Published online 22 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W77–W83***
doi:10.1093/nar/gkp389

WebGMAP: a web service for mapping and aligning cDNA sequences to genomes

Chun Liang[1,2,*], Lin Liu[1] and Guoli Ji[3]

[1]Department of Botany, [2]Department of Computer Science and Systems Analysis, Miami University, Oxford, OH 45056, USA and [3]Department of Automation, Xiamen University, Xiamen, Fujian 361005, China

Received February 28, 2009; Revised April 15, 2009; Accepted April 29, 2009

ABSTRACT

The genomes of thousands of organisms are being sequenced, often with accompanying sequences of cDNAs or ESTs. One of the great challenges in bioinformatics is to make these genomic sequences and genome annotations accessible in a user-friendly manner to general biologists to address interesting biological questions. We have created an open-access web service called WebGMAP (http://www.bioinfolab.org/software/webgmap) that seamlessly integrates cDNA-genome alignment tools, such as GMAP, with easy-to-use data visualization and mining tools. This web service is intended to facilitate community efforts in improving genome annotation, determining accurate gene structures and their variations, and exploring important biological processes such as alternative splicing and alternative polyadenylation. For routine sequence analysis, WebGMAP provides a web-based sequence viewer with many useful functions, including nucleotide positioning, six-frame translations, sequence reverse complementation, and imperfect motif detection and alignment. WebGMAP also provides users with the ability to sort, filter and search for individual cDNA sequences and cDNA-genome alignments. Our EST-Genome-Browser can display annotated gene structures and cDNA-genome alignments at scales from 100 to 50 000 nt. With its ability to highlight base differences between query cDNAs and the genome, our EST-Genome-Browser allows biologists to discover potential point or insertion-deletion variations from cDNA-genome alignments.

INTRODUCTION

Sequencing is now underway on the genomes of thousands of organisms (1). At least in eukaryotes, a genome sequencing project is usually accompanied by a cDNA sequencing project, in which cDNA is created by reverse transcription of mature mRNA (or transcript) and then often sequenced from its both ends to obtain Expressed Sequence Tag (EST) sequences. As the active, transcribed portions of various genomes, cDNAs or ESTs represent experimental evidence of the transcription for genes expressed under specific conditions. Aligning a cDNA sequence to the locus from which it was presumably transcribed (i.e. *cis* alignment), or to a homologous locus from the same or another species (i.e. *trans* alignment), provides invaluable information about the existence and exon–intron structure of genes (2). Many tools such as BLAT (3), GeneSeqer (4), ECgene (5), GMAP (6), PALMA (7) and Spaln (8) are available to map and align spliced cDNAs to genome sequences with reasonable accuracy, and each of these programs has its own strengths and limitations.

Despite a great deal of effort through the years, it is still challenging to annotate genomes accurately with correct exon–intron structures and protein-coding regions (2), not to mention non-coding RNAs that temporally and spatially regulate gene expression (9). For example, in EGASP—an assessment of the best gene finders on the human ENCODE regions, only 40–50% of transcript isoforms were predicted correctly by the best programs using all available information (10). Even for the well-studied worm *Caenorhabditis elegans*, the best gene finders when using large numbers of ESTs were only able to predict half of gene models correctly (11). It is clear that complexity among genes, transcripts, and proteins has been underestimated, and that current cDNA and EST collections are far from providing a complete catalogue of transcripts for any eukaryote (2). Nevertheless, ESTs remain the fastest growing DNA sequence resource in GenBank. During the past year, the number of sequence reads in the major public EST repository dbEST (12) has increased from 49 127 466 to 59 861 825, and the number of species covered has increased from 1470 to 1695 (releases 11 January 2008 and 30 January 2009). Moreover, sequences in dbEST are being generated increasingly by next-generation sequencing technologies like 454 pyrosequencing, see (13) for example, which avoids the need for cloning and

*To whom correspondence should be addressed. Tel: +1 513 529 2336; Fax: +1 513 529 4243; Email: liangc@muohio.edu

increases the depth and coverage of sequencing novel and rare transcripts significantly compared with conventional Sanger sequencing (13,14).

