

# WebTraceMiner: a web service for processing and mining EST sequence trace files

Chun Liang<sup>1,\*</sup>, Gang Wang<sup>1</sup>, Lin Liu<sup>1</sup>, Guoli Ji<sup>2</sup>, Yuansheng Liu<sup>2</sup>, Jinqiao Chen<sup>2</sup>, Jason S. Webb<sup>1</sup>, Greg Reese<sup>3</sup> and Jeffrey F. D. Dean<sup>4</sup>

<sup>1</sup>Department of Botany, Miami University, Oxford, Ohio 45056, USA, <sup>2</sup>Department of Automation, Xiamen University, Xiamen, Fujian, 361005, China, <sup>3</sup>Research Computing Support Group, Miami University, Oxford, Ohio 45056, USA and <sup>4</sup>Warnell School of Forestry and Natural Resources, University of Georgia, Athens, Georgia 30602, USA

Received January 9, 2007; Revised March 30, 2007; Accepted April 14, 2007

## ABSTRACT

Expressed sequence tags (ESTs) remain a dominant approach for characterizing the protein-encoding portions of various genomes. Due to inherent deficiencies, they also present serious challenges for data quality control. Before GenBank submission, EST sequences are typically screened and trimmed of vector and adapter/linker sequences, as well as polyA/T tails. Removal of these sequences presents an obstacle for data validation of error-prone ESTs and impedes data mining of certain functional motifs, whose detection relies on accurate annotation of positional information for polyA tails added posttranscriptionally. As raw DNA sequence information is made increasingly available from public repositories, such as NCBI Trace Archive, new tools will be necessary to reanalyze and mine this data for new information. WebTraceMiner ([www.conifergdb.org/software/wtm](http://www.conifergdb.org/software/wtm)) was designed as a public sequence processing service for raw EST traces, with a focus on detection and mining of sequence features that help characterize 3' and 5' termini of cDNA inserts, including vector fragments, adapter/linker sequences, insert-flanking restriction endonuclease recognition sites and polyA or polyT tails. WebTraceMiner complements other public EST resources and should prove to be a unique tool to facilitate data validation and mining of error-prone ESTs (e.g. discovery of new functional motifs).

## INTRODUCTION

Expressed sequence tags (ESTs) are obtained from single-pass sequencing of cDNA clones prepared from mRNAs.

They remain a dominant approach for characterizing the active, protein-coding portions of genomes for a variety of species. As of December 2006, there were 40 227 021 EST entries deposited in GenBank dbEST (1). While these data represent by far the most abundant information resource for expressed portions of various genomes, they also present serious challenges for data quality control due to the inherent deficiencies of ESTs, particularly for those from species without available genome sequences. Many public EST resources, including NCBI UniGene (2) and TIGR Gene Indices (3), have been developed to address the quality limitations, as well as the issues of redundancy and the less-than-full-length nature of ESTs. Unfortunately, the power of these resources relies heavily on the veracity of sequence features that have been submitted to dbEST (4). Many software packages are available for processing raw EST traces prior to GenBank dbEST submission (5–13). Before such submission, EST sequences are typically screened and trimmed of vector, adapter/linker sequences and insert-flanking restriction endonuclease recognition sites, as well as polyA or polyT tails; however, available sequence processing packages are limited in their ability to detect and trim such sequences correctly (13,14). Loss of the information in these trimmed sequences can present an obstacle for data validation of error-prone ESTs, and can impede data mining for certain functional motifs, such as polyadenylation signals, whose detection requires accurate annotation of positional information for polyA tails added posttranscriptionally. To alleviate limitations imposed by trimmed sequences, both NCBI Trace Archive ([www.ncbi.nlm.nih.gov/Traces/trace.cgi](http://www.ncbi.nlm.nih.gov/Traces/trace.cgi)) and Ensembl Trace Server ([trace.ensembl.org](http://trace.ensembl.org)) were established as public repositories for raw DNA sequencer trace files, including those for ESTs. Unfortunately, these databases have so far remained mostly untapped resources for general biologists due to a shortage of freely available, easy-to-use bioinformatics tools with which to identify and highlight sequence

