CellPress

Science & Society

Microbiome Data Science: Understanding Our Microbial Planet

Nikos C. Kyrpides,[1,*] Emiley A. Eloe-Fadrosh,[1] and Natalia N. Ivanova[1]

Microbiology is experiencing a revolution brought on by recent developments in sequencing technology. The unprecedented volume of microbiome data being generated poses significant challenges that are currently hindering progress in the field. Here, we outline the major bottlenecks and propose a vision to advance microbiome research as a data-driven science.

Bottlenecks in Microbiome Research

The vast increase in sequencing output during the last decade [1] has not been matched with analogous scaling and democratization in computational resources, either in the form of available computational capacity for data processing or data integration. Although unprocessed microbiome data are deposited in INSDC (International Nucleotide Sequence Database Collaboration) centers, there are currently no funded efforts to process and integrate all the microbiome data. This has resulted in the majority of microbiome sequences being 'single-use', that is, they have limited, if any, data reuse beyond the original scope of the study. This phenomenon has led to a certain 'compartmentalization' of microbiome studies whereby the data are stored in an *ad hoc* manner, and are often inaccessible to other scientists who want to reproduce the results of the study or mine the data for other applications. It also prevents systematic review and meta-analysis of the data using newly developed strategies and tools that have the potential to dramatically increase the power of individual studies and generate valuable insights. This approach has greatly hindered discovery by wasting significant resources and limiting interpretation of results by underutilizing the available sequence information. Further, this has resulted in a high degree of duplicated efforts since every analysis group has to partially redevelop data storage, integration, and analysis platforms.

In the light of the problems listed above, we have identified the following major bottlenecks (all of which are related to funding) currently impeding progress in microbiome research.

(a) Lack of a grand vision in supporting Microbiome Data Science. There is currently a sharp contrast between what is needed and what is available and/or financially supported in the field. The need for a national and unified international microbiome effort was proposed several years ago [2] and renewed interest is beginning to gain broader community support [3,4]. However, it is still unclear whether these calls will gain traction to drive the effort from a conceptual idea to realization. Even with today's small-scale data (relative to what is expected 10 years from now), there is a profound lack of a grand vision in appropriate funding to support the extraction of knowledge from big data (i.e., across studies). Most microbiome projects currently have their data analyzed in the context of their own study and largely do not incorporate datasets from other publicly available studies. Current research efforts work well for small- to medium-scale projects, but fail to support and promote larger endeavors at global multifaceted analysis that may require processing and integration of all relevant publicly available data.

Furthermore, the reference data needed to contextualize the myriad microbiome samples is sorely lacking. A prime example is the fact that less than 20% of the bacterial and archaeal type strains have been sequenced, despite evidence for the scientific value of generating these basic data [5,6]. While recent trends emphasize hypothesis-driven science and a shift away from exploratory sequencing, we argue that part of a grand vision for microbiome data science necessitates the continued generation of reference data. These data are fundamental for interpretation of how microbiomes function in a community context, and how they interact within the environments and hosts they inhabit. Systematic decoding of microbes and their environments to fill in the gaps in our databases is a key step towards hypothesis-driven science and enabling a better understanding of microbial life.

(b) Inefficient funding mechanism. The increasingly interdisciplinary approach to biology has enabled us to reach the point where scientific progress can be hindered by the insulation of individual funding agencies. This is especially evident in the segregation of funding from individual agencies supporting big data integration and analysis, for example, the EarthCube (http://earthcube.org/) initiative by the National Science Foundation (NSF), the Human Microbiome initiative (https://commonfund.nih.gov/hmp/index) by the National Institutes of Health (NIH), and several other initiatives by the Department of Energy (DOE). Rather than joining forces to create interagency funding models to face the grand challenges of big data ahead (following up on existing recommendations from the scientific community) [7], agencies each support separate smaller-scale efforts. Furthermore, support for big data integration and analysis requires long-term commitment, which is required for microbiome research but has been nearly impossible to obtain due to the limited funding period for databases responsible for large data integration and analysis.

(c) Insufficient data standards and interoperability. Although international consortia for the establishment and propagation of standards have already formed [8], they are either limited in scope [9] or their adoption lacks the appropriate mandate from funding agencies and publishers alike. As a result, the lack of standards for all the steps from preparing the samples to the end point of processing and comparing data is currently impeding the community's ability to perform efficient comparative analysis.

