Fixing the front end

One bottleneck in next-generation sequencing is genomic sample selection. As research groups tackle the problem, companies are seizing a market opportunity. Ken Garber reports.

© 2008 Nature Publishing Group http://www.nature.com/naturebiotechnology

The transformative effect of next-generation sequencers has not been matched by an equal transformation at the front end. Until whole genome sequencing becomes affordable (the current cost exceeds $200,000 per genome), most researchers will need to select portions of the genome to resequence to approach disease genetics. Such selection, now done with PCR, is a serious bottleneck. To sequence all the genes in the X chromosome, for example, requires about 4,800 distinct 500-base-pair PCR products, well beyond the means of even the largest labs, which effectively negates the benefits of next-generation sequencing for anything but small projects.

Now, with little fanfare, several groups have tackled the problem. Three papers in the November, 2007 issue of *Nature Methods* demonstrate target-capture proof of principle[1–3]. In April this year, Roche NimbleGen of Madison, Wisconsin, launched the first commercial product based on the approach, and demand has been "overwhelming," according to NimbleGen product manager Luke Dannenberg. Other commercial products are in development, and users seem eager to take advantage of them. "Everyone's going to want to be doing this," says John McPherson of the Ontario Institute for Cancer Research (OICR) in Toronto, which is launching a project to resequence the pancreatic cancer genome in samples from 500 patients. "I think it's a very transformative type of technology. It reduces the cost significantly. And I think until whole-genome sequencing can compete with it on cost that's the way it's going to be done."

But the methods are still very new, and published results have fallen short in terms of performance. Questions remain about these systems' ability to efficiently and consistently capture all desired genomic regions with high fidelity. And with the field still wide open for innovation, the ultimate winners cannot yet be predicted.

The $5,000 exome

All the new methods work, in principle, the same way. A set of oligonucleotide probes is used to capture the desired sequences from total human genomic DNA. These captured sequences are then amplified in a single PCR reaction using common linkers or adaptors, originally attached either to the probes or to the genomic DNA, as primers. (The unwanted sequences are discarded.) What's left is DNA 'enriched' for the desired target sequences, which are then sequenced using any of the next-generation sequencing machines.

These front-end methods have been given various names, including 'genome partitioning', 'enrichment' and 'genomic capture'. The process may be best described as 'sequencing target capture', although all these names fall short. "I often think about it and explain to people that it's actually 'differential loss'," says Michael Zwick of Emory University in Atlanta, senior author on one of the *Nature Methods* papers, "trying not to lose the fragments you're interested in."

Target-capture methods match up well with next-generation sequencing in cost and speed, bringing the unimaginable to small labs.

Target-capture methods should empower individual labs to do new kinds of genetics. For example, megabase-size candidate genomic regions, identified through linkage studies, can be sequenced to find the exact disease-causing genes and mutations. Entire gene families, such as kinases, can be sequenced in large groups of individuals to find disease variants. Finally, all human exons—the human 'exome'—can be scanned in unbiased ways to find variants contributing to disease. The much-publicized whole-genome association (WGA) studies, which use single-nucleotide polymorphisms (SNPs) to point to disease-susceptibility genes, are powered only to detect variants that contribute to at least 5% of cases—a serious limitation for complex diseases, which are often genetically heterogeneous. The new methods offer a way around this limitation. The $1,000 genome is probably at least five years away, but the $5,000 exome, using the new capture methods, is imminent.

There are now at least ten groups, some in academia and some companies, developing front-end capture methods (**Table 1**). NimbleGen is first to market, but the competitive landscape will become very crowded during the next year, as companies jostle for position to serve a major but ultimately temporary need. "It's a narrow window that's going to close soon," says Robert Lyons, who directs the sequencing core facility at the University of Michigan medical center in Ann Arbor. "They're all trying to jump in and make their mark."

Two ways to capture

How can users distinguish the different offerings? To begin with, capture methods can be solid-phase or solution-phase. Solid-phase methods, like NimbleGen's, capture genomic DNA on probes placed on microarrays. Fragmented total human genomic DNA is ligated to an adaptor sequence, then applied to the microarray surface. The targeted genomic DNA hybridizes to the probes. Then, the hybridized sequences are eluted from the microarray and amplified in a single round of PCR. Finally, they're sequenced.

