Shotgun resequencing

By Lauren Martz, Staff Writer

The complete sequencing of a plethora of genomes has spurred the development and marketing of new DNA resequencing technologies aimed at identifying genetic variations underlying disease and other aberrant phenotypes. However, the cost and complexity of such technologies has limited their use.¹

Karolinska Institute researchers now suggest shotgun sequencing by hybridization (shotgun SBH) as a potentially cost-effective and accurate resequencing technique compared with marketed methods.² However, like all other current resequencing techniques, this method relies on the pre-existence of a reference genome sequence; therefore, shotgun SBH will not resolve the issue of how to quickly and cost-effectively perform *de novo* genome sequencing.

As described in *Nature Biotechnology*, shotgun SBH combines SBH and tiling probes (overlapping nucleotide probes that cover the entire genome)³ for genomic resequencing based on reference genomes. The team was led by Sten Linnarsson, assistant professor of medical biochemistry and biophysics at Karolinska and corresponding author on the paper.

Conventional SBH requires hybridization of fluorescently labeled, unknown DNA sequences to a microarray of single-stranded, known DNA sequences. Fluorescence signals indicate where the unknown DNA has bound to the complementary known strand, thus revealing the unknown sequence. A computational approach is then used to piece together the identified fragment sequences.

In contrast, shotgun SBH involves the sequencing of randomly cleaved genomic DNA fragments using a set of 582 universal fluorescent pentameric probes that cover all possible 5-base-pair nucleotide combinations.

Circularized, single-stranded shotgun fragments of about 200 base pairs in length were first immobilized on a microscope slide. *In situ* rolling-circle amplification then generated multiple copies of each fragment that remained associated to each other, generating submicron structures easily detectable with fluorescence microscopy following hybridization with fluorescent probes.

Images generated by sequential hybridization to the 582 pentamers were finally analyzed against a reference genome to digitally reassemble the fragments.

As initial proof of concept, the Karolinska researchers resequenced both the 48,502-base-pair Bacteriophage λ genome and the 4.6-million-base-pair *Escherichia coli* genome. The resequencing yielded a throughput of about 320 million base pairs per day and an overall accuracy of at least 99.93%.

Although the consensus accuracy of about 99.93% “looks good,” Jay Shendure, assistant professor of genome sciences at the University of Washington, noted that “if you consider that you are sequencing many millions of bases, the percentage actually indicates that it results in a lot of false positives for SNP detection.”

“The real question to ask is: What happens when they start pushing parameters?”
—Bud Mishra, New York University

An improvement in cost

Linnarsson told *SciBX* there are two major differences between shotgun SBH and existing resequencing methods. First, he said, is simplicity—the process uses only hybridization of short probes and involves no enzymatic reactions. Second, he said, is that the use of long read lengths of 200 base pairs results in high throughput and simplifies genome assembly. Both aspects can reduce cost.

The researchers estimated that shotgun SBH costs about \$0.50 per megabase, which includes crude reagent costs and the amortized cost of equipment. By comparison, the paper noted that two similar short-read methods in development cost \$110 and \$200 per megabase.

Table 1. Next-generation sequencing. At least five next-generation DNA sequencing technologies have reached the market over the last four years. The technologies can be used for DNA resequencing as well as other genetic analyses.

Company	Instrument	Technology
Applied Biosystems Group (NYSE:ABI)	SOLiD	Sequencing-by-ligation technology; Applied Biosystems obtained the technology through its 2006 acquisition of Agencourt Personal Genomics Inc.
Danaher Motion	Polonator G.007	Polymerase colony (polony) sequence-by-ligation technology; developed in collaboration with George Church's laboratory at Harvard Medical School
Helicos BioSciences Corp. (NASDAQ:HLCS)	HeliScope	True Single Molecule Sequencing (tSMS) technology
Illumina Inc. (NASDAQ:ILMN)	Genome Analyzer	Uses sequencing-by-synthesis technology developed by Solexa Inc., which was acquired by Illumina in 2007
Roche (SWX:ROG)	Genome Sequencer 20; Genome Sequencer FLX	454 technology using the sequencing-by-synthesis approach; developed by 454 Corp., which was acquired by Roche in 2007

"Synthetic oligonucleotides are much less expensive than fluorescently labeled dNTPs [dideoxynucleotides]," noted Paul Lizardi, professor of pathology at **Yale University School of Medicine**. "Even though 582 different compounds are required, these compounds will be very inexpensive and extremely stable when made in bulk quantities."

According to Shendure, "technically the method is pretty good because the accuracy is very high given that it strictly uses hybridization for the sequencing part, which might be favorable in terms of cost. It is impressive that no enzymes are involved."

But if shotgun SBH might offer a cost advantage over existing methods, Shendure noted other methods came to market first, potentially raising a barrier to new approaches.

