MICROBIOLOGY

The strain in metagenomics

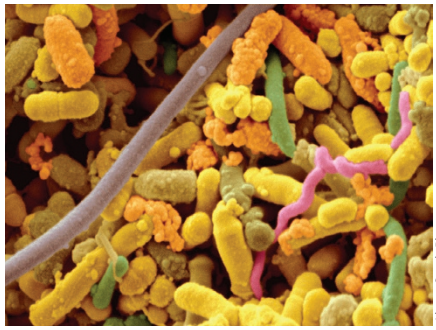
Two computational tools extract strain-level information from reams of microbial sequence data.

Doctors are well aware that differences between strains of the same bacterial species can have consequences at opposite ends of the health-and-disease spectrum. "If you only look at a species-level annotation, you may end up missing the fact that there are virulence genes in some *E. coli* strains," says Ramnik Xavier of Massachusetts General Hospital, the Broad Institute and Harvard University. Identifying strains, along with the genetic potential of their genomes, has become a key goal for computational biologists, but the challenges are compounded by enormous data sets, missing reference sequences and rare species.

At the Broad Institute, Dirk Gevers, Xavier and their colleagues began thinking about the problem when they analyzed the first Human Microbiome Project shotgun data. "We detected a person-specific profile at the subspecies level but could not deconvolute it back then," says Gevers, now at the Janssen Human Microbiome Institute. It took postdoc Chengwei Luo to ultimately drive development of their ConStrains strain-detection software.

In principle, unique patterns of single-nucleotide polymorphisms (SNPs) can identify strains, but SNP-based approaches rely on reference sequences or are limited to the SNPs that appear in a short sequence read. In the ConStrains approach, reads are recruited using universal core genes present in all species with tenfold sequence coverage. The software then uses a reference-independent strategy to cluster and concatenate the SNPs into strain-specific fingerprints. ConStrains brings quantitative metagenomics to the strain rather than the species level, without the need for additional types of data (Luo *et al.*, 2015).

Because fingerprints are identified from single samples, entire cohorts need not be analyzed at once, which eases computation. Gevers and Xavier say that ConStrains shines when it comes to longitudinal studies. The researchers used it to track strains in large published data sets, including five strains of *Bifidobacterium longum* in infant guts associated with distinct functions that were missed in species-level analyses.



Alamy Stock Photo

Microbes from the wild are now being studied at the strain level.

Nearby at the Massachusetts Institute of Technology, Eric Alm was addressing similar questions with single-cell sequencing when he was pulled back into metagenomic analysis. "It was almost accidental," he says. A chance e-mail introduced him to Brian Cleary, just off a post-college stint developing algorithms on Wall Street. Cleary was working on an algorithm that searched for words across documents, which Alm realized was similar to finding patterns in sequence reads. "Within a couple of weeks he came back and he had something that was starting to actually work for metagenome assembly," says Alm.

Reference-free genome assembly combines short reads into longer stretches called contigs—a computationally intensive step when starting from complex microbial mixtures. One way to then assign contigs to species 'bins' is to track their abundance across samples; those that vary in the same way probably belong to the same species.

Cleary's insight was to assign reads to species bins first, then assemble contigs from these much smaller piles of data. His Latent Strain Analysis (LSA) assesses the covariation of short '*k*-mer' sequences found in reads, on the basis of a hashing function from his search algorithm that represents all *k*-mer abundance patterns in a few gigabytes of fixed memory, regardless of data size (Cleary *et al.*, 2015). LSA also speeds up the covariation-detection and assembly steps.

The advantages are manifold. LSA can discriminate highly related strains, and it has worked efficiently on four terabytes of gut microbiome data, whereas other approaches top out at around 100 gigabytes. It also found microbes that are missed by traditional assembly because they contribute few reads to any single sample. "If you could get all those reads put into one box, that box may have 30-fold coverage," Alm explains. One family they detected made up just one–10-millionth of the gut data.

LSA assembled up to 90% of some genomes, enabling a deeper functional understanding. Very similar regions that cannot be distinguished are often interpreted as separate genomes, however. "That's a neat opportunity for someone to develop a new tool" to put together scattered genomes—something they are working on, says Alm. He estimates that LSA works best with 100 or more gut samples, a study size that is becoming more commonplace.

The researchers will use their tools to understand how strains function in their environments. "We want to do *de novo* assembly-based strain calling," and to link strains to function, says Gevers. After census-taking and correlation, "the next step would be…to get to causality," says Xavier.

Tal Nawy

RESEARCH PAPERS

Luo, C. *et al.* ConStrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* **33**, 1045–1052 (2015).

Cleary, B. *et al.* Detection of low-abundance bacterial strains in metagenomic data sets by eigengenome partitioning. *Nat. Biotechnol.* **33**, 1053–1060 (2015).