Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W459–W462***
doi:10.1093/nar/gkl066

EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments

Ali Masoudi-Nejad[1,2,*], Koichiro Tonomura[1], Shuichi Kawashima[3], Yuki Moriya[1], Masanori Suzuki[4], Masumi Itoh[1], Minoru Kanehisa[1,3], Takashi Endo[2] and Susumu Goto[1]

[1]Laboratory of Bioknowledge Systems, Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho Uji, Kyoto 611-0011, Japan, [2]Laboratory of Plant Genetics, Division of Applied Bioscience, Kyoto University, Kyoto 606-8502, Japan, [3]Laboratory of Genome Database, Human Genome Center, University of Tokyo, Tokyo 108-8639, Japan and [4]Hitachi Government & Public Corporation System Engineering, Ltd. 4-18, Tokyo 2-Chome, Koto-ku, Tokyo 135-8633, Japan

Received January 19, 2006; Revised February 19, 2006; Accepted March 1, 2006

ABSTRACT

Expressed sequence tag (EST) sequencing has proven to be an economically feasible alternative for gene discovery in species lacking a draft genome sequence. Ongoing large-scale EST sequencing projects feel the need for bioinformatics tools to facilitate uniform EST handling. This brings about a renewed importance for a universal tool for processing and functional annotation of large sets of ESTs. EGassembler (http://egassembler.hgc.jp/) is a web server, which provides an automated as well as a user-customized analysis tool for cleaning, repeat masking, vector trimming, organelle masking, clustering and assembling of ESTs and genomic fragments. The web server is publicly available and provides the community a unique all-in-one online application web service for large-scale ESTs and genomic DNA clustering and assembling. Running on a Sun Fire 15K supercomputer, a significantly large volume of data can be processed in a short period of time. The results can be used to functionally annotate genes, to facilitate splice alignment analysis, to link the transcripts to genetic and physical maps, design microarray chips, to perform transcriptome analysis and to map to KEGG metabolic pathways. The service provides an excellent bioinformatics tool to research groups in wet-lab as well as an all-in-one-tool for sequence handling to bioinformatics researchers.

INTRODUCTION

Expressed sequence tags (ESTs) are partial sequences of expressed genes prepared by reverse transcribing mRNA and cloning the cDNA fragments into a plasmid (1). ESTs have proven to be an extremely valuable resource for high-throughput gene discovery, annotating the genome's drafts by providing sequence information to identify novel genes, gene location and intron–exon boundaries within genomic sequence assemblies (2,3). EST sequencing has also proven to be an economically feasible alternative for gene discovery in species lacking a draft genome sequence. It is a cost-effective way to survey the expressed portions of the genome, especially in plants with extremely large genomes (e.g. 16 000 Mbp in wheat).

Large-scale EST sequencing projects, which are conducted by a consortium of laboratories, require bioinformatics tools to facilitate the uniform handling of ESTs. The importance of EST clustering and assembly has been well established as evidenced by the number of databases currently available, such as TIGR gene indices (4), STACK (5) and UniGene (6). This proliferation of online resources demonstrates the need for a universal tool for processing and functional annotation of large sets of ESTs.

EGassembler is a single web server, which provides an automated as well as a user-customized analysis tool for cleaning, repeat masking, vector trimming, organelle masking and assembling of ESTs and genomic fragments. It is also designed to serve as a stand-alone web application for each one of the processes.

*To whom correspondence should be addressed. Tel: +81 90 2195 4417; Fax: +81 774 38 3269; Email: amasoudi@kais.kyoto-u.ac.jp
Correspondence may also be addressed to Susumu Goto. Tel: +81 774 3271; Fax: +81 774 3269; Email: goto@kuiicr.kyoto-u.ac.jp

