ESTpass: a web-based server for processing and annotating expressed sequence tag (EST) sequences

Byungwook Lee[1,2], Taehui Hong[1], Sang Jin Byun[3], Taeha Woo[1] and Yoon Jeong Choi[1,*]

[1]Korean BioInformation Center, KRIBB, Daejeon 305-817, Korea, [2]Department of BioSystems, KAIST, Daejeon 305-701, Korea and [3]Bioinformatics team, Bioneer, 49-3 Munpyeong-dong, Daedeok-gu, Daejeon 306-220, Korea

ABSTRACT

We present a web-based server, called ESTpass, for processing and annotating sequence data from expressed sequence tag (EST) projects. ESTpass accepts a FASTA-formatted EST file and its quality file as inputs, and it then executes a back-end EST analysis pipeline consisting of three consecutive steps. The first is cleansing the input EST sequences. The second is clustering and assembling the cleansed EST sequences using d2_cluster and CAP3 programs and producing putative transcripts. From the CAP3 output, ESTpass detects chimeric EST sequences which are confirmed through comparison with the nr database. The last step is annotating the putative transcript sequences using RefSeq, InterPro, GO and KEGG gene databases according to user-specified options. The major advantages of ESTpass are the integration of cleansing and annotating processes, rigorous chimeric EST detection, exhaustive annotation, and email reporting to inform the user about the progress and to send the analysis results. The ESTpass results include three reports (summary, cleansing and annotation) and download function, as well as graphic statistics. They can be retrieved and downloaded using a standard web browser. The server is available at http://estpass. kobic.re.kr/.

INTRODUCTION

Expressed sequence tag (EST) sequences are generated by single-pass 5′ or 3′ DNA sequencing of clones randomly picked from cDNA libraries (1). EST represents a partial description of the transcribed portions of genomes, and thus can provide insight into transcribed genes in a variety of organisms. EST sequences are widely used in rapid and cost-effective methods for discovering genes, and as a useful resource for gene mapping and cDNA array construction (2). The utility of EST is also illustrated by the phylogenetic diversity of organisms represented in dbEST, an EST database (3,4).

However, identifying encoded genes from EST sequences presents a number of challenges (5). EST may contain low-complexity sequences, relatively frequent chimeric sequences, repeat sequences and contaminant sequences such as vectors and adaptors. They should be trimmed or masked before further analysis. Because EST sequences are partial fragments of cDNA, they should be assembled and reconstructed into mRNA transcripts to be used for identifying encoded genes. However, this reconstruction is often hampered by the presence of chimeric EST sequences, which are created by the joining of two or more different fragments during cDNA cloning or EST sequencing (6). Such chimeric EST sequences cause misassembly, which leads to incorrect gene annotation. Therefore, removing chimeric EST is essential for reconstructing reliable transcripts from EST sequences.

Several EST processing systems have been developed to cope with these challenges such as EST analysis pipeline (ESTAP) (7), ESTAnnotator (8), EST pipeline system (9), PartiGene (10) and ParPEST (11). Although they each have their own objectives, these systems commonly provide automated or semi-automated pipelines for cleansing EST sequences and annotating them using public databases (12). However, most of these pipelines require local installation and maintenance of the latest versions of the tools and databases, and provide simple annotation functions. Moreover, they are only capable of removing chimeric EST that contains contaminant sequences, such as vectors and adaptors.

Here we present a web-based server, called ESTpass, which provides an automated pipeline for cleansing and annotating user-inputted EST sequences according to user-specified options. The use of ESTpass does not require application installation or testing steps. Instead, the user simply uploads EST data and chooses the appropriate analysis tools and parameters on a web browser.

METHODS

The main function of the ESTpass server is a back-end EST annotation pipeline, whose procedures can be divided into three consecutive steps: cleansing, clustering and assembling, and annotation. A schematic of the pipeline workflow is depicted in Figure 1.

Cleansing step

EST sequences may contain various types of contaminants that should be removed before the sequences are used. The cleansing performed in the first step is fundamental to obtaining high-quality sequences from raw sequence data. The cross_match program is used to identify and mask vector sequences and contaminant sequences, such as *Escherichia coli* sequences, at the 5′ and 3′ ends. These masked regions are removed by an ESTpass trimming tool. If adaptor, primer or other contaminant sequences are uploaded by the user, ESTpass searches for and trims them from the both ends of EST sequences. ESTpass also trims low-quality end sequences based on a user-inputted quality file and a minimum quality score. Low-complexity regions in each EST sequence are masked using the RepeatMasker program (A.F.A. Smit, R. Hubley and P. Green RepeatMasker at http://repeatmasker.org) with a user-selected repeat database.

