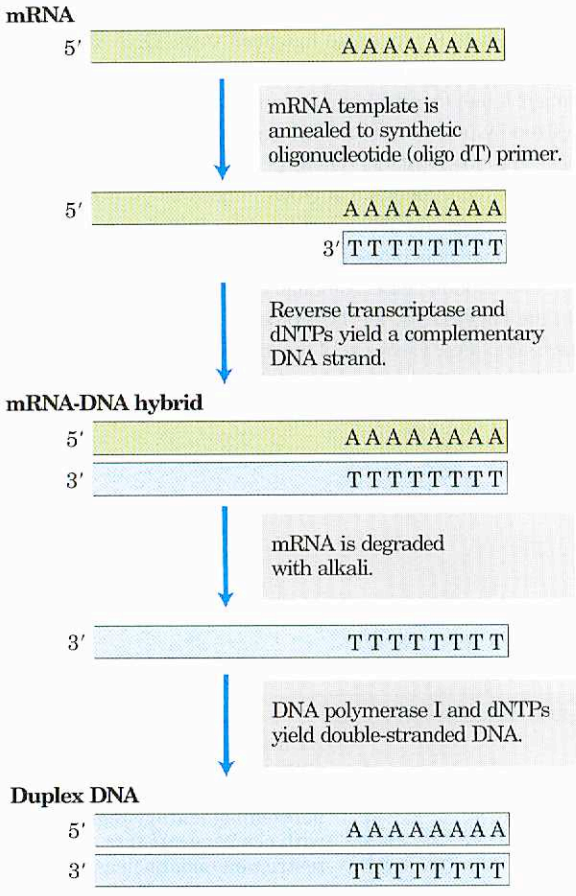


FIGURE 9–14 Construction of a cDNA library from mRNA. A cell's mRNA includes transcripts from thousands of genes, and the cDNAs generated are correspondingly heterogeneous. The duplex DNA produced by this method is inserted into an appropriate cloning vector. Reverse transcriptase can synthesize DNA on an RNA or a DNA template (see Fig. 26–29).

gene fused with a gene for **green fluorescent protein (GFP)** generates a fusion protein that is highly fluorescent—it literally lights up (Fig. 9–15a). Just a few molecules of this protein can be observed microscopically, allowing the study of its location and movements in a cell. An **epitope tag** is a short protein sequence that is bound tightly by a well-characterized monoclonal antibody (Chapter 5). The tagged protein can be specifically precipitated from a crude protein extract by interaction with the antibody (Fig. 9–15b). If any other proteins bind to the tagged protein, those will precipitate as well, providing information about protein-protein interactions in a cell. The diversity and utility of specialized DNA libraries are growing every year.

The Polymerase Chain Reaction Amplifies Specific DNA Sequences

The Human Genome Project, along with the many associated efforts to sequence the genomes of organisms of every type, is providing unprecedented access to gene sequence information. This in turn is simplifying the process of cloning individual genes for more detailed biochemical analysis. If we know the sequence of at least the flanking parts of a DNA segment to be cloned, we can hugely amplify the number of copies of that DNA segment, using the **polymerase chain reaction (PCR),** a process conceived by Kary Mullis in 1983. The amplified DNA can be cloned directly or used in a variety of analytical procedures.

The PCR procedure has an elegant simplicity. Two synthetic oligonucleotides are prepared, complementary to sequences on opposite strands of the target DNA at positions just beyond the ends of the segment to be amplified. The oligonucleotides serve as replication primers that can be extended by DNA polymerase. The 3′ ends of the hybridized probes are oriented toward each other and positioned to prime DNA synthesis across the desired DNA segment (Fig. 9–16). (DNA polymerases

