Published online 25 October 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D391–D395***
doi:10.1093/nar/gkp918

Megx.net: integrated database resource for marine ecological genomics

Renzo Kottmann[1,*], Ivalyo Kostadinov[1,2], Melissa Beth Duhaime[1,2], Pier Luigi Buttigieg[1,2], Pelin Yilmaz[1,2], Wolfgang Hankeln[1,2], Jost Waldmann[1] and Frank Oliver Glöckner[1,2]

[1]Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen and [2]Jacobs University Bremen gGmbH, D-28759 Bremen, Germany

Received September 15, 2009; Accepted October 8, 2009

ABSTRACT

Megx.net is a database and portal that provides integrated access to georeferenced marker genes, environment data and marine genome and metagenome projects for microbial ecological genomics. All data are stored in the Microbial Ecological Genomics DataBase (MegDB), which is subdivided to hold both sequence and habitat data and global environmental data layers. The extended system provides access to several hundreds of genomes and metagenomes from prokaryotes and phages, as well as over a million small and large subunit ribosomal RNA sequences. With the refined Genes Mapserver, all data can be interactively visualized on a world map and statistics describing environmental parameters can be calculated. Sequence entries have been curated to comply with the proposed minimal standards for genomes and metagenomes (MIGS/MIMS) of the Genomic Standards Consortium. Access to data is facilitated by Web Services. The updated megx.net portal offers microbial ecologists greatly enhanced database content, and new features and tools for data analysis, all of which are freely accessible from our webpage http://www.megx.net.

INTRODUCTION

Over the last years, molecular biology has undergone a paradigm shift, moving from a single experiment science to a high-throughput endeavour. Although the genomic revolution is rooted in medicine and biotechnology, it is currently the environmental sector, specifically the marine, which delivers the greatest quantity of data. Marine ecosystems, covering >70% of the Earth's surface, host the majority of biomass and significantly contribute to global organic matter and energy cycling. Microorganisms are known to be the 'gatekeepers' of these processes and insights into their lifestyle and fitness will enhance our ability to monitor, model and predict future changes.

Recent developments in sequencing technology have made routine sequencing of whole microbial communities from natural environments possible. Prominent examples in the marine field are the ongoing Global Ocean Sampling (GOS) campaign (1,2) and Gordon and Betty Moore Foundation Marine Microbial Genome Sequencing Project (http://www.moore.org/microgenome/). Notably, the GOS resulted in a major input of new sequence data with unprecedented functional diversity (3). The resulting flood of sequence data available in public databases is an extraordinary resource with which to explore microbial diversity and metabolic functions at the molecular level.

These large-scale sequencing projects bring new challenges to data management and software tools for assembly, gene prediction and annotation—fundamental steps in genomic analysis. Several new dedicated database resources have recently emerged to tackle the current need for large-scale metagenomic data management, namely CAMERA (4), IMG/M (5) and MG-RAST (6).

Nevertheless, it is increasingly apparent that the full potential of comparative genome and metagenome analysis can be achieved only if the geographic and environmental context of the sequence data is considered (7,8). The metadata describing a sample's geographic location and habitat, the details of its processing, from the time of sampling to sequencing and subsequent analyses are important, e.g. modelling species' responses to environmental change or the spread and niche adaptation of bacteria and viruses. This suite of metadata is collectively referred as contextual data (9).

Megx.net is the first database to integrate curated contextual data with their respective genes, genomes and metagenomes in the marine environment (10). Now, the

*To whom correspondence should be addressed. Tel: +49 421 2028974; Fax: +49 421 2028580; Email: rkottman@mpi-bremen.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

extended megx.net database resource allows post factum retrieval of interpolated environmental parameters, such as temperature, nitrate, phosphate, etc. for any location in the ocean waters based on profile and remote sensing data. Furthermore, the content has been significantly updated to include prokaryote and marine phage genomes, metagenomes from the GOS project (2) and all georeferenced small and large subunit ribosomal RNA (rRNA) sequences from the SILVA database project (11).