Vision to Advance Microbiome Research: Enabling Data Science

Key to moving forward in the face of these bottlenecks is a vision for transforming the deluge of data from a problem to a solution, by enabling the research community to utilize and explore the data produced worldwide. To achieve this, it is imperative to develop a long-term strategy that will support the anticipated data growth, and that will ensure that the data revolution will not become disruptive for the field through the 'balkanization' (i.e., fragmentation) of microbiome data generation and analysis, as is currently the case. The development of this strategy requires a major cultural and conceptual transformation whereby the generation of vast amounts of biological data is no longer considered the goal or the end result of funded studies, but rather, the most important tool needed in order to efficiently address fundamental biological questions critical to human health, biotechnology, energy, food, and environmental sustainability. Analogous to the telescope for astronomy and the particle accelerator for high-energy physics, biological sequence data should be considered an instrumental tool for the study of biological systems. Tools like the Hubble telescope or CERN's particle accelerator required several years for construction, multibillion dollar funding efforts, and very large and distributed research networks. Funding or development of data-science-related tools of that scale are currently not available for microbiome research, even though it is well appreciated that we live on a microbial planet [10] and that attempts to understand biological phenomena based on incomplete data only lead to erroneous conclusions [11].

Establishing a Distributed National Microbiome Data Center

Although biology is rapidly moving towards a holistic view of life, we are witnessing an increase in funding awards that are regressing biology towards individual, partially redundant, and largely disconnected efforts, instead of preparing it to fulfill its destiny as a quantifiable science akin to physics. To this end, there is great demand for creating a distributed national microbiome data center that would organize, process, and serve all available environmental genomic data. Significant improvements in computational methods for data processing and high performance in distributed data-management systems, coupled with the ability to utilize high-performance computing (HPC), are now rendering such an endeavor possible. The key objective of this center would be to develop and maintain a state-of-the-art data-management system integrating all available environmental genomic data. This would enable efficient handling and processing (i.e., assembly and annotation) of all publicly available primary microbiome data and metadata generated around the globe for downstream interpretation and discovery. In this respect this facility should also serve as an international microbiome data center. The envisioned data center would support both grand vision projects as well as smaller studies by providing the ability to conduct effective comparative analysis in an integrated context. Efficient data handling and interpretation rests on three major pillars, all of which are profoundly interconnected and interdependent:

(a) Comparative analysis. This represents the hallmark of data interpretation. It is well known that the single most important tool for interpreting genomic and metagenomic sequences is their analysis on a comparative level. The larger the size of the dataset used in the comparison, discovery becomes more likely and more accurate. In principle, we do not know upfront which data are the most relevant to compare, and because life has no divisions or boundaries along the lines of the funding agencies or application areas, comparative analysis should not be restricted to individual organisms, individual environments, or funding focus scope.

(b) Data integration. The success of the comparative analysis is directly dependent on the efficiency of data integration, which, in turn, depends on the breadth of the integration, the quality of the integrated data, and the underlying structure of the integration. Each of these parameters is critical for building successful data-integration platforms. The breadth of the integration here refers not only to the number and diversity of the integrated datasets, but also to the various types of 'omics' data as they become available. As the microbiome field moves towards more holistic approaches, and the emerging technologies enable exploration of whole systems (e.g., human or plant microbiome), it becomes essential to integrate a wide array of data types across all domains of life. The quality of the integration directly depends on the quality of the data, as reflected by the level of data contamination, coherence of annotations, availability of metadata, and the overall level of detail in identifying accuracy and completeness of the integrated data. Finally, the underlying structure should not only enable integration of a wide range of interdisciplinary data, it should also support vigorous data visualization and sustain an unprecedented growth in data.

(c) Data standards. Standardizing the description and processing methods of microbiome data is critical for comparisons across different samples and studies that have adopted incompatible recommendations from different

international bodies promoting standards in microbiome research.

A number of large data-management systems are currently available for supporting the comparative analysis of assembled [12] or unassembled [13] microbiome data and their associated metadata [14], as well as systems designed for predictive modeling (https://kbase.us/) and cyberinfrastructures [15]. Similar successful systems with existing and dedicated long-term funding should be an integral part of such a distributed national microbiome data center.

Concluding Remarks

Future endeavors in microbiome research are expected to lead us to a new age of holistic understanding of microbial life, develop novel therapeutic strategies to treat infectious diseases, identify solutions for protecting the environment, and ultimately understand and harness the power of the most abundant natural resources on our planet. To achieve these endeavors and enable the vision described above, the research community requires a major restructuring in the current research-funding policies through the development of innovative funding mechanisms that will provide long-term support for microbiome data science. Examples of such mechanisms can be drawn from existing models such as the Brain Initiative (https://www.whitehouse.gov/share/brain-initiative), a grand challenge research effort to revolutionize our understanding of the human brain. At the dawn of the third decade of microbial genomics, and well into the information age, the time is ripe to embark on the greatest endeavor to understand Earth's microbiome. Microbiome data science, through the establishment of a national microbiome data center, can pave the way.

