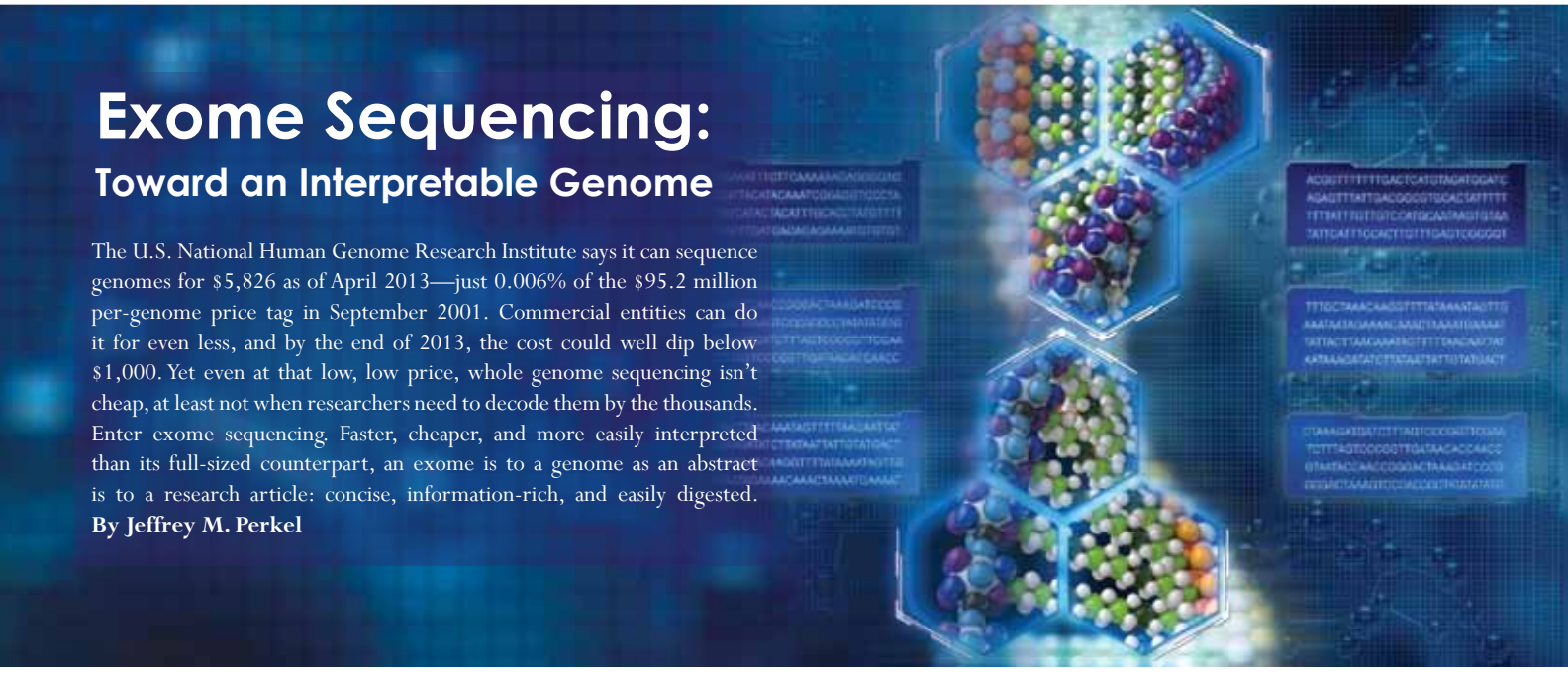
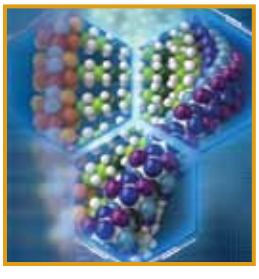
Exome Sequencing:
Toward an Interpretable Genome



The U.S. National Human Genome Research Institute says it can sequence genomes for $5,826 as of April 2013—just 0.006% of the $95.2 million per-genome price tag in September 2001. Commercial entities can do it for even less, and by the end of 2013, the cost could well dip below $1,000. Yet even at that low, low price, whole genome sequencing isn't cheap, at least not when researchers need to decode them by the thousands. Enter exome sequencing. Faster, cheaper, and more easily interpreted than its full-sized counterpart, an exome is to a genome as an abstract is to a research article: concise, information-rich, and easily digested.

By Jeffrey M. Perkel

"The exome right now is the part of the genome we know how to interpret."



