genomeweb

International Barcode of Life Project Plans to Expand, Switches from Sanger to PacBio Sequencing

Oct 03, 2017 | Julia Karow

Premium

NEW YORK (GenomeWeb) – Having reached their initial goal of generating DNA barcode sequences for 500,000 animal, plant, and fungal species in 2015, the organizers of the International Barcode of Life (iBOL) project are working on further expanding their barcode reference library.

Along with the expansion comes a shift in sequencing technology — from Sanger sequencing to Pacific Biosciences — at the Centre for Biodiversity Genomics at the University of Guelph in Canada, which has generated the majority of the data so far, a move that the organizers say will cut their overall project cost several-fold.

Last month, researchers led by Paul Hebert, the director of the center, which has about 100 staff members, published a preprint on the *BioRxiv* server that described a pilot project that tested the PacBio Sequel platform for generating barcode sequences in up to almost 10,000 samples in a single run.

The goal of iBOL, a collaboration that involves more than a dozen countries and is spearheaded by the Centre for Biodiversity, is to establish a genetic catalog of life on Earth that uses short standardized gene sequences, so-called DNA barcodes, to uniquely identify each species. The project has received funding from the Canadian government and a number of other sources over the years.

For most animals, iBOL uses an approximately 650 base pair region in the mitochondrial cytochrome c oxidase 1 gene (COI) as its barcode. For plants, the barcode comes from two chloroplast genes, and fungi use yet another barcode sequence.

DNA barcodes are collected in the Barcode of Life Data Systems (BOLD) database, along with other data and images of the specimens. Scientists can use BOLD as a reference library to molecularly identify samples they collect in the field.

The first phase of the iBOL project formally launched in late 2010 and completed its initial goal – to register about 5 million specimens from 500,000 species in BOLD – in the summer of 2015. About 70 percent of the data was generated by the Canadian Centre for DNA Barcoding (CCDB), the sequencing core facility of the Centre for Biodiversity Genomics, and the remaining 30 percent by other facilities around the world.

At a meeting in 2015, Hebert said, the researchers formulated an even more ambitious goal, to be conducted under a new project name, the Planetary Biodiversity Mission, which would register DNA barcodes for all multicellular species on Earth, an estimated 10 million to 20 million, and which would involve barcode sequencing for several hundred million samples.

At the time, Hebert said, the estimated price tag for this project was at least $2 billion – a number he and his colleagues thought could be reduced by taking advantage of new sequencing technologies, "because a lot of the cost for the long amplicons we are chasing was involved with sequencing," he said.

Up until then, the CCDB had been using Sanger sequencing to generate the DNA barcodes, utilizing five ABI 3730xl machines and an automated workflow for DNA extraction and sample preparation. According to the facility's website, it is also equipped with an Ion Torrent PGM and an Ion S5 sequencer, both from Thermo Fisher Scientific. "It was all done on Sanger because we were just not happy with the quality of the sequences and the problems we encountered with pseudogenes when we moved onto some of the shorter-read platforms," Hebert said. "So, we stuck with Sanger, and the price tags were pretty high for some of the projects we were running here in Canada."

Looking at the PacBio Sequel platform, he said he was initially skeptical because the ultralong reads seemed overkill for the 650 to 900 base pair amplicons his group was interested in for DNA barcoding, and because the single-read accuracy of the reads was low. However, when they found out about the circular consensus sequencing (CCS) capability of the platform, which enhances the accuracy by sequencing the same DNA stretch several times, they became more interested and decided to explore how many different samples they could multiplex in a single Sequel run, and how the data compares to Sanger sequencing.

For their recently published pilot, they generated four amplicon sequencing libraries from more than 20,000 arthropod specimens in total: one from about 100 samples, one from almost 1,000 samples, and the last two from almost 10,000 samples each. For each sample, they generated a single amplicon of a 658 base pair COI gene region, which varied in GC content between species.

Hebert said that even without special efforts to normalize the DNA amount for each sample or to clean up the PCR products, they were able to obtain sequence data for about 90 percent of the samples from the two highly multiplexed libraries. The fidelity of the sequences was similar to, and sometimes even better than, that of bidirectional Sanger sequencing, he said.

In addition, the Sequel worked equally well for sequences with high and low GC content and was able to "bulldoze through homopolymer tracts that defeated Sanger," he said.

Also, with Sanger sequencing, there was always the need to look at some trace files manually to get a full read, he said, which is not necessary for Sequel data. "We're very buoyant on that technology and are convinced that it's a major breakthrough," he said.

One issue is that the number of CCS reads currently generated by the Sequel is modest compared to other next-gen sequencing platforms – Hebert said they typically obtained 300,000 reads with the Sequel – which means that rarer amplicons in the mix don't get picked up. The plan is to pool those amplicons and sequence them in a second round, which should bring the recovery rate to almost 100 percent, he said.

