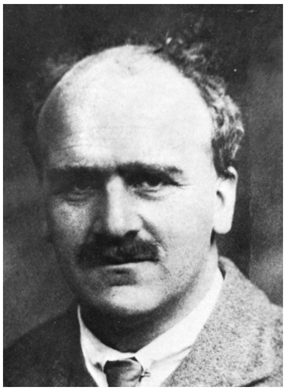
CHAPTER 1

# Genome as Information

Hans Winkler

(Reprinted, with permission, from Brabec F. 1955. Berichte der Deutschen Botanischen Gesellschaft 68: 27, ©Wiley-Blackwell; courtesy of Hunt Institute for Botanical Documentation, Carnegie Mellon University, Pittsburgh.)



J.B.S. Haldane

(Reprinted from Clark RW, 1969. *J.B.S.: The life and work of J.B.S. Haldane,* ©Coward-McCann.)

The term genome was coined in 1920 by the German botanist Hans Winkler. A combination of the words *gen*e and chromos*ome*, a genome is the set of genes, located on one or more chromosomes, that defines a living organism. The human genome, for example, is composed of ~25,000 genes that encode proteins needed to carry out the processes of life within the several trillion cells that make up our bodies.

Also working at the beginning of the 20th century, the American geneticist Thomas Hunt Morgan provided an enduring mental picture of the genome as a collection of genes arranged on each chromosome like beads on a string. With the realization that a chromosome is a linear DNA molecule, the concept of "genome" has been expanded to mean the entire sequence of DNA nucleotides or "letters" (A, T, C, and G) that compose the haploid (half set) of chromosomes of an individual. With advances in DNA sequencing technology during the past 30 years, we can now rapidly determine the entire nucleotide sequence of any organism. For humans, this amounts to ~3.2 billion "letters" in the set of chromosomes inherited from one's mother or father.

Although much work is focused on decoding genes that specify proteins, genes also specify several types of RNA molecules that are not translated into proteins. Many unexpected RNA genes have been identified in the past decade, and more unusual sorts of genes may be found in the future. Moreover, almost 99% of the human genome is composed of "spaces" within and between protein-coding genes whose purpose is not fully understood. Included in the non-protein-coding portion of the genome are regulatory sequences that control how genes express their protein products at different times and places. Almost half of the human genome is occupied by so-called "jumping genes" or their remnants, some of which move about using a mechanism that is shared with human immunodeficiency virus (HIV) and other retroviruses. During the evolution of higher living things, genomes have been extensively remodeled by the duplication of individual chromosomes and the exchange of pieces between chromosomes.

Genomes are thus considerably more complicated than originally envisioned by Winkler. We are coming to understand that the genome is both a dynamic structure that changes through evolutionary time and a dynamic concept that changes with our increasing knowledge. For most higher organisms, including humans, Morgan's analogy of a genome should now be envisioned as strings of different sorts spliced together into a very long strand, with many bits of beads scattered here and there.

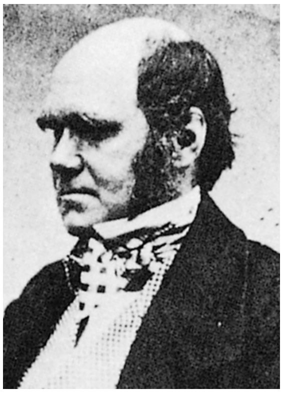
This book is designed to provide the conceptual and experimental background needed to participate in the new science of genomes. The geneticist J.B.S. Haldane once famously said "The universe is not only queerer than we suppose but queerer than we *can* suppose." With its black holes, curved space, and unaccounted-for dark matter, Haldane's prediction for the universe has certainly come true. The exploration of the human and other genomes is just beginning. They, too, are turning out to be queerer and more exciting than previously imagined and promise to reveal many more surprises in the future.

## ESTABLISHING THE PHYSICAL BASIS OF HEREDITY

Living things preserve their own lineages through reproduction—by creating offspring that carry on their inheritance through successive generations. Humans intuitively understand that they pass on some of their physical traits to their children. Thus, men and women endeavor to select vigorous, healthy mates who can give birth to vigorous, healthy children. Over time, the production of vigorous offspring contributed to the development of humans with traits that adapted them to live in a variety of environments. The extension of this concept to other organisms led to the domestication of plants and animals.

During the last 150 years, scientists have sought an increasingly explicit explanation of the hereditary process that allows traits to be passed from one generation to another. Let us start by taking a brief look back to the history of the quest to understand the physical basis of heredity, which is the foundation of genome science.

In his 1859 book *On the Origin of Species*, the Englishman Charles Darwin described how heredity operates in populations of organisms, enabling them to adapt to different environmental conditions. In the process of evolution by natural selection, members of the same and different species compete for limited resources needed for survival. The fittest members of a population are more likely to reproduce. On rare occasions, a random physical change in an individual increases its ability to adapt to environmental conditions or exploit new food resources. This "adaptive" change increases the individual's chance to survive and reproduce. Adaptive changes are more likely passed on to offspring, who, in turn, are fitter than their peers; they also have a greater chance of surviving to pass on their physical characteristics to succeeding generations. In this way, beneficial traits accumulate within a population of organisms. Through the process of adaptive radiation, populations expand into new environments and evolve to exploit specialized food resources, thus limiting competition and increasing their chances for survival.



Charles Darwin (ca. 1859)
(Courtesy of the American Museum of Natural History Library.)

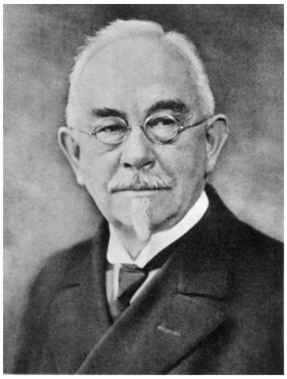
Gregor Mendel (ca. 1860)
(Courtesy of the Austrian Press and Information Service.)

Although Darwin proposed an incorrect mechanism of heredity, termed "pangenesis," he did not know the physical source of individual variation upon which his evolutionary processes acted or how it was passed on to successive generations. In his paper "Experiments in Plant Hybridization," published in 1865, the Moravian monk Gregor Mendel described the hereditary process at the level of the individual organism and provided a mechanism to drive evolution. From the results of controlled crosses of garden peas, he showed that traits are inherited in a predictable manner as "factors," which we now call genes. Mendel related each plant trait to a pair of genes, one of which is inherited from each parent. Although common sense suggests that offspring are a mixture of parental traits, Mendel showed that the parental genes governing each trait do not blend. Instead, each parental gene is maintained as a discrete bit of hereditary information, unchanged through generations.

Some pea traits examined by Mendel. Album Bernay (1876–1893) shows some of the pea traits that Mendel examined.

(The John Innes Archives, courtesy of the John Innes Foundation.)

Mendel's notion that a trait is determined by a pair of genes presented a potential problem. If parents pass on both copies of a gene pair, then their offspring would end up with four genes for each trait. This doubling of genetic material would continue in ensuing generations. Mendel deduced that parents contribute only half of their gene set to their offspring. He hypothesized that the gene number is reduced during gametogenesis, so that each gamete (sex cell) receives one copy of each gene pair. During fertilization, the male and female gametes then fuse to restore each pair of genes in the offspring.

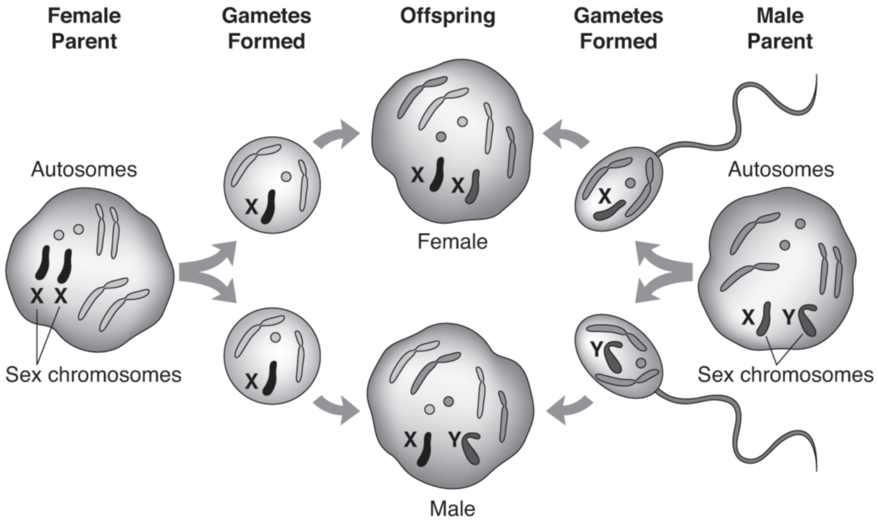


Wilhelm Johannsen

(Courtesy of Hunt Institute for Botanical Documentation, Carnegie Mellon University, Pittsburgh, Pennsylvania.)

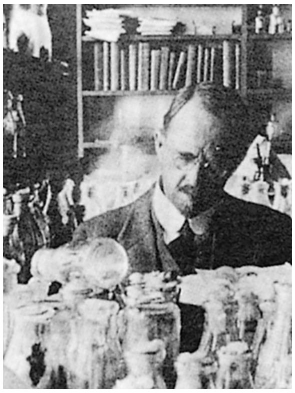
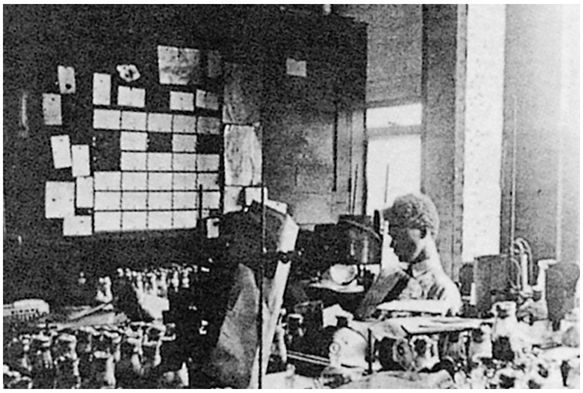
Mendel's work went essentially unnoticed for 35 years. Then, in 1900, the Dutchman Hugo de Vries, the German Carl Correns, and the Austrian Erich von Tschermak-Seysenegg rediscovered Mendel's paper and published research data that confirmed his earlier work. de Vries realized that Mendel's "factors" were the same entities that he called "pangens," which he had derived from Darwin's "pangenesis." In 1909, Wilhelm Johannsen shortened the term to "gene" and also coined the words "genotype" and "phenotype" to refer to an organism's genetic composition (genes) and its observable characteristics (traits).

In 1902, Theodor Boveri, at the University of Würzburg, and Walter Sutton, a student at Columbia University, were the first to directly relate heredity to chromosome behavior. Boveri found that a sea urchin egg fertilized by two sperm produces daughter cells that divide asymmetrically and have incomplete sets of chromosomes. Sutton found that the genetic material of the grasshopper *Brachystola* consists of 11 pairs of chromosomes and that gametes formed during meiosis receive only one chromosome from each pair. Then, independent work in 1905 by Nettie Stevens and Edmund Wilson (Sutton's mentor at Columbia) showed that sex is determined by separate X and Y chromosomes, with females having two X chromosomes (XX) and males having a single X and Y chromosome (XY). During meiosis, each egg receives a single copy of an X chromosome, whereas each sperm receives either an X or a Y chromosome. These behaviors exactly paralleled the segregation of Mendel's hereditary factors into parental gametes and suggested that genes are physically located on the chromosomes.



Segregation of X and Y chromosomes in *Drosophila*.

Conclusive evidence that genes are located on chromosomes became available during the second decade of the 20th century. During this period, Thomas Hunt Morgan and his bright cadre of students at Columbia University—Alfred Sturtevant, Calvin Bridges, and Hermann Muller—established the physical basis of heredity. Working with the common fruit fly *Drosophila melanogaster* in 1910, Morgan's group identified a mutation that produces white-colored eyes (as opposed to the normal red color). First, their Mendelian analyses showed that white eyes were confined to males in most crosses, suggesting that white eye color is a sex-linked recessive trait. This meant that the gene for eye color was located on the X chromosome. Next, they identified more than 80 additional mutants and showed that sets of genes are "linked" or inherited together as if they are a single physical unit. All genes sorted into four linkage groups, which corresponded to the number of *Drosophila* chromosomes seen under a microscope.

Thomas Hunt Morgan (ca. 1917), *left*, Courtesy of the American Society of Zoologists; Calvin Bridges in the "fly room" at Columbia University (ca. 1926), *middle*, courtesy of the American Society of Zoologists; and Alfred Sturtevant, *right*, courtesy of the American Philosophical Society.

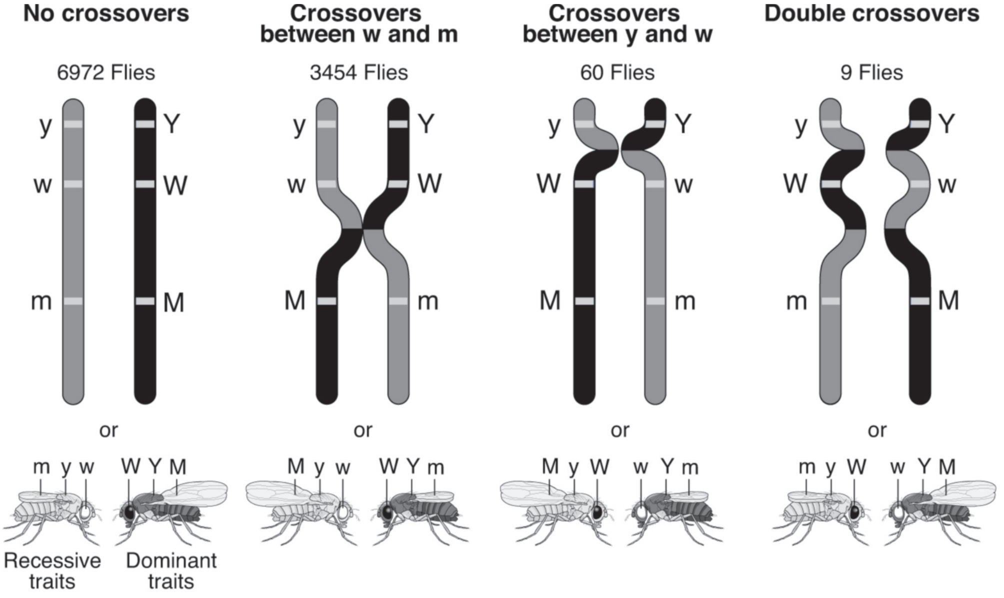
Frans Alfons Janssens

(Courtesy of the Centre of Microbial and Plant Genetics, K.U. Leuven.)

Working at the Catholic University of Leuven in 1909, Frans Alfons Janssens found that, early in meiosis, homologous chromosomes intertwine and exchange pieces—a process that became known as "crossing-over." Morgan realized that crossing-over could provide a measure of the relative distance between two genes. He reasoned that closely linked genes will rarely be separated by crossing-over, but genes that are far apart will be frequently separated. Therefore, the lower the crossover frequency between two genes, the closer together they should be on the chromosome. Alfred Sturtevant provided support for this concept in his 1913 doctoral thesis, when he made a map of the relative locations of three genes on the *Drosophila* X chromosome.

