Overcoming hurdles in sharing microbiome data

Increasing research on microbial communities has resulted in massive amounts of data being generated and shared, yet data accessibility, accuracy and thoroughness remain problematic and can be a substantial obstacle for scientists looking to explore existing datasets.

Metagenomics has now been commonplace for over a decade and continuous improvements in sampling efforts, computing power, and bioinformatics pipelines have resulted in the microbiome research field collectively generating vast amounts of data, with the aim to better understand the complex microbial communities associated with different environments. The data are not limited to nucleic acid sequence alone, but also the accompanying metadata and outputs from other technologies including metabolomics, metatranscriptomics, proteomics and imaging. Making such data publically available can help promote the transparency of scientific research, but even more importantly it allows members of the scientific community to replicate, confirm and build on the results of others. The reuse of these datasets for meta-analyses and data mining can enable microbiologists to ask new questions and develop new bioinformatics tools. Accordingly, consortia efforts such as the US Human Microbiome Project, MetaHIT, the Earth Microbiome Project and *Tara* Oceans, as well as a plethora of laboratory-level projects, have generated extensive pools of data, which have been used to validate findings in independent studies. However, generation, sharing and reuse of multiple datasets does not come without complications. For example, the range of extraction methods, technologies and analytical pipelines available for processing similar samples makes comparison between studies incredibly difficult. Steps have been taken to promote standardization of sample processing, sequencing and data analysis across the microbiome field. However, achieving global standardization is a distant aim at the moment and will not be appropriate or possible for all types of microbiome research given the rapid emergence of new technologies and specific methods needed for different samples[1].

It is currently standard practice for data to be shared through government-sponsored databases such as the Sequencing Read Archive or the European Nucleotide Archive, which are the 'go-to' repositories for many researchers and mandated by various journals, including *Nature* research journals. Other, more specific databases also exist for microbial community data, for example MG-RAST and Qiita, and for specific methods, such as the Global Natural Products Social Molecular Network for metabolomics data[2]. A list of recommended repositories curated by *Scientific Data* can be found here: https://www.nature.com/sdata/policies/repositories. Cloud-based storage is becoming increasingly common with the NIH cloud Commons, Illumina BaseSpace and Figshare. Although still in relative infancy, many cloud platforms attempt to bypass the tedious and resource-draining step of downloading data from traditional repositories onto local servers and have integrated analysis tools. This is a particularly pressing concern given the increasing size of data sets, and if funded long term, these platforms could well become the preferred method for sharing raw data and the associated analyses.

Given the size, complexity and diverse formats that come with generating massive data, sharing these records is an onerous task and one that has several ongoing issues. Making data available is not only time-consuming, but there is no single, straightforward pipeline or database for sharing it and no simple mechanism to ensure that data release is coordinated with a research paper. Furthermore, lack of standardization can result in poor records of datasets, with missing samples, incomplete descriptions of associated technical and biological metadata, and in some cases no information regarding the analyses that were performed, such as detailed and informative bioinformatics scripts. These issues are not limited to a handful of datasets, but are widespread, creating limitations on their value for the field as a whole. Indeed, *Nature Microbiology* has run into such hurdles over the past couple of years, where data have not been released in a suitably timely manner to coincide with publication of studies. Like most journals, we do have procedures in place to try and prevent occurrences of this type, but unfortunately such issues have arisen on occasion, especially when large datasets have been involved. We remain committed to working together with our authors and the data repositories to identify additional steps by which we can coordinate release of datasets concurrently with our research papers, and to ensure that data and associated metadata are comprehensive and accurate when released. Community-led standards exist for ensuring data are accompanied by thorough and useful metadata, for example the Minimum Information about any (x) Sequence checklist[3]. Although, for some human cohorts there are additional obstacles preventing the distribution of informative metadata associated with patient information, which usually require Institutional Review Board approval. Furthermore, the publication of Data Descriptors in *Scientific Data* can be used to provide a far richer account of a dataset, helping others to better reuse the data and to credit those researchers who share their data. Compliance with community-agreed minimum reporting standards and publication of a data descriptor to support a dataset and research article are, however, a long way off from universal adoption.

What else can be done to stimulate data sharing in the microbial field? Encouraging and incentivising public data sharing is vital, including the adoption of minimum reporting standards, and this can be achieved by working more closely with databases and journals to promote their recognition. Training for life scientists in bioinformatics and promoting an overall increased awareness of the different databases available for the range of data types and appropriate analyses to use is another way to improve this. There will of course be other solutions to achieve the goal of improved microbiome data sharing, and we are certainly not the first to enter the discussion surrounding such issues. However, promoting discourse between researchers and journals is one way that we can help the field and to this end, we would encourage all readers to share their thoughts with us on what could and should be done by journals, authors and repositories alike to overcome these hurdles and promote greater and better data sharing. ❐

Published online: 24 November 2017
https://doi.org/10.1038/s41564-017-0077-3

References

1. *Nat. Microbiol.* **7**, 16112 (2016).
2. Wang, M. *Nat. Biotechnol.* **8**, 828–837 (2016).
3. Field, D. *Nat. Biotechnol.* **5**, 541–547 (2008).