Microbiology: the road to strain-level identification

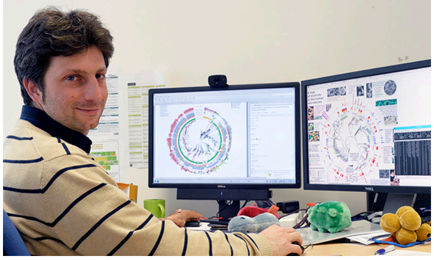
Vivien Marx

Tools are emerging to help labs trawl for sequences that reveal microbial strains and their functional potential in deep pools of metagenomic data.

'Who is there' and 'what are they doing' are typical questions in military intelligence and in metagenomics. Both fields sift through big data for signals. Speed is crucial when averting terrorist threats or hunting the cause of food contamination or an infectious disease outbreak. In calmer times, both communities want to keep learning about their subjects of interest. Both communities also benefit from sophisticated tools. Metagenomics researchers analyze genetic information from microbes in different environments, including the human body, using high-throughput sequencing and computational methods.

In metagenomics, researchers might analyze sequencing reads to find out how the gut microbiome differs between individuals with and without Crohn's disease; others capture the teeming bacterial diversity in samples of soil removed at different times from the same location or in ocean water samples taken at varying time points or depths. Analyses reveal slews of invisible microbial species. But addressing the question 'who is there' calls for more than species-level identification: researchers want to identify microbes on the strain level.

Some, but not all, microbial strains spell trouble. US officials recently determined that Shiga-toxin-producing *Escherichia coli* (STEC O26) caused food poisoning in customers of a particular restaurant chain. In the summer of 2011, the food-borne pathogen enterohemorrhagic *E. coli* (EHEC) O104:H4 led to illness and deaths in Germany.



Alessio Coser, University of Trento

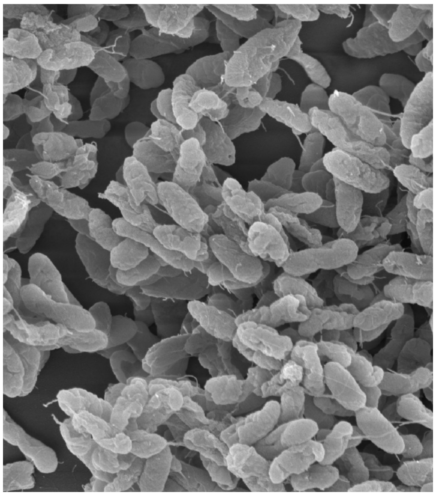
Analysis of both DNA and RNA gives a fuller view of a microbial community, says Nicola Segata.

Successful strain-level identification is part of a larger metagenomics trend. Over the past five to ten years, scientists have continuously improved the resolution of their microbiome data analysis, says Nicola Segata, a metagenomics methods developer at the University of Trento, Italy. Scientists kept moving down the taxonomic ranks as they distinguished phyla, families, genera and species. The discovery of microbial diversity below the species level led them further still. "Differences in different strains of the same species are crucial for a lot of tasks," says Segata.

It has been difficult to extract strain-level insight from short-read sequence data. Now it's becoming more routinely possible to extract not just species genomes but also strains from the metagenome data contained in multiple samples, says Christopher Quince, a microbiome researcher and methods developer at University of Warwick Medical School.

Within the same species, microbes can differ genetically and perform quite different functions. Such differences might help to explain the emerging contradictions in the burgeoning microbiome literature. Different labs find varying types of association of microbial phyla with certain diseases, says Alice McHardy, a computational biologist at the Helmholtz Center for Infection Research in Braunschweig, Germany. Strain-level analysis may help resolve these seeming contradictions.

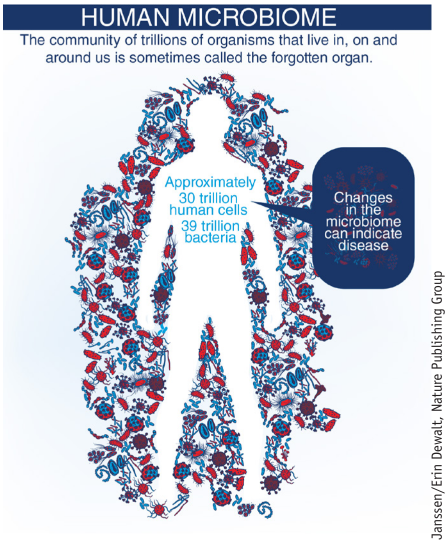
One method, the 'bag of genes' approach, looks at the functions encoded by all genes in the entire metagenome sample. In contrast, strain-level reconstructions shed light on which genomic properties, such as particular gene functions or even single-nucleotide polymorphisms, are distinctive for particular microbial and host phenotypes, says McHardy. Such reconstructions help scientists explore metabolic changes in strains over time or under varying conditions. Strain-level information could contribute to research about pathogen–host interaction, she says, which can help with treatment decisions and improve understanding of a pathogen's biology. Analysis can reveal microevolution aspects and phenomena such as horizontal gene transfer or mutational hot spots.