All these solid-phase methods trace their conceptual lineage back to Southern blots, which use hybridization to DNA probes on a membrane to capture target DNA fragments. The new capture methods are made possible by the development in the nineties of glass slide microarrays densely coated with long oligonucleotide probes (385,000 on a single NimbleGen microarray, with much larger arrays soon to come) that can be custom designed, one chip at a time.

Solution-phase methods, on the other hand, use probes in solution to capture the target DNA. The basic hardware required is inexpensive standard laboratory equipment—microtiter plates, not custom microarrays. Most solution-phase capture methods involve circularized DNA and trace their lineage back to the 1994 development of 'padlock probes' by Ulf Landegren, Mats Nilsson and colleagues at Uppsala University in Sweden[4]. These researchers created probes consisting of two consecutive target-complementary sequences connected by a common linker. The two ends of the linear probe are brought together by hybridization to the target sequence and then joined by a DNA

Tecan Ex. 2029, Page 1 of 4
QIAGEN Sciences, LLC v. Tecan Genomics - IPR2025-00028

Table 1 Target capture methods for next-generation sequencing

Group	Method	Reference	Development status
Roche NimbleGen, Madison, Wisconsin	Solid-phase microarray hybridization capture	2	Service and delivery products both available commercially
Agilent Technologies, Santa Clara, California	Oligo library synthesis (OLS) array platform supports two protocols: solution-phase capture method from Broad Institute using biotinylated RNA, and microarray-based capture derived from Cold Spring Harbor method	3,[a]	Solid phase microarray hybridization–based method to be released late summer 2008; Broad Institute solution phase–based method to be commercially available in the late fall as kit and possibly as service
Chad Nusbaum group, Broad Institute of MIT and Harvard, Cambridge, Massachusetts	Solution-phase using biotinylated RNA capture protocol, using Agilent oligo library synthesis array platform	[a]	Capture method licensed nonexclusively to Agilent Technologies
Richard McCombie group, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York	Solid-phase microarray hybridization, using Agilent chips and Solexa (Illumina, San Diego, California) adaptors	8	Did not file for patent. Intend to make widely available, including to companies (Agilent is using modified version.)
George Church, Harvard; Kun Zhang, University of California San Diego; Jay Shendure, University of Washington, Seattle	Solution-phase, modified molecular inversion probe (MIP) capture	3	In discussions with Affymetrix, Santa Clara, California, about possible commercialization. Affymetrix holds MIP intellectual property
Olink Genomics, subsidiary of Olink Bioscience, Uppsala, Sweden, in collaboration with Rudbeck Laboratory, Uppsala University	Solution-phase protocol using selector (circularized DNA) technology	10	Full commercial launch (service and kit) scheduled for 2009
Michael Zwick group, Emory University, Atlanta	Solid-phase microarray hybridization capture with NimbleGen chips	1	Informal collaborations underway; patent applications filed, and method available for licensing by companies
Ron Davis group, Stanford Genome Technology Center, Palo Alto, California	Spacer multiplex amplification reaction (SMART); variation of padlock probe, using long padlock probes	7	Patent application filed; Stanford University is actively seeking licensee
febit, Heidelberg, Germany	Microfluidic oligo synthesis platform with solid-phase microarray hybridization capture protocol	None	Service product to be available beginning first quarter of 2009
FlexGen, Leiden, Netherlands, and Leiden University Medical Center	Solid-phase microarray hybridization capture	None	Undisclosed

[a]Agilent Technologies press release, 13 May 2008 (http://www.agilent.com/about/newsroom/presrel/2008/13may-ca08032.html).

© 2008 Nature Publishing Group http://www.nature.com/naturebiotechnology

ligase to create a circle. The circle latches onto the target DNA like a padlock, and the common linker sequence can be used for amplification with a single PCR primer pair.

