Spotlight

De novo Assembly Vastly Expands the Known Microbial Universe

Samuel S. Minot[1,*]

The study of the human microbiome relies heavily on the genomes of bacterial isolates that can be grown in culture. A recent study (Pasolli *et al. Cell* 2019;176:649–662) of stool microbiome samples generated over 150 000 microbial genomes without any culture, vastly expanding our knowledge of the biases in existing reference databases.

The study of microbes in the environment (including microbial ecology, host–microbe interactions, etc.) faces a perennial challenge that can be difficult to explain fully to people outside of that field – that the microbes present in the environment only partially resemble the microbes available in reference databases. A recent publication by Pasolli *et al*. [1] used a database-free approach to try to estimate the size and shape of this missing-data problem, and generated some really intriguing results on just how little we know.

For people who study a single microbe, or no microbes at all, it can be difficult to comprehend the immense diversity of the microbial world. Some microbes have a high mutation rate and start to resemble new lineages that persist for months and years within a single host [2,3]. Some microbes are naturally competent and regularly ingest DNA from the environment for potential integration into their genomes [4]. Some microbes are continually assaulted by genomic parasites that attempt to manipulate their genome and move between cells via pili or viral capsids [5]. On top of that are all of the microbes that we have never cultivated and do not even know how little we know about them. All combined, for every piece of DNA isolated and sequenced from the environment (which is a common approach to microbial ecology and microbiome science) we may (i) have the right idea, (ii) have no idea, or (iii) have the completely wrong idea about what organism it came from. As much as we might want to know the ratios of (i), (ii), and (iii), we rarely have an appropriate gold standard ground truth to benchmark against. To side-step this fundamental challenge, Pasolli *et al.* [1] took the approach of assembling as many genomes as they could from scratch, *de novo* assembly, and then used that set of uncultivated genomes to map the contours of our missing knowledge.

As other groups have taken similar approaches in recent years [6] it can be instructive to point out the technological innovations that have motivated and enabled the study of uncultured microbial genomes. In this case there are two complementary technologies that have enhanced our ability to assemble genomes *de novo* (from scratch) out of uncultured samples. One technology is the set of computational algorithms for *de novo* genome assembly, which have advanced in recent years in large part due to the efforts of groups using genome-resolved metagenomics to study the dynamics of microbial communities [7]. The other is sequencing technologies (such as Hi-C and PacBio) which provide additional long-range structure for generating highly contiguous assemblies of bacterial genomes [8]. These complementary advances in computational and sequencing technologies have increased the sensitivity and precision of researchers attempting to identify complete microbial genomes from sequencing data.

By reanalyzing public datasets with metagenomic whole-genome ‘shotgun’ (WGS) sequence data from human microbiome samples, Pasolli *et al*. were able to generate >150 000 complete and partial bacterial genomes, orders of magnitude more than has been generated by recent culture-based sequencing studies [9, 10], or even other recent culture-free approaches [6]. While it is true that many of those microbial genomes are missing from existing reference databases, it would be a mistake to take that as the primary conclusion. In fact, the most striking aspect of these results is how clearly they show that existing reference databases are biased in their representation, both taxonomically and biogeographically. In other words, the genomes that are missing from the reference databases follow a distinct pattern that is a result of the biases used to create them. Given the central importance of reference databases in the study of the microbiome, it is crucial to understand any biases that may arise from their use.

The taxonomic bias of microbial genome identification has been acknowledged qualitatively, but it has never been described in such detail. Put simply, some species, genera, families, etc., of bacteria are more amenable to cultivation, and therefore more likely to have been sequenced and deposited in reference databases. Pasolli *et al.* provide a compelling demonstration of this effect within the Ruminococcaceae, showing entire species groups that are sister clades to *Faecalibacterium* and *Ruminococcus*, and yet have never been sequenced before. These results provide compelling evidence that there are species of bacteria in the gut microbiome that are found at high abundance across many individuals, and yet have never been studied in the context of any reference genome database.

The biogeographic bias of microbial reference databases is demonstrated with the finding that the previously unknown microbial species from this study make up a larger proportion of the microbiome in non-Westernized populations from regions including Madagascar, Fiji, and Tanzania, compared with Westernized populations from around the world. Similar results have been presented by other groups that also used a culture-free approach for microbial genome reconstruction [6]. Operationally, this means that we (as a scientific community) have been doing a worse job at characterizing the microbiome of individuals living 'non-Western' lifestyles, so much so that the genomes from this study roughly doubles the amount of information we can capture from those microbiome samples. These results also suggest that microbial reference databases may be similarly biased against other socioeconomic groups that are less likely to have access to research studies, even within industrialized countries. It is unlikely that anyone would dispute the general idea that non-Westernized populations have been understudied, but it is remarkable to see how large the effect of that bias has been, representing almost half of those microbial communities.

Technological advances have dramatically expanded our understanding of the microbiome in recent years, and in so doing have revealed the biases in our previous gaps of knowledge. However, we should assume that, even if those biases have been reduced, they likely still exist. Therefore, our microbial reference databases are likely still biased by taxonomy and biogeography, and any research that is built on those databases is similarly influenced. Having been forewarned, we are now forearmed; we have the opportunity use this expanded universe of microbes to augment our understanding of the human microbiome and how it influences the health of all people.

Disclaimer Statement

S.S.M. holds financial interest in Reference Genomics, Inc.

[1]Microbiome Research Initiative, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

*Correspondence:
sminot@fredhutch.org (S.S. Minot)

https://doi.org/10.1016/j.tim.2019.02.007

© 2019 Elsevier Ltd. All rights reserved.

References

1. Pasolli, E. *et al.* (2019) Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 176, 649–662.e20.
2. Garud, N.R. *et al.* (2019) Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biol.* 17, e3000102.
3. Minot, S. *et al.* (2013) Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U. S. A.* 110, 12450–12455.
4. Redfield, R.J. *et al.* (2006) Evolution of competence and DNA uptake specificity in the Pasteurellaceae. *BMC Evol. Biol.* 6, 82.
5. von Wintersdorff, C.J.H. *et al.* (2016) Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.* 7, 173.
6. Almeida, A. *et al.* (2019) A new genomic blueprint of the human gut microbiota. *Nature* Published online February 11, 2019. https://doi.org/10.1038/s41586-019-0965-1.
7. Sieber, C.M.K. *et al.* (2018) Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* 3, 836–843.
8. Stewart, R.D. *et al.* (2018) Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* 9, 870.
9. Forster, S.C. *et al.* (2019) A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat. Biotechnol.* 37, 186–192.
10. Zou, Y. *et al.* (2019) 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* 37, 179.