

RNA direct

By Gaspar Taroncher-Oldenburg, Executive Editor

Researchers at **Helicos BioSciences Corp.** have devised a way to directly sequence short RNA strands without having to first make a full DNA copy of the molecules—an obstacle that has long plagued RNA analysis by causing inaccurate results.¹ The company is already optimizing the direct RNA sequencing protocol for use on its HeliScope single-molecule sequencer and expects to offer the technology to customers in 2010.

Further out, Helicos hopes to develop the direct RNA sequencing technology for applications ranging from single-cell transcriptional analysis to RNA-based diagnostics.

Current approaches to RNA sequencing are based on first generating cDNAs, which are DNA copies of the RNA strands. That step requires complex sample processing and the use of reverse transcriptase. Circumventing the use of the enzyme would be an advantage because virtually all of the relevant IP is held by a single company, **Life Technologies Corp.**

In addition, cDNA-based sequencing has technical limitations such as RNA degradation due to handling, low efficiency and the error-proneness of the reverse transcriptases. Together, these shortcomings can produce inaccurate readouts that compromise expression analysis.^{2,3}

Thus, a Helicos team led by VP and CSO Patrice Milos sought to bypass the use of reverse transcriptase entirely.

First, the group captured polyadenylated RNAs onto a solid support. A DNA polymerase capable of using RNA as a template was then used to attach one of Helicos' four fluorescent virtual terminator (VT) nucleotides—dubbed VT-A, VT-C, VT-G and VT-T—to the first nucleotide of the RNA strand. Fluorescent signals, indicating attachment, were read and the fluorescent tags cleaved to prepare the attached VT nucleotides for polymerization with the VT nucleotide interacting with the next position on the RNA strand.

The prototype process consisted of 120 cycles of nucleotide complementation, fluorescent readout and tag cleavage. The result was a series of strings of As, Cs, Gs and Ts representing the sequences of every RNA captured on the solid support.

As proof of concept, the team sequenced the mRNA of *Saccharomyces cerevisiae*. The protocol provided reads with an average length of 28.7 nucleotides, representing about 91% of known *S. cerevisiae* genes. The team used about 2 ng of *S. cerevisiae* mRNA, which is an improvement of 1–2 orders of magnitude over other transcriptome analysis methods.

Milos told *SciBX* that, for the first time, a group has been able to “directly sequence RNA providing an unbiased transcriptome view, which is both quantitative and qualitative. The direct RNA sequenc-

ing technology requires minimal amounts of starting RNA and has the potential to drive the method to single-cell measurements.”

“I think that it is extremely cool that they are able to directly sequence single molecules of RNA,” noted Jay Shendure, assistant professor of genome sciences at the **University of Washington**. “The most important aspect is that it has the potential to allow for unbiased quantification from very small amounts of starting material.”

Technological hurdles

Kevin McKernan, distinguished scientific fellow at Life Technologies, was less convinced that Helicos' approach removes bias from RNA sequencing. McKernan is co-inventor of SOLiD (Sequencing by Oligonucleotide Ligation and Detection), Life Technologies' second-generation DNA sequencing platform, which can also be applied to whole-transcriptome analysis.

“The paper is an exciting first step, but many large leaps still exist and even further leaps to clinical settings,” he said.

For example, noted McKernan, avoiding the use of reverse transcriptase comes at the cost of employing a DNA polymerase to incorporate very large, bulky synthetic nucleotides on RNA, a non-native substrate for the enzyme. “One wonders with the sequencing readout being so error prone if they are really making a net benefit on bias with the method,” he said.

Milos countered that the error rates were “consistent with the error rate seen with our current single-molecule DNA sequencing technology. We will continue to optimize the chemistry as we are doing for DNA sequencing—all consistent with driving the per nucleotide error rate below 2%.”

McKernan also was concerned about RNA stability. “Although the authors mention that scaling this to 125 million reads should be possible, the run times are likely to increase, and questions about the stability of the RNA on the instrument come to mind,” he said.