The generation of chimeric EST during the cDNA library construction or the EST sequencing may cause problems in subsequent analysis steps. Thus, ESTpass detects them if the trimmed EST contains internally inserted contaminants. These chimeric EST sequences are not used in the next process. After cleansing, EST sequences shorter than a user-specified length (100 bases by default) are discarded. The cleansed EST sequences and their cleansing information are stored in the ESTpass database.

Clustering and assembling step

The second step involves clustering and assembling the cleansed EST sequences. This step is the key to identifying the expressed genes of a cDNA library. The d2_cluster (13) and CAP3 (14) programs were used to reconstruct putative transcripts from the cleansed EST sequences.

EST data sets can be contaminated with genomic DNA, such as intron and intergenic regions, and unknown contaminants. However, these types of chimerism cannot be identified in the cleansing step. These types of chimeric EST sequences are removed by using the chimer-detection method that employs the sequence alignments outputted by the CAP3 program. First, ESTpass detects a putative chimeric EST sequences if multiple alignments in this output have a chimerism spot, which represents a single EST region surrounded by two flanking sequence segments containing four or more ESTs. Its occurrences result in a barbell-shaped contig EST alignment (Figure 2). Second, to evaluate the degree of chimerism in the detected chimeric EST, EST sequences containing the chimerism spot are searched against the nr database using BLASTX (15). The putative chimerism is disproved if the sequence matches a protein of the nr database and their alignments spans its chimerism spot in the putative chimeric EST, whereas it is confirmed if both sides of the chimerism spot in the putative chimeric EST sequence match different proteins of the nr database. If any chimerism is found, ESTpass will recluster and reassemble the EST sequences after excluding the confirmed chimeric EST sequences. Two examples of chimeric ESTs are given in a 'Supplementary Data' section. From the assembly
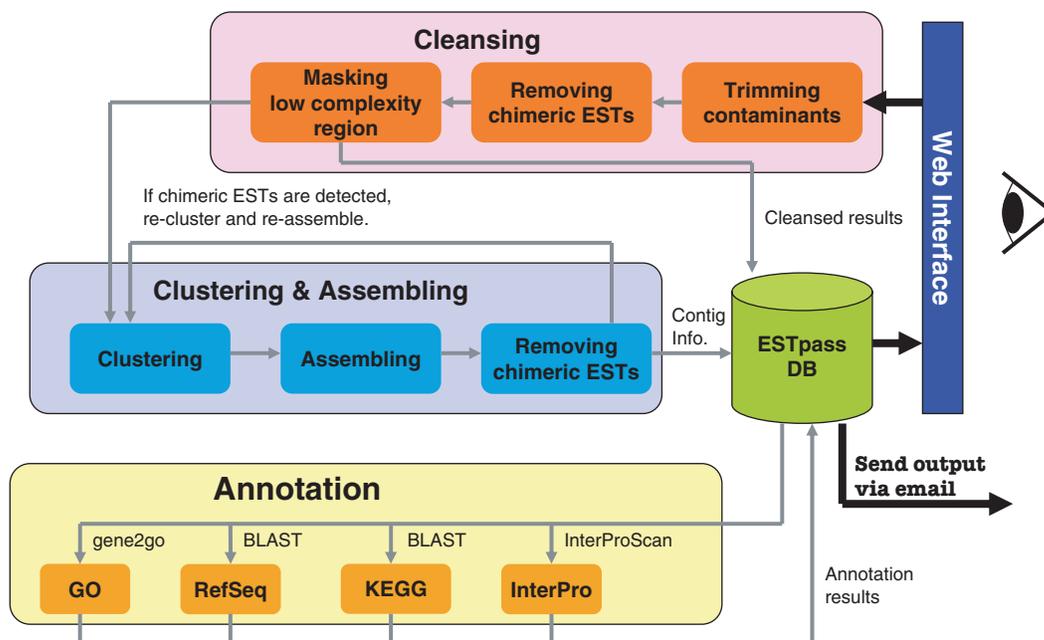

Figure 1. Schematic of the ESTpass workflow. The ESTpass pipeline consists of three major steps: cleansing, clustering and assembling, and annotation. ESTpass output is sent to the user via email and can be retrieved using a standard web browser.

results, the consensus sequences of contigs, singletons and singlets are chosen as putative transcripts, which are subjected to the annotation process.

