Published online 22 April 2013 *Nucleic Acids Research, 2013, Vol. 41, Web Server issue* ***W29–W33***
doi:10.1093/nar/gkt282

BLAST: a more efficient report with usability improvements

Grzegorz M. Boratyn, Christiam Camacho, Peter S. Cooper, George Coulouris, Amelia Fong, Ning Ma, Thomas L. Madden*, Wayne T. Matten, Scott D. McGinnis, Yuri Merezhuk, Yan Raytselis, Eric W. Sayers, Tao Tao, Jian Ye and Irena Zaretskaya

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 45, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received January 30, 2013; Revised March 28, 2013; Accepted March 31, 2013

ABSTRACT

The Basic Local Alignment Search Tool (BLAST) website at the National Center for Biotechnology (NCBI) is an important resource for searching and aligning sequences. A new BLAST report allows faster loading of alignments, adds navigation aids, allows easy downloading of subject sequences and reports and has improved usability. Here, we describe these improvements to the BLAST report, discuss design decisions, describe other improvements to the search page and database documentation and outline plans for future development. The NCBI BLAST URL is http://blast.ncbi.nlm.nih.gov.

INTRODUCTION

Sequence alignments often provide the first connection between newly sequenced DNA or protein and already categorized sequences. Basic Local Alignment Search Tool (BLAST) (1) is one of the more popular choices for searching and aligning sequences. BLAST takes a nucleotide or protein sequence as input and searches it against a database of nucleotide or protein sequences. BLAST can translate nucleotide sequences as needed; therefore, BLAST can search a nucleotide query against a protein database or a protein query against a nucleotide database. BLAST uses heuristics to accelerate searches. BLAST also provides statistics that estimate the likelihood of a match occurring by chance. The National Center for Biotechnology Information (NCBI) provides BLAST both as a stand-alone application (2) and through its website (3–5). The BLAST website produces results with links to other resources at the NCBI. It also offers fast indexed megaBLAST (6) searches against the human and mouse genomes, as well as the nucleotide collection (nt), which is a database comprising 44 billion bases of annotated GenBank and NCBI RefSeq nucleotide sequences. Finally, BLAST offers sensitive protein–protein searches using PSI-BLAST (1,7) and DELTA-BLAST (8). The website uses a specialized queueing system that spreads the search across 10 3 GHz cores and returns results quickly. It gives priority to users who have few outstanding requests, as they are typically the interactive users. The searches are generally fast. For interactive users, the median search time for the indexed megaBLAST service for queries of moderate length (<2000 bases) against the nucleotide collection is typically ~4 s, and overall the median search time for the website is typically 14 s.

Much research has been performed on the algorithmic and queueing aspects of BLAST (1,6–8). Johnson *et al*. (3) presented an extensive redesign of the BLAST submission pages. Less work has been done at the NCBI on improving the presentation of BLAST results. The PowerBLAST client (9) introduced the graphical presentation that eventually became the BLAST Graphical Overview in the NCBI BLAST report. Ye *et al*. (5) presented improvements to the alignment display. However, these improvements did not address a number of usability issues in the BLAST reports at the NCBI website.

Here, we report on improvements to the BLAST website. We describe a redesigned BLAST report that is easier to use, more flexible and fixes many usability issues. Additionally, we report on other improvements to the website.

Later in the text, we refer to the redesigned BLAST report as the 'new' report and the previous one as the 'old' report.

THE BLAST REPORT

The HTML BLAST report at the NCBI website is based on a text report for a stand-alone program, which consists of several sections. First, the header lists the search performed, the query and database and the BLAST version. Second, the table of descriptions summarizes the results and presents the subject sequence identifier (accession), title and statistics about the match. Finally, the alignment

*To whom correspondence should be addressed. Tel: +1 301 435 5987; Fax: +1 301 480 0814; Email: madden@ncbi.nlm.nih.gov

section presents the full sequence title, additional accessions and titles for redundant sequences in the database, the length of the subject sequence, information about the score of the match, as well as the actual alignment. The HTML version of the report presented at the NCBI website is a modified version of the text report. It includes links to other reports such as 'taxonomy reports' and a 'distance tree' immediately after the header, followed by the BLAST Graphical Overview. A table contains the subject sequence descriptions, as well as subject sequence identifiers (hyperlinked to other NCBI resources), and links to the alignments further down the report.