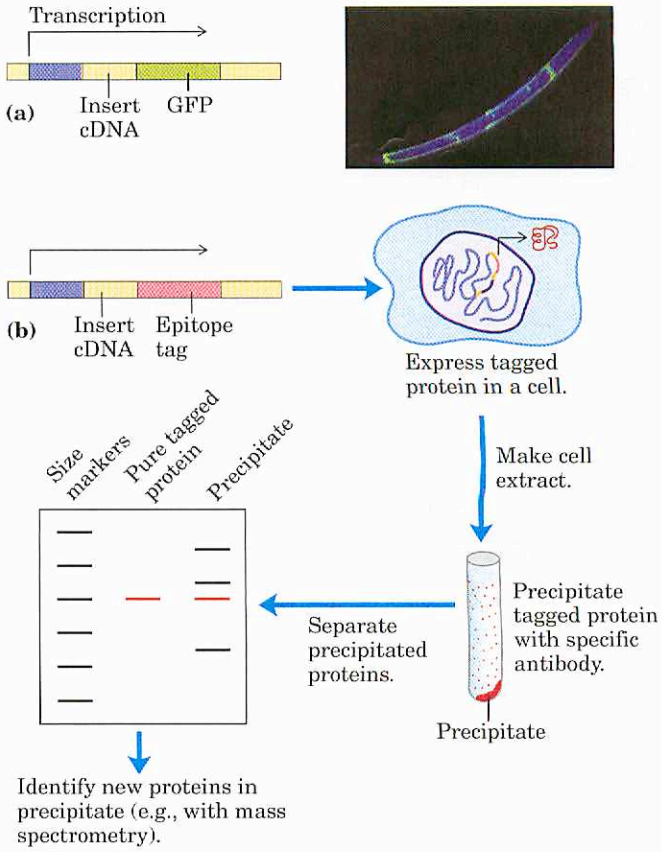
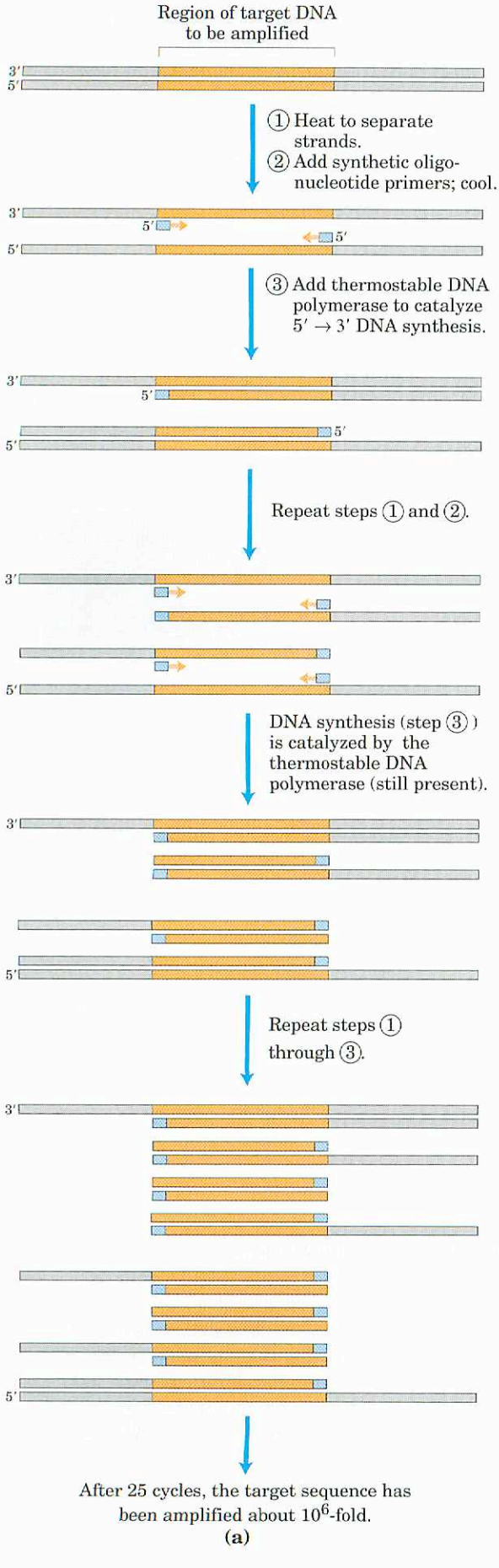


