Conceptualizing a Genomics Software Institute (GSI)

Jack A. Gilbert[1,2*], Charlie Catlett[1,3], Narayan Desai[1,3], Rob Knight [4,5], Owen White[6], Robert Robbins[7], Rajesh Sankaran[1,3], Susanna-Assunta Sansone[8], Dawn Field[8,9], Folker Meyer[1,3]

[1]Argonne National Laboratory, Argonne, IL, USA
[2]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA
[3]Computation Institute, University of Chicago, , Chicago, IL USA
[4]Department of Chemistry and Biochemistry, University of Colorado at Boulder, Boulder, CO, USA.
[5]Howard Hughes Medical Institute, Boulder, CO, USA
[6]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, USA
[7]University of California, San Diego, La Jolla, California, USA
[8]Oxford e-Research Centre, University of Oxford, Oxford, UK
[9]Centre for Ecology & Hydrology, Natural Environment Research Council, Crowmarsh Gifford, Wallingford, Oxon, UK

*__Corresponding author__: Jack A. Gilbert (gilberjack@gmail.com)

Microbial ecology has been enhanced greatly by the ongoing 'omics revolution, bringing half the world's biomass and most of its biodiversity into analytical view for the first time; indeed, it feels almost like the invention of the microscope and the discovery of the new world at the same time. With major microbial ecology research efforts accumulating prodigious quantities of sequence, protein, and metabolite data, we are now poised to address environmental microbial research at macro scales, and to begin to characterize and understand the dimensions of microbial biodiversity on the planet. What is currently impeding progress is the need for a framework within which the research community can develop, exchange and discuss predictive ecosystem models that describe the biodiversity and functional interactions. Such a framework must encompass data and metadata transparency and interoperation; data and results validation, curation, and search; application programming interfaces for modeling and analysis tools; and human and technical processes and services necessary to ensure broad adoption. Here we discuss the need for focused community interaction to augment and deepen established community efforts, beginning with the Genomic Standards Consortium (GSC), to create a science-driven strategic plan for a Genomic Software Institute (GSI).

Introduction

The importance of the microbial world has long been recognized in all aspects of humanity, from global biogeochemical cycles to food chains, and from agriculture to animal and human diseases [1]. More recently, the advent of detailed, high-throughput multi-omic analyses used in metagenomics, metatranscriptomics, metaproteomics, and metametabolomics has enabled an unprecedented degree of resolution in the examination of these systems. By applying high-resolution characterization of microbial communities, efforts such as the Earth Microbiome Project [2] have revealed the dynamic, tightly interconnected relationship between microbes and their environment. Yet these new sequencing and analysis capabilities have also created significant challenges such as integrating the information extracted from large volumes of data into useful and accessible knowledge, specifically concerning the dynamics of multiscale, complex systems. The lack of standard data and metadata formats, application programming interfaces, and frameworks for finding and sharing even modest sized data sets, prevents these results from being readily shared and used. Moreover, most research groups do not have sufficient resources to build the capabilities to perform such characterization, and thus focus on the composition of a particular, individual microbial community (Figure 1).