It was not until 1931, however, that Barbara McClintock and Harriet Creighton, at Cornell University, obtained direct cytological proof of genetic crossing-over. Working in maize, they related the phenotypes caused by gene crossovers to the coinheritance of a visible chromosome "knob." In the same year, Curt Stern, at the University of Berlin, used a similar approach to study the X chromosome of *Drosophila*. Taken together, these experiments conclusively proved that genes reside on chromosomes and are arrayed at specific points along their length.

A clear understanding of the physical basis of heredity came with the discovery that DNA is the genetic material. The first clue came from experiments conducted in

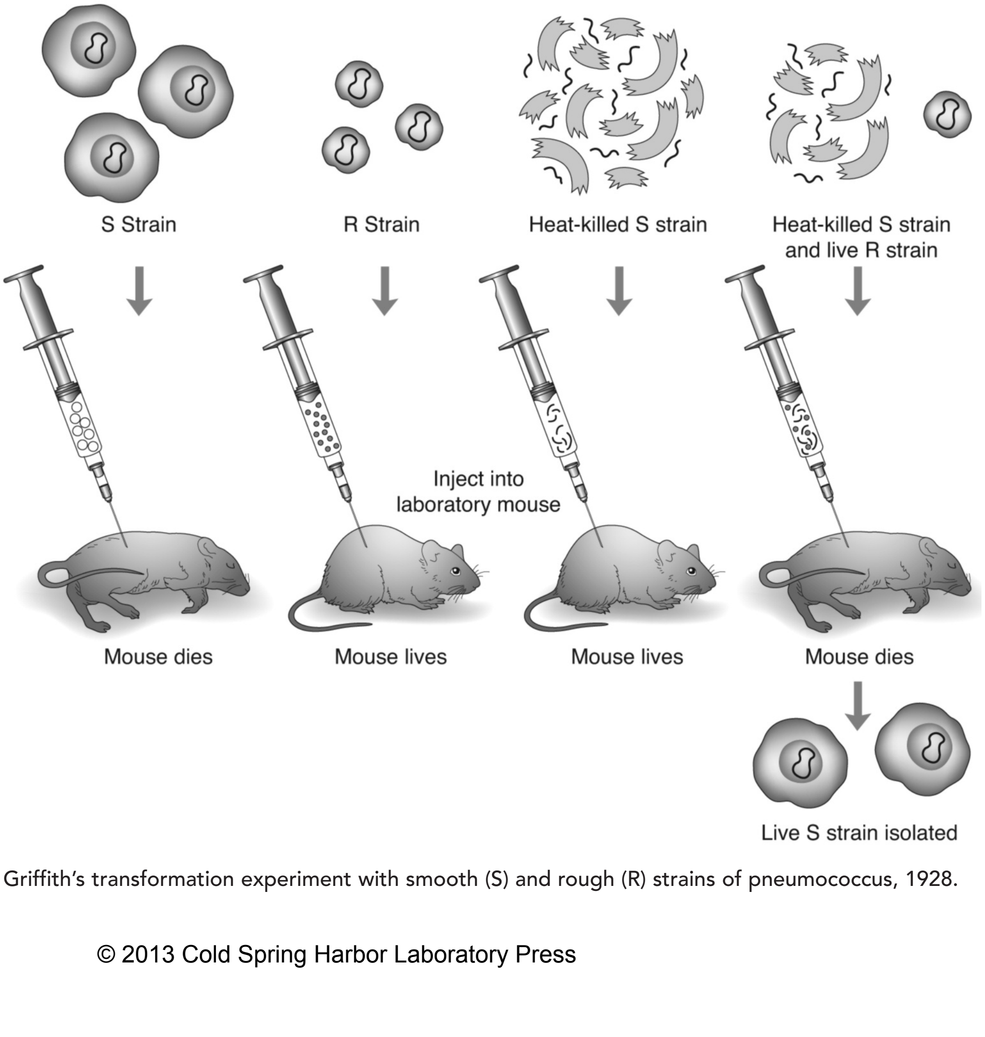


Sturtevant's linkage experiment in *Drosophila*, 1913. Sturtevant examined the X-linked inheritance of three recessive traits: yellow body (*y*), white eyes (*w*), and miniature wings (*m*). He crossed recessive males (y,w,m) with heterozygous females having recessive genes on one X chromosome (*y,w,m*) and dominant genes on the other (*Y,W,M*). Because a male parent can only contribute recessive genes on its single X chromosome, the phenotypes of both male and female offspring are due entirely to the inheritance of the maternal X chromosomes. Mendelian analysis predicts that all of the 10,495 offspring in Sturtevant's experiment would show either a purely dominant phenotype, normal body/eye color/wings, or a purely recessive phenotype, yellow body/*white* eyes/miniature wings. However, offspring inherited various mixtures of dominant and recessive traits. Sturtevant deduced that the mixed phenotypes were caused by genetic exchange between a female's two X chromosomes during gamete formation. The frequency of exchange is a measure of the distance between two genes located on the same chromosome.

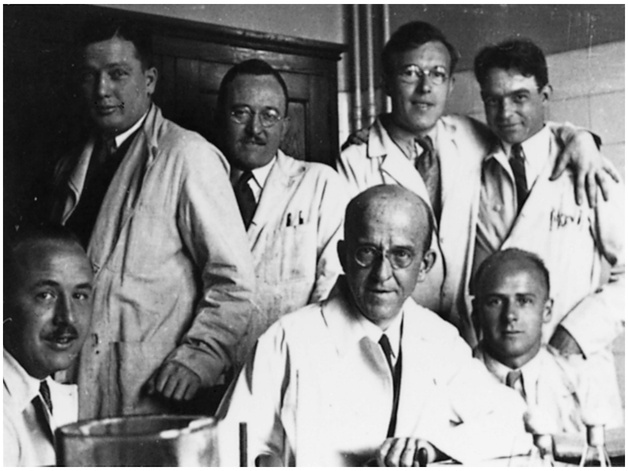
Fred Griffith
(From www.wikipedia.org.)

1928 by the English microbiologist Fred Griffith with two strains of pneumococcus bacteria. A virulent smooth (S) strain possesses a smooth polysaccharide capsule that is essential for a pneumonia infection, whereas a nonvirulent, rough (R) strain lacks this outer capsule. Following injection with the S strain, mice succumb in several days to pneumonia. Although neither living R strain nor heat-killed S strain caused illness when injected alone, Griffith found that coinjecting the two produced a lethal infection. Furthermore, he retrieved virulent S strains from mice infected with this mixture of bacteria. He concluded that some principle from the dead S bacteria had "transformed" the innocuous R strain, allowing it to produce the polysaccharide capsule required for virulence.

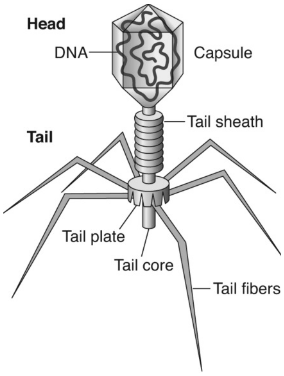
Although Griffith's experiment hinted at an involvement of metabolism, genes were still known only by their outward manifestation as visible traits. However, in 1941, George Beadle and Edward Tatum at Stanford University finally showed that the job of a gene is to produce a specific enzyme (protein). Their experiment used the simple red bread mold *Neurospora*, which is able to synthesize amino acids and vitamins from simple components (sucrose, salts, and biotin). After exposing *Neurospora* to X rays, they identified strains that grew only when supplemented with a specific amino acid or vitamin. They concluded that, for each deficient strain, irradiation had mutated a single gene that produces an enzyme needed to synthesize one amino acid or vitamin.



Griffith's transformation experiment with smooth (S) and rough (R) strains of pneumococcus, 1928.



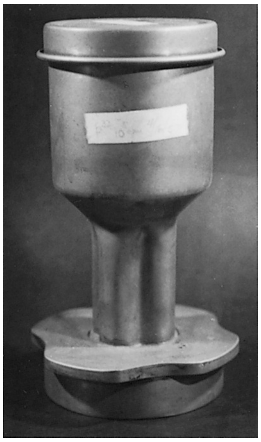
Oswald Avery (center foreground) and associates, 1932. (Seated, *left to right*) Thomas Francis Jr., Avery, and Walther F. Goeble; (standing) Edward E. Terrell, Kenneth Goodner, Rene J. Dubos, and Frank H. Babers. (Courtesy of the Rockefeller Archive Center.)



Bacteriophage. The phage particle is essentially a protein capsule surrounding a core of DNA.

In the meantime, Oswald Avery, Collin MacLeod, and Maclyn McCarty followed up on Griffith's transformation experiments at the Rockefeller Institute. They purified the "transforming principle" from killed S bacteria that had readily induced R bacteria to synthesize the outer capsule. Transforming activity was unaffected by treatment with trypsin and chymotrypsin (which digest protein) and ribonuclease (RNase, which digests RNA). However, deoxyribonuclease (DNase, which digests DNA) destroyed all transforming activity, and analysis of molecular composition and weight indicated that the active fraction was primarily DNA. In 1944, they concluded that "The inducing [transforming] substance has been likened to a gene, and the capsular antigen which is produced in response to it has been regarded as a gene product." Thus, the Rockefeller group provided conclusive evidence that a gene is made of DNA.

Lingering dogma that protein was the genetic material prevented most scientists from focusing on DNA until the so-called "blender" experiment was conducted in 1952 by Alfred Hershey and Martha Chase at the Carnegie Department of Genetics at Cold Spring Harbor. They used a bacterial virus, or bacteriophage (phage), which is simply composed of an outer capsule of protein and an inner core of DNA. They attached different radioactive labels to the phage protein and DNA, allowed the phage time to infect bacteria, and then agitated the culture in a Waring blender to detach the phage particles from the bacteria. After centrifuging to separate the detached phages



Waring blender used in the Hershey-Chase experiment.

Martha Chase and Alfred Hershey, 1953. (Courtesy of Cold Spring Harbor Laboratory Archives.)

from the bacterial cells, they found that the radioactive DNA that remained with the bacterial fraction was sufficient to produce a new generation of phages. This work further strengthened the concept that DNA is the hereditary material that comprises genes.

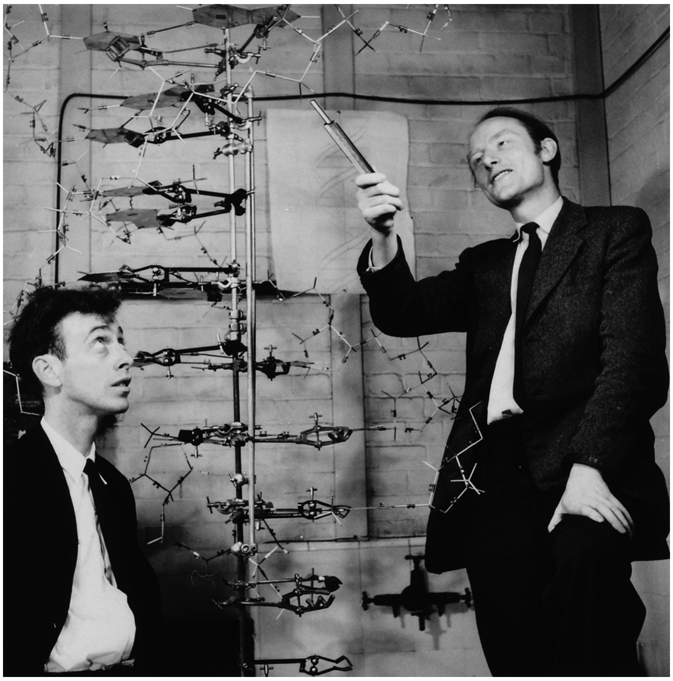
## DNA AS INFORMATION

Friedrich Miescher

(From www.wikipedia.com.)

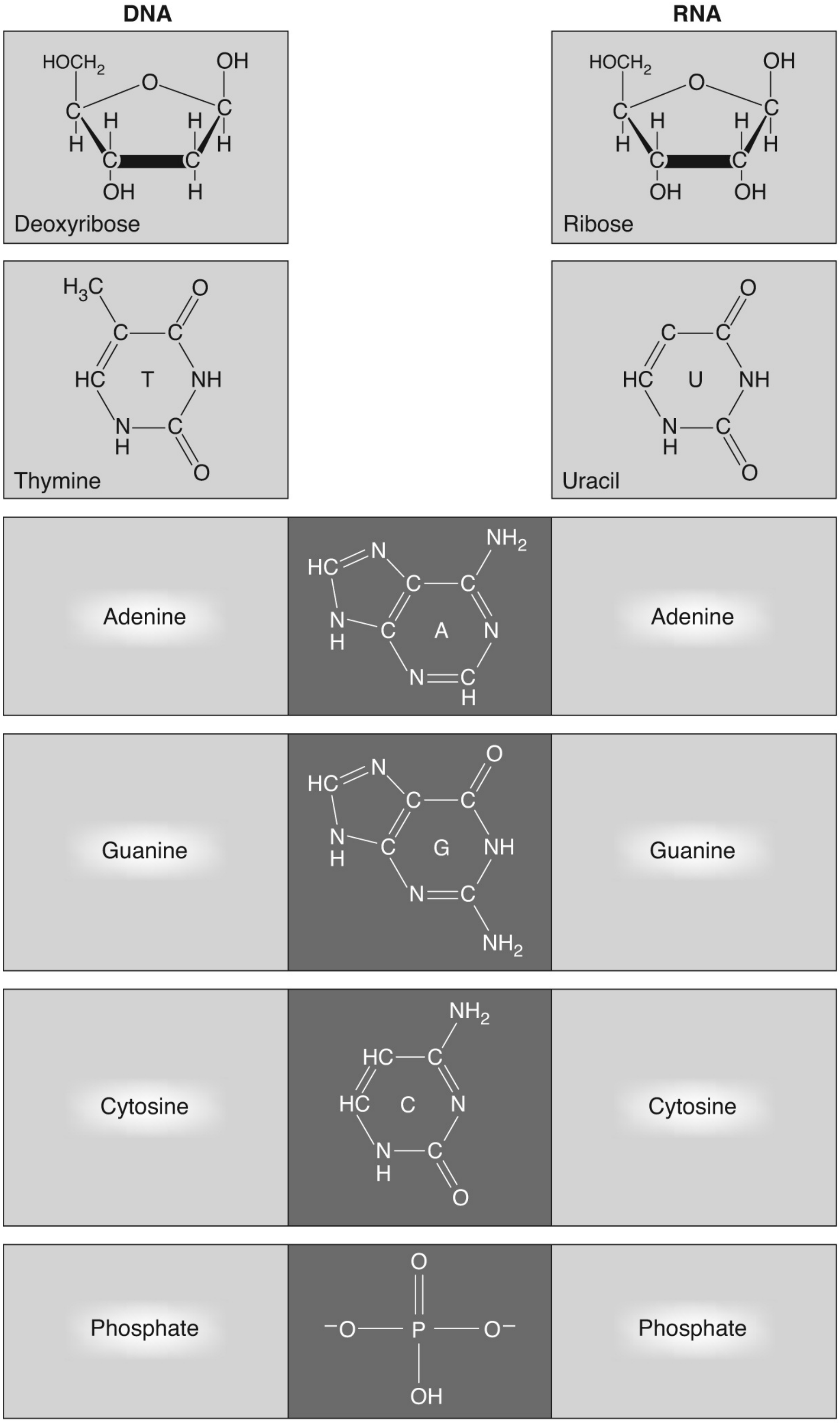
Ironically, DNA was discovered in 1869, only 10 years after the publication of Darwin's *On the Origin of Species* and 4 years after Mendel's "Experiments in Plant Hybridization." A Swiss doctor, Friedrich Miescher, isolated a substance he called "nuclein" from the large nuclei of white blood cells. His source of cells was pus from soiled surgical bandages. Building upon Miescher's observation that the substance was rich in phosphorus and nitrogen, by 1900, it had been determined that nuclein was a long molecule composed of three distinct chemical subunits: an acidic phosphate, five types of nitrogen-rich bases (adenine, thymine, guanine, cytosine, and uracil), and a five-carbon sugar. By the 1920s, two forms of nucleic acids were differentiated by virtue of their sugar composition: ribonucleic acid (RNA), based on ribose sugar, and deoxyribonucleic acid (DNA), based on deoxyribose sugar. These forms were also found to differ slightly in base composition; thymine is found exclusively in DNA, whereas uracil is found only in RNA.

