# A comparison of expressed sequence tags (ESTs) to human genomic sequences

**Tyra G. Wolfsberg and David Landsman***

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, Room 8N-807, Bethesda, MD 20894, USA

Received October 11, 1996; Revised and Accepted February 18, 1997

## ABSTRACT

**The Expressed Sequence Tag (EST) division of GenBank, dbEST, is a large repository of the data being generated by human genome sequencing centers. ESTs are short, single pass cDNA sequences generated from randomly selected library clones. The ~415 000 human ESTs represent a valuable, low priced, and easily accessible biological reagent. As many ESTs are derived from yet uncharacterized genes, dbEST is a prime starting point for the identification of novel mRNAs. Conversely, other genes are represented by hundreds of ESTs, a redundancy which may provide data about rare mRNA isoforms. Here we present an analysis of >1000 ESTs generated by the WashU-Merck EST project. These ESTs were collected by querying dbEST with the genomic sequences of 15 human genes. When we aligned the matching ESTs to the genomic sequences, we found that in one gene, 73% of the ESTs which derive from spliced or partially spliced transcripts either contain intron sequences or are spliced at previously unreported sites; other genes have lower percentages of such ESTs, and some have none. This finding suggests that ESTs could provide researchers with novel information about alternative splicing in certain genes. In a related analysis of pairs of ESTs which are reported to derive from a single gene, we found that as many as 26% of the pairs do not BOTH align with the sequence of the same gene. We suspect that some of these unusual ESTs result from artifacts in EST generation, and caution researchers that they may find such clones while analyzing sequences in dbEST.**

## INTRODUCTION

Discovery of novel genes has traditionally been a laborious task requiring months or even years of work at the bench. In the current era of large scale genome sequencing, however, identifying new genes can require as little time as the few minutes it takes to perform a computer-driven search of a sequence database (1). The database with the highest rate of growth has been a division of GenBank called dbEST, the database of Expressed Sequence Tags (2) (http://www.ncbi.nlm.nih.gov/dbEST). Expressed sequence tags, or ESTs, are short sequences, a few hundred base pairs in length, which are derived by partial, single pass sequencing of the inserts of randomly selected cDNA clones (3). Although ESTs from many organisms, including mouse, rice, *Arabidopsis thaliana* and *Caenorhabditis elegans* are all present in dbEST, we focus in this report on the human ESTs (4), which comprise at present ~75% of the sequences in dbEST. It has been estimated that 40–80% of the total number of human genes are represented in dbEST (M.S.Boguski, personal communication). The generation of these human ESTs was seen as a crucial step in the progress of the human genome project (5). In fact, ESTs have enabled the recent mapping of 16 000 genes in the human genome (6).

The Washington University Genome Sequencing Center, under contract with Merck & Co., has produced 76% of the human ESTs in dbEST to date. The WashU-Merck EST project uses oligo(dT)-primed, directionally-cloned cDNA libraries from a variety of human tissues (4). Many of these libraries are normalized to bring the frequency of occurrence of both highly expressed and rare messages within a narrow range (7). Clones are selected randomly from these libraries, and their cDNA inserts are sequenced at the 5′ and 3′ ends. Many, but not all, inserts are sequenced from both ends, yielding two ESTs for each clone. The EST derived from the 3′ end of the insert often aligns with sequence in the 3′ untranslated region of the gene, while that from the 5′ end of the insert derives from sequence further upstream. Depending on the length of the clone insert, the two ESTs may overlap. Some genes are represented by only one EST, whereas others, such as serum albumin, are represented by more than 1000 (6). However, because of normalization and other technical aspects, the frequency of representation of a gene in dbEST should not be used to predict its expression level (8).

It is not surprising that, due to the protocols used for their rapid generation, ESTs can contain sequence and annotation inaccuracies. Most ESTs are generated by automated fluorescent sequencing methods and consist of a single read of one strand of the cDNA. Although bacterial, mitochondrial and vector sequences are removed from the dataset before it is submitted to dbEST, little manual editing of individual sequences is performed (4). Thus, some EST sequences are of low quality, and contain unknown or incorrect nucleotides, insertions or deletions, particularly near the trailing (distal) end of a sequencing read. Furthermore, although cDNAs are directionally cloned into the vectors, it has been reported that some inserts appear in the reverse orientation (4,9). Some of these sequences may represent undocumented transcription

*To whom correspondence should be addressed. Tel: +1 301 496 2477; Fax: +1 301 480 9241; Email: landsman@ncbi.nlm.nih.gov of the complementary DNA strand. In other cases, the annotation which indicates whether the EST derives from the 5′ or 3′ end of the clone insert may provide incorrect information about the orientation of the EST on the cDNA itself.

dbEST currently contains ~415 000 partial human cDNA sequences which derive both from known and from yet-uncharacterized human mRNAs. In contrast, the non-EST portion of GenBank, which includes cDNA, genomic DNA and RNA sequences derived by traditional functional and positional cloning methods, contains ~54 000 human sequences. The UniGene project, the result of large-scale sequence comparisons among selected human ESTs and known genes, has grouped these sequences into ~50 000 clusters, or likely genes. The majority of these potential genes (~91%) are represented only by ESTs (6). Motivated by the large number of previously undiscovered cDNAs present in dbEST, investigators are turning to this database as a powerful tool to identify novel genes (4,5,10–13). In this study, we show that ESTs may also be a useful tool for analyzing the splicing patterns of previously characterized cDNAs, as we have found a significant number of intron sequences in dbEST. However, we caution that some data in dbEST may contain annotation errors, as we have noted that certain pairs of ESTs, which are annotated as deriving from the same cDNA clone, are likely the product of two different genes.

