

Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream

Nikos C Kyrpides

As we approach the completed sequencing of 1,000 microbial genomes, the field of microbial genomics is poised at a crossroads. The future holds great promise for far-reaching advancements in microbiology as well as in diverse, related sciences. But realizing that potential will require meeting the challenges that have accompanied the rapid development of the underlying technology and the exponential growth of data. New technologies provide unprecedented opportunities but also call for conceptual shifts. Experience gained in the first decade of genomics can guide the improved approaches now needed for the selection of genome sequencing projects and their funding, for genome publication and annotation, as well as for data analysis and access. Equipped with these new tools and policies, microbiologists will have a unique opportunity for unprecedented exploration of our microbial planet.

The dramatic advancements in sequencing technology achieved during the past decade have mediated a rapid transition from single-gene to whole-genome studies. In so doing, they also transformed what had been an almost purely experimental discipline into a predominantly theoretical and predictive one¹. Although not the driving force *per se* behind the development of the technology, microbial genomics was in the forefront of this transition from the very beginning and paved the 'genomics way'². Indicative of this leading role, more than two-thirds of the 4,800 currently reported genome projects³ are microbial (<http://genomesonline.org/>), and the same percentage is observed for microbial proteins in the public archive sequence databases from genome sequencing projects⁴.

During its first decade, the newly defined field of genomics adopted the fundamentally reductionist perspective that marked twentieth century biology⁵. Accordingly, genome projects were initiated almost exclusively on the basis of potential practical applications for the selected organism, often in the fields of medicine (e.g., pathogenicity or drug targets) or biotech (e.g., bioenergy, agriculture, environmental remediation or industrial production of microbial products). Indeed, just as the human genome project originally set the tone for all genomics, direct practical exploitation of microorganisms set the stage for the microbial sequencing program.

When the boundaries of science are shattered by the sheer force of technological innovation in the absence of a guiding vision, there usually follows a period of time before the affected scientific community

realizes that something has gone awry. Such realizations are now arising. Here I present some of the underlying problems and myths that I believe substantially hinder additional growth of the field and, even more importantly, compromise the ability of biologists to use and interpret the available data. Where possible, a solution is proffered, often one whose implementation will necessitate action by the entire community, including scientific journals, sequencing centers and the funding agencies.

Genome publication and data release policy

The policy for publication of complete genomes that we witnessed for most of the first genomics decade has violated a longstanding precept in all scientific fields. Namely, the vast majority of genome papers have been submitted and accepted for publication long before the public release of the sequenced data. As a result, reviewers, unable to examine the actual data, could evaluate such papers only on faith and trust, thus undermining the peer-review process in this field. Although no genome paper has been retracted, subsequent analyses have often revealed fundamental flaws in both the derived conclusions and the data itself^{6,7}.

Granted, annotation is generally considered a never-ending bioinformatics adventure; closure is achieved only when all the functions and all the genes of an organism have been experimentally verified. Still, a fine degree of separation exists between what might be considered an acceptable error due to the incomplete adventure⁸ and what is essentially an incorrect and misleading conclusion. Charging the reviewers with the responsibility for making this distinction calls for a strict policy from the publishing journals requiring the actual sequence data to be publicly released well in advance of publication of the genome. Moreover, simply providing the sequence files would not suffice, as most scientists cannot make much sense out of a GenBank file. Rather, the data should be provided to the community in a meaningful way that facilitates cogent analysis and evaluation. This can be accomplished only through data management systems that support comparative genome analysis. Several such systems are already freely available in the community⁹.

Over time, as the substantial benefits of prepublication release of genome data have been recognized, many funding agencies and most of the large sequencing centers now adhere to the rapid data release policy set forth as the Bermuda Principles in 1996 and renewed in 2003 (<http://www.genome.gov/page.cfm?pageID=10506376>). Thus, over the past few years, we have witnessed an increasing number of complete genomes released in GenBank without accompanying publications¹⁰.

Owing to the exponential increase in the number of completed genome sequences, coupled with frequent phylogenetic redundancy of

Genome Biology Program, DOE Joint Genome Institute, Walnut Creek, California, USA. Correspondence should be addressed to N.C.K. (nckyrpides@lbl.gov).

