Sequencing Genomes and Beyond

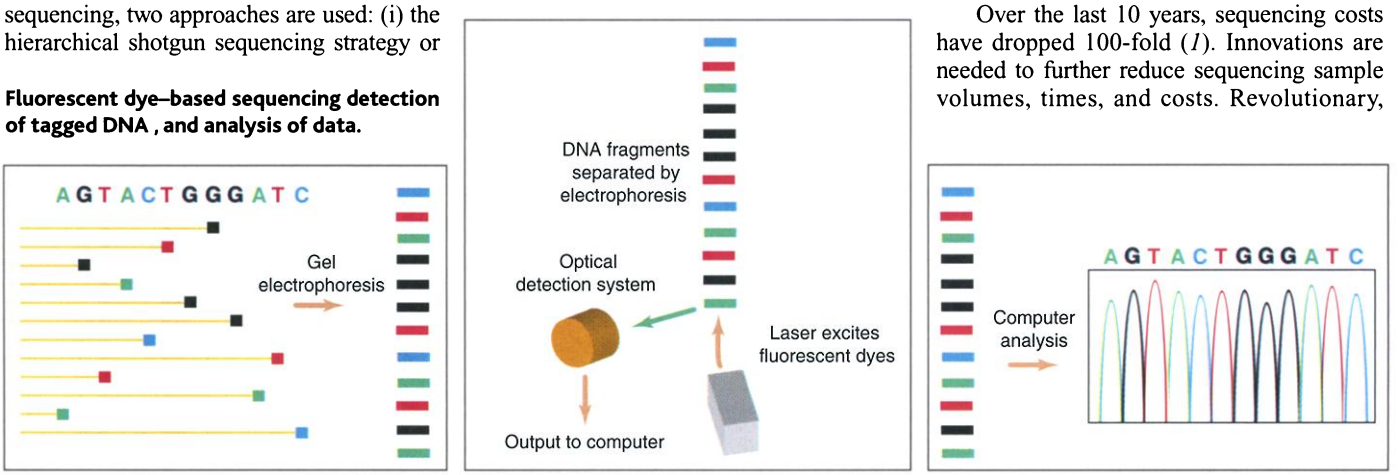
Deirdre R. Meldrum

The completion of the working draft sequence of the human genome, an achievement as significant as landing man on the moon, was announced 26 June 2000 and published 16 February 2001 (*1*, *2*). This feat was made possible by technological advances in DNA sequencing, the process of determining the order of the individual chemical bases (A, T, G, and C) in a DNA molecule.

Three key milestones—the invention of sequencing reactions (*3*, *4*), the polymerase chain reaction (PCR) (*5*), and automated fluorescent DNA sequencers (*6*, *7*)—made it possible to streamline and automate most of the processes required for DNA sequence analysis. The recent introduction of capillary sequencers, which can process 96 samples in parallel, has been central in producing most of the working draft sequence. With these systems, sequencing capacity and output rose approximately eightfold in 8 months to a rate exceeding 1000 nucleotides per second, 24 hours a day, seven days a week (*1*).

DNA sequence analysis is a multistage process that includes the preparation of DNA, its fragmentation and base analysis, and the interpretation of the resulting sequence (*8*). For large-scale genome sequencing, two approaches are used: (i) the hierarchical shotgun sequencing strategy or "clone-by-clone" approach used by the International Human Genome Sequencing Consortium (*1*) and (ii) the whole-genome shotgun approach used by Celera Genomics (*2*).

Fluorescent dye–based sequencing detection of tagged DNA , and analysis of data.



CREDIT: ADAPTED FROM FIGURES PROVIDED BY E. GREEN, NATIONAL HUMAN GENOME RESEARCH INSTITUTE

In the clone-by-clone approach, the genome is subdivided and replicated in segments of about 150,000 bases using bacterial artificial chromosomes (BACs). Genome mapping methods determine the physical placement of each clone in the genome. Next, the clones are fragmented into smaller pieces of about 500 to 1000 bases for sequencing. By contrast, in the whole-genome shotgun method small clones are prepared directly from genomic DNA rather than from mapped BACs. This approach is potentially faster than the clone-by-clone method but necessitates a more challenging assembly process after sequencing is done (*8*).

The author is in the Department of Electrical Engineering, University of Washington, Seattle, WA 98195–2500, USA. E-mail: deedee@kant.ee.washington.edu

In both the clone-by-clone approach and the whole-genome shotgun method, gel electrophoresis is used to determine the order of bases on a strand. DNA molecules (see figure below) that differ in size by one base are labeled with nucleotide-specific fluorescent dye and loaded onto a slab of acrylamide gel. An applied voltage causes the DNA fragments to separate by size as they migrate through the gel. At the bottom of the apparatus, a detector, such as a charge-coupled device (CCD) or photomultiplier tube (PMT), reads the color of the fluorescent dye at the end of each DNA fragment (*6*, *7*).

