Metagenomic Sequencing Identifies a Pathogen that Other Technologies Missed

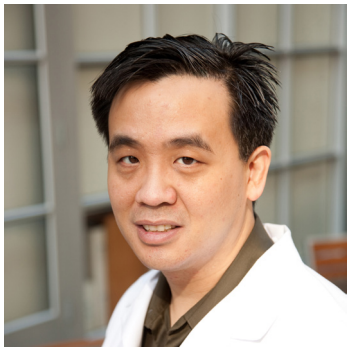
The MiSeq® system and a new bioinformatics pipeline enable researchers to identify a *Leptospira* strain that was overlooked using conventional methods.

Introduction

Charles Chiu, M.D., Ph.D. is a pathogen hunter. As Director of the UCSF-Abbott Viral Diagnostics and Discovery Center, he understands how the complexity and diversity of bacteria and viruses enable them to evade routine clinical detection. Dr. Chiu and his team focus on out-maneuvering pathogens by advancing metagenomic-based technologies to identify and characterize them rapidly. His team's objectives are clear: apply unbiased sequencing methods to identify emerging pathogens and develop these strategies into the next generation of infectious disease diagnostics.

The team's early work with targeted microarrays has expanded to include next-generation sequencing (NGS) systems that offer a whole-genome view of the pathogenic landscape. Dr. Chiu's team is using the MiSeq system to conduct its research, leveraging the desktop sequencer's ability to perform rapid sequencing with high accuracy. They've developed a sequence-based ultrarapid pathogen identification (SURPI) bioinformatics pipeline to analyze results in an actionable timeframe[1]. Recently, SURPI and the MiSeq System were put to the test, generating and analyzing sequencing data that identified the presence of a pathogen that had eluded all conventional testing techniques[2].

iCommunity spoke with Dr. Chiu about SURPI and the development of NGS as a potential tool for pathogen detection.



Charles Chiu, M.D., Ph.D., is Director of the University of California, San Francisco (UCSF)-Abbott Viral Diagnostics and Discovery Center, Associate Director of the UCSF Clinical Microbiology Laboratory, and an Assistant Professor of Laboratory Medicine and Medicine, Infectious Diseases at UCSF.

Q: What does metagenomic sequencing offer that microarrays can't provide in pathogen detection?

Charles Chiu (CC): Unbiased metagenomic sequencing offers a coverage depth that you can't get with microarrays. Unlike microarrays, which employ specific probes to detect a defined set of targets, metagenomic sequencing doesn't target individual genes or pathogens. Instead, it enables interrogation of all the DNA present in a sample at one time.

Sequencing also provides information that is as specific as you can get for purposes of microbial identification. While microarrays reveal a pattern of hybridization intensities depending on which probes light up, NGS yields sequences that you can then assemble into longer contigs (contiguous sequences) or even use to recover complete genomes. In addition to providing a much greater breadth of information, metagenomic sequencing also delivers more specificity. That becomes very important when we're trying to find a needle in a haystack, which is the case with a complex sample. We're trying to identify the small subset, usually a tiny fraction of the raw sequence reads that may correspond to pathogens. We need to be able to do that in a rapid fashion to yield results that are actionable in infectious disease. For this method to be a competitive microbiological assay, the sample-to-answer turnaround time must be within hours as opposed to days or weeks.

Q: What was the inspiration behind the development of SURPI?

GL: Traditional algorithms for performing metagenomic analysis—the classification of microbes, or more specifically pathogens in the diagnosis of infectious disease--are too slow for clinical diagnostics. The gold standard BLAST algorithm has been around since the early 1980s, and many algorithms out there for pathogen detection still use BLAST as the computational backbone. This algorithm is incredibly insensitive, very slow, and is not tractable computationally for most clinical laboratories.

We realized we could develop a better approach by taking advantage of two algorithms that were publicly available as open-source. A very talented pathology resident in my laboratory, Dr. Taylor Sittler, in collaboration with groups at UC Berkeley and Microsoft Corp, designed a nucleotide

aligner called SNAP that is about 1,000–10,000 times faster than BLAST. We combined SNAP with in-house developed software and other available open-source algorithms to create SURPI. SURPI can analyze MiSeq system runs in 10 minutes to a few hours, depending on the type of sequencing and analysis being performed. SURPI can also be run on a single computational server or even in the cloud. That's tractable in a clinical laboratory setting and was the motivation in creating SURPI.

