ESTviewer: a web interface for visualizing mouse, rat, cattle, pig and chicken conserved ESTs in human genes and human alternatively spliced variants

Feng-Chi Chen¹,² and Trees-Juen Chuang¹,*

¹Genomics Research Center and ²Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan

Received on December 14, 2004; revised January 29, 2005, accepted February 14, 2005
Advance Access publication February 18, 2005

ABSTRACT

Summary: ESTviewer is a web application for interactively visualizing gene structures, with emphasis on mammalian and avian expressed sequence tags (ESTs) that are conserved in the human genome and alternatively spliced (AS) variants. AS variants from the UCSC, Vega and PSEP annotations are presented in this application for comparison. EST data from six species, human, mouse, rat, cattle, pig and chicken, are mapped to the human genome to show cross-species EST conservation in annotated exonic and intronic regions. Cross-species EST conservation is evolutionarily and functionally important because it represents the effects of selection pressure on genic regions and transcriptome over evolutionary time. Emphatically, ESTviewer provides a convenient tool to compare highly conserved non-human ESTs and human AS variants. The application takes human gene accession ids or coordinates of genomic sequences as inputs and presents annotated gene structures and their AS variants. In addition, the lengths and percentages of human genic regions covered by ESTs are displayed to show the level of EST coverage of different species. The percentages of the UCSC, Vega and PSEP annotated exons covered by ESTs of the six studied species are also displayed in the interface.

Availability: The ESTviewer web interface is publicly accessible at http://www.gate.sinica.edu.tw/~trees/ESTviewer/ESTviewer.htm

Supplementary information: Detailed documentation and the data sets, including the whole human genome annotation of PSEP and 6-species ESTs conserved in the human genome, can be found on the ESTviewer home page.

INTRODUCTION

Expressed sequence tags (ESTs) represent one of the richest genetic information sources. The abundance of EST data gives researchers a powerful tool to probe various topics. ESTs have been widely used for gene structure identification (Chuang et al., 2003, 2004; Brent and Guigo, 2004; Eyras et al., 2004; Birney et al., 2004), expression regulation analysis (Wasserman and Sandelin, 2004), developmental studies (Oimoto et al., 2004; Yu et al., 2004) and other important scientific issues. Evolutionarily, cross-species conservation of ESTs of homologous genes implies preservation of gene structure and/or gene function through evolutionary time. On the other hand, given adequate amounts of EST data from compared species, the absence of certain ESTs suggests the occurrence of evolutionary events that either inactivated or altered the expression pattern of certain genes (Modrek and Lee, 2003). Furthermore, EST data can also be applied to alternatively spliced (AS) variants analysis. AS variants play critical roles in development process, disease pathology and evolution of new protein functions. The difference in AS variants between closely related species can have profound influences on phenotype evolution.

In ESTviewer, interactive visualization of whole human genome annotation from three sources is provided: the University of California, Santa Cruz (UCSC, ttp://www.genome.ucsc.edu), Vega (http://vega.sanger.ac.uk) and PSEP (Chuang et al., 2004). The UCSC annotation includes a large number of annotated genes and transcripts. On the other hand, the Vega annotation, which is reannotated from the Ensembl annotation, contains well-annotated AS forms on nine human chromosomes (chromosomes 6, 7, 9, 10, 13, 14, 20, 22 and X). Meanwhile, PSEP annotates human transcripts and AS variants using human and mouse genomic data and EST data. It was shown that mouse ESTs were good references to predict human transcripts (Kan et al., 2004). Therefore, it is reasoned that EST data from mammalian species other than mouse can also be informative in analyzing human gene structures and AS variants. Consequently, ESTViewer employs EST data from three mammalian species in addition to human and mouse; and chicken, of which abundant EST information is available, is also included for comparison. As PSEP is capable of performing comparative analysis, conserved EST patterns across six species are displayed in the interface. A previous study pointed out high levels of AS conservation between human and mouse (Thanaraj et al., 2003). Therefore, cross-species EST conservation data can be very useful for studying gene function and AS evolution. In addition, novel AS forms predicted for certain human disease-related genes can be new research targets for unveiling so far obscure disease mechanisms or developing novel diagnosis or treatment tools.