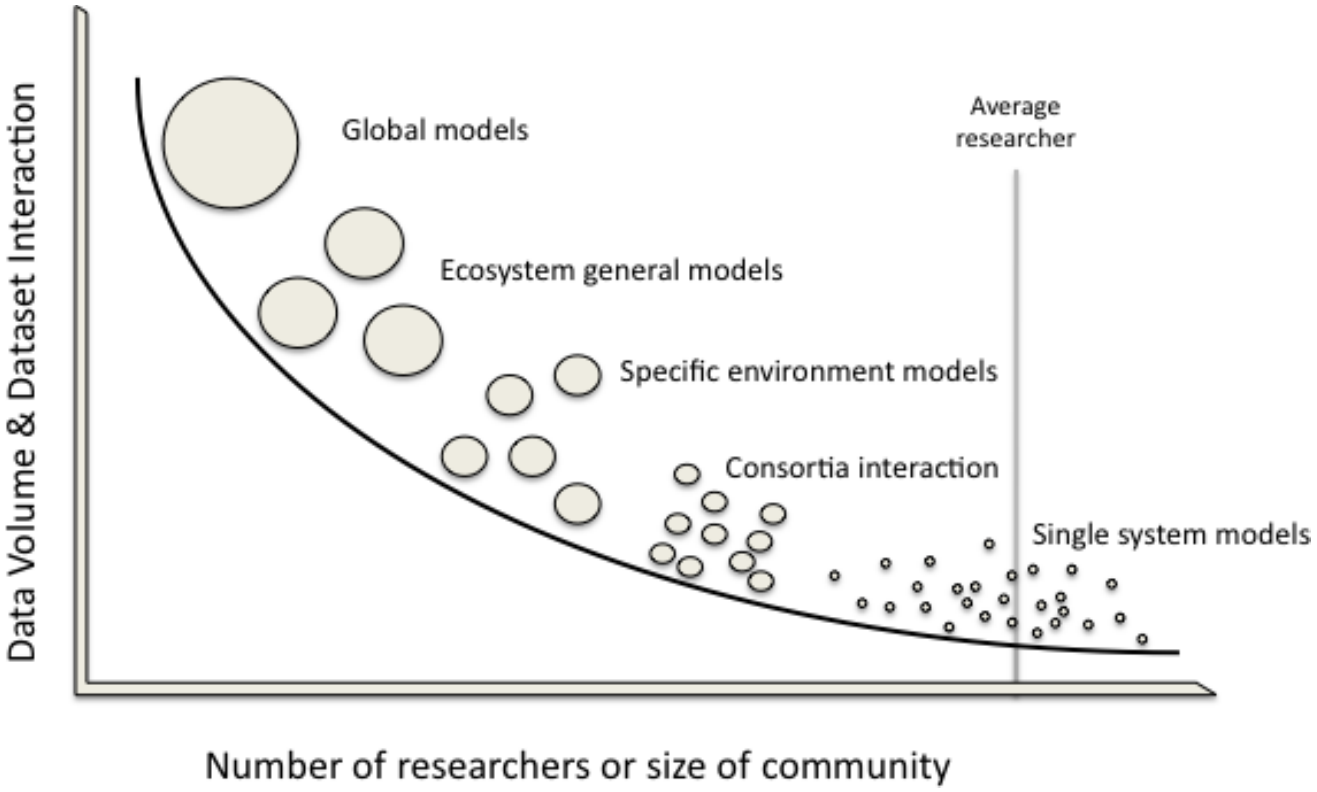

Figure 1. As access to data increases, researchers will be able to integrate a greater volume of data and more data types. This will result in a shift in the scale of addressable questions from the current position near single system models, to the goal near the left of large-scale questions and global model creation.

The characterization of a single microbial ecosystem is no longer sufficient, as the deeply interconnected nature of these systems leads to many direct and indirect impacts virtually everything that is performed and produced by human society. Consequently, understanding the metabolic dynamics, ecological interactions and evolutionary processes present in and among microbial ecosystems is of critical importance to our ability to predict the impact of human activity on these systems. Indeed, while human activities impact on these systems, understanding the dynamics is also foundational to affecting sustainable, and ultimately constructive, human engagement to create a mutually beneficial relationship.

While biological systems research has benefited immensely from technological advances in the acquisition of data and its interpretation into knowledge, this success has created new challenges. The plethora of data generated by current (much less projected) sequencing technologies, operated by a rapidly growing number of groups, and combined with the resulting analysis products, greatly outstrips our cognitive capacity and technical capabilities to effectively utilize it [3,4]. Concurrently, the volume and scale of these data sources and products overpower traditional analysis and curation tools and techniques. Consequently there is an urgent need to develop scalable methods to effectively and efficiently transform biological data into usable and accessible information, and to make such data and information broadly available to enhance discovery. Compounding these challenges, the number of possible hypotheses extracted from that data is overwhelming the ability of researchers to process, integrate and utilize that potential *knowledge.* Performing *in silico* experiments will both expand and accelerate the processes of examination and evaluation of hypotheses, identifying plausible subsets of hypotheses and suggesting future real world experiments to further refine those hypotheses. However, for many biologists the full range of mathematical and computational methods available remains elusive due to the absence of widely used data access and transparency mechanisms and an increasingly fragmented body of opaque, non-interoperable data.