> "There are many patients where the pathogen can't be identified despite extensive conventional clinical testing."

Q: How does SURPI identify pathogen DNA in the midst of all the metagenomic sequencing data?
GL: SNAP was written originally to align human genomes for resequencing applications in cancer and genetic testing. We modified and customized it for the extensive metagenomic databases provided by the National Center for Biotechnology Information (NCBI) that include sequences from pathogens and hosts. The SURPI pipeline uses SNAP to first identify and then subtract out human host sequences. It then aligns reads to reference sequences in NCBI databases, including GenBank, for comprehensive identification of all microorganisms: bacteria, viruses, fungi, and parasites. We added several programs to enable this data analysis to be performed in parallel.

Q: How long did it take you to develop SURPI?
GL: Developing the pipeline, and getting it validated, published, and made publicly available took about a year and a half.

Q: Why did you choose to use the MiSeq system to generate the metagenomic sequence for SURPI analysis?
GL: UCSF began using Illumina sequencing systems starting with the Genome Analyzer™ system and later upgraded to a HiSeq® system. When the MiSeq system became available, we migrated over to it for two reasons. It's a portable, desktop sequencer that's small enough to fit into a laboratory. The MiSeq system also generates a sizeable number of sequences (10–30 million) in 6–24 hours, a turnaround time that is suitable for infectious disease investigations.

Q: What library preparation kit did you use for this research study?
GL: We used TruSeq® kits, but recently switched to Illumina Nextera® library prep kits for this type of analysis. That's because Nextera kits are much faster.

Q: How did the MiSeq system perform in the study?
GL: I think the MiSeq system performed very well, producing 150 base pair (bp) reads that enabled us to make an unambiguous identification. Changes were made from the normal MiSeq system protocol so we could perform single-read sequencing, immediately pull the data from the instrument, and begin the analysis while the second paired end was being generated. By interrupting the sequencing run, we obtained data faster and that was critical. We're now working with Illumina to see if we can simultaneously analyze data in real time as it is generated by the instrument.

Q: What were the results of metagenomic sequencing and SURPI data analysis?
GL: In a cerebrospinal fluid (CSF) sample, we identified 475 of 3,063,784 sequence reads corresponding to Leptospira infection. This was a very convincing result because the sequences spanned the entire genome and *Leptospira* was the only credible pathogen DNA that was detected. When we sent the sample off for confirmatory testing to the CDC, it initially came back negative by antibody testing and the gold standard PCR for *Leptospira*. We later showed that the reason the CDC PCR results were negative was because the PCR assay they used had not been fully validated with that particular *Leptospira* species, *Leptospira santarosai*, and thus was not sensitive enough to detect the bacterial pathogen.

So it turned out that NGS was probably the only way the pathogen could have been identified, at least in an actionable timeframe. No clinical test available at the time would have been able to make the diagnosis, even if leptospirosis infection had been considered *a priori*.

Q: How much faster was the MiSeq system/SURPI method than current technologies?
GL: For this particular case, the sample-to-answer turnaround time was 48 hours. After the case report was published, we've decreased the turnaround time to less than 24 hours. We are ultimately aiming for an 8-hour turnaround time to make the NGS test fit within a single laboratory shift and be competitive with other molecular tests that are now performed routinely (multiplex PCR tests for virus detection, mass spectrometry-based methods for multiplex detection, and specific PCR-based tests to identify unusual organisms).

Q: What tests had been performed on this patient before enrolling him in your research study?
GL: The patient had been hospitalized three times over 4 months. Physicians had ordered an extensive infectious disease workup, including diagnostic laboratory testing, MRIs, CSF analysis, and a brain biopsy. All testing results were negative, although the spinal fluid and brain biopsy profiles showed prominent inflammation and strongly suggested the possibility of an infection.

> "I think the MiSeq system performed very well, producing 150 base pair (bp) reads that enabled us to make an unambiguous identification. "

Q: How often are physicians faced with cases of encephalitis, meningitis, fever, etc. where they can't determine the cause?
GL: There are many patients where the pathogen can't be identified despite extensive conventional clinical testing. In any intensive care unit in the country about 20 to 30% of all pneumonias and basic lung infections, and 20% of unknown fevers are not diagnosed. Over 50% of encephalitis cases in the U.S. are undiagnosed. It's thought that many of these cases are probably infectious, yet we are unable to make a diagnosis due to limitations in existing tests.