Published online 8 July 2009; doi:10.1038/nbt.1552

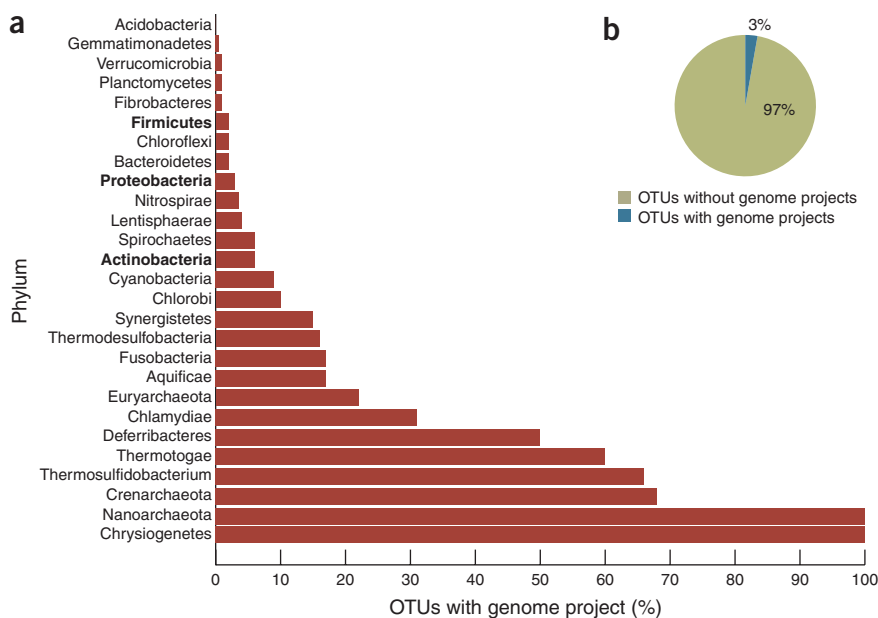


Figure 1 Phylogenetic distribution of sequencing projects. All available 16S RNAs sequenced were assigned to Operational Taxonomic Units (OTUs) that were then classified into phyla. (a) Percentage of OTUs with genome projects for each phylum. (b) Percentage of OTUs with genome projects. OTU data are from the GreengenesDatabase (<http://greengenes.lbl.gov/>); genome project data are from the GenomesOnLine database (<http://genomesonline.org/>).

this diversity were funded by both the National Science Foundation (<http://www.nsf.gov/od/lpa/news/02/pr0294.htm>) and the Department of Energy (DOE); <http://genomebiology.jgi-psf.org/Content/Archaea.htm>.

Nevertheless, the first systematic attempt to coordinate efficient and balanced sampling of the phylogenetic tree was only recently launched through the DOE-funded Genomic Encyclopedia of Bacteria and Archaea (<http://www.jgi.doe.gov/programs/GEBA/index.html>).

The goal of this effort is to sequence representatives for every phylogenetic node of the microbial tree of life. The results are anticipated to fill in our incomplete view of the available microbial functional repertoire and, even more importantly, to spearhead our attempt to understand the complexity of entire microbial communities.

Maintenance of biological databases

Arguably, one of the forces driving innovation and advancement in twenty-first century biology is the close interaction with other scientific fields. However, the current interdisciplinary synthesis has posed one the most daunting challenges biologists face today. Specifically, funding in computational biology and bioinformatics (and by extension other closely related fields of biology) has been heavily weighted toward computer scientists and engineers who, as expected, have proceeded to transform the field into engineering disciplines. During the past decade, funding agencies and foundations have been promoting and supporting the development of more new biological databases, while simultaneously withdrawing support for existing ones. In consequence, an unprecedented number of new databases have been created¹⁴, many of which were abandoned when the initial funding period expired (usually 3–5 years). Continuing this same funding policy will likely lead to the same outcome in the future.

