Microbiota meet big data

Big Data analytic tools will be invaluable for extracting meaning from microbiome data to enable new solutions to global health problems and provide new insights into microbiology.

Despite weighing in at less than eight ounces, the total number of microbial cells that are in a human body outnumbers our own cells by a factor of ten. Less than 1% of these foreign species within can be cultured, so their identification has relied on next-generation sequencing techniques used, for instance, on PCR-amplified 16S rRNA contained within the various niches of the human body and nearly everywhere microbes exist. The collection and storage of the resultant data require significant infrastructure and have greatly benefited from the use of cloud computing. What are needed now are robust analysis tools that can mine the data to find trends and gain new mechanistic insights and connections among members of the microbiota, informing on their chemistry and biology.

The global information storage capacity has roughly doubled every 40 months since the 1980s, and larger and larger data sets are being amassed. The Human Microbiome Project (http://commonfund.nih.gov/hmp/index/), a US National Institutes of Health (NIH) program, brought microbiomes to the spotlight and completed its first phase last year—its characterization of the microbial compositions from a cohort of 300 healthy adult human subjects within five regions of the body, including nasal passages, the gastrointestinal tract and the urogenital tract. There are a number of repositories available for depositing microbiome data collected outside of the Human Microbiome Project, including QIIME (http://www.microbio.me/qiime/) and MG-RAST (http://metagenomics.anl.gov/). These databases make sequence reads and other primary data available to researchers and represent new insights waiting to be found.

Early insights from these efforts are numerous, particularly in relation to health and disease, with small numbers of healthy and problematic microbiomes having been defined so far. For instance, gut microbiota and the metabolites that they generate have an immense impact on host physiology by modulating the chemistry of the gut (*Nat. Chem. Biol.* **10**, 416–424, 2014). As well, a causative connection has been established between the reduced diversity of the microbiota found in fecal samples and the inflammatory bowel diseases Crohn's disease and ulcerative colitis; disturbances in the bacterial microbiota of the gut result in dysregulation of adaptive immune cells underlying these disorders (*Nat. Rev. Immunol.* **9**, 313–323, 2009).

Similarly, research from Jeffrey Gordon's group at Washington University has found causative relationships between the microbiome of children in impoverished countries and malnutrition (*Science* **339**, 548–554, 2013; *Nature* **510**, 417–421, 2014). Going beyond the use of probiotics as a strategy to restore a healthy microbiome, this work has led to the proposal of generating therapeutic foods designed to promote growth of the microbiota characteristic of that of healthy children. The importance of the microbiome composition in children is highlighted by the recent discovery that antibiotic treatment in newborns is a risk factor for growth of toxin-producing *Clostridium difficile*, which colonizes the gut of infants more than that of adults.

These and other advances, such as the successes of fecal transplantation for antibiotic-associated diarrhea caused by *C. difficile*, illustrate that we have reached a point where decisions need to be made about which microbiota are to be considered optimal and healthy. Microbiome information promises to help doctors to predict patients' risks of developing various conditions. Recognizing the power of Big Data as a research tool, hospitals and research centers are changing the face of healthcare (http://gigaom.com/2012/07/15/better-medicine-brought-to-you-by-big-data/).

Beyond their compositions, it is also now increasingly important to determine how different microbes interact with one another and with the human host within their niches (*Nature* **500**, 16–17, 2013). Big Data—also called 'found data' because it represents the data that you did not know existed—has the wherewithal to help. Big Data, in its essence, represents large amounts of data for which you do not have the computational tools to query thoroughly or current tools make analysis impossibly time consuming. Big Data engines that relax search rules provide the advantage of quicker and more flexible queries along with real-time data analytics. It is in this mining of the data that Big Data engines shine. As an example of the concept of using Big Data to uncover new insights, web browser search queries were used to track influenza-like illness within the US population (*Nature* **457**, 1012–1014, 2009). New ways to manage and analyze data are welcome as they offer important complements to more traditional data collection and analysis, but an important goal is to ensure that these methods are well validated and are used to answer hypothesis-driven questions.

Beyond storing data, microbiome data need deeper mining tools to retrieve the found data that will inform on the mechanistic connections among individual microbes. Compared to the challenges of data storage and mining that came along with the Human Genome Project, an arguably hypothesis-free endeavor, the Human Microbiome Project involves random sampling of sequences for which the source organisms are not known but need to be inferred in some way. New query languages allow for the parallelization that such big data sets require. Indeed, the Human Microbiome Project data are currently searchable via Amazon web services (https://aws.amazon.com/datasets/1903160021374413/) and other cloud services that query the central repository for all Human Microbiome Project data at the NIH Data Analysis and Coordination Center (http://hmpdacc.org/).

The second phase of the Human Microbiome Project that began in 2013 is focused on developing data sets and tools for evaluating which biological properties of the microbiome and host will be most informative in understanding health and disease. Indeed, a slew of new data is in the offing, which already requires careful consideration of how they are collected, stored and mined to extract meaning. A suite of data collection and search tools and the mechanistic mindset inherent in chemical biology have now converged to answer questions about host-microbe interactions, microbial communication, antibiotic development and human physiology and health. ■