FIGURE 9–15 Specialized DNA libraries. (a) Cloning of cDNA next to a gene for green fluorescent protein (GFP) creates a reporter construct. RNA transcription proceeds through the gene of interest (insert DNA) and the reporter gene, and the mRNA transcript is then expressed as a fusion protein. The GFP part of the protein is visible in the fluorescence microscope. The photograph shows a nematode worm containing a GFP fusion protein expressed only in the four "touch" neurons that run the length of its body. Reporter Constructs **(b)** If the cDNA is cloned next to a gene for an epitope tag, the resulting fusion protein can be precipitated by antibodies to the epitope. Any other proteins that interact with the tagged protein also precipitate, helping to elucidate protein-protein interactions.



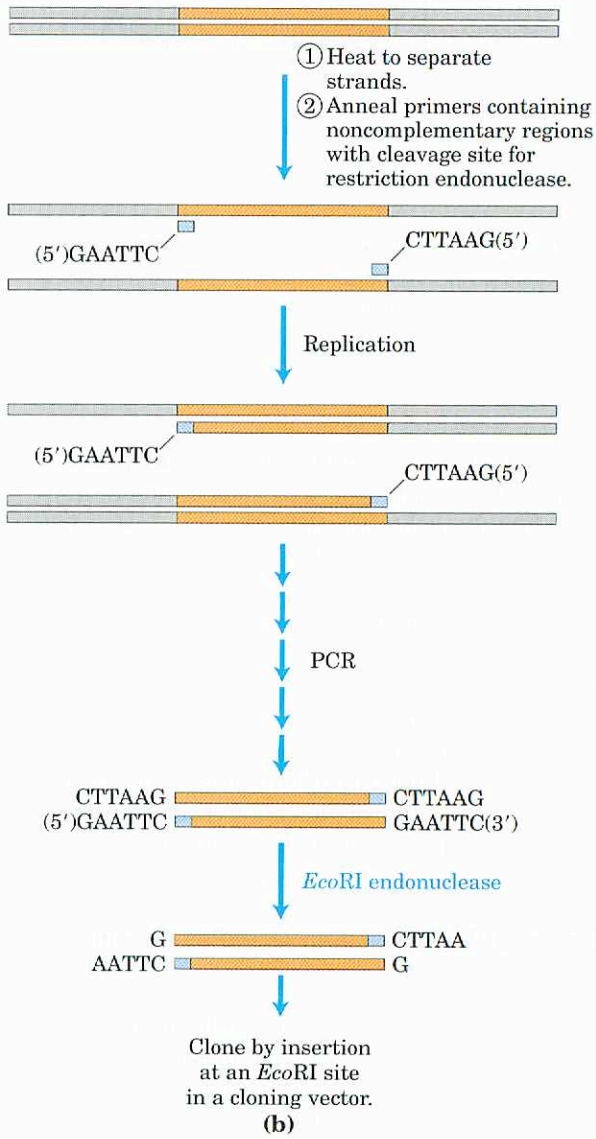


FIGURE 9-16 Amplification of a DNA segment by the polymerase chain reaction. **(a)** The PCR procedure has three steps. DNA strands are ① separated by heating, then ② annealed to an excess of short synthetic DNA primers (blue) that flank the region to be amplified; ③ new DNA is synthesized by polymerization. The three steps are repeated for 25 or 30 cycles. The thermostable DNA polymerase *Taq*I (from *Thermus aquaticus*, a bacterial species that grows in hot springs) is not denatured by the heating steps. **(b)** DNA amplified by PCR can be cloned. The primers can include noncomplementary ends that have a site for cleavage by a restriction endonuclease. Although these parts of the primers do not anneal to the target DNA, the PCR process incorporates them into the DNA that is amplified. Cleavage of the amplified fragments at these sites creates sticky ends, used in ligation of the amplified DNA to a cloning vector. Polymerase Chain Reaction