M. Rohde, Helmholtz Center for Infection Research

Bacteria of one species can appear alike, but different strains can actually differ markedly from one another and have different functions.

The field's progress has led to microbiome-focused biotech ventures, and larger companies are taking notice. What he finds exciting, says Dirk Gevers, who directs the Janssen Human Microbiome Institute


HUMAN MICROBIOME
The community of trillions of organisms that live in, on and around us is sometimes called the forgotten organ.
Approximately 30 trillion human cells 39 trillion bacteria
Changes in the microbiome can indicate disease

Janssen/Erin Dewalt, Nature Publishing Group

(JHMI), is that new types of microbiome-based products could be more natural interventions; they might be used just as disease develops and could potentially deliver more complex and subtle signals to the host.

The JHMI, which is part of Janssen Research and Development, one of the Janssen Pharmaceutical Companies of Johnson & Johnson, is pursuing collaborations with entrepreneurs and with academic and clinical labs as it ramps up microbiome-specific in-house expertise. Gevers was previously at the Broad Institute of Harvard and MIT and worked on the National Institutes of Health Human Microbiome Project, a large-scale project devoted to characterizing the human microbiome's role in health and disease[1]. Products based on microbes cannot be an indiscriminate mix; they will contain specific strains selected for desired qualities and functions, says Gevers. Strain-level insight is only slowly emerging; new analytical technologies are starting to change that.

Who's there?

Amplicon sequencing has long helped researchers analyze samples both of cultivable microbes and of those that fail to grow in the lab. The technique leverages small differences in a distinctive taxonomic marker: the small-subunit ribosomal RNA (16S) gene locus that is highly conserved across microbes. 16S rRNA sequencing remains a "powerful and cost-effective tool" for many experiments, says Segata, yet it is not as useful for studying aspects in which strain-level differences matter. In his view, and especially for microbes that cannot be cultivated in the lab, shotgun metagenomics is "the only way to go for strain-level profiling."

Shotgun metagenomic sequencing lets scientists explore 'who is there'. A sample's jumble of microbial DNA is sequenced and analyzed to shed light on taxonomic identities. A sample from a person's gut inevitably includes bacterial, fungal, archaeal, viral and human DNA, says Segata. Add this to the overall microbial diversity and it's a seemingly intractable analytical challenge. Quince offers a thought experiment to describe how labs assign sequence reads to the strains from which they came (see **Box 1**, 'What the monk saw').

Computational tools, sequencing technology with longer reads, and lower error rates will let scientists move toward comprehensively characterizing all microbes in samples at the strain level from shotgun metagenomes, says McHardy.

What motivates the community is that strain-level identification connects sequencing studies based on human clinical samples to functional testing in animal models, says Julie Segre, an investigator at the National Human Genome Research Institute. She is also part of the Human Microbiome Project, now in its second phase. "You can't take a bunch of base pairs of DNA from a sequencer and use them to colonize a mouse," says Segre. That would not successfully create a

BOX 1 WHAT THE MONK SAW: A METAGENOMIC TALE

Dramatis personae: **libraries:** microbial samples; **book**: a microbial species; **book versions**: strains; **words and letters**: genomic data; **reconstructions**: continguous genomic segments (contigs); **shared colors**: metagenomic binning, which uses shared frequencies across samples in order to assign contigs to species or strains.

A bibliophile medieval king wants to survey Europe's libraries by copying all the books. Some libraries hold many copies of one title, such as the Bible or *Aesop's Fables*. The king seeks out a monk who is a renowned, fast copyist, albeit a little unfocused.

The monk travels far and wide to monastery libraries. He randomly pulls out a book and copies 50 words onto a parchment snippet. He repeats this task a million times, sometimes with the same, sometimes with a different book. He puts the snippets in a sack and travels onward. Over time, he enlists other copyists to help with the copying tasks.

The monk returns home, empties the sacks and sits amid mountains of paper snippets. He matches them up, finds overlap and reconstructs a few longer texts. But he is stymied by identical book sections. He can't tell which book they came from when the copied text on the paper pieces is shorter than the longest, repeated book text segments. There is seemingly no way to join the many partial reconstructions. But he has a plan.

In each library he used parchment of a different hue. When he counts how many times each reconstruction appears in each color, he notices that some reconstructions share a color pattern. He surmises that reconstructions with the same color pattern all came from the same book, as each book has a unique set of frequencies across libraries. He is able to link the reconstructions to yield the book contents, even though the order of the partial reconstructions is still jumbled.

