Cell

Whole-Genome Sequencing Breaks the Cost Barrier

The ever-decreasing cost of sequencing is ushering in a new era in genomic medicine. Laura Bonetta reports on new studies that provide a glimpse of the opportunities ahead.

Only 5 years ago, it was a multimillion-dollar proposition. But today, the 3 billion base pairs of a human genome can be sequenced for about $50,000, and prices are expected to drop by more than half by year's end. It's still a hefty price tag, especially if you add to it the time and resources needed to interpret the data. But the drop in costs now makes it feasible to sequence the entire genome of patients and their families to find elusive disease-causing mutations.

Researchers currently rely on microarrays containing common human single-nucleotide polymorphisms (SNPs) to quickly scan the genomes of hundreds or thousands of individuals at once to pull out variations associated with disease. But despite an explosion of data from these so-called genome-wide association studies, the variations identified only account for a small proportion of the genetic risk of developing a disease. Whole-genome sequencing promises to uncover this hidden genetic risk. By sequencing the entire genome of an individual and then comparing the sequence to that of a reference human genome, researchers can catalog every single variation in the DNA code of that particular individual. Bioinformatics tools can then home in on the variations associated with a particular trait or condition. Once enough genome sequences have been amassed, whole-genome sequencing then could be used to diagnose disease or to inform the best therapeutic options.

As sequencing costs continue to plummet and bioinformatics tools mature, whole-genome sequencing is poised to replace other tools of the genomics trade. "It's definitely the way forward," says Richard Gibbs of the Baylor College of Medicine in Texas, whose group has just reported a pioneering study using whole-genome sequencing to identify mutations that cause the hereditary disorder Charcot-Marie-Tooth disease (*N. Engl. J. Med.*, 2010, *362*, 1181–1191).

Going after Mendelian Diseases

Since coming on the market in 2005, next-generation sequencing methods have greatly increased sequencing speeds and have driven down costs, compared to the original Sanger sequencing method used by the Human Genome Project. The immediate application of whole-genome sequencing using next-generation technologies is to search for mutations responsible for single-gene (Mendelian) diseases or monogenic traits. Gibbs sequenced the whole genome of his colleague James Lupski, a prominent medical geneticist who has Charcot-Marie-Tooth disease, a neurological disorder that causes muscle weakness. Mutations in over 30 different genes, many of them identified by Lupski himself, give rise to this condition, but Lupski did not carry any of them.

Comparison of Lupski's full genome sequence to the Human Genome Project reference sequence revealed many differences that greatly exceeded the number of known SNPs and other variations. As a result, Gibbs and Lupski had to determine which of these variations were disease-associated mutations, harmless variations, or just errors in sequencing. "This is going to be a common problem for all sequencing projects," says Gibbs. "We are going to find many more mutations than are currently in databases, so you need to find a way to narrow down the hits."

So, the researchers focused on select variations found in genes previously linked to Charcot-Marie-Tooth disease and other nerve disorders. They found two compelling mutations, one in each allele, of a gene called *SH3TC2*, which is expressed in the Schwann cells that wrap the myelin sheath around nerves. Of Lupski's seven siblings, the three with Charcot-Marie-Tooth disease carried both mutations and the four without the disease did not, confirming that the disease-causing mutations had been identified. "This is the first demonstration of using full blown sequencing in a clinically relevant approach," says Gibbs.

Similarly, Leroy Hood and David Galas at the Institute for Systems Biology in Seattle analyzed the entire genomes of four members of one family in which two children had different single-gene recessive diseases (Miller syndrome and primary ciliary dyskinesia; *Science*, 2010, *328*, 636–639). By comparing the four sequences, "we found that sequencing errors are about 1000 times more prevalent than true mutations," says Galas. "Distinguishing errors from variations in the sequence is more difficult to do if you are just sequencing unrelated individuals and comparing them to a reference genome."

Once they had eliminated sequencing errors and obtained a list of possible genetic variations in each individual, the group looked specifically for rare missense mutations (which would cause amino acid changes) in the coding sequences of genes located within chromosome segments inherited by both siblings. They reduced the number of possible variations responsible for the two conditions to four. "We used a simple model of rare SNPs in exons that cause missense mutations and found that applying that model eliminates all but four genes," says Galas. "We were amazed at how well it worked."

They also obtained the first direct estimate of how much the genome changes from one generation to the next and found the mutation rate (1.1×10^{-8} mutations per position per generation per haploid genome) to be about half the predicted number. Galas and colleagues are now gearing up to sequence 30 individuals over 4 generations to obtain a

more precise estimate. "Then we could ask things like, 'What does the mutation rate depend on?'" says Galas.

