Published online 1 December 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D599–D604***
doi:10.1093/nar/gkr1105

BacMap: an up-to-date electronic atlas of annotated bacterial genomes

Joseph Cruz[1], Yifeng Liu[1], Yongjie Liang[1], You Zhou[2], Michael Wilson[1], Jonathan J. Dennis[2], Paul Stothard[3], Gary Van Domselaar[4] and David S. Wishart[1,2,*]

[1]Department of Computing Science, [2]Department of Biological Sciences, [3]Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB, Canada T6G 2E8 and [4]National Microbiology Laboratory—Public Health Agency of Canada, 1015 Arlington Street, Winnipeg, MB, Canada R3E 3R2

Received September 15, 2011; Revised October 29, 2011; Accepted November 5, 2011

ABSTRACT

Originally released in 2005, BacMap is an electronic, interactive atlas of fully sequenced bacterial genomes. It contains fully labeled, zoomable and searchable chromosome maps for essentially all sequenced prokaryotic (archaebacterial and eubacterial) species. Each map can be zoomed to the level of individual genes and each gene is hyperlinked to a richly annotated gene card. The latest release of BacMap (http://bacmap.wishartlab.com/) now contains data for more than 1700 bacterial species (~10× more than the 2005 release), corresponding to more than 2800 chromosome and plasmid maps. All bacterial genome maps are now supplemented with separate prophage genome maps as well as separate tRNA and rRNA maps. Each bacterial chromosome entry in BacMap also contains graphs and tables on a variety of gene and protein statistics. Likewise, every bacterial species entry contains a bacterial 'biography' card, with taxonomic details, phenotypic details, textual descriptions and images (when available). Improved data browsing and searching tools have also been added to allow more facile filtering, sorting and display of the chromosome maps and their contents.

INTRODUCTION

When the first bacterial genome was completed in 1995 it took more than a year of sequencing effort and cost nearly $2 million (1,2). Today it is possible to sequence, assemble and even annotate an entire bacterial genome in less than a day, at a cost of just a few hundred dollars (3). The ease with which bacterial genomes can be sequenced has led to an explosion of microbial sequences being assembled and deposited into various databases. Currently, GenBank (4) lists more than 7000 prokaryotic genomes with 1790 (as of 27 October 2011) fully completed bacterial and archaebacterial genomes and 5230 genomes marked as 'in progress' (with ~1/3 of these having draft sequences available). Never before has so much genome-scale information been available about so many different bacterial species. A growing challenge, therefore, is to find ways to better manage, display and compare this mountain of sequence data.

Over the past decade a number of excellent visualization tools have been developed for these purposes, such as CGView (5), BaSys (6) and DNAPlotter (7). These programs can create colorful, annotated, interactive circular genome maps that are ideal for bacterial genome maps. In addition, tools such as Circos (8) and Bluejay (9) have been developed to allow users to create colorful comparative genome maps. At the same time that these visualization tools were being developed, several superb whole-genome resources emerged that nicely integrated gene, genome, phenotypic and taxonomic information together. Some of these databases include the GenBank Genome Database (4), KEGG Genomes (10), PEDANT (11), Integr8 (12), Ensembl Genomes (13), TIGR's CMR (14), BioCyc (15), GOLD or the Genomes Online Database (16), PATRIC (17) and our own BacMap (18).

Originally released in 2005, BacMap was quite unique compared to most whole-genome databases as it was designed to serve more as an electronic atlas rather than a pure genome database. As an atlas, BacMap's primary role was to provide tools and resources to enable users to interactively select, display and manipulate bacterial genome maps. BacMap proved to be quite popular with many researchers in the microbiology community as it allowed facile, platform-independent viewing of both the structure and genomic content of many popular microbial genomes. Indeed, a number of the visualization tools

*To whom correspondence should be addressed. Tel: +780 492 0383; Fax: +780 492 1071; Email: david.wishart@ualberta.ca