The Applied Biosystems PRISM 3700 and the Molecular Dynamics MegaBACE 1000 capillary sequencers have enabled the rapid DNA analysis used for the determination of the human sequence. These instruments, which incorporate capillary tubes to hold the sequence gel, automate sample loading, separation matrix loading and removal, data collection and analysis, and provide significant improvement over the manual gel preparation and lane tracking required in slab-gel sequencers. Sequencing time per sample is about 10 hours in slab gels but only 3 hours in capillaries. These sequencers maintain efficiency with high-throughput, upstream automated sample preparation and sample tracking throughout the process to avoid mix-up, and real-time quality control.

Knowing the DNA sequence of humans—3.2 gigabases in all—is just the beginning of our endeavor to understand disease and transform medicine. Knowledge of gene function in one species will likely be applicable to homologous genes in other species.

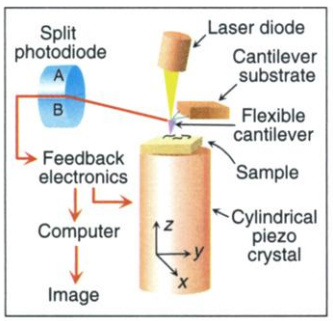
Over the last 10 years, sequencing costs have dropped 100-fold (*1*). Innovations are needed to further reduce sequencing sample volumes, times, and costs. Revolutionary, non–gel-based methods, such as mass spectrometry, single-molecule sequencing, and nanopore technology, hold great promise for simplifying the method and improving our ability to determine sequences (*9*). However, each has shortcomings that must be overcome to enable determination of sequences 700 to 1000 nucleotides long.

Matrix-assisted laser desorption ionization (MALDI) time-of-flight (TOF) mass spectrometry has been adapted primarily for detecting sequence variations between individuals. It uses the chemistry established for conventional sequencing but replaces the size separation of DNA in gels with mass-dependent strand separation of gas phase ions in a vacuum. The chemistry specifies the base at the end of a fragment, and the mass of the fragment specifies the location of the corresponding base in the sequence. This approach, much like electrophoresis in a vacuum (*10*), permits simultaneous analysis of many DNA strands in seconds. Although the technique is fast, the sequence length that can be achieved currently is only about 100 bases.

Pyrosequencing is a nonelectrophoretic DNA sequencing method used primarily for mutation analysis and genotyping. It enables sequencing of DNA strands up to 200 nucleotides long and is currently one of the fastest methods for analyzing a primed DNA strand. The technique takes advantage of four enzymes [DNA polymerase, sulfurylase (from yeast), luciferase (from the firefly), and apyrase (from potato)] cooperating in a single tube to signal the incorporation of a nucleotide to a growing DNA strand. Detection is based on the visible light produced by coupling the pyrophosphate released during nucleotide incorporation with the enzymes sulfurylase and luciferase. By sequentially adding nucleotides and observing the flash of light each addition causes, the sequence of the template can be determined as the DNA strand is copied (*11*).

An important consideration in increasing sequencing efficiency is reducing fluid volumes of the sequencing and separation steps to reduce costs and processing speeds. The Genomation Laboratory at the University of Washington (http://rcs.ee.washington.edu/GNL/genomation.html) is developing an instrument that prepares 5000 microliter samples in 8 hours for high-throughput large-scale DNA sequencing. The reaction preparation module of this automated, multipurpose, fluid-handling system aspirates submicroliter DNA samples, delivers precise picoliter volumes of reagents using piezoelectric dispensers, and mixes fluids (*9*).

The next generation of electrophoresis-based sequencers will most likely comprise microchannel plates (*12*) or "microchips." The advantages of microchip sequencers over capillary electrophoresis systems include more efficient sample injection; higher speed, resolution, throughput, and detection sensitivity; lower costs and sample and reagent consumption; and the potential integration on a single chip with upstream sample processing steps (*9*). Mathies's (*13*) rotary design runs 96 samples in parallel on individual channels in a radial configuration, like the spokes of a wheel. Samples are detected in the center of the microplate by a laser-excited rotary confocal scanner and reduces sequencing times from hours to minutes.

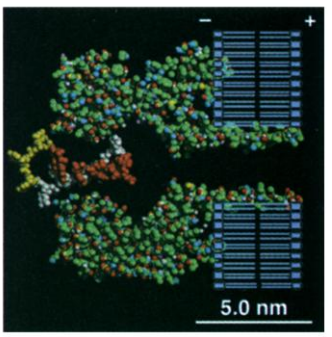


Schematic of AFM.