Regrettably, the problem seems to be rooted in two deficiencies within the funding agencies and the programs that support the launching of these projects: first, lack of a biologically grounded guiding vision; and second, failure to recognize that unlike typical biology research projects, biological databases require ongoing support beyond the usual 3–5 year period.

Clearly, any biological database requires ongoing data curation. Current funding policies pose a fundamental barrier to progress in biology, a field that profoundly depends on the conscientious curation of biological data. Failure to realize the significance of this soon enough will likely result in an increasing amount of wasted public funds and will transform one of the most vibrant and promising landscapes of modern biology into a cemetery of abandoned projects and lost opportunities. It is therefore imperative to create innovative funding mechanisms and policies that will provide balanced support for both the maintenance of key biological databases (through data curation) and the continued creation of new ones.

the genomes selected⁴, individual genome studies have become increasingly difficult to publish and, in several cases, have been rendered superfluous. Yet, the complete sequencing of an organism requires a tremendous technical effort across several coordinated teams and constitutes the landmark availability of its complete genetic code. As such, it warrants a citable publication of its own. A new scientific journal named *SIGS* (Standards in Genomic Sciences) is about to be launched to provide a home for this type of genome report (<http://standardsingenomics.org/>)¹¹. Notably, this will also facilitate a clear differentiation between a genome report accompanying the data release and a publication providing genome interpretation and insights into the biology of the organism.

Genome project selection bias

The first decade of genome sequencing programs provided us with more than 1,000 microbial genome sequences, including both complete and draft sequences. Such a collection could be expected to provide a reasonably good sampling of the functional and evolutionary diversity of microbial life, especially as fewer than 2,000 bacterial and archaeal genera have been characterized (<http://www.bacterio.cict.fr/number.html#total>) and fewer than 8,000 unique organisms have been validly named (<http://www.namesforlife.com/>). Yet, the actuality falls far short of that. The universal tree of life¹¹, which provided us with a powerful framework for understanding the relationships among all organisms and made us aware that the preponderance of life's diversity lies within the microbial world, now suggests the biased selective sampling to date.

Of the 3,000 reported bacterial genome projects, as many as 82% focus on just three major phylogenetic lineages (Proteobacteria, Firmicutes and Actinobacteria), and within these lineages, perhaps 10% of their currently estimated species have been sequenced (Fig. 1)^{4,12,13}. Furthermore, of the bacterial genome projects currently known to be in progress, only 30% break new ground by sequencing an organism from a different genus. Granted, comprehensive reconstruction of the biology of an organism from its genome sequence calls for efficient sampling of its close phylogenetic neighborhood. But environmental genomics has provided ample evidence that the extant functional and evolutionary diversity is vast, and mostly still waiting to be explored¹². Earlier small-scale efforts aimed at exploring

One way to move rapidly in this direction would be to recognize—and appropriately fund—the two fundamental and distinct phases in the evolution of any biological database. A first funding step would support the launching of new biological databases. The current funding mechanisms and typical duration of 3–5 years would be appropriate for this phase. During this period, it would be the responsibility of those developing the database to build their user community and make a strong case for their warranting long-term support. The second funding step calls for the development of a new program that would provide renewable, long-term support, provided that certain criteria for service to larger user communities are met. This new program would ideally represent a joint effort across several funding agencies, a reflection of the interdisciplinary and cross-agency nature of several of the large-scale integration databases.

Community annotations

One of the underlying myths concerning biological databases, and arguably one of the main assumptions leading to the situation just described, is the notion that once a database is designed, the data will somehow miraculously be curated. Much has been written about ‘community annotation’ as a promising mechanism of database curation, but no evidence suggests that it will ever happen. Millions of dollars have been invested in the development of large databases that facilitate comparative analysis and annotation of genomes⁸. One would expect these to be the natural places where the community would gather to contribute to the available knowledge. Ironically, it is the far simpler, wiki-type web applications that hold the best hope for community input^{15,16}.