Most research groups need to move from characterizing an individual microbial community to its ecosystem, and to access and leverage the results of groups who are characterizing other microbial ecosystems. Those groups working at the ecosystem granularity are beginning to pursue research questions that require a jump from characterizing an ecosystem in isolation to understanding its interaction with other ecosystems [5-7]. The entire community has recognized that forward progress along this continuum can be realized through a collaborative, focused initiative comprising several foundational capabilities. First, combining data from multiple sequencers, from multiple microbial communities, and from multiple ecosystems can only be achieved through widely adopted, science-driven community standards. Second, application programming interface (API) standards are essential to harnessing these data and metadata standards and to making results accessible. This white paper calls for a community-wide, interdisciplinary effort to identify key requirements for improved biological data access, analysis, and curation to enable environmental microbiological research at macro scales, and to facilitate the characterization and understanding of the biodiversity and interdependence of microbial communities on this planet. The goal of a Genomic Software Institute (GSI) would be to democratize microbial sciences by significantly reducing the threshold for researchers to access and combine data from different sources, allowing the use of cutting-edge computational modeling and simulation tools and techniques to enhance and accelerate the iterative cycle of observation, analysis, and experiment. Creating a GSI will require an interdisciplinary community effort to identify and prioritize requirements for a "data centric architecture" whose target is a scalable, science-driven cyberinfrastructure that anticipates next-generation data volumes and computational capabilities, positioning the science community to explore the microbial kingdom at realistic scales.

Metagenomics and the dynamics of an emerging field

Molecular analysis of microbial communities is a relatively young field; prior to 1975 there was no sequence data associated with microbes. The first data from microbial communities was generated during the next decade, but it took another decade to confirm that microbial communities were extremely diverse, and that the available cultured representatives poorly represented their taxonomy and functional potential [8]. Improvements in sequencing, storage, and computing technologies during the last ten years have converged to create the new field of metagenomics. The rapid acceleration of this new field has resulted in a profusion of incompatible tools, data formats, and analysis products. In part this is attributable to competition among independent research teams. Equally at issue, however, is that the extent of microbial diversity allows research teams to steadily produce results without venturing beyond a given microbial community to explore the context of their microbe or function outside of its immediate interactions. However, a central motivation for understanding individual microbial communities is in fact to discover the dynamics of their interactions.

In the last five years, next-generation sequencing technologies have catalyzed a rapid development and democratization of large scale sequencing efforts for exploring a myriad of microbial ecosystems with exceptional depth and breadth. The explosion of data in terms of both the number of sources and the sheer volume generated by these sources has begun to outpace the tools available for data curation, access, analysis, and dissemination. Consequently, the community is realizing that facilitating the next 10 years of microbial ecological analysis will require a shift in how we attack the problem.

Despite the fragmentation of data and tools, the community itself has developed a remarkably strong foundation of integrative collaborations and shared resources. A key development has been the Genomic Standards Consortium (GSC) [9], which has acted as community-glue, providing a valuable forum for interaction between biologist and computational researchers, and the development of many new software and standards solution products via community consensus. Other community-led informatics efforts including the many data sharing activities now unified under the BioSharing banner (i.e. Ibvestigation /Study/Assay [10]), GSC, etc), large software efforts that are advancing the field of microbial ecology (Department of Energy's KnowledgeBase [20] Quantitative Insights into Microbial Ecology (QIIME [21]), MG-RAST [11]) and efforts to bring suites of software together (Open Microbiome Initiative (OMI – [22]), BioCode Commons [23], etc.) provide a growing framework on which to build a

Genomics Software Institute (GSI). Most recently, the Data-Enabled Life Sciences Alliance (DELSA) has provided an additional interaction forum for thinking about how to deal with data intensive areas of science. Such an institute is needed and required by the community to reduce the hurdles individual researchers must leap to access data, reformat data, and generate ecosystem and organism level models. The biological community has many hubs, which are beginning to self-organize under the auspices of community-led initiatives. These initiatives will act as conduits to enable interaction within the proposed conceptualization stage.

The community is now realizing that the next decade of microbial ecological analysis will require a fundamental shift in data management and sharing. The future will see a virtuous cycle of [1] hypothesis driven data generation [2], *in silico* coding of the generated knowledge [3], mining the data to improve our encoded models [4], and using these models to generate iterative hypotheses to accelerate knowledge discovery.