There are a number of issues with the old HTML BLAST report. The linking to other NCBI resources is inconsistent. Sequence identifiers in the descriptions and alignments sections normally link to a GenBank or GenPept report in Entrez, but for assembled genomes they link to the Mapviewer. The old BLAST report uses one-letter icons to link to other NCBI resources, such as 'G' for Gene or 'U' for UniGene, but these icons are not obvious to users. Some users are more familiar with BLAST than the rest of the NCBI website; therefore, they might not know what information is provided by Gene or the difference between Gene and UniGene. The title is often truncated, especially for longer titles. There are almost no navigational links in the alignments section of the report and no convenient way to move to the next alignment or return to the top of the page. It is important to present the report as quickly as possible, but in some cases, formatting the alignments can delay loading of the page and can consume substantial resources on the user's desktop. Users also often look at only a few alignments. Because of these considerations, the old BLAST report prints all of the descriptions and only half of the alignments by default. To see all the alignments, the user needs to reformat the report. Users may not know how many alignments they want to examine until they start looking at the report. They could initially format either too many or too few alignments. Users have also requested the ability to conveniently download FASTA for subject sequences, as well as XML or BLAST reports. Additionally, the old report does not include links to the newly developed graphical sequence viewer that can be used to display BLAST alignments. Despite the limitations discussed here, users are familiar with the basic format of the BLAST report and find it useful.

A new BLAST report addresses the aforementioned issues without changing the basic structure of the report. Later in the text, Figures 1–3 use a megaBLAST search of the genomic region for the gulonolactone (L-) oxidase gene of *Rattus norvegicus* (bases 48 898 799–48 921 150 of NC_005114.3) against the nucleotide collection (nt) to demonstrate new report features. This search uses default BLAST parameters, except that rodent repeat filtering is enabled. The header and BLAST Graphical Overview are unchanged from the old report. The table of descriptions, presented in Figure 1, is different from the old report. For most sequences, as the title is more informative than the accession, the title is in the leftmost column of the table. The title is followed by statistics describing the quality of the match. Because a subject sequence may have multiple separate alignments to the query, both the highest scoring alignment and the total score of all alignments are presented (max score and total score). Query coverage describes what percentage of the query length matches the subject sequence. The expect value describes the statistical significance of the match and 'Max ident' the per cent identity of the match with the highest identity. Per cent identity is calculated from the number of identical letters divided by the alignment length, where the alignment length is the number of matching letters plus the number of gaps for either the query or the subject. Finally, on the right-hand side is the accession of the subject sequence (hyperlinked to the GenBank or GenPept style report). The table of descriptions can be sorted by clicking on numerical column headers. Columns can also be hidden using the gear icon on the right side of the table of descriptions (Figure 1). The table of descriptions has been optimized to show more of the title for the subject sequence, as the browser window is widened. The new report provides additional download options, such as FASTA, for the full or aligned portion of subject sequences, GenBank reports and various BLAST reports. It also provides links to other resources at the NCBI, such as the graphical sequence viewer and the distance tree. For example, Figure 2 shows the graphical sequence viewer display of a query sequence and selected aligned subject sequences available from the descriptions table.

To address the issues previously mentioned with alignment loading, the new report uses asynchronous JavaScript (AJAX) to format the alignments as needed. In most cases, the new report quickly loads alignments for the top five subject sequences. A user may obtain more alignments in several ways. The first way is to simply scroll down the page. The BLAST formatter will print alignments as needed. A user may also select any title in the table of descriptions. The BLAST formatter then prints the alignments for that subject sequence, as well as four before and after it, and moves the focus to the first match for the selected sequence. If the user scrolls up or down from this alignment, the formatter prints other alignments as needed. Finally, a user can select a bar in the BLAST Graphical Overview. The behavior is the same as the selection of the title in the table of descriptions. In the old report, an attempt to select an alignment (through the BLAST Graphical Overview or the table of descriptions) that had not been formatted would have simply resulted in no action.

