When the World Wide Web launched in 1993, it seemed to catch on and spread overnight, unlike most new technologies, which typically take at least a decade to move from first “proof of concept” to broad acceptance. But the Web did not really emerge in a single year. It built on infrastructure, including the construction of the Internet between 1965 and 1993, as well as a sudden recognition that resources, such as personal computers, had passed a critical threshold.

Vision and market forces also push the development and spread of new technologies. The space program, for example, started with a government vision, and only much later did military and civilian uses for satellites propel the industry to commercial viability. Looking forward to the next technological revolution, which may be in

Next-generation technologies that make reading DNA fast, cheap and widely accessible are coming in less than a decade. Their potential to revolutionize research and bring about the era of truly personalized medicine means the time to start preparing is now.

By George M. Church
biotechnology, one can begin to imagine what markets, visions, discoveries and inventions may shape its outcome and what critical thresholds in infrastructure and resources will make it possible.

In 1984 and 1985, I was among a dozen or so researchers who proposed a Human Genome Project (HGP) to read, for the first time, the entire instruction book for making and maintaining a human being contained within our DNA. The project’s goal was to produce one full human genome sequence for $3 billion between 1990 and 2005.

We managed to finish the easiest 93 percent a few years early and to leave a legacy of useful technologies and methods. Their ongoing refinement has brought the street price of a human genome sequence accurate enough to be useful down to about $20 million today. Still, that rate means large-scale genic sequencing is mostly confined to dedicated sequencing centers and reserved for big, expensive research projects.

The “$1,000 genome” has become shorthand for the promise of DNA-sequencing capability made so affordable that individuals might think the once-in-a-lifetime expenditure to have a full personal genome sequence read to a disk for doctors to reference is worthwhile. Cheap sequencing technology will also make that information more meaningful by multiplying the number of researchers able to study genomes and the number of genomes they can compare to understand variations among individuals in both sickness and health.

“Human” genomics extends beyond humans, as well, to an environment full of pathogens, allergens and beneficial microbes in our food and our bodies. Many people attend to weather maps; perhaps we might one day benefit from daily pathogen and allergen maps. The rapidly growing fields of nanotechnology and industrial biotechnology, too, might accelerate their mining of biomes for new “smart” materials and microbes that can be harnessed for manufacturing or bioremediation of pollution.

The barrier to these applications and many more, including those we have yet to imagine, remains cost. Two National Institutes of Health funding programs for “Revolutionary Genome Sequencing Technologies” challenge scientists to achieve a $100,000 human genome by 2009 and a $1,000 genome by 2014. An X Prize–style cash reward for the first group to attain such benchmarks is also a possibility. And these goals are already close. A survey of the new approaches in development for reading genomes illustrates the potential for breakthroughs that could produce a $20,000 human genome as soon as four years from now—and brings to light some considerations that will arise once it arrives.

Reinventing Gene Reading

With any sequencing method, the size, structure and function of DNA itself can present obstacles or be turned...
advantages. The human genome is made up of three billion pairs of nucleotide molecules. Each of these contains one of four types of bases—abbreviated A, C, G and T—that represent a genomic alphabet encoding the information stored in DNA. Bases typically pair off according to strict rules to form the rungs in the ladderlike DNA structure. Because of these pairing rules, reading the sequence of bases along one half of the ladder reveals the complementary sequence on the other side as well.

Our three-billion-base-long genome is broken into 23 separate chromosomes. People usually have two full sets of these, one from each parent, that differ by 0.01 percent, so that an individual’s personal genome can really be said to contain six billion base pairs. Identifying individual bases in a stretch of the genome requires a sensor that can detect the subnanometer-scale differences between the four base types. Scanning tunneling microscopy is one physical method that can visualize these tiny structures and their subtle distinctions. For reading millions or billions of bases, however, most sequencing techniques rely at some stage on chemistry.