A Genomics Software Institute

Conceptualizing a GSI for the life sciences community will require a new depth of collaboration between biologists, computational scientists, and computer scientists. Without a substantive *co-design* approach, tools, data formats, APIs, and other critical aspects and components of cyber-infrastructure may be optimized for technology, or for general use-cases, yet fall short of catalyzing truly new science. Similarly, the lack of sufficient technical collaboration among life science teams has led to many overlapping, duplicate efforts, none of which has the critical mass required to address the scale and complexity necessary for advancing the field. Currently, the ability to move data from one community to another, to combine tools into complex workflows, or to apply tools across multiple databases (where possible at all) is limited to data volumes that are orders of magnitude too small.

Breaking this cycle, then, must begin with a conceptualization effort that can work with the biological community to determine and focus on a set of common challenges and opportunities that are achievable through a collective, coordinated effort and at the same time offer revolutionary advances within and between life sciences communities.

The Investigation/Study/Assay ISA commons community [10,12] provides an illustration of the feasibility of this approach. ISA is a growing exemplary ecosystem of data curation and sharing solutions built on the ISA metadata-tracking framework. It provides tools and resources to harmonize metadata descriptions of disparate datasets, enabling data commoning through invisible compliance with the community standards described in the BioSharing catalogue [24]. These collaborative groups are, in essence, on the path to building a data commons, serving an increasingly diverse set of domains including environmental health, environmental genomics, metabolomics, (meta)genomics, proteomics, stem cell discovery, systems biology, transcriptomics, toxicogenomics, and communities working to characterize nucleic acid structures and to build a library of cellular signatures. The ISA commons illustrates the potential of the synergistic approach and the horizontal integration that transcends individual life science domains and assay- or technology-focused communities.

BioSharing works at the global level to build stable linkages, especially between journals, funders, implementing data sharing policies, and well-constituted standardization efforts in the biosciences domain. Its goal is to (i) address overlaps and duplication of efforts that hamper the wider uptake of standards and interfere with the creation of standards-compliant tools, and (ii) expedite the production of an integrated standards-based framework for the capture and sharing of high-throughput genomics and functional genomic bioscience data. The web-based BioSharing catalogue is a "one-stop shop" for those seeking data-sharing policy documents and information about the standards and technologies that support them. It exposes core information on well-constituted, community-driven standardization efforts and links to their standards, documentation, training material and point of contact.

Biology has benefited from an unprecedented growth in sequence generation capabilities, which outstrips even Moore's Law that states the processing speed of computational technology doubles every 18 months [13]. While this data represents a profound opportunity for the community, its rapid arrival has left substantial gaps in the data culture and practice in biology. These gaps include infrastructure for data stewardship, re-use, and sharing. Moreover, the rapid change in costs for sequence data, as well as the algorithmic complexity of analysis has shifted the costs into analysis and away from data production, making analysis results the most valuable data the community has. While there has

been some progress made by the GSC in several of these areas such as metadata for re-use and the sharing of computational results [19], this field remains a nascent yet critically important one [9].

The growth in sequence data has resulted from a combination of decreasing costs and widespread deployment of next-generation sequencing technologies. These changes have resulted in a large-scale democratization of sequence data production, where an increasing share of data production occurs at moderate scale across the community. In this sort of data production regimen, it has become apparent that monolithic archives are unsustainable. Moreover, data growth has resulted in substantial analysis challenges for the community. Resources providing scalable analysis capabilities to the community have become de facto clearinghouses for high quality data. As an example of these two observations, the RAST server [14], which is a 'private' workbench for the community to analyze genomic data on, hosts nearly 11,000 novel bacterial genomes. These genomes, while not publicly available, greatly outnumber those available from the GenBank archive. This organizational legacy originates from a time when sequencing was costly.

However, all of the comments above merely highlight problems with handling "raw" sequence data. With sequencing becoming cheaper by a factor of 10 every year [15], the community needs to establish ways of sharing computational results and also increasing the efficiency of the algorithms and workflows used. By allowing results to be shared between groups, the community can reduce the overall usage of computing simply by downloading an analysis that has already been done. This raises a number of technical challenges, including data set identification, data transfer, provenance tracking, etc.

