Source code:

Putting metagenomics to the test in the clinic

By Carrie Arnold

By July 2013, pediatrician James Gern had diagnosed hundreds of children at the University of Wisconsin Hospital in Madison, with ailments ranging from prosaic infections such as strep throat to emerging diseases such as West Nile virus. But one patient, a 14-year-old boy with an inherited immunodeficiency condition, who had been in the hospital for 32 days with encephalitis, stumped him. Three months before, the boy had complained of headaches and fever, which prompted a visit to Gern and a prescription of a steroid (prednisone) to reduce swelling, as well as an antibiotic (ciprofloxacin). But his condition continued to deteriorate. After intense seizures began wracking his thin frame, he was hospitalized. A brain biopsy failed to reveal a cause, and doctors placed the boy in a medically induced coma to halt the unrelenting and intensifying seizures.

"If we didn't figure out what was wrong and get him treatment, I knew his infection would likely be fatal," Gern says.

Gern contacted his collaborators Joseph DeRisi and Charles Chiu, microbiologists at the University of California, San Francisco (UCSF), to tap into their expertise. DeRisi and Chiu had been waiting for this type of phone call. They had developed a new platform that could be used for infectious diseases that defied diagnosis with standard protocols—perfect for Gern's patient. Instead of testing a sample of cerebrospinal fluid for one or two pathogens at a time, as Gern had been doing, the UCSF team used a technique called metagenomics to sequence all of the DNA in Gern's sample in one go. Software called sequence-based ultra-rapid pathogen identification (SURPI) analyzed the results and compared the DNA sequences in the sample to those found in publicly available genome databases. Within 48 hours, the UCSF platform, termed Precision Diagnosis of Acute Infectious Diseases, discovered the causative organism: a bacterium called *Leptospira santarosai*, which the patient had acquired on a trip to Puerto Rico the year before[1].

"We hadn't thought to look for *Leptospira*," Gern says, "but as soon as we started high-dose penicillin, he rapidly improved."

Over the past decade, metagenomics has freed microbiologists from the time- and labor-intensive need to culture organisms in a dish to identify them. The technique has opened new doors on efforts to catalogue and study microbes in the soil, air and water, giving scientists the ability to study prokaryotes that won't grow in the lab. It has also paved the way for the Human Microbiome Project and other efforts to map the range of commensal microbes growing in and on our bodies.

In the past few years, however, a small group of scientists has pioneered the use of metagenomics for the diagnosis of infectious disease. In March, Chiu and his team at UCSF launched a major clinical trial of a metagenomics diagnostic test for encephalitis and meningitis, which they say will forever change infectious-

disease diagnosis. Half of the thousands of Americans diagnosed each year with one of these conditions never learn what pathogen is causing their illness. Chiu and colleagues intend to change this by sequencing genetic material from pathogens in cerebrospinal fluid. Other commercially available metagenomic tests using blood samples have also entered the market in the past two years.

Their ambitions, however, must contend with the realities of DNA sequencing. To make a diagnosis, Chiu's team must compare a patient's results to a large DNA database of microbial sequences. The largest such database, RefSeq, which is run by the US National Institutes of Health (NIH), contains whole-genome sequences of only 5,000 bacteria, out of a total of 150,000–200,000 known bacterial species. If the pathogen identified isn't one of the 5,000 fully sequenced organisms, scientists won't have the genetic data required to definitively trace the source of infection or determine whether it carries antibiotic-resistance genes—two crucial pieces of information for physicians. Yet, sorting through the hundreds of thousands of pathogen sequences needed to overcome these hurdles could create an avalanche of data, which threatens to overwhelm clinicians. If this mountain of information buries researchers, it will hinder their ability to rapidly identify the cause of a patient's disease.

Still, say Chiu and others, the confluence of improved data analytics and advances in genetic-sequencing technologies within the past several years has helped scientists to begin to dig out of the data deluge and parse them accurately enough to distinguish one microbe from another. This, in turn, has pushed metagenomic diagnostic tests from an aspiration into an increasingly workable diagnostic strategy that has begun to emerge in commercial applications. "This changes the entire diagnostic paradigm. Instead of detecting targets a priori, you can go in with an open mind," Chiu says.

Capturing more

Since the mid-nineteenth century, microbiologists have studied bacteria by growing them in cultures. Until the early 2000s, when researchers developed the ability to study microbes on the basis of their DNA alone, most of the tiny, single-celled organisms on Earth remained undiscovered by science. Ten years ago, the advent of next-generation sequencing enabled scientists to step away from the world of culture flasks and Petri dishes. Metagenomics allowed microbiologists to sequence the DNA of numerous microorganisms in a sample of soil, a droplet of rain or even a swab of the skin's surface.