Figure 3 presents a set of alignments from a search. A shaded bar begins the presentation of alignments for a subject sequence. Below the bar is information about the subject sequence as well as the match. There is a download menu as well as a link to Genbank and the Graphical sequence viewer. In the middle of the shaded bar there is a pull-down menu that controls the sorting of the matches for subject sequences with more than one match. To the right of the alignment the words next, previous and descriptions (along with some arrows) serve as navigation aids. Below these navigation aids, there are links to

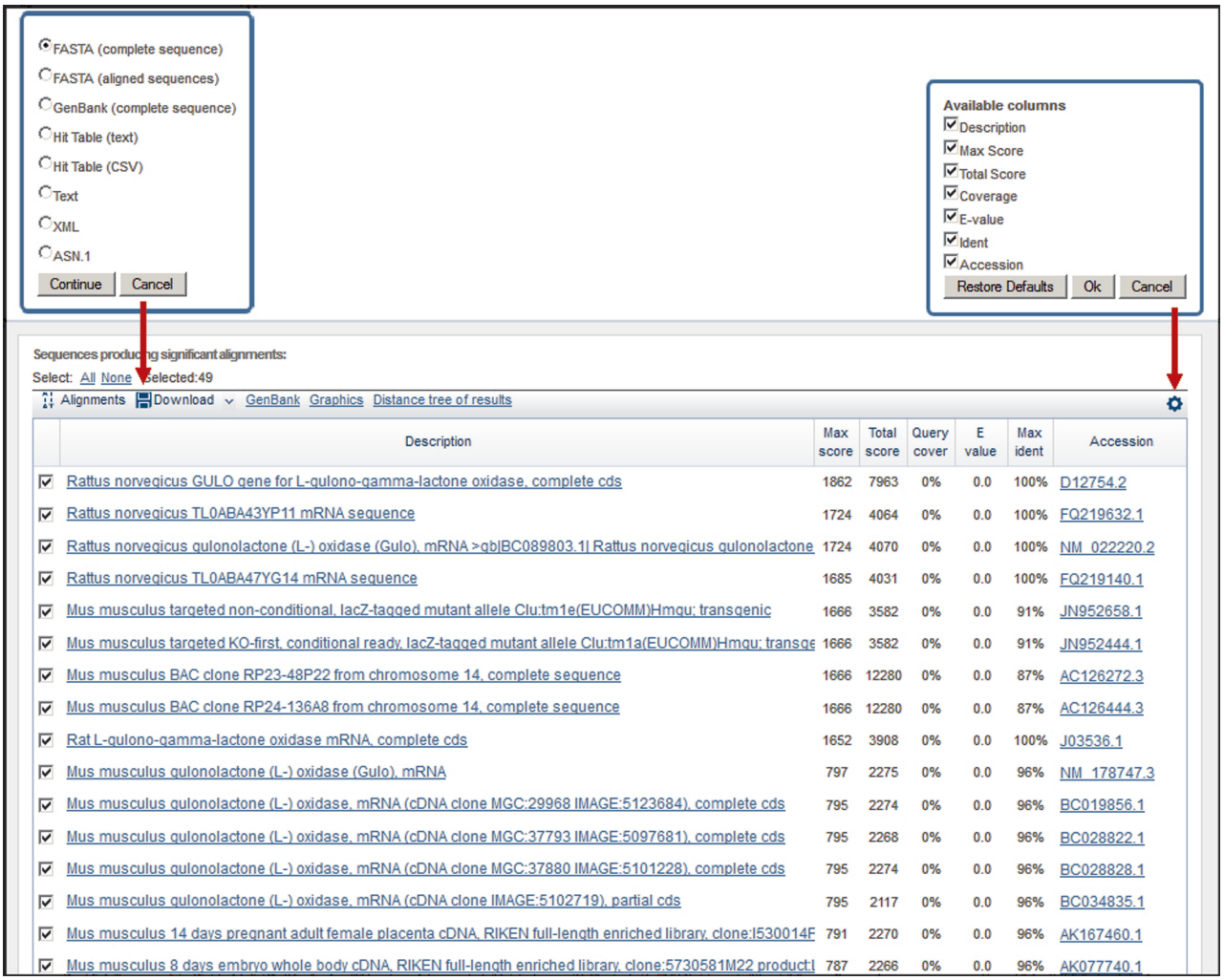

Figure 1. Table of descriptions for a search of the genomic region for the gulonolactone (L-) oxidase gene of *Rattus norvegicus* (bases 48 898 799–48 921 150 of NC_005114.3) against nt. Selecting the title loads the alignments for that sequence (if needed) and moves the focus to that alignment. Selecting an accession on the right opens a GenBank report on that sequence. The download menu (left side), as well as the GenBank, graphics and distance tree views, can be enabled by selecting checkboxes on the left side. So as not to cover the table, the download menu is shown above where it would actually open. The gear icon on the right side can be used to control the columns shown. Its menu is shown above where it actually opens.

