

ASAP: a resource for annotating, curating, comparing, and disseminating genomic data

Jeremy D. Glasner*, Michael Rusch, Paul Liss, Guy Plunkett III, Eric L. Cabot, Aaron Darling, Bradley D. Anderson, Paul Infield-Harm, Michael C. Gilson and Nicole T. Perna

Genome Center of Wisconsin, University of Wisconsin, Madison, WI, USA

Received September 27, 2005; Revised and Accepted November 2, 2005

ABSTRACT

ASAP is a comprehensive web-based system for community genome annotation and analysis. ASAP is being used for a large-scale effort to augment and curate annotations for genomes of enterobacterial pathogens and for additional genome sequences. New tools, such as the genome alignment program Mauve, have been incorporated into ASAP in order to improve display and analysis of related genomes. Recent improvements to the database and challenges for future development of the system are discussed. ASAP is available on the web at <https://asap.ahabs.wisc.edu/asap/logon.php>.

INTRODUCTION

ASAP is a database that supports annotation and curation of genomic data by a distributed community of users through a web interface. The system permits users to upload genome sequence data, annotations and experimental data and it provides an environment for initial annotation of a genome or for updating, viewing and downloading existing annotations and experiments. The database currently provides data for both eukaryotes and bacteria, although the taxonomic focus is on members of the bacterial family Enterobacteriaceae. ASAP was originally created to facilitate the annotation and analysis of the *Erwinia chrysanthemi* genome by an international group of researchers (1). Since its debut, ASAP was used to complete the initial annotation of this genome and is now used for a number of ongoing genome projects and for maintenance of a larger number of complete genomes. A key feature of the system is a hierarchical curation procedure that provides rapid access to preliminary data that is subject to review by an expert curator. Extensive tracking of annotation data is provided. The system requires that annotators provide the evidence used to support their claims, which enables

downstream users to assess the quality of each piece of information.

EXPANDED DATABASE CONTENT

More taxa

The number of genome sequences contained in ASAP has expanded rapidly in recent years to include all published genomes from enterobacteria (currently 15), two mosquito expressed sequence tag (EST) projects (2), two Mycobacterium genome sequences and sequences from a metagenomics study (3) (Table 1). Additional datasets of high-throughput functional genomics data have also been added to the system. The current database contents demonstrate the ability of the system to accommodate diverse data, ranging from unfinished or partial genome sequences to completely annotated genomes with associated experimental data. The choice of genomes to be included in the database is driven by the user community. For several genome projects, particularly the phytopathogenic enterobacteria, ASAP is used as the primary system for genome annotation. For human pathogenic enterobacteria, ASAP is used as part of an NIH-funded project aimed at identifying targets for vaccines, diagnostics and therapeutics by integrating genomic information about these organisms. The EST sequencing project, metagenomics study and Mycobacterium genomes were entered in ASAP by request from user groups who desired to use the functionality available in ASAP for their specific annotation and analysis needs.

More options for access

ASAP was designed to provide access to genome annotation and analysis tools to distributed communities of participants in collaborative genome-scale sequencing and functional genomics experiments. Although the primary focus of the system is to make data available to a wide public audience, there are important practical reasons to restrict access to