The extended megx.net portal is the first resource of its kind to offer access to this unique combination of data, including manually curated habitat descriptors for genomes, metagenomes and marker genes, their respective contextual data and additionally integrated environmental data. See the megx.net online video tutorial for a guided introduction and overview at http://www.megx.net/portal/tutorial.html (Supplementary Data).

NEW DATABASE STRUCTURE AND CONTENT

The Microbial Ecological Genomics DataBase (MegDB), the backbone of megx.net, is a centralized database based on the PostgreSQL database management system. The georeferenced data concerning geographic coordinates and time are managed with the PostGIS extension to PostgreSQL. PostGIS implements the 'Simple Features Specification for SQL' standard recommended by the Open Geospatial Consortium (OGC; http://www.opengeospatial.org/), and therefore offers hundreds of geospatial manipulation functions.

MegDB is comprised of (i) MetaStorage, which stores georeferenced DNA sequence data from a collection of genomes, metagenomes and genes of molecular environmental surveys, with their contextual data, and (ii) OceaniaDB, which stores georeferenced quantitative environmental data (Figure 1).

Contextual and sequence data content

Sequences in MetaStorage are retrieved from the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org/). However, as of September 2009, GOLD reported 5776 genome projects, of which, only 1095 were finished and published (http://www.genomesonline.org/gold.cgi). As most of the sequenced functional diversity is contained in these draft and shotgun datasets, megx.net was extended to host draft genomes and whole genome shotgun data. Currently, MegDB contains 1832 prokaryote genomes (940 incomplete or draft) and 80 marine shotgun metagenomes from the GOS microbial dataset. Marine viruses are a missing link in the correlation of microbial sequence data with contextual information to elucidate diversity and function. Consequently, megx.net now incorporates all sequenced marine phage genomes in MegDB, the first step towards a community call for integration of viral genomic and biogeochemical data (12).

In an effort towards integrating microbial diversity with specific sampling sites, megx.net has been extended to include georeferenced small and large subunit rRNA sequences from the SILVA rRNA databases project

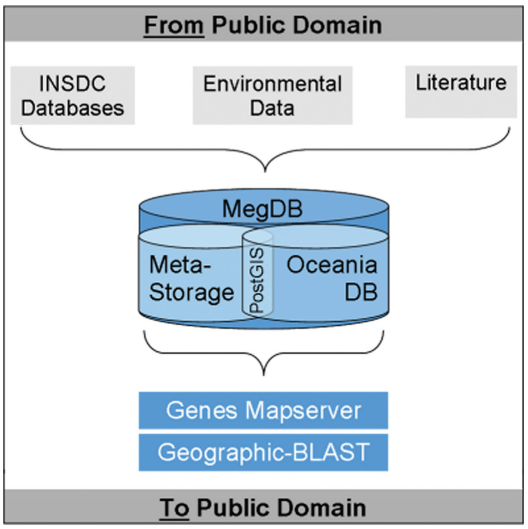

Figure 1. General architecture of megx.net: DNA sequence data (from INSDC) is integrated with contextual data from diverse resources (i.e. manual literature mining and the GOLD database) and interpolated environmental data. MegDB integrates the data conforming to OGC standards and MIGS/MIMS specification. The core megx.net tools, Genes Mapserver and Geographic-BLAST access the MegDB content.

(11). Currently, only 9% (16S/18S) and 2% (23S/28S) of over 1 million sequences in SILVA SSUParc (16S/18S) and LSUParc (23S/28S) databases are georeferenced. With the implementation of the Minimal Information about an Environmental Sequence (MIENS) standard for marker gene sequences (http://gensc.org/gc_wiki/index.php/MIENS), efforts are ongoing to significantly improve this situation.

All genomic sequences in megx.net are supplemented by contextual data from GOLD (13) and NCBI Genome Projects (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html). The database is designed to store all contextual data recommended by the Genomics Standards Consortium, and is thus compliant with the Minimum Information about a Genome Sequence (MIGS) standard and its extension, Minimum Information about a Metagenome Sequence (MIMS) (7,9).

