

# **Bioinformatic Resources for Marine Microbial Genomics**

Renzo Kottmann,<sup>1</sup> Andreas Wilke,<sup>2</sup> Folker Meyer,<sup>3</sup> Frank Oliver Glöckner<sup>1,4</sup>

<sup>1</sup> Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany

<sup>2</sup> Computation Institute, University of Chicago/Argonne National Laboratory, Chicago, IL 60637

<sup>3</sup> Argonne National Laboratory, Mathematics and Computer Science Division, 9700 South Cass Avenue,  
Argonne, IL 60439

<sup>4</sup> Jacobs University Bremen gGmbH, D-28759 Bremen, Germany

\* to whom correspondence should be addressed

Max Planck Institute for Marine Microbiology

Celsiusstrasse 1

D-28359 Bremen, Germany

Phone: +49 421 2028

FAX: +49 421 2028580

Email: @mpi-bremen.de

Keywords: genomics, marine, bioinformatics, integration, infrastructure



## **Abstract**

Large-scale marine sequencing projects promise insights into the catalytic potential of marine genomes or entire communities and have the potential to enhance our ability to monitor, model, and predict changes in the marine ecosystem. The number of base pairs generated in sequence projects is growing with each generation of sequencing platforms. Thus, bioinformatics and computational resources are becoming critical factors in the success of marine genomics projects.

However, only a few such resources currently are dedicated to marine genomics. Moreover, there exists a tremendous need for community-based bioinformatic infrastructures that meet twofold challenges of (1) efficiently processing and managing large amounts sequence data and (2) consistently integrating domain-specific contextual data to facilitate large-scale comparative genomic studies.

## **Introduction**

Molecular biology has undergone a paradigm shift in the past few years. Data-driven high-throughput studies are revolutionizing many research areas. The genomic revolution is rooted in medicine and biotechnology, but marine genomics currently delivers a great quantity of data in its own right. At the time of publication, the marine metagenome sequencing of the Global Ocean Sampling (GOS) campaign doubled the content in the public sequence repositories (Yooseph, et al., 2007) and confirmed the astonishing diversity of microbes. Currently, the Gordon and Betty Moore Foundation Marine Microbial Genome Sequencing Project, founded in 2004, has sequenced nearly 180 marine microorganisms, of which 80% are already published. The project is motivated by the fact that marine ecosystems cover more than 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. Therefore, insights into the genomic basis of their catalytic activities and interaction with the environment will enhance our ability to monitor, model, and predict changes in the marine ecosystem.

The impressive number and size of marine genome and metagenome projects are driven by astonishing advancements in sequencing technologies. Current and predicted trends in the development of new sequencing technologies show that the sheer pace of sequence data growth is unlikely to slow (Gupta, 2008, Hall, 2007, Metzker, 2010, Shendure and Ji, 2008, ten Bosch and Grody, 2008). Thus, genomics – including ecological genomics – is being transformed into a data-intensive science with an exponential increase of data (Szalay and Gray, 2006).

The rapid development of platforms for high-throughput experiments at lower costs can be observed in the fields of transcriptomics, proteomics, and metabolomics as well, providing scientists with a more holistic view of microbes in their natural, environmental context through multiomic studies.

Furthermore, these multiomics studies are extended to metatranscriptomics and metaproteomics, involving analysis of entire microbial communities. Indeed, multiomic studies not only significantly increase the size and complexity of genomic data; they demand the integration of diverse data to maximize scientific insights. The paradigm shift toward high-throughput experiments also shifts the workload toward bioinformatics and computational resources, which have become a critical factor for success. Indeed, the rate of sequence data generation is far outpacing the rate of increase in CPUs, and the cost of analyzing large datasets produced by, for example, Solexa, already exceeds the cost of generating them (Metzker, 2010, Meyer, 2006, Nature, 2009, Wilkening, et al., 2009). This situation is often characterized by fear-inducing metaphors such as “data tsunami,” “data avalanche,” and “data deluge.” Rather than being a threat to humankind, however, the technological improvements open challenging, but excellent opportunities for marine biology and biotechnology. The current exponential data production is a universal fact in biology and will enable a new kind of research limited only by our

computing power and bioinformatic capacity (Committee on Metagenomics: Challenges and Functional Applications, 2007, Szalay and Gray, 2006). The metagenomic approach may open a wide door to the rich metabolic and enzymatic repertoire of bacterial and archaeal communities for research in molecular ecology, ecological genomics, and marine biotechnology. However, the ability to make scientific use of the raw sequencing data heavily depends on the bioinformatic resources available to the marine genomics community.