However, the real question here is, Why would anyone expect—or even worse, depend on—a community annotation effort? Imagine investing millions of dollars into state-of-the-art sequencing facilities, and then expecting volunteers from the community to stop by and run the sequencing machines. One might argue that this analogy is not valid because running a sequencing facility requires well-trained personnel, standardized protocols, clear procedures, quality controls and, most of all, tight coordination. Yet, the same professional standards are required for data curation, and it is precisely these aspects that are rarely achieved through a community contribution approach. Community annotation should be encouraged and facilitated, but the curation of biological data cannot depend solely on volunteer work. High standards and quality implies professionalism, and this, in turn, requires investing in dedicated professionals. Until this is done, data curation—and consequently the whole field of microbial genomics—will not move beyond the amateur stage.

Lack of standards in genomics

Like molecular biology, genomics has been fueled by the innovative energy of many interdisciplinary activities. Unlike molecular biology, which has thrived on the principle of standardized methods and protocols, genomics has progressed without regard for the critical importance of shared standards. Now, 14 years since the first complete genome was published¹⁷ and with more than 900 genome sequences finished, it is astonishing to observe the lack of standards for so many critical procedures in the field, ranging from simple data exchange to gene finding, function prediction and metabolic pathway description.

For example, several ontology schemas and control vocabularies for the representation of gene function exist. Even so, there is still no consensus in the community regarding their use, resulting in multiple terms denoting essentially the same function. There is not even a consensus term for proteins of unknown function, with instead an extremely long list of permutations of names referring

to the same thing. As a consequence, we currently find over 400,000 distinct protein ‘functions’ assigned to the genes predicted from the genome projects, according to the product names available through the Integrated Microbial Genomes system (<http://img.jgi.doe.gov/cgi-bin/pub/main.cgi?section=ImgStatsOverview>)¹⁸, a number that is estimated to be at least one order of magnitude larger than the number of currently estimated unique functions.

The heart of the problem may lie in the lack of widely accepted and/or adopted ontology standards for gene function (particularly in the case of microbial genomes). Nevertheless, it has been substantially augmented and perpetuated by each sequencing center or annotation group employing a different standard operating procedure (SOP) for genome annotation. This dearth of cooperation among the various annotation groups has created a tremendous amount of apparent genomic variability in the public databases that is, in actuality, computational rather than genetic. The complete lack of a generally accepted definition of a gene or a pseudogene (even for the ‘simpler’ microbial genomes) poses an even greater problem, and is perhaps the chief obstacle limiting the accuracy and overall quality of comparative genome analyses.

As an example, we compared the genomes of two closely related organisms, *Burkholderia mallei* ATCC 23344 (ref. 19) and *Burkholderia pseudomallei* K96243 (ref. 20), sequenced in 2004 by The Institute for Genomic Research (Rockville, MD, USA) and the Wellcome Trust Sanger Institute (Hinxton, UK), respectively. Although the organisms are closely related, *B. mallei* is an obligate parasite of horses, whereas *B. pseudomallei* is free living. As expected, the free-living organism has a substantially larger genome size (7.2 Mb versus 5.8 Mb). We then looked for genes present in the smaller genome of the parasite that are not present in the larger. Using the Integrated Microbial Genomes system¹⁸, we identified 548 genes in *B. mallei* that are absent from *B. pseudomallei* and are potentially related to their different lifestyles. Manual curation of those 548 genes revealed that, in fact, 497 of them are also in the *B. pseudomallei* genome, but there they had not been identified as ‘real’ genes. The reason for this discrepancy? The two sequencing centers used different gene finding methods. The consequence was an almost 90% error rate in the results of our comparison.

Similar inaccuracies are observed whenever one compares the predicted genes from closely related genomes that have been annotated by any combination of different analysis groups. Obviously, this creates a huge barrier preventing efficient and accurate automated comparisons between publicly available genomes. One would expect this situation to be gradually improving as the different annotation pipelines come into closer agreement. Unfortunately, this is not the case. We are still witnessing a discrepancy rate of up to 20% in gene prediction when different annotation pipelines are applied to the same genome sequence (<http://genepriimp.jgi-psf.org/>).