An exome is simply the protein-coding content of the genetic code, some 1%–2% of the genome in all. Since sequencers can read only so many bases per run, researchers sequencing exomes can produce more of them more quickly, at greater resolution and lower cost than they can whole genomes.

Joris Veltman, professor of translational genomics at **Radboud University Medical Center** in Nijmegen, the Netherlands, uses **Life Technologies**' SOLiD instruments for both research and clinical sequencing applications. His lab has the capacity to sequence a couple thousand exomes annually, he estimates, but just 50 or so whole genomes. "Throughput is … a major reason why we choose [to do] exomes," he says.

Plus, there's interpretability. Whole-genome sequencing inarguably produces more data, including the noncoding bases and structural and haplotype phasing information that exomes miss. But what most of those nucleotides do remains a mystery, as are the functional consequences of changing them. The implications of a missense mutation in a protein-coding gene, though, are more easily grasped. "The exome right now is the part of the genome we know how to interpret," says Stacey Gabriel, director of the genomics platform at the **Broad Institute of Massachusetts Institute of Technology** (MIT) and **Harvard University.**

As a result, researchers have been sequencing exomes at a blistering pace. Only a handful of papers on exome sequencing had been published by early 2010; today, there are more than 1,600 listed in PubMed.

"In excess of one-hundred-some thousand exomes" have been decoded at the Broad Institute, says Gabriel. (Compared with the approximated 17,000 whole human genomes that have been sequenced to date, by Harvard University geneticist George Church's estimate.) Gabriel says her facility probably sequences four exomes for every whole genome and has a capacity of some 2,000 exomes per week. That's thanks to a fleet of 50 or so **Illumina** HiSeq 2000s and 2500s that were installed largely for the National Heart, Lung, and Blood Institute's Grand Opportunity Exome Sequencing Project, a gene-discovery project for which the Broad Institute and the **University of Washington** collectively sequenced 7,500 exomes.

Yale University's 10 HiSeqs decoded 12,000 exomes just in the past year, says **Howard Hughes Medical Institute** Investigator Richard Lifton, chair of the university's Department of Genetics, who published one of the first examples of exome sequencing in the clinic in 2009 and 15 additional exome papers since. "We're in a very productive phase right now," Lifton says. "We're now sequencing exomes for an all-in cost, including all of the instrument amortization, of $500 flat." By comparison, he says, his lab has done "very few" whole genomes.

From the sequencer's point of view, though, it's six-of-one—both exomes and whole genomes are decoded the same way. The difference lies upstream, during the target capture and library preparation steps that precede sequencing.

Upcoming Features

Cell Culture: Scaling Up—December 6
RNA Technologies—January 17
Proteomics—February 21

HYBRIDIZATION STRATEGIES

Exome sequencing is simply a special form of target enrichment, a sequence preparation strategy that pulls out genetic elements of interest prior to decoding them. Researchers do this to stretch dollars and maximize efficiency: If you don't need the entire genome, why sequence it? At the same time, by sequencing fewer bases per sample, researchers can sequence more of those samples at once, and at higher coverage.

Coverage describes the number of times a given base is read by the sequencer. But such figures are statistics, not absolutes. Thirty-fold coverage means that each base is read 30 times on average; some are read more frequently, others less so. For many applications that's sufficient. But when it comes to finding rare variants on which clinical decisions might depend—a key mutation in a heterogeneous tumor, for example—the more passes at a given nucleotide, the better.

Some researchers use target enrichment to select a handful or a few hundred genes for sequencing. In exome sequencing, the target is the entire exon complement of the human genome. Typically, that's about 30 Mb of sequence, though users can supplement that pool with 5' and 3' untranslated regions, microRNAs, long noncoding transcripts, and other selected custom regions, which can significantly expand the amount of captured material. Illumina's Nextera Rapid Capture Exome kits, for instance, come in two flavors: a basic kit that captures 214,405 exons totaling 37 Mb and an "expanded" kit that adds UTRs and microRNAs for a total of 62 Mb.

Several of the earliest exome studies, including seminal 2009 reports by Lifton and Jay Shendure, associate professor of genome sciences at the University of Washington, employed array-based hybridization for target capture, a strategy then commercialized by **Agilent Technologies** and **Roche NimbleGen.** But solution hybridization (which uses a pool of biotinylated oligonucleotides that are pulled down post-hybridization with streptavidin beads) has now supplanted that approach.