synthesize DNA strands from deoxyribonucleotides, using a DNA template, as described in Chapter 25.) Isolated DNA containing the segment to be amplified is heated briefly to denature it, and then cooled in the presence of a large excess of the synthetic oligonucleotide primers. The four deoxynucleoside triphosphates are then added, and the primed DNA segment is replicated selectively. The cycle of heating, cooling, and replication is repeated 25 or 30 times over a few hours in an automated process, amplifying the DNA segment flanked by the primers until it can be readily analyzed or cloned. PCR uses a heat-stable DNA polymerase, such as the *Taq* polymerase (derived from a bacterium that lives at 90 °C), which remains active after every heating step and does not have to be replenished. Careful design of the primers used for PCR, such as including restriction endonuclease cleavage sites, can facilitate the subsequent cloning of the amplified DNA (Fig. 9–16b).

This technology is highly sensitive: PCR can detect and amplify as little as one DNA molecule in almost any type of sample. Although DNA degrades over time (p. 293), PCR has allowed successful cloning of DNA from samples more than 40,000 years old. Investigators have used the technique to clone DNA fragments from the mummified remains of humans and extinct animals such as the woolly mammoth, creating the new fields of molecular archaeology and molecular paleontology. DNA from burial sites has been amplified by PCR and used to trace ancient human migrations. Epidemiologists can use PCR-enhanced DNA samples from human remains to trace the evolution of human pathogenic viruses. Thus, in addition to its usefulness for cloning DNA, PCR is a potent tool in forensic medicine (Box 9–1). It is also being used for detection of viral infections before they cause symptoms and for prenatal diagnosis of a wide array of genetic diseases.

The PCR method is also important in advancing the goal of whole genome sequencing. For example, the mapping of expressed sequence tags to particular chromosomes often involves amplification of the EST by PCR, followed by hybridization of the amplified DNA to clones in an ordered library. Investigators found many other applications of PCR in the Human Genome Project, to which we now turn.

Genome Sequences Provide the Ultimate Genetic Libraries

The genome is the ultimate source of information about an organism, and there is no genome we are more interested in than our own. Less than 10 years after the development of practical DNA sequencing methods, serious discussions began about the prospects for sequencing the entire 3 billion base pairs of the human genome. The international Human Genome Project got underway with substantial funding in the late 1980s. The effort eventually included significant contributions from 20 sequencing centers distributed among six nations: the United States, Great Britain, Japan, France, China, and Germany. General coordination was provided by the Office of Genome Research at the National Institutes of Health, led first by James Watson and after 1992 by Francis Collins. At the outset, the task of sequencing a 3×10^9 bp genome seemed to be a titanic job, but it gradually yielded to advances in technology. The completed sequence of the human genome was published in April 2003, several years ahead of schedule.

This advance was the product of a carefully planned international effort spanning 14 years. Research teams first generated a detailed physical map of the human genome, with clones derived from each chromosome organized into a series of long contigs (Fig. 9–17). Each contig contained landmarks in the form of STSs at a distance of every 100,000 bp or less. The genome thus mapped could be divided up between the international sequencing centers, each center sequencing the mapped BAC or YAC clones corresponding to its particular segments of the genome. Because many of the