At first, researchers cataloging the range of living organisms in a sample sequenced only one specific gene, known as 16S rRNA, in bacteria and their prokaryotic cousins, Archaea. This provided a broad census of what organisms were there, but to figure out what these microbes ate and how they survived, scientists needed to look at their entire genomes. However, older DNA-sequencing technology could read a genome only in 100–300 base-pair chunks. Samples often had tens or hundreds of microbial species, and these sequencing reads were jumbled together, as if pieces from different jigsaw puzzles had been tossed together on the floor. Sorting these pieces into their requisite puzzles and then assembling them into a final picture—a process known as binning—required even more brute-force computing power. But better algorithms and software developed within the last five years have greatly simplified the process.

Formally moving metagenomics from the study of the world around us into the clinical realm presented researchers with several hurdles. The 16S rRNA sequences could identify microbes at the genus level, but often could not identify individual species or strains. Given that pathogens such as anthrax-causing *Bacillus anthracis* had 16S sequences that were nearly identical to those of harmless *Bacillus* species, this distinction was crucial. Making this distinction required sequencing the entire genome. Scientists also needed to speed up the time-intensive process of sequencing and data analysis. Physicians wanted answers within hours or days, not weeks or months, explains Alexandra Trkola, a virologist at the University of Zurich's Institute for Medical Virology in Switzerland.

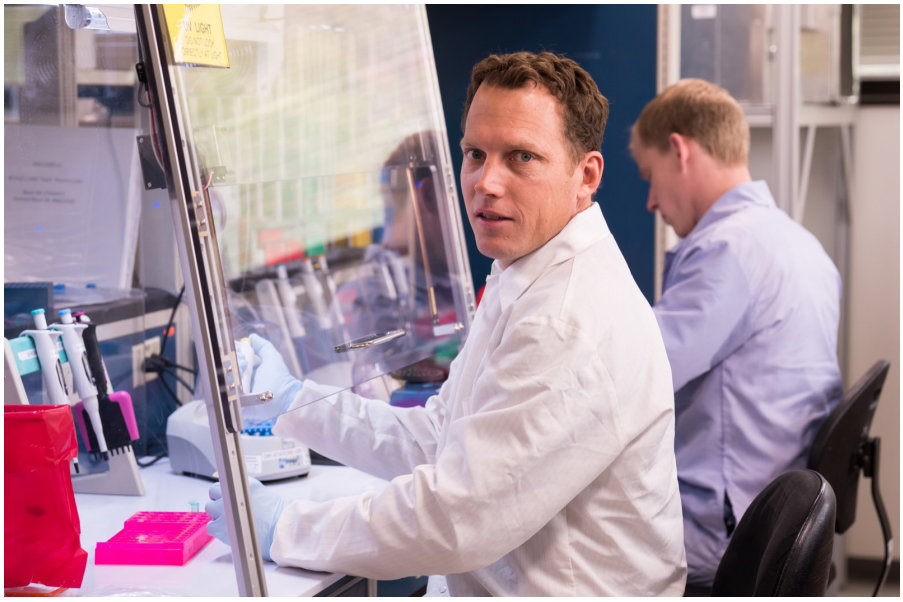
One of the first diagnostic tests of metagenomics took place in 2013, when microbial geneticist Nick Loman at the University of Birmingham, UK, and his colleagues wanted to see whether the strategy could identify the cause of a large foodborne

Michael Huber

Prep step: Processing samples for study.

Luke Tallon

Data drive: FDA-ARGOS's Luke Tallon, Lisa Sadzewicz and the PacBio Sequel System.



Robert Schlaberg

The tax man: Robert Schlaberg helped develop the Taxonomer.

outbreak in Germany[2]. Conventional microbiology had identified the organism as an unusual strain of *Escherichia coli*, but the process took months. Using metagenomics, Loman's team identified the bacterium in under a week. In 2015, Chiu's group at UCSF used the technique to decipher the cause of an unknown infection that was affecting an individual in the UK who had received a bone marrow transplant. Within 96 hours of receiving the sample, the researchers identified the pathogen as an astrovirus[3].

As their confidence in the potential of metagenomics as a diagnostic tool grew, scientists also had to grapple with the issue of microbial contamination in their samples. Microbes are everywhere—on the body, in the environment, even in the reagents used for sequencing. Metagenomic DNA amplification methods are so sensitive, Chiu said, that they can pick up even minute snippets of DNA from technicians that have contaminated the reagents. Once sequenced, these contaminants can seem as if they came from the patient sample. Running large numbers of control samples and identifying sequences that appear without patient samples helps SURPI to recognize potential contaminants. It's also one of the reasons Chiu and his team began testing SURPI in patients with meningitis or encephalitis, because samples of cerebrospinal fluid are more sterile than stool or respiratory fluid. They are now beginning to test SURPI on blood samples.