## MATERIALS AND METHODS

### Identification of ESTs

We used the Entrez browser (14) to identify 15 full length human genomic and cDNA sequences which are represented in dbEST by five or more ESTs from the WashU-Merck EST Project. The gene names and accession numbers of the genomic DNAs are as follows. Desmin: M63391; Corticotropin releasing factor (Corticoliberin): V00571; Osteopontin: U20758; Aldolase C: X05196/X07292; Vitronectin: X05006; Alpha-fetoprotein (AFP): M16110; HMG-14: M21339; HMG-17: X13546; Glutathione *S*-transferase (GST): X08058; Lecithin-cholesterol acyltransferase (LCAT): X04981; Ornithine decarboxylase (ODC): X16277; Splicing factor, arginine/serine-rich 7 (SFRS7): L41887; Cytochrome P450IIE1 (Cyt p450): J02843; Cystatin B: U46692; Serum albumin: M12523. We masked Alu and other repetitive elements present in these genomic sequences by performing a BLASTN (15) search with each sequence against the Alu database at the NCBI, and then using XBLAST (16) to filter out the Alu and other repeats from the genomic sequence. We identified the ESTs which derived from these 15 genes by performing BLASTN searches of dbEST with the masked genomic DNA sequence. We selected all hits with a *P* value of $<10^{-87}$, and hand selected additional matching ESTs with higher *P* values. We sorted the ESTs from each gene by clone number. If both the 5′ and 3′ ESTs from a clone were not obtained in the BLASTN search, we retrieved the missing EST from dbEST. In some cases, however, the missing EST was not available from the database. Only those clones which were represented by two ESTs, one from the 5′ end of the insert and one from the 3′ end of the insert, were analyzed.

### Characterization of ESTs

ESTs were scored for their alignment with other sequences as well as for their splicing patterns. ESTs were aligned to each other and to the full length genomic and cDNA sequences from which they derived using the gapped alignment programs Sequencher (Gene Codes Corporation) and sim2aln (17). Splicing patterns were analyzed with the gapped alignment program tsim (18) with a gap extension penalty of 0. Alignments generated by sim2aln and tsim were viewed using the program musk/chromoscope (19). We performed additional BLASTN searches against dbEST and nr (the non-redundant set of GenBank, EMBL and DDBJ database sequences) with all unmatching ESTs, i.e., those which do not align with the genomic or cDNA sequences. All data presented in this paper are current as of June, 1996. Specific data are available from the authors upon request.

## RESULTS AND DISCUSSION

### Classification of normal and aberrant clones in dbEST

While scanning dbEST for new members of gene families, we identified a set of unusual clones. These clones were represented by two ESTs, one derived from sequence at the 5′ end of the insert, and one from sequence at the 3′ end. One EST of the pair aligned well with previously characterized sequences, while the other did not. As these discrepancies did not appear to result from simple sequencing errors, we were curious whether the clones represented novel members of the gene family, contained intron sequences, or were the result of sequencing artifacts. We thus undertook a more thorough analysis of a sample of clones in dbEST by characterizing ESTs whose ‘correct’ sequence could be easily verified, that is, ESTs which derived from genes whose full-length genomic sequences had already been deposited into GenBank. We identified the ESTs by performing a sequence similarity search using the genomic DNA sequence to query dbEST. Since we were interested in determining whether the two ESTs from a given clone derived from the same gene, we analyzed only those clones whose inserts had been sequenced at both the 5′ and 3′ ends. In some cases, the EST representing the 5′ or 3′ end of an insert did not appear in the sequence similarity search. We retrieved those sequences from dbEST using the clone number. We limited our search to sequences generated by the WashU-Merck project, as these ESTs are well annotated and represent 76% of the human ESTs in dbEST. We examined 545 clones (i.e., 1090 ESTs), from 15 different human genes, in which at least one EST from the clone overlapped with known genomic sequence.

We studied the 1090 ESTs by comparing their sequence to the sequence of the parental genomic DNA. We found that the ESTs could be grouped into two categories, matching and unmatching. ***Matching*** ESTs align with the parental genomic sequence, while ***unmatching*** ESTs align either with other sequences in the database or with nothing, but not with the genomic sequence. We observed four different types of matching ESTs, types A–D. In the example shown in Figure 1, matching ESTs align with the genomic sequence of the ‘blue’ gene. We say that matching ESTs of types A and B are of a ***known transcript type***, since they are made up of sequences which are known to be transcribed. Type A matching ESTs show no evidence of splicing, as they align with sequences in the middle of a single exon. Matching ESTs of type B, which span two or more exons, are derived from mRNAs which have been spliced at previously documented intron–exon boundaries. Conversely, we use the term ***unreported transcript type*** to refer to matching ESTs which contain intron sequence(s) and/or novel intron–exon boundaries (types C and D). Type C