Table 1. Comparison of bacterial and archebacterial genome resources

Database name	BacMap[a]	NCBI genome[b]	KEGG genomes[c]	Integr8[d]	JCVI CMR[e]	PEDANT[f]
No. of Bacterial genomes (as of 27 October 2011)	1671	1671	1370	2592 (incl. draft seqs)	672	811
No. of Archaea genomes (as of 27 October 2011)	119	119	116	106	48	57
Includes taxonomy	Yes	Yes	Yes	Yes	Yes	Yes (via NCBI)
Sequencing center/source	Yes	Yes	No	No	Yes	No
Includes references	Yes	Yes	Yes	Yes	Yes	Yes (via NCBI)
Genome statistics	Yes	Yes	Yes	Yes	Yes	No
Statistical charts	Yes	No	No	Yes	No	No
Bacterial descriptions	Yes	Some	No	Some	No	Some (via NCBI)
Genome map	Yes	No	No	No	Yes	No
tRNA/rRNA map	Yes	No	No	No	No	No
Prophage map	Yes	No	No	No	No	No
Zoomable maps	Yes	No	No	No	No	No
Sortable views	Yes	Yes	No	No	Yes	No
Phenotype filter	Yes	No	No	No	No	No
Data fields per Gene/Prot	63	7	10	10	16	16
BLAST query	Yes	Yes	No	Yes	Yes	Yes
Text search	Yes	No	Partial	Yes	Yes	Yes
Precomputed alignments	No	Yes	No	Yes	Yes	No
Analytical tools	No	Yes	No	No	Yes	No
Pathway information	Yes	No	Yes	No	Yes	No

[a]http://bacmap.wishartlab.com.
[b]http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi.
[c]http://www.genome.jp/kegg/catalog/org_list.html.
[d]http://www.ebi.ac.uk/integr8.
[e]http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi.
[f]http://pedant.gsf.de/.

developed for BacMap have become widely used in the microbial genomics community (5,6). However, as the number of sequenced genomes grew and as our access to computer resources waned, it became increasingly difficult to keep all components of BacMap current. Fortunately, additional computer resources have recently become available and this has allowed us to substantially update and upgrade BacMap over the past year.

Here, we describe the major improvements and changes made to BacMap, including the expansion of the database (by 10× over the 2005 release), the addition of new genome visualization tools (for displaying prophage and tRNA/rRNA genes), the construction of thousands of new bacterial 'biography' pages and the redesign of the website to improve the ability of users to query, sort or select genes, genomes, pathway, taxonomic and/or phenotypic information from the database. With these new enhancements along with our improved ability to semi-automatically maintain and update this resource, we believe BacMap has now become one of the most complete, current and comprehensive bacterial genome resources available (see Table 1 for a detailed comparison between BacMap with other commonly used microbial genome resources). BacMap is available at http://bacmap.wishartlab.com.

WHAT'S NEW IN BACMAP?

Details relating to BacMap's overall architecture, layout, general querying capabilities, and annotation protocols have been described previously (18) and will not be reviewed here. Instead, we shall focus primarily on describing the changes and enhancements made to BacMap since the last release. More specifically, we will describe: (i) the growth and enhancements made to BacMap's existing content; (ii) changes to the BacMap interface and layout; and (iii) improvements to BacMap's data querying and filtering capabilities.

CONTENT GROWTH AND ENHANCEMENT

The first release of BacMap contained fully annotated gene/protein maps from just 177 bacterial species (18). The latest release of the BacMap database (as of 27 October 2011) contains pre-calculated genome from 1790 completed eubacterial and archaebacterial species or strains, consisting of more than 2880 chromosomes and plasmids (or replicons). Overall, this represents a ~10-fold increase in the number of bacterial species in the database. In the previous version of BacMap, only one type of genome map (a gene/protein map) was available for each species, which translated to about 300 different chromosome or plasmid maps. Now each bacterial species in BacMap is associated with three different kinds of genome maps: (i) a gene/protein map; (ii) a prophage map and (iii) a tRNA/rRNA map. Consequently BacMap now contains more than 5300 pre-calculated bacterial chromosome (>400 kb) maps and more than 3300 pre-calculated bacterial plasmid (<400 kb) maps. The decision to include both prophage and tRNA/rRNA maps in the latest release of BacMap was motivated by a number of user requests. It was also based on several emerging trends in microbial genomics where information about prophage 'species' and 16S

rRNA sequences is being routinely used to help understand bacterial evolution, phylogeny and gene transfer. Certainly, the identification and mapping of prophage sequences (which can occupy up to 20% of some bacterial genomes) is an oft-ignored component to many bacterial annotation efforts.

