Better Metagenomics Methods Offer Hope for Studies with Ultra-Low DNA

For low-input DNA studies such as viral metagenomics, there is no room for error in library protocols. Scientists at the University of Arizona have optimized methods for sequencing sample prep — and emphasize the need for efficient, precise size selection.

Behind almost every great biological discovery is a league of scientists who spent years honing methods, testing technologies, and polishing protocols to build a solid foundation upon which important and reliable research can happen.

Matthew Sullivan knows this better than most. An assistant professor in the departments of Ecology and Evolutionary Biology as well as Molecular and Cellular Biology at the University of Arizona, Sullivan is a biologist whose research centers on viral genomes in ocean environments. His lab is currently studying thousands of ocean samples taken from global expeditions to elucidate the viral genomes in different locales, such as oxygen minimum zones.

But as a participant in the nascent field of metagenomics for viruses, Sullivan has seen firsthand how desperately the community needs more robust, quantitative protocols to ensure that the findings being made are meaningful. That's why he has spent the last three years establishing a sample-to-sequence pipeline — putting each step of the viral metagenomics process through its paces, figuring out ways to make each protocol as reliable and reproducible as possible.

"We test out all these things, from concentration and purification of the viral particles to the molecular biology of DNA sample preparation," Sullivan says. "I firmly believe it's fundamental to rigorous science, and the field needed a quantitative baseline." This is a technical tradition required for any new scientific field. "At some point we stop doing descriptive, qualitative work and get rigorously quantitative and eventually predictive," he adds.

"This means that you can make a library from 0.01 nanograms and be as comfortable with it as you are with 10 nanograms," Sullivan says. "That was both surprising and comforting, since we are often down in the 0.01 nanogram range."

Biosphere2, a 3-acre self-contained research facility in Arizona, includes a 6,000-square-foot ocean that Matt Sullivan uses to hone metagenomics methods.

The Sullivan lab has tested and improved a number of steps in viral metagenomics work; a paper published in the September 2012 edition of *Environmental Microbiology* reports on optimization of the linker amplification method. (Duhaime et al., "Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method," available with open access.) The paper focused on ways to reduce amplification bias, including precise DNA size selection, enzyme choice, and optimizing cycle number.

Call of the Ocean

Sullivan has a self-professed "passion for the ocean," so it's no surprise that his biological studies focused on marine work. He became a seaweed ecologist and it was during a research stint at the Scripps Institute of Oceanography that he originally worked with phages, which put him on the path to virology.

That path would not be an easy one. Studying viruses in the ocean offers any number of challenges, the most obvious being getting enough of a sample to study. While there may be a million viruses living in 1 milliliter of sea water, it could take 20 liters of that same water to yield just 1 picogram to 1 nanogram of viral DNA.

Then, of course, there's access to ocean samples — which wouldn't be easy for a landlocked researcher in Arizona. Fortunately for Sullivan, though, he has the next best thing: the Biosphere. Built in the late 1980s in Oracle, Ariz., and taken over by the University of Arizona in 2007, Biosphere2 is a unique, self-contained research facility meant to mimic various biomes from all over the world. The 3-acre structure is used to study climate change and other forms of ecology.

Part of Biosphere2 is a 6,000-square-foot ocean, originally started with more than 150,000 gallons of water taken from the Pacific Ocean near Scripps Pier in La Jolla, Calif, and mixed with local water treated to introduce qualities similar to sea water. Other elements were contributed as well: corals from the Caribbean, fish from Hawaii, rocks from the desert. While Sullivan's biological research studies require samples from a natural ocean, his methods work does not. "All I needed it to be was complex, and the Biosphere ocean is a complex community," he says. "At the level of microbes and viruses, it actually looks a lot like southern California water." For improving metagenomics techniques, the local ocean was just what Sullivan needed.

The Question of Bias

In this latest effort, Sullivan and his team homed in on ways to reduce amplification bias for cases where available DNA is minimal. They started with the linker-amplified shotgun library (LASL) approach, adapting it for next-generation sequencers to boost throughput. The two areas targeted as most important in reducing bias were the size selection and amplification processes.

Size selection was critical because PCR is known to preferentially amplify shorter reads, skewing what's represented in the final library compared to the starting material. "If you can minimize the size variance, then that's one of the PCR biases you've just controlled," Sullivan says. The other reason to focus on size selection was the overwhelming need for good sample recovery in projects that might have less than a nanogram to begin with. "If you enzymatically digest gel slices, you lose tons of DNA — on the order of 70 percent," he adds. Automated size selection was a promising alternative.

"Of the three sizing fractionation methods tested for target recovery efficiency (fraction recovered DNA in target 400–600 bp size range), the Pippin Prep was the most efficient and reproducible (94–96% of input DNA), with the tightest, most specific sizing."