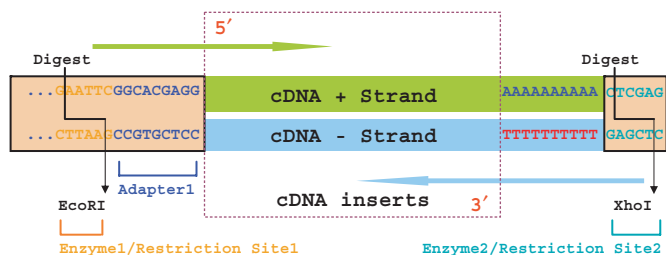
\*To whom correspondence should be addressed. Tel: +1 513 529 2336; Fax: +1 513 529 4243; Email: [liangc@muohio.edu](mailto:liangc@muohio.edu)

features after raw trace processing. To the best of our knowledge, preAssemble (14) is the only freely available sequence processing pipeline service that allows biologists to upload and process raw EST traces online without difficult software installation and configuration procedures. Like other programs, preAssemble creates trimmed sequences, but it produces a quality assessment of sequencing data in color-coded HTML pages that facilitate online data exploration.

Like preAssemble, WebTraceMiner's aim is to be a public sequence processing service that permits biologists to upload their own traces as well as those downloaded from NCBI Trace Archive. However, WebTraceMiner differs from preAssemble, as well as most other packages, by focusing on EST sequence features that characterize the 3' and 5' termini of cDNA inserts and by providing robust data visualization and mining capability of these features. WebTraceMiner is designed to first detect all putative sequence features as single, multiple and/or concatenated occurrence(s) in each sequence read after basecalling, allowing for both perfect and imperfect (i.e. mismatch, deletion and insertion) matching patterns. Currently implemented sequence features include vector fragments, adapter/linker sequences, restriction endonuclease recognition sites and polyA or polyT runs. Based on common canonical models for directional cDNA library construction using oligo-dT primers, as shown in Figure 1, WebTraceMiner then evaluates the number, location, order, fidelity and orientation of the putative features to determine *in silico* verified features that should in theory occur at the 3' and/or 5' termini of all cDNA inserts in a given library. Both putative and verified features can easily be explored and compared using sophisticated data querying and visualization web interfaces. All data can be saved locally and selectively, either in FASTA, tab-delimited or XML formats, for future exploration if necessary. Unambiguous identification and authentication of 3' and 5' termini of cDNA inserts through ESTs provides for the creation of cleaner EST clusters, improves annotation of gene position in genomic sequence, facilitates identification of 3' and 5' untranslated regions (UTRs), and aids detection of potential functional motifs embedded in abundant ESTs.