As with previous versions of BacMap the gene/protein maps generated via CGView (5) and annotated using BASys (6). The new tRNA/rRNA maps were generated using existing genome annotations and supplemented with information from tRNAscan (19) while BacMap's prophage maps were generated using PHAST (20). Both the tRNA/rRNA maps and prophage maps are displayed using a different and somewhat more sophisticated Google Map style of graphics. Both of these new BacMap display tools, which require Adobe Flash, support both circular and linear genomic views as well as interactive browsing and dynamic image labeling. Additional details about the display capabilities, methodology and the accuracy of PHAST's prophage predictions are available in the PHAST manuscript or on the PHAST website (20).

In addition to the significant growth in the number of genomes (10×) and map types (3×), there has also been significant growth in the number (from 80 to 1790), depth (5 data fields to 38 data fields) and proportion (from 50% to 100%) of bacterial species with bacterial 'biographies'. These biography cards contain information on the bacterium's name(s), accession numbers, taxonomy, subspecies/strain, date of genome release, sequencing center, completeness, sequencing center, sequencing quality, sequencing depth, sequencing method, isolation site/country, number of replicons, chromosome shape, plasmid shape, gram stain, shape, motility, flagellar presence, number of membranes, oxygen requirements, optimal temperature, temperature range, habitat, biotic relationship, host name, cell arrangement, sporulation properties, metabolism, energy source, associated diseases (if any) and pathogenicity. A brief textual description of the organism covering its physiology, general characteristics, ecological niche, source, relevance to human or animal disease and related references is also given. Additionally an image of the organism (if available) is provided. This information was mined from Integr8 (12), the NCBI BioProjects (4), HAMAP (21), Wikipedia, Microbewiki, Karyn's Genomes, various bacterial genome home pages, Google Images as well as other sources and manually edited. Each BacMap biography page or 'BioCard' has two other tabs that also contains a list of metabolic pathways that occur or are though to occur in that organism and a list of references or database hyperlinks.

Each gene in BacMap is linked to a gene card that provides detailed information about that gene/protein. The original release of BacMap provided just 11 data fields for each gene or protein. The latest version now has and average of 63 data fields, covering a wide range of information on gene features, protein features, protein functions, subcellular locations and other relevant data. The rich annotation for the completed genomes in BacMap was primarily derived or calculated from BASys (6). In addition to these BASys annotations, COG and PEDANT functional classifications (where available) have been extracted from their respective online databases (4,11). Overall, the amount or 'depth' of gene/protein specific data in BacMap has grown by more than a factor of 5. Given that there are ~5 million genes in the new release of BacMap (compared to approximately 500 000 genes in the original release), this represents a nearly 50× increase in the quantity of sequence annotation data.

INTERFACE CHANGES

The growth in BacMap's content and size has necessitated a number of changes in its interface. The database is still easily browseable, but to increase the number of entries per page a more compact, tabular display has been adopted. As seen in Figure 1, each row in the BacMap genome table has nine columns covering: (i) the organism name/species; (ii) replicon type (chromosome/plasmid); (iii) release date; (iv) number of replicons; (v) GenBank identifier; (vi) replicon length; (vii) GC content; (viii) Maps; and (ix) Tools. Under the Maps column users may click on either the Gene/Protein map (a circle icon), the tRNA/rRNA map (a tRNA icon) or the Prophage map (a T4 phage icon) to generate the desired interactive genome map. Under the Tools column users may select the genome statistics link (a bar-graph icon), the genome-specific BLAST search (a magnifier icon with a chromosome), the genome-specific text search (a magnifier icon with a T) or the download link (a green arrow). Clicking on the organism name will launch BacMap's 'bacterial biography' card or BioCard, yielding detailed taxonomic and phenotypic information about the organism of interest. BacMap's genome table may be navigated by clicking on page numbers or Previous/Next arrows marked above the table. Additionally, by clicking on specific columns, users may be able to sort the display according to organism name (alphabetically—which is the default display), replicon type, date of release, number of replicons, identifier number, replicon/genome length and GC content.

Clicking on the map links, stats links, BLAST and text searches will open up new windows which will produce genome images or fillable text boxes that are very similar to those seen in the original version of BacMap. The BLAST and text search tools associated with each chromosome or plasmid are specific to that plasmid or that chromosome. In other words, the searches are limited to the data in that replicon's BacMap genome cards. The same zooming and map navigation tools are still used for the gene/protein maps while the tRNA/rRNA and prophage maps are navigated the same way as those described for PHAST. The graphs and charts formats for the BacMap stats links are essentially unchanged from the previous version of BacMap.