Annotation step

The last step involves annotating the putative transcripts created in the previous step. ESTpass provides five annotation facilities. The first is homology searching, in which the putative transcript sequences are compared with the RefSeq protein database (16) using BLASTX. The BLASTX results are filtered using a user-specified cutoff e-value (1E-04 by default), and the top-five hits and their alignment results are stored in the ESTpass database. The second annotation facility is Gene Ontology (GO) assignment, in which the sequences are annotated with GO terms using both gene2go and gene2refseq files downloaded from Entrez gene (17). The third facility is pathway analysis, in which the sequences are BLASTed against the KEGG gene database (18), and the top-hit KEGG IDs are reported. The fourth annotation facility is motif/domain finding, in which the sequences are translated in all six frames and their translation products are queried against the InterPro database (19) using the InterProScan program (20). The fifth annotation facility is the identification of the full length of the putative transcript sequences. There are several algorithms (21–23) for identifying translation initiation sites in EST sequences. Among them, we used the TargetIdentifier (23) algorithm that does not require 'training' and uses the BLASTX output. Although its original algorithm classifies the full length into six classes, ESTpass provides only a 'full-length' class since it is most evident among the six.

IMPLEMENTATION

The ESTpass web server comprises three major components: ESTpass web interfaces, a set of back-end pipeline programs and a relational database (MySQL). The web interfaces are implemented in static HTML pages and Java Server Pages programs (http://java.sun.com/products/jsp/). MySQL is used to store input EST,

intermediate data of the pipeline and the cleansed and annotated results. The database schema is available at the ESTpass website. The pipeline consists of several program modules written in Perl, Python or Java, and an Apache Ant (http://ant.apache.org) script controls the configuration and operation of these pipeline modules. The back-end system is a Linux machine with four dual-core AMD Opteron 875 CPUs (8 cores) and 16 GB of RAM. The ESTpass server has a queuing system to control user-submitted projects. ESTpass simultaneously runs three projects; any remaining projects will be put into a job queue.

INPUT AND OUTPUT

Input

The ESTpass web interfaces allow the user to submit EST sequences and their quality scores, and contaminants such as vectors and adaptors. All the EST sequences need to be prepared in FASTA format and saved as a single text file before being uploaded. Although most EST projects produce a large number of chromatogram files, ESTpass cannot accept chromatogram files due to file-size limitations of web-based uploading. Accordingly, chromatogram files should be converted into DNA sequence files using a base-calling program such as phred (24,25). The maximum number of input EST sequences in a single submission is 10 000 EST sequences. ESTpass treats the first word in the description line of an EST sequence as its name, and checks for sequence name duplication and the consistency between the sequence file and the quality file (if provided).

Output

The ESTpass output is stored in a MySQL database and its access URL is sent to the user-specified email address. The output largely consists of three reports: summary, cleansing, and annotation (Supplementary Figure S1). The summary report describes the statistics of cleansing, clustering, assembly, and annotation. In addition, detailed statistics on putative transcripts and their annotation
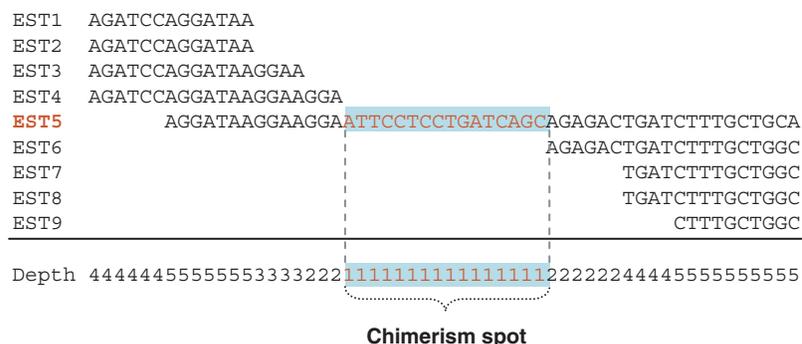

Figure 2. Illustration of the detection of a chimeric EST sequence in the alignment output of the CAP3 program. A putative chimeric EST sequence is detected if it has chimeric spots, which is represented by a stretch of EST sequences with both a depth of one and being surrounded by an alignment depth of four or more, and is dumbbell-shaped. In this example, EST5 is a candidate chimeric EST sequence. The chimerism of the EST sequence containing the chimerism spot is confirmed by comparison with the nr database using BLASTX. In the BLASTX output, the putative chimeric sequence matches a protein and its alignment spans the chimerism spot, which disproves its chimerism. In contrast, both sides of the chimerism spot in the putative chimeric EST sequence can match different proteins of the nr database, its chimerism is confirmed.

results are represented as graphs. The cleansing report presents detailed information about the cleansing results of each EST sequence such as the EST length and trimming information. It provides links to input and cleansed EST sequences. The annotation report presents annotation results about the putative transcript sequences such as the full length information, RefSeq number, GO ID, KEGG ID and InterPro ID with their detailed information. The user can also download the three reports and components ESTs of clusters and putative transcripts as tab-delimited text files from the download menu of the access URL. After finishing the user submitted projects, the user can further analyze the final output using public software or web-based servers, e.g., finding ORF (26) regions in the putative transcript sequences. The results will be kept for 1 month and then deleted.