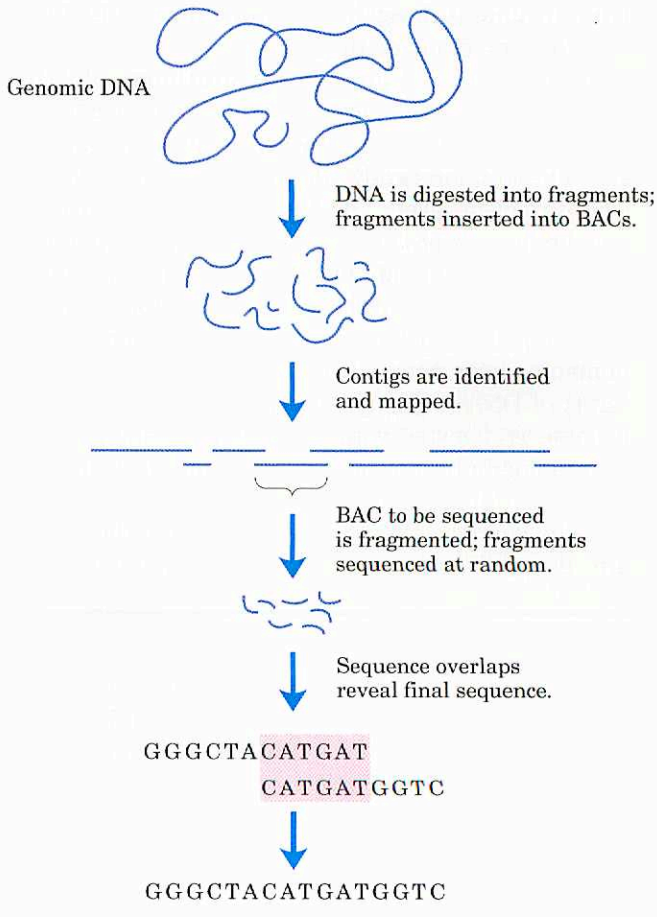


FIGURE 9–17 The Human Genome Project strategy. Clones isolated from a genomic library were ordered into a detailed physical map, then individual clones were sequenced by shotgun sequencing protocols. The strategy used by the commercial sequencing effort eliminated the step of creating the physical map and sequenced the entire genome by shotgun cloning.

BOX 9–1 WORKING IN BIOCHEMISTRY

A Potent Weapon in Forensic Medicine

Traditionally, one of the most accurate methods for placing an individual at the scene of a crime has been a fingerprint. With the advent of recombinant DNA technology, a more powerful tool is now available: **DNA fingerprinting** (also called DNA typing or DNA profiling).

DNA fingerprinting is based on **sequence polymorphisms,** slight sequence differences (usually single base-pair changes) between individuals, 1 bp in every 1,000 bp, on average. Each difference from the prototype human genome sequence (the first one obtained) occurs in some fraction of the human population; every individual has some differences. Some of the sequence changes affect recognition sites for restriction enzymes, resulting in variation in the size of DNA fragments produced by digestion with a particular restriction enzyme. These variations are **restriction fragment length polymorphisms (RFLPs).**

The detection of RFLPs relies on a specialized hybridization procedure called **Southern blotting** (Fig. 1). DNA fragments from digestion of genomic DNA by restriction endonucleases are separated by size electrophoretically, denatured by soaking the agarose gel in alkali, and then blotted onto a nylon membrane to reproduce the distribution of fragments in the gel. The membrane is immersed in a solution containing a radioactively labeled DNA probe. A probe for a sequence that is repeated several times in the human genome generally identifies a few of the thousands of DNA fragments generated when the human genome is digested with a restriction endonuclease. Autoradiography reveals the fragments to which the probe hybridizes, as in Figure 9–9.

The genomic DNA sequences used in these tests are generally regions containing repetitive DNA (short sequences repeated thousands of times in tandem), which are common in the genomes of higher eukaryotes (see Fig. 24–8). The number of repeated units in these DNA regions varies among individuals (except between identical twins). With a suitable probe, the pattern of bands produced by DNA fingerprinting is distinctive for each individual. Combining the use of several probes makes the test so selective that it can positively identify a single individual in the entire human population. However, the Southern blot procedure requires relatively fresh DNA samples and larger amounts of DNA than are generally present at a crime scene. RFLP analysis sensitivity is augmented by using PCR (see Fig. 9–16a) to amplify vanishingly small amounts of DNA. This allows investigators to obtain DNA fingerprints from a single hair follicle, a drop of blood, a small semen sample from a rape victim, or samples that might be months or even many years old.