The structure of the DNA molecule was solved in 1953 by James Watson and Francis Crick, working at the Cavendish Laboratory in Cambridge, England. They constructed a metal model that showed DNA to resemble a twisting ladder—with the



James Watson and Francis Crick with their DNA model in Cambridge, England, 1953.

(From A. Barrington Brown, Photo Researchers, Inc.)



Components of DNA and RNA molecules.

(Art concept developed by Lisa Shoemaker.)

rails formed of alternating units of deoxyribose sugar and phosphate and the rungs formed of nitrogenous bases. Each rung is composed of a two nitrogenous bases, a base pair, where adenine (A) always pairs with thymine (T) and guanine (G) always pairs with cytosine (C).

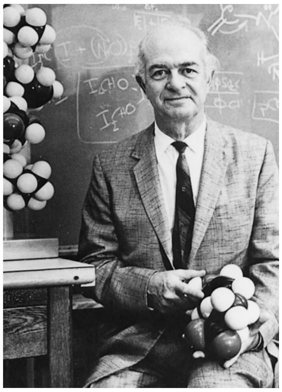
The Watson-Crick model was based on critical information that had accumulated quickly since 1950. The base-pair rule came from work by Erwin Chargaff of Columbia University, who found a consistent one-to-one ratio of adenine to thymine and guanine to cytosine in DNA samples from a variety of organisms. Linus Pauling, Robert Corey, and Herman Branson at California Institute of Technology provided the atomic dimensions of the α-helix configuration of protein, in which amino acids form a helical structure. Finally, the sharp X-ray diffraction photographs of DNA taken by Maurice Wilkins and Rosalind Franklin at Kings College, London, resembled the patterns of the protein helix—strongly suggesting that DNA is also an α-helix.

As the only biologist of this group, Watson had the greatest insight into how the DNA molecule must function to provide the physical basis of heredity. He understood that life ultimately depends on the perpetuation and amplification of a DNA sequence through time. A successful organism must survive and pass on its genome to succeeding generations, and the bearers of this successful genome will increase in number over time. On the one hand, the DNA molecule must be sufficiently stable so that a sequence is inherited with enough fidelity to maintain the identity of each species. On the other hand, the DNA molecule must be sufficiently plastic—mutable—to allow species to evolve and change over time.

Watson later came to the point with this simple definition: "DNA is information." A DNA molecule is capable of encoding information in its nucleotide sequence—the order in which the nucleotides A, T, C, and G follow one another along one strand of the molecule. The balance of this chapter will explore how information is encoded in a DNA sequence and how the DNA sequence of a genome is analyzed. Bioinformatics is the science of understanding the information encoded in DNA and other biological molecules, and genomics is the science of understanding the structure and function of the set of DNA molecules that distinguish each species.

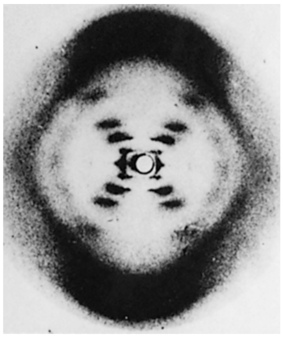
Erwin Chargaff, 1947

(Courtesy of Cold Spring Harbor Laboratory Archives.)



Linus Pauling, ca. 1950

(Courtesy of the Archives, California Institute of Technology.)



Rosalind Franklin's X-ray diffraction photograph of DNA, 1953

(Reprinted, with permission, from Franklin RE, Gosling RG. 1953. *Nature* 171: 740–741, ©Macmillan; photo courtesy of Cold Spring Harbor Laboratory Archives.)

Rosalind Franklin, 1948

(Courtesy of Anne Sayre.)

Maurice Wilkins, ca. 1955

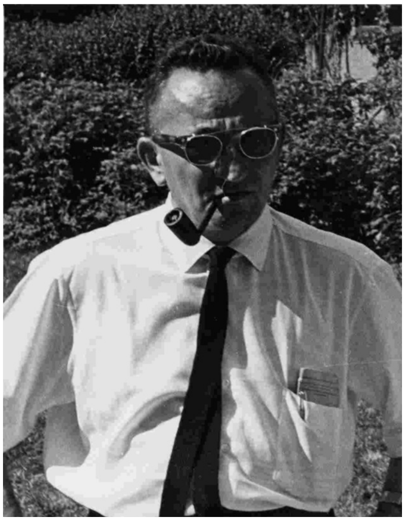
(Courtesy of Cold Spring Harbor Laboratory Archives.)

## THE GENETIC CODE

Work in the 1950s and 1960s showed how DNA encodes information and provided the mechanism for Beadle and Tatum's hypothesis that one gene makes one protein. First, Paul Zamecnik, at Massachusetts General Hospital, established that protein synthesis takes place on protein/RNA conglomerates located in the cytoplasm, which we now know as ribosomes. Then, Jerard Hurwitz, at New York University School of Medicine, and Samuel Weiss, at University of Chicago, independently identified RNA polymerase as the enzyme that synthesizes RNA by adding complementary nucleotides to a DNA template. Subsequently, three different RNA polymerases (I, II, and III) were identified in higher organisms. RNA polymerase I synthesizes ribosomal RNA (rRNA), RNA polymerase II synthesizes messenger RNA (mRNA), and RNA polymerase III synthesizes transfer RNA (tRNA) and one small rRNA (5S rRNA).

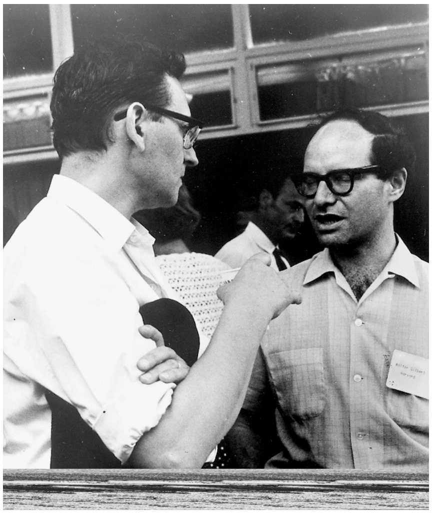
In addition, Benjamin Hall and Sol Spiegelman, at the University of Illinois, showed that complementary RNA and DNA sequences bind together to form a stable heteroduplex. Collaborators Sydney Brenner (MRC Laboratory), François Jacob (Institut Pasteur), and Matthew Meselson (Harvard University) and a team composed of James Watson, François Gros, and Walter Gilbert (Harvard University) independently showed that immediately after a bacteriophage infects a bacteria, RNA is synthesized and associates with ribosomes. Moreover, the newly synthesized RNA only lasts for several minutes inside the bacterial cells. Taken together, these experiments illuminated the first step of protein production—transcribing the DNA code (a gene) into a complementary RNA code (a messenger RNA).

The next step was to work out how the RNA code is translated into an amino acid code. Whereas RNA is made up of only four different nucleotides (A, C, G, and U), proteins are composed of 20 different amino acids. So it was immediately apparent that a combination of several nucleotides would be required to encode each amino acid. A two-letter code would only have 16 combinations—not enough to specify all 20 amino acids. However, a three-letter code provided more than enough combina-



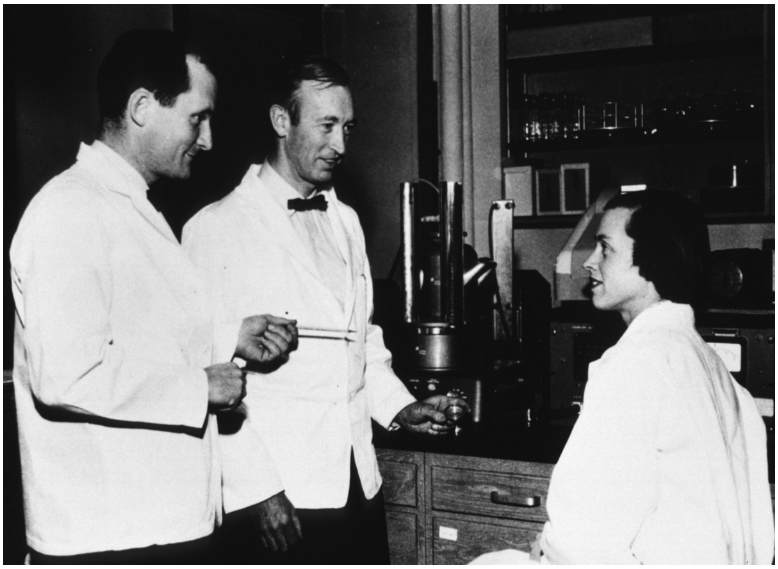
Sol Spiegelman, ca. 1963

(Courtesy of Cold Spring Harbor Laboratory Archives.)

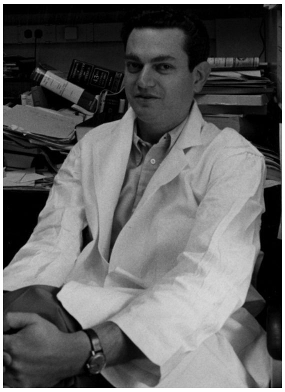


François Gros and Walter Gilbert, ca. 1970

(Courtesy of Cold Spring Harbor Laboratory Archives.)



Mahlon B. Hoagland, Paul C. Zamecnik, and Mary L. Stephenson, ca. 1956
(Courtesy of the National Library of Medicine.)



Marshall W. Nirenberg, ca. 1962
(Courtesy of the National Institutes of Health.)

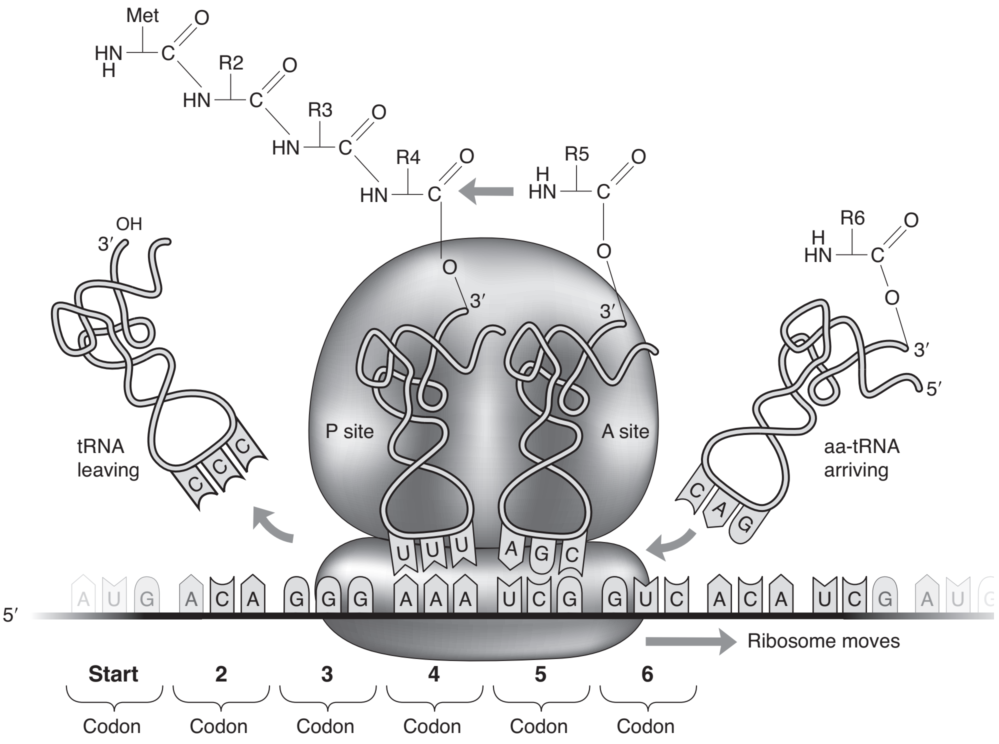
Har Gobind Khorana, ca. 1966
(Courtesy of Cold Spring Harbor Laboratory Archives.)

tions (64), and Francis Crick and Sydney Brenner referred to this nucleotide triplet as a codon. At the same time, Francis Crick realized that some sort of "adaptor" molecule was needed to link a codon to a corresponding amino acid.

Robert Holley, working at Cornell University, discovered that the adaptor molecule is a class of small RNA molecules, 70–80 nucleotides in length, that covalently bind amino acids; these are tRNAs. Then Paul Zamecnik and Mahlon Hoagland, at Massachusetts General Hospital, discovered a class of enzymes called aminoacyl tRNA synthetases that attach a specific amino acid to a specific tRNA. Each tRNA contains a loop structure with a unique three-nucleotide-long sequence—the anticodon—that binds to a complementary codon in mRNA. This aligns the specified amino acid at the ribosome for addition to a polypeptide chain.

By 1966, the laboratories of Marshall Nirenberg (the National Institutes of Health) and Har Gobind Khorana (University of Wisconsin) had broken the genetic code through which mRNA instructs tRNAs to add specific amino acids at the ribosome. Both researchers synthesized RNA molecules composed of repeating units of a single codon, added the synthetic mRNA to a cell-free extract containing all the required tRNAs bound to amino acids, and then monitored the composition of proteins that were synthesized. Initially, Nirenberg found that polyuracil (making codons UUU-UUU-UUU...) produced a protein made up solely of the amino acid phenylalanine. Eventually, all possible codon combinations were tried, yielding a complete genetic "dictionary" for the translation of mRNA into amino acids. Nearly all proteins begin with the amino acid methionine (Met); scientists quickly realized that its codon (AUG) represents the "start" signal for protein synthesis. Three codons for which there are no naturally occurring tRNAs—UAA, UAG, and UGA—are "stop" signals that terminate translation.