A method developed by Frederick Sanger in the 1970s became the workhorse of the HGP and is still the basis of most sequencing performed today. Sometimes described as sequencing by separation, the technique requires several rounds...
of duplication to produce large numbers of copies of the genome stretch of interest. The final round yields copy fragments of varying lengths, each terminating with a fluorescently tagged base. Separating these fragments by size in a process called electrophoresis, then reading the fluorescent signal of each terminal tag as it passes by a viewer, provides the sequence of bases in the original strand [see box on preceding two pages].

Reliability and accuracy are advantages of Sanger sequencing, although even with refinements over the years, the method remains time-consuming and expensive. Most alternative approaches to sequencing therefore seek to increase speed and reduce costs by cutting out the slow separation steps, miniaturizing components to reduce chemical volumes, and executing reactions in a massively parallel fashion so that millions of sequence fragments are read simultaneously.

Many research groups have converged on methods often lumped together under the heading of sequencing by synthesis because they exploit high-fidelity processes that living systems use to copy and repair their own genomes. When a cell is preparing to divide, for example, its DNA ladder splits into single strands, and an enzyme called polymerase moves along each
AMPLIFICATION

Because light signals are difficult to detect at the scale of a single DNA molecule, base-extension or ligation reactions are often performed on millions of copies of the same template strand simultaneously. Cell-free methods (a and b) for making these copies involve PCR on a miniaturized scale.

**Polonies**—polymerase colonies—created directly on the surface of a slide or gel each contain a primer, which a template fragment can find and bind to. PCR within each polony produces a cluster containing millions of template copies.

**Droplets** containing polymerase within an oil emulsion can serve as tiny PCR chambers to produce bead polonies. When a template fragment attached to a bead is added to each droplet, PCR produces 10 million copies of the template, all attached to the bead.

MULTIPLEXING

Sequencing thousands or millions of template fragments in parallel maximizes speed. A single-molecule base-extension system using fluorescent-signal detection, for example, places hundreds of millions of different template fragments on a single array (below left). Another method immobilizes millions of bead polonies on a gel surface for simultaneous sequencing by ligation of fragments on a single array (below right). Detection, for example, places hundreds of millions of different template copies unnecessary.

One cell-free amplification method, developed by Eric Kawashima of the Serono Pharmaceutical Research Institute in Geneva, Alexander Chetverin of the Russian Academy of Sciences, and Mitra when he was at Harvard, creates individual colonies of polymerase—polonies—freely arrayed directly on the surface of a microscope slide or a layer of gel. A single template molecule undergoes PCR within each polony, producing millions of copies, which grow rather like a bacterial colony from the central original template. Because each resulting polony cluster is one micron wide and one femtoliter in volume, billions of them can fit onto a single slide.

A variation on this system first produces polonies on tiny beads inside droplets within an emulsion. After the reaction millions of such beads, each bearing copies of a different template, can be placed in individual wells or immobilized by a gel where sequencing is performed on all of them simultaneously.

These methods of template amplification and of sequencing by base extension or by ligation are just a few representative examples of the approaches dozens of different academic and corporate research groups are taking to sequencing by synthesis.

Still another technique, sequencing by hybridization, also uses fluorescence to generate a visible signal and, like sequenc-
Like electrophoresis, this technique draws DNA toward a positive charge. To get there, the molecule must cross a membrane by going through a pore whose narrowest diameter of 1.5 nanometers will allow only single-stranded DNA to pass \( [a] \). As the strand transits the pore, nucleotides block the opening momentarily, altering the membrane’s electrical conductance, measured in picoamperes \( [pA] \). Physical differences between the four base types produce blockades of different degrees and durations \( [b] \). A close-up of a blockade event measurement shows a conductance change when a 150-nucleotide strand of a single base type passed through the pore \( [c] \).

Refining this method to improve its resolution to single bases could produce a sequence readout such as the hypothetical example at bottom \( [d] \) and yield a sequencing technique capable of reading a whole human genome in just 20 hours without expensive DNA copying steps and chemical reactions.