DESCRIPTION OF THE WEB INTERFACE

The ESTviewer presents human gene structures and AS forms from three different annotation sources (UCSC, Vega and PSEP) and cross-species EST matches against these annotations. It supports two query
schemes (Fig. 1A), and can accept gene accession IDs from UCSC, NCBI (National Center for Biotechnology Information) or Ensembl. As the UCSC annotations use the same gene ID for multiple-copy genes (genes that have the same transcripts but different genomic locations), a query result of more than one gene copy will be shown if a multiple-copy gene ID is submitted (Fig. 1B). The viewer can also accept coordinates that specify a certain region of the genome. In this second query scheme, all the UCSC-annotated genes located within the specified region will be shown in small-to-large genomic coordinate order (Fig. 1C). Users can select one of the UCSC genes (Fig. 1C) and the corresponding genic sequence (Fig. 1D); also, the UCSC-, Vega- and PSEP-annotated gene structures and AS variants will be illustrated (Fig. 1E). Followed by the visualizations of the three annotations, the EST fragments from human, mouse, rat, cattle, pig and chicken that are conserved in the specified human genomic region will also be presented (Fig. 1F). By pointing to the annotated exons or matched EST fragments, the corresponding coordinates on the matched human chromosome will be shown in a window. Note that all the coordinates of the interface are on the basis of the human genome. Also note that all the genes shown in this interface are based on UCSC annotations. Therefore, genes predicted by PSEP but not by UCSC are not presented.

In addition, as shown in Figure 1C, the ESTviewer also displays a global view of gene locations and related statistical information of AS variants and cross-species EST conservation in a chromosome. Figure 1C.1 presents the genomic locations of genic regions, where the intergenic regions are displayed as fixed-length intervals between genic regions, the displayed lengths of which are approximately proportional to the real lengths. For each human genic region, Figure 1C.2 illustrates the numbers of AS variants identified by UCSC, Vega and PSEP. Figure 1C.3 indicates lengths and percentages of human genomic regions covered by ESTs of the six species studied; and Figure 1C.4 shows the ratios of the total lengths of ESTs that respectively overlap with the UCSC-, Vega- and PSEP-annotated exons divided by the lengths of such exons. The cross-species conservation of expressed sequences is very useful and important information for AS studies in evolution and for judging whether an alternative exon is a major form or not (Boue et al., 2003). In summary, Figure 1C provides information regarding relative lengths of genic regions (Fig. 1C.1), comparison of the numbers of annotated AS from three sources (Fig. 1C.2), the degrees of EST coverage of six species on genic regions in the human genome (Fig. 1C.3) and the levels of cross-species EST conservation in human exonic regions (Fig. 1C.4).

**METHODS AND IMPLEMENTATION**

The cross-species EST-to-genome alignments are based on the CRASA aligner, which can exactly match EST data to genomic sequences with high efficiency, filter out potentially aberrant EST matches (such as disordered matches between ESTs and genomic sequences) and automatically patch incomplete EST matches (Chuang et al., 2003). In this web interface, all the EST fragments (cross-species ESTs) with ≥ 20 bases in length that are conserved in the human genome and that satisfy the CRASA alignment rules stated above are displayed.

The web interface was written as Java™ applets (http://java.sun.com) under Sun microsystems Java runtime environment (J2RE) and the programs of PSEP annotation were written in the PERL language. The EST-to-genome alignments were executed on the CRASA server (Chuang et al., 2003) with a 64-node PC cluster at http://big.pcf.sinica.edu.tw/service/tool.php. All the programs were compiled and executed on the Linux operating system. The interface has been tested and works well on Internet Explorer 6.0, Firefox 1.0, KKMann 3.0 and RealPlay10-RealNetworks for the Windows XP operating system; Firefox 1.0 for Macintosh under OS X 10.3; and Mozilla and Firefox 1.0 for X-windows under the Linux system.