Indeed, NimbleGen has since exited the microarray business altogether; Agilent still sells arrays, but mostly for customers who want to continue using the same enrichment tools they used earlier in a project, says Yong Yi, Agilent's marketing director for next generation sequencing products. "The vast majority of exomes generated today have been through solution hybridization," says Shendure.

Shendure was one of the first researchers to tap exomes for analyzing Mendelian disorders, and since 2009, by his count, "at least 100 new disease genes have been identified" using the technology. "It's kind of exploded pretty remarkably," he says, adding that there's nothing magic about exomes over whole genomes; they simply provide "a cost-accessible way of getting at most of what you wanted [from the genome] for a lot of the questions that are reasonable to ask."

In their exome work, Shendure and his colleagues use NimbleGen's solution-based SeqCap EZ Human Exome Library v3.0. So, too, does Lifton, whose lab has used the approach to identify genes that contribute to hypertension, congenital heart disease, autism, and thrombosis, among other conditions.

Illumina's Nextera Rapid Capture Exome kit (a high-speed method that uses transposons and optimized hybridization steps to simplify and shorten the protocol from several days to just a day and a half) is also based on solution hybridization, as is Agilent's SureSelect, a tool developed (and still used) at the Broad Institute that captures target sequences in solution using 120-mer biotinylated RNA baits—882,000 of them in the case of the SureSelect Human All Exon V5+UTRs panel.

CREDIT: © ISTOCKPHOTO.COM/LUCHSCHEN

"When you're interested in sequencing a modest number of genes in a very large cohort of people, the padlock approach is a great one."

MOLECULAR PADLOCKS

Targeted capture can also be accomplished using standard PCR (for instance, using **RainDance Technologies'** droplet-based approach). Or, users can try so-called molecular inversion probes (MIPs), or "padlock" probes.

A padlock probe, Church explains, is "basically two PCR primers that are joined at the hip." Linked in a single oligonucleotide, these two primers capture either end of the targeted sequence, forming a molecular semicircle (like an open padlock) that flanks a gap. The lock is closed using DNA polymerase and ligase, and the captured material amplified and sequenced, while nontargeted sequences are destroyed by exonucleases.

The approach offers a significant advantage over plain PCR: ease of multiplexing. "You don't get into the N-squared problem that you have from putting together multiplex PCR," Church explains, where "every primer can in principle interact with every other primer and all their extension products."

As a result, vast probe panels can be combined in a single reaction. In 2007, Church and his then-postdoc Shendure multiplexed 55,000 MIPs capable of capturing some 10,000 exons; in 2009, Shendure refined the pool to target 50,000 exons.

Today, Shendure favors SeqCap EZ. But he uses MIPs, too. In December 2012, he and his colleague Evan Eichler used them to capture 44 genes potentially related to autism spectrum disorders from 2,446 patient samples. Boston-based startup Pathogenica (co-founded by Church) also uses MIPs for bacterial strain typing and viral drug resistance analyses.

"When you're interested in sequencing a modest number of genes in a very large cohort of people, the padlock approach is a great one," Shendure says.

Agilent commercializes a somewhat related strategy in its HaloPlex Exome kit. In HaloPlex, long biotinylated oligos with capture sequences at either end capture targeted genomic fragments by hybridization, producing a circle that can be PCR amplified to generate a sequencing library in essentially a single step. "It's a combination of both worlds," Yi says. "It combines the advantages of hybridization with the simplicity of PCR."

EXOMES IN THE CLINIC

The power of exome sequencing was beautifully illustrated in a 2011 Pulitzer Prize-winning feature in the Milwaukee *Journal Sentinel*, which detailed the efforts of a group of researchers at the Medical College continued>

Featured Participants

Agilent Technologies
www.agilent.com

Broad Institute
www.broadinstitute.org

Harvard University
genetics.med.harvard.edu

Illumina
www.illumina.com

Life Technologies
www.lifetechnologies.com

Pathogenica
www.pathogenica.com

Radboud University Medical Center
www.ru.nl/english

RainDance Technologies
raindancetech.com

Roche NimbleGen
www.nimblegen.com

University of Washington Genome Sciences
www.gs.washington.edu

Wellcome Trust Sanger Institute
www.sanger.ac.uk

Yale University
medicine.yale.edu/genetics

Additional Resources

International Cancer Genome Consortium
www.icgc.org

Milwaukee *Journal Sentinel* story
www.jsonline.com/features/health/111224104.html

NHLBI Grand Opportunity Exome Sequencing Project
esp.gs.washington.edu/drupal

The Cancer Genome Atlas
cancergenome.nih.gov

of Wisconsin to diagnose and treat a young boy with an inexplicable and exceptionally severe case of inflammatory bowel disease. In what turned out to be the first clinical use of exome sequencing, researchers identified a single point mutation in the X-linked inhibitor of apoptosis (XIAP) gene. That information suggested a possible therapeutic strategy, umbilical cord blood transplant. Since then, the college has applied the approach to 25 additional cases, obtaining a "definitive diagnosis" in 27% of them.