Interestingly, only two amino acids, methionine and tryptophan, are specified by a single codon; all other amino acids are specified by two or more different codons.

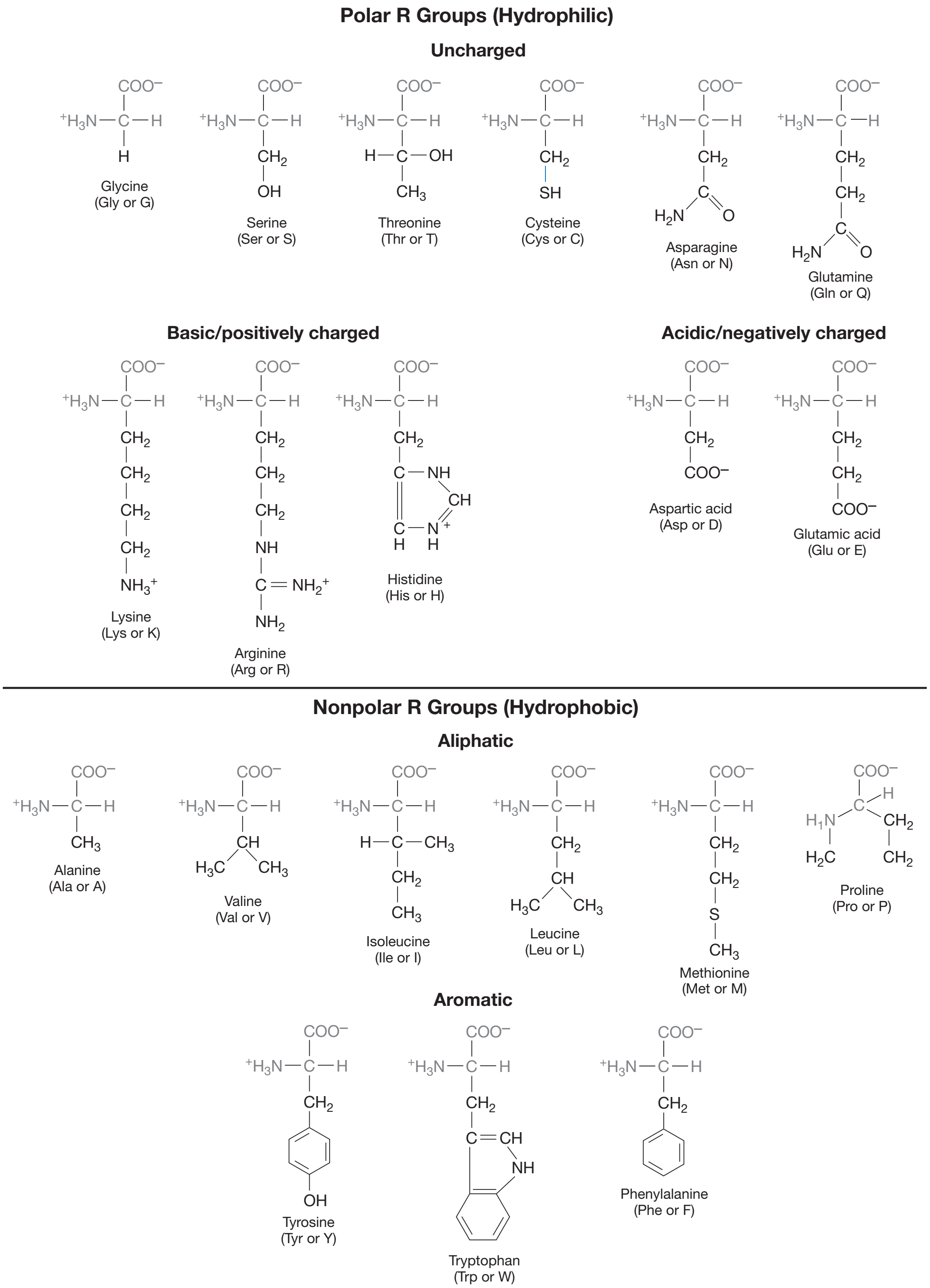


Translation of mRNA codons into amino acids.

| 1st position of codon (5′ terminus) | 2nd position of codon | | | | 3rd position of codon (3′ terminus) |
|---|---|---|---|---|---|
| | U | C | A | G | |
| U | UUU Phe | UCU Ser | UAU Ty | UGU Cys | U |
| | UUC Phe | UCC Ser | UAC Tyr | UGC Cys | C |
| | UUA Leu | UCA Ser | UAA Stop | UGA Stop | A |
| | UUG Leu | UCG Ser | UAG Stop | UGG Trp | G |
| C | CUU Leu | CCU Pro | CAU His | CGU Arg | U |
| | CUC Leu | CCC Pro | CAC His | CGC Arg | C |
| | CUA Leu | CCA Pro | CAA Gln | CGA Arg | A |
| | CUG Leu | CCG Pro | CAG Gln | CGG Arg | G |
| A | AUU Ile | ACU Thr | AAU Asn | AGU Ser | U |
| | AUC Ile | ACC Thr | AAC Asn | AGC Ser | C |
| | AUA Ile | ACA Thr | AAA Lys | AGA Arg | A |
| | AUG Met | ACG Thr | AAG Lys | AGG Arg | G |
| G | GUU Val | GCU Ala | GAU Asp | GGU Gly | U |
| | GUC Val | GCC Ala | GAC Asp | GGC Gly | C |
| | GUA Val | GCA Ala | GAA Glu | GGA Gly | A |
| | GUG Val | GCG Ala | GAG Glu | GGG Gly | G |

The genetic code.

Because of this redundancy—also referred to as degeneracy or wobble—single-nucleotide mutations in DNA are often of no functional consequence. Changing a single nucleotide in a degenerate codon to another triplet coding for the same amino acid has no effect on the amino acid sequence of a protein. For example, any codon beginning with GG specifies the amino acid glycine regardless of the nucleotide in the third position (GGU, GGC, GGA, or GGG).



Twenty naturally occurring amino acids grouped by properties. The side chains (gray) determine the characteristic properties of each amino acid.

## FINDING SIMPLE PATTERNS IN DNA SEQUENCE

Any sequence of characters (letters or numbers) may be randomly generated or encoded with meaningful information. A receiver might reject a message that is random or, if the sender uses a consistent set of rules to convey meaning, a receiver can decode a message. If we think of hereditary information stored in DNA as a language or code, we can use the English language as a model to introduce some principles of DNA sequence analysis.

It can sometimes be difficult to determine whether a sequence encodes information. Although it is not immediately evident, the meaning of the sequence of 1064 letters (below) becomes clear when we add in the conventions of word spacing and punctuation.

Understanding DNA sequence is not nearly as simple as looking at a language we have learned over a lifetime. English-language speakers may intuitively know that English has meaning, but we can use simple statistics to show, on another level, how a sequence of letters conveys meaning. Frequency analysis evaluates the occurrence of characters to determine if a sequence is random or potentially conveys meaning. First, let's compare

ThemainchallengeinbiologywastounderstandgenereplicationandthewayinwhichgenescontrolproteinsynthesisItwasobviousthattheseproblemscouldbelogicallyattackedonlywhenthestructureofthegenebecameknownThismeantsolvingthestructureofDNAThenthisobjectiveseemedoutofreachtotheinterestedgeneticistsButinourcolddarkCavendishlabwethoughtthejobcouldbedonequitepossiblywithinafewmonthsOuroptimismwaspartlybasedonLinusPaulingsfeatindeducingthealphahelixWealsoknewthatMauriceWilkinshadcrystallineXraydiffractionphotographsfromDNAandsoitmusthaveawelldefinedstructureTherewasthusananswerforsomebodytogetDuringthenexteighteenmonthsuntilthedoublehelicalstructurebecameelucidatedwefrequentlydiscussedthenecessitythatthecorrectstructurehavethecapacityforselfreplicationAndinpessimisticmoodsweoftenworriedthatthecorrectstructuremightbedullThatisitwouldsuggestabsolutelynothingandexciteusnomorethansomethinginertlikecollagenThefindingofthedoublehelixthusbroughtusnotonlyjoybutgreatreliefItwasunbelievablyinterestingandimmediatelyallowedustomakeaseriousproposalforthemechanismofgeneduplication

The main challenge in biology was to understand gene replication and the way in which genes control protein synthesis. It was obvious that these problems could be logically attacked only when the structure of the gene became known. This meant solving the structure of DNA. Then this objective seemed out of reach to the interested geneticists. But in our cold, dark Cavendish lab, we thought the job could be done, quite possibly within a few months. Our optimism was partly based on Linus Pauling's feat in deducing the alpha helix... We also knew that Maurice Wilkins had crystalline X-ray diffraction photographs from DNA and so it must have a well-defined structure. There was thus an answer for somebody to get. During the next eighteen months, until the double helical structure became elucidated, we frequently discussed the necessity that the correct structure have the capacity for self-replication. And in pessimistic moods, we often worried that the correct structure might be dull. That is, it would suggest absolutely nothing and excite us no more than something inert like collagen. The finding of the double helix thus brought us not only joy but great relief. It was unbelievably interesting and immediately allowed us to make a serious proposal for the mechanism of gene duplication.

El principal desafio en la biologia fue de comprender replica de gene y la manera en las que genes controlan sintesis de proteina. Fue obvio que estos problemas podrian ser atacados logicamente solo cuando la estructura del gene llego a ser conocida. Este destinado resolviendo la estructura de ADN. Entonces este objetivo parecio fuera de alcance a los genetistas interesados. Pero en nuestro frio, laboratorio oscuro de Cavendish, nosotros pensamos que el trabajo podria ser hecho, bastante posiblemente dentro de unos pocos meses. Nuestro optimismo fue basado en parte en la proeza de Linus Pauling a deducir la helice alfa... Nosotros tambien supimos que Maurice Wilkins tenia fotografias de cristal de difraccion de radiografia de ADN y tan debe tener una estructura bien definida. Habia asi una respuesta para alguien conseguir. Durante los proximos dieciocho meses, hasta que la doble estructura helicoidal llegara a ser aclarada, nosotros discutimos con frecuencia la necesidad que la estructura correcta tiene la capacidad para la auto-replica. Y en humors pesimistas, nosotros a menudo preocupamos que la estructura correcta quizas sea languida. Eso es, sugeriria absolutamente que nada y no nos emociona más que algo inerte quiere colageno. El hallazgo de la doble helice asi nos trajo no solo alegria pero gran alivio. Fue in-creiblemente interesante e inmediatamente nos permitio hacer una propuesta grave para el mecanismo de duplicacion de gene.

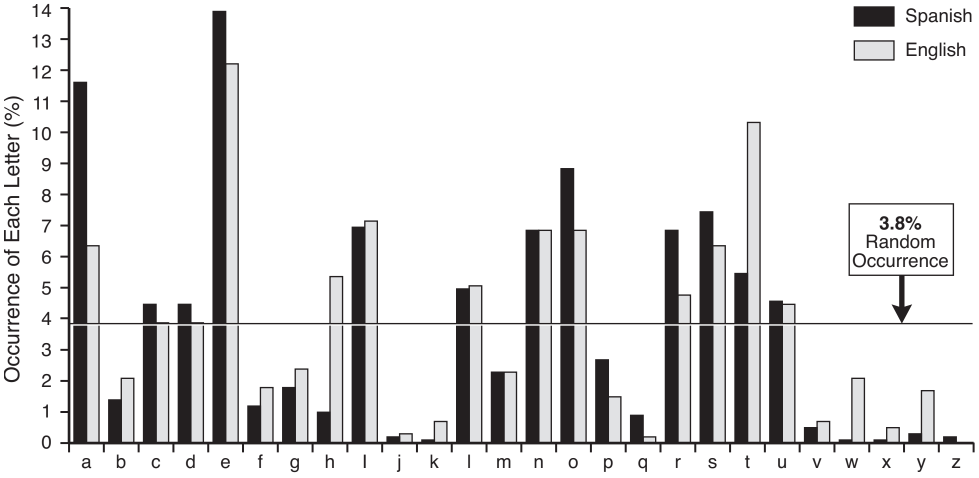
Excerpt from James D. Watson's Nobel lecture, December 11, 1962.

the frequency of each letter of the alphabet in the excerpt from Watson's Nobel lecture (see p. 15). If each of the 26 letters of the alphabet occurs equally, we would expect each letter to occur 41 times (1/26 × 1064) or to comprise 3.8% of the text. The graph below clearly shows that the rules used to encode meaning (words) in the English or Spanish language create a bias in the use of characters, such that some letters occur more frequently than expected by chance and others less frequently.

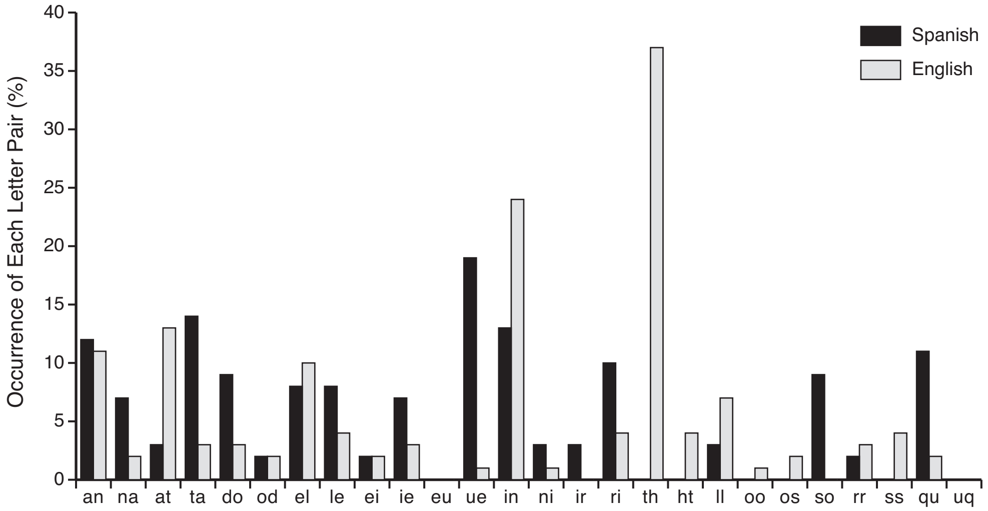
We can potentially uncover additional meaning if we analyze several two-character sequences and their inverses. If each letter pair is equally probable, we would expect 1.36 examples of each combination in the text (1/26 × 1/26 × 1064). However, any fan of crossword puzzles or *Wheel of Fortune* can say that certain letter pairs are much more frequent in the English language. Analysis of individual letters and letter pairs of a Spanish translation of the same text illustrates that different languages encode the same meaning with a different bias in letter use (see the graph at the top of p. 17).

Now, let's turn to DNA sequences composed of the nucleotides A, T, C, and G. Like a language, the genome evolved to convey meaning in DNA sequence—the directions to encode and regulate genes. Because this DNA language is not intuitive to us, merely looking at the three 1064-nucleotide sequences (see p. 18) will not distinguish which was taken from a mammalian protein-coding gene, which was from a mammalian noncoding (intergenic) region, and which was randomly generated by a computer. However, analysis of dinucleotide frequencies clearly shows that the two mammalian genomic sequences have nonrandom distributions—providing evidence of evolutionary selection (see the graph at the bottom of p. 17).