To address the general lack of standards, the Genomics Standards Consortium²¹, an international scientific consortium, was formed in 2005. Its goal is to promote mechanisms that standardize the description of genomes and thus facilitate the exchange and integration of genomic data. Although this effort is the first well-organized attempt to achieve this goal, and in some respects is the community’s best hope for eventually obtaining a wider implementation of standards, it is not the only one. More recently, the National Institutes of Health (NIH) launched a major undertaking to sequence and characterize the human microbiome (<http://nihroadmap.nih.gov/hmp/>). This project is unique in microbial genomics for the unprecedented volume of sequence data that will be generated but more importantly for the distribution of the work across several large-scale sequencing facilities (that is, the J. Craig Venter

Institute (Rockville, MD, USA), Washington University (St. Louis), Baylor College of Medicine (Austin, TX, USA) and the Broad Institute (Cambridge, MA, USA)). By organizing the project in this manner—reminiscent of that used for the human genome effort—NIH has created a unique opportunity for synergy and collaboration between some of the world's leading sequencing and analysis centers, thus in effect mandating the standardization of their sequencing, finishing and analysis pipelines. This stands as a milestone marking the first systematic attempt by any funding agency to promote the standardization of data and methods through collaboration.

The second decade of microbial genomics

Genomics, now in its second decade, can move away from the extreme reductionism of studying single organisms alone. Genetic diversity is so vast that we cannot even begin to decipher life if we restrict our studies to only model organisms²². Furthermore, even model organisms cannot be fully understood in isolation from their environment. The advent of environmental genomics, or metagenomics²³, is providing the methods to explore these more complex communities²⁴. One of the most profound realizations derived from these studies is that we live on a microbial planet²⁵. Understanding the microbial world—whether that of our external environment or in our own body—is arguably a far more complex and challenging endeavor than the sequencing and understanding of the human genome. This was best said by Carl Woese, the microbiologist who defined Archae as a phylogenetic kingdom: “Genome sequencing has come of age, and genomics will become central to microbiology’s future. It may appear at the moment that the human genome is the main focus and primary goal of genome sequencing, but do not be deceived. The real justification in the long run is microbial genomics”²⁶.

In the remainder of this article, I argue why microbial genomics, empowered by new technologies and metagenomics, is poised to become the driving force in genomics.

Economies of scale

It is worth noting that genomics technology used today is fundamentally different from that which birthed the field and brought it to this point. The increase in the speed and efficiency of the sequencing technology

over the last decade has been accompanied by more than a 90% reduction in the cost. Thus, it would now take but a small fraction of the time (and an even smaller fraction of the funds) to repeat all of the work done to date.

Looking ahead, one would conservatively estimate that the next decade will bring at least 10,000 more complete genomes, with tens of thousands more in various stages of completion, thus providing us with hundreds of millions of new genes. This expectation poses totally new challenges for the development of commensurate data handling procedures. We are rapidly approaching the point of having more data than can be analyzed. We will need intuitive interfaces and databases that can scale to this level and still provide simple navigation for efficient data mining, and we will need computational methods for large-scale comparative analysis. Likewise, computational methods for comparative analysis on this grand scale must be implemented.

The power of bioinformatics rests upon the strength of comparative analysis, which in turn depends on the availability of data. However, with the technological advances in sequencing, we are rapidly approaching the point of having more data than can be analyzed. The current bottleneck has been created by a combination of technological limitations and conceptual barriers.

The most widely used technology for identifying homologous gene families is the BLAST²⁷ all-versus-all comparison. Because this approach scales n^2 (where n is the number of genes), it is obvious that even with a linear rate of increase in data, this approach will soon become unusable. High-performance computing and parallel implementation applications, such as ScalaBLAST²⁸ may help in the near future. Even so, it is a mathematical certainty that the all-versus-all approach cannot scale to match the anticipated exponential increase in data. Although rapid technological advances have transformed sequencing into a commodity easily affordable by the average university or research institute, the ability to analyze the sequence data will become increasingly expensive, soaring out of reach of most institutions.

Changing of the guard: from genomes to pangenomes

The most promising approach for alleviating the data analysis bottleneck involves a conceptual change—namely, the realization that effective comparative analysis need not compare all genes with all other genes. As not all genes have a sequence similarity to all others, methods for limiting BLAST analyses to groups with detectable similarity could, in principle, considerably reduce the computational demand of comparative analysis. The development of such methods is already underway in several large sequencing and analysis centers.