As preparing and sequencing cDNAs becomes a routine task in many labs, there is an obvious need for general biologists to access and take advantage of the complete and draft genomes and their annotations, even though some of the gene models might be incomplete, incorrect or missing due to the aforementioned limitations. By mapping and aligning their cDNAs to the desired genomes, biologists can not only examine correct gene structures of interesting genes themselves, but also can explore other important biological processes like post-transcriptional regulation of genes and genetic variations revealed by individual ESTs that are mapped to the same loci. Although many different software tools are available for cDNA-genome mapping, GMAP has been shown to provide higher-quality alignments and better gene structures than BLAT (3) and other programs [e.g. SIM4 (15), MGAlign (16) and Spidey (17)], even in the presence of substantial polymorphisms and sequencing errors (6). GMAP is being used widely by research communities for determining gene structures and exploring alternative splicing (18), alternative polyadenylation (19) and genetic polymorphisms (20,21). GMAP is a standalone program that has a command-line interface and requires knowledge of Unix to install, which limits its usage among general biologists. Here, we create a purely web-based service for cDNA-genome mapping that makes GMAP available to a wider community and integrates many in-house tools as add-ons to facilitate data visualization and mining. The goal of our web service is to provide biologists with an easy-to-use bioinformatics platform that takes advantage of complete and draft genomes and their up-to-date genome annotations and facilitates exploration of important biological, molecular and evolutionary processes.

WEB SERVICES

WebGMAP takes a single sequence file in FASTA format as input, which can be uploaded by a user into our server. Currently, if the file contains more than 200 sequence reads, the server processes up to the first 200 valid sequences. A valid sequence is defined as a DNA sequence that contains bases *A*, *T*, *G*, *C* and *N* (ambiguity code), with a length between 50 and 2500 nt. All the bases other than *A*, *T*, *G* and *C* will be replaced with *N*. After inserting valid sequences into a database residing on our web server, WebGMAP presents a tabulated sequence list that allows users to sort, filter, and search individual sequences by sequence name, GC content, and length, as well as a *seq identifier* generated by the server (see Figure 1A). To examine individual sequences and conduct some routine analysis tasks, users can use a web-based sequence viewer with many useful functionalities, including nucleotide positioning, six-frame translations, reverse complementation and fuzzy motif detection and subsequent sequence alignments of identified motifs. As shown in Figure 1B, nucleotide positional information can be revealed clearly by checking the *Space Separator* checkbox and then clicking the *Redraw* button. Similarly, checking the *Reverse Complement* checkbox and clicking the *Redraw* button will display the reverse complement of the sequence. Selecting the *Protein Translation* checkbox will activate the frame selection for translation. For any given frame (i.e. +1, +2, +3, or –1, –2, –3 if the *Reverse Complement* checkbox is checked), a user can obtain the relevant amino acid sequences by clicking the *Redraw* button. *Motif Search* allows the user to search short sequence patterns in both perfect and imperfect (or fuzzy) match. By default, *Motif Max Error* is set to be 0 for searching perfect matches. For imperfect matching, *Motif Max Error* needs to be a positive number that indicates the maximum error base(s) allowed, regardless of insertions, deletions, mismatches or a mixture of them. After the user provides a target motif sequence pattern in the *Motif Search* text field and then clicks the *Redraw* button, the motif(s), if detected in the query sequence, will be displayed in red. If the *Alignment* checkbox is selected before clicking the *Redraw* button, users can visualize and examine the sequence alignment between the target motif and the motif detected in the query sequence in details (see Figure 1B).

Although our web software and framework can be used for any genome, our web site currently focuses on plant species. We provide different plant genomes for users to choose for cDNA-genome mapping. The default GMAP settings are highly recommended for general users, and GMAP itself has the ability to modify its internal alignment parameters based on its initial assessment of alignment quality. Nevertheless, users are allowed to modify parameters, such as the *max length for one intron* and *max total intron length*. They can also add other parameter options to meet their specific needs if they know how to fine-tune the GMAP standalone program. Our service also has the ability to provide pre-defined sets of parameters that may be useful for particular alignment tasks, such as cross-species or evolutionarily distant alignments. Upon the completion of the mapping process, which usually is the most time-consuming part in our web service, users can download the raw GMAP output directly from our web site onto their local computers. However, users may also choose to continue processing the results by clicking on '*GMAP Result To DB*', in order to use our seamlessly integrated data visualization and mining tools.