One of several companies looking to offer blood diagnostics using metagenomics is a startup based in Silicon Valley, California, called Karius, which offers a Clinical Laboratory Improvement Amendments (CLIA)-certified test. The company looks at cell-free DNA in the bloodstream to diagnose infections. "We can find fragments of DNA, not from the host but from a microbe," explains Karius cofounder and CEO Mickey Kertesz. But getting there required the development of a proprietary data-analytics and molecular-biology platform that could sift through the overwhelming amount of host DNA to identify the genetic snippets from invading pathogens, he adds. "Getting that signal from the noise took two and a half years of work."

Dynamic databases

A metagenomics diagnostic test, however, requires more than just finding the proverbial needle in a haystack. The test also must identify the species and strain of pathogen to which the DNA belongs, a process that requires comparing the genetic sequence to a reference database to find the culprit. A pilot study of metagenomic diagnostics for brain infections illustrates the need for dynamic databases. Of the six patients tested, researchers at Johns Hopkins University identified culprits in five cases. When they checked their results one year later, they got a hit for the sample that had previously come up as an unknown. A newly discovered bacterium called *Elizabethkingia meningoseptica* had caused the sixth patient's disease. When the researchers first ran the test, it hadn't yet been sequenced and entered into a database such as the NIH's RefSeq.

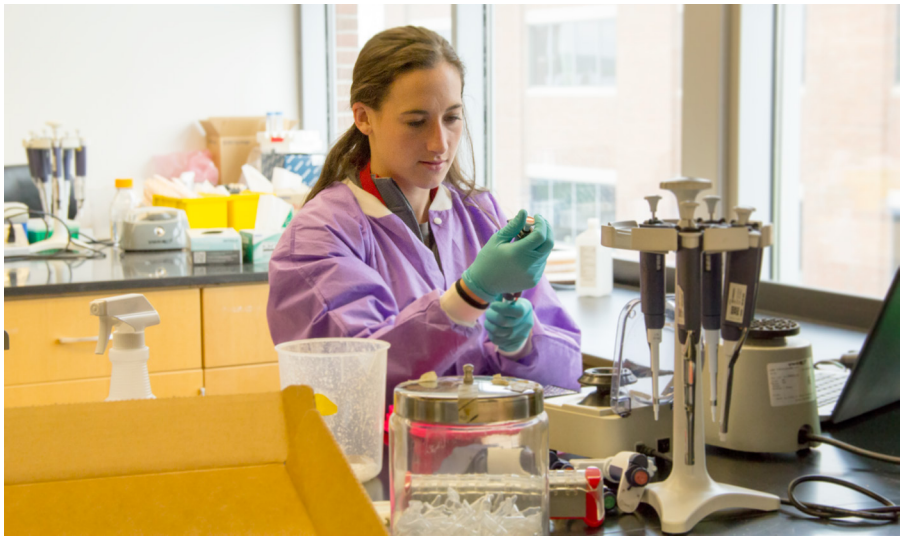
In databases such as RefSeq, more than 97% of species are identified by only a single gene sequence, such as 16S rRNA. This information lacks the detail and specificity needed for many clinicians to make a diagnosis, says Charles Langelier, an infectious-disease physician at UCSF.

With sequences for 400 million genes, the NIH's GenBank is the largest DNA database in the world. But repositories such as this aren't perfect. Their strength—namely, that anyone can deposit sequences—is also a weakness, because not all sequences are accurate and well annotated. This becomes especially problematic when GenBank contains only one DNA sequence for a particular organism, given that it hasn't been verified with genetic data from other microbes from the same species. "If you don't have a database that's well curated and accurate, you will generate unreliable data," says Chiu. "You don't want to make a diagnosis based on one entry."

In 2013, the US Food and Drug Administration recognized that approving metagenomic and other molecular diagnostic tests would require a better reference database with which to verify results. An effort led by Heike Sichtig, a researcher at the agency, launched the FDA–ARGOS (FDA dAtabase for Reference Grade micrObial Sequences) in May 2014. The database currently contains 2,000 complete bacterial sequences, and its curators—microbiologists at the FDA—plan to add several thousand more over the next few years, on the basis of data from patient infections as well as the needs of test developers. "To get the right diagnosis, you need the right data. And if you don't

Michael Huber

Going viral: Sequencing at the University of Zurich.



Heike Sichtig

Sequence of events: Holly Rousey of the FDA-ARGOS team prepares samples.

have the right data, you can't make that call," Sichtig says.