In their paper, the researchers estimated that pooling amplicons from about 10,000 samples on the Sequel reduces sequencing costs approximately 40-fold compared to Sanger, from $6 to $0.15 per sample. This does not include instrument amortization, Hebert said, but because of the high throughput of the Sequel, that would not add much per sample. Cost per sample could be further reduced with additional multiplexing, especially once the output of the Sequel increases, which PacBio has promised will happen next year, he said.

The quoted costs also do not include DNA extraction and PCR to generate the amplicons, but Hebert said his team is confident that the total cost of generating a barcode sequence could be driven down to less than $1 per specimen. One way to reduce sample preparation costs might be to outsource it to collaborators in countries with lower labor costs, he suggested.

As a result of the pilot project, the CCDB is going to switch from Sanger to PacBio sequencing for generating reference barcodes. "We're now taking our Sanger sequencers off warranty and off service contracts, and I would say by next year, we will be fully committed to Sequel with all of our reference library building," Hebert said. The facility already has one Sequel installed and plans to bring in at least one other instrument in 2018.

Switching technology also means that the DNA barcode standard – which is currently defined as a bidirectional Sanger read – will need to be adjusted, he added, which will require approval from the barcoding community.

Overall, the reduction in sequencing costs and greater automation is making the Planetary Biodiversity Mission project much more feasible, cutting the price tag from $2 billion to around $500 million, Hebert said. He hopes that as a result of the reduced costs, other barcode sequencing core facilities will spring up elsewhere in the world, for example, in Europe. However, additional funding will be needed to support the expansion of the barcode reference library.

Up until now, the Canadian government has invested about C$100 million in the center's core facility, he said, and in 2016, it received another C$21 million as part of a larger grant to the University of Guelph under the Canada First Research Excellence Fund in order to support its ongoing work.

Already, the center is preparing to increase the capacity of the BOLD database – originally laid out for 10 million records, and already holding 6 million – so it can take at least 100 million records.

Furthermore, it has built an informatics platform to store and analyze high-throughput sequencing data from DNA barcoding projects, called Multiplex Barcoding Research and Visualization Environment, or mBRAVE, that is currently available to beta users. The plan is to move both BOLD and mBRAVE to Compute Canada's cloud next year, Hebert said.

In the meantime, the center is also using the Sequel for other types of projects. For example, the platform lends itself well to sequencing degraded DNA, for example, from museum collections that can be more than 100 years old. "We did a lot of validation and we are finding that for degraded DNA templates, it's quite an interesting machine," he said. "We had developed protocols on both Illumina and Ion platforms, and we've now moved away from that approach on those platforms because we're just seeing so much more elegant recovery of sequences from PacBio."

Hebert said his group has also been testing Oxford Nanopore's MinIon platform, which, like PacBio's instrument, is a single-molecule sequencer that generates long reads. He said that the single-read error rate is currently still an issue, so multiple reads are required to obtain sufficient accuracy, and "it's certainly not cost-competitive with the PacBio today for the application that we're doing."

"Of course, we love the fact that it's portable and you use it in the field, and that's one of the visions of the barcode community, that you pick up something in the field and read its identity," he said. His group is currently using the MinIon in house and is including it in a project to build a one-cubic-foot box that contains devices for DNA extraction, PCR, sequencing, and informatics. "We're not going to take a Sequel into the field," he said.

Also, besides expanding the reference barcode library, the Centre for Biodiversity Genomics is getting involved in biomonitoring or environmental surveillance projects that use high-throughput DNA barcode sequencing on short-read platforms to track species in a certain area over time – for example, which insect species enter a field of crops. "That's where we see the growth in the future," Hebert said. "Not necessarily building the reference library – that's going to get done in the next 20 years – but after that, it will be ongoing biosurveillance programs."

Filed Under Sequencing Informatics University of Guelph CE sequencing NGS single-molecule barcoding Pacific Biosciences

We recommend

International Barcode of Life Project Sees Room for Sanger, New Sequencing Techs
GenomeWeb, 2008

Following Feasibility Study, UK Registry Plans to Implement PacBio for HLA Typing by Year's End
GenomeWeb, 2015

Researchers Advance 454-Based Multiplexing As Roche Preps its Own Sets for Fall Launch
GenomeWeb, 2007

NGS Used in Conservation Efforts, Biodiversity Monitoring
GenomeWeb, 2013

PacBio Outlines Expected Commercial Performance for RS at AGBT as Users Talk About their Experience
GenomeWeb, 2011

PacBio Bets Customer Data Showing Improved Performance Following Sequel Upgrades Will Drive Sales
GenomeWeb, 2016