Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Current Opinion in
Systems Biology

The microbiome and big data

Jose A. Navas-Molina[1], Embriette R. Hyde[2],
Jon G. Sanders[2] and Rob Knight[1,2,3]

Abstract

Microbiome datasets have expanded rapidly in recent years. Advances in DNA sequencing, as well as the rise of shotgun metagenomics and metabolomics, are producing datasets that exceed the ability of researchers to analyze them on their personal computers. Here we describe what Big Data is in the context of microbiome research, how this data can be transformed into knowledge about microbes and their functions in their environments, and how the knowledge can be applied to move microbiome research forward. In particular, the development of new high-resolution tools to assess strain-level variability (moving away from OTUs), the advent of cloud computing and centralized analysis resources such as Qiita (for sequences) and GNPS (for mass spectrometry), and better methods for curating and describing “metadata” (contextual information about the sequence or chemical information) are rapidly assisting the use of microbiome data in fields ranging from human health to environmental studies.

Addresses

[1] Department of Computer Science and Engineering, The University of California San Diego, 9300 Gilman Drive MC0736, La Jolla, CA, 92093, USA
[2] Department of Pediatrics, The University of California San Diego, 9300 Gilman Drive MC0763, La Jolla, CA, 92093, USA
[3] The Center for Microbiome Innovation, The University of California San Diego, La Jolla, CA, USA

Corresponding author: Knight, Rob (robknight@ucsd.edu)

Current Opinion in Systems Biology 2017, **4**:92–96

This review comes from a themed issue on **Big data acquisition and analysis (2017)**

Edited by **Pascal Falter-Braun and Michael A. Calderwood**

For a complete overview see the Issue and the Editorial

Available online 11 July 2017

http://dx.doi.org/10.1016/j.coisb.2017.07.003

2452-3100/© 2017 Published by Elsevier Ltd.

Keywords

Microbiome, Big data, Sequencing, Metabolomics.

From cells to bits: what is big data in microbiome research?

Since the term “microbiome” was coined by Joshua Lederberg in 2001 [1], the microbiome research field has exploded both in terms of the heterogeneity of the data produced and in the amount of data generated. Early approaches to characterizing the microbiome were based on targeted detection techniques in the laboratory, such as culturing and assays based on the Polymerase Chain Reaction (PCR), and assessed limited numbers of subjects (on the order of tens) [2]. The introduction of sequencing technologies revolutionized the field, enabling investigators to characterize microbial communities directly from primary samples. Historically, the 16S rRNA gene, a marker gene that exists in all bacteria and archaea as an essential part of the ribosome, has been targeted for these sequence-based profiling efforts. Its ubiquity among bacteria and archaea and the low cost of the approach has made it the most widely used for microbiome profiling of samples. Similarly, amplification and sequencing of the 18S rRNA gene and the internal transcribed spacer (ITS) permit investigators to profile the eukaryotic and fungal communities present in a sample using similar techniques. Since the introduction of Next Generation Sequencing, technologies have evolved from generating a few hundred thousand reads per run (454 GS) to tens of million reads (Illumina MiSeq) or even a few billion reads per run (Illumina HiSeq) [3]. Benchmarked protocols, such as those used by the Earth Microbiome Project and widely adopted by researchers around the globe, facilitate meta-analyses of unprecedented size-investigators can combine studies, each with hundreds to thousands of samples, into a single large analysis effort.

The precipitous drop in sample processing and sequencing costs associated with new technology development is enabling researchers to move beyond simple taxonomy and abundance-based work to species and strain level profiling as well as descriptions of functional pathways through whole genome shotgun metagenomics sequencing. As a result, researchers are able to ask more critical questions of their samples and are utilizing other technologies, such as detection of small molecules via mass spectrometry, to confirm or refute hypotheses driven by functional pathway and gene abundance information obtained from shotgun sequencing data.