Furthermore, megx.net is the first resource to provide a manually annotated collection of genomes using terms from EnvO-Lite (Rev. 1.4), a subset of the Environment Ontology (EnvO) (14). An EnvO-Lite term was assigned to each genome project, identifying the environment where its original sample material was obtained. The annotation can be browsed on the megx.net portal using, e.g. tag clouds, and may be used as a categorical variable in comparative analyses.

Environmental data content

OceaniaDB was added to MegDB to supplement the georeferenced molecular data of MetaStorage with interpolated environmental parameters. When sufficient date, depth and location measurements are provided, any 'on site' contextual data taken at a sampling site can

be supplemented by environmental data describing physical, chemical, geological and biological parameters, such as ocean water temperature and salinity, nutrient concentrations, organic matter and chlorophyll.

The environmental data is retrieved from three sources:

(1) World Ocean Atlas: a set of objectively analysed (one decimal degree spatial resolution) climatological fields of *in situ* measurements (http://www.nodc.noaa.gov/OC5/WOA05/pr_woa05.html);
(2) World Ocean Database: a collection of scientific, quality-controlled ocean profiles (http://www.nodc.noaa.gov/OC5/WOD05/pr_wod05.html); and
(3) SeaWIFS chlorophyll *a* data (http://seawifs.gsfc.nasa.gov).

These data are described at 33 standard depths for annual, seasonal and monthly intervals. Together, the location and time data (*x, y, z* and *t*) serve as a universal anchor, and link environmental data to the sequence and contextual data in MetaStorage (Figure 1). As such, megx.net integrates biologist-supplied sequence and contextual data (measured at the time of sampling) with oceanographic data provided by third-party databases. All environmental data are compatible with OGC standards (http://www.opengeospatial.org/standards) and are described with exhaustive meta-information consistent with the ISO 19115 standard.

Moreover, based on the integrated environmental data, megx.net provides information to aid biologists in grasping the ocean stability, on both global and local scales. For all environmental parameters, the yearly standard deviations of the monthly values can be viewed on a world map, for easy visualization of high and low variation sample sites. Furthermore, for each sample site, users can view trends in numerous parameters.

USER ACCESS

Genes Mapserver

The Genes Mapserver (formerly Metagenomes Mapserver) offers a sample-centric view of the georeferenced MetaStorage content. Substantial improvements to the underlying Geographic Information System (GIS) and web view have been made. The website is now interactive, offering user-friendly navigation and an overlay of the OceaniaDB environmental data layers to display sampling sites on a world map in their environmental context. Sample site details and interpolated data can be retrieved by clicking the sampling points on the map (Figure 2).

The GIS Tools of the Genes Mapserver allow extraction of interpolated values for several physicochemical and biological parameters, such as temperature, dissolved oxygen, nitrate and chlorophyll concentrations, over specified monthly, seasonally or annually intervals (Figure 2f).

Geographic-BLAST

The Geographic-BLAST tool queries the MegDB genome, metagenome, marine phages and rRNA sequence data using the BLAST algorithm (15). The results are reported according to the sample locations (when provided) of the database hits. With the updated Geographic-BLAST, results are plotted on the Genes Mapserver world map, where they are labeled by number of hits per site (Figure 2). Standard BLAST results are shown in a table, which also provides direct access to the associated contextual data of the hits.

Software extensions to the portal

In addition to the services directly provided by megx.net, the project serves as a portal to software for general data analysis in microbial genomics.

MetaBar (http://www.megx.net/metabar) is a tool developed with the aim to help investigators efficiently capture, store and submit contextual data gathered in the field. It is designed to support the complete workflow from the sampling event up to the metadata-enriched sequence submission to an INSDC database.

MicHanThi (http://www.megx.net/michanthi) is a software tool designed to facilitate the genome annotation process through rapid, high-quality prediction of gene functions. It clearly out-performs the human annotator in terms of accuracy and reproducibility.

JCoast [http://www.megx.net/jcoast; (16)] is a desktop application primarily designed to analyze and compare (meta)genome sequences of prokaryotes. JCoast offers a flexible graphical user interface, as well as an application programming interface that facilitates back-end data access to GenDB projects (17). JCoast offers individual, cross genome and metagenome analysis, including access to Geographic-BLAST.