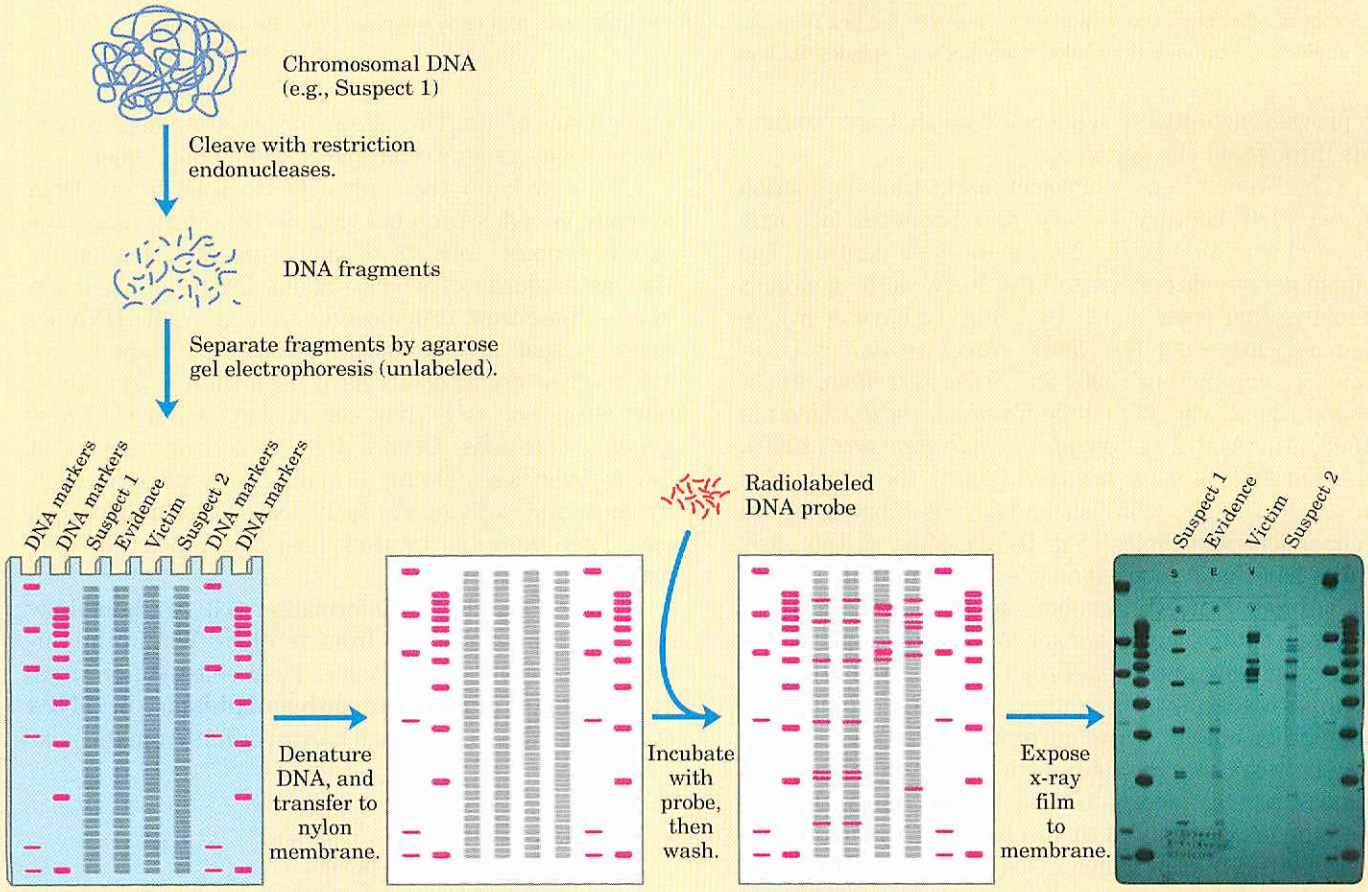
These methods are proving decisive in court cases worldwide. In the example in Figure 1, the DNA from a semen sample obtained from a rape and murder victim was compared with DNA samples from the victim and two suspects. Each sample was cleaved into fragments and separated by gel electrophoresis. Radioactive DNA probes were used to identify a small subset of fragments that contained sequences complementary to the probe. The sizes of the identified fragments varied from one individual to the next, as seen here in the different patterns for the three individuals (victim and two suspects) tested. One suspect's DNA exhibits a banding pattern identical to that of a semen sample taken from the victim. This test used a single probe, but three or four different probes would be used (in separate experiments) to achieve an unambiguous positive identification.