Resequencing methods on the market include the Solexa sequencing technology from **Illumina Inc.**, the 454 technology from **Roche**, SOLiD sequencing by ligation from **Applied Biosystems Group** and True Single-Molecule sequencing from **Helicos BioSciences Corp.** (See **Table 1**, "Next-generation sequencing").

Pushing throughput

To carve out room for shotgun SBH in this competitive space, Linnarsson told *SciBX* that the "next step is to dramatically increase throughput. One way to do this, which is being explored, is to use a bead-based platform, where microbeads can be packed in a high-density monolayer."

Compared with the rolling-circle products that make up the arrays in the paper, Linnarsson noted beads "can be packed much more closely because they are solid, hard objects that don't merge when placed near each other."

He expects this step could increase throughput by more than 20-fold. With closer packing, more DNA fragments can appear in a single image, thus lowering the total number of necessary images and increasing speed. Linnarsson's lab already has generated single images containing 120,000 beads, compared with those generated by rolling-circle products, which contain only 20,000 beads.

Bud Mishra, professor of computer science, mathematics and cell biology at **New York University's** Courant Institute and School of Medicine, was impressed by the method, but agreed that throughput should be improved. He told *SciBX* that the current level of throughput is sufficient for genomes such as *E. coli* but will need to be enhanced if researchers want to scale up to larger eukaryotic genomes.

To do so, Mishra said, increasing probe size and read length would both be logical options, but he questioned whether the technique is designed to incorporate such changes.

"They are using 582 cycles because their probes are five base-pairs long. If they were using six-base-pair probes, it would make the process four times as long," he said. "Between washing and reimaging the molecules, it is possible that they could degrade in the longer process. The real question to ask is: What happens when they start pushing parameters?"

Such issues are posed when applying shotgun SBH to the human

genome, of which about 43% is comprised of highly repetitive regions. This is a problem because multiple positions in the genome could correspond to a sequenced fragment. Lengthening the fragments would lessen the likelihood that there would be multiple identical sequences of that size in the human genome.⁴

Shendure thinks repetitive sequences could pose a problem for shotgun SBH, but noted "this is not a competitive disadvantage because it is a problem across the board."

At 200-base-pair fragments, Linnarsson argued the Karolinska approach provided a clear advantage over the shorter read lengths used by other methods. "Note that what is a repeat at 35 base pairs may very well not be a repeat at 200 base pairs, hence the importance of read length," he said.

Linnarsson said his group is working to extend the technique's base-calling method, which is used to identify differences in sequence between the target and reference genomes. Currently, the method can detect substitutions in nucleotide bases. The next step is to modify the method to recognize insertions, deletions and copy-number variations, which would increase the accuracy of the sequencing technique.

Because the method can already identify substitutions, "an attractive application would be resequencing of human exons to identify mutations and SNPs," Yale's Lizardi told *SciBX*.

On the other hand, Mishra noted, "none of the current methods on the market do whole-genome sequencing. The holy grail is to be able to be doing whole-haploid genome sequencing, and everyone seems to be stuck at the same point of *de novo* sequencing: many short reads but no strategy to assemble [them]."

Genizon Svenska AB, a Swedish subsidiary of **Genizon BioSciences Inc.**, has rights to the technology described in the *Nature Biotechnology* paper. Genizon said it is looking to either license the technology to a major instrument company or to create a spinout company.

REFERENCES

1. Bentley, D. *Curr. Opin. Genet. Dev.* **16**, 545–552 (2006)
2. Pihlak, A. *et al. Nat. Biotechnol.*; published online May 25, 2008; doi:10.1038/nbt1405
Contact: Sten Linnarsson, Karolinska Institute, Stockholm, Sweden
e-mail: sten.linnarsson@ki.se
3. Liu, X.S. *PLoS Comput. Biol.* **3**, e183; published online Oct. 26, 2007; doi:10.1371/journal.pcbi.0030183
4. Whiteford, N. *et al. Nucleic Acids Res.* **33**, e171; published online Nov. 7, 2005; doi:10.1093/nar/gni170

COMPANIES AND INSTITUTIONS MENTIONED

Applied Biosystems Group (NYSE:ABI), Foster City, Calif.
Genizon BioSciences Inc., St. Laurent, Quebec, Canada
Genizon Svenska AB, Stockholm, Sweden
Helicos BioSciences Corp. (NASDAQ:HLCS), Cambridge, Mass.
Illumina Inc. (NASDAQ:ILMN), San Diego, Calif.
Karolinska Institute, Stockholm, Sweden
New York University, New York, N.Y.
Roche (SWX:ROG), Basel, Switzerland
University of Washington, Seattle, Wash.
Yale University School of Medicine, New Haven, Conn.