## INPUT AND OUTPUT

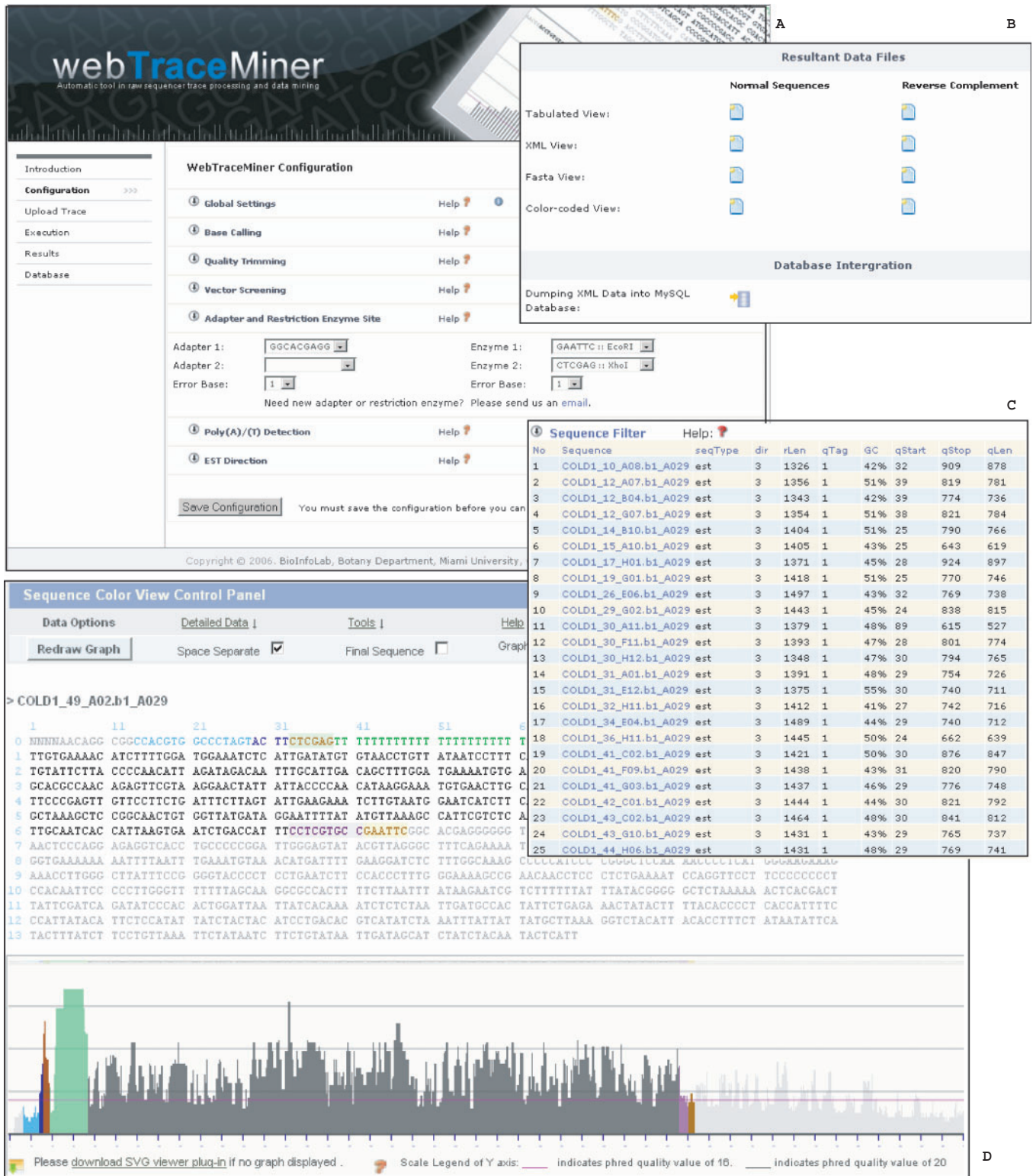
WebTraceMiner takes as input raw EST trace file(s), either a single trace or a zipped file containing multiple traces. As shown in Figure 2 Panel A, the configuration interface allows a user to define and characterize one trace processing event, which consists of several procedures or steps that are common among existing DNA sequence processing packages (5–13). For general users, default settings should be adopted for the most procedures or steps. The only exception is that 'Adapter and Restriction Enzyme Site' configuration requires users to enter the sequences as in the positive cDNA strand for 'Adapter 1', 'Adapter 2', 'Enzyme 1' and 'Enzyme 2', all of which are mandatory for defining appropriately a canonical model for the user-specific cDNA library according to its



**Figure 1.** Example of a canonical model for cDNA library construction. In some cases, a second adapter/linker sequence (Adapter2) might be applied between the polyA/T tails and Restriction site2.