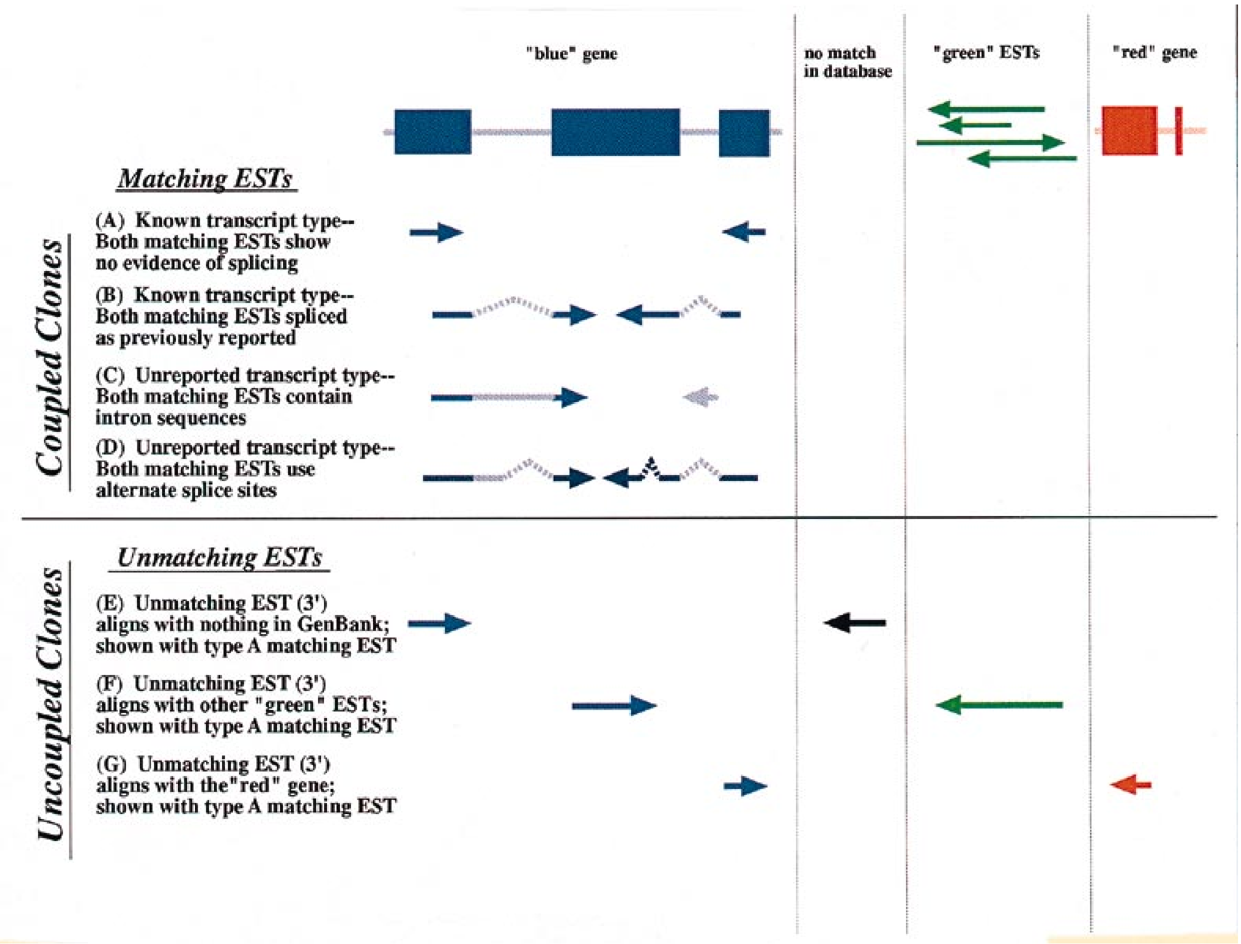
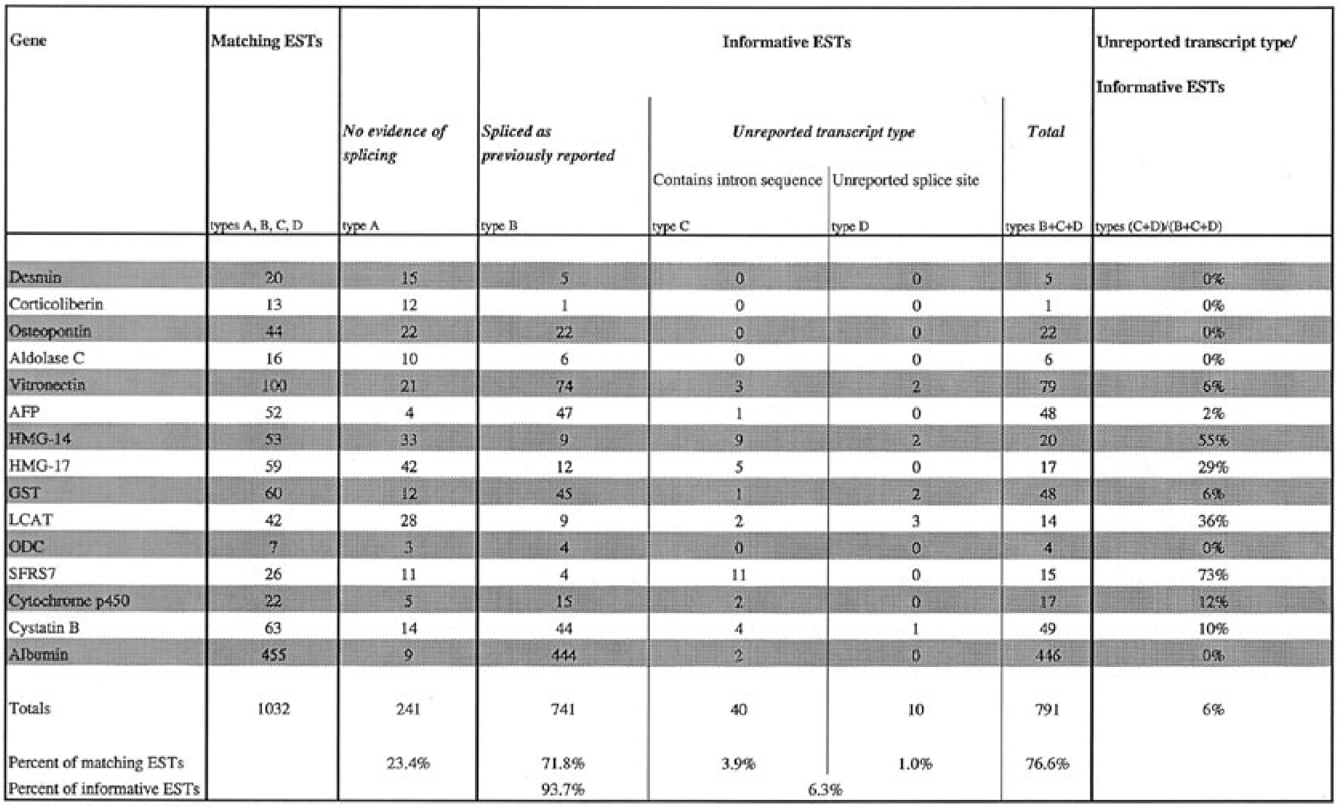