The original padlock probes were not designed for sequencing, but newer forms of circularized DNA have this ability. In 2003, Ron Davis's group at Stanford University in Palo Alto, California, developed 'molecular inversion probes' (MIPs) featuring a one-nucleotide gap between the probe ends that is filled by a DNA polymerase in the presence of the matching nucleotide on the sample, for use in SNP identification[5]. In 2007, George Church's group at Harvard in Cambridge, Massachusetts, modified MIPs by enlarging the gap to between 60–190 nucleotides, enabling capture of large sequences[3]. Earlier, in 2005, Nilsson, Fredrik Dahl and colleagues created 'selector' probes[6], which do not covalently bind to the target DNA, but instead act as a template for the circularization of target sequences, which can then be amplified using a single PCR primer pair. (A selector-based capture method is under commercial development by Olink Genomics in Uppsala, Sweden.) Most recently, Davis's group created modified padlock probes to capture sequences up to 500 base pairs long[7]. Another solution-phase capture method, from Chad Nusbaum's group at the Broad Institute of MIT in Cambridge, Massachusetts, uses biotinylated RNA probes to capture target sequences.

Agilent Technologies in Santa Clara, California, has licensed the Broad capture protocol and will combine it with its own chip-based oligonucleotide synthesis platform to create products (both a kit and a service) for commercial launch late this fall. At the same time, Agilent is on the verge of launching a solid-phase, microarray-based capture product that will compete directly with NimbleGen's. The other methods are at various stages of technical refinement and commercial development. Although researchers may be able to directly adapt some systems created in academia, this would likely involve the in-house creation of custom instrumentation and trial-and-error redevelopment of capture protocols and quality control measures. Purchase from companies, which offer user-ready products as well as training and customer support, will be the preferred option.

The diploid challenge

None of the methods described above is perfect in terms of performance. Indeed, according to the publications describing them, most aren't even close. The biggest problem is that some DNA sequences are far harder to capture than others, for reasons that remain unclear. Although the numbers aren't directly comparable, one group reported 53% coverage of their targets[8], another just 20% (ref. 3). And there was wide variability in the enrichment level of the target sequences—one group reported a 2–3 log range difference in abundance of captured targets (ref. 3). Adequate coverage is "the real technical challenge," says Zwick. Lost sequences may result in missing disease-causing mutations, negating the whole purpose of resequencing. Even with near-complete coverage, variable abundance may result in some captured variants not being detected during sequencing because the signal is too low.

All groups are working hard to improve coverage. NimbleGen's Tom Albert and Luke Dannenberg, in an e-mail, claim that the company can achieve >99% coverage of target regions, for smaller sequences. "The real challenge is to normalize coverage so that the entire target region(s) is captured with roughly equal efficiency," they write. "We have developed strategies to address this and are having considerable success."

Tecan Ex. 2029, Page 2 of 4
QIAGEN Sciences, LLC v. Tecan Genomics - IPR2025-00028

© 2008 Nature Publishing Group http://www.nature.com/naturebiotechnology

Uniform coverage is crucial for the detection of heterozygous variants. It's vital to capture both alleles (one from each parent) in sufficient abundance so that such variants can be detected. "This is…a technical challenge that any of these technologies are going to have to improve on," says Zwick. The Human Genome Project, which delivered a haploid, not diploid, genome, completely sidestepped this problem. "Sequencing diploid regions of the genome is actually much more challenging," says Zwick. "So the issue of coverage becomes even more important."

Church's group at Harvard, for example, reported correctly identifying only 25 of 90 heterozygous variants, or 31%, when captured sequences were checked against HapMap SNPs[3]. "We have improved that a lot" since that time, says Kun Zhang, a former Church lab postdoc who's now at the University of California San Diego. "Now we can detect 96% of those heterozygous [variants]." Technical improvements in the capture reaction, says Zhang, led to this substantial improvement. And the overall coverage shortfall has been completely reversed.

NimbleGen also claims to have largely solved the problem. "We have shown that once you achieve [~10-fold] coverage, virtually no allele bias is observed, and we can readily identify hets [heterozygotes] in HapMap samples with 99% accuracy," Albert writes. Stacey Gabriel of the Broad Institute says that, with enough coverage, detecting heterozygous variants hasn't been a problem with the Broad's protocol, either.

