Genomes from Metagenomics

Itai Sharon and Jillian F. Banfield

Science **342**, 1057 (2013);
DOI: 10.1126/science.1247023

passes through a filter that retains red and white blood cells to layers of paper containing colorimetric reagents. Fluid flow within the paper is directed by patterned hydrophobic barriers that are easily created using a wax-based printer and heat source. Results can be documented, analyzed, and transmitted with a cell phone camera, and devices can be safely disposed of by incineration. Paper-based tests to amplify and detect nucleic acids have been reported recently (*15*).

Health care innovation should be available to all the world's citizens, but technical and economic barriers remain. Low-resource settings present challenging constraints that require design for context, safety, reusability, and reparability. The current landscape of appropriate technologies reflects a reaction to economic incentives, largely shaped by philanthropic efforts rather than market forces. Often these funding mechanisms favor technical innovation over simplicity, and resulting technologies are too costly and difficult to maintain at scale. Alternative approaches that explicitly design technologies to function in settings that lack resources for consumables, effective distribution systems, supply chains, and technology management programs and that incorporate early private sector partnerships are needed. In parallel, efforts to develop and support innovators in low-resource settings must be strengthened. Partnerships that focus on developing and disseminating integrated packages of technologies that address focused areas (such as technologies for the neonatal unit) can navigate technical and implementation barriers more efficiently than single technologies.

References and Notes

1. World Health Organization, Medical Devices: Managing the Mismatch. An Outcome of the Priority Medical Devices Project (World Health Organization, Geneva, 2010); http://whqlibdoc.who.int/publications/2010/9789241564045_eng.pdf.
2. S. R. Sinha, M. Barry, *N. Engl. J. Med.* **365**, 779 (2011).
3. M. T. Glynn, D. J. Kinahan, J. Ducrée, *Lab Chip* **13**, 2731 (2013).
4. P. Matthews, L. Ryan-Collins, J. Wells, H. Sillem, H. Wright, Africa-UK Engineering for Development Partnership, A Summary Report: Engineers for Africa: Identifying Engineering Capacity Needs in Sub-Saharan Africa (Royal Academy of Engineering, London, 2012); www.raeng.org.uk/news/publications/list/reports/RAEng_Africa_Summary_Report.pdf.
5. G. Msemo *et al.*, *Pediatrics* **131**, e353 (2013).
6. H. L. Ersdal, N. Singhal, *Semin. Fetal Neonatal Med.* 10.1016/j.siny.2013.07.001 (2013).
7. J. Brown *et al.*, *PLOS ONE* **8**, e53622 (2013).
8. J. Brown *et al.*, *Low-Cost, Highly Effective Respiratory Support: Reducing Neonatal Death in Rural Malawi*, Saving Lives at Birth Development Exchange, Washington, DC, 19 to 21 July 2013.
9. M. Mwau *et al.*, *PLOS ONE* **8**, e67612 (2013).
10. I. V. Jani *et al.*, *Lancet* **378**, 1572 (2011).
11. M. Bergeron *et al.*, *PLOS ONE* **7**, e41166 (2012).
12. C. C. Boehme *et al.*, *Lancet* **377**, 1495 (2011).
13. G. Meyer-Rath *et al.*, *PLOS ONE* **7**, e36966 (2012).
14. N. R. Pollock *et al., Sci. Transl. Med.* **4**, 152ra129 (2012).
15. B. A. Rohrman, R. R. Richards-Kortum, *Lab Chip* **12**, 3082 (2012).

Acknowledgments: We apologize to those authors whose works we could not cover owing to space limitations. R.R.-K. is supported by the NIH (U54A1057156-10, R03EB013973-03, R21CA156704-02), and the Bill and Melinda Gates Foundation; R.R.-K and M.O. are supported by the U.S. Agency for International Development (AID-OAA-A-13-00014). The device reported in (*7*) is the subject of U.S. Provisional Patent Application 61/730,353 (Bubble Continuous Positive Airway Pressure), filed on 27 November 2012.

10.1126/science.1243473

MICROBIOLOGY

Genomes from Metagenomics

Metagenomic approaches are rapidly expanding our knowledge of microbial metabolic potential.

Itai Sharon and Jillian F. Banfield

Evaluation of the functional capacities of microorganisms long relied on laboratory cultivation of individual species. About a decade ago, recovery of draft genomes for a few uncultivated bacteria and archaea from natural communities opened the way for physiological prediction of their environmental roles. Further development of the metagenomics methods used in those early studies now allows the rapid delivery of accurately reconstructed microbial genomes from diverse environmental samples. The resulting knowledge has the potential to revolutionize our understanding of the topology of the tree of life and the metabolic capacities distributed across it. Advances in bioinformatics promise a new era in which comprehensive genetic characterization is sufficiently rapid to find application in diagnostics for medicine, agriculture, forensic science, and biotechnology.