Mere data and sharing results however cannot replace a sustained effort to broaden the number of groups and individual researchers that are capable of contributing models and algorithms for data analysis and data creation. The community is "creativity" limited as much as it is currently analysis limited. Easier access for researchers at a broad range of institutions will allow them to recruit more talent for the process of providing more informative analyses, more computationally affordable analyses, and higher quality ecosystem models.

One interesting side effect of the establishment of a common distributed data ecosystem that provides easy access to data and exchangeability of data between centers will be the weakening of current data silos. However, these existing systems, like MG-RAST [11] and IMG/M [16], have been developed in isolation and are now attempting to create exchangeable data products. However, two are not enough. Bringing the community together to create a GSI will accelerate this transformation and enable exchange formatting to match the increased availability of data. The data products that can be shared today in e.g. metagenomics are merely the output of the sequencing platforms. The efforts of the Genomic Standards Consortium's M5 group [17], namely the "Metagenomic Transfer Format", have been focused on facilitating the exchange of quality controlled data between annotation pipelines (e.g. MG-RAST, IMG/M, CAMERA, etc.) saving hundreds of thousands of hours of computation at these facilities.

The data tide is rising, being driven by an increase in the number and size of biological data sets; however, the computational budgets available for biological research are not rising to match these changes. The efforts that are being developed are isolated and lack the interoperability needed by the wider community. A GSI would coordinate these isolated efforts to facilitate the interactivity that will enable the community to share higher-level data sets, e.g. ecosystem models instead of "raw" sequences. But for this to work, all participants need to trust the data products being transferred, have the ability to track provenance, and have access to the data analysis level they want to enable re-analysis. Through this mandate the GSI would create a community established on trust.

What might a GSI look like?

The fragmented, non-interoperable ensemble of tools and data sources available to the metagenomics community today is in large part similar to the state of technology that motivated today's commercial cloud services. These challenges are inevitable byproducts of rapid growth through entrepreneurial culture and constructive competition, and any solutions addressing the challenges must do so while retaining these cultural and procedural engines for progress.

Amazon Web Services is a relevant, worked example of similar challenges, with solutions that went

beyond retaining the progress and growth of a particular community, and led to acceleration and further expansion. Within Amazon, several hundred independent groups provide specific services that collectively populate the web pages used by customers (sales rank, recommendations, pricing, etc.). Each group is responsible for a specific service including the development and reliable operation of the underlying infrastructure. This includes developing and maintaining software, and originally it also meant purchasing, provisioning, and maintaining servers, along with related work to maintain operating systems, back up data, etc. Amazon determined that each group was spending 70% of their resources on generic infrastructure (identical in functionality to that provided by all service providers) and only 30% of their resources on the particular service they uniquely provided [18].

Certainly this had an impact on cost, but moreover it slowed progress to the point that continued growth would either become non-linearly more expensive, or would be simply unsustainable. Neither of these options was acceptable for Amazon, and this is in fact the current state of the metagenomics community.

One traditional approach to addressing such crippling redundancy and inefficiency would have been to work with each group to increase resources, deploying newer technologies and increasing staffing. Another might have been consolidating into fewer but larger groups to amortize common services, or perhaps centralizing a traditional set of infrastructure services (big database, supercomputer, sophisticated website, etc.), possibly outsourcing to a major IT company or group of outside "experts" to design a generalized solution.

Instead, Amazon engaged their community of service providers to determine what foundational standards and services could be shared and, of critical importance, what application programming interfaces and performance metrics would be required for adoption of those services to enable growth and innovation. According to Dr. Werner Vogels, Chief Technology Officer of Amazon Web Services, "Two key requirements in the design of these infrastructure services markedly changed the way resources are managed: the services are fully self-service, allowing engineers to start using them with minimal friction; and resources can be managed dynamically, giving engineers the power to acquire and release resources immediately." [18].

The microbiology community is in precisely the state that led Amazon to bring its service providers together to revolutionize their enterprise. There is recognition that the many independent efforts, groups, and services must change their approach in order to maintain progress, deal with rapidly increasing scale, and moreover to seize the new opportunities that cannot be achieved with current methods and tools.