construction (see Figures 1 and 2, Panel A). After a user saves the configuration and finishes uploading raw trace(s) online, WebTraceMiner then performs automatic trace processing, produces sequence reads by basecalling and identifies in each sequence read all putative sequence features (i.e. vector fragments, adapter/linker sequences, restriction sites and polyA or polyT runs). The resultant data are presented as four views—'Tabulated View', 'XML View', 'Fasta View' and 'Color-coded View' for both normal and reverse complement sequences, as shown in Figure 2 Panel B. Using 'Tabulated View' (see Figure 2, Panel C), a user can obtain a quick overview of sequencing performance and various putative sequence features in perfect and/or imperfect matching patterns. Sequencing performance can be determined easily by the relative proportion of passed (i.e.  $qTag=1$ ) versus failed (i.e.  $qTag=0$ ) sequence reads according to whether or not the sequence reads meet the minimum length threshold of the good quality region for each sequence read after quality trimming. The default threshold is 100 bases and can be customized in terms of users' requirement in the 'Quality Trimming' configuration (see Figure 2, Panel A). The tabulated results in 'Tabulated View' can be sorted in either descending or ascending order by clicking individual column heads in a toggle manner, and they also can be filtered and searched using the expandable 'Sequence Filter' displayed above the tabulated results (see Figure 2, Panel C). 'XML View' presents all processed data in the XML format to enhance data exchangeability and interoperability. Using 'Fasta View', a user can retrieve either whole raw sequence reads or trimmed high-quality portions of sequence reads in the FASTA format. Like 'Fasta View', 'Color-coded View' also furnishes FASTA-format results for all raw sequence reads in a HTML page, but with color-coded bases to indicate putative sequence features, as well as regions have high or low Phred (15) quality scores.

After trace processing and preliminary characterization, a user may choose to proceed with optional database integration by clicking the icon for 'Dumpling XML Data into MySQL Database' (see Figure 2, Panel B). Subsequently, all putative sequence features, with raw sequence reads and corresponding Phred quality values, are populated into the database. Based on the number, location, order, fidelity and orientation of all putative features, sequence features that fit the canonical model of cDNA library construction are then determined for each



**Figure 2.** WebTraceMiner Web Interfaces. Panel (A): The configuration interface allows users to save customized information for how to process EST traces. For general users, the only required information is entered into the ‘Adapter and Restriction Enzyme Site’ section to define a specific canonical model for directional cDNA library construction, such as that illustrated in Figure 1. Panel (B): After processing, resultant data are available as four views—‘Tabulated View’, ‘XML View’, ‘Fasta View’ and ‘Color-coded View’. With satisfactory processing, users can proceed with optional database integration by clicking the database icon. Panel (C): Tabulated results are available from ‘Tabulated View’. The results can be sorted by clicking relevant column heads. They can also be searched and filtered using the expandable ‘Sequence Filter’ on the top. By clicking a sequence name link, a new panel, such as that shown in Panel (D), will pop up automatically. Panel (D): An EST sequence comprising a complete cDNA insert with *in silico* authenticated 3’ and 5’ termini. Both the SVG graph and the color-coded sequence can be redrawn by applying different parameters available in the ‘Sequence Color View Control Panel’.

sequence read and inserted into the database automatically. New 'Tabulated View', 'XML View', 'Fasta View' and 'Color-coded View' will be then available for biologists to query and visualize both putative and *in silico* verified features. Because of the database integration, these newly created views provide far more advanced functionalities than views based on resultant data files (see Figure 2, Panel B). For example, when a user clicks a sequence name link displayed in 'Tabulated View', an HTML page containing the corresponding color-coded sequence read with a scalable vector graphics (SVG) graph is created on-the-fly (see Figure 2, Panel D). This page allows users to inspect individual nucleotides and their associated quality values for putative and/or verified sequence features. The SVG graph can be redrawn with different zooming scales from 25, 50, 75 to 100%, while the color-coded sequence read can also be recreated with or without space separators, to facilitate searching and text capture for other tools, such as BLAST (16). As shown in Figure 2, Panel D, the 'Sequence Color View Control Panel' furnishes several integrated data options that users can use to explore various data. Within the 'Detailed Data' link, the menu items 'All Putative Features' and 'Verified Features' allow users to toggle between web pages for all putative sequence features and for verified sequence features. Within the pages for putative sequence features, users can selectively view each type of sequence features in perfect, imperfect or both matching patterns. In particular, the 'Sequence Feature Table' menu item provides users detailed information about each sequence feature, either putative or verified, including start and stop positions, length, identity percentage (i.e. 100 stands for perfect matches) and matching orientation (i.e. D indicates normal direction, P represents palindrome or reverse complement and 0 is indeterminate). By applying the menu items 'Sequence Quality' or 'Reverse Complement' within the 'Detailed Data' link, users can obtain individual Phred quality scores for all bases or retrieve relevant reverse complement sequences.