**THE AUTHOR**

George M. Church is professor of genetics at Harvard Medical School and director of the Harvard-Lipper Center for Computational Genetics, U.S. Department of Energy Genome Technology Laboratory, and the National Institutes of Health Centers of Excellence in Genomic Science. His research spans and integrates technologies for analyzing and synthesizing biomolecules and cells. He holds 10 U.S. patents and has been scientific adviser to more than 20 companies.
amplification followed by base-extension sequencing with pyrophosphate detection in an array of wells. Both groups read about the same amount of sequence, 30 million base pairs, in each sequencing run. Our system read about 400 base pairs a second, whereas 454 read 1,700 a second. Sequencing usually involves performing multiple runs to produce a more accurate consensus sequence. With 43-times coverage (43x)—that is, 43 runs per base—of the target genome, 454 achieved accuracy of one error per 2,500 base pairs. The Harvard group had less than one error per three million base pairs with 7x coverage. To handle templates, both teams employed capture beads, whose size affects the amount of expensive reagents consumed. Our beads were one micron in diameter, whereas 454 used 28-micron beads in 75-picoliter wells.

THE PERSONAL GENOME PROJECT

Every baby born in the U.S. today is tested for at least one genetic disease, phenylketonuria, before he or she leaves the hospital. Certain lung cancer patients are tested for variations in a gene called EGFR to see if they are likely to respond to the drug Iressa. Genetic tests indicating how a patient will metabolize other drugs are increasingly used to determine the drugs’ dosage. Beginnings of the personalized medicine that will be possible with low-cost personal genomes can already be glimpsed, and demand for it is growing.

Beyond health concerns, we also want to know our genealogy. How closely are we related to Genghis Khan or to each other? We want to know what interaction of genes with other genes and with the environment shapes our faces, our bodies, our dispositions. Thousands or millions of data sets comprising individuals’ whole genome and phenome—the traits that result from instructions encoded in the genome—that will make it possible to start unraveling some of those complex pathways.

Yet the prospect of this new type of personal information suddenly becoming widely available also prompts worries about how it might be misused—by insurers, employers, law-enforcement agents, friends, neighbors, commercial interests or criminals. No one can predict what living in an era of personal genomics will be like until the waters are tested. That is why my colleagues and I recently launched the Personal Genome Project (PGP). With this natural next step after the Human Genome Project, we hope to explore possible rewards and risks of personal genomics by recruiting volunteers to make their own genome and phenome data openly available.

GEORGE M. CHURCH, shown with images of fluorescent polonies, is one of a group of volunteers planning to open their genomes to public scrutiny.

These resources will include full (46-chromosome) genome sequences, digital medical records, as well as information that could one day be part of a personal health profile, such as comprehensive data about RNA and proteins, body and facial measurements, and MRI and other cutting-edge imagery. We will also create and deposit human cell lines representing each subject in the Coriell repository of the National Institute of General Medical Sciences. Our purpose is to make all this genomic and trait information broadly accessible so that anyone can mine it to test their own hypotheses and algorithms—and be inspired to come up with new ones.

A recent incident provides a simple example of what might happen. A few PGP medical records—my own—are already publicly available online, which prompted a hematologist on the other side of the country to notice, and inform me, that I was long overdue for a follow-up test of my cholesterol medication. The tip led to a change in my dose and diet and consequently to a dramatic lowering of at least one type of risk. In the future this kind of experience would not rely on transcontinental serendipity but could spawn a new industry of third-party genomic software tools.

The PGP has approval from the Harvard Medical School Internal Review Board, and like all human research subjects, participants must be informed of potential risks before consenting to provide their data. Every newly recruited PGP volunteer will also be able to review the experience of previous subjects before giving informed consent. The project’s open nature, including fully identifying subjects with their data, will be less risky both to the subjects and the project than the alternative of promising privacy and risking accidental release of information or access by hackers.