SELECTION AND QUERYING ENHANCEMENTS

The large number and diversity of bacterial genomes in BacMap has also necessitated other changes to the interface and to the querying tools. BacMap now has

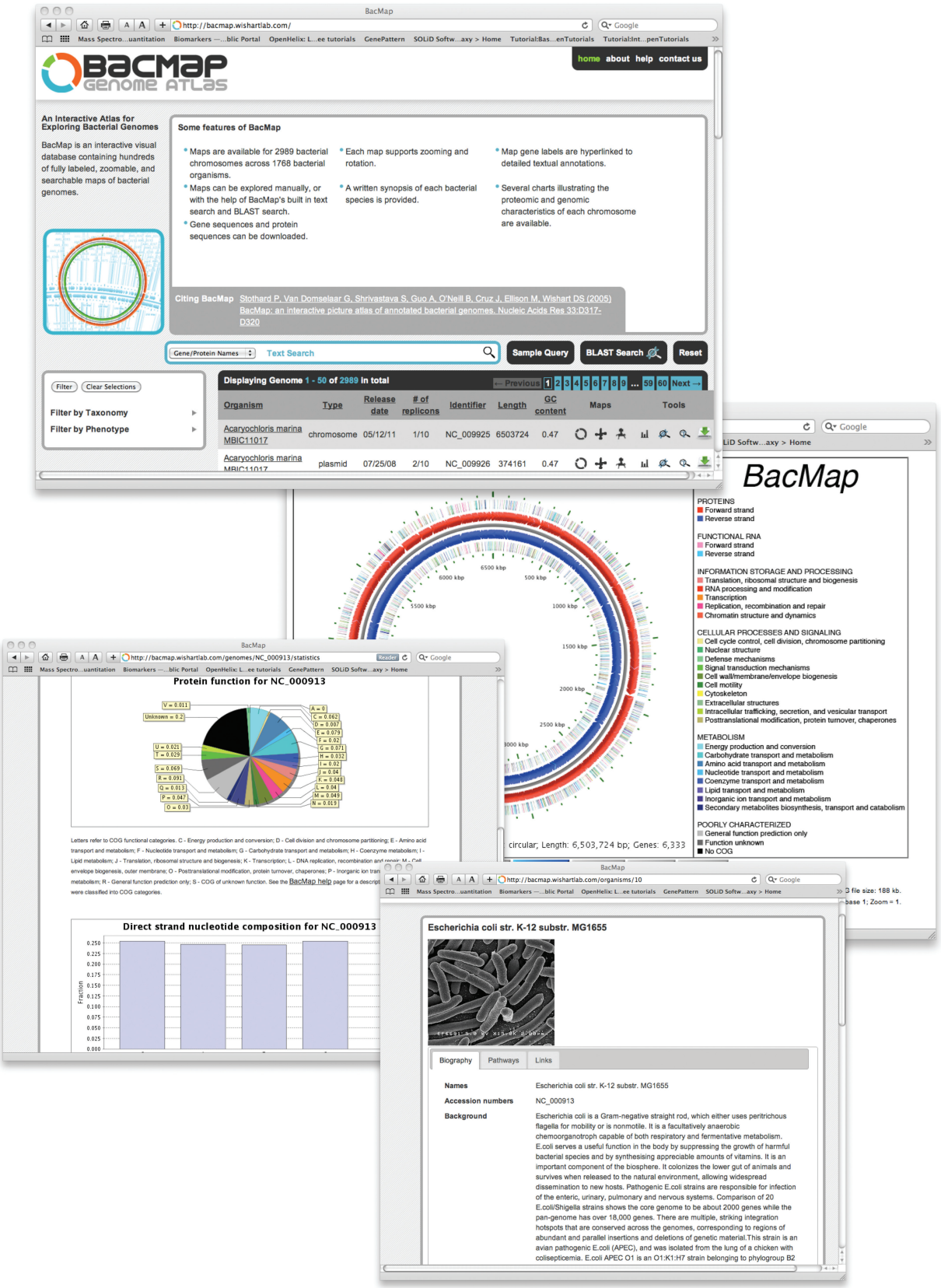

Figure 1. A screenshot montage of the BacMap database showing the different display, browsing and filtering tools.