"blue" gene
no match in database
"green" ESTs
"red" gene
Coupled Clones
Matching ESTs
(A) Known transcript type-- Both matching ESTs show no evidence of splicing
(B) Known transcript type-- Both matching ESTs spliced as previously reported
(C) Unreported transcript type-- Both matching ESTs contain intron sequences
(D) Unreported transcript type-- Both matching ESTs use alternate splice sites
Uncoupled Clones
Unmatching ESTs
(E) Unmatching EST (3') aligns with nothing in GenBank; shown with type A matching EST
(F) Unmatching EST (3') aligns with other "green" ESTs; shown with type A matching EST
(G) Unmatching EST (3') aligns with the "red" gene; shown with type A matching EST

**Figure 1.** Schematic illustration of the types of ESTs and clones observed in dbEST. ESTs from the WashU-Merck EST Project which share sequence identity with a previously reported genomic DNA sequence (in this example, from the 'blue' gene) were identified as described in the Materials and Methods. The cDNA library clones from which the ESTs derived were determined. ESTs which derived from the same clones, but which did not appear in the BLAST search, were retrieved from dbEST. Only those clones which were represented by two ESTs, one from the 5′ end of the insert and one from the 3′ end of the insert, were analyzed. Library clones were classified into two groups, coupled and uncoupled. Coupled clones are composed of two matching ESTs which both derive from sequence in the 'blue' gene. Uncoupled clones are composed of one matching EST, which derives from sequence in the 'blue' gene, and one unmatching EST, which does not align with sequence in the 'blue' gene. Previously reported exons in the blue gene are shown in dark blue, previously reported introns are in light blue, and sequences removed by splicing are shown as dotted lines. There are four types of matching ESTs: (A) matching EST aligns with the sequence from one exon of the 'blue' gene; (B) matching EST derives from at least two exons of the 'blue' gene which have been spliced as previously reported; (C) matching EST contains intron sequence from the 'blue' gene, either alone or in conjunction with exon sequence; (D) matching EST contains a splice junction which has not been previously reported for the 'blue' gene. Coupled clones can contain any combination of types of matching ends. There are three types of unmatching ESTs: (E) unmatching EST (black) aligns to no sequences in the sequence databases; (F) unmatching EST shares significant sequence similarity with other ESTs ('green' ESTs) ; (G) unmatching EST shares sequence identity with a different, previously characterized gene (the 'red' gene). Although all unmatching ESTs shown in this example are 3′ ESTs, in our analysis of 15 human genes (Table 2), unmatching ESTs derived from both the 5′ and the 3′ ends of clone inserts.

ESTs contain sequences previously described as introns. Some appear to be unspliced as they contain only intron sequences, or intron sequences along with adjacent exons. Others appear to be partially spliced; for example, one intron sequence is present, while another is absent. Type D ESTs contain a previously unreported splice site. Most contain a fragment of an intron placed between two exons, but in a few cases, sequences within the middle of an exon are spliced out. We hesitate to call any of these splice sites alternative before their existence is confirmed experimentally. We observed three different types of unmatching ESTs (types E–G), ESTs which do not align with the genomic sequence (Fig. 1). A type E unmatching EST, shown in black, is a unique sequence that does not share a significant stretch of sequence identity with any other human sequence in GenBank. A type F unmatching EST shares sequence identity with one or more other ESTs, shown in green, but not with any characterized sequences. A type G unmatching EST shares sequence identity with an independent, well characterized cDNA in GenBank, in this case, the 'red' gene.

Based on our characterization of the individual ESTs as matching or unmatching, we were able to classify clones as coupled or uncoupled. We use the term ***coupled*** to refer to clones which are made up of two matching ESTs, that is, both ESTs align with the genomic sequence. We use the term ***uncoupled*** to refer to clones in which one EST is matching and one is unmatching, that is, one EST aligns with the genomic sequence and one does not. All four types of matching ESTs (A–D) were observed in coupled clones. Theoretically, the matching end of an uncoupled

**Table 1.** Analysis of splicing of all matching ESTs from coupled and uncoupled clones



| Gene | Matching ESTs | | Informative ESTs | | | | Unreported transcript type/ Informative ESTs |
|---|---|---|---|---|---|---|---|
| | | *No evidence of splicing* | *Spliced as previously reported* | *Unreported transcript type* | | *Total* | |
| | | | | Contains intron sequence | Unreported splice site | | |
| | types A, B, C, D | type A | type B | type C | type D | types B+C+D | types (C+D)/(B+C+D) |
| Desmin | 20 | 15 | 5 | 0 | 0 | 5 | 0% |
| Corticoliberin | 13 | 12 | 1 | 0 | 0 | 1 | 0% |
| Osteopontin | 44 | 22 | 22 | 0 | 0 | 22 | 0% |
| Aldolase C | 16 | 10 | 6 | 0 | 0 | 6 | 0% |
| Vitronectin | 100 | 21 | 74 | 3 | 2 | 79 | 6% |
| AFP | 52 | 4 | 47 | 1 | 0 | 48 | 2% |
| HMG-14 | 53 | 33 | 9 | 9 | 2 | 20 | 55% |
| HMG-17 | 59 | 42 | 12 | 5 | 0 | 17 | 29% |
| GST | 60 | 12 | 45 | 1 | 2 | 48 | 6% |
| LCAT | 42 | 28 | 9 | 2 | 3 | 14 | 36% |
| ODC | 7 | 3 | 4 | 0 | 0 | 4 | 0% |
| SFRS7 | 26 | 11 | 4 | 11 | 0 | 15 | 73% |
| Cytochrome p450 | 22 | 5 | 15 | 2 | 0 | 17 | 12% |
| Cystatin B | 63 | 14 | 44 | 4 | 1 | 49 | 10% |
| Albumin | 455 | 9 | 444 | 2 | 0 | 446 | 0% |
| Totals | 1032 | 241 | 741 | 40 | 10 | 791 | 6% |
| Percent of matching ESTs | | 23.4% | 71.8% | 3.9% | 1.0% | 76.6% | |
| Percent of informative ESTs | | | 93.7% | 6.3% | | | |