Closer inspection shows a general trend that represents the lowest level of genome organization. The relative abundance of the CG dinucleotide in the genic regions of mammalian genes—and its virtual absence in intergenic regions—can help to focus research efforts on gene-rich regions of the genome. The CG dinucleotide is properly termed CpG to indicate that it is linked by a phosphate on the same strand of the DNA molecule—as opposed to a C ≡ G base pair. The biological explanation for the relationship between "CpG islands" and gene enrichment is that the CpG dinucleotide is

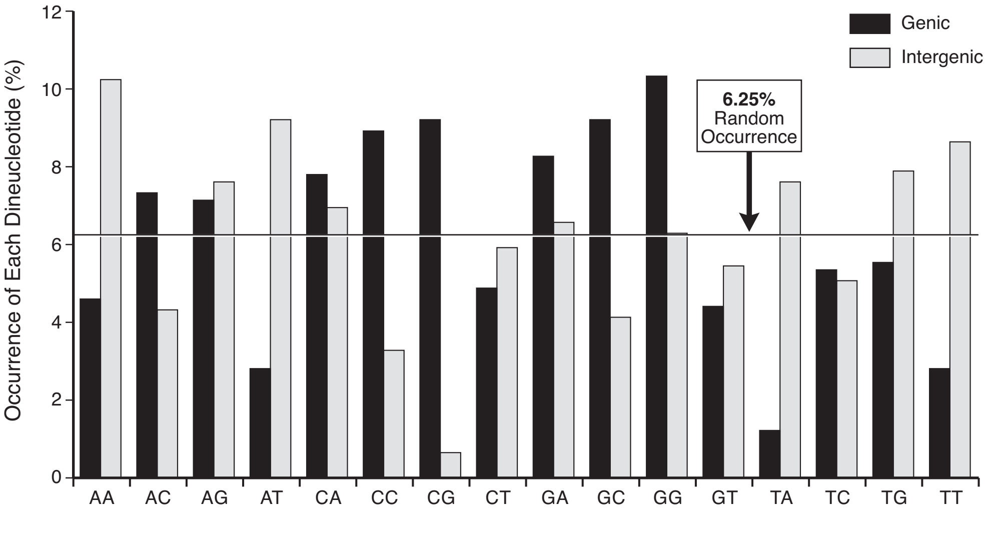


Frequency of letters in excerpt from Watson's Nobel lecture (see p. 15) and Spanish translation.



Frequency of some two-letter pairs in excerpt from Watson's Nobel lecture (see p. 15) and Spanish translation.

common in the 5′ promoter sequence that defines the beginning of many plant and animal genes. Promoters are recognized by RNA polymerase and other proteins that bind to the DNA to transcribe a gene sequence into mRNA. Thinking back to the evolution of cells, the cell membrane provided a means to sequester molecules from the environment and increase the concentration of reactants needed to efficiently assemble DNA molecules of sufficient length to encode proteins. At the same time, simple DNA sequences (such as the CpG dinucleotide) may have evolved to aggregate proteins needed to transcribe DNA into mRNA. Thus, the CpG dinucleotide may have been part of the early selection of gene regulators, perhaps analogous to the cosmic background radiation as a remnant of the cosmic dust clouds from which galaxies evolved after the Big Bang.



Dinucleotide frequencies in 1064-nucleotide sequences.

| Random | Genic | Intergenic |
|---|---|---|
| TTGAATGCACTCAGTCGTCGCGGCACC | CCTTTCTGCACTGGGTGGCAACCAGGTC | GAACTGACATGAAAACATGCCTCAGATA |
| AGTTTCCTCGGACGTTCAAGGAGTTTCC | TTTAGATTAGCCAACTAGAGAAGAGAAG | TATTGTTGAGGGGAAAAGCAAGTTATAA |
| TCAAGAGCTAGAAGTGTTATATCTCCAA | TAGAATAGCCAATTAGAGAAGTGACATC | ACTAGCATATGCTTTTGATTTTATATAT |
| GATAAGTAGCACCGATACACTCGTAAGA | ATGTTGACTCTAACTCGCATCCGCACTG | GTATAAAAACATGTGTATACACATATAT |
| GGGACGGCCAGCGAGCCGTGAACATAAG | TGTCCTATGAAGTCAGGAGTACATTTCT | CCTATATTTAAATGAAAAGACAAATTCC |
| TTAACAACTTGTGTCAGATTCTAGTAGG | GTTCATTTCAGTCCTGGAGTTTGCAGTG | GCAATGGTAATTCATGAAACTGATAACA |
| ATACTGATCACTCATTAGTGCCATCTAT | GGGTTTCTGACCAATGCCTTCGTTTTCT | GTCTTTCCCTCTGGGAAACAGCCTAGGC |
| GTTAATCTTGCGCGCATAGCGGTGAGCG | TGGTGAATTTTTGGGATGTAGTGAAGAG | CTAGGGACATTGATGATCAATGAGAATA |
| TGTGGAGGGGACGATCGTGTGCACAAAC | GCAGGCACTGAGCAACAGTGATTGTGTG | TTTCTCTATAGGAGATGAATCCTCTTAC |
| TAAAGAGTGCAGCACTAAATATCCCGTC | CTGCTGTGTCTCAGCATCAGCCGGCTTT | TGCAATAATATATTCATGTTCACAGTTG |
| ACAGTGAACGATCCAGGACTTTGGACTA | TCCTGCATGGACTGCTGTTCCTGAGTGC | CAAATTGTGGTCTCCTTATCATTAAAGT |
| TCTAGGAGCGTTTCGGCTCAGAGCGTGC | TATCCAGCTTACCCACTTCCAGAAGTTG | CTTTCATTCCCTGGAAGAATCAGAAAGC |
| AGAGCGCAAAAGGTTTGAACTTAACTTA | AGTGAACCACTGAACCACAGCTACCAAG | TTGAGTTTATCTTTCAGTAGTTACAGTC |
| TGGGTGTCAGAACGCTTGTGGGATATAT | CCATCATCATGCTATGGATGATTGCAAA | TGTGCTAATGGGGGAATATTTTTTATTC |
| CTCCCACCAGCAGTTGGATCCAATTCGG | CCAAGCCAACCTCTGGCTTGCTGCCTGC | ACTCAAGTATACCAAGATTGACAAAGCG |
| CACCGGCGACTCTGCTGTCTCACCTTCT | CTCAGCCTGCTTTACTGCTCCAAGCTCA | CATCTAAGGTATCAGGTACGCTAGTAGG |
| AGCTCTGTGCTCCTCTCAGCCCCCACTG | TCCGTTTCTCTCACACCTTCCTGATCTG | TACCAAGAGAAGTAAATGAAAAGCTCTC |
| TCACGACAGCCGTGAAAGGTTTAAGAAG | CTTGGCAAGCTGGGTCTCCAGGAAGATC | TTTATTTGGAAGAGCTCACCATCTTGGG |
| TCAATAGTCGCGTCCCTGTGGTGGTTAC | TCCCAGATGCTCCTGGGTATTATTCTTT | TGGTGGGAGGTAAGACATTTACACAATT |
| CATCTCTTATCGCCCTACGTAGAGCCTA | GCTCCTGCATCTGCACTGTCCTCTGTGT | AAATAGTTCAATCTATGCAACAAATGCT |
| CTGTACTGTTCTAACTAGCGTAAGAGCG | TTGGTGCTTTTTTAGCAGACCTCACTTC | ATTATTTCTAGTTTTTCATCCAACAAAT |
| GACGGTTTGGCCTACGTGGATGCCTGAG | ACAGTCACAACTGTGCTATTCATGAATA | ATTTCCTGAGCACCTGCAGGGCCCAGGC |
| TATACGCCGCCGTGTTCACTAGTACTGT | ACAATACAAGGCTCAACTGGCAGATTAA | TTTGAGTCATGCACTAAGGATGTGCATG |
| AAATAACGGGCAGAGGGATGTCAAATCC | AGATCTCAATTTATTTTATTCCTTTCTC | GTTAAATACTTTTCTGCCCTTGAGAAAC |
| TACTGTTTCCACCTCGTAGCGGCTGCTA | TTCTGCTATCTGTGGTCTGTGCCTCCTT | TCACCTATGTTGCTTGTCTGGTGCATGG |
| ATAGGTGGAATCGATCTCCGAGAAGTCA | TCCTATTGTTTCTGGTTTCTTCTGGGAT | CCCAGGGCAAAAACCATATCTTACTTAC |
| ACATTAGCTTGATTAGCTCAGGCGACGG | GCTGACTGTCTCCCTGGGAAGGCACATG | CTCTTTACCCACTGGAGCATCCAGTACC |
| GACCGTTAAGCCGTGATCTTAGTACAAA | AGGACAATGAAGGTCTATACCAGAAACT | ATGCTTTGTGCATATCAATGGCAGAAGG |
| GTCTTCGCCCCTGAACAGGCGTACTTGT | CTCGTGACCCCAGCCTGGAGGCCCACAT | TGCACTGCCAGGGTGGGGGTGAATGGAG |
| GGGCCTGAGACAGATATGGTGTACTCAC | TAAAGCCCTCAAGTCTCTTGTCTCCTTT | GAGGTGAATGAGAGGGAAGAGACACGAA |
| GATGGTAGAGTTCAGGGTGACCGATGTA | TTCTGCTTCTTTGTGATATCATCCTGTG | GGCGTATAGAATTTCTAACTCGAGTGGC |
| CAGCCGCCATCAATATTGATCAGGGTGA | CTGCCTTCATCTCTGTGCCCCTACTGAT | TACAGGAAAGTTCAACTTTGTTCATTTT |
| GAATCGGTTTTACTTATTTTGACTAGGA | TCTGTGGCGCGACAAAATAGGGGTGATG | TAATGTGACGCATGTGCCTGGTAAACAA |
| TTGCATATTCTGTGCCCGAGGGGTTCGT | GTTTGTGTTGGGATAATGGCAGCTTGTC | GTAGTTAGAAAAATAAGTTTGGGTGTCA |
| GTAGGGGAAGTCCGTCAAATTGGGTAGT | CCTCTGGGCATGCAGCCATCCTGATCTC | TCTTAGGGCAAGAATTTAGAAATAAAGA |
| GTGTTTTCAATTTATGCGATGCTCGGAT | AGGCAATGCCAAGTTGAGGAGAGCTGTG | CTGGGGAGTCCTAAGCATGGTGTTGCAG |
| CGGACGTGCTGTCTCTAAGGCACAAGGA | ATGACCATTCTGCTCTGGGCTCAGAGCA | CCATGGAAGTGAATGTGAATCCTAAGGA |
| TACTTAACGCTTCACTGAGTTGTCTTGA | GCCTGAAGGTAAGAGCCGACCACAAGGC | GTGAGGGAGAGAAGGGCTATTGATGAAG |

Three 1064-nucleotide sequences.

## DNA DIRECTIONALITY AND READING FRAMES

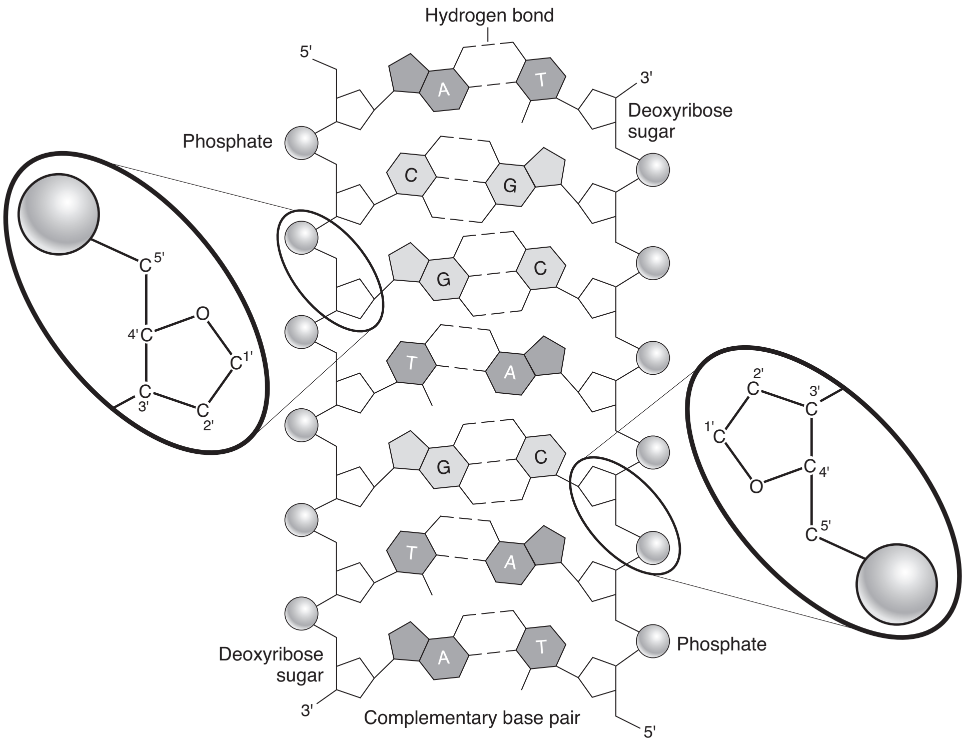
Of course, the triplet codons that specify each of the amino acids are the most direct way to search for protein-coding genes. However, to understand the properties of genes that can be discovered by computer algorithms, we first need to consider some fine points of the genetic code and DNA structure.

We previously defined triplet codons as a property of mRNA, translated at the ribosome into amino acids. However, from a bioinformatics standpoint, it is more useful to deal with the genetic code as it exists in the DNA molecule. By convention, the strand of the DNA molecule that encodes a gene is termed the "sense" strand, and the complementary strand is termed the "antisense" strand. The antisense strand is used as the template to produce mRNA, which means that the mRNA then carries the same code as the sense strand, with uracil (U) in mRNA replacing thymine (T) in DNA. In this way, the genetic code carried by a DNA sequence and one that is carried by an mRNA sequence are completely interchangeable.

Each strand of a DNA molecule has a directionality based on the way in which the phosphate–sugar "rails" of the DNA "ladder" are joined when DNA is synthesized (replicated). Each deoxyribose sugar is composed of five carbons, labeled 1′ to 5′. The phosphates that link adjacent nucleotides form covalent bonds between the 3′ and 5′ carbons of adjacent sugars. (These are technically termed phosphodiester linkages, because the hydroxyl groups of phosphoric acid are replaced by oxygen linkages to two deoxyribose molecules.) In every organism ever studied, an incoming nucleotide is always added onto the 3′ carbon during DNA synthesis. Thus, DNA synthesis occurs in a 5′ to 3′ direction, and a DNA strand is referred to as running 5′ to 3′. The sequence of nucleotides forming a gene is also read in a 5′ to 3′ direction. If one considers genetic information as flowing downhill, like a river, 5′ is considered "upstream" and 3′ "downstream."