a sophisticated data filtering system located on the left side of its genome table (Figure 1). Using this tool, users may filter or select genomes based on their taxonomy and/or phenotype. BacMap's taxonomy and phenotype filters may be toggled to be displayed or hidden by clicking on their respective headings (the default is to display all filter options). To enable the selection of different taxonomic groupings, users may select any combination of 78 yes/no/NA check boxes under 2 different kingdoms (Bacteria or Archaea) or 24 different phyla/classes. Choosing one (say Archaea) will reformat the standard BacMap browsing table and display only the known archaebacteria in the table. This table may then be sorted, as described earlier, by clicking on the appropriate column headings. To enable the selection of different phenotypic groupings in BacMap, users may select any combination of yes/no/NA check boxes under 16 different phenotypic headings including: (i) Flagella; (ii) Human pathogen; (iii) Motility; (iv) Number of membranes; (v) Number of chromosomes; (vi) Chromosome shape; (vii) Plasmid shape; (viii) Cell shape; (ix) Cell arrangement; (x) Gram stain; (xi) Temperature range; (xii) Oxygen requirements; (xiii) Biotic relationship; (xiv) Habitat; (xv) Energy source; and (xvi) Sporulation. The default is to leave all check boxes cleared, which allows the full BacMap genome table to be viewed. If one or more check boxes is selected, the BacMap genome table will be reformatted to display only those organisms with the selected phenotype. Once a phenotype has been selected, the BacMap browsing table will be reformatted and will display only those organisms with the selected phenotype. As before, the resulting table may be sorted by clicking the appropriate column headings. Using these sorting and filtering tools, it is now relatively easy for a user to perform a query such as: 'Find all Archaebacteria that are hyperthermophiles with >55% GC content' or 'Find all Proteobacteria that are gram positive and that have genome sizes greater than 6 megabases'.

BacMap now supports both genome-specific BLAST (22) queries, filtered database BLAST queries and whole database BLAST queries. The genome-specific BLAST queries are accessed by clicking on the BLAST hyperlink for a specific organism or genome in BacMap's genome table (the last column). The whole BacMap database or filtered database BLAST queries can be accessed by clicking on the 'BacMap BLAST' link, located at the top right of the BacMap genome table. This will produce a fillable text box for standard protein or gene sequence queries. If the database has not been filtered (via taxonomy or phenotype) prior to clicking the BacMap BLAST link, then the entire database will be searched. If the database has been filtered, then the BLAST search will be limited to only those genomes displayed in the BacMap genome table. The same model also works for BacMap's text search tool (located adjacent to the BacMap BLAST link), where users may perform genome-specific text queries, filtered database text queries or whole database text queries.

To help with the speed and specificity of the text searches, there is now an added filter in the newest release of BacMap. In particular, users can now choose to search selected components of the database covering only the: (i) Genus/Species/Strain names; (ii) Gene/Protein Names; (iii) text in the bacterial biography cards ('Biocards'); or (iv) text in the metabolic pathway descriptors ('Metabolic Pathway'). This selection is done by using the pull-down menu located in the text search box. Using these text searching tools, users may first filter the database to select 'all Archaea that are hyperthermophiles' and then from this subset, search for the term 'methanogen' from the biography card. The result would be a list of sequenced Archaea that are hyperthermophilic methanogens. This flexibility in searching and filtering should make BacMap particular useful for a wide range of microbiologists and metagenomics specialists.

CONCLUSION

To summarize, BacMap is a richly annotated, easily queried and highly interactive electronic atlas containing data from more than 1700 fully sequenced bacterial genomes. The latest release of BacMap builds on earlier strengths but also adds tremendously to the number of annotated genomes, the number and quality of visual displays, the amount of phenotypic or organism-specific data and the ability for users to display, sort, search and filter the data. As shown in Table 1, these changes and enhancements have made the latest release of BacMap quite comparable to a number of more widely known or widely used bacterial genome resources. Furthermore, with its unique focus on using easily understood, rapidly accessible, interactive visual displays to transmit genome-scale information, we anticipate that BacMap may have particularly broad appeal among students, educators and scientists. With a growing interest in bacterial genomics and the growing ease with which bacterial genomes can be sequenced, we believe that an up-to-date resource such as BacMap is a timely addition and a useful contribution to the field.