## **Bioinformatics Resources**

The value of sequences is realized only through its annotation, which expresses a scientific understanding of the raw data. Today, a range of annotation systems exists for single-genome analysis. These systems support the management and integration of data from computational analyses using diverse sets of bioinformatic algorithms and software tools (Médigue and Moszer, 2007). Examples include sequence assembly (Scheibye-Alsing, et al., 2009), gene finding, protein domain prediction (Finn, et al., 2008), protein function assignment (Juncker, et al., 2009, Rentzsch and Orengo, 2009), prediction of gene expression, and gene regulation accompanied by data from laboratory studies.

## **Large-Scale Sequence Analysis**

While the computational needs for single-genome analysis are solved and can be performed on today's commodity hardware, the new large-scale metagenomic sequencing projects – which generate 2,000-3,000 genome equivalents of sequence information per project – bring new challenges. On the one hand, the challenge is the sheer amount of sequence data per metagenome. On the other hand, all metagenome sequences are mere fragments of unknown organismal origin. Taken together, these challenges demand further development of software for assembly, gene calling, and annotation. Recently, several new and dedicated data processing and database resources have emerged to address

the current need for large-scale metagenomic data analysis and management, for example, CAMERA (Seshadri, et al., 2007), IMG/M (Markowitz, et al., 2008), and the MG-RAST platform (Meyer, et al., 2008). However, just a “simple” automatic annotation based on BLAST sequence similarity searches poses a severe computational bottleneck. Consider the following example: in November 2009, the MG-RAST server processed 278 metagenomes with an average of 33 Mbp of sequence data per project (see Figure 1). On a single high-end server (Xeon E5540, 8 cpu’s, 2.53 GHz, 16 GB ram) a complete BLAST analysis of a single project against a 2 Gbp nonredundant reference database would take three days. Hence, it would take more than two years to calculate all metagenome projects that were submitted to MG-RAST in a single month, or 30 such servers. Because BLAST computing time grows linearly with the reference database size, and the reference databases double every year, a comparable growth in the number of servers is needed. In other words, any institution that aims to keep pace with the growth of sequence data needs to *double* its budget each year. Even when accounting for the 15% yearly increase in CPU speed, a 1.7 factor budget growth is needed. This simplified example does not yet incorporate concomitant costs incurred as a result of increases in memory consumption, storage capacities, and network bandwidth for data exchange, as well as power consumption and cooling. The issue becomes even more severe when one considers that, unlike BLAST, several analysis tools have nonlinear computational time consumption and the growth of metagenomic datasets continues steadily (see Figure 2).

## **From Sequence Data to the Environmental Context**

It is increasingly apparent that the isolated analysis of genes and genomes provides limited information with respect to gaining deeper insight into the function of ecosystems. Environmental sequence data needs to be analyzed on the basis of geographic and environmental context (Field, 2008, Field, et al., 2008). The contextual data can describe the geographic location and habitat, the processing details,

from the time of sampling up to sequencing, and subsequent analyses of any sequence. Each such contextual data item expands the number of dimensions available in comparative genomics and downstream hypothesis testing (Hughes Martiny and Field, 2005) and is therefore important for hypotheses about functions of predicted genes and, on higher levels, for modeling species' responses to environmental change or the spread and niche adaptation of marine microbes. The addition of contextual data to sequences is therefore becoming as valuable as the four nucleotides that make up the sequences. An important prerequisite for successful usage of contextual data is the availability and storage of such data in an accurate, structured, and accessible fashion.

## **Integrating Sequence and Contextual Data**

Unfortunately, even conceptually simple contextual data-driven requests, such as “Give me the temperature at the sampling site of the microbial isolate of interest” or “Give me all unknown genes sampled at temperatures greater than 80 degrees Celsius,” are far from trivial.

The reason is as simple as it is profound. From the time of field sampling to final sequence analysis, a range of diverse data is produced among different scientific communities, where individual researchers process the samples with different protocols in different time frames. Often, the data is not deposited in public resources at all, or the submitters have the choice of up to a dozen repositories. Therefore, current molecular, environmental, and diversity data is fragmented, imprecise, or lost in lab books or proprietary private archives. Moreover, environmental data, such as temperature, cannot be stored consistently in the current records of sequences in the INSDC databases (Benson, et al., 2010, Leinonen, et al., 2010); and nor can a genome sequence be stored in databases dedicated to environmental data (compare Figure 3).

An important source of contextual data is the data taken in the field (on site), such as the geographic and environmental origin of the sample, as well as information about subsequent processing to obtain the DNA, and the sequencing itself. Given that each DNA sequence from the marine environment is properly geo-referenced, with geographic location, depth in the water column or sediment, and time of sampling, the environmental context can be significantly complemented with data from environmental databases.

### ***Environmental Databases***

Environmental databases collect a variety of geo-referenced observations of different bio- and physicochemical conditions. Hundreds of such databases exist worldwide. European databases targeting the ocean include Pangaea, the British Oceanographic Data Centre (BODC), and SeaDataNet, which store a wide range of environmental measurements, mostly from cruise expeditions.