Like the free data access policy established by the HGP, the openness of the PGP is designed to maximize potential for discovery. In addition to providing a scientific resource, the project also offers an experiment in public access and insurance coverage. In its early stages, private donors will help to insure a diverse set of human subjects against the event that they experience genetic discrimination as a consequence of the PGP. This charity-driven mechanism has the advantage of not needing to be profitable at first, but insurance companies may nonetheless be very interested in its outcome. —G.M.C.

Details of the PGP can be found at http://arep.med.harvard.edu/PGP/
The best available electrophoresis-based sequencing methods average 150 base pairs per dollar for “finished” sequence. The 454 group did not publish a project cost, but the Harvard team’s finished sequence cost of 1,400 base pairs per $1 represents a ninefold reduction in price.

These and other new techniques are expected very soon to bring the cost of sequencing the six billion base pairs of a personal genome down to $100,000. For any next-generation sequencing method, pushing costs still lower will depend on a few fundamental factors. Now that automation is commonplace in all systems, the biggest expenditures are for chemical reagents and equipment. Miniaturization has already reduced reagent use relative to conventional Sanger reactions one billionfold from microliters to femtoliters.

Many analytic imaging devices can collect raw data at rates of one billion bytes (a gigabyte) per minute, and computers can process the information at a speed of several billion operations a second. Therefore, any imaging device limited by a slow physical or chemical process, such as electrophoresis or enzymatic reaction, or one that is not tightly packed in space and time, making every pixel count, will be correspondingly more costly to operate per unit DNA base determined.

Another consideration in judging emerging sequencing technologies is how they will be used. Newer methods tend to have short read-lengths of five to 400 base pairs, compared with typical Sanger read-lengths of 800 base pairs. Sequencing and piecing together a previously unknown genome from scratch is therefore much harder with the new techniques. If medicine is the primary driver of widespread sequencing, however, we will be largely resequencing the human genome looking for minute variations in individuals’ DNA, and short read-lengths will not be such a problem.

Accuracy requirements will also be a function of the applications. Diagnostic uses might demand a reduction in error rates below the current HGP standard of 0.01 percent, because that still permits 600,000 errors per human genome. At the other end of the spectrum, high-error-rate (4 percent) random sampling of the genome has proved useful for discovery and classification of various RNA and tissue types. A similar “shotgun” strategy is applied in ecological sampling, where as few as 20 base pairs are sufficient to identify an organism in an ecosystem.

Raising Value

Beyond developing these new sequencing technologies, we have much work to do in a short amount of time to get ready for the advent of low-cost genome reading. Software will be needed to process sequence information so that it is manageable by doctors, for example. They will need a method to derive an individualized priority list for each patient of the top 10 or so genetic variations likely to be important. Equally essential will be assessing the effects of widespread access to this technology on people.

From its outset, the HGP established a $10-million-a-year program to study and address the ethical, legal and social issues that would be raised by human genome sequencing. Participants in the effort agreed to make all our data publicly available with unprecedented speed—within one week of discovery—and we rose to fend off attempts to commercialize human nature. Special care was also taken to protect the anonymity of the public genomes (the “human genome” we produced is a mosaic of several people’s chromosomes). But many of the really big questions remain, such as how to ensure privacy and fairness in the use of personal genetic information by scientists, insurers, employers, courts, schools, adoption agencies, the government, or individuals making clinical and reproductive decisions.

These difficult and important questions need to be researched as rigorously as the technological and biological discovery aspects of human genomics. My colleagues and I have therefore initiated a Personal Genome Project [see box on preceding page] to begin exploring the potential risks and rewards of living in an age of personal genomics.

When we invest in stocks or real estate or relationships, we understand that nothing is a sure thing. We think probabilistically about risk versus value and accept that markets, like life, are complex. Just as personal digital technologies have caused economic, social and scientific revolutions unimagined when we had our first few computers, we must expect and prepare for similar changes as we move forward from our first few genomes.

MORE TO EXPLORE