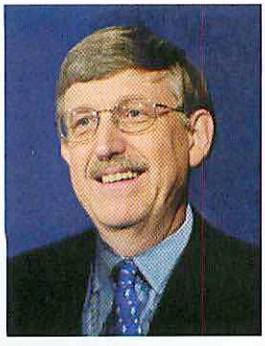
clones were more than 100,000 bp long, and modern sequencing techniques can resolve only 600 to 750 bp of sequence at a time, each clone had to be sequenced in pieces. The sequencing strategy used a shotgun approach, in which researchers used powerful new automated sequencers to sequence random segments of a given clone, then assembled the sequence of the entire clone by computerized identification of overlaps. The number of clone pieces sequenced was determined statistically so that the entire length of the clone was sequenced four to six times on average. The sequenced DNA was then made available in a database covering the entire genome. Construction of the physical map was a time-consuming task, and its progress was followed in annual reports in major journals throughout the 1990s—by the end of which the map was largely in place. Completion of the entire sequencing project was initially projected for the year 2005, but circumstances and technology intervened to accelerate the process.

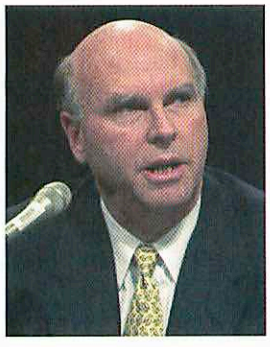
A competing commercial effort to sequence the human genome was initiated by the newly established Celera Corporation in 1997. Led by J. Craig Venter, the Celera group made use of a different strategy called "whole genome shotgun sequencing," which eliminates the step of assembling a physical map of the genome. Instead, teams sequenced DNA segments from through-

Such results have been used to both convict and acquit suspects and, in other cases, to establish paternity with an extraordinary degree of certainty. The impact of these procedures on court cases will continue to grow as societies agree on the standards and as formal methods become widely established in forensic laboratories. Even decades-old murder mysteries can be solved: in 1996, DNA fingerprinting helped to confirm the identification of the bones of the last Russian czar and his family, who were assassinated in 1918.

FIGURE 1 The Southern blot procedure, as applied to DNA fingerprinting. This procedure was named after Jeremy Southern, who developed the technique.




Francis S. Collins


J. Craig Venter

out the genome at random. The sequenced segments were ordered by the computerized identification of sequence overlaps (with some reference to the public project's detailed physical map). At the outset of the Human Genome Project, shotgun sequencing on this scale had been deemed impractical. However, advances in computer software and sequencing automation had made the approach feasible by 1997. The ensuing race between the private and public sequencing efforts substantially advanced the timeline for completion of the project. Publication of the draft human genome sequence in 2001 was followed by two years of follow-up work to eliminate nearly a thousand discontinuities and