An integrated, microfabricated system or "lab-on-a-chip" under development by researchers at the University of Michigan prepares DNA samples and sequences them. It includes a system to control fluid motion, an air-driven fluid pump, a temperature-controlled reaction chamber, and an electrophoresis and fluorescence detection system. All of the components are microfabricated on the same wafer and require no external lenses, heaters, or mechanical pumps (*14*). Plastic electrophoresis chips have also been constructed on polycarbonate substrates and have successfully separated double-stranded DNA. This technology has the potential for low-cost fabrication of disposable, single-use electrophoresis devices (*15*).

Caliper Technologies recently released their AMS 90 or "LabChip" that takes 96 DNA samples directly from microplates and performs electrophoretic size-based separation on a microchip. Small (10 nl) samples can be processed in 5 to 30 s. An Agilent Technologies Bioanalyzer 2100 detects the separated DNA bands using laser-induced fluorescence (*16*).

DNA sequencing by size fractionation is generally a slow and indirect method of obtaining sequence data. In the future, detection and analysis of single molecules may provide a very rapid, low-cost DNA sequencing option. One technique being developed may allow direct analysis of individual DNA molecules at rates of up to 1000 bases per second. Bases on a single-stranded DNA molecule are forced, under an electric potential difference, single-file through a nanopore less than 2 nm in diameter (see figure, right). An integral detector in the pore translates the characteristic physical and chemical properties of a base or sequence of bases into an electrical signature. When no DNA is present, the pore exhibits an ionic current of 120 pA at 120 mV potential. When the pore is occupied by polynucleotides, the ionic conductance decreases according to the nucleotide composition of the DNA. To achieve single-nucleotide sequence resolution, several improvements are needed, not least of which includes the coupling of more sensitive detectors to the pores to discern single nucleotides in the signature (*16–19*).



DNA hairpin molecule in an alpha-hemolysin channel nanopore detector.

Another single-molecule analysis technique, atomic force microscopy (AFM), involves scanning a nanometer-scale tip across a surface (*20*) to read the surface of DNA, much as a blind person reads Braille (*10*). Intermolecular forces between the tip and surface move a flexible cantilever up and down (see figure, left). The corresponding deflections are measured with a laser, and a topographic map of the surface is generated. The ultimate goal of AFM is to scan single-stranded DNA with single-nucleotide resolution.

Nanopore and AFM sequencing have the potential to sequence very long DNA molecules, obviating the need for cloning, fragmenting, and reassembling used with current approaches. Thus, they would also dramatically reduce the need for preparation of large numbers of samples.

DNA sequencing throughput and analysis will continue to accelerate as technologies are developed. Currently, electrophoretic methods yield the longest sequence reads and will continue to be used for de novo sequencing (*19*). Information produced by newly developing technologies—microchips, biochips, and nonelectrophoretic methods—will enhance genetic medicine by providing rapid diagnoses, individually tailored treatments, and new ways to prevent disease. With the sequence of the human genome in silico, we are at the same stage that chemistry was with the unveiling of the Periodic Table. The Biological Periodic Table provides an inventory of all genes used to assemble a living creature (*21*). Global analysis of this information has just begun; exciting revelations lie ahead.

References

1. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
2. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
3. A. M. Maxam, W. Gilbert, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 560 (1977).
4. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977).
5. K. B. Mullis *et al.*, *Cold Spring Harbor Symp. Quant. Biol.* **51**, 263 (1986).
6. L. M. Smith *et al.*, *Nature* **321**, 674 (1986).
7. T. Hunkapiller, R. J. Kaiser, B. F. Koop, L. Hood, *Science* **254**, 59 (1991).
8. M. V. Olson, *Nature* **409**, 816 (2001).
9. D. R. Meldrum, *Genome Res.* **10**, 1081 (2000); ———, *Genome Res.* **10**, 1288 (2000).
10. C. R. Cantor, C. L. Smith, *Genomics* (John Wiley & Sons, New York, 1999).
11. M. Ronaghi, *Genome Res.* **11**, 3 (2001).
12. L. Koutny *et al.*, *Anal. Chem* **72**, 3388 (2000).
13. Y. Shi *et al.*, *Anal. Chem.* **71**, 5354 (1999).
14. M. A. Burns *et al.*, *Science* **282**, 484 (1998).
15. J. R. Webster, M. A. Burns, D. T. Burke, C. H. Mastrangelo, in *Micro total analysis systems '98*, D. J. Harrison, A. van den Berg, Eds. (Kluwer, Boston, 1998), pp. 249–252.
16. J. Boguslavsky, *Drug Discovery Devel.* **January**, 71 (2001).
17. J. Kasianowicz, E. Brandin, D. Branton, D.W. Deamer, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13770 (1996).
18. D. W. Deamer, M. Akeson, *Trends Biotechnol.* **18**, 147 (2000).
19. A. Marziali, M. Akeson, *Annu. Rev. Biomed. Eng.*, in press.
20. J. Binnig *et al.*, *Phys. Rev. Lett.* **56**, 930 (1986).
21. E. Lander, *Nature Genet.* **21** (Suppl.), 3 (1999).