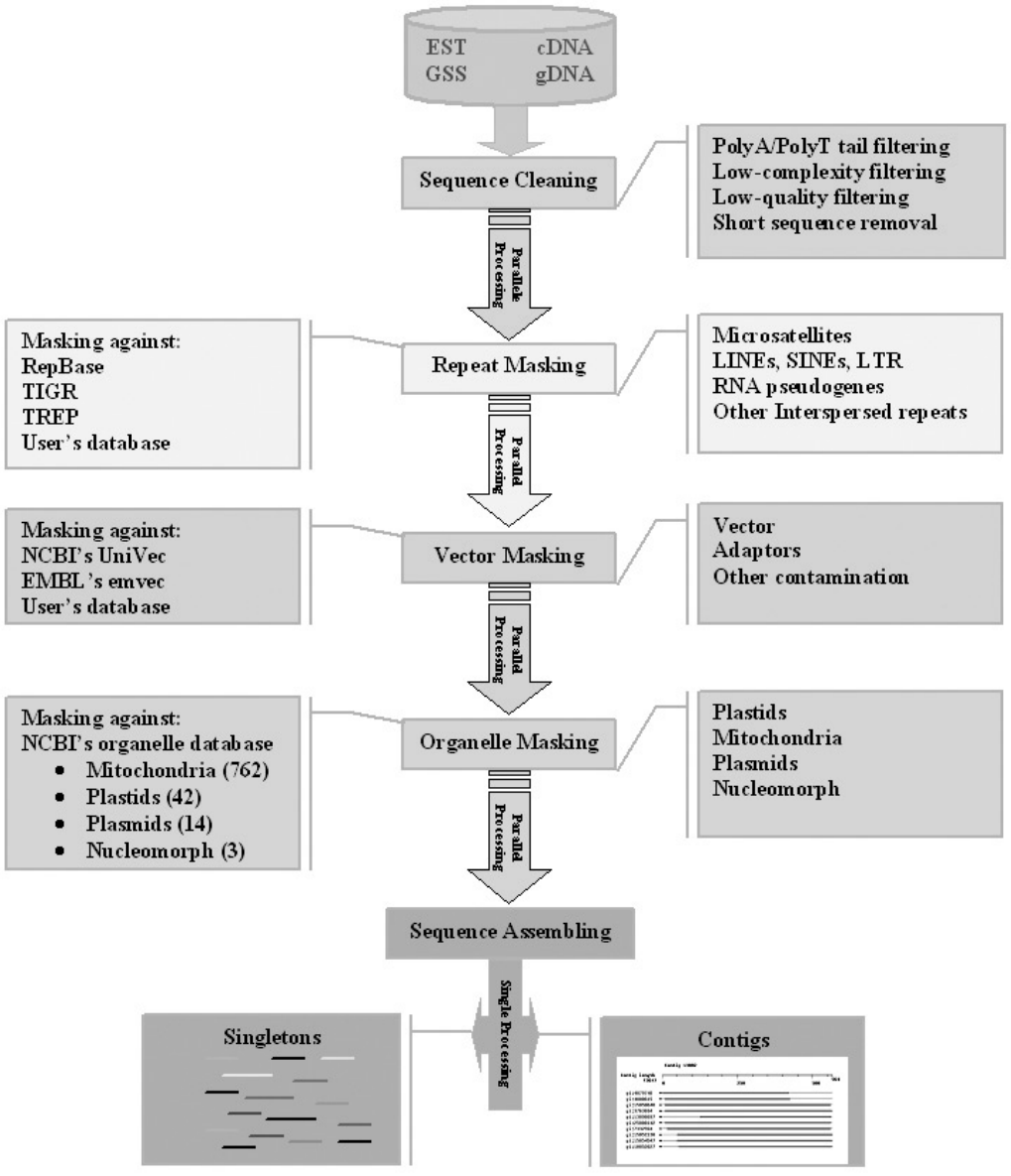

Figure 1. EGassembler data flow. The flowchart shows the pipeline used in the EGassembler web server. The Middle portion shows the process and running modes (parallel or single). The right side shows each process action and the left side shows the databases used by each process for masking.

DESCRIPTION

EGassembler consists of a pipeline of five components, each using highly reliable open-source tools (4,7–9) and a non-redundant custom-made database of repeats (EGrep) and vectors (EGvec) covering almost all publicly available vectors and repeats databases. The EGrep is a non-redundant repeats database covering latest release of the RepBase (10), TREP (11) (http://wheat.pw.usda.gov/ITMI/Repeats/index.shtml), TIGR plant repeats (12) and thousands other publicly available repeat sequences on the Internet. The EGrep was constructed by combining and assembling repetitive elements using PHRAP and CAP3 assembling programs into one single database. EGvec was made by assembling the NCBI's UniVec and EMBL's emvec vector/adaptor library and other vector sequences using CAP3 program.