From the international community, the World Ocean Atlas provides a set of objectively analyzed (one decimal degree spatial resolution) climatological fields of *in situ* measurements (World Ocean Atlas, 2005). The World Ocean Database is a collection of scientific, quality-controlled ocean profiles (World Ocean Database, 2005). SeaWiFS provides chlorophyll *a* data based on remote sensing (SeaWiFS, 2009).

Together they provide a rich set of biotic and abiotic data for the marine ecosystem that can be used for integrated analysis.

### ***Toward More Contextual Data in Bioinformatic Resources***

Only recently have bioinformatic resources begun to invest in better management and integration of contextual data. CAMERA hosts the fully geo-referenced GOS data set. IMG/M integrates rich details about the hosted genomes and metagenomes. Megx.net is the first resource to provide a comprehensive annotation of the environment of microbial genomes (Kottmann, et al.). The Barcode of

Life initiative (Ratnasingham and Hebert, 2007) successfully collaborated with INSDC to store latitude and longitude in the public sequence repositories.

## **From Single Systems to a Network of Resources**

A rich landscape of hundreds of bioinformatic resources exists (Cochrane and Galperin, 2010). Each of these resources has its own data processing approach and data model according to its biological research focus. IMG/M, MG-RAST, and CAMERA can be grouped as processing resources for genomes and metagenomes. They host large-scale genomic and metagenomic datasets and are backed with large-scale computing resources for a variety of sequence analysis tasks. CAMERA and megx.net are the only resources that explicitly support marine microbiology, though megx.net focuses on the integration of marine environmental and sequence data and should not be considered a processing resource.

More technically, these resources differ significantly in the nature of the data stored, the way the data is presented to users, and the way the data is published and made available to the public. Consequently, there exists a plethora of different data formats, storage systems, and data access methods.

For an individual user it is a laborious task to ascertain the best way to make use of the resources. To get the “big picture,” each user must locally download all data from all resources needed. This ad hoc integration of data is time consuming, error prone, and often infeasible because local resources are insufficient in terms of bioinformatic skills as well as computing and storage capabilities.

Integrated community resources are urgently needed. They would share the burden of data processing and provide each user with a single, easily accessible view on complete, quality-controlled data. This goal can be fulfilled only if resources join in a federated, but community coordinated, network of resources to work together and share processing resources, common data models, data formats,

synchronized data exchange, and data presentation. In order to build such a networking system, several issues must be taken into account.

## ***Interoperability***

Interoperability describes the capabilities of computing systems to exchange data. Given the multitude of databases in genomics, metagenomics, biodiversity, and environmental research, integrated access is of fundamental importance to exchange data. Today the World Wide Web is key to such exchanges of scientific data.

The possibility to browse the World Wide Web is based on dozens of low-level communication protocols and data format specifications that are hidden from the casual user. What the user sees is the graphical representation of text documents in HyperText Markup Language (HTML) conveyed using the Hypertext Transfer Protocol (http). The World Wide Web not only allows users to browse hyperlinked web pages ("human-web") but also allows software programs to automatically access and exchange data ("machine-web"). While users surf the web by means of web browsers, software programs communicate and remotely process data via web services. Several different architectures and frameworks for building web services exist. While SOAP (Simple Object Access Protocol) enjoyed popularity in the past, the REST (Representational State Transfer) approach (Fielding, 2000) has gained increasing attention recently. The popularity of REST is based on its simplicity. REST is merely a guide to how to best use the http protocol for machine-machine communication.

These technologies provide the functionality now essential to everyday scientific research. Biology, especially molecular biology, is said to be one of the first sciences to co-evolve with the digital repositories and now completely relies on the Internet. To achieve interoperability, however,

bioinformatic resources must decide on a common web service framework, a common data model, and a specified data format.

### ***Semantics and Ontologies***

The query “Give me all metagenomes sampled at temperatures higher than 80 degrees Celsius from the marine environment” introduces the problem of semantics encoded in the data. Submitting this query to two different resources can lead to different answers, since the two resources have different understandings of the term “marine environment.” Is it the ocean water column? Or does it include all environments influenced by the ocean, such as mangroves, tidal areas, and estuaries? This semantic problem is called homonymy, where same terms have different meanings. Another frequent issue of semantics is known as synonymy, whereby two different terms have the same meaning; for example, one resource uses the term “marine environment” and another resource uses “maritime environment.”

Today, ontologies are perceived as the best approach for solving semantic issues in data management. Ontologies are elaborated, controlled vocabularies that attempt to capture and hierarchically structure the semantics of main concepts in a knowledge domain (Gruber, 1993, Rubin, et al., 2008, Schulze-Kremer, 2002).