Acknowledgments

We thank Victor Markowitz, Torben Nielsen, and Heather Maughan for critical reading and suggestions on the manuscript. This work was conducted by the U. S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, under Contract No. DE-AC02-05CH11231.

[1]Prokaryotic Super Program, Department of Energy Joint Genome Institute, Walnut Creek, CA, USA

*Correspondence: nckyrpides@lbl.gov (N.C. Kyrpides).

http://dx.doi.org/10.1016/j.tim.2016.02.011

References

1. Koboldt, D.C. *et al.* (2013) The next-generation sequencing revolution and its impact on genomics. *Cell* 155, 27–38
2. Kyrpides, N.C. (2009) Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nat. Biotechnol.* 27, 627–632
3. Alivisatos, A.P. *et al.* (2015) A unified initiative to harness Earth's microbiomes. *Science* 350, 507–508
4. Dubilier, N. *et al.* (2015) Microbiology: Create a global microbiome effort. *Nature* 526, 631–634
5. Wu, D. *et al.* (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462, 1056–1060
6. Kyrpides, N.C. *et al.* (2014) Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS Biol.* 12, e1001920
7. Gilbert, J.A. *et al.* (2014) Meeting report: Ocean 'omics science, technology and cyberinfrastructure: current challenges and future requirements (August 20–23, 2013). *Stand. Genomic Sci.* 9, 1252–1258
8. Field, D. *et al.* (2011) The Genomic Standards Consortium. *PLoS Biol.* 9, e1001088
9. Field, D. *et al.* (2011) Genomic standards consortium projects. *Stand. Genomic Sci.* 9, 514–526
10. National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications (2007) *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet,* National Academies Press (US)
11. Ioannidis, J.P.A. (2005) Why most published research findings are false. *PLoS Med.* 2, 696–701
12. Markowitz, V.M. *et al.* (2014) IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res.* 42, D568–D573
13. Wilke, A. *et al.* (2013) A metagenomics portal for a democratized sequencing world. *Meth. Enzymol.* 531, 487–523
14. Reddy, T.B. *et al.* (2015) The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.* 43, D1099–D1106
15. Goff, S.A. *et al.* (2011) The iPlant Collaborative: cyberinfrastructure for plant biology. *Front. Plant Sci.* 2, 34

Spotlight

Engineering Coronaviruses to Evaluate Emergence and Pathogenic Potential

Susanna K.P. Lau[1,2,3,4,5,*] and Patrick C.Y. Woo[1,2,3,4,5,*]

A recent study provides a platform for generating infectious coronavirus genomes using sequence data, examining their capabilities of replicating in human cells and causing diseases in animal models, and evaluating therapeutics and vaccines. Similar approaches could be used to assess the potential of human emergence and pathogenicity for other viruses.

The severe acute respiratory syndrome (SARS) epidemic in 2003 and the Middle East respiratory syndrome (MERS) epidemic in the last 3 years have shown that coronaviruses (CoVs) have the capability to cause major epidemics. For the SARS epidemic, a total of >8000 laboratory-confirmed cases with >800 deaths were observed (http://www.cdc.gov/sars/about/fs-sars.html). This horrific epidemic was followed by the publication of >7500 scientific papers on CoVs visible in PubMed, which represents two-thirds of the total number of publications on CoVs in Pubmed. Despite the numerous studies on CoVs, it is still difficult to predict which CoV may have the potential to emerge as the next culprit. A recent study in *PNAS* by Menachery *et al.* [1] and another similar study in *Nature Medicine* published in December 2015 by the same group [2] reported the use of existing sequence data with reverse genetics to engineer SARS-related CoVs and evaluate their potential of emergence and pathogenicity.

Shortly after the emergence of SARS-CoV, SARS-related CoVs were found in civets [3]. However, multiple lines of evidence showed that the civets are just the intermediate or amplification hosts for SARS-CoV. Through intensive surveillance studies in various mammals in Hong Kong, Lau *et al.* reported the presence of SARS-related CoVs in Chinese horseshoe bats in Hong Kong [4]. A similar observation was also reported by another group in mainland China [5]. Since then, numerous SARS-related CoV sequences were observed in different