A general user can utilize our service anonymously. All anonymous input and output data are stored in our server with a unique identifier. The data files, as well as relevant data entries in our database, are maintained on the server for 7 days. No user data is disclosed, either internally or externally. In the current release (1.0), however, we cannot necessarily guarantee complete confidentiality of any users' data. In the future, we plan to add additional data protection for registered users.

## DESIGN AND IMPLEMENTATION

WebTraceMiner consists of three components: an object-oriented Perl pipeline, a MySQL 5.0 relational database and a PHP web application that communicates with both the pipeline and database. The core of WebTraceMiner is a Perl pipeline, which contains a few classes that have been designed as extensions of Bioperl modules (17). Config.pm contains all configuration information required for executing a trace processing event. It is utilized as a parameter by

almost all other classes, except some utility classes. Config.pm parses a XML configuration file saved by the user and acts as a driving force for all processing procedures. Trace.pm represents individual traces as targets for processing, whereas TraceGroup.pm stands for a grouping of trace(s) from any given processing event. In WebTraceMiner, trace files are processed using Phred (15) or other basecallers to conduct basecalling and create sequence reads. In addition, WebTraceMiner provides several options for vector screening, including CROSS\_MATCH (bozeman.mbt.washington.edu/phrap.docs/phrap.html) and Vmatch (www.vmatch.de). Correspondingly, WebTraceMiner possesses some wrapper classes, including Phred.pm, CrossMatch.pm and Vmatch.pm, which help with executions of the above-mentioned thirty-party software. QualityTrim.pm is a class that characterizes the high-quality region determination for individual sequence reads. HomopolymerDetect.pm is a class representing the detection process of homopolymers, such as polyA/T. SeqRead.pm defines each sequence read obtained from basecalling procedures. As the counterpart of TraceGroup.pm, SeqReadGroup.pm represents the grouping of individual sequence reads, facilitating procedures performed on a group of sequence reads (e.g. writing tab-delimited text and XML output files). The design and implementation of the Perl pipeline provides for easy modification and expansion in the future.

WebTraceMiner adopts open-source MySQL 5.0 (www.mysql.com) as its relational database component. With five tables in its current database schema, the database was designed for simplicity, efficiency, portability and scalability. CONFIG table contains all configuration information required for one trace processing event. SEQUENCE table stores all processed results for a group of traces using a specific configuration, while SEQ\_BASE table houses relevant sequence reads and corresponding quality values. FEATURE table holds all properties (i.e. start and stop positions, length, identity percentage and matching orientation) for putative features, whereas SEQUENCE\_VER houses information for verified sequence features.

The web application of WebTraceMiner was implemented using PHP (PHP: Hypertext Preprocessor, 5.0). We adopted the Smarty Template Engine (smarty.php.net) to speed development cycles, enhance software usability and simplify maintenance. The Smarty Template Engine is a PHP template engine that separates PHP, as a business logic, from HTML, as a presentation logic, to facilitate cleaner coding and more flexible modification of codes. In particular, it generates web contents by placing special 'Smarty' tags within documents and then pre-compiles these codes for faster runtime execution.

## DISCUSSION

While EST analysis remain a dominant approach for characterizing the protein-coding portions of various genomes, freely available and easy-to-use bioinformatics tools for processing EST traces without the hassles of

software installation and configuration are rare. Like preAssemble (14), WebTraceMiner provides a valuable public sequence processing service that allows biologists to process their own EST traces or traces downloaded from NCBI Trace Archive. WebTraceMiner, however, provides greater flexibility in trace processing and sophisticated functionality of data querying and visualization, which are not available in preAssemble. Although general users are only required to furnish sequence information for the canonical model of directional cDNA library construction, advanced users have many different options for configuration to customize their trace processing events. For example, different criteria have been adopted in other software for the detection of polyA/T tails in ESTs (8,10,13,18,19). To investigate the impact of these different criteria, a user might want to compare polyA/T detection using two different criteria, e.g., a minimum length of 12 bases with 2 base maximum internal errors versus a minimum length of 8 continuous bases without any error base. Using two different web sessions, a user can save these two criteria separately in two configuration files, each of which stands for an independent trace processing event, and then upload the same set of traces independently to do comparative trace processing. With robust functionalities in data sorting, filtration, search and visualization, a user can easily appreciate the differences resulting from use of these two criteria. Clearly, WebTraceMiner provides advanced users a handy 'play-and-try' research tool that can be customized in terms of individual research needs.