CONCLUSIONS

ESTpass provides more rigorous chimeric EST detection and exhaustive annotation facilities, compared to other EST pipelines (Supplementary Table S1). EST analysis is generally time-consuming due to the large number of EST sequences—it may take more than 1 day depending on the number of EST sequences (Supplementary Table S2). Therefore, all the results are sent the user via email. Among the three steps, the annotation process requires the longest time, especially finding motif/domains of putative transcripts. Thus, the 'motif/domain annotation' on the annotation options should be unchecked if the user want to receive the results more quickly.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

REFERENCES

1. Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.
2. Parkinson,J. and Blaxter,M. (2004) *Parasite Genome Protocols* Humana Press, Totowa, NJ.
3. Parkinson,J., Guiliano,D.B. and Blaxter,M. (2002) Making sense of EST sequences by CLOBBing them. *BMC Bioinformatics*, **3**, 31.
4. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST–database for "expressed sequence tags". *Nat. Genet.*, **4**, 332–333.
5. Pertea,G., Huang,X., Liang,F., Antonescu,V., Sultana,R., Karamycheva,S., Lee,Y., White,J., Cheung,F. *et al.* (2003) TIGR Gene Indices Clustering Tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.
6. Kunne,C., Lange,M., Funke,T., Miehe,H., Thiel,T., Grosse,I. and Scholz,U. (2005) CR-EST: a resource for crop ESTs. *Nucleic Acids Res.*, **33**, D619–D621.
7. Mao,C., Cushman,J.C., May,G.D. and Weller,J.W. (2003) ESTAP–an automated system for the analysis of EST data. *Bioinformatics*, **19**, 1720–1722.
8. Hotz-Wagenblatt,A., Hankeln,T., Ernst,P., Glatting,K.H., Schmidt,E.R. and Suhai,S. (2003) ESTAnnotator: a tool for high throughput EST annotation. *Nucleic Acids Res.*, **31**, 3716–3719.
9. Xu,H., He,L., Zhu,Y., Huang,W., Fang,L., Tao,L., Zhu,Y., Cai,L., Xu,H. *et al.* (2003) EST pipeline system: detailed and automated EST data processing and mining. *Genom. Proteo. Bioinform.*, **1**, 236–242.
10. Parkinson,J., Anthony,A., Wasmuth,J., Schmid,R., Hedley,A. and Blaxter,M. (2004) PartiGene–constructing partial genomes. *Bioinformatics*, **20**, 1398–1404.
11. D'Agostino,N., Aversano,M. and Chiusano,M.L. (2005) ParPEST: a pipeline for EST data analysis based on parallel computing. *BMC Bioinformatics*, **6**(Suppl. 4), S9.
12. Nagaraj,S.H., Gasser,R.B. and Ranganathan,S. (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief. Bioinform.*, **8**, 6–21.
13. Burke,J., Davison,D. and Hide,W. (1999) d2_cluster: a validated method for clustering EST and full-length cDNAsequences. *Genome Res.*, **9**, 1135–1142.
14. Huang,X. and Madan,A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
15. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
16. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
17. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.
18. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
19. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
20. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
21. Nishikawa,T., Ota,T. and Isogai,T. (2000) Prediction whether a human cDNA sequence contains initiation codon by combining statistical information and similarity with protein sequences. *Bioinformatics*, **16**, 960–967.
22. Nadershahi,A., Fahrenkrug,S.C. and Ellis,L.B. (2004) Comparison of computational methods for identifying translation initiation sites in EST data. *BMC Bioinformatics*, **5**, 14.
23. Min,X.J., Butler,G., Storms,R. and Tsang,A. (2005) TargetIdentifier: a webserver for identifying full-length cDNAs from EST sequences. *Nucleic Acids Res.*, **33**, W669–W672.
24. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
25. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
26. Min,X.J., Butler,G., Storms,R. and Tsang,A. (2005) OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res.*, **33**, W677–W680.