One such approach takes advantage of the high sequence identity shared by many of the >5 million microbial genes presently in public genome databases. One could, for example, create protein families from genes that share >80% identity, thus reducing the overall data size by at least one-third. Including only a single representative from each family in all-versus-all comparisons would reduce the number of comparisons to about 50%.

The same principle could be applied to the increasing number of microbial genome projects. Of the 1,155 bacterial genomes listed in the Integrated Microbial Genomes database v2.7 (ref. 18), only 691 belong to separate species and just 339 to distinct genera. Accordingly, a method for collapsing all the sequenced strains of a species to a single representative genome for that species, or all the sequenced species of a genus to a representative genome for that genus, would reduce the number of genome projects by 40% and 70%, respectively.

Ideally, this data reduction methodology would be based on the concept of the ‘pangenome’²⁹, defined as all of the different genes present in a set of genomes. At the species level, the pangenome constructed from all sequenced strains of a species encompasses the sum of the genetic repertoire found in any of those strains. Genes present in more than one strain are counted only once, and only one representative sequence of the gene is included in the pangenome. Thus, the pangenome of a species consists of the core genome found in all the isolates plus the ‘flexible’ genes that are present in some but not all. Several case studies have revealed that pangenomes of different species differ with respect to the relative proportion of core and flexible genes. Those with a high percentage of core genes are called ‘closed’ pangenomes, those with a high percentage of flexible genes are termed ‘open’^{29,30}. The degree of ‘openness’ of the pangenome generated from those strains can reveal the evolutionary dynamics of that species and indicate how many additional strains may need to be sequenced to adequately characterize the organism. The large-scale application of this approach to all sequenced genomes, combined with the shift from single genome-based to pangenome-based systems for comparative analysis, may fundamentally alter the way we view and analyze an organism’s physiology and coding potential.

Eventually, this approach may lead to a new understanding of our microbial planet, fulfilling microbiology’s dream of the systematic study and

comparative analysis of microorganisms but now redefined as dynamic communities that may be computationally represented as pangenomes. Looking back at the breakthroughs that have brought genomics to where it stands today, we find that in 1960–1990, the era of ribosomal RNA, we were building the tree of life and establishing the framework for the genomics revolution of 1990–2010, when we were growing the tree of life. The next decade (2010–2020) will be marked as the era of pangenomics, defined as finally understanding the tree of life.

New technologies, new ways forward

The greatest challenge to increasing our genomic coverage of microbial diversity lies in obtaining the DNA to sequence. More than 99% of the currently known microbial diversity resides in unculturable organisms. Of those that can be cultured, many are difficult to grow or grow only very slowly. Some present hindrances to DNA extraction. Growing the organisms for even a hundred sequencing projects consumes huge resources and requires much infrastructure. Most importantly, unlike DNA sequencing and data analysis, provisioning of DNA does not seem to be scaling up to expedite the process.

Community metagenomics cannot fill this gap, as discrete genomes cannot be assembled from the metagenomic data obtained from most environments. Therefore, our best hope for the future may lay in a new direction: single-cell genomics³¹. Already, current technology can provide ~70% coverage of a microbial genome by sequencing the DNA from an individual microbial cell³¹. It has been predicted that coverage will increase to ~95% within the next 3–5 years, owing to intense technology development. Even at the current coverage, this approach constitutes a major breakthrough that has opened a window into vast, previously inaccessible realms of unculturable microbial diversity.

Community metagenomics can be partnered with single-cell genomics, an approach that will likely become common for metagenomic projects. In parallel with sampling and sequencing the metagenome for an environment of medium complexity, single-cell techniques can be used to sequence several of the individual cell types present. Even at the current 70% coverage, this would provide representative reference genomes for that environment and lead to a more holistic understanding of the community and its individual members.