The original EST databases were generously provided by The Institute for Genome Research (TIGR, http://www.tigr.org/tdb/tgi/). The human, mouse, rat, cattle, pig and chicken EST databases used in this study are HGI Release 14 with 494 MB for 889 669 sequences, MGI Release 13 with 390 MB for 718 586 sequences, RGI Release 12 with 89 MB for 132 639, BTGI Release 10 with 69 MB for 95 817 sequences, SGGI Release 9 with 60 MB for 84 858 sequences and GGGI Release 7 with 111 MB for 118 873 sequences. The original human genomic data is version hg16 (or NCBI Human Build 34) and is downloaded from UCSC at http://hgdownload.cse.ucsc.edu/downloads.html.

**DISCUSSION AND FUTURE WORK**

In this web interface, we present not only the gene structures and AS variants of three different annotations (UCSC, Vega and PSEP), but also the conservation between the human genome and ESTs of six species (human, mouse, rat, cattle, pig and chicken). For gene structures and AS variants, the interface provides a convenient display for researchers to study different AS variants among three annotations. It is important for researchers to take into account different annotations for comprehensive AS studies because the three annotations sometimes give very different numbers of AS variants. On
the other hand, for cross-species EST conservation, we find that some non-human EST fragments are highly conserved in human genomic regions (Fig. 1F), even though the coverage ratios of the non-human ESTs on the human genome are relatively small (Fig. 1C.3 and 1C.4). Moreover, some of the non-human EST matches do not overlap with any human ESTs. These ESTs may represent lost transcripts (expressed at low level or in tissues not yet sampled) in human evolution, or real human transcripts that are not yet identified. These conserved non-human ESTs are potentially important in understanding mammalian evolution and the complexity of human transcriptome. We expect that the coverage ratios will rapidly increase with the accumulation of EST data. Therefore, more cross-species EST matches are soon to be identified.

An example of ESTviewer’s application in biomedical studies is to identify potentially novel human exons or transcripts from EST matches of other organisms on the human genome. It has been shown that AS is highly relevant to disease and therapy (Faustino and Cooper, 2003; Garcia-Blanco et al., 2004). Different AS transcripts may have completely different or even antagonistic functions. However, the AS variants of certain disease-related genes are yet to be discovered. Through the cross-species approach adopted in this application, researchers may be able to identify important disease-related AS forms that otherwise cannot be found. For evolutionary studies, ESTviewer can identify conserved AS forms between organisms of interest. In other words, given adequate numbers of ESTs, we will be able to distinguish between conserved and lineage-specific AS forms along the phylogenetic tree and infer the evolutionary significance of AS changes in the studied organisms. Such analyses may also provide some hint to the evolution of transcriptome complexity from simple organisms to complex organisms. Still another application of ESTviewer will be identifying AS forms of non-human mammals using human and mouse ESTs. AS information of mammals other than that of human and mouse is very limited. Since the numbers of EST entries of human and mouse are larger than those of all the other mammals, and cross-species AS conservation is generally recognized (Boue et al., 2003; Kan et al., 2004), genome-wide analysis of AS variants of other mammals using ESTs from these two species is deemed meaningful and important.

In the future, we will identify and visualize AS variants of other model organisms, such as Arabidopsis thaliana, Saccharomyces cerevisiae, Caenorhabditis elegans and Drosophila melanogaster. It will be interesting to explore the evolution of AS variants in orthologous genes across species. We will also keep updating the interface as EST data accumulate.

ACKNOWLEDGEMENTS

This work is supported by the Genomics Research Center, Academia Sinica, Taiwan; the National Science Council, Taiwan, under contract NSC 93-2213-E-001-023-; and the National Health Research Institutes (NHRI), Taiwan, under contract NHRI-EX94-9408PC. We thank UCSC Genome Browser for freely downloaded data, Meng-Yuan Chou for assistance in execution of PSEP; Jason Huang, Sheng-Shun Wang and Chuang-Jong Chen for assistance in programming; and ASCC (Academia Sinica Computing Center) for computational resources. We are also thankful to all the contributing sequencing centers and scientists who provided the public sequence data and annotation results that made this work possible.

REFERENCES