The Human Transcript Database: a catalogue of full length cDNA inserts

John Bouck*, Michael P. McLeod, Kim Worley and Richard A. Gibbs
Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA

Received on April 2, 1999; revised on August 26, 1999; accepted on September 10, 1999

Abstract
Summary: Full length cDNA sequences are an important resource for the research community but are currently intermingled with other sequences. We have identified the human full length insert cDNA sequences in GenBank and placed them in a single location, the Human Transcript Database.

Availability: The Human Transcript Database is available at http://www.hgsc.bcm.tmc.edu/HTDB/.
Contact: John Bouck: jbouck@bcm.tmc.edu

Introduction
The public sequence databases continue to be flooded with data generated from a variety of large- and small-scale sequencing efforts throughout the world. Over 3 million sequences are contained in the GenBank repository (Benson et al., 1999) and providing public access to this data set is a formidable task. The nucleotide sequences present in GenBank are derived from different sources and serve different purposes. Expressed sequence tags (ESTs) are the most abundant sequences, consisting of a single sequencing read from the end of a cDNA clone. These sequences are usually less than 500 bases in length and, as the name suggests, were initially intended to identify the presence of a transcript in a cDNA library (Adams et al., 1992). If the cDNA insert is more thoroughly sequenced with multiple overlapping reads, then it is labeled as either a cDNA or a mRNA sequence. The length of cDNA inserts are often 1–2 kb and occasionally much longer (>10 kb). Genomic sequences are generally produced by large-scale sequencing projects and range in size from 30 kb to >200 kb. Smaller segments of genomic sequences and other more fragmented data are also submitted to GenBank including sequence tagged sites (STS) that are utilized for mapping and entries describing polymorphisms that may contain fewer than several dozen bases.

Although much of the data contained within GenBank is easily accessible and very comprehensive, sometimes relatively small sets of data are obscured by the more abundant sequences. Presently there are a number of groups that are producing full-length cDNA insert sequences, some of which cover entire genes (Ishikawa et al., 1998; Nagase et al., 1998; Yu et al., 1997). These sequences are then submitted to GenBank where they are mixed with many other sequences. Further complicating the databases is the fact that the system for annotating cDNAs has changed over time. Due to the difficulty of identifying full-length cDNA insert sequences in the primate division of GenBank, we have generated a complete database containing all human non-EST transcript sequences and then clustered these sequences into groups of putative genes. Here we present the Human Transcript Database (HTDB) that is a useful supplement to the current databases by providing a thorough representation of the high-quality transcript sequences that are available to the public.

Construction of the human transcript database
Sequences that were derived from RNA were identified using a set of keyword filters we developed to screen each GenBank entry. These filters were developed by parsing all of the annotated fields and grouping the entries that contain identical annotations. Examination of the groups revealed additional filter candidates. Sequential parsing, grouping and filtering led to the set of filters eventually utilized for construction of the database. Several sequences have been found that escape the filtering process and are incorrectly identified as mRNAs. These entries usually have minimal annotation and evade the filters due to a lack of associated information. We have manually identified these non-mRNA sequences that are not identified by our filter process and eliminated them. Therefore, we label this database as semi-curated because the few troublesome sequences (currently numbering 170) are dealt with individually.

Subjecting the 91 565 sequences in the primate division of GenBank (release 110.0) to the filtering process iden-
that is maintained at Baylor College of Medicine and can be searched by a Blast similarity search server for the nucleotide sequence. Additionally, the sequences are maintained in a relational database separate from the EST database, which contains duplicates. We therefore identified a smaller number of entries (95% identity over 95% of length). The non-redundant version of the database contains over 13,000 entries. The average length of the 22,000 entries is just over 2 kb, and less than one-third of all entries are larger than this. The number of sequences over 10 kb is even less (66). As expected, many of the entries are short, which might indicate that the majority of mRNA sequences are incomplete and do not represent the entire gene or open reading frame.

**Clustering of sequences into putative gene groups**

Several cDNA sequences might be derived from the same gene but fail to be identified as nearly identical in our previous grouping. In order to discover the number of genes that have been sequenced, we grouped sequences together that contained a significant amount of overlap (>5% length at 95% identity). We call these groups Putative Genes (Pugs) of which we have identified over 10,000 in the HTDB. If the genome contains 80,000 to 100,000 genes then only 10–15% of the genes have corresponding cDNA sequence. Clearly many more cDNAs must be identified to obtain complete representation of expressed sequences.

**Accessing the human transcript database**

The information in the HTDB has been made available on the World Wide Web (http://www.hgsc.bcm.tmc.edu/htdb/) and is accessible through two different search modes, text-based keyword searches or sequence-based similarity searches. The text-based search allows one to identify cDNAs by querying many annotated features such as author, chromosomal location and/or gene name. These fields are derived directly from the GenBank entry and are maintained in a relational database separate from the nucleotide sequence.

The sequence information is also available on the Web and can be searched by a Blast similarity search server that is maintained at Baylor College of Medicine—Human Genome Sequencing Center. Sequences are compared to the entire database with the NCBI Blast2.0 program and strong hits are displayed (Altschul et al., 1990). Due to the smaller total number of sequences and the low redundancy of the HTDB, large numbers of hits rarely occur and gene information is quickly extracted. Information from either text- or sequence-based searches is displayed in table format and contains links to a number of other databases as well as links to obtain more information that is maintained in the HTDB. Other sequence similarity searches of the HTDB that display the alignment data are also available from the BCM search Launcher (http://www.hgsc.bcm.tmc.edu/SearchLauncher/).

**Acknowledgements**

We thank James H. Gorrell and Andrew Arenson for helpful comments and programming assistance. J.B. and R.G. are supported by grant number HG01459 from the National Human Genome Research Institute at the National Institutes of Health. K.W. is supported by grant number DE-FG03-95ER62097 from the Human Genome Project at the Department of Energy.

**References**