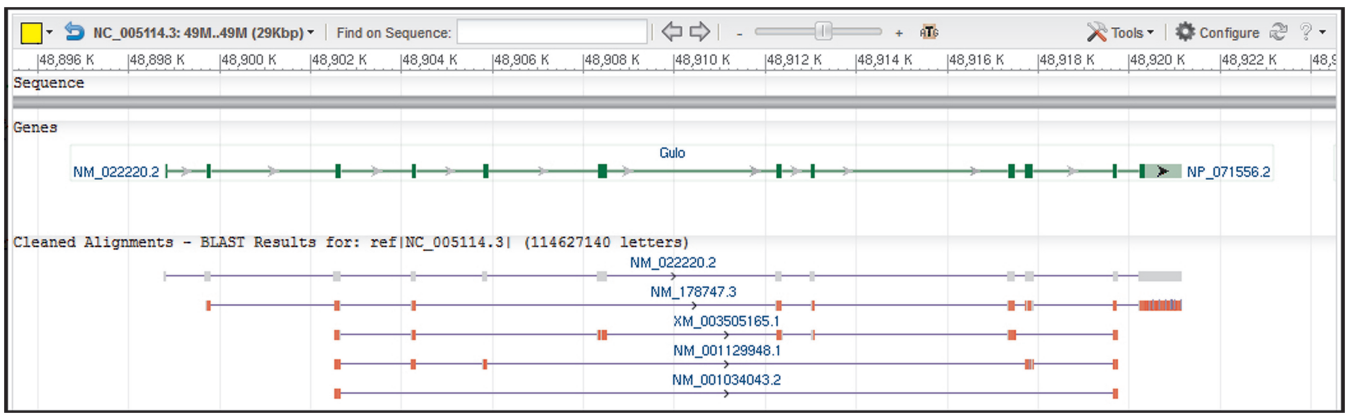

Figure 2. Example view of query and selected subject sequences in the graphical sequence viewer. The subject sequences are selected mRNAs found by a megaBLAST search against nt with the genomic region for the gulonolactone (L-) oxidase gene of *Rattus norvegicus* (bases 48 898 799–48 921 150 of NC_005114.3) as the query. The viewer shows the query sequence at the top of the figure. Next, the Gulo gene feature on the query sequence is shown in green. This track was enabled with the 'configure' feature of the viewer. Finally, the subject sequences are shown as 'cleaned alignments', which presents an overview of the alignments that groups matches from the same subject sequence with a thin line and adds an accession label. The red color in the subject sequences represents mismatches to the query.