The GMAP results are provided in a tabulated list that allows users to sort, filter, and search individual alignments by several criteria, including chromosome name, cDNA sequence name, mapped chromosome start and end positions, and mapped cDNA sequence start and end positions. By following the web link on a particular alignment, a user can take advantage of our EST-Genome-Browser in which all valid alignments will be automatically anchored (or aligned) to the annotated genomes for visualization, in resolution from 100 (i.e. the highest resolution at nucleotide level) to 50 000 nt. Using the EST-Genome-Browser, biologists can carefully examine both the existing community annotation of gene structures and the resulting partial or complete

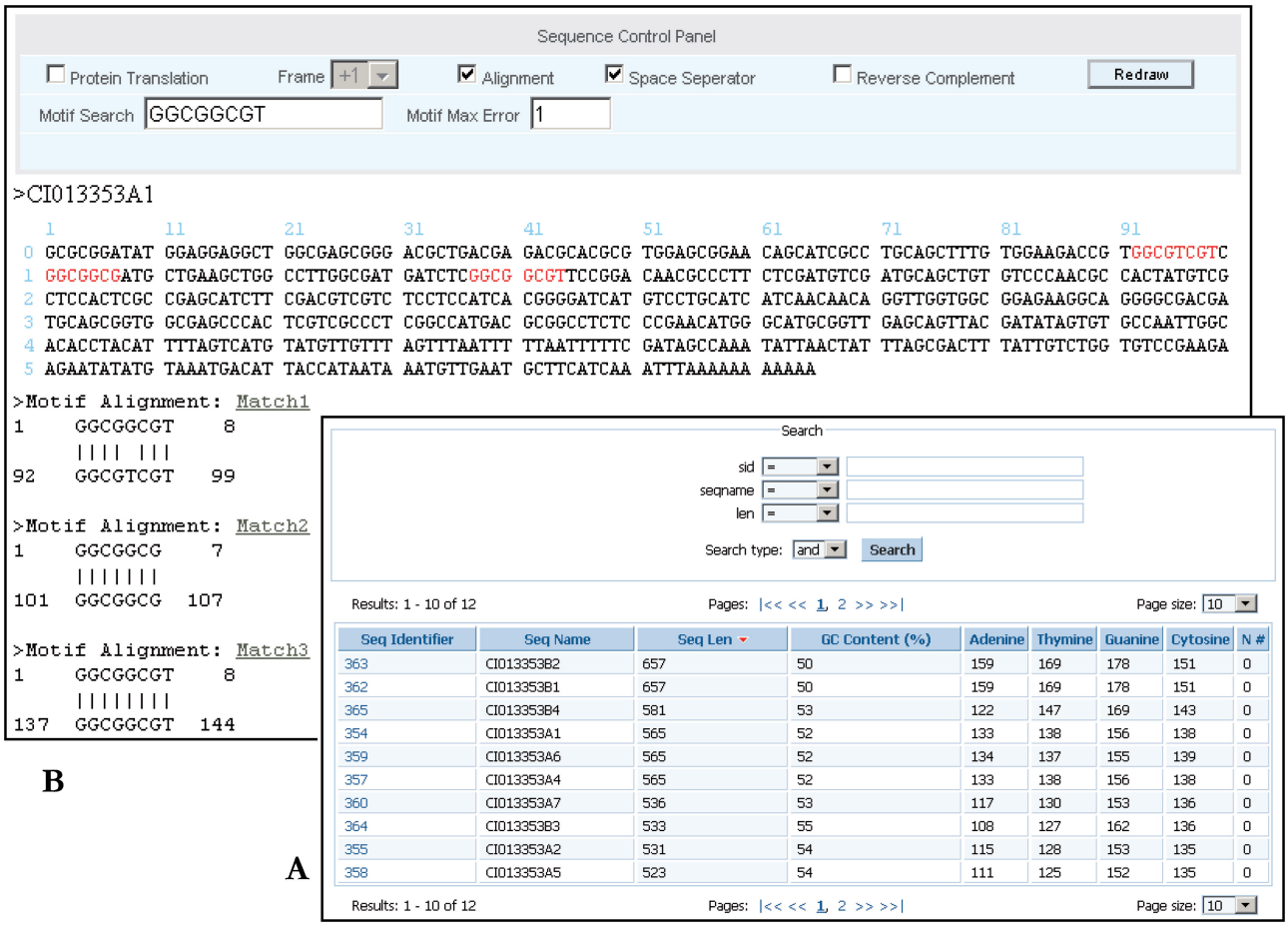

Figure 1. WebGMAP web interfaces for a sequence list and sequence viewer. (**A**) A tabulated sequence list allows users to sort, filter and search among individual valid sequences accepted by WebGMAP. (**B**) A sequence viewer with functionalities like nucleotide positioning, six-frame translations, sequence reverse complementation, and imperfect motif detection and alignment.

gene structures derived from their query cDNA sequences. As shown in Figure 2 (Upper panel), individual EST-genome alignments suggest two putative gene models, one with and one without an intron retention. Because of its capability to distinguish insertion, deletion and mismatch bases with different colours, our EST-Genome-Browser can be used to explore genetic polymorphisms of either single (i.e. SNPs) or multiple nucleotides revealed by individual ESTs that are mapped to the same loci in the genome. For example, Figure 2 (Lower panel) shows one insertion, one mismatch and one deletion, which might be potential SNPs, at