*To whom correspondence should be addressed. Email: glasner@svm.vetmed.wisc.edu

Table 1. Genome data housed in the ASAP database

Sequence project	Project type	Taxonomic group	Status	Availability
<i>Aedes aegypti</i> , bacteria-inoculated hemocyte	EST	Eukaryote	Ongoing data collection	Public/Private
<i>Armigeres subalbatus</i> , bacteria-inoculated hemocyte	EST	Eukaryote	Ongoing data collection	Public/Private
<i>Buchnera aphidicola</i> (<i>Baizongia pistaciae</i>)	Genome	Enterobacteriaceae	Complete	Public
<i>B.aphidicola</i> strain Sg (<i>Schizaphis graminum</i>)	Genome	Enterobacteriaceae	Complete	Public
<i>Buchnera</i> sp. APS	Genome	Enterobacteriaceae	Complete	Public
<i>Candidatus Blochmannia floridanus</i>	Genome	Enterobacteriaceae	Complete	Public
Environmental BAC clone	Metagenome	Uncharacterized	Unfinished	Private
<i>Erwinia amylovora</i> strain Ea273	Genome	Enterobacteriaceae	Unfinished	Private
<i>Erwinia carotovora</i> subsp. atroseptica strain SCRI1043	Genome	Enterobacteriaceae	Complete	Public
<i>Erwinia chrysanthemi</i> strain 3937	Genome	Enterobacteriaceae	Complete	Public
<i>Escherichia coli</i> K-12 strain MG1655	Genome	Enterobacteriaceae	Complete, under curation	Public
<i>E.coli</i> O157:H7 strain EDL933	Genome	Enterobacteriaceae	Complete, ERIC curation	Public
<i>E.coli</i> O157:H7 strain RIMD 0509952	Genome	Enterobacteriaceae	Complete, ERIC curation	Public
<i>E.coli</i> strain CFT073	Genome	Enterobacteriaceae	Complete	Public
<i>E.coli</i> strain RS218	Genome	Enterobacteriaceae	Complete, undergoing annotation	Private
<i>Mycobacterium avium</i>	Genome	Actinobacteria	Complete	Private
<i>M.avium</i> subsp. paratuberculosis K-10	Genome	Actinobacteria	Complete	Private
<i>Pantoea stewartii</i> DC283	Genome	Enterobacteriaceae	Unfinished	Private
<i>Photorhabdus luminescens</i> subsp. laumondii TTO1	Genome	Enterobacteriaceae	Complete	Public
<i>Salmonella Choleraesuis</i> str. SC-B67	Genome	Enterobacteriaceae	Complete, ERIC curation	Public
<i>Salmonella Paratyphi A</i> str. ATCC 9150	Genome	Enterobacteriaceae	Complete, ERIC curation	Public
<i>Salmonella enterica</i> serovar Typhi plasmid R27	Plasmid	Enterobacteriaceae	Complete, ERIC curation	Public
<i>S.enterica</i> serovar Typhi strain CT18	Genome	Enterobacteriaceae	Complete, ERIC curation	Public
<i>S.enterica</i> subspecies enterica serovar Typhi Ty2	Genome	Enterobacteriaceae	Complete, ERIC curation	Public
<i>Salmonella typhimurium</i> LT2	Genome	Enterobacteriaceae	Complete, ERIC curation	Public
<i>S.flexneri</i> 2a strain 2457T	Genome	Enterobacteriaceae	Complete, ERIC curation	Public
<i>S.flexneri</i> 2a strain 301	Genome	Enterobacteriaceae	Complete, ERIC curation	Public
<i>S.flexneri</i> virulence plasmid pWR100	Plasmid	Enterobacteriaceae	Complete, ERIC curation	Public
<i>S.flexneri</i> virulence plasmid pWR501	Plasmid	Enterobacteriaceae	Complete, ERIC curation	Public
<i>Wigglesworthia brevipalpis</i>	Genome	Enterobacteriaceae	Complete	Public
<i>Yersinia pestis</i> KIM	Genome	Enterobacteriaceae	Complete, ERIC curation	Public
<i>Y.pestis</i> biovar Medievalis strain 91001	Genome	Enterobacteriaceae	Complete, ERIC curation	Public
<i>Y.pestis</i> strain CO92	Genome	Enterobacteriaceae	Complete, ERIC curation	Public
<i>Yersinia pseudotuberculosis</i> IP 32953	Genome	Enterobacteriaceae	Complete	Public

In the status column, projects referred to as 'complete' are finished, annotated genome projects open for community annotation, 'unfinished' projects are incomplete genome sequences undergoing completion and annotation, 'complete, ERIC curation' refers to complete genome projects that are being curated by members of the Enteropathogen Resource Integration Center (ERIC, <http://www.ericrc.org/>).

certain information in the database to a subset of users. As indicated in Table 1 the majority of projects in ASAP allow any guest user to access all data and allow anyone to sign up as a community annotator. Annotators, as opposed to guests, need a user name and password to log on to the system. This is simply so that we can record the identity of the annotator alongside their contributions, which provides a means of assigning credit and creates a mechanism for users or curators to request additional information about particular annotations. There are important reasons for restricting access to some projects. For example we have used ASAP in classrooms to teach genomic analysis to students. Students were provided with a copy of a genome project and they used ASAP to analyse and annotate gene functions. By restricting access to students in the class we can store the project in the same database as the public data without affecting the information available to a typical user. In other situations, such as with the *Mycobacterium* sequences, the data are used only by select users for specific purposes and are not made available to the general public (that has access to these sequences through other sources).