The full length genomic sequences of these 15 human genes were retrieved using Entrez. The accession numbers and full gene names are listed in the Materials and Methods. The ESTs associated with these genes, either as matching or unmatching ESTs, were obtained as described in the Materials and Methods. The matching ESTs from both coupled and uncoupled clones were then classified as described in Figure 1. This table documents the observed types and numbers of matching ESTs. We report novel splicing patterns only if they occur within the coding region of the gene. Less than one third of the ESTs from serum albumin were classified.

clone could also be any of the four types. In practice, however, all uncoupled clones had matching ESTs of type A or B.

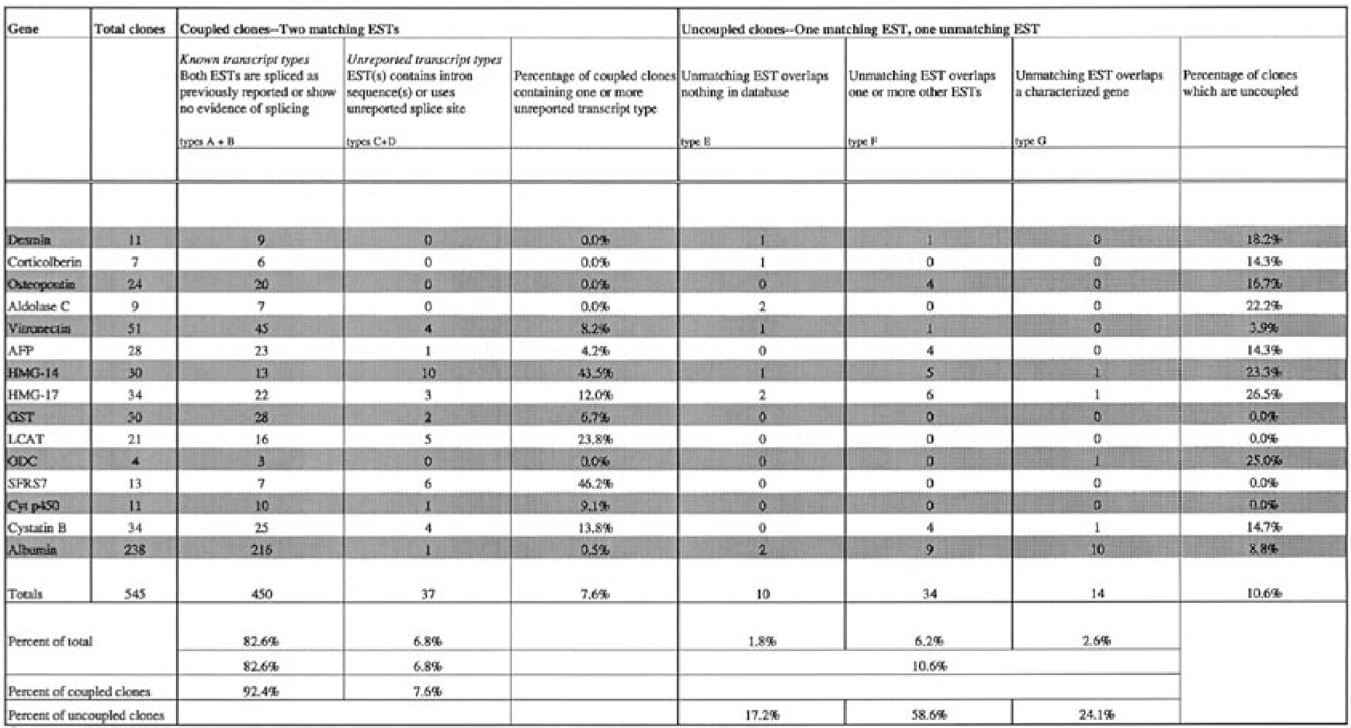
## Identification of novel mRNA splice patterns in dbEST

The first phase of our analysis of clones in dbEST was to classify the individual matching ESTs from the 545 coupled and uncoupled clones according to the four types of splicing patterns described in Figure 1. Our analysis of the 1032 matching ESTs is shown in Table 1.

As we were unsure whether transcript splicing at the extreme 5′ and 3′ ends of genes had been fully documented, we counted only those splicing events which occurred within the coding sequence. Overall, 23% of the ESTs show no evidence of splicing (type A matching ESTs in Fig. 1), although further sequencing of these clones would allow for better categorization. This percentage varies widely between genes because of differences in sizes and numbers of exons. For example, in corticoliberin, which has only two exons, 92% of the ESTs show no evidence of splicing, but in serum albumin, which has 14 exons, only 2% of the ESTs show no evidence of splicing. For the purpose of analyzing splicing, the more informative ESTs are the 77% which provide some information about intron usage, that is, matching ESTs of types B, C and D. Although 94% of these informative ESTs are spliced in ways that have been previously reported in GenBank (type B), 6% are transcript types that were previously unreported (types C and D). Of these unreported transcript types, 80% of the ESTs contain intron sequences, and appear to be either unspliced or partially spliced (type C), and 20% are spliced, but use splice sites which were not previously reported for that gene (type D). We note that none of the unreported transcript types derive from ESTs which have either remarkably low quality sequence or shorter than average length for dbEST entries, and thus these sequences are not likely due to sequencing or library artifacts.