Figure 1 shows a flow chart of the EGassembler process. The web server accepts any type of DNA sequences in FASTA format (EST, GSS, cDNA, gDNA). The sequence cleaning process involves basic procedures such as, removing the polyA/polyT tail, clipping low-quality ends (the ends rich in undetermined bases) and discarding those that are too short (shorter than 100) or which appear to be mostly low-complexity sequences. The repeat masking process compares the query sequence against one or more files of FASTA sequences (library for masking). Masking vectors and organelles is performed using the program Cross_Match (9) where is a general-purpose utility for comparing any two sets of DNA sequence. It is used to compare query sequences to a set of vector or organelle sequences and produce vector/organelle masked versions of the input sequences. The

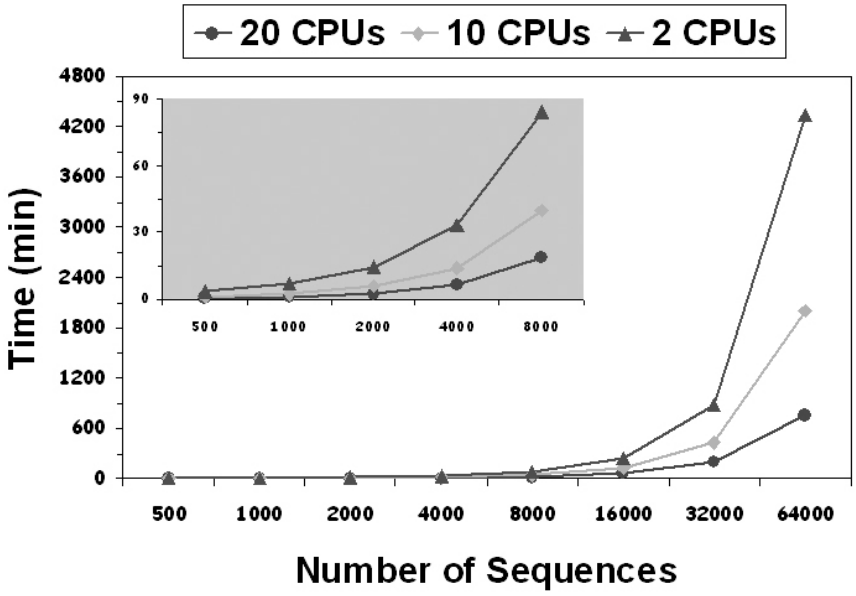

Figure 2. EGassembler performance. The large plot shows the EGassembler performance under different sequence loads and different numbers of CPUs. The inset displays the performance with ⩽8000 sequences.

sequence assembling process uses the CAP3 program (7) for Clustering and assembling the sequences into contigs and singletons. CAP3 assembles ESTs from the same gene under more stringent criteria compared with other approaches, and is able to distinguish gene family members while tolerating sequencing error.

All of the processes in the pipeline, except the assembling step, run in parallel using all CPU resources available on the server. Those programs that were originally written as serial programs, using only one CPU, are now executed in parallel by implementation of a new algorithm using the Perl thread module. This implementation is especially valuable for trimming the vector and masking the organelle sequences. Using the original program on a single CPU required several days depending on input sequences, but now it takes only a few hours. Figure 2 shows a diagram of the EGassembler performance under different loads.

The main menu on EGassembler interface has three submenus providing users with the following processing options.

One-click assembling

All the components in the pipeline are run consecutively with their default options. After uploading the sequences, choosing the libraries for trimming and masking, assembling results can be obtained in one-click. The results of all steps are available to users for downloading as both URL addresses in one single-zipped file and as separate files for each step. The URL addresses of results are valid for access by users for one week after completion.

Step-by-step assembling

Users run all the components outlined in the pipeline interactively and have the opportunity to run each one of them with advanced options. The output of each step of the process is automatically used as the input to the next step of the pipeline; users can also jump into any step at anytime with the previous results.

Stand-alone processing

Users can use each one of the components alone with all options available. Web-interface displays the default parameters of the original programs, any of which users can choose/change for each program.

APPLICATION

Using the One-Click Assembling option, we used EGassembler server to analyze 386 515 rice ESTs deposited in NCBI's dbEST database. By searching the Nucleotide database of GenBank using the term 'oryza sativa AND gbdiv_est[PROP]', all the deposited ESTs (386 515) were downloaded in FASTA format and used as the input file. From 386 515 EST sequences, 125 404 reads were trimmed and 11 553 sequences discarded through the sequence cleaning process. The repeat masking process identified 345 SINE, 83 LINE, 273 Copia and 1668 Gypsy belonging to the retroelements group and 191 hobo-activator, 1581 TC1-pogo, 398 En-Spm, 268 MuDr and 951 Tourist, all belonging to the DNA transposons group. The number of simple sequence repeats and low-complex elements were 43 293 and 23 412, respectively. Total repetitive elements masked were 5 216 297 bp, about 2.7% of the query sequence. Vector and organelle sequence matches were found in 17 980 (1 300 270 bp, 0.65%) and 2958 (1 064 453 bp, 0.54%) sequences, respectively. CAP3 assembling results in 73 555 singletons (reads that are not used in assembly) and 25 193 contigs. The EGENES database of KEGG (release 34.0, April 2005), which is the transcriptome-based plant database of genes with metabolic pathway information, has also been developed using the pipeline described here.