genomic coordinates 11042, 11048, and 11050–11051 respectively. Moreover, our EST-Genome-Browser provides the useful ability to sort, filter, and search alignments. Clicking the *Filter* checkbox at the top of EST-Genome-Browser activates the *Genome*, *Gene* and *EST* tabs. These tabs provide users with multiple ways to sort, filter, and search various tracks of chromosomes, genes, and ESTs displayed in the browser. Using *Gene* Tab, for instance, a user can selectively examine all cDNA sequences that are mapped to the positive or negative strand of a genome. For another example, *EST* Tab allows a user to filter and focus on a single cDNA sequence for further scrutiny.

A user can utilize WebGMAP anonymously without any login requirement. Each submission of an uploaded sequence file that is aligned against a user-specified genome with a set of user-specified parameters is defined as a single WebGMAP project. There are no limitations on the number of projects that a user can create. Once a sequence file is uploaded, a project with a name like *guest@muohio.edu_ymmdd_hhmmss* (i.e. guest@muohio.edu_090217_115231) will be created automatically. The project name is unique and serves as a key for storing and accessing all associated data residing on our server and database. Due to our limited resources, all data associated with a specific project are kept for 1 week. Currently, we can not guarantee complete confidentiality of any data uploaded by a user, although we are committed not to disclose users' data either internally or externally.

DESIGN AND IMPLEMENTATION

WebGMAP consists primarily of web front-end applications, a back-end database, and a few Perl scripts that populate data into the database. *Seq2DB.pl* is a Perl script that filters out invalid sequence reads within