More diverse projects

A key reason that ASAP is able to support such diverse projects is the capacity of the database to handle multiple

sequences associated with a project, such as short sequences, assembled contigs, ESTs or complete genomes that may have multiple chromosomes and plasmids. We have expanded the query tools available to users to search for annotations across genomes and to restrict queries to specific subsets of annotations by nearly any characteristic stored in the database. ASAP queries produce lists of database objects (features) matching the search criteria. Each feature corresponds to a span of nucleotide coordinates in a given sequence within the genome.

The inclusion of all available genomes from enterobacteria in ASAP reflects our desire to create a resource for updating and comparing genomic data from this phylogenetically cohesive bacterial family. The enterobacteria are phenotypically diverse organisms that include a number of medically and agriculturally important pathogens. As described below, ASAP has a number of enhancements that facilitate comparative genomic analyses and standardization of annotations across genomes.

More ontologies

Uniformity in the type, quality and display of genome information makes comparisons across genes and organisms more useful. We have incorporated several ontologies that make standardization of annotations easier. For descriptions of gene functions and products we currently employ the Gene

Ontology (4) and Multifun (5) classification systems. Other ontologies can easily be imported into ASAP for use in annotation. To indicate the evidence used by annotators to describe genome features we have developed an ontology of terms that shows the nature of the information used to make the assessment. When published experimental or comparative analyses suggest annotation information, a hypertext link to the publication is provided. A distinction is made between evidence arising from analysis of the precise organism/strain in ASAP and data from closely related strains/species. The evidence codes in ASAP support evidence from a number of other databases and search tools, but do not attempt to subdivide experimental data into precise methods used for functional characterization. Detailed descriptions of methods and data associated with study of a genome feature can be entered as annotation notes or comments and their evidence would indicate that it came from experimental work with a link provided to the article. A more detailed ontology of evidence codes for experimental work such as http://obo.sourceforge.net/cgi-bin/detail.cgi?evidence_code could be incorporated into the existing codes to provide annotators with a richer choice of terms to describe experimental work.

Enhancements for comparative genomics

ASAP provides extensive support for analyses of related genome sequences. We maintain curated lists of relationships between the annotated proteins (orthology, paralogy, etc.). There is a semi-automatic implementation of the reciprocal best hits BLAST algorithm for identifying potential orthologs. PHP-scripts are used to download sequences, run the searches, parse the results, import the results, detect best hits and add ortholog candidates. Adding orthologs from the list of reciprocal best hits is done using an interface that provides statistics to a user and allows the user to add orthologs individually or by user-defined criteria, such as *E*-value and percent identity. Since proteins encoded by orthologous genes are likely to perform equivalent or related functions across species, there are interfaces that facilitate propagation of annotations between orthologous features.

Reciprocal best BLAST analyses are prone to predicting false orthologs and missing real orthologs. As expert curators review relationships between genomes they correct and augment the automated predictions of relationships between genes through an interface in ASAP. Often changing the assignment of homology between a pair of genomes necessitates that a curator reassess all of the relationships associated with the particular gene. We are developing additional algorithms and interfaces that will expedite the propagation of information across related sequences to reduce the time required to annotate orthologous features. However, even with the best possible automated solutions to these problems, the complexity of biological systems will still require ultimate review of annotations by expert curatorial staff.