Metagenomics is a cultivation-independent method for studying microbes sampled directly from the natural environment. DNA is extracted and sequenced from one or a series of samples, and the resulting data is analyzed using computational tools. The approach addresses two important needs: It enables analysis of the 99% of microbes in nature that have not yet been cultivated, and it facilitates the study of organisms in the context of their community.

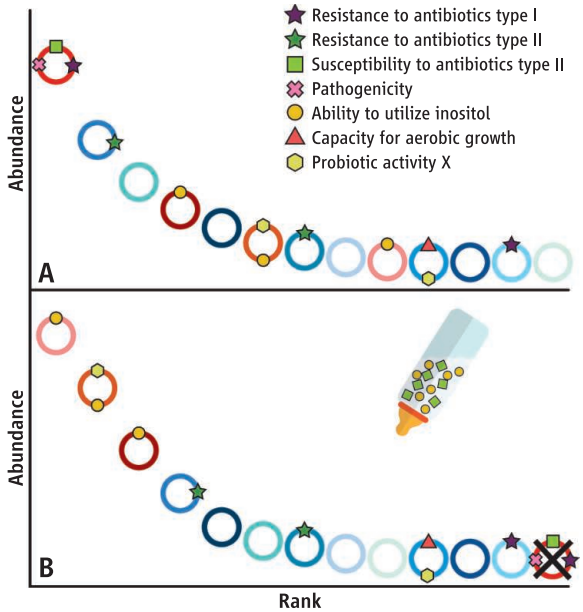
Because the DNA originates from multiple populations, the recovery of genomes from metagenomic data is a complex task. Until recently, genomes were reconstructed only from relatively simple environments with a few abundant genotypes (*1*). The advent of high-throughput DNA sequencing has enabled genomic sampling of much less abundant organisms and characterization of communities with relatively even species abundance levels, but the complexity of data analysis has increased greatly. Newly developed computational tools allow data assembly (*2*) and accurate assignment of genome fragments to specific organisms (*3*, *4*), a process termed binning.

In 2012, Wrighton *et al*. reconstructed 49 genomes with varying completeness levels for bacteria from at least five phyla for which there was almost no prior genomic information (*5*). The authors used a binning method that combines time-series abundance information with sequence compositional characteristics. More recently, Albertsen *et al*. (*6*) used information from multiple samples—an approach similar to that used for analysis of human infant gut microbial consortia (*4*)—to reconstruct 31 genomes with an average estimated genome completeness of 80% from DNA sequence information for an activated sludge bioreactor community. They were able to assemble the complete genome of an organism from the TM7 candidate bacterial phylum (lacking cultivated representatives) into a single contiguous sequence. Complete genomes for organisms that constitute ~1% of the community have also been reconstructed from environments such as the ocean (*3*) and, very recently, from adult human gut (*7*) and sediment (*8*, *9*). These examples demonstrate that metagenome-based genome recovery can now be applied to very complex systems.

Uncertainty about accuracy currently limits wide acceptance of metagenomics-derived

Department of Earth and Planetary Science, University of California, Berkeley, CA 94720, USA. E-mail: jbanfield@berkeley.edu

genomes. Of most concern are the possibilities of incorrect assembly of regions from different genomes into a single sequence and false assignment of assembled fragments to genomes during the binning process. Assembly errors can be addressed using automatic and manual procedures for error correction. Ultimately, the same criterion traditionally used to validate genomes from clonal cultures should be applied: consistent, unambiguous mapping of paired reads across the final sequence. In some cases, the credibility of the derived genomes also can be evaluated by comparison to published genomes for closely related organisms (*4*) or by additional sequencing using long, high-quality DNA sequences (*10*). Binning is an error-prone process that requires special care. Use of information from multiple sources, particularly unique patterns of organism abundance across samples, can reduce error levels substantially (*4*, *5*).



Treatment by genome-informed microbial selection. Medical treatments could benefit from knowledge of how microbial capacities are combined in individual organisms. In this hypothetical example, organisms are ranked in order of decreasing abundance. Circles represent genomes. Symbols on the circles indicate traits such as antibiotic resistance or substrate metabolic capacities. Prior to treatment (**A**), organisms with probiotic activity are low in abundance; the most abundant organism is pathogenic. Genome-informed choice of an infant formula containing appropriate antibiotics favors organisms with probiotic activity and eliminates the pathogenic organism (crossed out genome) (**B**).