While Table 1 shows a detailed distribution of the numbers of known and unreported transcript types among the matching ESTs, the left columns of Table 2 present an overview of the numbers of known and unreported transcript types among the coupled clones. In 92% of the coupled clones, both ESTs are of a known transcript type (types A or B). However, in 8% of the coupled clones, one or both of the matching ESTs is an unreported transcript type (types C or D). In a separate analysis of 755 clones from dbEST, Hillier *et al.* found that 0.53–2.25% of EST clones derive from intronic or intergenic sequences (4). As these authors examined ~40% more clones than we did, the range they report may reflect a more global average of the number of unreported transcript types in dbEST. However, Hillier *et al.* used different criteria to analyze their clones, and may have missed some intronic or intergenic sequences because they made the assumption that if two or more ESTs match the same genomic DNA sequence, these ESTs derive from mRNA. We did not make a similar assumption; in fact, we show that this hypothesis is likely incorrect (see below).

**Table 2.** Analysis of coupled and uncoupled clones based on types of matching and unmatching ESTs



| Gene | Total clones | Coupled clones--Two matching ESTs | | | Uncoupled clones--One matching EST, one unmatching EST | | | |
|---|---|---|---|---|---|---|---|---|
| | | *Known transcript types* Both ESTs are spliced as previously reported or show no evidence of splicing<br>types A + B | *Unreported transcript types* EST(s) contains intron sequence(s) or uses unreported splice site<br>types C+D | Percentage of coupled clones containing one or more unreported transcript type | Unmatching EST overlaps nothing in database<br>type E | Unmatching EST overlaps one or more other ESTs<br>type F | Unmatching EST overlaps a characterized gene<br>type G | Percentage of clones which are uncoupled |
| Desmin | 11 | 9 | 0 | 0.0% | 1 | 1 | 0 | 18.2% |
| Corticolberin | 7 | 6 | 0 | 0.0% | 1 | 0 | 0 | 14.3% |
| Osteopontin | 24 | 20 | 0 | 0.0% | 0 | 4 | 0 | 16.7% |
| Aldolase C | 9 | 7 | 0 | 0.0% | 2 | 0 | 0 | 22.2% |
| Vitronectin | 51 | 45 | 4 | 8.2% | 1 | 1 | 0 | 3.9% |
| AFP | 28 | 23 | 1 | 4.2% | 0 | 4 | 0 | 14.3% |
| HMG-14 | 30 | 13 | 10 | 43.5% | 1 | 5 | 1 | 23.3% |
| HMG-17 | 34 | 22 | 3 | 12.0% | 2 | 6 | 1 | 26.5% |
| GST | 30 | 28 | 2 | 6.7% | 0 | 0 | 0 | 0.0% |
| LCAT | 21 | 16 | 5 | 23.8% | 0 | 0 | 0 | 0.0% |
| ODC | 4 | 3 | 0 | 0.0% | 0 | 0 | 1 | 25.0% |
| SFRS7 | 13 | 7 | 6 | 46.2% | 0 | 0 | 0 | 0.0% |
| Cyt p450 | 11 | 10 | 1 | 9.1% | 0 | 0 | 0 | 0.0% |
| Cystatin B | 34 | 25 | 4 | 13.8% | 0 | 4 | 1 | 14.7% |
| Albumin | 238 | 216 | 1 | 0.5% | 2 | 9 | 10 | 8.8% |
| Totals | 545 | 450 | 37 | 7.6% | 10 | 34 | 14 | 10.6% |
| Percent of total | | 82.6% | 6.8% | | 1.8% | 6.2% | 2.6% | |
| | | 82.6% | 6.8% | | 10.6% | | | |
| Percent of coupled clones | | 92.4% | 7.6% | | | | | |
| Percent of uncoupled clones | | | | | 17.2% | 58.6% | 24.1% | |

The full length genomic sequences of these 15 human genes were retrieved using Entrez. The accession numbers and full gene names are listed in the Materials and Methods. The ESTs associated with these genes, either as matching or unmatching ESTs, were obtained as described in the Materials and Methods. The ESTs, as well as the clones from which they were derived, were then classified as described in Figure 1. This table documents the number of clones represented by the indicated type of matching and unmatching ESTs. The matching ends of all uncoupled clones were of types A or B.

Although the total number of unreported transcript types in dbEST may be low, the differences between the numbers of unreported transcript types for individual genes is particularly striking (Table 1). No unreported transcript types were found for six of the 15 genes, and only 2–15% of spliced ESTs from another five genes have undocumented splice patterns. However, among the four remaining genes, SFRS7, HMG-14, LCAT and HMG-17, 73%, 55%, 36% and 29%, respectively, of the transcripts which provide information about intron usage are of a previously unreported type. It is important to note, however, that not all of the ESTs represent individual transcripts. For example, in SFRS7, the 11 ESTs which contain intron sequences derive from only six different clones, or cDNAs, while in HMG-14, the 11 unreported transcript types derive from 10 different clones (Table 2). Furthermore, the sequence of certain clones may appear to be duplicated in dbEST, as we found five examples in which pairs of clones from the same library had nearly identical sequences. These putative duplicate clones could result from amplification procedures used during library construction, or because some library clones were inadvertantly sequenced more than once.