Whole-genome multiple alignments

Annotation and visualization of genomes have been greatly enhanced by integrating Mauve, a whole-genome alignment and viewing system (6). The Mauve aligner is used to construct multiple alignments of relevant genomic sequences that are uploaded into the ASAP system. All pre-built multiple

alignments related to a particular genome are presented to a user viewing that genome. For example, a user viewing one of the *Escherichia coli* O157:H7 genomes is presented with options for viewing an alignment of four related *E.coli* genomes or an alignment of *E.coli* with other enterobacteria. When browsing annotations for a particular gene a user can choose to view the multiple alignment by launching a java applet that displays a visualization of the genome alignment centered on the selected gene (Figure 1). Users can zoom in and out, scroll through the alignment and browse genome annotations through the applet. Selecting an annotated feature will bounce the user to the ASAP annotation page with more detailed information about the entity. The tool is particularly useful for identifying conserved blocks of sequences between genomes and regions corresponding to genomic islands. Curators often use the tool when reviewing homology relationships between genes. Since the alignments generated by Mauve are based on nucleotide sequences they are also useful for analyzing regions other than just protein-coding gene sequences, such as conserved DNA binding sites and functional RNAs.

Handling experimental data

ASAP serves as a repository for experimental data associated with genomes in the database. There are examples of microarray hybridization data, high-throughput phenotypic data and results from IVET experiments (7) in ASAP currently, and the system is flexible enough to accommodate most other forms of experimental information associated with genome-scale functional characterization. Nearly any genomic data that can be represented in a tabular format can be imported into the system. To comply with standards such as MIAME for microarray experiments, users can attach nearly any sort of metadata associated with an experiment when uploading results. This can include information such as detailed protocols, array design files, analytical methodology, and links to additional resources.

There are two main ways in which ASAP users can interact with the experimental data. The most direct route is to select a genome and view a list of available experimental data sets. These data are organized into sets by the depositors, where a set is a collection of experiments related by a common theme, such as data from a single publication. After selecting a set, a user is presented with a table describing each of the experiments in the set along with the associated metadata. When experiments are selected, the user can choose from several options that specify the format of the data requested and can query for results for specific genes. The second route to the data is through the gene annotations. When there is experimental data in ASAP corresponding to a particular gene, users are shown a list of relevant experimental datasets on the gene annotation page. Selecting a set will return all of the data for that gene from the experiments in the selected set.

Comparison of ASAP to other software tools

*coli*BASE (8) is a web-based database for comparative genome analysis of *E.coli* and related species. It includes many of the same types of data as ASAP including genome alignments and predictions of orthologous genes and a number of