Expanding at an exponential rate, NCBI dbEST (1) currently houses 40 227 021 entries of ESTs. However, EST data quality is highly variable and suffers from deficiencies inherent to both sequencing (e.g. partial, single-pass sequencing of variable quality) and cDNA library construction procedures (e.g. concatenated adapters/linkers, chimeric genes or inversely inserted cDNAs). The value of these EST sequences can be enhanced tremendously by removing redundancy and grouping them into clusters, as has been done in creating the NCBI UniGene (2) and TIGR Gene Indices (3) resources. However, data quality of individual sequence reads is critical for correct EST clustering and, thus, impacts results from applications using the resultant clusters. Unfortunately, to our best knowledge, there has been until now no public EST resource that allows biologists to validate individual sequence reads by highlighting features in the 3' and/or 5' ends of cDNA inserts.

Before GenBank dbEST submission, ESTs are typically trimmed of vector and adapter/linker sequences, as well as polyA or polyT tails. Unfortunately, existing packages are also limited in their ability to cleanly detect and trim such sequences. For example, preAssemble, which utilizes Pregap4 of the open-source STADEN package (5), proves to be problematic sometimes in detecting polyA/T tails shorter than 10 bases (14). For adapter/linker detection, MAGIC-SPP (13) does a better job than other existing packages, including TIGR Lucy (6), but has problems with inserts having more than two concatenated adapters. Adopting a different approach, WebTraceMiner

first detects in an unbiased fashion all vector fragments, restriction sites, adapter/linker sequences and polyA or polyT runs in each sequence read as putative features. The putative features can be identified in single/multiple occurrence(s), in independent/concatenated status and with perfect/imperfect (i.e. mismatch, insertion or deletion) matching patterns. Based on the canonical model for directional cDNA library construction, WebTraceMiner then examines the number, location, order, fidelity and orientation of the putative features and identifies *in silico* verified sequence features to characterize 3' and/or 5' termini of cDNA inserts. Currently, authenticated polyA/T tails are defined as those immediately adjacent to an appropriate restriction site in correct orientation (e.g. either PolyA + XhoI (CTCGAG) or XhoI (CTCGAG) + PolyT), with an allowance for a minimal number of low-quality bases between the polyA/T tails and the restriction site (see Figure 1). Meanwhile, adjacency between a restriction site and an adapter sequence, both of which must also be in correct orientation (e.g. either EcoRI(GAATTC) + Adapter1(GGCA CGAGG) or Adapter1(CCTCGTGCC) + EcoRI(GAA TTC)) with the same allowance for low-quality bases between restriction sites and adapters, is mandatory for annotating the other terminus of the cDNA insert (see Figure 1). Although *in silico* verified 3' and 5' termini might not be all true in every case because of imperfections in molecular biology manipulations (e.g. oligo-dT primer mispriming from a stretch of adenosyl residues internal to the mRNA transcript), they enable biologists to better create cleaner EST clusters, annotate gene positions, explore 3' and 5' UTRs, and detect functional motifs, among other things.

WebTraceMiner represents a unique service that provides great flexibility in EST trace processing and robust functionality in visualizing and mining both putative and verified sequence features. Through easy-to-use web interfaces, biologists can inspect and validate their data and identify artifacts, such as chimeric cDNAs and adapter concatemers. Clearly, WebTraceMiner can fill the gaps between NCBI dbEST and Trace Archive, and alleviate the limitation in EST data mining imposed by trimmed sequences in dbEST. Focusing on sequence features that characterize the 3' and 5' termini of cDNA inserts, WebTraceMiner provides a unique yet complimentary tool to facilitate data validation and data mining of enormous, error-prone ESTs. Efforts are currently underway to add more functionality and options to WebTraceMiner. For example, where base calls and associated quality values obtained from applying ABI or KB Basecaller (docs.appliedbiosystems.com/pebi docs/04362968.pdf) are embedded in traces, sequence reads will be retrievable via the Staden Package modules (5), as an alternative to using Phred (15). In a near future release, WebTraceMiner will also provide users an option of utilizing BLAST (16) to screen vector and other contaminants, as well as to annotate their sequences. In particular, we are now in the process of improving our classifier to more accurately identify verified features from putative sequence features.