Solid versus solution

Assuming that sufficient coverage and minimal variability can be achieved, which approach is superior: solid phase or solution phase? Each has its own merits. Solid-phase protocols, unlike most solution-phase methods, do not require any enzymatic reactions, so they're relatively simple, once the custom arrays are in hand. "You just take the array and do the capture," says Richard McCombie of the Cold Spring Harbor Laboratory in Cold Spring Harbor, New York. On the other hand, solid-phase methods are inherently less scalable, because roboticization is harder than for solution phase, where the capture reaction takes place in easily manipulated microtiter plates.

The biggest obstacle to solid-phase automation is the need to separately manipulate many individual microarrays through the hybridization and elution steps. With chips, which are not cheap, costs can add up quickly. To sequence a gene set from 1,000 patients, for example, ten 96-well plates would suffice in a solution-phase protocol, notes Zhang, whereas 1,000 microarrays would in theory be needed in solid phase. NimbleGen will not disclose its pricing, but at $1,000 a chip (a reasonable guess) that would total $1 million, even before sequencing. NimbleGen's chips can be reused once, though, and Albert did note in a seminar last April that the company is working on ways to pool samples for capture. These measures would cut the need for chips by as much as 90%, in theory.

The biggest practical drawback to solid-phase methods is the sheer quantity of DNA required on the front end. Each probe on a microarray occupies a vanishingly small space, yet genomic DNA must be spread over the entire microarray surface. So an excess of sample DNA is required up front to enable individual probes to contact and capture enough for enrichment and sequencing. This can be a problem when sample DNA is scarce, as when using paraffin-embedded tissue samples (most primary tumor samples), in which DNA breaks down easily.

In solution phase, on the other hand, far less DNA should be needed, because higher probe concentration can drive the capture reaction faster. "You can add an excess of probes, compared [with] the DNA, and basically the probes can scan the whole volume for targets," explains Olink's CEO Olle Ericsson. "That way you can have a more efficient enrichment."

NimbleGen has required researchers to supply 20 micrograms of DNA for its service product. (Albert writes that NimbleGen is making progress on reducing this.) Most solution-phase methods require less, and some, like the Broad Institute's, far less. "We're routinely running at half a microgram," Nusbaum says. "I think you could do it with less. We just haven't pushed that." Nusbaum credits these results to his single-strand RNA probes, which tightly bind target DNA and which can be added in high concentration to the capture reaction, enhancing efficiency.

Agilent claims that its unique ability to synthesize high-quality libraries of long oligonucleotide probes (up to 200 base pairs) on its chips also reduces the amount of input DNA needed, even in solid phase. "Since we have more DNA on the surface of each probe, we have higher capacity to capture the [genomic DNA] sample," says Emily LeProust, Agilent's R&D chemistry manager. "Which means that we may need less input DNA than other microarray protocols." NimbleGen, whose oligo probes are <100 base pairs long, denies that Agilent has a technical edge. "We have shown that our arrays are not a significant bottleneck for the process," the company writes.

What's certain is that most solution-phase methods are inherently better than solid phase when it comes to capture specificity. That's because they target not one but two DNA sequences in each captured molecule. Enzymatic reactions, used to close the circle, act as checkpoints. Methods that rely solely on hybridization are more likely to capture DNA outside the regions of interest, leading to wasted sequencing.

So is solution-phase better than solid? Not necessarily. Given the low cost of sequencing, tight specificity may not be that important. As sequencing prices drop, "if half your reads map someplace else, or aren't directly mappable to your region of interest, maybe you don't care—you just throw them away," says Zwick, who adds the microarray probes might even help coverage by capturing targeted sequences that are much longer than the probes themselves. More importantly, the specificity of DNA circle-based solution phase methods have a possible drawback. These methods are very good at capturing sequences with single-base-pair variations, but they may miss sequences with deletions or insertions because of the close homology required between probe and target. "They might just not bind," says Zwick, who adds that this concern is so far theoretical. "I don't think there's any data yet showing one way or another."

And the relative complexity of most solution-phase protocols is unsettling to some. "There is a robustness issue," says McCombie. "The more complicated things get, the more skeptical I am of their ability to scale."

All these concerns are theoretical, because the methods themselves are new, unproven and still being perfected. Ultimately, their success will be determined by their usefulness to the research community, which has yet to try any of them, except for NimbleGen's. "It's pretty new stuff," says McCombie. "There will be a handful of approaches that have different strengths and weaknesses, and will shake out as they become more clear in the next six to twelve months."