The Environmental Ontology (EnvO) (EnvO Consortium, 2009) defines and organizes the semantics of environment descriptions. For example, EnvO defines the term “marine habitat” as “a habitat that is in or on a sea or ocean containing high concentrations of dissolved salts and other total dissolved solids (typically >35 grams dissolved salts per litre).” Given that EnvO is shared by several resources, it guarantees that a user asking to retrieve information from “marine habitat” gets information based on a shared common understanding. Moreover, because EnvO, like all ontologies, is hierarchically structured, a resource implementing EnvO will return information for all items classified as “freshwater habitat” and

“marine habitat” if it is asked for information items from “aquatic habitat.” By incorporating an “is-a” relationship between the various expressions, it knows semantically that both freshwater and marine habitats are aquatic habitats.

The availability of a controlled vocabulary and the hierarchical structure to model “is-a” relationships can greatly facilitate semantic coherent data exchange. Because ontologies are themselves technically encoded data, however, successful usage of ontologies again depends on efficient solutions to interoperability. Thus, ontologies not only promise better semantic consistency of data but also contribute to the complex issue of interoperability and data management.

### ***Standards, Standards, and Standards***

Standards are everywhere. Every product we use is subject to some quality, processing, or production standard. These standards are either de facto standards, resulting from the wide acceptance of a common specification, or official standards resulting from the work of experts in national, international, or industry standards organizations. The Internet would not function without the standards defined by the World Wide Web Consortium (W3C). Scientific activities are similarly based on standards ranging from the International System of Units (SI), which defines the standard metric system and system of measurement, to standardized laboratory protocols for specialized experiments. Molecular biology and marine sciences further prosper through the use of standardized methods and protocols. Astonishingly, genomics so far has progressed without standards for critical analyses, such as assembly, gene finding and function predictions. Moreover, it lacks standards for the storage and exchange of raw sequence data and their respective analysis results (Kyrpides, 2009).

In 2005, the Genomics Standards Consortium (GSC) was established. Its goal is to promote mechanisms that standardize the description of genomes and thus facilitate the exchange and integration of genomic

data. This is the first well-organized effort to establish a constant community debate on standardization issues in genomics. Although the GSC is mostly a grass-roots community of volunteers, it has already established several successful projects, including the EnvO-Lite project, which established a manageable subset of terms from EnvO, especially accessible to non-ontology experts. The *Standards in Genomic Sciences (SIGS)* is an open-access, standards-supportive journal that seeks to rapidly disseminate concise genome and metagenome reports in compliance with GSC standards (Garrity, et al., 2008). In 2008, the GSC published the "Minimum Information about a (Meta-) Genome Sequence standard" (MIGS/MIMS), a specification of the minimum set of contextual data to accompany a genome sequence, which will aid comparative analyses (Field, et al., 2008). With the development of the Genomic Contextual Data Markup Language, the GSC also provides an XML implementation of MIGS/MIMS for use in web services-based data exchange (Kottmann, et al., 2008). Currently, the GSC further extends MIGS/MIMS to specify a checklist for the "Minimum Information about an ENvrionmental Sequence" (MIENS) for better description of marker genes from cultured organisms or the environment.

These standards are integral components in achieving an interoperable bioinformatics world (Stein, 2002, Stein, 2003). Of course, standards alone do not guarantee success (Ball, 2006, Brooksbank and Quackenbush, 2006, Burgoon, 2006). The value of standards in solving interoperability, integration, and semantic data issues depends on the quality of the standard itself and its acceptance and adoption in the scientific community.

## **Cyberinfrastructures and e-Infrastructures**

The advances in sequence technologies have led to petabyte-scale raw sequence data at an accelerating pace (Metzker, 2010). Already, this development makes the bioinformatic analysis a serious bottleneck. It is becoming apparent that no single computing infrastructure, not even a supercomputing center, can

keep pace in providing the computing power for the basic analysis tasks of sequence assembly, gene calling, and automatic annotation.

The first challenge is to create better algorithms and new strategies for the bioinformatic analysis of the raw sequence data. The second challenge is to supplement sequence data with contextual data to facilitate analyses in dimensions beyond gene annotation and comparative sequence studies. The key to successful and effective use of these datasets is interdisciplinary and international collaboration among computer scientists, software engineers, statisticians, theorists, and field researchers. The development of bioinformatic infrastructures is now a priority: biology has become a data-driven science, with computers and the Internet as essential for scientists as the laboratories in which they work.

To address these needs, European and U.S. funding agencies are developing strategies to better fund infrastructures, termed “cyberinfrastructures” in the U.S. and “e-infrastructures” in Europe. In the U.S., current infrastructures in biology include the Cancer Bioinformatics Grid (caBIG); the Biomedical Informatics Research Network (BIRN); and iPlant, a \$50 million program to tackle the biggest computational challenges in plant biology.