Importantly, the two strands of any DNA molecule are antiparallel, i.e., they run in opposite 5′ to 3′ orientations. Although the sequence of the paired strands is complementary, each is "read" in the opposite direction. Thus, each strand of a DNA double helix carries a unique coding sequence. Furthermore, the codons on each strand can be "read" in any of three different ways, depending on whether one starts at the first, second, or third nucleotide position of a DNA sequence. Each of these different ways of "spelling out" codons is termed a "reading frame." There are three reading frames on each strand of DNA, making a total of six reading frames per double-stranded DNA molecule. In any given region of the genome, several genes may be present in any of these six reading frames—on opposite strands or even overlapping in different frames of the same strand.

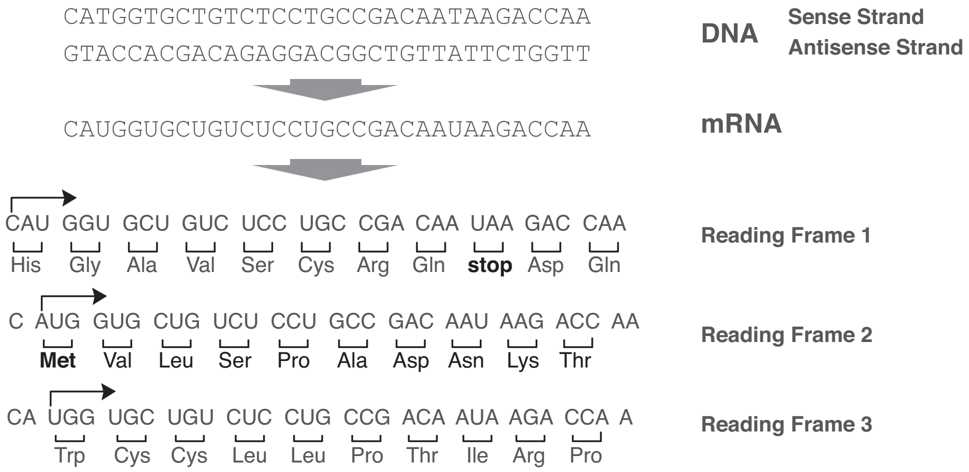
The first details of gene structure and function were analyzed in simple organisms. Bacteria are examples of prokaryotes ("pro," before and "karyon," kernal or nucleus) that lack an organized nucleus. The average prokaryotic protein has ~250 amino acids, so in any given reading frame, there are ~250 codons. However, according to chance alone, each of the three stop codons will occur once in every 64 codons of a given reading frame, so that on average there will be one stop codon in every 21 codons of a reading frame. Any reading frame with frequent stop codons is "closed" to the possibility of a functioning gene, whereas a reading frame with hundreds of contiguous codons that potentially encode amino acids is "open." A prokaryotic gene is a



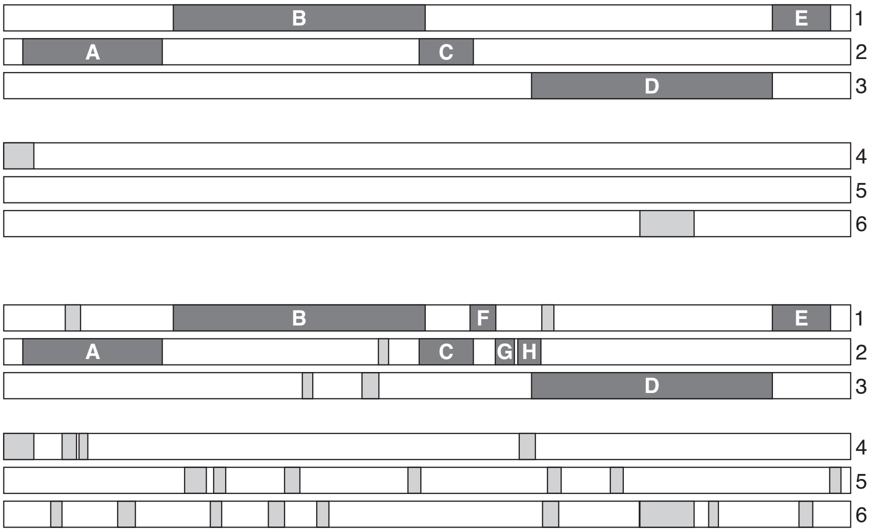
Chemical structure of DNA. Insets show the antiparallel orientation of 5′ and 3′ carbons on the complementary strands.

simple structure—an open reading frame (ORF) beginning with a start codon (ATG), followed by a long string of triplet codons that encode amino acids, and ending with a stop codon (TAA, TAG, or TGA).

ORF prediction programs are straightforward to use, with the main adjustable parameter being the minimum number of consecutive codons that is scored as an


CATGGTGCTGTCTCCTGCCGACAATAAGACCAA
GTACCACGACAGAGGACGGCTGTTATTCTGGTT
DNA
Sense Strand
Antisense Strand
CAUGGUGCUGUCUCCUGCCGACAAUAAGACCAA
mRNA
CAU GGU GCU GUC UCC UGC CGA CAA UAA GAC CAA
His Gly Ala Val Ser Cys Arg Gln stop Asp Gln
Reading Frame 1
C AUG GUG CUG UCU CCU GCC GAC AAU AAG ACC AA
Met Val Leu Ser Pro Ala Asp Asn Lys Thr
Reading Frame 2
CA UGG UGC UGU CUC CUG CCG ACA AUA AGA CCA A
Trp Cys Cys Leu Leu Pro Thr Ile Arg Pro
Reading Frame 3

mRNA transcription and reading frames. By convention, mRNA is transcribed from the antisense strand of DNA. The mRNA (coding sequence) has the same sequence as the sense strand, except that thymine in DNA is replaced by uracil in RNA. Each reading frame is a series of triplet codons specifying amino acids. Each of the three reading frames is offset by one nucleotide, yielding a different set of codons.



Open reading frame predictions in the HIV genome. (*Top*) Set to a minimum threshold of 300 nucleotides, an ORF prediction program correctly identifies five large genes (*A–E*), as well as two false positives (gray) in six reading frames. (*Bottom*) Lowering the threshold to 100 nucleotides identifies three additional genes (*F–H*), as well as 25 false positives. Shorter sequences—the *rev* gene (91 nucleotides) and part of the *tat* gene (48 nucleotides)—are missed. (A) *gag*, (B) *pol*, (C) *vif*, (D) *env*, (E) *nef*, (F) *vpr*, (G) *tat*, (H) *vpu*.

ORF. This consideration balances sensitivity—failing to detect real genes—and specificity—misidentifying ORFs that are not real genes. If the cutoff is set too high, many smaller genes are not identified as ORFs. If the cutoff is too low, many of the smaller ORFs are not true genes. Although a typical cutoff of 100 codons detects a high percentage of genes in a bacterial genome, it still identifies a substantial number of random ORFs that are not real genes.

## 3′- AND 5′-UNTRANSLATED REGIONS AND PROMOTERS

A gene is more than a simple open reading frame of codons transcribed into mRNA and translated into protein. During the process of transcription and translation, a number of protein and RNA molecules are recruited to bind directly with DNA and mRNA to regulate the expression of a gene. Because the regulatory sequences carried by mRNA molecules are also encoded in genomic DNA, a computer can search genomic DNA for sequences that regulate both DNA transcription and mRNA translation.

Evolutionary conservation of a common set of regulatory molecules—maintained across a range of organisms—has created a bias toward the use of certain nucleotide combinations in protein-binding sites. However, most protein-binding sites are not identical among species or even among different genes within a species. Rather, each binding site is represented by a consensus sequence of six to ten nucleotides, which is the most frequent combination of nucleotides found at the binding site. Functional binding sites may have combinations that differ by several nucleotides from the consensus. Some consensus sequences have several invariable nucleotide positions in combination with several variable positions. Computer algorithms translate these patterns into a scoring matrix, ratcheting along a DNA sequence in overlapping "windows" and evaluating each against the consensus sequence.

It is important to remember that transcription and translation occur separately and at different times and places in the cell. The start and stop codons that define the ends of a protein do not define the beginning and end of an mRNA. In fact, mRNAs typically include sequences that extend upstream (5′) of the start codon and downstream (3′) from the stop codon. Because these sequences are not translated into amino acids, they are termed untranslated regions (UTRs). Thus, *translation* begins with the start codon (ATG) and ends with a stop codon (TAA, TAG, or TGA), but *transcription* begins with the 5′ UTR and ends with the 3′ UTR.

The majority of eukaryotic mRNAs have a distinctive 3′ feature—a poly(A) tail composed of a long tract of adenine nucleotides. The poly(A) tail stabilizes the mRNA; over time, the tail shortens, and the mRNA is degraded when the poly(A) tail reaches a critical length. The poly(A) tail is not part of the gene sequence, but rather, it is added post-transcriptionally, after the mRNA has been generated. Polyadenylate polymerase cleaves the mRNA 11–30 nucleotides 3′ of a consensus poly(A) signal—A(A/U)UAAA—then adds a string of tens or hundreds of adenine residues. Thus, identifying the poly(A) signal in the DNA—A(A/T)TAAA—helps to define the end of the 3′ UTR in eukaryotes.

Although the poly(A) signal conveniently defines the 3′ end of a gene, the 5′ end is more difficult to define. The 5′ start site can be inferred from promoter sequences that position RNA polymerase at the transcription start site. RNA polymerase II (Pol II), the eukaryotic polymerase that transcribes DNA, cannot initiate transcription on its own but requires the assistance of a number of accessory proteins called transcription factors (TFs), of which TFIIB and TFIID are the most well studied. The core, or basal, promoter provides binding sites for TFs that provide a maintenance level of transcription. Additional elements, termed transcriptional activators, bind promoter sequences with one surface and interact with transcription factors with another surface. In this way, a number of protein molecules work together to recruit Pol II to the transcription start site. Thus, there are a number of binding sites for transcription factors and activators in the 5′ promoter region and in the 5′-coding region. The positions of promoter and activator sequences are denoted by negative numbers to indicate their positions upstream of the transcription start site (position 1).

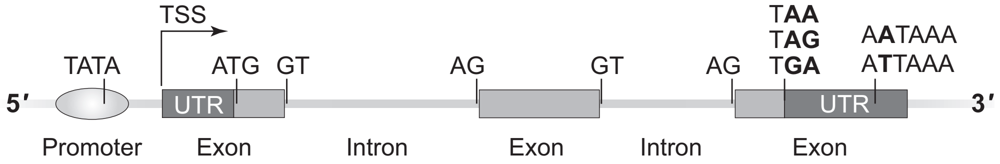
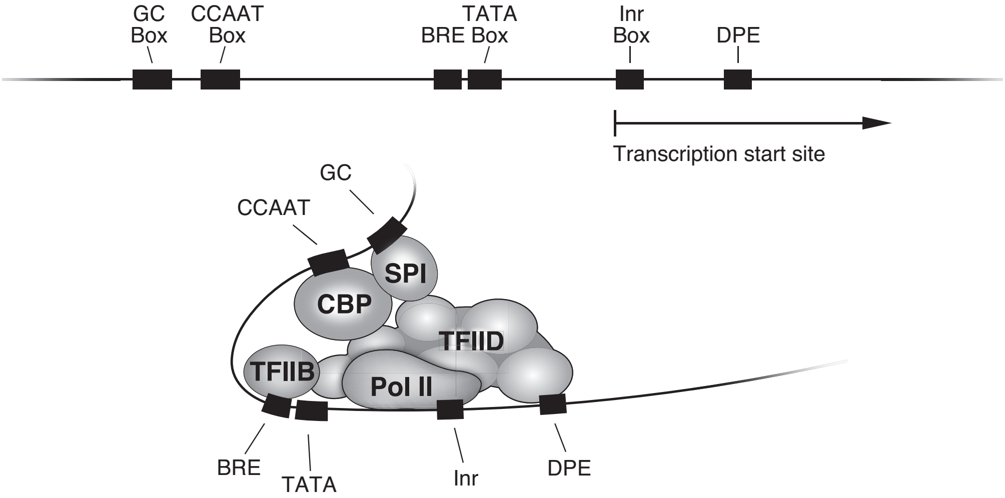


Diagram of a eukaryotic protein-coding gene. The DNA code is "read" from 5′ to 3′, as defined by the orientation of carbon atoms in the deoxyribose backbone. The core, or basal, promoter, located 5′ of the gene, provides binding sites for proteins that position RNA polymerase II at the transcription start site (TSS). Transcription begins approximately 30 nucleotides after the TATA box, one of four major elements that compose the core promoter. Exons (filled boxes) alternate with introns that are spliced out of the pre-mRNA (thinner lines) and whose boundaries are defined by GT and AG sequences. Translation of amino acids begins with the start codon ATG (methionine) and ends with one of three stop codons (TGA, TAG, or TAA). Contrary to popular belief, all exons are not translated into amino acids; instead, the first exon begins and the final exon ends with an untranslated region (UTR; dark box). It is not uncommon for the 3′ UTR to span several exons. The end of the 3′ UTR is defined by the recognition signal for polyadenylate polymerase (A[A/T]TAAA), which cleaves about 20 nucleotides from the end of the transcript and adds a string of adenine residues, the poly(A) "tail."



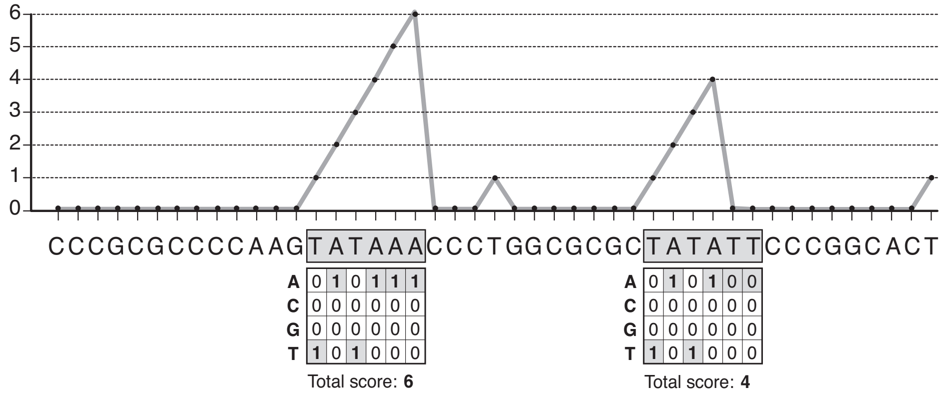
Eukaryotic promoter elements and transcription factors. Promoter elements provide binding sites for transcription factors that help to position RNA polymerase II (Pol II) at the transcription start site. Different genes may have different combinations of promoter sequences that recruit varied sets of transcription factors.

The eukaryotic core promoter is composed of four major sequence elements: TATA box, initiator (Inr) box, downstream promoter element (DPE), and TFIIB recognition element (BRE). Although many genes have a consensus site for only one of these elements, there is strong evidence that additional transcription factors may bind to promoters nonspecifically.