Making a match

Each metagenomics platform uses its own, generally proprietary, algorithm to search quickly through the heaps of sequences in available databases. Initial metagenomics studies had to search through DNA databases gene by gene—one of the reasons why those scientists had sequenced only the 16S rRNA gene. Identifying a pathogen using this strategy would take days or weeks, making it impractical for clinical use. Even when not sifting through millions of genes one at a time, many metagenomics platforms assemble a microbe's entire genome piece by piece, which still requires hours, according to Crystal Icenhour, cofounder and CEO of Aperiomics, a next-generation sequencing company in Ashburn, Virginia.

Instead of painstakingly stitching a genome together from small fragments, Aperiomics has developed a strategy that prioritizes unique information found in sequencing data and assigns priority to the analysis of these sequences in their samples. By focusing on these data-rich segments of the genome, Aperiomics can sidestep the time-intensive need to compile an entire genome. "As long as we get the right piece, we don't need to put together the whole puzzle," Icenhour says.

Even with the best database in the world, making that call isn't always straightforward. A microbial sequence from a patient sample almost never matches a reference sequence with 100% accuracy, which is why most metagenomics labs have a full-time bioinformatics specialist to help determine how closely sequences need to match to be confirmed as the same species or strain. Clinical labs, however, don't have this in-house expertise, which limits their ability to interpret metagenomic tests.

"Someone needs to be able to quickly assess the quality of sequencing data, and have validated cut-offs for how close is close enough to make an ID," says Robert Schlaberg, a microbial geneticist at the University of Utah in Salt Lake City.

With this goal in mind, Schlaberg has helped to build an application called Taxonomer, software that automates some bioinformatics tools. The online interface provides results for physicians, similar to other clinical tests, and classifies 99.6% of test reads in nine minutes—a feat that would take SURPI four hours—according to 2016 results published in *Genome Biology*[4]. When Schlaberg and colleagues combined Taxonomer with RNA sequencing, they report in a study published in the *Journal of Clinical Microbiology*[5], they were able to gather more information about respiratory pathogens than an existing commercial test would have allowed. They also identified 12 viruses with sequences that differed too much from any in the database to be recognized by existing molecular tests. The software doesn't eliminate the need for expert medical interpretation, but it does help to reduce the need for bioinformatics specialists. In May 2016, Schlaberg and colleagues announced that they had licensed Taxonomer to the bioinformatics startup IDbyDNA for use in its diagnostic tests.

> "This changes the entire diagnostic paradigm."

"Managing these novel and divergent pathogens, where they don't quite match previous sequences, is going to be a challenge for metagenomics. It really takes an expert to figure out what's going on," Loman says.

At present, metagenomic tests are more expensive and time-consuming than other types of molecular tests. Although universities and private companies have started to roll out metagenomic diagnostics, Gern says, they are most useful once doctors have exhausted all other options.

The advent of smaller DNA sequencers, such as Oxford Nanopore's MinION, promises to bring metagenomic tests to a patient's bedside, whether in a well-equipped research hospital or a rural clinic. To Schlaberg, however, the most revolutionary aspect of metagenomic testing may be in how it forces scientists to rethink the nature of infectious disease itself. Since Louis Pasteur and Robert Koch pioneered the field 150 years ago, microbiologists have viewed infectious diseases as caused by a single microbe. Existing diagnostic tests continue to confirm this line of thought; physicians stop testing once they get a positive result. "But not all infections may be caused by a single microbe acting alone. There may be a whole host of organisms contributing to [a patient's] disease," Schlaberg says.

Because metagenomics sequences nearly all of the DNA detected in a sample, researchers may finally be able to identify the panoply of pathogens at the root of disease. Multi-microbial infections could explain why some types of bacterial disease seem to withstand treatment. No one knows how to treat these infections, if treatments exist. It's possible that double doses of existing antibiotics might work, or sequential treatment of existing drugs. But perhaps, newer, better antimicrobials are needed. "It's a whole new way of thinking about infectious disease," Schlaberg says.

Carrie Arnold is a freelance science writer based in Virginia.

1. Wilson, M.R. *et al. N. Engl. J. Med.* **370**, 2408–2417 (2014).
2. Loman, N.J. *et al. JAMA* **309**, 1502–1510 (2013).
3. Naccache, S.N. *et al. Clin. Infect. Dis.* **60**, 919–923 (2015).
4. Flygare, S. *et al. Genome Biol.* **17**, http://dx.doi.org/10.1186/s13059-016-0969-1 (2016).
5. Graf, E.H. *et al. J. Clin. Microbiol.* **54**, 1000–1007 (2016).