An alternative to metagenomic approaches is single-cell genomics, in which single cells or cell concentrates are obtained directly from the environment, without cultivation, and sequenced. Recently, Rinke *et al*. (*11*) separated single cells from a variety of natural samples, amplified and sequenced the DNA, and recovered 201 partial genomes, with an estimated average genome completeness of 40%. To our knowledge, no complete genome has yet been achieved for a single cell alone (or by aggregating results from multiple single cells). Genomes generated by single-cell methods are typically very fragmented. In the study of Rinke *et al*., the best-assembled genome, estimated to be 100% complete, remained in 10 pieces; another genome, estimated to be 99% complete, remained in 137 pieces.

For the many genomes from metagenomic data that are initially incomplete (typically because of undersampling), completeness can be improved by additional sequencing. For single-cell genomes, additional sequencing is unlikely to improve the assembly. Metagenomics also does not entail the time-consuming cell manipulation or sorting required in single-cell genomics studies. However, the methods provide subtly different information. A genome derived from metagenomic data represents a population. Currently used assembly algorithms are essentially strain-specific, but there will likely be some nucleotides for which polymorphic variants are detected. Furthermore, a subset of individual sequencing reads will only partially match the consensus sequence if the sampled cell has an inserted or deleted gene. This is advantageous if the researcher is interested in overall population metabolic potential, population structure, diversity, or evolutionary dynamics. On the other hand, single-cell genomics can provide gene variant linkage information that is lost in metagenomic analyses.

As both single-cell genomics and metagenomics gain widespread acceptance, we advocate the use of clearly defined terms to describe genome completeness so as to accurately communicate the success or limitations of the analysis (*12*). A genome should be described as complete or finished only if it is assembled into a single contiguous sequence with no ambiguities or gaps, after careful checking for errors. Genomes assembled into multiple pieces where the fragment order cannot be resolved because of repeated sequences may be termed essentially complete. Following Chain *et al*. (*13*), genomes assembled into multiple pieces without all scaffold connections resolved should be defined as standard draft genomes. In such cases, completeness is typically calculated according to inventories of single-copy genes (*6*, *10*, *11*, *14*). Single-copy genes generally constitute less than 10% of the genes and are unevenly distributed across the genome, thus providing only a rough estimation of completeness.

More robust methods for assessing the completeness of draft genomes are needed. One improvement may involve the use of marker genes that do not tend to cluster together on the genome. Better sampling of microbial genomes from uncultivated organisms will enable refinement of the set of genes considered to be universal. These new genomes will also improve identification of single-copy genes that diverge from known sequences.

Given the important developments in sequencing speed, accuracy, and informatics, high-throughput metagenomic approaches have the potential to revolutionize fields where rapid, accurate, and strain-specific diagnostics are essential. One can imagine, for example, an age of personal microbiomics in which antibiotics are prescribed on the basis of accurate and fast screening of the resistance gene repertoire of the pathogen population (see the figure). Metagenomic insight could enable selective stimulation of desirable microbial populations so as to address medical conditions such as chronic diarrhea or obesity (*15*). More broadly, our understanding of life and its evolutionary history will be dramatically advanced by access to genomes from the numerous, previously unstudied parts of the tree.

References

1. G. W. Tyson *et al*., *Nature* **428**, 37 (2004).
2. Y. Peng, H. C. Leung, S. M. Yiu, F. Y. Chin, *Bioinformatics* **28**, 1420 (2012).
3. V. Iverson *et al*., *Science* **335**, 587 (2012).
4. I. Sharon *et al*., *Genome Res.* **23**, 111 (2013).
5. K. C. Wrighton *et al*., *Science* **337**, 1661 (2012).
6. M. Albertsen *et al*., *Nat. Biotechnol.* **31**, 533 (2013).
7. S. C. Di Rienzi *et al*., *eLife* **2**, e01102 (2013).
8. C. J. Castelle *et al*., *Nat. Commun.* **4**, 2120 (2013).
9. R. S. Kantor *et al*., *mBio* **4**, 5 (2013).
10. A. Voskoboynik *et al*., *eLife* **2**, e00569 (2013).
11. C. Rinke *et al*., *Nature* **499**, 431 (2013).
12. E. Mardis *et al*., *Genome Res.* **12**, 669 (2002).
13. P. S. Chain *et al*., *Science* **326**, 236 (2009).
14. B. K. Swan *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **110**, 11463 (2013).
15. V. K. Ridaura *et al*., *Science* **341**, 1241214 (2013).

10.1126/science.1247023