The existence of these unreported transcript types allows us to speculate on some interesting biological implications. We note that the unreported transcript types were observed in 11 different cDNA libraries, so the phenomenon is not limited to a single, error-prone library. All or many of these ESTs may represent real, but rare transcript types which have not been previously identified because of their low abundance. Such transcripts could encode for proteins with alternate sequence. Support for this hypothesis comes from the finding that in some genes, particular intron sequences are found in multiple ESTs. For example, four ESTs from HMG-14 include the same intron sequence. It is puzzling and unfortunate that none of the unreported transcript types from any of the 15 genes contains a strikingly long alternate open reading frame (data not shown). However, since the full length sequences of the clones are not present in dbEST, a detailed analysis of all different open reading frames is not possible. Our preliminary analysis also raises the alternate, intriguing possibility that cells may contain a number of transcripts which lack an open reading frame and thus do not code for protein. In a traditional small scale sequencing project, cDNAs lacking open reading frames might be classified as library preparation or sequencing artifacts and the data would be ignored. Any potential ‘unusual’ sequences are much more likely to be detected in a highly redundant, unedited database such as dbEST. Additionally, the unreported transcript types may provide information on the rate of splicing of certain genes. One striking observation is that some genes have high proportions of unspliced transcripts, while others have none. Thus, perhaps some genes are spliced to completion more quickly than others. Furthermore, within an individual gene, there may be a preference for the removal of certain introns, as some intron sequences are more likely than others to be represented by an EST (data not shown). This observation could reflect the order with which introns are excised from the nuclear hnRNA transcript. Neither the location nor the size of the intron appears to have an effect on its appearance in dbEST.

A somewhat less interesting explanation for the high number of unreported transcript types is that the libraries being used to generate ESTs contain unspliced cDNA or genomic DNA. We consider this a less likely possibility, as the unreported transcript types derive from 11 of the 26 libraries used by the WashU-Merck project for EST generation, and these 11 libraries were generated by at least four of the methods described in (7). However, perhaps some common step of the mRNA preparation technique is leading to an excessive contamination of cytoplasmic mRNA by genomic DNA or nuclear hnRNA.

## Identification of uncoupled clones in dbEST

The second phase of our analysis of clones in dbEST was to classify the unmatching ESTs from the uncoupled clones according to the three categories described in Figure 1. Table 2 documents our analysis of 545 clones. As we found 58 unmatching ESTs among the 15 genes, a total of 11% of the clones are uncoupled. None of the uncoupled clones also contain an unreported transcript type. Overall, in 17% of the uncoupled clones, the unmatching EST did not share significant sequence similarity with any other sequence in GenBank (type E). As most of these ESTs are composed of a few hundred base pairs of high quality sequence data, this lack of similarity is not likely due to errors in the sequences themselves. These uncoupled clones are the most innocuous as the unmatching EST might actually derive from yet unsequenced 5′ or 3′ untranslated regions. A total of 59% of the uncoupled clones have an unmatching EST which shares sequence identity with other ESTs, but not with any full length sequence in GenBank (type F). These clones are more problematic. Again, the unmatching ends could derive from unsequenced 5′ or 3′ untranslated regions, especially in cases where the ends overlap with few other ESTs. However, we think it is more likely that such unmatching ends, especially those that overlap with many ESTs, are actually derived from independent but uncharacterized genes. The unmatching ends of 24% of the uncoupled clones appear to be derived from a previously characterized gene (type G). The matching and unmatching ends of these clones likely derive from two independent genes, and the reporting of the two ESTs as ends of a single clone is probably due to annotation errors or cloning artifacts (see below). In a few cases, AFP, HMG-14, HMG-17 and Cystatin B, two type F unmatching ESTs align with each other as well as with other ESTs, and in albumin, five type G unmatching ESTs align with α-globin (not shown).

A number of the uncoupled clones which we observed are probably due to technical complexities. ESTs could receive incorrect clone numbers if samples were switched during preparation, or if human error led to the entering of incorrect information into the database which tracks the association between EST and clone number. Errors in lane tracking during sequencing gel runs may also be responsible for some of the uncoupled clones. Uncoupled clones are found in 11 of the 26 libraries being sequenced by the WashU-Merck EST project. These 11 libraries were generated by at least four separate methods (7). It is possible that some step of library preparation, such as normalization, has generated chimeric clones which contain sequences derived from more than one gene. However, our data imply that some uncoupled clones may be the result of yet undocumented biological events. Although our study of 545 clones indicates that, on average, 11% of dbEST clones are uncoupled, two recent studies, by Hillier *et al.* of 5000 (4) and by Aaranson *et al.* of 10 000 (9) human clones from dbEST, have shown much lower frequencies of uncoupled clones, ~1%. The authors of these two manuscripts performed automated sequence comparisons between ESTs and databases of mRNA sequences. Each manuscript utilized different analysis stragegies and mRNA databases. On the other hand, we manually compared genomic DNA sequences to a database of ESTs. Differences in the experimental approaches partially explain our 10-fold higher estimate of the number of uncoupled clones. Other differences may reflect the fact that we analyzed fewer clones, as well as our finding that the occurrence of uncoupled clones appears to be gene specific. In particular, while no uncoupled clones were observed for GST, LCAT, SFRS7 and cytochrome P-450, 20–26% of the clones from aldolase C, HMG-14, HMG-17 and ODC are uncoupled (Table 2). Uncoupled clones may result from unusual biological events which are evident only when redundant sequence datasets are analyzed. For example, a separate analysis of ESTs suggests that there are pairs of human genes which overlap at their 3′ ends and which are transcribed in opposite orientations (20). However, this phenomenon does not appear to explain the existence of any of the uncoupled clones which we analyzed. Spurious ligations between otherwise distinct cDNAs might be an indication of an RNA chemistry of which we are not yet aware (i.e., rearrangement activity).