IMPLEMENTATION

EGassembler is written in Perl CGI and uses suites of open-source programs. The web server runs on a Sun Fire 15K supercomputer, located in the Human Genome Center at the University of Tokyo. While processing, the web server refreshes the results page every 30 s for small sets of data (less than 1000 sequences). For larger data set it provides instead a hyperlink for downloading the results. A user manual for each program and tutorial is available on the web server to provide assistance on using the interface.

FUTURE PLANS

Recently many new algorithms have been introduced for sequence clustering that provides more flexibility and advancement for large-scale projects (13–15). We are planning to validate and use new algorithms to improve the quality of the pipeline on this web server. In the near future there will be an option for users who want to annotate their assembling results based on the KEGG pathway database (16). The results of assembling, including contigs and singletons will be mapped to the pathways in KEGG by transferring the results to another server for automatic functional annotation based on KEGG (KAAS; http://www.genome.jp/kegg/kaas/). We will also continue collecting new repeats and vector sequences from public resources to enrich our custom database for filtering the sequences.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to all the developers of the software which they used to make EGassembler available to the public and to Dr Craig Wheelock and Dr Kiyoko F. Aoki-Kinoshita for critical reading of the manuscript. JSPS Post Doctoral award to A.M-N. is acknowledged. The authors would also like to thank the peer reviewers for their valuable comments. This study was supported in part by a Grant-in-Aid for Scientific Research (A) (No. 123066001) and Grant-in-Aid for Scientific Research on priority areas (No. 17020005 and 17017019) from the Ministry of Education, Science, Sports and Culture, Japan. Funding to pay the Open Access publication charges for this article was provided by Bioinformatics Center of Kyoto University.

Conflict of interest statement. None declared.

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., Moreno,R.F. *et al*. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*., **252**, 1651–1656.
2. Dias,N.E., Correa,R.G., Verjovski-Almeida,S., Briones,M.R., Nagai,M.A., Wilson,D.S., Zago,M.A., Bordin,S., Costa,F.F., Goldman,G.H. *et al*. (2000) Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl Acad. Sci. USA*, **97**, 3491–3496.
3. Hillier,L.D., Lennon,G., Becker,M., Bonaldo,M.F., Chiapelli,B., Chissoe,S., Dietrich,N., DuBuque,T., Favello,A., Gish,W. *et al*. (1996) Generation and analysis of 280 000 human expressed sequence tags. *Genome Res*., **6**, 807–828.
4. Lee,Y., Tsai,J., Sunkara,S., Karamycheva,S., Pertea,G., Sultana,R., Antonescu,V., Chan,A., Cheung,F. and Quackenbush,J. (2005) The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res*., **33**, D71–D74.
5. Christoffels,A., Van Gelder,A., Greyling,G., Miller,R., Hide,T. and Hide,W. (2001) STACK: sequence tag alignment and consensus knowledgebase. *Nucleic Acids Res*., **29**, 234–238.
6. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schrimi,L.M., Sequeira,E., Tatusova,T.A. *et al*. (2003) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **31**, 28–33.
7. Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res*., **9**, 868–877.
8. Smit,A.F.A., Hubley,R. and Green,P. (2004) RepeatMasker Open-3.0. 1996–2004.
9. 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.
10. Jurka,J. (2000) RepBase Update: a database and an electronic journal of repetitive elements. *Trends Genet*., **9**, 418–420.
11. Wicker,T., Matthews,D. and Dubcovsky,J. (2004) TREP, the Triticeae Repeat Sequence Database.
12. Ouyang,S. and Buell,C.R. (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res*., **32**, D360–D363.
13. Konno,H., Fukunishi,Y., Shibata,K., Itoh,M., Carninci,P., Sugahara,Y. and Hayashizaki,Y. (2001) Computer-based methods for the mouse full-length cDNA encyclopedia: real-time sequence clustering for construction of a nonredundant cDNA library. *Genome Res*., **11**, 281–289.
14. Ptitsyn,A. and Hide,W. (2005) CLU: a new algorithm for EST clustering. *BMC Bioinformatics*., **15**, S3.
15. Shibuya,T., Kashima,H. and Kanagawa,A. (2004) Efficient filtering methods for clustering cDNAs with spliced sequence alignment. *Bioinformatics*., **20**, 29–39.
16. Kanehisa,M., Goto,G., 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.