The TATA box is the most well-studied promoter element, and it is found in 20%–50% of eukaryotic genes. Located at positions –23 to –33 before the transcription start site, the eukaryotic consensus sequence, TATAAA, is slightly different from the prokaryotic consensus, illustrating why gene prediction programs must be tuned for different organisms. The TATA box binds to a subunit of TFIID, appropriately called the TATA-binding protein (TBP). The TBP is the first protein to bind DNA to initiate transcription and binds the promoter region even in genes that do not contain a TATA box. TBP binding introduces a kink in the DNA molecule and stresses hydrogen bonds in the region, causing the helix to open slightly. (The double hydrogen bonds in this A = T–rich region denature more easily than G ≡ C triple bonds.) The denatured region allows easier access for RNA polymerase to begin transcription.

The Inr box, which binds TFIID, is the most frequent core element, and it is found in 40%–65% of eukaryotic genes. This pyrimidine (C and T)–rich sequence, (C/T)(C/T)$A_{+1}$N(A/T)(C/T)(C/T), straddles the transcription start site. DPE also binds TFIID and, in some species, may work with Inr to function as the core promoter when the TATA sequence is absent. DPE is located in the 5′ coding region, +23 to +33, with the consensus (A/G)G(A/T)(C/T)(A/C/G). TFIIB binds to a B-recognition element (BRE) at –42 to –32, with the consensus (C/G)(C/G)(A/G)CGCC.

The CCAAT box, with the consensus GGNCAATCT, is an activator sequence found in about half of vertebrate genes. This sequence binds a number of different CCAAT-box binding proteins (CBPs). Although the CCAAT box is usually located at –40 to –100, it can be located near the Inr box or DPE in promoters without TATA boxes. The transcriptional activator Sp1 binds the GC box located at –40 to –100, which has the consen-



|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
| A | 0 | 1 | 0 | 1 | 1 | 1 |
| C | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 0 | 0 | 0 | 0 | 0 |
| T | 1 | 0 | 1 | 0 | 0 | 0 |

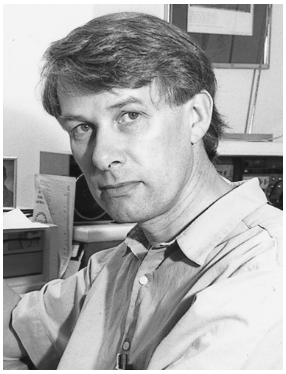
|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
| A | 0 | 1 | 0 | 1 | 0 | 0 |
| C | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 0 | 0 | 0 | 0 | 0 |
| T | 1 | 0 | 1 | 0 | 0 | 0 |

Scoring matrix for TATA box. Computer algorithms use a scoring matrix to search genomic sequence for a close match to a DNA regulatory element, such as the TATA box. The computer ratchets along a DNA sequence, analyzing the positions in each successive "window" of six nucleotides. Each nucleotide that matches its corresponding position in the consensus receives a score of 1; nonmatches score 0. The TATAAA consensus (first gray box) receives a perfect score of 6. Not all TATA boxes perfectly match the consensus sequence; the promoter of the human albumin gene scores 4 with the sequence TATATT.

sus sequence GGGCGG. Sp1 regulates the expression of many "housekeeping" genes that are essential for key cellular functions in vertebrates. The GC box, which may be present in multiple copies in the promoter region, is thought to be the source of the association between elevated G + C levels and CpG dinucleotide enrichment in genic regions.

Enhancers, which further increase transcription, are poorly understood. They may be located hundreds or thousands of nucleotides upstream of the transcription start site, within the transcribed gene or an adjacent gene, or on the opposite DNA strand. Regardless of the enhancer's position, enhancer-binding proteins interact with the enhancer as well as transcription factors assembled at the promoter. The enhancer does not operate from a distance; rather, a loop in the chromosome brings it into proximity with the core promoter.

## EXONS AND INTRONS

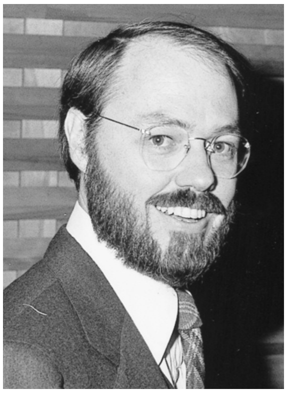


Richard Roberts

(Courtesy of Cold Spring Harbor Laboratory Archives.)

In 1977, Richard Roberts, at Cold Spring Harbor Laboratory, and Philip Sharp, at the Massachusetts Institute of Technology, independently discovered that most eukaryotic genes are not ORFs composed of a contiguous sequence of codons. Using adenovirus, both groups created heteroduplex molecules by hybridizing an mRNA to single-stranded genomic DNA. Electron microscopy revealed that the mRNA hybridized to discontinuous regions of the DNA, throwing out loops of DNA that were not represented in the mRNA. Their explanation was that adenovirus genes are "split," with protein-coding regions (exons) interrupted by nonprotein-coding regions (introns) that are not represented in mRNA. During transcription, the entire gene is copied into a precursor mRNA (pre-mRNA), which includes exons and introns. Subsequent "RNA processing" removes the intervening introns to form a contiguous coding sequence. This "mature" mRNA passes out of the nucleus and attaches to a ribosome for translation.

Later work explained the mechanism of RNA splicing at the spliceosome, a nuclear complex of numerous proteins and five *s*mall *n*uclear RNAs (snRNAs) aver-

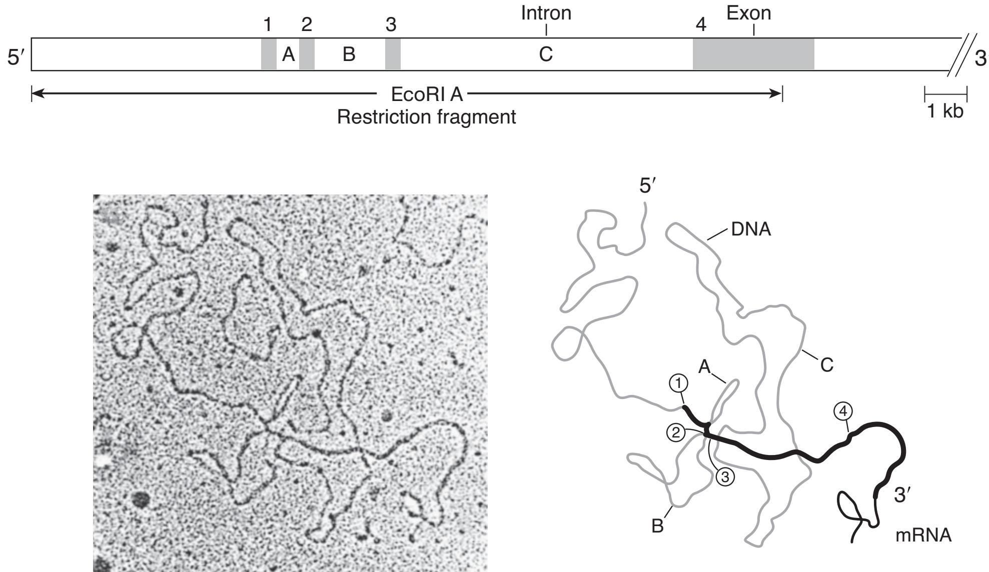


Philip Sharp
(Courtesy of Cold Spring Harbor Laboratory Archives.)

aging ~150 nucleotides in length. The spliceosome assembles on pre-mRNA, aligning 5′ and 3′ splice sites to loop out the intron to form a "lariat." The lariat is then cleaved, and the 5′ and 3′ exon junctions are ligated together. Complementary sequences in snRNAs recognize consensus sequences that define adjacent exon/intron borders, bringing them into proximity for the splicing reactions to occur. The mRNA consensus sequence at the 5′ splice junction is CAG/GUAAGU, whereas the 3′ consensus is UUUUCCCUCCAG/GU. Notably, GU and AG nucleotides define the 5′ and 3′ ends of virtually every eukaryotic intron. At the DNA level, introns invariably begin with GT and end with AG. Using this fact, in combination with other nucleotides in the consensus, a scoring matrix can efficiently identify introns and exons in DNA sequence. After introns are eliminated by splicing, every eukaryotic mRNA is, in fact, an ORF bounded by a start and stop codon.

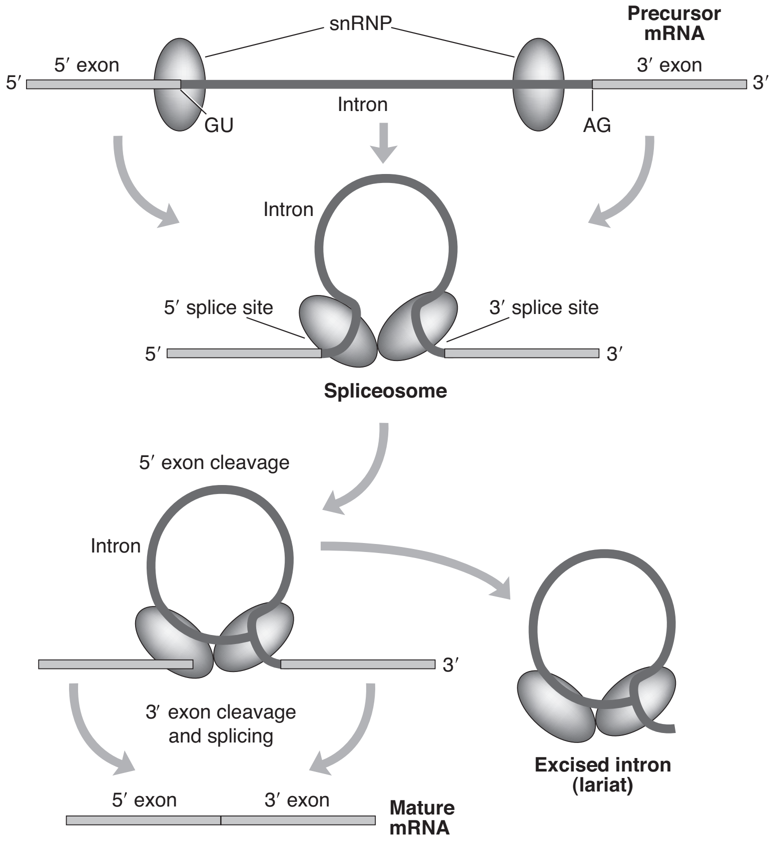
The simple situation in prokaryotes—where an ORF in the genomic DNA sequence is identical to a contiguous mRNA sequence—is complicated in eukaryotic genes. The "average" human gene has eight exons, each ~135 nucleotides in length (1080 nucleotides in total) and seven introns of ~2200 nucleotides each (15,400 nucleotides in total). It is rare that more than several exons of a gene are "read" by the transcription machinery in the same contiguous reading frame. Rather, the reading frame shifts over the length of a gene to read different exons in different reading frames.

Stop codons are of no consequence inside introns, because these sequences are removed from the pre-mRNA before translation at the ribosome. A shift to one reading frame avoids any stop codons that may be present in others. Frame shifting is actually dictated by intron length. A eukaryotic reading frame shifts after any intron whose length is not divisible by three, compensating +1 or +2 to maintain the triplet codons that define

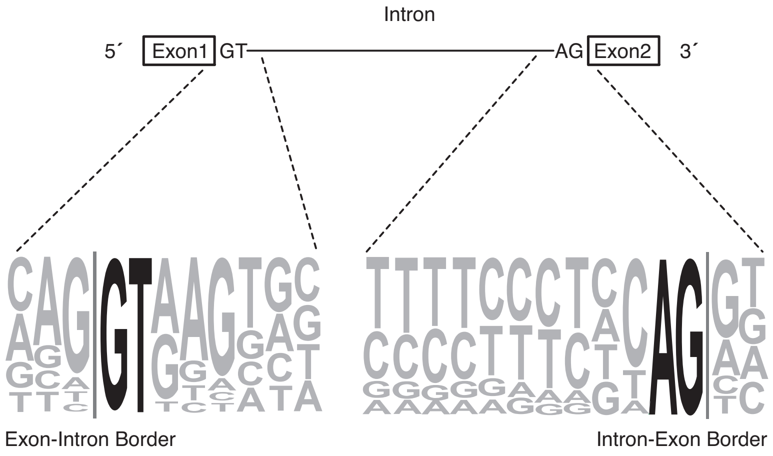


Electron microscopic evidence for RNA splicing. An EcoRI restriction fragment of adenovirus genomic DNA was hybridized to its corresponding mRNA (*bottom left*). In the diagram at right, mRNA (black) and genomic DNA (gray) form a double-stranded molecule in complementary coding regions (1, 2, 3, 4 in gene diagram). Introns A, B, and C are thrown out as loops.

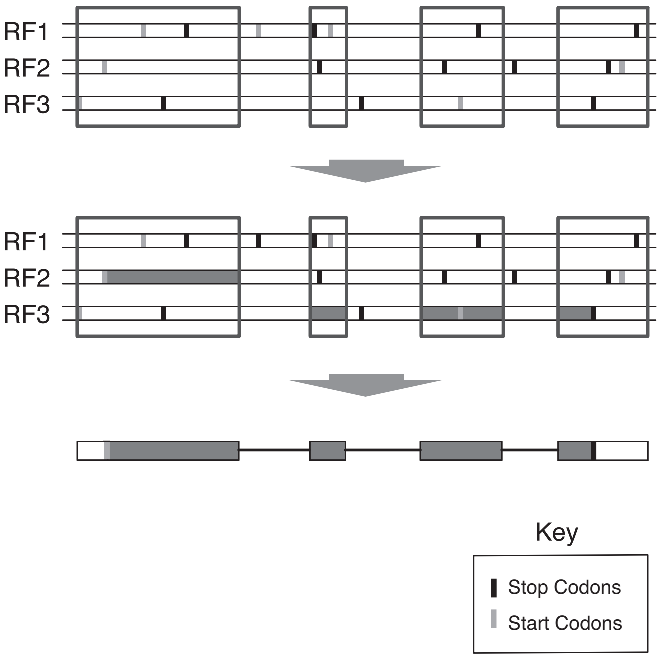
(*Bottom*: Reprinted, with permission, from Berget SM, Moore C, Sharp PA. 1977. *Proc Natl Acad Sci* 74: 3171–3175.)



Mechanism of RNA splicing. Conserved GU and AG residues define the intron borders and align a precursor mRNA in the spliceosome. First, the 5′ intron/exon junction is cleaved and an unusual 5′-2′ phosphodiester bond joins the 5′ end of the intron to an adenine residue at the 3′ end of the intron. The resulting loop structure resembles a lariat. The 3′ intron/exon junction is then cleaved to release the lariat, and the two exons are joined.



“Sequence logo” of consensus sequences at intron/exon boundaries. Introns are demarcated in genomic sequence by GT and AG. However, because these sequences are common in genic regions, merely searching for GT and AG will turn up many false exons, as well as a few real ones. Computer algorithms use a scoring matrix to evaluate the consensus sequence surrounding the GT and AG dinucleotides. The sequence logo visually summarizes these characteristics, with the height of each letter showing its probability at that position in validated splice junctions.