User test case

To demonstrate the interpretation of genomic content in environmental context, consider a test case with the marine phages. Marine phage genomes (18) and 'viral' classified GOS scaffolds (19) have revealed host-related metabolic genes involved in, i.e. photosynthesis, phosphate stress, antibiotic resistance, nitrogen fixation and vitamin biosynthesis. Geographic-BLAST can be used to investigate the presence of PhoH (accession YP_214558), a phosphate stress response gene, among the sequenced marine phages. The search results can then be interpreted in their environmental context, either as (i) average annual phosphate measurements, or (ii) stability of phosphate concentrations in terms of monthly SD (Figure 2c and d). A closer look at a single genome sample site reveals that *in situ* temperature was not originally reported (Figure 2e), whereas the interpolated data supplements this parameter, among others (Figure 2f).

Web Services

The newly extended version of megx.net offers programmatic access to MegDB content via Web Services, a powerful feature for experienced users and developers. All geographical maps can be retrieved via simple web requests, as specified by the Web Map Service (WMS) standard. The base URL for WMS requests is http://www.megx.net/wms/gms, where more detailed

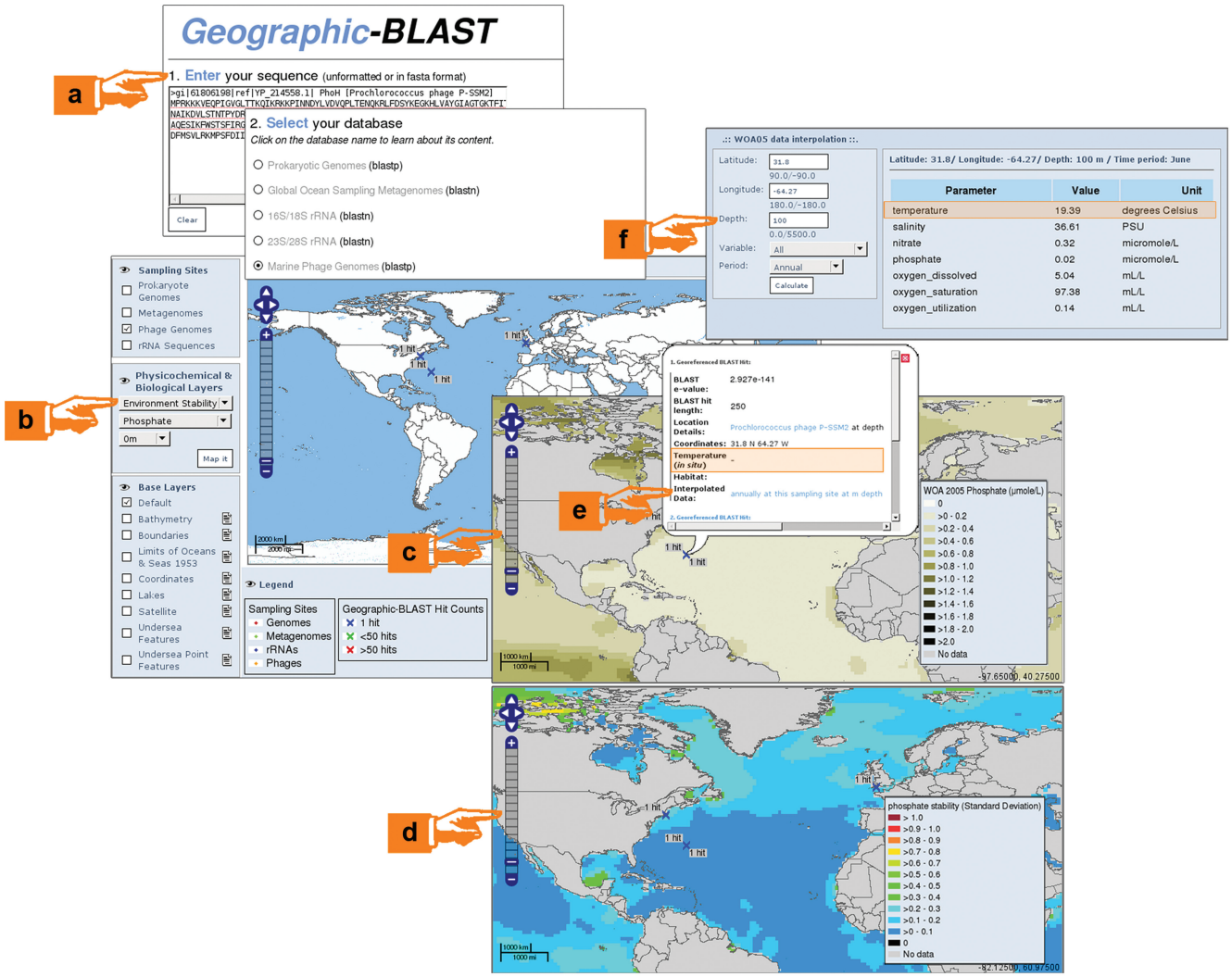

Figure 2. User test case: (**a**) BLAST sequence against the marine phage genomes to see the results on the Genes Mapserver. (**b**) View the BLAST hits with underlying environmental data, such as (**c**) average annual phosphate values, or (**d**) stability of phosphate concentrations in terms of monthly standard deviations. (**e**) BLAST result information can be displayed in a pop-up window, (**f**) where you can link out to megx.net's GIS data interpolator.

information on how to use this service can be found. Megx.net also provides access to MIGS/MIMS reports in Genomic Contextual Data Markup Language (GCDML) XML files for all marine phage genomes through similar HTTP queries, e.g. http://www.megx.net/gcdml/Prochlorococcus_phage_P-SSP7.xml (7,9).

Other changes