For those culturable organisms for which complete genome sequences can already be obtained, greater insights will emerge from bridging the gap between genotype and phenotype as expected from the integration of transcriptomics and proteomics with genomics. For the most part, genes in sequenced microbial genomes are computationally predicted based on the location of start and stop codons within the sequence. Thus, gene prediction is essentially protein prediction, and there is little known about the transcribed but untranslated regions (UTRs) at either end. Coordinating a genome with its companion transcriptome and proteome can provide experimental confirmation of the accuracy of those predictions and can reveal genes missed by computational approaches³². Transcriptomes can extend known protein-coding sequences to include the UTRs, thus identifying the locations where transcription starts and stops. Overall, the advent of new sequencing technologies is opening entire new worlds of possibilities in microbial genomics, ranging from the identification of novel small regulatory RNAs³³ to elucidation of the mechanisms underlying the generation of genetic diversity. Indeed, as sequencing technology becomes cheaper, faster and more accurate, resequencing, and by effect, studies on the origins of mutations and population variability, are finally within our reach.

National and international initiatives: a MEGA approach

Although one of the greatest challenges ahead lies in managing the current exponential growth in sequence data, it is ironic that the

Table 1 Estimating the magnitude of microbial diversity

Number of bacteriophages on Earth	10^{31}
Number of microbes on Earth	5×10^{30}
Number of stars in the universe	7×10^{21}
Number of microbes in all humans	6×10^{23}
Number of humans	6×10^9
Number of microbial cells in one human gut	10^{14}
Number of human cells in one human	10^{13}
Number of microbial genes in one human gut	3×10^6
Number of genes in the human genome	2.5×10^4
Combined length of all bacteriophages on Earth	10^8 Ly
Diameter of the Milky Way	10^5 Ly

primary factor limiting the understanding of our microbial planet is, in fact, the need for even larger quantities of data. In a remarkable achievement, the Sorcerer II Global Ocean Sampling Expedition³⁴ sequenced over six million microbial genes, almost doubling the size of GenBank at the time. However, viewed from the perspective of the actual extent of microbial diversity³⁵, such efforts are, and will remain, extremely small scale. The remarkable number of microbes (Table 1)—already estimated to be several orders of magnitude greater than the number of stars in the universe—urgently calls for a transition from random, anecdotal and small-scale surveys toward a systematic and comprehensive exploration of our planet.

This cannot be achieved by the efforts of individual researchers but requires the establishment of effective national and international collaborations. For comparison, space and planetary exploration could never have been realized by a single researcher or even a small network. To achieve those goals, a National Aeronautics and Space Administration (NASA; Houston, TX, USA) was formed in the United States, with similar national efforts introduced in several other countries. The success of NASA can serve as a model here.

It is imperative to see the formation of national Microbial Environmental Genomics Administrations (MEGA) launched around the globe. Current ongoing international efforts include the International Census for Marine Microbes (ICoMM) (<http://www.coml.org/descrip/icommm.htm>) and the International Soil Metagenome Sequencing Project, or so-called ‘Terragenome’ (<http://terrigenome.org/>). National initiatives include the Australian Genome Alliance (<http://www.genomealliance.org.au/>) and the MikroBioKosmos initiative in Greece (<http://www.mikrobiokosmos.org/>).

Clearly, efforts of this magnitude require substantial investment. To explore and seek to understand how the Earth breathes, grows, evolves, renews and sustains life—all essentially the work of the microbial world—is the great adventure now beckoning to us. Microbial genomics paves the way forward.

Note: Supplementary information is available on the Nature Biotechnology website.

Published online at <http://www.nature.com/naturebiotechnology/>
Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>

ACKNOWLEDGMENTS

I would like to thank C. Woese, P. Hugenholz and C. Ouzounis for their critical reading and helpful suggestions, and M. Youle for her excellent editorial assistance. Special thanks to the members of the Genome Biology Program at the Joint Genome Institute for keeping me constantly in a most challenging and stimulating environment.