FUNDING

Genome Alberta (a division of Genome Canada); Canadian Institutes of Health Research (CIHR). Funding for open access charge: CIHR.

Conflict of interest statement. None declared.

REFERENCES

1. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496–512.
2. Fraser,C.M., Eisen,J.A., Nelson,K.E., Paulsen,I.T. and Salzberg,S.L. (2002) The value of complete microbial genome sequencing (you get what you pay for). *J. Bacteriol.*, **184**, 6403–6405.
3. Nagarajan,N. and Pop,M. (2010) Sequencing and genome assembly using next-generation technologies. *Methods Mol. Biol.*, **673**, 1–17.
4. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2011) GenBank. *Nucleic Acids Res.*, **39**, D32–D37.

5. Stothard,P. and Wishart,D.S. (2005) Circular genome visualization and exploration using CGView. *Bioinformatics*, **21**, 537–539.
6. Van Domselaar,G.H., Stothard,P., Shrivastava,S., Cruz,J.A., Guo,A., Dong,X., Lu,P., Szafron,D., Greiner,R. and Wishart,D.S. (2005) BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res.*, **33**, W455–W459.
7. Carver,T., Thomson,N., Bleasby,A., Berriman,M. and Parkhill,J. (2009) DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics*, **25**, 119–120.
8. Krzywinski,M., Schein,J., Birol,I., Connors,J., Gascoyne,R., Horsman,D., Jones,S.J. and Marra,M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.
9. Soh,J., Gordon,P.M., Taschuk,M.L., Dong,A., Ah-Seng,A.C., Turinsky,A.L. and Sensen,C.W. (2008) Bluejay 1.0: genome browsing and comparison with rich customization provision and dynamic resource linking. *BMC Bioinformatics*, **9**, 450.
10. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
11. Walter,M.C., Rattei,T., Arnold,R., Güldener,U., Münsterkötter,M., Nenova,K., Kastenmüller,G., Tischler,P., Wölling,A., Volz,A. *et al.* (2009) PEDANT covers all complete RefSeq genomes. *Nucleic Acids Res.*, **37**, D408–D411.
12. Sterk,P., Kulikova,T., Kersey,P. and Apweiler,R. (2007) The EMBL nucleotide sequence and genome reviews databases. *Methods Mol. Biol.*, **406**, 1–21.
13. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
14. Davidsen,T., Beck,E., Ganapathy,A., Montgomery,R., Zafar,N., Yang,Q., Madupu,R., Goetz,P., Galinsky,K., White,O. *et al.* (2008) The comprehensive microbial resource. *Nucleic Acids Res.*, **38**, D340–D345.
15. Caspi,R., Altman,T., Dale,J.M., Dreher,K., Fulcher,C.A., Gilham,F., Kaipa,P., Karthikeyan,A.S., Kothari,A., Krummenacker,M. *et al.* (2008) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **38**, D473–D479.
16. Liolios,K., Chen,I.M., Mavromatis,K., Tavernarakis,N., Hugenholtz,P., Markowitz,V.M. and Kyrpides,N.C. (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **38**, D346–D354.
17. Gillespie,J.J., Wattam,A.R., Cammer,S.A., Gabbard,J.L., Shukla,M.P., Dalay,O., Driscoll,T., Hix,D., Mane,S.P., Mao,C. *et al.* (2011) PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect. Immun.*, **79**, 4286–4298.
18. Stothard,P., Van Domselaar,G., Shrivastava,S., Guo,A., O'Neill,B., Cruz,J., Ellison,M. and Wishart,D.S. (2005) BacMap: an interactive picture atlas of annotated bacterial genomes. *Nucleic Acids Res.*, **33**, D317–D320.
19. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
20. Zhou,Y., Liang,Y., Lynch,K.H., Dennis,J.J. and Wishart,D.S. (2011) PHAST: a fast phage search tool. *Nucleic Acids Res.*, **39**, W347–W352.
21. Lima,T., Auchincloss,A.H., Coudert,E., Keller,G., Michoud,K., Rivoire,C., Bulliard,V., de Castro,E., Lachaize,C., Baratin,D. *et al.* (2009) HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res.*, **37**, D471–D478.
22. 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.