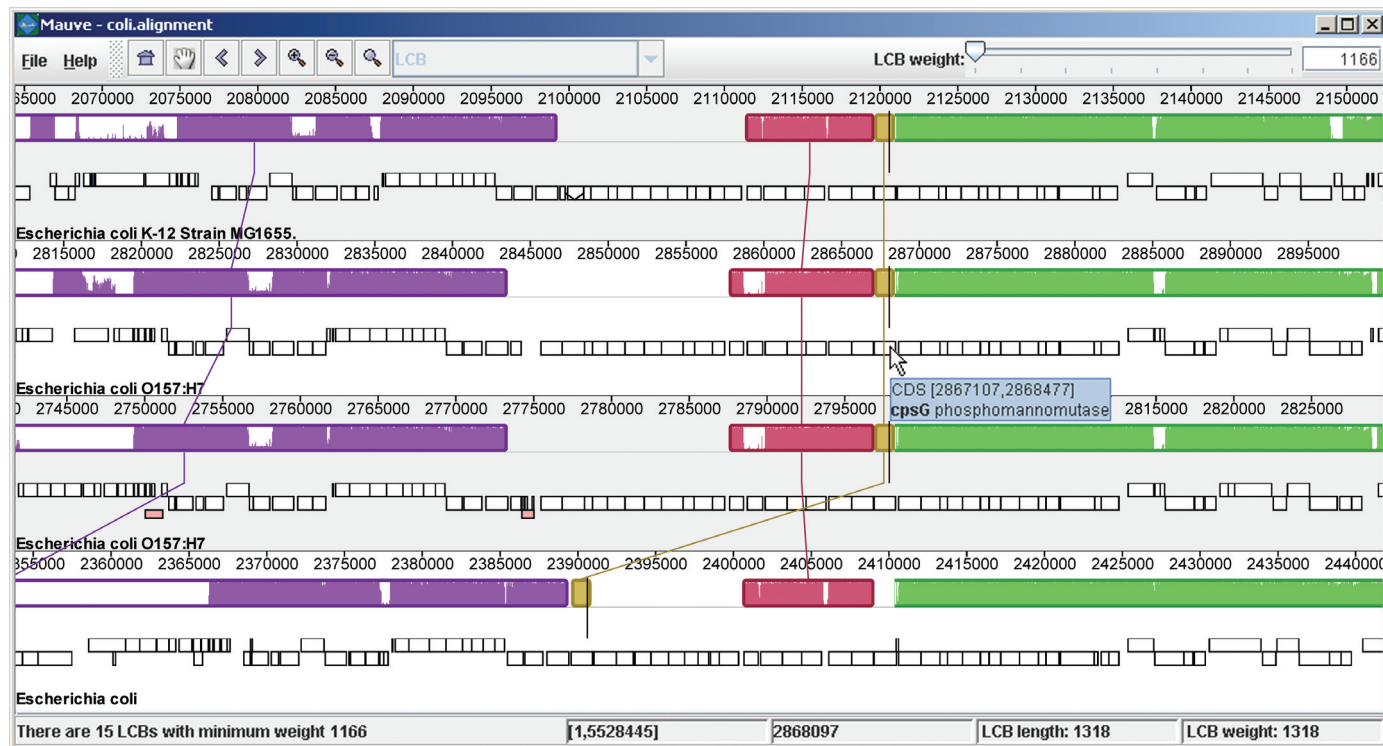


Figure 1. A portion of an alignment of four *E. coli* genomes visualized using the Mauve interface in ASAP. Large blocks with colored borders indicate rearrangement-free regions conserved across genomes. Within each block, a colored similarity plot indicates the level of sequence identity. Smaller white boxes with black outlines drawn below these blocks indicate annotated features (genes) in each genome.

additional automated predictions. Key features that distinguish the ASAP system include a focus on providing users with tools for contributing to the genome annotations, detailed descriptions of evidence used to infer gene functions and support for curators to evaluate and correct automated predictions and contributed information. *coliBase* is not packaged for redistribution although it is also used for databases of other microbes.

Artemis (9) and the Artemis Comparison Tool (ACT) (10) are stand-alone applications that are widely used for genome annotation, viewing and comparison. Artemis is particularly useful for initial annotation of a complete microbial genome sequence and, in fact, several of the genomes in the ASAP database were initially annotated using this tool. Artemis does not include a database and relies on flat files for input. ACT uses components of Artemis and provides an interactive visualization of comparisons of complete genomes based on BLAST (11) or MUMmer (12) searches to identify conserved segments. As a web-based application, ASAP provides an environment for multiple users to collaborate on genome annotation simultaneously, a feature lacking in the Artemis system. Additionally ASAP supports genome projects in progress, which may consist of many individual sequence files, and provides means to track features across multiple versions of a project, tasks that are not easy to accomplish using Artemis.

GenColors (13) is a new web-based system for annotation of prokaryotic genomes that considers data from related genomes and genome comparisons. It offers integration of data from ongoing sequencing projects and annotated genomic sequences obtained from GenBank. The database

does not require annotators to record the evidence for new input and does not offer any support for high-throughput experimental data associated with genome projects. Since ASAP will export genome annotations as Genbank flat files, it is simple to create files that are suitable for input into either Artemis or Gencolors.

Interfacing with the community of users

Since its initial release the ASAP user community has grown and their needs have diversified. ASAP is used for curatorial review of complete annotated bacterial genomes, primary annotation of incomplete sequences and annotation of several eukaryotic EST projects. The latest implementation of ASAP provides tools that allow users to customize the look and feel of the interface. A user logged onto the system is provided with a set of custom links to database contents on their front page, and new links to any database content can be added while browsing. The general look of the ASAP display (colors, logos, etc.) can be modified for individual genomes, since some user communities prefer that the style of the ASAP interface reflect their unique content or personal tastes.

Uploading data into ASAP is facilitated by a number of scripts. New genomes can be entered by importing GenBank flat files or FASTA files. Information about the genome, such as coordinates of predicted genome features, annotations, external links, coordinate updates, orthologs and experimental data can be uploaded using tab-delimited text files. The specific formats required for uploading data are documented on the individual upload pages.