On the amplification front, Sullivan and his team examined a number of possibilities, from polymerase choice to number of cycles run to reconditioning. The polymerase selection was the most pressing issue: the viral metagenomics community had tacitly agreed on phi29 as the whole genome amplification method of choice, and it was creating problems. Phi29 is associated with 1,000-fold to 10,000-fold bias and results are not reproducible, Sullivan says. Even worse, phi29 is known for a systematic bias that overamplifies circular and single-stranded DNA templates, which are common elements in viral genomes. "Among viruses, that means you'll get totally different representation in your resulting nucleic acids than the natural sample," he adds.

He and his colleagues screened about 10 high-fidelity polymerases against a handful of notoriously difficult samples. A hot-start polymerase from TaKaRa amplified the most and was chosen for the rest of the project.

With the optimal polymerase chosen, Sullivan moved on to cycle number — expected to be a key factor in minimizing the amplification bias. But as it turned out, amplified genomes at 15 cycles looked a lot like amplified genomes at 30 cycles. They were both different from unamplified genomes, but given that Sullivan and others working with low-input DNA do not have the option of avoiding amplification, the important point was that amplified samples were comparable regardless of cycle number.

"This means that you can make a library from 0.01 nanograms and be as comfortable with it as you are with 10 nanograms," Sullivan says. "That was both surprising and comforting, since we are often down in the 0.01 nanogram range."

The team also looked at reconditioning, a protocol for refreshing the PCR reaction before it reaches saturation to minimize problematic PCR amplicons such as heterodimer formations. For some experiments, this step is important in keeping the reaction clean and more representative of the original, Sullivan says. But for the samples his team tested, recondition-

ing did not materially change bias — "probably because it's such a complex mixture to begin with," he says.

The Pippin Recommendation

The test of size selection method pitted manual gel extraction against two automated alternatives: Solid Phase Reversible Immobilization (SPRI) from Beckman Coulter Genomics and the Pippin Prep from Sage Science. Sullivan and his team were looking for both extremely narrow sizing and excellent sample recovery.

In the paper, they sum up the results: "Of the three sizing fractionation methods tested for target recovery efficiency (fraction recovered DNA in target 400–600 bp size range), throughput (ease of applicability to numerous samples simultaneously), and risk of cross-sample contamination, Pippin Prep, an automated optical electrophoretic system that does not require gel extraction, was the most efficient and reproducible (94–96% of input DNA), with the tightest, most specific sizing."

They note that SPRI was also high-throughput with low risk of contamination, but was the least efficient in recovery of the three methods tested, yielding 46-50 percent of the targeted size fraction after shearing.

Manual gel extraction, by contrast, offered better recovery, but "efficiency varies greatly with researcher proficiency and, further, the size selection takes orders of magnitude more time and risks cross-sample contamination," the scientists report in the paper.

"You can tune up your sizing with the Pippin Prep, that's the beauty of it," he says. "It will adjust with these new sequencing technologies."

"Based on this comparative analysis, we recommend the Pippin Prep automated electrophoretic system to prepare samples for [next-gen sequencing] libraries," they conclude.

Sullivan says that the Pippin platform is "critical for this ultra-low DNA work we do." The instrument rapidly became a lab favorite "because of the tight sizing conditions and zero chance of contamination," he adds.

One thing Sullivan really likes about Pippin sizing — especially as his lab begins to work with Illumina sequencing after years of expertise with 454 — is that it will be just as useful regardless of the sequencing platform. "You can tune up your sizing with the Pippin Prep, that's the beauty of it," he says. "It will adjust with these new sequencing technologies."

Up Next

With these steps nailed down and published, Sullivan is moving ahead with a few more things to optimize for the sample-to-sequence pipeline — but he says the lion's share of this methods work is complete. For the linker amplification method reported in this paper, Sullivan's lab has posted a detailed protocol online, which they keep updated in response to feedback from the community. He is also hosting an environmental virology workshop to disseminate this information in person.

Thanks in large part to the heavy lifting done by Sullivan's lab, "the field has basically moved from the whole genome amplification method to what I'll call a first-generation quantitative amplification method," he says. Now, his team is finalizing the last few steps, including finer detail on polymerase design with an eye toward processivity. "We're working on a second-generation method in collaboration with Genoscope in France," he adds. "Some very slight tweaks of Illumina library prep allow you to get down to that couple of nanogram range for input DNA."

Of course, Sullivan knows better than to assume that he will ever be done with methods work. "As Oxford Nanopore and PacBio come online, they offer a whole other level of sequencing read lengths and throughput and library prep challenges," he says. But in the meantime, he's looking forward to getting back to the biological discoveries that his lab is geared toward — discoveries that will be far more reliable and trustworthy now that the methods foundation has been built.

For more information:

The Sullivan lab sample-to-sequence protocol website: http://eebweb.arizona.edu/Faculty/mbsulli/protocols.htm

Environmental Virology workshop, January 2013: http://eebweb.arizona.edu/Faculty/mbsulli/viral_ecology_workshop.htm

Suite 3150
500 Cummings Center
Beverly, MA 01915
978.922.1832
www.sagescience.com

sage science