As outlined earlier, current interactions within the community reveal two interdependent, high-level technical design processes that must be pursued in parallel with a governance and stewardship process (and that are discussed in the following section):

- Access and Scale. Defining the appropriate API and distributed storage framework for the data and models;
- Interoperation. Continuing the discussions on standardizing data and metadata representations (through existing and continuing efforts with the GSC and ISA [9,11].)

Scale in particular is of urgent importance for the metagenomics community. New advances in sequencing technology have led to rapid growth with the adoption of next generation sequencing hardware. These sequencers are capable of producing genome data in the range of a 0.5-1 trillion base pairs in a week, and are being employed by hundreds of biologists all over the world. The major challenges presented by the increasing scale—the size of data per sequencer, the number of sequencers, and the parallel nature of data generation and retrieval—are cross-cutting of the entire storage hardware and software. Lessons can be borrowed from other scientific communities with data challenges. The experimental physics community manages many Terabyte-scale data streams from a small number of large instruments, which are shared by many sub-communities, each of which collaboratively refines and interprets the data. On the other end of the spectrum is the sensor networking community, where experiments can involve thousands of individual sensors, each generating megabyte-scale data streams. On the scale of sources of data and the size of data, the data streams from hundreds to thousands of sequencing machines in the microbial community will straddle the space between the physics and sensor network efforts.

To complicate matters even more, these sequencers produce data in varying formats and resolution, and are often stored by the biologists in distinct data silos that often employ diverse mechanisms for storing and retrieving data. To maximize utilization of both the sequenced data and the sequencing hardware, the need for a radically different ecology of hardware and software is evident. Because of the rapidly expanding number of groups generating and analyzing data sets, searching for key pieces of information is expensive, particularly given that these data are distributed, stored, and retrieved in non-uniform ways. Some of these data silos are community-operated while individual groups who store data based on their requirements or based on some usage and age metrics maintain others.

To provide search and access across this expanding and fragmented corpus, frameworks and associated middleware must be designed that take these nuances into consideration. This begs for a new paradigm, including APIs and data and metadata format standards, including maintenance and curating provenance. These and other key design threads conceptually create a "data layer" that allows for integration and interoperation of data and tools, enabling the community to share common solutions to general issues such as scale, search, or data movement.

In contrast to the challenges faced by Amazon, today there are many architectural and technical options ranging from commercial cloud services to open source tools to various NSF-funded cyberinfrastructure frameworks. It will be particularly important to change some of the technology-driven design methodology that often results in unused, or underused, solutions. The design of the data sharing mechanisms, formats, and semantics have traditionally benefited from the experience of the computer systems and computational science community, but are often not adequately informed and are supported by scientists viewed as "customers" rather than "partners." There are certainly important lessons, technologies, tools, and architecture to be adopted as appropriate from other communities; and therefore this conceptualization effort will also pursue ideas and advice from other research domains including physics, astronomy and engineering, as well as from the private sector where many emerging companies have challenges similar to those illustrated earlier. The order of the day is to bring the community together to evaluate the current state, analyze and prioritize requirements, and evaluate architectures and approaches to creating a cyberinfrastructure for this community, such as could be subsequently implemented through a Genomics Software Institute.

Conclusions

Based on an extensive set of discussions that have already taken place among community leaders, there is general consensus that the task at hand for the community amounts to implementing three synchronized processes to conceptualize a GSI:

1. Access and Scale. Defining the appropriate API and distributed storage framework for the data and models;

2. Interoperability. Continuing the discussions on standardizing data and metadata representations (through existing and continuing efforts with the GSC and BioSharing);

3. Implementation and Sustainability. Establishing community-driven processes for governance, for the development and evolution of the software layer, and for fostering innovation in developing a diverse new generation of metagenomics scientists and educators.

The community both needs and is capable of creating a system to ameliorate issues of data volume, access, and interpretation. However, while the seeds of such a revolution already exist in many self-organized community initiatives, bringing these together will still take considerable coordination. Such coordination requires funding and appropriate infrastructure to enable inclusive interaction and to reduce the potential for alienation and the development of splinter groups. To create such an ecosystem that fosters creativity and interaction, while creating standardization is difficult, but by no means impossible. The time is right and action must be taken.