The rate at which these technologies are increasing their data output is faster than our computational power is growing [4], effectively shifting the costs of a research study from the sequencing pipeline to the data analysis pipeline. Additionally, as researchers utilize larger and larger datasets, they are able to design large-scale studies to ask (and answer) complex questions. The metadata associated with samples, therefore, is becoming an increasingly large contributor to microbiome big data and the challenges associated with streamlining data analysis. Standards such as MIMARKs [5] have helped investigators format their metadata to facilitate data analysis and data upload to repositories such as the European Bioinformatics Institute's European Nucleotide Archive (EBI ENA). Nevertheless, as samples are increasingly processed in parallel with multiple different protocols (i.e., 16S, 18S, ITS, shotgun, metabolomics, etc.), correct formatting of metadata to capture this information and facilitate multi-omics correlative analyses will require careful attention and appropriate implementation of tools capable of handling hundreds to thousands of columns of data for hundreds to thousands of samples. Tools such as Qiita (qiita.microbio.me) are being developed to address the challenges associated with analyzing large numbers of samples, processed via multiple different protocols, and with complex metadata-and these tools rely on both the availability and effective usage of large-scale compute resources. The ability to apply tools such as QIIME in the cloud; e.g., using Amazon Web Services [6], has broadened these capabilities far beyond the original user base, and enabled users in developing countries such as Bangladesh to use these tools without operating their own large-scale compute infrastructure. These techniques are now being applied in the United States through Illumina's BaseSpace (https://basespace.illumina.com/home/index) and NIH's Cloud Pilot (https://commonfund.nih.gov/bd2k/commons).

From bits to knowledge: how is big data moving microbiome research forward?

Initial efforts to characterize and understand the healthy human microbiome using next generation sequencing techniques [7,8] raised more questions than answers, and led to the explosion of microbiome research that has identified associations between the microbiome and diseases as varied as obesity, inflammatory bowel disease, cardiovascular disease, and autism (among many others). Most of these studies have simply identified associations and the question of causation or simple association remains unknown. Key studies, such as the obesity work done by Jeffrey I. Gordon and his team at Washington University [9–11] and the personalized nutrition work done by Eran Segal of the Weizmann Institute [12] are coming closer to answering the question of causality versus association. However, it is becoming increasingly clear that integrating DNA sequence data with other 'omics techniques such as metatranscriptomics (sequencing the RNA), proteomics (sequencing the proteins), and metabolomics (characterizing the metabolites) will be key for advancing microbiome research. An example of the power of combining multiple techniques for assessing the microbiome is the National Institutes of Health's (NIH) Human Microbiome Project (HMP), the largest human microbiome sequencing effort at the time of its publication in 2012. 16S rRNA gene amplicons were generated from total of 4788 samples collected from 242 healthy adults [7] and sequenced using 454 pyrosequencing. Additionally, a whole genome shotgun sequencing on the paired-end Illumina platform was performed on a subset of 681 samples, generating 2.9 Gigabases per sample (close to 2 terabytes of data for the entire dataset).

The HMP shotgun metagenomics data revealed a key observation: while no taxon was observed in all individuals (i.e., no "core" healthy microbiome was identified), the functional pathways inferred from the shotgun data were evenly distributed across individuals and body sites. While this was an important observation, the addition of other data types, such as RNA-seq or metabolomics would have provided precise information regarding the actual activity of the microbial community and which small molecules were present, respectively, further exemplifying importance of combining different -omics techniques for generating hypotheses that ultimately lead to studies designed to obtain a more complete picture of a given microbial community (and the significance of its presence). For example, as reported by Bouslimani et al. [13], using a paired sequencing-mass spectrometry approach allowed the investigators to identify correlations between *Propionibacterium* genera and the presence of oleic acid, palmitic acid, mono-oleic, and palmitic acylated glycerols on human skin. Hypothesizing that *Propionibacterium* mediates the hydrolyzation of triacylglycerides or diacylglycerides from human acylated glycerols, Bouslimani et al. cultured *Propionibacterium acnes* in a medium supplemented with the triglyceride triolein and examined the resulting metabolic products, ultimately confirming their hypothesis.

Microbiome citizen science initiatives such as the American Gut Project (AGP; americangut.org) have made significant contributions to the field by "democratizing" microbiome research and thus providing large-scale datasets that can be used as comparative frameworks for other studies. Citizens support the science by sending samples from their bodies, their pets, or their environment as well as the necessary funds to cover the sample processing. These projects face the challenge of dealing with large numbers of samples; while most current microbiome studies contain hundreds or a few thousand samples, these citizen science efforts contain a continually growing number of samples that in some cases are on the order of over ten thousand samples, pushing the limits of the current computational tools. Furthermore, this democratization is not free: subject data is self-reported, and at times, significant amounts of data are necessary to correctly characterize the sample source. The American Gut Project currently collects up to 400 variables about study participants, including