European development of research infrastructures is consolidated in the ESFRI (European Strategy Forum on Research Infrastructures) Roadmap. This roadmap comprises more than 30 infrastructure projects in research fields across Europe. ESFRI clearly defines e-infrastructures as critical to all projects. Notable e-infrastructure projects include Lifewatch (Science and Technology Infrastructure for Biodiversity Data and Observatories) for environmental sciences and ELIXIR (European Life-Science Infrastructure for Biological Information) for biological science.

These infrastructures will provide digital grid computing-based research environments that combine data and software for scientific communities. Furthermore, these infrastructures allow the scientific

communities to run sophisticated computational analyses on their domain specific data; facilitate visualization and interpretation; and allow exchange, sharing, and publication of the knowledge arising from those analyses.

## **Domain-Specific e-Infrastructure for Marine Genomics**

The most important reason current e-infrastructures are developed for specific domains of knowledge is that they are most applicable and accessible by users when developed by and for a specific research community. In marine genomics, “Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis” (CAMERA) is established in the U.S. Currently, there is no such infrastructure for marine genomics in Europe.

In Europe, megx.net is the first resource to provide access to geographically integrated information on microbial genes and genomes in their marine environmental context. It makes available contextual data on several hundreds of genomes and metagenomes from prokaryotes and phages, as well as over a million small and large subunit ribosomal RNA sequences. In addition to storing all available “on-site” data describing sampling time, location, and field measurements of genomic sampling events, megx.net allows *post factum* retrieval of interpolated environmental parameters, such as temperature and pH, for any location in the ocean waters based on global profile and remote sensing data from environmental databases.

The ever growing mass of sequence and associated data, however, demands bioinformatic skills and resources often exceeding the available computing resources in marine laboratories. Additionally, the accurate and consistent handling of the complex contextual data needed for high-dimensional analysis and modeling demands dedicated standardization, software, and integrated database development. A European bioinformatic e-infrastructure for marine genomics would need to integrate bioinformatic

predictions with environmental data from the ocean, seas, and European coasts to gain knowledge about the genomic basis of microbial lifestyles, adaptations, and fitness in response to the marine environment (see Figure 3). Notably, the integration, visualization, analysis, and interpretation of the data are domain-specific tasks that can be done only by a community of marine experts. Such a specific e-infrastructure, implemented as an integral component of ELIXIR, would gather a virtual community across disciplines and borders and would allow and safeguard vital open access to integrated marine specific data.

## **Conclusions**

Marine genomics greatly benefits from rapid development in sequencing technologies. Marine large-scale sequencing projects promise insights into the catalytic potential of marine genomes or entire communities. Full use of the sequence data can be made, however, only if appropriate community-based bioinformatic infrastructures exist that meet twofold challenges of (1) efficiently processing and managing the ever growing amount of sequence data and (2) consistently integrating domain-specific contextual data to facilitate large-scale comparative studies of different marine samples. Such infrastructure would allow better and cost-effective use of computing resources that already exceed the laboratory costs of sequence generation. A standard information infrastructure integrating diversity and genomics of ecosystem functioning is the basis for an international and interdisciplinary community resource in marine genomics. Tools for the visualization and analysis of the integrated information allow researchers to effectively develop better approaches to monitor, model, and predict changes in the marine ecosystem. Such newly generated knowledge will help us to make better, sustainable use of our largest ecosystem on earth.

## Acknowledgments

This work was supported by the Max Planck Society and in part by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract DE-AC02-06CH11357.

## References

- Ball, C. A. (2006) Are we stuck in the standards? *Commentary* **24**: 1374-1376.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2010) GenBank. *Nucleic Acids Res* **38**: D46-51.
- Brooksbank, C., and Quackenbush, J. (2006) Data standards: a call to action. *OMICS* **10**: 94—99.
- Burgoon, L. D. (2006) The need for standards, not guidelines, in biological data reporting and sharing. *Nat Biotechnol* **24**: 1369—1373.
- Cochrane, G. R., and Galperin, M. Y. (2010) The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucleic Acids Res* **38**: D1-4.
- Committee on Metagenomics: Challenges and Functional Applications, N. R. C. (2007) *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. The National Academies Press, Washington DC.
- EnvO Consortium (2009) Environmental Ontology (EnvO). In. Morrison, N. (ed).
- Field, D. (2008) Working together to put molecules on the map. *Nature* **453**: 978.
- Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., et al. (2008) The minimum information about a genome sequence (MIGS) specification. *Nat Biotech* **26**: 541-547.
- Fielding, R. T. (2000) *Architectural Styles and the Design of Network-based Software Architectures*. In. Irvine: University of California.
- Finn, R. D., Tate, J., Mistry, J., Coghill, P. C., Sammut, S. J., Hotz, H.-R., et al. (2008) The Pfam protein families database. *Nucl. Acids Res.* **36**: D281-288.
- Garrity, G. M., Field, D., Kyrpides, N., Hirschman, L., Sansone, S.-A., Angiuoli, S., et al. (2008) Toward a standards-compliant genomic and metagenomic publication record. *Omics* **12**: 157-160.
- Gruber, T. R. (1993) *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. In.: Knowledge System Laboratory, Stanford University.
- Gupta, P. K. (2008) Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol* **26**: 602-611.