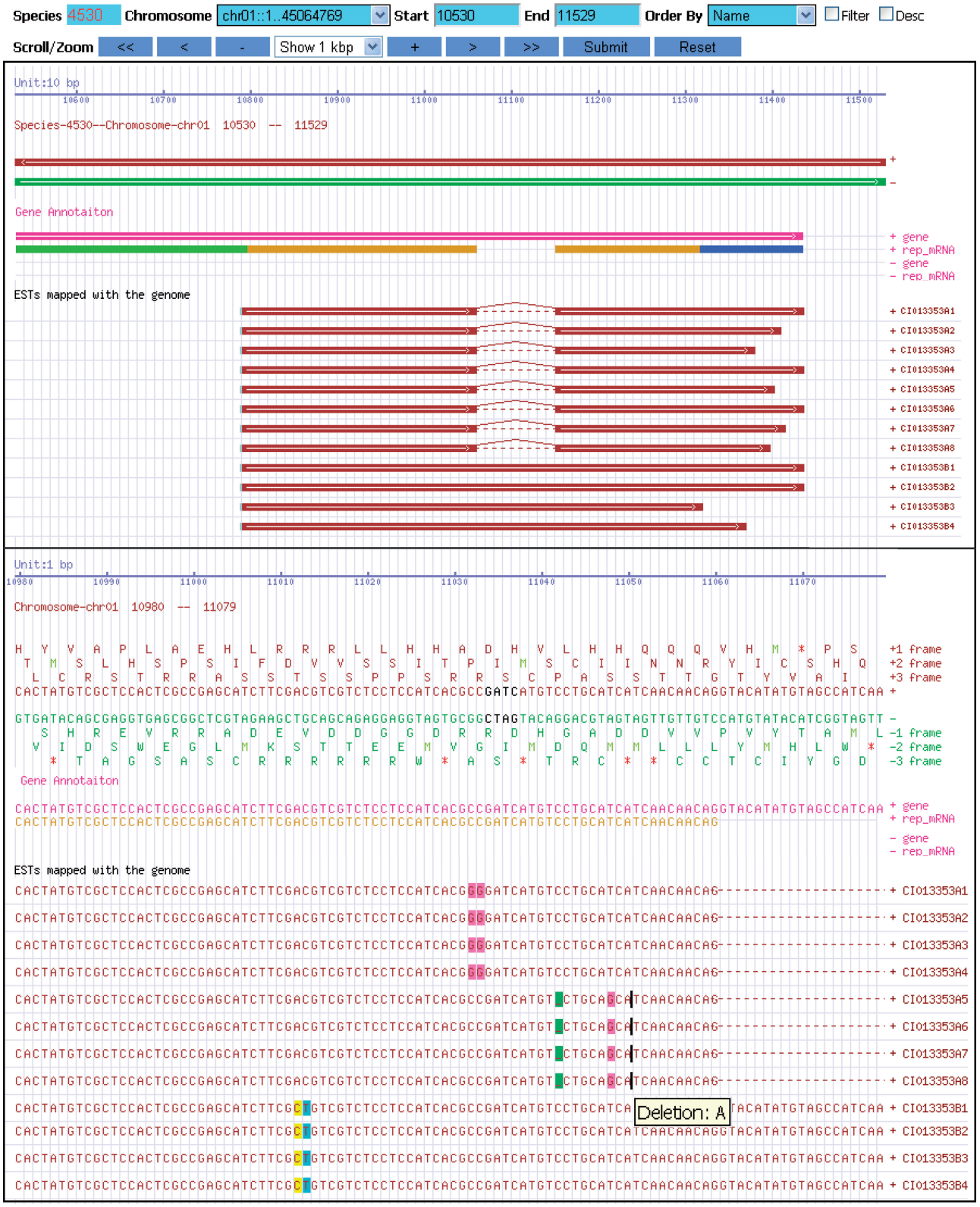

Figure 2. Snapshots of EST-Genome-Browser that reveal potentially different gene models and putative genetic polymorphisms. Upper panel: EST-genome alignments suggest two putative gene models, one with and one without an intron retention. The 5′ UTR, CDS and 3′ UTR of the representative gene model of a given locus from the community genome annotation are highlighted in green, orange and blue respectively. Lower panel: Detailed EST-genome alignments show the exon–intron boundaries and potential polymorphisms of single (SNP) or multiple nucleotides. Insertions are displayed as '-' in green background colour, deletions as '|', and mismatches in background colours that are configurable in terms of nucleotide type and matched orientation. When a user moves the cursor over a particular deletion position, the actual base(s) in the deletion will be displayed.

a given FASTA sequence file, extracts the sequence name, nucleotide contents and other information from valid reads, and inserts the data into the database. *GMapParser.pl* is a Perl script that parses GMAP output files and saves the resultant alignment and gene structure information into the database. *GffParser.pl* is a Perl script for parsing community genome annotation in GFF format into the database for a given genome.