Q: What are your next steps in advancing metagenomics sequencing into the clinical laboratory?
GL: We're focusing on validating the method in the clinical laboratory for a number of applications related to infectious disease. We hope to start with the identification of viral and bacterial cultures. That's a big need in the clinical lab. Often, we grow up something in culture, either a bacterium or a virus, and we don't know exactly what it is. The secondary testing we have to perform to identify the pathogen can be very costly and time consuming. A metagenomics method that identifies the bacteria or virus immediately would be useful clinically.

From there I think that we are targeting three specific diseases: unexplained meningitis/encephalitis, pneumonia in the intensive care unit, tick-borne illnesses, and other febrile disease such as sepsis. Part of what makes the NGS approach attractive is the fact that it potentially has utility across many different types of applications.

Q: What applications can metagenomic sequencing be used for beyond infectious disease testing?
CC: There are a number of public health applications, such as outbreak investigation and pathogen or disease surveillance. We're working with the California Department of Public Health and the CDC on developing those applications.

I'm also working with the American Red Cross on developing a blood bank screening application. To ensure the safety of blood transfusions, multiple tests have to be performed individually to detect a broad range of pathogens. A single test that is sufficiently sensitive and accurate could be used to screen blood and prevent pathogens from being transmitted through blood transfusions, which would be very useful from a public health perspective.

Q: How important is having updated reference databases of bacterial/viral genomes for the success of this method?
CC: As you might expect, GenBank is a "sequence zoo". There are lots of sequences being sent in from many different sources and unfortunately some of them are not well-annotated or even misannotated. Using GenBank as a reference becomes problematic, especially for clinical testing. We're currently working with the NCBI and FDA to establish a clinical reference database for pathogens. Our hope is that we can generate a highly curated, well-annotated microbial reference database that can be used specifically for NGS sequencing applications.

Q: What needs to happen before NGS is used routinely in the clinical laboratory?
CC: One of the barriers to widespread use of NGS in the clinical laboratory is the computational or bioinformatics challenge. SURPI and other software programs are only beginning to meet this challenge.

Several other challenges remain, including issues of clinical validation and regulatory approval. The question is how the FDA will respond to such a highly multiplexed assay. With metagenomic NGS, we're looking at a huge, if not infinite number of targets. We're having discussions with the FDA about this now. It's something to resolve before we can see NGS methods being used routinely in the clinical laboratory.

There are also the issues of reducing NGS testing cost, turnaround time, and the computational burden of data analysis. With solutions such as the MiSeq System and SURPI, I think we can solve the first two issues. However, for the third issue, it will be critical for us to design a user interface for SURPI that's easy to use for non-bioinformaticians, including clinicians, lab personnel, and lab directors. We are working on the design of a simple user interface that could be integrated into HIPAA-protected electronic medical record (EMR) platforms in hospitals and accessible from the web through a graphical user interface.

These challenges are not insurmountable. I think microbiology has lagged behind cancer and genetic testing, but I don't think it's going to be long before we start seeing an NGS-based approach being used on a widespread scale for infectious diseases.

> "There's definitely a strong clinical need for fast NGS data analysis, especially in the areas of pathogen detection and infectious disease."

Q: What changes are you going to be making to SURPI in the future?

CC: We're working on optimizing SNAP and the algorithms used in SURPI to produce even faster turnaround times. Ultimately, I'd like SURPI to be capable of aligning sequence data to every single read in GenBank, and to perform that within a few minutes.

Since this research study was completed, we've made SURPI widely available. Several groups, including the CDC, have downloaded it and are now testing it. We're hoping to gain additional feedback about how to improve it. We want it to become a dynamic piece of software.

There's definitely a strong clinical need for fast NGS data analysis, especially in the areas of pathogen detection and infectious disease. It's pretty amazing that we might be able to help patients by using this technology.

References

1. Naccache SN, Federman S, Veeraraghavan N, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res 2014: 24: 1180–1092.
2. Wilson MR, Naccache SN, Samayoa E, et al. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med 2014: 370: 2408–2417.

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY.

illumina®