Finding answers about basic biological processes, such as mutation rates or disease mechanisms, is another application of whole-genome sequencing. "Individual genome sequencing so far has been primarily descriptive. It has been used to look at the features and landscape of the genome," says Stephen Kingsmore of the National Center for Genome Resources in Santa Fe. Kingsmore recently used whole-genome sequencing to examine why in monozygotic (identical) twins discordant for multiple sclerosis—an autoimmune disease that destroys the myelin sheath protecting nerve cells—one twin has the disease and the other does not (*Nature*, 2010, *464*, 1351–1356).

His group sequenced the complete genomes of a monozygotic twin pair discordant for multiple sclerosis to look for variations that might explain the difference in disease manifestation but did not find any. "We did not find any sequence differences between twins in any unique genome sequence," says Kingsmore. "This was a pioneering study to show how sequencing can be used to examine hypotheses." Kingsmore also used sequencing to analyze differences in epigenetic markers and in transcripts between the genomes of this twin pair and again did not find any telltale differences. "A new paradigm in genome analysis is to include the functional genome and epigenome in relevant tissues," he says, "in addition to just sequencing the canonical, static genome."

More Sequence for Your Buck

Whole-genome sequencing requires tremendous computational power—for example, one complete human genome at 30-fold coverage requires 90 Gb of computer space—and sophisticated analyses. "The paper we published makes it look fairly straight forward. But that work involved large teams of people with expertise in biology and bioinformatics, and we spent a lot of time analyzing the data," says Gibbs.

Another challenge is that most high-speed, next-generation instruments sequence DNA in very short segments, or reads, of about 100 base pairs, according to Jeff Schloss of the National Genome Research Institute. "This is a huge problem for understanding the structure of the genome." Short reads make it impossible to independently assemble a genome sequence from scratch. Thus, researchers need to overlay their sequence data on a reference genome. But the sequence data cannot be sorted into the two sets of parental chromosomes, or haplotypes. And short reads make it difficult to detect variations other than single base-pair changes or very small deletions and insertions. "There are new technologies that promise longer reads and are probably not going to be much more costly," says Schloss. "Many people will want to get their hands on them." Pacific Biosciences and Life Technologies plan to market these so-called third-generation sequencing machines within the next year or so. These instruments provide single-molecule resolution—in other words, sequence information from a single DNA strand—and can sequence reads of a thousand bases or more. In addition, some third-generation sequencing instruments should allow detection of methylation changes and the ability to directly sequence RNA.

Another problem is that the reference genome is not optimal for comparisons that involve complex datasets. In addition, databases of genome variations, such as SNP databases, are not sufficiently well populated to enable researchers to find matches for the thousands of changes they find in their sequencing efforts. Also, algorithms for interpreting sequence data are still not sufficiently mature to handle data from large numbers of individuals with many different diseases.

Given these challenges and the cost of whole-genome sequencing, some labs have opted only to sequence the exons, or exome, of an individual. Exome sequencing (about 1% of the whole genome) costs $5000 and requires simpler analysis. "We are interested in monogenic disorders and most mutations are found in coding regions of genes," says Michael Bamshad at the University of Washington School of Medicine in Seattle. "At the moment, we think exome sequencing is a better use of resources." Bamshad and colleague Jay Shendure sequenced the exomes of four individuals from three families that had a child with Miller syndrome. They mapped the mutation to the gene *DHODH* (which encodes an enzyme involved in pyrimidine biosynthesis) that was also identified by Hood and Galas (*Nat. Genet.*, 2010, *42*, 30–35).

Bamshad predicts that soon there will be many more reports of disease mutations identified by exome sequencing. "We and others have found additional genes using this method that will be published soon," he says. Of course, exome sequencing will miss mutations that are in noncoding regions of the genome or in coding regions that have not yet been annotated. "We will probably get 1 in 2 or 1 in 3 mutations. This is because a lot of mutations are too subtle or they are not in what is known to be a transcript," says Kingsmore.

Nonetheless, Kingsmore's group has taken advantage of exome sequencing to develop a sequence-based test to identify people who carry mutations in a panel of 500 genes that predispose to several childhood diseases. "It will be used for carrier testing, much like was done for Tay Sachs and other genetic diseases," he says. Launching at the end of the year, it will be one of the first gene sequencing-based tests to come on the market. "These types of panels will be a great way to usher in genomics-based medicine," he says.

Tackling Complex Diseases

One of the most exciting applications of whole-genome sequencing is to decipher the molecular basis of complex diseases such as diabetes, cardiovascular disease, autism, and schizophrenia, which are thought to be caused by combinations of several mutations and environmental factors.

Genome-wide association studies have picked up many of the more common variants associated with particular diseases. "But these studies have told us that the vast majority of the genetic risk of most common diseases is not explained by common variants and that most disease-related common variants have very small effects on disease risk," says Richard Lifton at Yale University in Connecticut. "These findings suggest an alternative model in which individually rare variants with relatively large effects will play a substantial role."