We use MySQL 5.0 (http://www.mysql.com) to implement our back-end database, which has six tables in its schema. The CHROMOSOME table contains information about species, chromosome numbers, chromosome lengths and the file location of the nucleotide contents for each chromosome. We do not save chromosome sequences explicitly in our database, but use indexed flat files instead for fast data retrieval. The GENE_ANNOTATION table houses genomic coordinates of community gene structure annotations including 5′-UTR, CDS and 3′-UTR. The SEQBASE table stores the nucleotide contents and identifiers of cDNA sequences entered by users for subsequent display and filtering of those sequences in the web interfaces shown in Figure 1. Three tables store GMAP results at three different levels: the GMAP_HIT table holds information about individual valid cDNA-genome hits, including mapped chromosome strand, mapped chromosome start and end positions, mapped cDNA sequence start and end positions, and total number of HSPs (High Score Pairs) per hit; the GMAP_HSP table stores properties of individual HSP for a given hit; and the GMAP_DETAIL table records positional base differences, either indels or mismatches, between query cDNA sequences and the genome sequence. We have found our database design to be effective for extracting genetic polymorphism information for further data mining and for facilitating data visualization.

Using the Smarty Template Engine (http://www.smarty.net) and other open-source tools, we have implemented our web portal using PHP 5.0 (http://www.php.net). The Smarty Template Engine is a PHP template engine that separates PHP from HTML to facilitate cleaner programming and flexible modification of source code. Our EST-Genome-Browser is a CGI-based web application coded using C++. Utilizing MySQL++ package (http://tangentsoft.net/mysql++/) and RudeCGI™ C++ CGI Library (http://rudeserver.com/cgiparser/), it takes advantage of the C++programming language for efficient database connections and fast data retrieval over the Internet. Its responsiveness and its capability to handle indels and mismatches benefit from our database design which stores GMAP results as alignment differences rather than the full alignments.

DISCUSSION

According to the Genomes OnLine Database (GOLD) (1) as of January 2009, there are about 900 completely sequenced genomes and over 3000 genomes being sequenced (http://www.genomesonline.org/). One of the great challenges in bioinformatics is to make these genomic sequences and associated genome annotations accessible in a user-friendly manner to general biologists to address interesting biological questions.

Many cDNA-genome mapping programs have provided online versions of their software for use by the community. However, these web-based tools typically provide little or no integration with either existing genome annotations or data visualization and analysis tools. GeneSeqer@PlantGDB is a web server for gene structure prediction dedicated to plant species (22) (http://www.plantgdb.org/cgi-bin/GeneSeqer/PlantGDBgs.cgi). The unique strength of this web service is its integration of up-to-date public plant sequence data (e.g. ESTs, EST assemblies and full-length cDNA) in the PlantGDB database (http://www.plantgdb.org) with the spliced alignment capability of GeneSeqer software (4). Although it can produce preliminary, coarse graphical alignments for mapped cDNA or ESTs, its lack of integrated genome annotations or detailed multiple sequence alignments at the nucleotide level limits its popularity among general biologists. One exception to the generally poor utility of web-based genome tools is Human Blat Search (http://genome.ucsc.edu/cgi-bin/hgBlat), which presents the community with a useful web service for mapping and aligning cDNAs to genomes. This site integrates a selected set of metazoan genomes and relevant gene annotations with the UCSC Genome Browser to allow users to explore gene structures and SNP polymorphisms in query sequences in a highly interactive manner. Although multiple sequence alignment at nucleotide level is not provided for query sequences, UCSC Genome Browser is able to highlight the positions of mismatches, insertions and deletions in query ESTs relative to the reference genome sequence. Unfortunately, the web server is limited to processing submissions of only to 25 sequences at a time.