- Hall, N. (2007) Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol* **210**: 1518-1525.
- Hughes Martiny, J. B., and Field, D. (2005) Ecological perspectives on the sequenced genome collection. *Ecology Letters* **8**: 1334-1345.
- Juncker, A., Jensen, L., Pierleoni, A., Bernsel, A., Tress, M., Bork, P., et al. (2009) Sequence-based feature prediction and annotation of proteins. *Genome Biology* **10**: 206.
- Kottmann, R., Gray, T., Murphy, S., Kagan, L., Kravitz, S., Lombardot, T., et al. (2008) A Standard MIGS/MIMS Compliant XML Schema: Toward the Development of the Genomic Contextual Data Markup Language (GCDML). *Omic* **12**: 115-121.
- Kottmann, R., Kostadinov, I., Duhaime, M. B., Buttigieg, P. L., Yilmaz, P., Hankeln, W., et al. Megx.net: integrated database resource for marine ecological genomics. *Nucl. Acids Res.* **38**: D391-395.
- Kyrpides, N. C. (2009) Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nat Biotech* **27**: 627-632.
- Leinonen, R., Akhtar, R., Birney, E., Bonfield, J., Bower, L., Corbett, M., et al. (2010) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res* **38**: D39-45.
- Markowitz, V. M., Ivanova, N. N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., et al. (2008) IMG/M: a data management and analysis system for metagenomes. *Nucl. Acids Res.* **36**: D534-538.
- Médigue, C., and Moszer, I. (2007) Annotation, comparison and databases for hundreds of bacterial genomes. *Research in Microbiology* **158**: 724-736.
- Metzker, M. L. (2010) Sequencing technologies - the next generation. *Nat Rev Genet* **11**: 31-46.
- Meyer, F. (2006) Genome Sequencing vs. Moore's Law: Cyber Challenges for the Next Decade. *CTWatch Quarterly* **2**.
- Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386.
- Nature (2009) Metagenomics versus Moore's law. In: *Nat Meth*.
- Ratnasingham, S., and Hebert, P. (2007) bold: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Mol Ecol Notes* **7**: 355-364.
- Rentzsch, R., and Orengo, C. A. (2009) Protein function prediction – the power of multiplicity. *Trends in Biotechnology* **27**: 210-219.
- Rubin, D. L., Shah, N. H., and Noy, N. F. (2008) Biomedical ontologies: a functional perspective. *Brief Bioinform* **9**: 75-90.

Scheibye-Asling, K., Hoffmann, S., Frankel, A., Jensen, P., Stadler, P. F., Mang, Y., et al. (2009) Sequence assembly. *Computational Biology and Chemistry* **33**: 121-136.

Schulze-Kremer, S. (2002) Ontologies for molecular biology and bioinformatics. *In Silico Biology* **2**: 17.

SeaWIFS (2009). In.

Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P., and Frazier, M. (2007) CAMERA: a community resource for metagenomics. *PLoS Biol* **5**: e75.

Shendure, J., and Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol* **26**: 1135-1145.

Stein, L. (2002) Creating a bioinformatics nation. *Nature* **417**: 119—120.

Stein, L. D. (2003) Integrating biological databases. *Nat Rev Genet* **4**: 337-345.

Szalay, A., and Gray, J. (2006) 2020 Computing: Science in an exponential world. *Nature* **440**: 413-414.