Milos agreed that RNA stability is still an open question. “We are currently working on eukaryotic samples and we will be discussing these data in the future. We in fact do not have a definitive answer to all the questions related to RNA stability yet,” she said.

An overarching question for any transcriptome sequencing approach is whether it can handle the complexity of the human genome.

The sequencing of the human transcriptome with this technology is “very complex due to the Helicos technology's high indel error rates, and the short reads from the RNA sequencer will make it difficult to make sense of, for example, gene splicing,” said McKernan.

Indels are insertions and/or deletions of nucleotides that can lead to frameshifts during gene translation. The result is mutated forms of a protein or different gene splicing patterns. At the level of sequencing, high indel frequencies increase the need for more overlapping reads to accurately determine a sequence.

“This was done in yeast for a reason. It has a tiny genome and you get good specificity of reads that are 20bp and larger, even with a high indel rate,” said Sean Grimmond, associate professor at **The University of Queensland's** Institute for Molecular Bioscience. “This will not wash

so well with mammalian genomes—a significant proportion of the transcriptome will be ambiguous. Indel rate is a crucial factor if looking at clinical samples.”

Milos was less concerned about the indel issue. “All our current genomic sequencing, including DNA, cDNA and direct RNA, utilizes alignments to a reference such that our IndexDPGenomic aligner can fully support reads containing indels,” she said.

IndexDPGenomic aligner is an open-source software package developed by Helicos that maps sequence data to a reference sequence.

Accurate mapping of reads to a specific reference “is an important attribute of our aligner and allows us to maximize yields to the human genome as well as smaller references, including transcriptome references,” said Milos. “Thus we can readily utilize uniquely aligning reads obtained from single-molecule sequencing from 25 nucleotides and greater lengths.”

Into the clinic

In addition to the use of direct RNA sequencing in the research space, Milos thinks the technology also has clinical applications. For example, she said the approach could allow oncologists “to more fully elucidate the cellular state of a tumor, including circulating tumor cells, to better define treatment regimens for the personalization of healthcare.”

“For clinical and translational studies, the use of small amounts of material leads to a wide expanse of applications,” added Hanlee Ji, assistant professor at the **Stanford University School of Medicine** and senior associate director of the **Stanford Genome Technology Center**.

According to Ji, the applications include guided needle biopsies, sampling from sorted cell populations such as stem cells used in bone

marrow transplant and analysis of different tissue compartments in organs such as the kidney.

The key for any of these applications will be increasing the throughput of the direct RNA sequencing technology. Indeed, the technology will require the capability to obtain many more reads than the prototype version to deal with the size of the human transcriptome, noted Ji.

“Helicos obviously needs to have a publication demonstrating this protocol on their current commercial system with many orders of magnitude more sequence reads,” he told *SciBX*.

Indeed, Milos said Helicos’ key goal is to transition “the direct RNA sequencing methodology from our research instruments to the ultra-high throughput HeliScope single-molecule sequencer.”

Helicos said the direct RNA sequencing technology is available for licensing.

Taroncher-Oldenburg, G. *SciBX* 2(39); doi:10.1038/scibx.2009.1459
Published online Oct. 8, 2009

REFERENCES

- Ozsolak, F. *et al. Nature*; published online Sept. 23, 2009; doi:10.1038/nature08390
Contact: Patrice Milos, Helicos BioSciences Corp., Cambridge, Mass. e-mail: pmilos@helicosbio.com
- Haddad, F. *et al. BMC Biotechnol.* 7, 21 (2007)
- Roy, S.W. & Irimia, M. *Bioessays* 30, 601–605 (2008)

COMPANIES AND INSTITUTIONS MENTIONED

Helicos BioSciences Corp. (NASDAQ:HLCS), Cambridge, Mass.
Life Technologies Corp. (NASDAQ:LIFE), Carlsbad, Calif.
Stanford Genome Technology Center, Stanford, Calif.
Stanford University School of Medicine, Stanford, Calif.
The University of Queensland, Brisbane, Queensland, Australia
University of Washington, Seattle, Wash.