We believe that our WebGMAP will be a valuable addition to the community, not only by making the high-quality alignments and gene structures produced by GMAP more accessible, but also by providing users with the many valuable add-ons we have integrated to enhance the usability and functionality of our web service. As demonstrated in Figure 1, we provide biologists a sequence viewer that has many useful functions, including nucleotide positioning, 6-frame translations, sequence reverse complementation, and fuzzy motif detection and alignment. In addition, our WebGMAP has robust data sorting, filtration and search functions both for cDNA sequences and for cDNA-genome alignments, which should greatly facilitate data analysis. To our knowledge, none of the aforementioned web services provide biologists with such functionality. In particular, due to our special database design in which we only store alignment differences relative to the genome, our EST-Genome-Browser can display multiple-sequence alignments rapidly at the nucleotide level and distinguish genetic point and indel variations. Furthermore, our EST-Genome-Browser allows the user to manipulate, sort, and filter individual alignments, which can also facilitate data visualization and mining. For instance, all cDNA-genome alignments can be sorted by sequence name, mapped start or end positions in terms of genome, matched strand, and so

on. Users can selectively examine one strand of the genome, one gene and/or one EST at a time for gene structure or sequence alignment.

In contrast with the focus of Human Blat Search on vertebrate and other metazoan genomes, our web server is currently designed to concentrate on plant genomes while integrating other model organisms, and we strive to provide our users with the up-to-date plant genome annotations. For example, our server provides information on the TAIR8 Genome release for *Arabidopsis thaliana*, Rice Annotation Project (RAP) build 4.0 for *Oryza sativa*, and AUGUSTUS 5.0 for *Chlamydomonas reinhardtii*, all of which are widely accepted as the community gene annotations. Some data suggest that plants and animals might have different mechanisms for splice site recognition, and that splicing regulatory signals are less conserved in plants than in animals (23). Therefore, it may be more challenging to create accurate gene structures for plant genomes. It would be interesting to compare GMAP and other tools like BLAT on their performance in plant genomes. WebGMAP is designed to have flexibility in its methodology for cDNA-genome mapping. At this moment, WebGMAP provides users with the capability to perform cDNA-genomic alignment with different GMAP parameter settings and versions. In the future, we should be able to incorporate other alignment programs like BLAT so that users can compare cDNA-genome mappings using two different programs side-by-side. Such a capability might provide the community with an interactive research tool not only for *in silico* biological experimentation (e.g. exploration of alternative splicing in plants), but also for computational feedback to improve the accuracy of cDNA-genome mapping tools.

We are working hard to integrate the genome sequences and their annotations for more plant species as well as other model organisms like *C. elegans*. Efforts are also underway to improve and enhance functionalities of our EST-Genome-Browser to satisfy the needs of biologists. For example, using AJAX technique, current multiple sequence alignments in all zoom scales can be dragged and moved only vertically, not horizontally. Also our server currently handles only cDNA sequences with length greater than 50 nt, which excludes short 25–50 nt reads generated from Illumina sequencing. However, software tools, including one called GSNAP (Thomas D. Wu, personal communication) that is compatible with GMAP databases, are being produced for handling such short reads, and we should be able to incorporate such functionality in the future.

WebGMAP is intended to leverage our enormous wealth of existing genomic sequences and annotations by providing users with easy-to-use web-based tool for performing and analyzing cDNA-genome alignments. Our web service allows users to map their transcript data to the genome and to visualize, examine and mine sequence alignments to obtain accurate gene structures and infer genetic polymorphisms. Using our tools, more biologists should be able to contribute towards improving the annotation of particular genes or gene families. We are working hard to make our core programs such as our web-based sequence viewer and EST-Genome-Browser open source so that they can benefit the larger community.

ACKNOWLEDGEMENTS

The authors thank Yuansheng Liu, Min Dong and Jinqiao Chen for their participation in coding of EST-Genome-Browser. We greatly appreciate Thomas D. Wu for providing us new versions of GMAP and helpful discussions about the project as well as the manuscript. Special thanks go to Oliver Vallon, Sara Savage, Sumeta Sachdeva and Praveen Kumar Raj Kumar for testing the software and providing valuable suggestions to improve the system.

FUNDING

The new faculty start-up grant from Miami University and a grant award from the Ohio Plant Biotechnology Consortium to CL; National Natural Science Fund of China (Project #60774033), Natural Science Foundation of Fujian Province in China (Project #B0710031) and Specialized Research Fund for the Doctoral Program of Higher Education (Project #20070384003) awarded to GJ. Funding for open access charge: Department of Botany and OARS PREP Program, Miami University.