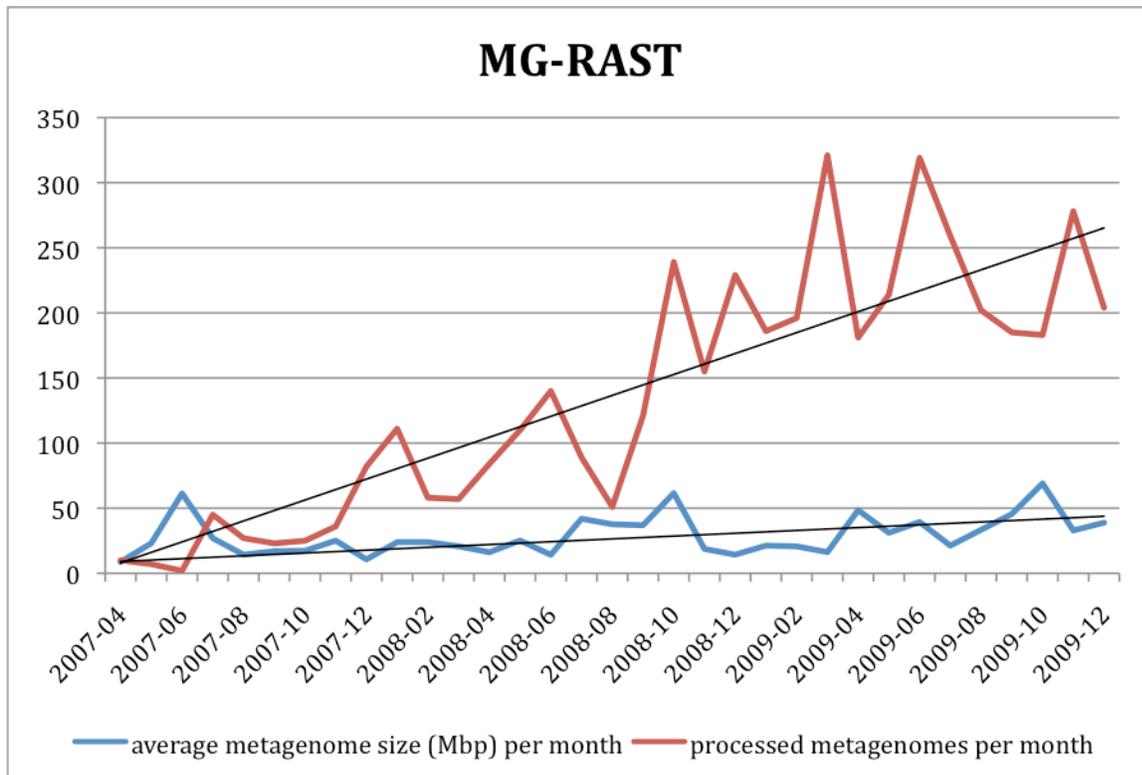
ten Bosch, J. R., and Grody, W. W. (2008) Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *J Mol Diagn* **10**: 484-492.

Wilkening, J., Wilke, A., N, D., and Folker, M. (2009) Using Clouds for Metagenomics: A Case Study. In: Proceedings IEEE Clouds.

World Ocean Atlas (2005). In.

World Ocean Database (2005). In.

Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., et al. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**: e16.



**Figure 1.** From 2007 until 2009 the MG-RAST has processed 4,429 metagenomic data sets, of which 420 have been made publicly available and can be accessed through the MG-RAST web interface. The average number of jobs per month has increased from 6 to 220 (as of December 2009). Additionally, the average size of a metagenomic dataset increased up to 40 Mbp with a maximum average size of 69 Mbp. The largest dataset has a size of 1.8 Gbp.

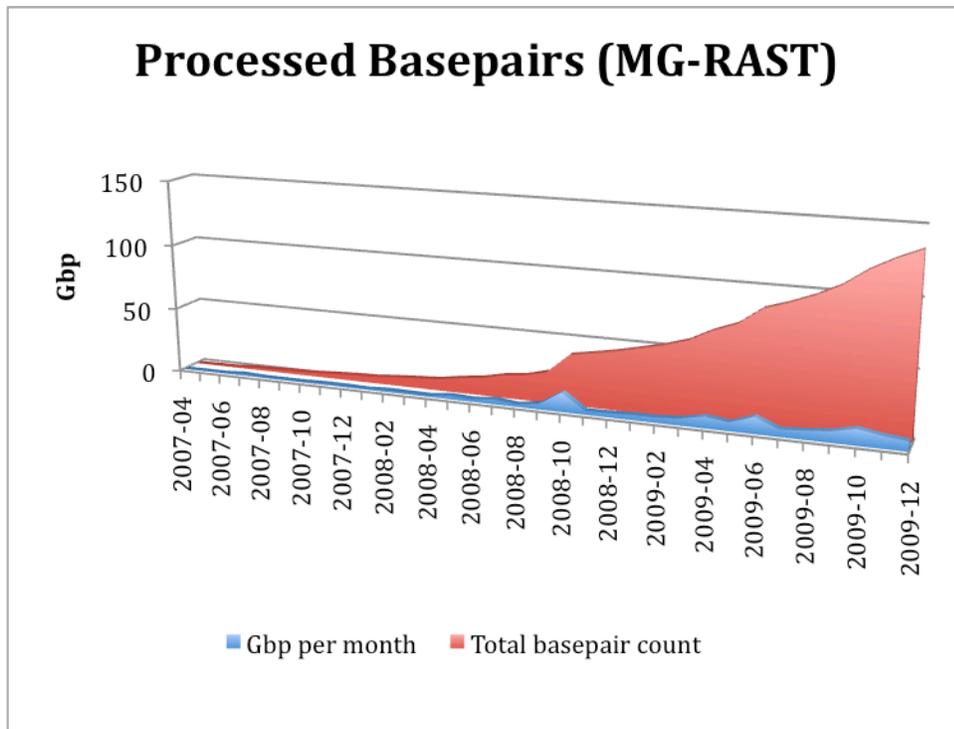


Figure 2. MG-RAST processed a total number of 120 Gbp by the end of 2009 (red) alongside the steady increase in Gbp processed per month.