Dorling Kindersley/ Getty Images

As the years pass, the monk realizes the copyists made small errors, even with the Bible. Words were changed, new passages were added, perhaps deliberately. Intrigued, he sets out to find all modified books across Europe. Again, color patterns save the day: he can link changed words together by following the pattern of colored fragments associated with each change. This approach allows him to reconstruct all the variants of an existing book. The monk—now gray, but limber—asks to see the king, who is overjoyed to hear the survey is complete. The king holds a week-long feast in the monk's honor.

Source: Christopher Quince, University of Warwick Medical School.

BOX 2 COMPETITIVE BUT SENSITIVE: TOOL BENCHMARKING IN METAGENOMICS

Alice McHardy from the Helmholtz Center for Infection Research, Alexander Sczyrba at the University of Bielefeld and Thomas Rattei at the University of Vienna launched the Critical Assessment of Metagenome Interpretation (CAMI) competition, and evaluation of the first CAMI challenge is under way. One CAMI category assesses tools that handle assembly, and another assesses taxonomic profiling. A third category assesses binning tools, which assign individual sequence reads into bins that ideally hold strain-level information but might contain other taxonomic categories.

Evaluations are sensitive business. Scientists all too often discover that another lab evaluated their tool in an unfavorable way, says McHardy. They might have used the simplest settings or ones likely to deliver a suboptimal performance. With competitions, developers can submit their tools and explain how they like them to be used. The community can help decide how performance should be assessed.

The team reached out to labs that chose not to participate and inquired about which tool settings to use in the competition. “This was really well received by some,” she says.

When CAMI results are published, McHardy knows some labs might be disappointed by their software’s performance. The plan is to highlight the pros and cons of tools for different questions and scenarios, she says. Her lab has two tools in the competition, and because analysis is anonymized, she does not know how either is faring. “We will see how all of this will work out—in a few months we will know more,” she says.

She hopes that developers will continue to integrate their tools into the CAMI platform’s so-called docker containers. This integration could help to semiautomatically benchmark these software tools in the future. In conversations with colleagues who run the Critical Assessment of Protein Structure Prediction (CASP) she heard that such regular benchmarking accelerated the protein prediction field. “It would be great if we can make that happen for meta’omics as well,” she says.

model of a bacterial infection by a specific strain. Strain-level identification, she says, connects scientists to clinical microbiology and to up-to-date medical knowledge.

Strain-level identification from shotgun data offers many opportunities to explore the functional capacity of strains, the ‘what are they doing’ question. Researchers want to learn which genes underlie traits of interest, says Segre.

Quince is happy about the emerging strain-level analysis options and says that “it has practically made single-cell sequencing redundant.” His software tool CONCOCT takes a pile of short reads, assembles them into fragments and can assign the fragments to the organism from which the genetic information came[2]. Other tools work similarly and, he says, “it is the principle that is more important than any one algorithm.” He has extended this software in *De Novo* Extraction of Strains from Metagenomes, or DESMAN, which infers strains and their abundances using patterns of nucleotide frequencies. The software then determines the nonshared contigs that are unique to individual strains. Such tools help researchers to begin to extract biologically relevant information directly from shotgun metagenome reads. He points to a recent study by Jillian Banfield at the University of California, Berkeley, and her team, as one that shows the strength of combining metagenomic and functional analysis[3]. For example, the group reports finding “unusual biology” in bacteria, including an unusual ribosomal structure and a lack of ribosomal proteins that had been considered universally present in bacteria.

Compute this

Mining metagenomes is computationally demanding, but solutions have emerged and more are coming, says Segata. As it stands, software tools can deliver “noisy,” partially incomplete results. At the same time, he says, “it is just fantastic that this is possible at all.”

During his postdoctoral fellowship in the lab of Curtis Huttenhower at the Harvard School of Public Health, Segata developed MetaPhlAn and MetaPhlAn2, which leverage a genetic signature to identify and track species, and in many cases strains. In practice, that can mean following around 200 genes per species, and a sample might contain thousands of species, such that the tools dig into around one million genes. But, he says, “it’s like distinguishing lions and tigers from their footprints: it works, it’s great, but it does not tell you much about the biology of lions and tigers.”

More recently, he and his Trento team, along with colleagues at the University of Massachusetts Medical School and the Perinatal Institute in Ohio, have developed Pangenome-based Phylogenomic Analysis (PanPhlAn)[4], which he finds more powerful given that it reveals the full gene repertoire of strains.