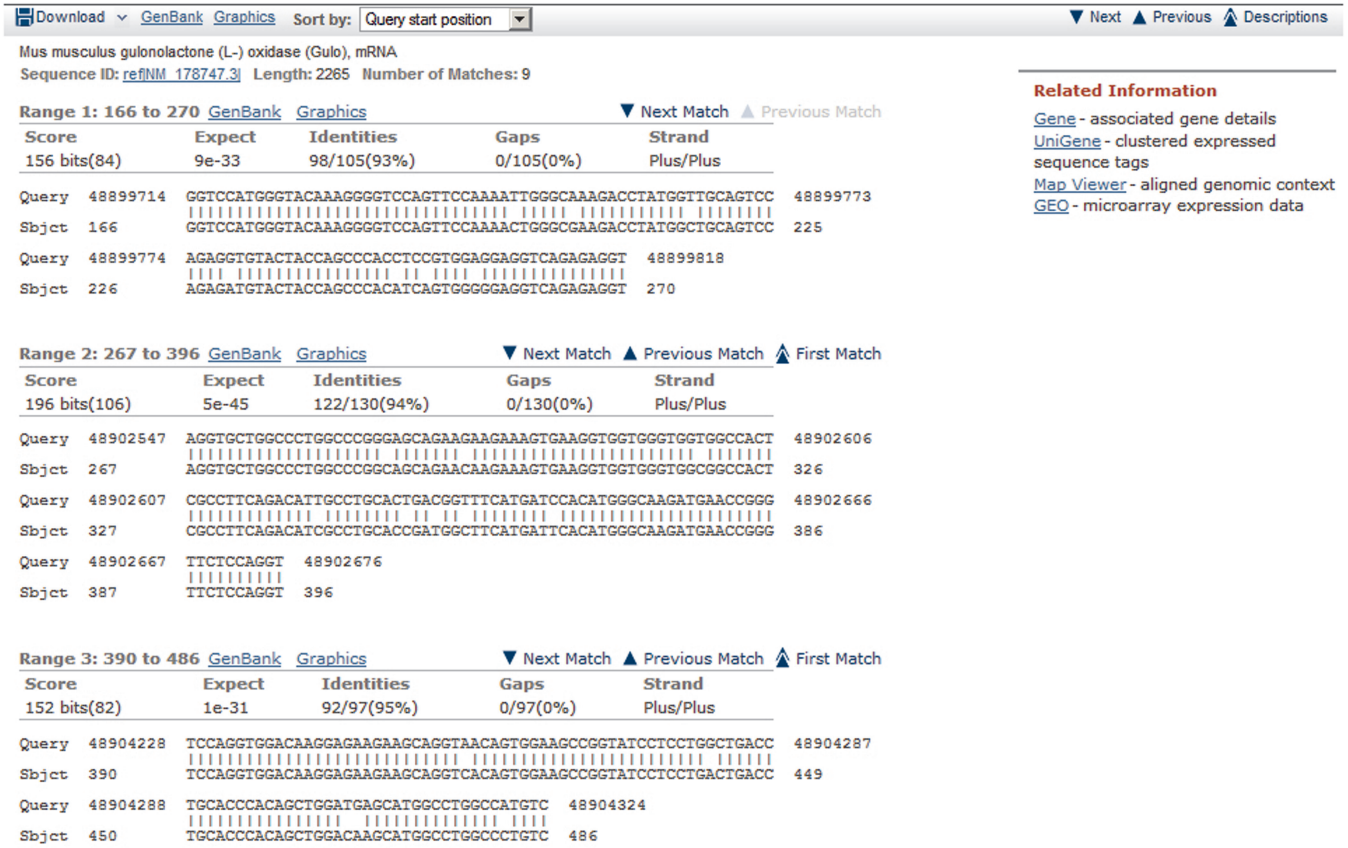

Figure 3. Alignments from a search of the genomic region for the gulonolactone (L-) oxidase gene of *Rattus norvegicus* (bases 48 898 799–48 921 150 of NC_005114.3) against nt. The download menu and links for GenBank and graphic apply only to this subject sequence, but they have the same behavior as described for the descriptions earlier in the text. New navigation aids are part of the alignment display. 'Next' ('previous') moves the focus of the report to alignments for the next (previous) subject sequences. 'Descriptions' takes the focus back to the line for the current subject sequence in the table of descriptions. For subject sequences with more than one match, 'next match' and 'previous match' move the focus of the report to the next or the previous match for that subject sequence.