## Conclusions

The database of expressed sequence tags, dbEST, is becoming a widely used tool in basic biological research as its users discover that its large collection of human cDNAs provides a well-stocked pool in which to search for novel gene sequences. Some researchers may find that the sequence redundancy in dbEST, the fact that many genes are represented by multiple ESTs, means that the database may reveal new features of well-characterized genes as well. For example, ESTs provide information about previously uncharacterized gene alleles, expression specificities of genes from multigene families, and RNA modifications (J.C.Wootton, personal communication). We show in this report that dbEST may also be a useful location in which to learn more about mRNA splicing, as, in a given gene, up to 73% of the spliced or partially spliced primary transcripts either contain intron sequences or are otherwise spliced at sites which have not been previously documented. It will take further research to determine whether these unreported transcript types have true biological significance, or if they are artifacts of EST sequence generation. We also present evidence that certain genes are represented by high numbers, up to 26%, of uncoupled clones, that is, clones in which the pairs of ESTs making up the 5′ and 3′ end sequences are likely to derive from two separate and unrelated genes. The existence of such clones may be due to errors in EST preparation or to unrecognized biological phenomena.

Some investigators, accustomed to analyzing full-length, carefully proofread, sequences in GenBank, will be mislead after they inadvertently discover some of the more ambiguous ESTs. However, other studies, as well our personal experience, indicate that the majority of sequences in dbEST are problem-free. Without knowing the full length genomic sequence from which it is derived, it is difficult to know *a priori* whether a given EST is unmatching or a given clone is uncoupled. We offer some basic suggestions to help researchers evaulate their ESTs of interest. First, information about the quality of individual ESTs is available. Although all ESTs can be retrieved from GenBank (http://www3.ncbi.nlm.nih.gov/Entrez), ESTs can also be retrieved directly from dbEST in a format which annotates the range of high quality sequence and includes up-to-date information about matches between the EST and other sequences (http://www.ncbi.nlm.nih.gov/dbEST). Questions about the sequence of a particular EST may be resolved by looking at the original sequence trace available from the WashU-Merck Project (http://genome.wustl.edu/est). Clones can also be purchased from distributors if more comprehensive sequence verification is necessary (http://www-bio.llnl.gov/bbrp/image/image.html). Second, the sequences of multiple overlapping ESTs can be aligned. The alignment may help to pinpoint unwanted unmatching ESTs and sequencing errors, and may highlight intron sequences or unreported splice sites, which could guide future experimental work. UniGene clusters, groups of overlapping ESTs and other GenBank sequences, provide a useful starting point for finding multiple overlapping sequences (http://www.ncbi.nlm.nih.gov/UniGene) (6). In conclusion, although we warn researchers about some potentially dubious EST entries, we also believe that some of these more ‘unusual’ sequences may lead to some very interesting experiments.

## ACKNOWLEDGEMENTS

We thank Drs D.J.Lipman and G.D.Schuler for their insightful comments, J.C.Wootton for communicating results before publication, and D.E.Bassett,Jr, M.S.Boguski and G.D.Schuler for their critical reading of the manuscript.

## REFERENCES

1 Tilghman,S.M. (1996) *Genome Res.*, **6,** 773–780.
2 Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) *Nature Genet.*, **4,** 332–333.
3 Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B., Moreno,R.F., *et al.* (1991) *Science*, **252,** 1651–1656.
4 Hillier,L., Lennon,G., Becker,M., Fatima Bonaldo,M., Chiapelli,B., Chissoe,S., Dietrich,N., DuBuque,T., Favello,A., Gish,W., *et al.* (1996) *Genome Res.*, **6**, 807–828.
5 Boguski,M.S. (1995) *Trends Biochem. Sci.*, **20**, 295–296.
6 Schuler,G.D., Boguski,M.S., Stewart,E.A., Stein,L.D., Gyapay,G., Rice,K., White,R.E., Rodriguez-Tomé,P., Aggarwal,A., Bajorek,E., *et al.* (1996) *Science*, **274**, 540–546.
7 de Fatima Bonaldo,M., Lennon,G. and Soares,M.B. (1996) *Genome Res.*, **6**, 791–806.
8 Bains,W. (1996) *Nature Biotechnol.*, **14**, 711–713.
9 Aaronson,J.S., Eckman,B., Blevins,R.A., Borkowski,J.A., Myerson,J., Imran,S. and Elliston,K.O. (1996) *Genome Res.*, **6**, 829–845.
10 McIntosh,J.R. and West,R.R. (1995) *J. Cell. Biol.*, **131**, 1361–1364.
11 Boguski,M.S., Tolstoshev,C.M. and Bassett,D.E.,Jr (1994) *Science*, **265**, 1993–1994.
12 Connelly,C. and Hieter,P. (1996) *Cell*, **86**, 275–285.
13 Gavin,K.A., Hidaka,M. and Stillman,B. (1995) *Science*, **270**, 1667–1671.
14 Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) *Methods Enzymol.*, **266**, 141–162.
15 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
16 Claverie,J.-M. and States,D.J. (1993) *Comput. Chem.*, **17**, 191–201.
17 Chao,K.M., Zhang,J., Ostell,J. and Miller,W. (1995) *Comput. Appl. Biosci.*, **11**, 147–153.
18 Huang,X.Q., Hardison,R.C. and Miller,W. (1990) *Comput. Appl. Biosci.*, **6**, 373–381.
19 Zhang,J., Ostell,J. and Rudd,K.E. (1994) In Hunter,L. (ed.), *27th Hawaii International Conference on System Sciences*. IEEE Computer Society Press, Maui, Hawaii, pp. 58–67.
20 Houlgatte,R., Mariage-Samson,R., Duprat,S., Tessier,A., Bentolila,S., Lamy,B. and Auffray,C. (1995) *Genome Res.*, **5**, 272–304.