags

SIR — The publication of 'expressed sequence tags' (ESTs) derived from randomly selected clones from commercially obtained brain complementary DNA libraries[1,2] has received much attention, in part because of its promise to help define hitherto unknown functions encoded by the genome, but also because the genes so defined are now the subject of a patent application by the National Institutes of Health. Indeed, this decision has forced the hand of other agencies in charge of similar projects elsewhere[3], not to mention the continuing doubts about the patentability of such sequences. We have now discovered standard cloning artefacts in some of the published ESTs by examining a limited set, suggesting there may be many more.

Initially, we examined one EST of interest to us (EST01828, similar to the *Drosophila* homeobox gene *otd*; ref.4), and found that it appears to contain an intron. Subsequently, it seemed that there were several ESTs reported by Adams *et al.*[2] that showed a high degree of similarity to known genes (>90%) but over a region much smaller than the mean sequence length of an EST, suggesting abrupt interruptions of the similarity. Of seven such ESTs selected by visual inspection, we found that three seem to contain cloning artefacts (see figure). One terminates its similarity at what appears to be an intron (splice acceptor) and two have apparently unrelated sequences (probably noncoding) joined to a smaller segment containing the region of similarity (see figure). One of these unrelated sequences seems to be a member of a repetitive sequence family. Adams *et al.*[2] state that they found no evidence that genomic DNA or unspliced cDNAs were major contaminants of their commercially obtained libraries; but given the data presented here we wonder how many entries in the full list are in fact genomic, unspliced or otherwise corrupted. It would be difficult to estimate this without exhaustive analysis, especially for entries with no sequence similarity, and it would not be an easy problem to solve using computer analysis alone. At least the presence of a poly(A) tail could establish whether a clone is transcribed.

Clones bearing introns may be either cDNAs from unspliced mRNA precursors, or genomic DNA. Certainly, cDNA libraries often contain genomic DNA clones as contaminants. There is thus a finite probability that any of the ESTs are not expressed at all (let alone in the brain). Perhaps, therefore, patent applications and the 'EST' label should be reserved for clones that have been better characterized.

Thomas R. Bürglin
Thomas M. Barnes
Department of Molecular Biology,
Massachusetts General Hospital and Department of Genetics,
Harvard Medical School,
Wellman 8, MGH, Boston,
Massachusetts 02114, USA



Comparison of ESTs with their reported[2] homologues. *a*, EST01828 compared with *Drosophila otd* (ref. 4). *b*, EST00642 compared with mouse *dilute* (accession no. X57377). *c*, EST02420 compared with human thrombospondin (accession no. M14326). *d*, EST01751 compared with mouse phospholipase C α (accession no. M73329) and a repeated sequence in the first intron of the human interferon α/β receptor gene (accession no. X60459).

a, The sequence of EST01828 with its conceptual translation in the region of similarity to *otd* is shown. Of all homeodomains in the database, EST01828 best matches *otd*. The similarity drops off markedly upstream of the VWF motif, and the conceptual translation product is in small capitals (thick bar, *otd* homeodomain). A consensus splice acceptor is present at the region of divergence. Other members of the *prd*-like class of homeodomains, for example *unc-4* (ref. 5) and *ceh-10* (ref. 6), have an intron in the same position. The sequence of EST01828 further upstream has stop codons and proline residues in all three reading frames in the region where helix 2 might be expected (not shown), which is incompatible with the structure of the homeodomain. *b*, The 286 bp of the EST show similarity to dilute over the last 86 bp (reverse orientation). The genomic structure of the mouse *dilute* gene in this region is not known, but EST00642 appears to be the human homologue with an unspliced intron or a genomic clone: a consensus splice acceptor occurs just at the point of sequence divergence, and the upstream sequence contains an in-frame stop codon (boxed). The putative intron has no significant similarity to any other DNA or protein sequence. *c*, The 333 bp of the EST show identity to human thrombospondin over the last 83 bp, except for one frameshift (probably a sequencing error). Before this, the sequences diverge at what does not seem to be a splice site. The nonidentical sequence is 70% (A+T), while the region of identity is close to 50%. It is thus probably noncoding, and the EST probably represents a co-ligation artefact of the cDNA library. The nonidentical sequence has no significant similarity to any other DNA or protein sequence. *d*, The 380 bp EST shows similarity to two sequences. The reported similarity to PLCα occurs over the first 126 bp of the EST. Again, the sequence diverges at what does not appear to be a splice junction. The remainder of the sequence has approximately 70% (A+T) content. Part of this sequence (the last 100 bp) is clearly a member of a human repetitive sequence element, found twice in the first intron of the interferon α/β receptor gene. Because it is only 79% identical, it does not originate from this locus, but presumably from some other dispersed location. Thus the clone is probably a co-ligation artefact. The sequence between these two similarities in the EST shows no further similarity to any DNA or protein sequence.

1. Adams, M. D. *et al. Science* **252**, 1651–1656 (1991).
2. Adams, M. D. *et al. Nature* **355**, 632–634 (1992).
3. Aldhous, P. *Nature* **356**, 98 (1992).
4. Finkelstein, R. *et al. Genes Dev.* **4**, 1516–1527 (1990).
5. Miller, D. M. *et al. Nature* **355**, 841–845 (1992).
6. Hawkins, N. C. & McGhee, J. D. *Nucleic Acids Res.* **18**, 6101–6106 (1990).

ADAMS *ET AL.* REPLY — The quality of a cDNA library is usually assessed pragmatically by whether it contains a particular clone of interest. As a result, incompletely spliced cDNAs and mitochondrial, ribosomal and polyadenylate clones are generally not a hindrance to use of a library. However, any cDNA library constructed from RNA isolated from whole cells, as opposed to the cytoplasm alone, will contain some cDNA to incompletely spliced mRNA, as well as mitochondrial and ribosomal transcripts. Library contamination by genomic DNA is rare, provided sufficient DNase treatment is included in the initial RNA purification step.

In general, a cDNA clone can be conclusively identified as either incompletely spliced or chimaeric only by complete sequencing and direct comparison to mRNA by hybridization, S1 nuclease protection or PCR, together with characterization of all expressed splicing patterns. Sequences that match the consensus patterns for splice sites are very common in exon, intron and intergenic DNA[1]; their presence in a cDNA does not provide strong evidence that it is incompletely spliced. While intron positions are conserved in some genes among mammals, they are not conserved universally. We have searched our EST data for known intron sequences, and