## ACKNOWLEDGEMENTS

The authors thank Quinn Li, Chris Wood and Feng Sun for providing valuable comments on the manuscript. This work was supported by the new faculty start-up grant and CFR Summer Research Award from Miami University to CL. Funding to pay the Open Access publication charges for this article was provided by OhioLINK, Georgia Traditional Industry Program-Pulp and Paper and Miami University Committee on Faculty Research.

*Conflict of interest statement.* None declared.

## REFERENCES

1. Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993) dbEST—database for “expressed sequence tags”. *Nat. Genet.*, **4**, 332–333.
2. Pontius, J.U., Wagner, L. and Schuler, G.D. (2003) UniGene: a unified view of the transcriptome. In *The NCBI Handbook*. Bethesda (MD): National Center for Biotechnology Information, 2003.
3. Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R. and White, J. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.*, **29**, 159–164.
4. Rudd, S. (2003) Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci.*, **8**, 321–329.
5. Staden, R. (1996) The Staden sequence analysis package. *Mol Biotechnol.*, **5**, 233–241.
6. Chou, H.-H. and Holmes, M.H. (2001) DNA sequence quality trimming and vector removal. *Bioinformatics*, **17**, 1093–1104.
7. Ayoubi, P., Jin, X., Leite, S., Liu, X., Martajaja, J., Abduraham, A., Wan, Q., Yan, W., Misawa, E. and Prade, R.A. (2002) PipeOnline 2.0: automated EST processing and functional data sorting. *Nucleic Acids Res.*, **30**, 4761–4769.
8. 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.
9. Paquola, A.C.M., Nishiyama, M.Y. Jr, Reis, E.M., da Silva, A.M. and Verjovski-Almeida, S. (2003) ESTWeb: bioinformatics services for EST sequencing projects. *Bioinformatics*, **19**, 1587–1588.
10. Scheetz, T.E., Trivedi, N., Roberts, C.A., Kucaba, T., Berger, B., Robinson, N.L., Birkett, C.L., Gavin, A.J., O’Leary, B. *et al.* (2003) ESTprep: preprocessing cDNA sequence reads. *Bioinformatics*, **19**, 1318–1324.
11. Aerts, J.A., Jungerius, B.J. and Groenen, M.A. (2004) POSA: perl objects for DNA sequencing data analysis. *BMC Genomics*, **5**, 60.
12. Parkinson, J., Anthony, A., Wasmuth, J., Schmid, R., Hedley, A. and Blaxter, M. (2004) PartiGene – constructing partial genomes. *Bioinformatics*, **20**, 1398–1404.
13. Liang, C., Sun, F., Wang, H., Qu, J., Freeman, R.M. Jr, Pratt, L.H. and Cordonnier-Pratt, M.M. (2006) MAGIC-SPP: a database-driven DNA sequence processing package with associated management tools. *BMC Bioinformatics*, **7**, 115.
14. Adzhubei, A.A., Laerdahl, J.K. and Vlasova, A.V. (2006) preAssemble: a tool for automatic sequencer trace data processing. *BMC Bioinformatics*, **7**, 22.
15. 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.
16. Altschul, S.F., Madden, T.L., Schäffer, 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.
17. Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
18. Zhang, H., Hu, J., Recce, M. and Tian, B. (2005) PolyA\_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res.*, **33**, D116–120.
19. Loke, J.C., Stahlberg, E.A., Strenski, D.G., Haas, B.J., Wood, P.C. and Li, Q.Q. (2005) Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures. *Plant Physiol.*, **138**, 1457–1468.