The pangenome is a catalog of genes for all sequenced strains in a species of interest, says Segata. It’s built by extracting genes from available reference genomes and then binning them in gene-family clusters. This is how the team calculated pangenomes for over 400 species. PanPhlAn characterizes strains by identifying which combinations of genes, from the pangenome of each species, are found in a sample.

Segata’s team tested the tool and discovered, for example, that some tested Chinese individuals harbor a set of *Eubacterium rectale* strains that are genomically distinct from the same species found in the gut of tested people in Europe and the United States. Segata believes many more such patterns can be found in metagenomic data, including some that correlate with disease risk.

Using PanPhlAn, researchers can also obtain transcriptional profiles of a microbial community. Knowing which genes are present in a sample speaks only to potential organismal functions; a bacterium with a virulence gene that is not expressed might not be a troublemaker. PanPhlAn, with its analysis of DNA and RNA, gives a fuller view of a microbial community, says Segata.

A computer scientist who gravitated toward biology, Segata is at his university’s Center for Integrative Biology. One-third of his team does wet-lab work, and the others are

Helmholtz Center for Infection Research

Being a tool developer in metagenomics can become fun again and more productive, says Alice McHardy.

Alessio Coser, University of Trento/ GIANTmicrobes www.giantmicrobes.com

In the Segata group, one-third of the team does wet-lab work and the others are computational scientists focusing on biology.

computational scientists working on biological problems. He integrates the group for discussions and tool testing. A different kind of integration—tool integration—will, Segata hopes, begin in the metagenomics community. Integration means, for example, having on hand data standards for tool input and output. But such standards are hard to establish in a rapidly changing field, and, he says, grant funders sometimes do not view tool-integration projects as favorably as projects about new tools.

Another community issue is tool benchmarking. McHardy says she and many other labs spend much time on performance benchmarking. She began her metagenomics journey developing tools—for example, PhyloPythia, which is a supervised taxonomic binner. Software tools in this class have exploded, she says, and the developer community has grown, too. She counts as many as one new metagenomic analysis tool published per month in 2015. The situation can be confusing when labs use different benchmark settings or data sets that might seem artificial. Along with colleagues, she created a community approach to tool benchmarking (see **Box 2**, 'Competitive but sensitive'). The hope is that analysis results will also be representative of real-life questions, she says. "I think being a tool developer in metagenomics can become fun again, and more productive," she says.

Janssen

Understanding an ecosystem, such as the human gut's bustling microbiome, at high resolution is the next big challenge, says Dirk Gevers.

Along the way to better tools, the community has some fundamentals to discuss, such as the definition of a strain, says Segata. For example, two organisms might differ on average in one nucleotide every 50,000 bases, and researchers will disagree on whether these two organisms are different strains. The number of such differences is likely to vary depending on the species. There is no threshold with biological meaning, he says, that defines the limit between 'same strain' and 'different strain.'

A long view

High-resolution analysis in metagenomics is a tall order, but strain-level insights are promising, says Gevers. This view can show organisms to be stable residents of an environment for several years or longer. Within a habitat, different species are competing, and so are different strains of the same species. The unit of microbial action is a strain, not a species, he says. Being able to differentiate between different strains matters greatly when scientists work on ways to intervene with a microbial community by introducing new strains, he says.

The next big challenge is understanding the dynamics of an ecosystem, such as the bustling microbiome in the human gut, and doing so at high resolution, he says. Dissecting a community into its different strains, understanding the differences in genetic content and knowing which strains are expressing which genes or producing which metabolites, secreting which peptides or proteins, "will be eye opening," says Gevers. Spatial distribution might also matter, he says. Are strains one big mix, or, he asks, is there a " 'method to the madness' "?

Emerging techniques reveal the many different strains of a particular species in a sample, strain stability across different samples and, ultimately, differences in genetic or functional composition within a species across samples. Gevers looks forward to developments that link complete genetic content to each strain and measure each strain's activity. He is hopeful about progress beyond capturing the genetic composition of a microbial community, bringing strain-level analysis and resolution to RNA, metabolite and protein levels.

Segata and many of his colleagues hope sequencing costs will continue to drop, allowing labs to improve the sequencing depth with which they probe microbiomes. Matters get complicated when there are closely related strains in a sample, or when strains are present in amounts too low to detect or below the level of sequence coverage needed for their reconstruction. Such challenges motivate him and others to keep refining computational methods.

1. The Human Microbiome Project Consortium *Nature* **486**, 207–214 (2012).
2. Alneberg, J. *et al. Nat. Methods* **11**, 1144–1146 (2014).
3. Brown, C. *et al. Nature* **523**, 208–211 (2015).
4. Scholz, M. *et al. Nat. Methods* **13**, 435–438 (2016).

Vivien Marx is technology editor for *Nature Methods* (v.marx@us.nature.com).