The Wisconsin researchers read their XIAP patient's exome to 34x. Veltman says he would prefer to sequence his clinical exomes to 1,000x, but cost and throughput limit him to typically 60-fold coverage, while most of the exomes Broad's Gabriel collects for her work on The Cancer Genome Atlas (TCGA) project are at 120-fold coverage. That's far deeper than the typical whole genome—the Broad Institute sequences those to 50x, Gabriel says—but for some applications, and especially if patients are involved, even 120-fold coverage won't do.

Pancreatic tumor samples, for instance, are notoriously difficult to obtain at high purity, Gabriel says, and require 300- to 400–fold coverage. For other applications, for instance, when looking for specific deleterious mutations in patient samples, physicians "want to have 500x coverage minimum," Gabriel says.

That high depth of coverage, combined with the fact that—even in the exome—relatively few gene mutations are actually actionable, has some researchers contemplating a scaled-down approach in the clinic.

"The reality is for every clinical exome, one would be able to sequence many, many more [samples] in the setting of a targeted gene screen," says Ultan McDermott, a principal investigator in the Cancer Genome Project at the **Wellcome Trust Sanger Institute,** adding, "It's almost certain that defining the mutational signatures that underpin biological outcomes such as drug response and survival will require an order-of-scale larger numbers than all of the previous international [next generation sequencing] efforts."

As a result, McDermott advocates using large-scale exome-sequencing projects like the International Cancer Genome Consortium to identify interesting variants and smaller-scale targeted gene sequencing of four or five hundred genes to actually test for in patients. Indeed, McDermott says that approach already is being implemented in a pilot study called SPECTAcolor now kicking off in Europe, wherein all colorectal cancer patients being considered for inclusion in clinical trials will have 400 or so genes sequenced at the Sanger Institute. "This information would instantaneously allow the patient to be stratified into any clinical trial that arises where you need to know a particular mutational event as a point of entry."

The flip side, of course, is that targeted studies are inherently biased. "You're obviously not going to find anything that you haven't specifically targeted," McDermott says.

Veltman's research is a case-in-point. Veltman studies severe intellectual disability. Perhaps 200 genes have been linked to that condition, he says, but "it's clear that we know perhaps only 10%" of them. As a result, a targeted approach that looks only at known genetic players might well miss something interesting.

In one 2010 study, Veltman's team sequenced 10 "trios"—an affected child and his or her unaffected parents—to 42-fold coverage, ultimately focusing on 10 "de novo" mutations (lesions not present in the parents). Three of these were known to be associated with intellectual disability, and three appeared irrelevant. Functional analysis of the remaining four—their observed behavior and interaction partners in model organisms, for instance—suggested they too could play a role in intellectual disability. A follow-up study in 2012 applied the same strategy to 100 patients, identifying de novo mutations in 10 genes known to be associated with intellectual disability and 19 candidate genes, and confirmed three additional novel genetic players in the condition (one of which was among the four picked up in the 2010 study).

The prevalence of de novo mutations in these two studies, Veltman says, flies in the face of the classical view of sporadic severe intellectual disability, which typically attributes the condition to autosomal recessive inheritance, and suggests a fundamental reassessment of the way geneticists think about rare diseases could be in order. "It turns out that these de novo mutations are a very common cause of intellectual disability," Veltman says.

And yet many of them would have been missed using a simple gene panel as they were not among the many genes then known to be associated with severe intellectual disability.

Of course, even more can be discerned from whole genome analyses, and Veltman (like many others) is pursuing these, too. Most researchers agree that once the cost and ease of sequencing and interpreting whole genomes matches exomes, the attraction of exome sequencing will fade.

But it won't disappear—if nothing else, large patient population studies may well require it. Says Shendure, "You should rationally do the study that makes the most sense for the question you're interested in. Sometimes it will be genomes, sometimes it'll be exomes. Sometimes it'll be much more targeted."

Jeffrey M. Perkel is a freelance science writer based in Pocatello, Idaho.

DOI: 10.1126/science.opms.p1300079