Acknowledgements

We would like to thank the board of the Genomic Standards Consortium for their valuable insight, especially Dr Renzo Kottman. This work was supported by the U.S. Dept. of Energy under Contract DE-AC02-06CH11357. The organizers gratefully acknowledge the support form the National Science Foundation grant RCN4GSC, grant DBI-0840989.

References

1. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* 1998; **95**:6578. PubMed http://dx.doi.org/10.1073/pnas.95.12.6578
2. Gilbert JA, Meyer F, Antonopoulos D, Balaji P, Brown CT, Brown CT, Desai N, Eisen JA, Evers D, Field D, *et al*. Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. *Stand Genomic Sci* 2010; **3**:243. PubMed http://dx.doi.org/10.4056/sigs.1433550
3. Delmont TO, Malandain C, Prestat E, Larose C, Monier JM, Simonet P, Vogel TM. Metagenomic mining for microbiologists. *ISME J* 2011; **5**:1837. PubMed http://dx.doi.org/10.1038/ismej.2011.61
4. Gilbert JA, Meyer F, Bailey MJ. The future of microbial metagenomics (or is ignorance bliss?). *ISME J* 2011; **5**:777. PubMed http://dx.doi.org/10.1038/ismej.2010.178
5. Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, Temperton B, Huse S, McHardy AC, Knight R, Joint I, *et al*. Defining seasonal marine microbial community dynamics. *ISME J* 2012; **6**:298-308. PubMed http://dx.doi.org/10.1038/ismej.2011.107
6. Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B, Weynberg K, Huse S, Hughes M, Joint I, *et al*. The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. *PLoS ONE* 2010; **5**:e15545. PubMed http://dx.doi.org/10.1371/journal.pone.0015545
7. Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM. Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol* 2011; **77**:1315. PubMed http://dx.doi.org/10.1128/AEM.01526-10
8. Hugenholtz P. Exploring prokaryotic diversity in the genomic era. *Genome Biol* **3.** *RE:view* 2002; **3**:S0003.
9. Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, Gilbert J, Glöckner FO, Hirschman L, Karsch-Mizrachi I, *et al*. The Genomic Standards Consortium. *PLoS Biol* 2011; **9**:e1001088. PubMed http://dx.doi.org/10.1371/journal.pbio.1001088
10. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, Field D, Harris S, Hide W, Hofmann O, *et al*. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* 2010; **26**:2354-2356. PubMed http://dx.doi.org/10.1093/bioinformatics/btq415
11. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, *et al*. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 2008; **9**:386. PubMed http://dx.doi.org/10.1186/1471-2105-9-386
12. Sansone S. Beyond Open Data: Commoning for the collection, curation, management and distribution of bioscience investigations. *Nat Genet* 2011; (In press).
13. Butte AJ. Challenges in bioinformatics: infrastructure, models and analytics. *Trends Biotechnol* 2001; **19**:159. PubMed http://dx.doi.org/10.1016/S0167-7799(01)01603-1
14. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, *et al*. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 2008; **9**:75. PubMed http://dx.doi.org/10.1186/1471-2164-9-75
15. Stein LD. The case for cloud computing in genome informatics. *Genome Biol* 2010; **11**:207. PubMed http://dx.doi.org/10.1186/gb-2010-11-5-207
16. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IM, Grechkin Y, Dubchak I, Anderson I, *et al*. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* 2007; **36**:D534. PubMed http://dx.doi.org/10.1093/nar/gkm869

17. Gilbert JA, Meyer F, Knight R, Field D, Kyrpides N, Yilmaz P, Wooley J. Meeting report: GSC M5 roundtable at the 13th International Society for Microbial Ecology meeting in Seattle, WA, USA August 22-27, 2010. *Stand Genomic Sci* 2010; **3**:235. PubMed http://dx.doi.org/10.4056/sigs.1333437

18. Vogels W. Beyond Server Consolidation. *Queue* 2008; **6**:20. http://dx.doi.org/10.1145/1348583.1348590

19. The Genomic Standards Consortium. http://gensc.org

20. Department of Energy's KnowledgeBase. http://www.systemsbiologyknowledgebase.org

21. Quantitative Insights into Microbial Ecology. QIIME – http://www.qiime.org

22. Open Microbiome Initiative. OMI – http://www.openmicrobiome.org

23. BioCode Commons. http://www.biocodecommons.org

24. BioSharing catalogue. http://www.biosharing.org