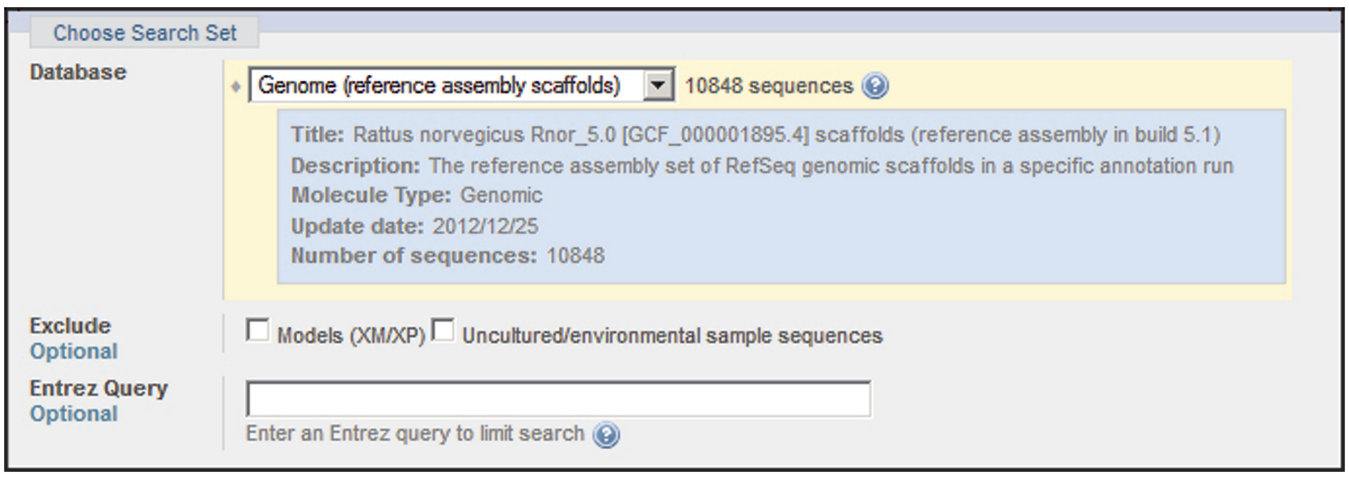

Figure 4. Automatically generated documentation for the rat genome database. The information is stored in the blastdb_info database and formatted on demand. Information, such as last update, number of sequences and type or sequence ('genomic'), is automatically generated when the database is constructed. The database curator provides the title and description.

'related information'. These links spell out the name of the resource and provide a short description.

The BLAST help tab, accessible from the BLAST home page at http://blast.ncbi.nlm.nih.gov, has links to a video (from the NCBI YouTube channel), as well as a document about using the new BLAST report.

OTHER IMPROVEMENTS

In the past, the NCBI has provided different BLAST query pages for assembled Refseq genomes and microbial sequences. Users needed to adjust to a different search page when they moved between the standard page [using the design of Johnson *et al*. (3)] and these genomic pages. Some features of the standard page, such as the ability to 'edit and resubmit' a search from the BLAST report or to save a search strategy for later execution, did not work on the genomic pages. We have converted many genomic pages during the past year to a design similar to the one of Johnson *et al*. As a result, users now find it much easier to move from one NCBI BLAST search page to another, and the cost to support these pages is lower for the NCBI.

We have also implemented a new system ('blastdb_info') to store metadata for BLAST databases. This metadata includes the specification of the database, the type of sequence (e.g. genomic or cDNA), organism information and comments added by the NCBI database curator. Figure 4 demonstrates the documentation available for the rat reference genome available at the pull-down menu for BLAST databases.

FUTURE DEVELOPMENT

We plan to continue to improve the NCBI BLAST website. For example, the descriptions table is a flexible design, and it would be possible to add new columns (such as one for taxonomy). We also plan to make blastdb_info searchable through a web interface, so that users can search for databases based on taxonomy, type of sequence or keywords. Finally, we plan to improve the integration with other resources at the NCBI. Priorities will depend on usage of the page and user feedback.

ACKNOWLEDGEMENTS

The authors acknowledge Rana Morris, Walter Ratzat, Anatoliy Kuznetsov, Mike DiCuccio and Liangshou Wu for supporting work and discussions on the projects described here. They also thank Eugene Yaschenko, Greg Schuler, Karl Sirotkin, Jim Ostell and David Lipman for helpful discussion and feedback. Many other people at the NCBI also contributed comments and tested the new report.

FUNDING

Intramural Research Program of the National Institutes of Health; National Library of Medicine. Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
3. Johnson,M., Zaretskaya,I., Raytselis,Y., Merezhuk,Y., McGinnis,S. and Madden,T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res*., **36**, W5–W9.
4. McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*., **32**, W20–W25.
5. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res*., **34**, W6–W9.
6. Morgulis,A., Coulouris,G., Raytselis,Y., Madden,T.L., Agarwala,R. and Schaffer,A.A. (2008) Database indexing for production MegaBLAST searches. *Bioinformatics*, **24**, 1757–1764.
7. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res*., **29**, 2994–3005.
8. Boratyn,G.M., Schaffer,A.A., Agarwala,R., Altschul,S.F., Lipman,D.J. and Madden,T.L. (2012) Domain enhanced lookup time accelerated BLAST. *Biol. Direct*., **7**, 12.
9. Zhang,J. and Madden,T.L. (1997) PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res*., **7**, 649–656.