Conflict of interest statement. None declared.

REFERENCES

1. Liolios,K., Mavromatis,K., Tavernarakis,N. and Kyrpides,N.C. (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **36**, D475–D479.
2. Brent,M.R. (2008) Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat. Rev. Genet.*, **9**, 62–73.
3. Kent,W.J. (2002) BLAT – the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
4. Usuka,J., Zhu,W. and Brendel,V. (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics*, **16**, 203–211.
5. Kim,N., Shin,S. and Lee,S. (2005) ECgene: genome-based EST clustering and gene modeling for alternative splicing. *Genome Res.*, **15**, 566–576.
6. Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.
7. Schulze,U., Hepp,B., Ong,C.S. and Rätsch,G. (2007) PALMA: mRNA to genome alignments using large margin algorithms. *Bioinformatics*, **23**, 1892–1900.
8. Gotoh,O. (2008) A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Res.*, **36**, 2630–2638.
9. The ENCODE Project Consortium. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
10. Guigó,R., Flicek,P., Abril,J.F., Reymond,A., Lagarde,J., Denoeud,F., Antonarakis,S., Ashburner,M., Bajic,V.B., Birney,E. *et al.* (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.*, **7(Suppl. 1)**, S2.1–31.
11. Coghlan,A., Fiedler,T.J., McKay,S.J., Flicek,P., Harris,T.W., Blasiar,D., the nGASP Consortium, and Stein,L.D. (2008) nGASP – the nematode genome annotation assessment project. *BMC Bioinformatics*, **9**, 549.

12. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST – database for "expressed sequence tags". *Nat. Genet.*, **4**, 332–333.
13. Emrich,S.J., Barbazuk,W.B., Li,L. and Schnable,P.S. (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.*, **17**, 69–73.
14. Cheung,F., Haas,B.J., Goldberg,S.M., May,G.D., Xiao,Y. and Town,C.D. (2006) Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics*, **7**, 272.
15. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.
16. Lee,B.T., Tan,T.W. and Ranganathan,S. (2003) MGAlignIt: A web service for the alignment of mRNA/EST and genomic sequences. *Nucleic Acids Res.*, **31**, 3533–3536.
17. Wheelan,S.J., Church,D.M. and Ostell,J.M. (2001) Spidey: a tool for mRNA-to-genomic alignments. *Genome Res.*, **11**, 1952–1957.
18. Haas,B.J. (2008) Analysis of alternative splicing in plants with bioinformatics tools. In Reddy,A.S.N. and Golovkin,M. (eds), *Nuclear pre-mRNA Processing in Plants: Current Topics in Microbiology and Immunology,* Vol. 326, Springer, Berlin Heidelberg, pp. 17–37.
19. Shen,Y., Liu,Y., Liu,L., Liang,C. and Li,Q.Q. (2008) Unique features of nuclear mRNA poly(A) signals and alternative polyadenylation in *Chlamydomonas reinhardtii*. *Genetics*, **179**, 167–176.
20. Guryev,V., Koudijs,M.J., Berezikov,E., Johnson,S.L., Plasterk,R.H., van Eeden,F.J. and Cuppen,E. (2006) Genetic variation in the zebrafish. *Genome Res.*, **16**, 491–497.
21. Kaminker,J.S., Zhang,Y., Waugh,A., Haverty,P.M., Peters,B., Sebisanovic,D., Stinson,J., Forrest,W.F., Bazan,J.F., Seshagiri,S. *et al.* (2007) Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Res.*, **67**, 465–573.
22. Schlueter,S.D., Dong,Q. and Brendel,V. (2003) GeneSeqer@PlantGDB: gene structure prediction in plant genomes. *Nucleic Acids Res.*, **31**, 3597–3600.
23. Barbazuk,W.B., Fu,Y. and McGinnis,K.M. (2008) Genome-wide analysis of alternative splicing in plants: opportunities and challenges. *Genome Res.*, **18**, 1381–1392.