Front end alignment

NimbleGen is first to market and therefore has an important advantage. And the company was aggressively courting customers, even before official product launch. "NimbleGen was willing to give us early access to their system," says the University of Michigan's Lyons. "That gives them an inside route." But Lyons adds that he's not wedded to the NimbleGen technology and is open to other systems. And NimbleGen's pricing is an issue. "If they don't get their price down now, they're going to find it hard to compete," Lyons says.

Even as companies prepare for product launch, the technologies are leaping forward.

Tecan Ex. 2029, Page 3 of 4
QIAGEN Sciences, LLC v. Tecan Genomics - IPR2025-00028

Whole-exome capture should soon be a reality. "I think it might happen in six months or a year," says Zhang. NimbleGen plans to soon offer a five-array human genome set, followed by a single-array system, possibly by the end of the year. The US National Institutes of Health are on the verge of awarding $12 million in grants to several groups to develop whole-exome capture systems, so a variety of systems should soon be available from both the private and public sectors[9]. If they're cheap enough and reliable enough, these should largely supplant WGA studies to discover rare disease variants.

But capturing the whole exome, strictly defined, in a single reaction won't be easy. "I think what we're going to find out is that there's certain exons that are just never going to be amenable to these sorts of approaches," says Zwick. For example, sequences that are very high (or low) in G+C base pair content are very difficult to capture. But researchers are working hard on capturing all the exons. "We're pushing for 100%," says McCombie. "We certainly want to go, if at all possible, well over 95%. But I think that's going to have to wait for more experimental work."

Whole-exome capture may be the biggest prize, but the new front-end systems will probably be used more often for lower-profile projects. "The real direct applications will actually be in human genetics where people have a few megabase-size regions. . .they need to sequence," says Zwick; for example, when a family pedigree gives a general location of a disease gene and resequencing can find the gene and identify the variants. "If you could tell someone that they could generate 300,000 base pairs of high-quality sequence very fast, there's a lot of people who are [now] limited by their ability to resequence that targeted region."

The new front-end technologies will bring large-scale resequencing to the laboratory masses, and an explosion of new discoveries should follow. But resequencing is only the first step in understanding and treating human disease, and not an end in itself. "The sequence is necessary to make progress. You have to find the variant," says Zwick. "If you're a geneticist, if you don't find the variant, you're done." But once variants are identified, Zwick adds, researchers must work out their function (if any) in biology and disease. The real importance of the new target capture methods is that, together with next-generation sequencing, they enable individual biologists to begin this crucial work.

Ken Garber, Ann Arbor, Michigan

1. Okou, D.T. *et al. Nat. Meth.* **4**, 907–909 (2007).
2. Albert, T.J. *et al. Nat. Meth.* **4**, 903–905 (2007).
3. Porreca, G.J. *et al. Nat. Meth.* **4**, 931–936 (2007).
4. Nilsson, M. *et al. Science* **265**, 2085–2088 (1994).
5. Hardenbol, P. *et al. Nat. Biotechnol.* **21**, 673–678 (2003).
6. Dahl, F., Gullberg, M., Stenberg, J., Landegren, U. & Nilsson M. *Nucleic Acids Res.* **33**, e71 (2005).
7. Krishnakumar, S. *et al. Proc. Natl. Acad. Sci. USA* **105**, 9296–9301 (2008).
8. Hodges, E. *et al. Nat. Genet.* **39**, 1522–1527 (2007).
9. US National Institutes of Health. Development and application of new technologies to targeted genome-wide resequencing in well-phenotyped populations (R01). *Funding Opportunity Announcement* <http://grants.nih.gov/grants/guide/rfa-files/RFA-HL-08-004.html> (2008).
10. Dahl, F. *et al. Proc. Natl. Acad. Sci. USA* **104**, 9387–9392 (2007).

© 2008 Nature Publishing Group http://www.nature.com/naturebiotechnology

Tecan Ex. 2029, Page 4 of 4
QIAGEN Sciences, LLC v. Tecan Genomics - IPR2025-00028