His group has demonstrated that rare heterozygous mutations in genes that cause recessive forms of low blood pressure fit this model. "Those alleles are found collectively in 2% of the popula-

tion, but each one reduced the risk of hypertension by about 60%," says Lifton. He thinks that the same principle will apply to more common types of disease. "Because these variants are rare in the population, you will need large cohorts to distinguish functional variants from those that have no consequence," he says. "That will be a significant challenge."

Matthew State, also at Yale, is applying exome sequencing to elucidate the mechanisms involved in autism and other neurodevelopmental disorders. "There is a lot of evidence that multiple rare mutations will account for a good deal of autism," says State. "Already the rare variants that have been identified so far have changed the way people think about autism." For example, some of the known variants are involved in the mechanism by which nerve impulses in a presynaptic neuron increase the probability that the postsynaptic neuron will fire an action potential.

Using a large DNA repository from 1700 families (comprising an autistic child, the parents, and an unaffected sibling), State and colleagues have started sequencing exomes. "Initially, we are sequencing trios—two parents and a child—and then picking out rare missense and nonsense mutations that arise de novo in the affected children," he says. "There is already a good deal of evidence that new mutations are important in autism and this approach is a great way of winnowing down the tremendous amount of rare variations we see in the exome."

Although State believes that rare variants will reveal a great deal about both the genetic predisposition for and the molecular mechanisms underlying autism, he is cautiously optimistic. "I also think there is a good chance that the genetic risks have been overestimated," he says. "We will probably have to take into account other mechanisms such as epigenetics and environmental contributions."

One of the biggest efforts to look for variants associated with complex traits is the Exome Sequencing Project funded by the National Heart, Lung and Blood Institute of the National Institutes of Health. The project aims to sequence about 8000 exomes of individuals with common cardiovascular and lung diseases like early myocardial infarction, stroke, and chronic obstructive pulmonary disease.

Whether rare alleles will individually account for a large proportion of genetic risk is debatable. "My expectation is that we will find many rare variants associated with common diseases but individually they will not account for a large proportion of cases of disease," says Kari Stefansson, president of the Icelandic company DeCode Genomics. "The epidemiological data does not support this conclusion." Instead, Stefansson predicts that common diseases will be caused by the "rare confluence of variants rather than just individual rare variants giving large effects," he explains.

DeCode plans to sequence the entire genomes of 2000–2500 Icelanders, a population for which the company has already gathered extensive genotypic and medical data. This sequencing project and other population-based studies such as the 1000 Genomes Project—an international collaboration between China, Germany, the UK, and the US to sequence the whole genomes of about 2000 individuals—will help to answer questions about the prevalence and role of rare variants in disease. "Once we have the sequence data, we won't have to postulate any more and I predict we will all be very surprised," says Stefansson.

Scrutinizing Cancer Genomes

Other whole-genome sequencing projects are looking at the sequences of different cancers. "In cancer, it is clear that each cancer cell will have thousands to tens of thousands of mutations. Most of them will be random, what we call passenger mutations," says Peter Campbell at the UK's Sanger Institute. "But some, and we don't know the number, will be driver mutations responsible for cancer development." Campbell's group is part of the International Cancer Genome Consortium, which plans to describe all genetic, transcriptional, and epigenetic changes in 50 different tumor types in 25,000 cancer samples.

"What whole-genome sequencing is providing is discovery of all mutations without bias. We were sequencing genes before, but based on candidate genes we already knew were probably involved in cancer. So the sequencing merely allowed us to delve deeper into the mutations," says John McPherson, at the Ontario Institute for Cancer Research in Toronto, whose group is sequencing the pancreatic cancer genome. "With unbiased sequencing we can find anything. We will find new genes. This will help in the basic understanding of cancer mechanisms but also will help us define new subtypes."

The picture emerging from cancer genomics is that specific cellular pathways are important in cancer development. "Driver mutations occur in certain genes at a higher frequency than they would only by chance, but they are still rare in the patient population. That is because one patient will have a mutation in one gene and another patient with the same type of cancer will have a different mutation in another gene, but the two genes are part of the same pathway," says Campbell. "Each mutation is only found in a few percent of patients. But the pathway accounts for a large proportion of patients."

There are still many challenges before whole-genome sequencing goes mainstream, but most geneticists agree that these are exciting times. "Five years ago, it was not clear what path these next-generation sequencing technologies would take because they had not been tested," says Lifton. "It is quite astounding to be at this point now, where we can routinely get whole genome sequences of patients." And as sequencing costs continue to plummet, more and more labs are getting into the sequencing game. "Large projects are still the domain of genome centers, but we are seeing more groups starting to sequence and providing valuable information," says Gibbs. "We are seeing a healthy proliferation of sequencing projects."

Laura Bonetta
Washington DC
DOI 10.1016/j.cell.2010.05.034