The massive influx of sequence data in the last years will out-compete the ability of scientists to analyze it (20). This development already pushes megx.net's capability to provide comprehensive pre-computed data to the limit. To better focus on integration of molecular sequence, contextual and environmental data, megx.net no longer offers pre-computed analyses, especially considering that other facilities, such as MG-RAST and CAMERA have emerged. Furthermore, the 'EasyGenomes Browser' has been replaced with links to the NCBI Genome Projects.

SUMMARY

Since its first publication (10), megx.net has undergone extensive development. The web design has been revamped for better user experience, and the database content greatly enhanced, providing considerably more genomes and metagenomes, marine phages and rRNA sequence data.

Megx.net's unique integration of environmental and sequence data allows microbial ecologists and marine scientists to better contextualize and compare biological data, using, e.g. the Genes Mapserver and GIS Tools. The integrated datasets facilitate a holistic approach to understanding the complex interplay between organisms, genes and their environment. As such, megx.net serves as a fundamental resource in the emerging field of ecosystem biology, and paves the road to a better understanding of the complex responses and adaptations of organisms to environmental change.

Database access

The database and all described resources are freely available at http://www.megx.net/.

Continuously updated statistics of the content are available at http://www.megx.net/content. A web feed for news related to megx.net is available at http://www.megx.net/portal/news/. Feedback and comments, the most effective springboard for further improvements, are welcome at http://www.megx.net/portal/contact.html and via email to megx@mpi-bremen.de.

Overall, it is important to note that the megx.net website does not fully reflect the content and search functionalities of MegDB. For any specialized data request, contact the corresponding author.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to acknowledge Timmy Schweer, Thierry Lombardot, Magdalena Golden and Laura Sandrine for their valuable input to megx.net, as well as David E. Todd for redesigning the web page.

FUNDING

FP6 EU project MetaFunctions (CT 511784); Network of Excellence 'Marine Genomics Europe'; Max Planck Society. Funding for open access charge: Max Planck Society.

Conflict of interest statement. None declared.