The results of all queries to ASAP can be downloaded as tab-delimited text files. When querying genome annotations, users can opt to download a text file containing the protein or DNA sequences or annotations for the set of features returned by the query. Downloads are available for complete genome sequences and annotation in a number of formats including complete Genbank formatted files or spreadsheets formatted for Genbank submission.

We encourage the installation of ASAP at other sites and have made the source code available under a GNU public license.

Future challenges

An ongoing challenge in genomic sciences is the standardization of data formats across different database systems. To facilitate exchange of sequences and annotations between different systems we are planning to implement the use of GFF3 format (<http://song.sourceforge.net/gff3.shtml>) as an option for download. We plan to support additional formats for download of data from ASAP as they develop.

To allow users another avenue for querying of the database we plan to implement a BLAST server integrated into the ASAP system. Users could enter query sequences, do searches against the ASAP contents and receive formatted results that would link back to entries in ASAP. Additional resources that we would like to add to ASAP include pre-built alignments and phylogenetic trees for related sequences. Results from other software to predict protein function such as Interpro scans (14) for each protein and PSORTB subcellular predictions (15) will be added in the near future.

As the number of complete genome sequences grows, the prospect of annotating and curating the sequences by careful manual inspection by a dedicated expert scientist, the current state of the art, becomes daunting. This problem is currently mitigated by solutions that attempt to automate much of the annotation procedure. However, there are many complexities in genome analyses that cannot be readily automated. Where possible we will develop and incorporate computational solutions to these knottier problems. At the same time we will augment and improve the tools and interfaces used by scientists to manually interact with the data.

ACKNOWLEDGEMENTS

We would like to thank the Handelsman lab for using ASAP to test annotation of their metagenomics BAC sequences and the Talaat lab for using the system in their research. We would like to thank the Blattner and Christiansen labs for contributing information to the database. Funding to pay the Open

Access publication charges for this article was provided by The University of Wisconsin Graduate School.

Conflict of interest statement. None declared.

REFERENCES

1. Glasner, J.D., Liss, P., Plunkett, G. III, Darling, A., Prasad, T., Rusch, M., Byrnes, A., Gilson, M., Biehl, B., Blattner, F.R. *et al.* (2003) ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.*, **31**, 147–151.
2. Bartholomay, L.C., Cho, W.L., Rocheleau, T.A., Boyle, J.P., Beck, E.T., Fuchs, J.F., Liss, P., Rusch, M., Butler, K.M., Wu, R.C. *et al.* (2004) Description of the transcriptomes of immune response-activated hemocytes from the mosquito vectors *Aedes aegypti* and *Armigeres subalbatus*. *Infect. Immun.*, **72**, 4114–4126.
3. Schloss, P.D. and Handelsman, J. (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol.*, **6**, 229.
4. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
5. Serres, M.H. and Riley, M. (2000) MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb. Comp. Genomics*, **5**, 205–222.
6. Darling, A.C., Mau, B., Blattner, F.R. and Perna, N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, **14**, 1394–1403.
7. Yang, S., Perna, N.T., Cooksey, D.A., Okinaka, Y., Lindow, S.E., Ibekwe, A.M., Keen, N.T. and Yang, C.H. (2004) Genome-wide identification of plant-upregulated genes of *Erwinia chrysanthemi* 3937 using a GFP-based IVET leaf array. *Mol. Plant Microbe Interact.*, **17**, 999–1008.
8. Chaudhuri, R.R., Khan, A.M. and Pallen, M.J. (2004) coliBASE: an online database for *Escherichia coli*, *Shigella* and *Salmonella* comparative genomics. *Nucleic Acids Res.*, **32**, D296–D299.
9. Berriman, M. and Rutherford, K. (2003) Viewing and annotating sequence data with Artemis. *Brief Bioinform.*, **4**, 124–132.
10. Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G. and Parkhill, J. (2005) ACT: the Artemis Comparison Tool. *Bioinformatics*, **21**, 3422–3423.
11. 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.
12. Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.
13. Romualdi, A., Siddiqui, R., Glockner, G., Lehmann, R. and Suhnel, J. (2005) GenColors: accelerated comparative analysis and annotation of prokaryotic genomes at various stages of completeness. *Bioinformatics*, **21**, 3669–3671.
14. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
15. Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C.J., Ester, M. and Brinkman, F.S. (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, **21**, 617–623.