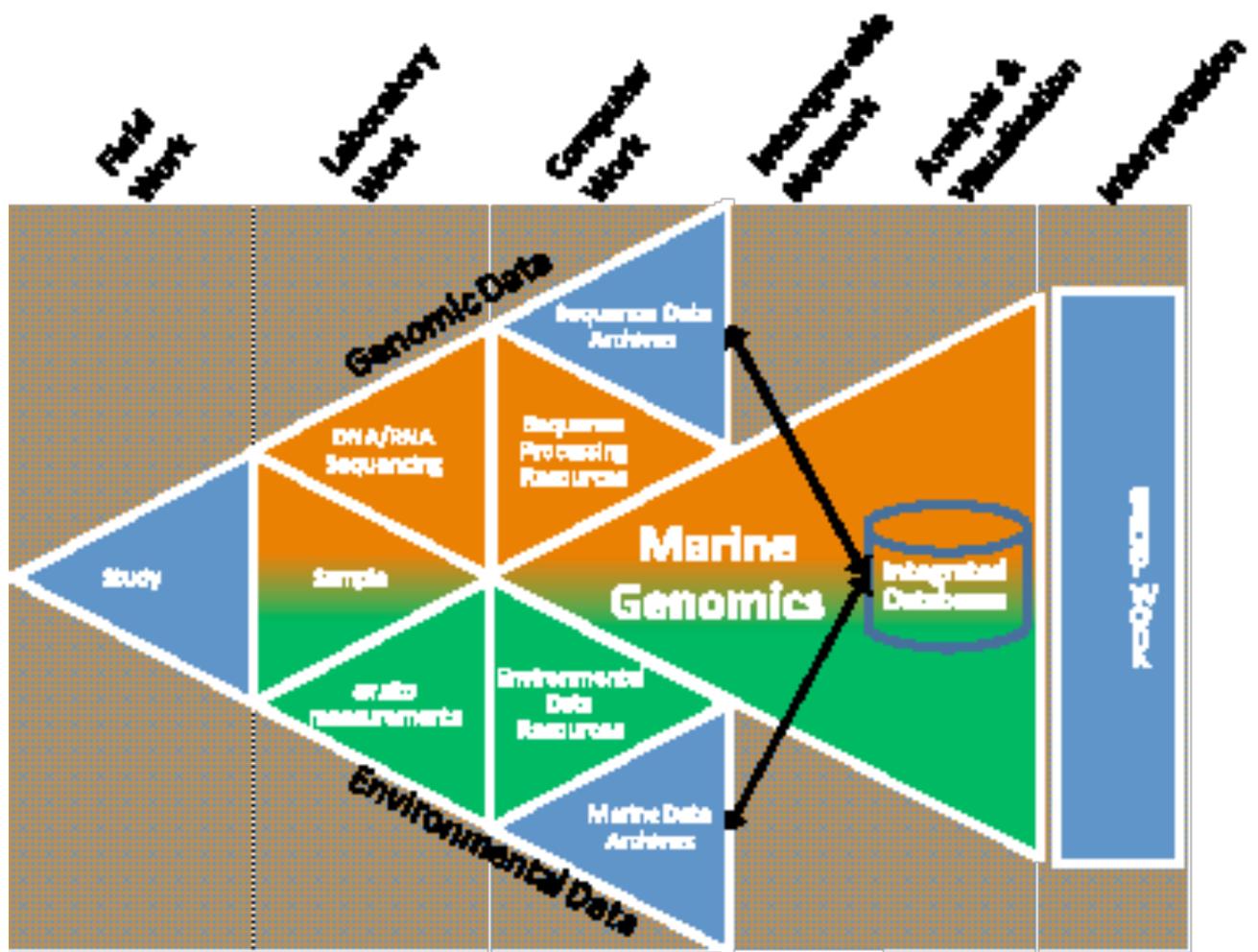


Figure 3. A model for marine genomics e-infrastructure. The positioning in the middle between genomic- and environmental science, both having own data flows, emphasizes the importance of contextual data integration.

The submitted manuscript has been created in part by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.