REFERENCES

1. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D.Y., Paulsen,I., Nelson,K.E., Nelson,W. *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
2. Rusch,D.B., Halpern,A.L., Sutton,G., Heidelberg,K.B., Williamson,S., Yooseph,S., Wu,D., Eisen,J.A., Hoffman,J.M., Remington,K. *et al*. (2007) The Sorcerer II Global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol*., **5**, e77.
3. Yooseph,S., Sutton,G., Rusch,D.B., Halpern,A.L., Williamson,S.J., Remington,K., Eisen,J.A., Heidelberg,K.B., Manning,G., Li,W. *et al*. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol*., **5**, e16.
4. Seshadri,R., Kravitz,S.A., Smarr,L., Gilna,P. and Frazier,M. (2007) CAMERA: a community resource for metagenomics. *PLoS Biol*., **5**, e75.
5. Markowitz,V.M., Ivanova,N.N., Szeto,E., Palaniappan,K., Chu,K., Dalevi,D., Chen,I.M.A., Grechkin,Y., Dubchak,I., Anderson,I. *et al*. (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acid Res*., **36**, D534–D538.
6. Meyer,F., Paarmann,D., D'Souza,M., Olson,R., Glass,E.M., Kubal,M., Paczian,T., Rodriguez,A., Stevens,R., Wilke,A. *et al*. (2008) The Metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.
7. Field,D., Garrity,G., Gray,T., Morrison,N., Selengut,J., Sterk,P., Tatusova,T., Thomson,N., Allen,M.J., Angiuoli,S.V. *et al*. (2008) The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol*., **26**, 541–547.
8. Field,D., Morrison,N., Glöckner,F.O., Kottmann,R., Cochrane,G., Vaughan,R., Garrity,G., Cole,J., Hirschman,L., Schriml,L. *et al*. (2008) Working together to put molecules on the map. *Nature*, **453**, 978.
9. Kottmann,R., Gray,T., Murphy,S., Kagan,L., Kravitz,S., Lombardot,T., Field,D., Glöckner,F.O. and Genomic Standards Consortium. (2008) A standard MIGS/MIMS compliant XML schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS*, **12**, 115–121.
10. Lombardot,T., Kottmann,R., Pfeffer,H., Richter,M., Teeling,H., Quast,C. and Glöckner,F.O. (2006) Megx.net–database resource for marine ecological genomics. *Nucleic Acid Res*., **34**, D390–D393.
11. Pruesse,E., Quast,C., Knittel,K., Fuchs,B.M., Ludwig,W.G., Peplies,J. and Glöckner,F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acid Res*., **35**, 7188–7196.
12. Brussaard,C.P.D., Wilhelm,S.W., Thingstad,F., Weinbauer,M.G., Bratbak,G., Heldal,M., Kimmance,S.A., Middelboe,M., Nagasaki,K., Paul,J.H. *et al*. (2008) Global-scale processes with a nanoscale drive: the role of marine viruses. *ISME J*., **2**, 575–578.
13. 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 Acid Res*., **36**, D475–D479.
14. Hirschman,L., Clark,C., Cohen,K.B., Mardis,S., Luciano,J., Kottmann,R., Cole,J., Markowitz,V., Kyrpides,N., Morrison,N. *et al*. (2008) Habitat-Lite: a GSC case study based on free text terms for environmental metadata. *OMICS*, **12**, 129–136.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
16. Richter,M., Lombardot,T., Kostadinov,I., Kottmann,R., Duhaime,M.B., Peplies,J. and Glöckner,F.O. (2008) JCoast - a biologist-centric software tool for data mining and comparison of prokaryotic (meta) genomes. *BMC Bioinformatics*, **9**, 177.
17. Meyer,F., Goesmann,A., McHardy,A.C., Bartels,D., Bekel,T., Clausen,J., Kalinowski,J., Linke,B., Rupp,O., Giegerich,R. *et al*. (2003) GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acid Res*., **31**, 2187–2195.
18. Sullivan,M.B., Coleman,M.L., Weigele,P., Rohwer,F. and Chisholm,S.W. (2005) Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol*., **3**, 790–806.
19. Williamson,S.J., Rusch,D.B., Yooseph,S., Halpern,A.L., Heidelberg,K.B., Glass,J.I., Andrews-Pfannkoch,C., Fadrosh,D., Miller,C.S., Sutton,G. *et al*. (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE*, **3**, e1456.
20. (2009) Metagenomics versus Moore's law. *Nat. Methods*, **6**, 623.