1. Roberts, R.J. Identifying protein function—a call for community action. *PLoS Biol.* **2**, E42 (2004).

2. Woese, C.R. A manifesto for microbial genomics. *Curr. Biol.* **8**, R781–R783 (1998).
3. Liolios, K., Mavromatis, K., Tavernarakis, N. & Kyrpides, N.C. The Genomes OnLine Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **36**, D475–D479 (2008).
4. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. Genbank. *Nucleic Acids Res.* **37**, D26–D31 (2009).
5. Woese, C.R. A new biology for a new century. *Microbiol. Mol. Biol. Rev.* **68**, 173–186 (2004).
6. Stanhope, M.J. *et al.* Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* **411**, 940–944 (2001).
7. DeLong, E.F. Microbial community genomics in the ocean. *Nat. Rev. Microbiol.* **3**, 459–469 (2005).
8. Kyrpides, N.C. & Ouzounis, C.A. Whole-genome sequence annotation: 'Going wrong with confidence'. *Mol. Microbiol.* **32**, 886–887 (1999).
9. Markowitz, V.M. Microbial genome data sources. *Curr. Opin. Biotechnol.* **18**, 267–272 (2007).
10. Garrity, G.M. *et al.* A new model of open access publishing: an ejournal for the Genomic Standards Consortium. *OMICS* **2**, 157–60 (2008).
11. Fox, G.E. *et al.* The phylogeny of prokaryotes. *Science* **209**, 457–463 (1980).
12. Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome Biol.* **3**, REVIEWS0003 (2002).
13. Hugenholtz, P. & Kyrpides, N.C. A changing of the guard. *Environ. Microbiol.* **11**, 551–553 (2009).
14. Galperin, M.Y. & Cochrane, G.R. *Nucleic Acids Research* annual database issue and the NAR online Molecular Biology Database Collection in 2009. *Nucleic Acids Res.* **37**, D1–D4 (2009).
15. Huss, J.W., III *et al.* A gene wiki for community annotation of gene function. *PLoS Biol.* **6**, e175 (2008).
16. Mons, B. *et al.* Calling on a million minds for community annotation in WikiProteins. *Genome Biol.* **9**, R89 (2007).
17. Fleischmann, R.D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
18. Markowitz, V.M. *et al.* The Integrated Microbial Genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res.* **36**, D528–D533 (2008).
19. Nierman, W.C. *et al.* Structural flexibility in the *Burkholderia mallei* genome. *Proc. Natl. Acad. Sci. USA* **101**, 14246–14251 (2004).
20. Holden, M.T. *et al.* Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. *Proc. Natl. Acad. Sci. USA* **101**, 14240–14245 (2004).
21. Field, D. *et al.* The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.* **26**, 541–547 (2008).
22. Parkhill, J. Time to remove the model organism blinkers. *Trends Microbiol.* **16**, 510–511 (2008).
23. Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J. & Goodman, R.M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, 245–249 (1998).
24. Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. & Hugenholtz, P. A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.* **72**, 557–578 (2008).
25. Handelsman, J. *et al.* *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet* (The National Academies Press, Washington, DC, 2007).
26. Woese, C. The quest for Darwin's grail. *ASM News* **65**, 260–263 (1999).
27. Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
28. Oehmen, C.S. & Nieplocha, J. ScalaBLAST: A scalable implementation of BLAST for high performance data-intensive bioinformatics analysis. *IEEE Trans. Parallel Dist. Sys.* **17**, 740–749 (2006).
29. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." *Proc. Natl. Acad. Sci. USA* **102**, 13950–13955 (2005).
30. Kettler, G.C. *et al.* Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet.* **3**, e231 (2007).
31. Ishoye, T., Woyke, T., Stepanauskas, R., Novotny, M. & Lasken, R.S. Genomic sequencing of single microbial cells from environmental samples. *Curr. Opin. Microbiol.* **11**, 198–204 (2008).
32. Armengaud, J. A perfect genome annotation is within reach with the proteomics and genomics alliance. *Curr Opin Microbiol.* **12**, 292–300 (2009).
33. Toledo-Arana, A. *et al.* The *Listeria* transcriptional landscape from saprophytism to virulence. *Nature* **459**, 950–956 (2009).
34. Yooseph, S. *et al.* The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* **5**, e16 (2007).
35. Whitman, W.B., Coleman, D.C. & Wiebe, W.J. Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. USA* **95**, 6578–6583 (1998).