Frame shifting in a eukaryotic gene. Layering start/stop codons in the three reading frames on one DNA strand illustrates the fact that the coding sequences of most eukaryotic genes "shift" between several reading frames. Each of the three reading frames has several stop codons, and none is "open" over the entire coding sequence. The four exons of the gene are boxed. In this case, the coding sequence begins in reading frame 2 for exon 1 and then shifts to reading frame 3 for exons 2, 3, and 4. Stop codons in introns are of no consequence.

the ORF. Thus, the location of splice sites is the most important factor in determining the exon structure of a eukaryotic gene. Put most simply, an ORF is a property of mRNA, whereas intron/exon boundaries are properties of DNA and pre-mRNAs.

## METHODS FOR FINDING GENES IN GENOMIC SEQUENCE

The raw DNA sequence is the starting point for understanding a genome. In some cases, the initial search for genes is narrowed by focusing on CpG islands and regions with high G + C content. Whether narrowed in this way or not, two types of computational methods are used to identify genes within DNA sequence. Pattern-based programs use algorithms to search for sequences associated with gene features, including start/stop codons, coding triplets, intron/exon boundaries, promoters, and poly(A) signals. This is termed ab initio (from the beginning) gene finding, because genes are predicted directly from DNA sequence. Comparative programs look for similarities between the genome sequence and independent sequence evidence from the organism under study and from related organisms. The best gene-finding programs now incorporate both pattern-based and comparative strategies, providing increasingly accurate gene models.

Most pattern-based programs are trained on representative genes to develop a "hidden Markov model (HMM)," which identifies the gene patterns "hidden" in DNA sequence. Notably, different organisms have preferences, or biases, among synony-

**ACAUUUGCUUCUGACACAACUGUGUUCACUAGCAACCUCAAACAGACACCAUGGUGCACCUGACUCCUGAGGAGAAC**
(met) val his leu thr pro glu glu lys
**1 5**

**UCUGCCGUUACUGCCCUGUGGGGCAAGGUGAACGUGGAUGAAGUUGGUGGUGAGGCCCUGGGCAG**GUUGGUAUCAAG
ser ala val thr ala leu trp gly lys val asn val asp glu val gly gly glu ala leu gly arg
**10 15 20 25 30**

GUUACAAGACAGGUUUAAGGAGACCAAUAGAAACUGGGCAUGUGGAGACAGAGAAGACUCUUGGGUUUCUGAUAGGC

ACUGACUCUCUCUGCCUAUUGGUCUAUUUUCCCACCCUUAG**GCUGCUGGUGGUCUACCCUUGGACCCAGAGGUUCUUU**
leu leu val val tyr pro trp thr gln arg phe phe
**31 35 40**

**GAGUCCUUUGGGGAUCUGUCCACUCCUGAUGCUGUUAUGGGCAACCCUAAGGUGAAGGCUCAUGGCAAGAAAGUG**
glu ser phe gly asp leu ser thr pro asp ala val met gly asn pro lys val lys ala his gly lys lys val
**45 50 55 60 65**

**CUCGGUGCCUUUAGUGAUGGCCUGGCUCACCUGGACAACCUCAAGGGCACCUUUGCCACACUGAGUGAGCUGCAC**
leu gly ala phe ser asp gly leu ala his leu asp asn leu lys gly thr phe ala thr leu ser glu leu his
**70 75 80 85 90**

**UGUGACAAGCUGCACGUGGAUCCUGAGAACUUCAGG**GUGAGUCUAUGGGACCCUUGAUGUUUUCUUUCCCCUUCUUU
cys asp lys leu his val asp pro glu asn phe arg
**95 100 104**

UCUAUGGUUAAGUUCAUGUCAUAGGAAGGGGAGAAGUAACAGGGUACAGUUUAGAAUGGGAAACAGACGAAUGAUUG

CAUCAGUGUGGAAGUCUCAGGAUCGUUUUAGUUUCUUUUAUUUGCUGUUCAUAACAAUUGUGUAUAACAAAAGGAAAU

AUCUCUGAGAUACAUUAAGUAACUAAAAAAAAACUUUACACAGUCUGCCUAGUACAUUACUAUUUGGAAUAUAUGUG

UGCUUAUUUGCAUAUUCAUAAUCUCCCUACUUUAUUUUCUUUUAUUUUAAUUGAUACAUAAUCAUUAUACAUAUUUAUG

GGUUAAAGUGUAAUGUUUUAAUAUGUGUACACAUAUUGACCAAAUCAGGGUAAUUUUGCAUUUGUAAUUUUAAAAAAU

GCUUUCUUCUUUUAAUAUACUUUUUUGUUAUCUUAUUUCUAAUACUUUCCCUAAUCUCUUUCUUUCAGGGCAAUAAUGA

UACAAUGUAUCAUGCCUCUUUGCACCAUUCUAAAGAAUAACAGUGAUAAUUUCUGGGUUAAGGCAAUAGCAAUAUUU

CUGCAUAUAAAUAUUUCUGCAUAUAAAUUGUAACUGAUGUAAGAGGUUUCAUAUUGCUAAUAGCAGCUACAAUCCAG

CUACCAUUCUGCUUUUAUUUUAUGGUUGGGAUAAGGCUGGAUUAUUCUGAGUCCAAGCUAGGCCCUUUUGCUAAUCAU

GUUCAUACCUCUUAUCUUCCUCCCACAG**CUCCUGGGCAACGUGCUGGUCUGUGUGCUGGCCCAUCACUUUGGCAAA**
leu leu gly asn val leu val cys val leu ala his his phe gly lys
**105 110 115 120**

**GAAUUCACCCCACCAGUGCAGGCUGCCUAUCAGAAAGUGGUGGCUGGUGUGGCUAAUGCCCUGGCCCACAAGUAU**
glu phe thr pro pro val gln ala ala tyr gln lys val val ala gly val ala asn ala leu ala his lys tyr
**125 130 135 140 145**

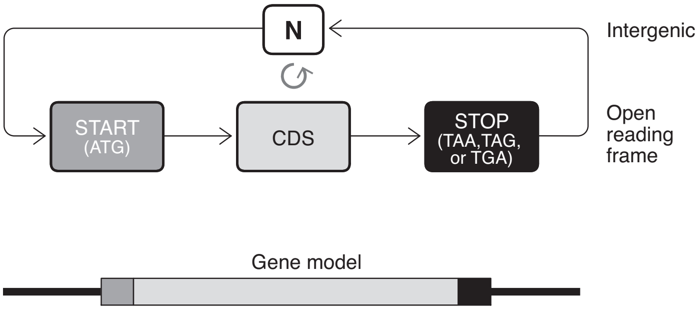
**CACUAAGCUCGCUUUCUUGCUGUCCAAUUUCUAUUAAAGGUUCCUUUGUUCCCUAAGUCCAACUACUAAACUGGGGG**
his **stop**

**AUAUUAUGAAGGGCCUUGAGCAUCUGGAUUCUGCCUAAUAAAAAACAUUUAUUUUCAUUUGC**

β-Globin gene structure. Three exons with encoded amino acids are in bold, and two introns are in plain font.

mous codons, intron and exon lengths, consensus sequences for promoter elements, intron/exon boundaries, and poly(A) signals. Although these overall sequence characteristics cannot be readily discerned by simply inspecting the DNA sequence, HMMs derive statistical information about these biases from the training set.

In the roundworm *Caenorhabditis elegans* and the fruit fly *Drosophila*, HMMs can correctly identify ~90% of individual exons and every exon in ~40% of genes. However, these figures drop to 70% and 20%, respectively, in human DNA. Gene prediction programs readily identify internal exons, which have two splice junctions (left and right) adjacent to two introns. The first and last exons are frequently missed because they have only half of the sequence information used in prediction. Furthermore, exon prediction generates a large number of false positives. Consensus splice site sequences are very common in introns, making "pseudo-exons" common. Additionally, HHMs cannot predict the correct start codon among several possible ATGs in the 5′ region. Often, definitive information on the translation start site can only be determined directly from protein sequence.



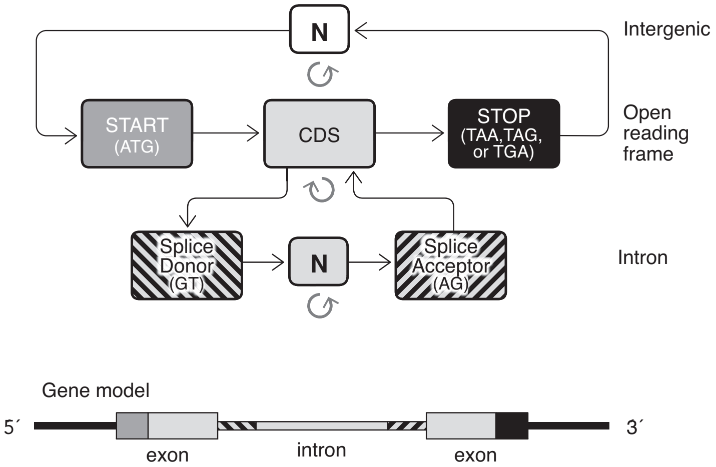
Flow diagram for predicting an ORF gene. A hidden Markov program scans nucleotides (N) until it encounters a potential start codon and then scans for a minimum number of triplet codons that define an ORF. After encountering a stop codon, the program cycles back to the intergenic state (N) and searches for another potential start codon.

Comparative (homology-based) methods find sequence similarities among DNA, RNA, or protein sequences. Homology refers to sequence similarity based on a common origin. Orthologs are similar genes in different species that have arisen due to descent from a common ancestor, whereas paralogs are similar genes within a species that have arisen by the duplication of a single gene. Growing data from parallel sequencing of model species—notably *E. coli*, *Saccharomyces cerevisiae*, *C. elegans*, *Schizosaccharomyces pombe*, fruit fly, mouse, and *Arabidopsis*—provide collections of previously identified genes. A match to a known mRNA sequence from the same species provides direct evidence that a DNA sequence is transcribed—and therefore is a coding sequence—whereas homology with a known gene or protein from another species provides indirect evidence. Homology-based methods identified ~60% of genes in the first draft of the human genome, and 40% of genes were identified ab initio.

Comparative algorithms such as BLAST (Basic Local Alignment Search Tool) or BLAT (BLAST-like Alignment Tool) align raw sequence or gene models to a database of known genes. BLAT is the fastest program for scanning an entire genome, because it stores in RAM memory an index of short DNA or protein sequences (11 or 4 mers) most relevant to the genome under investigation. The BLAT index can fit in the RAM memory of a personal computer, and it can handle long lists of queries simultaneously.

## GENE ANNOTATION

The output from a pattern- or homology-based program is termed a gene model, because it may or may not be an accurate representation of the actual gene. Annotation is the process of adding information about the structure and function of a predicted gene. Structural annotations improve the initial gene model by extending 5′- and 3′-noncoding regions, identifying alternative start and stop codons, sorting out exon structure, and finding alternative splice sites. Functional annotations identify conserved amino acid motifs and, therefore, functions that are shared with other organisms. Although early genome-sequencing efforts relied on annotations submitted by human curators, this time-consuming step is now automated. Considering the exponentially increasing rate at which new DNA sequences are being produced, it seems certain that the vast majority of new sequence data will never be carefully examined by human eyes.

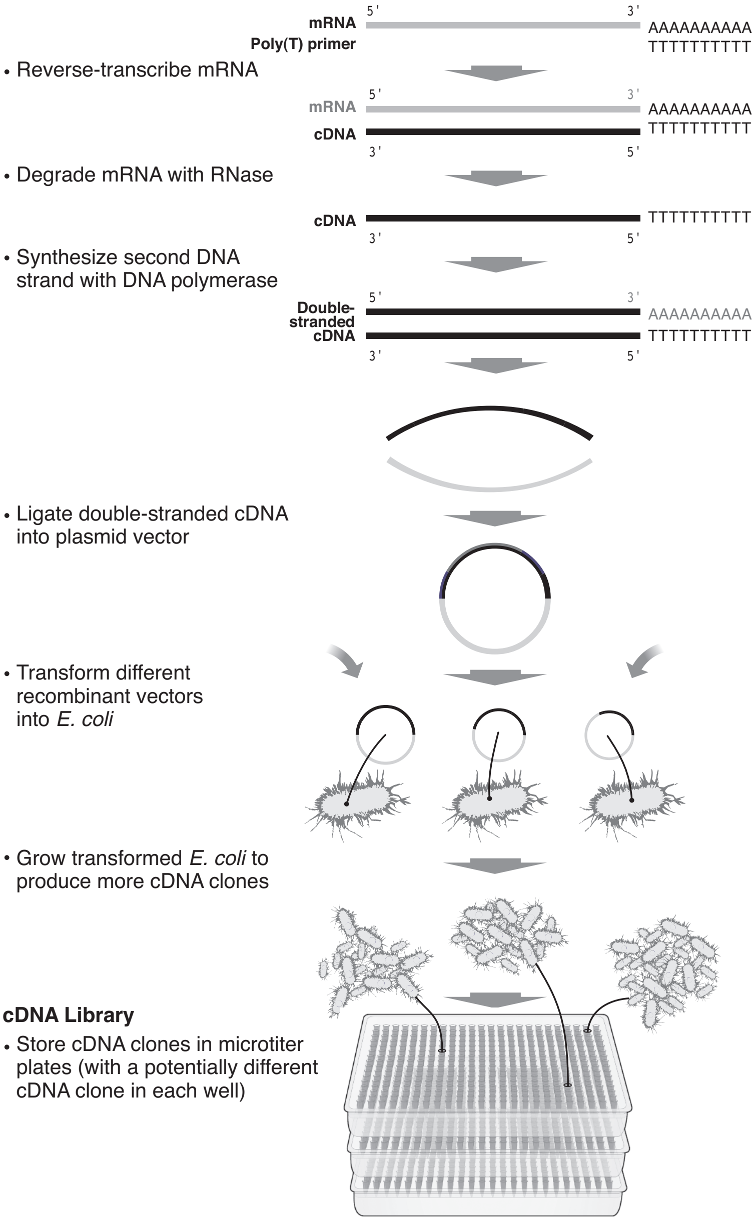


Flow diagram for predicting a gene with an intron. After identifying a potential start codon, a hidden Markov program cycles through triplet codons until it encounters a potential 5′ splice donor (GT) in the context of an appropriate consensus sequence. It then scans intron sequence (N) until it encounters a 3′ splice acceptor (AG) within a consensus sequence. The program then cycles through additional codons until it encounters another 5′ splice donor or a stop codon. When the program encounters a stop codon, it returns to the intergenic (N) state and searches for the next potential start codon.

Confirmation of a predicted gene and annotation of its structure and function must come from independent evidence, usually from homology-based matches to previously annotated genes or mRNA evidence from the species under study and its close relatives. BLASTn uses the predicted gene sequence to search a nucleotide database to discover homologs in closely related species. However, the redundancy of the genetic code usually makes it difficult to uncover more distant relationships at the DNA level. Therefore, BLASTx translates the coding sequence (assembled exons) into an amino acid sequence to search a protein database. BLASTx or BLASTp, which makes protein–protein searches, also highlights any conserved motifs (domains) within the predicted protein, including structures that bind to DNA or other proteins. Programs such as PHYLIP place the predicted protein in a phylogenetic tree that shows its evolutionary relationships to homologs from other organisms.

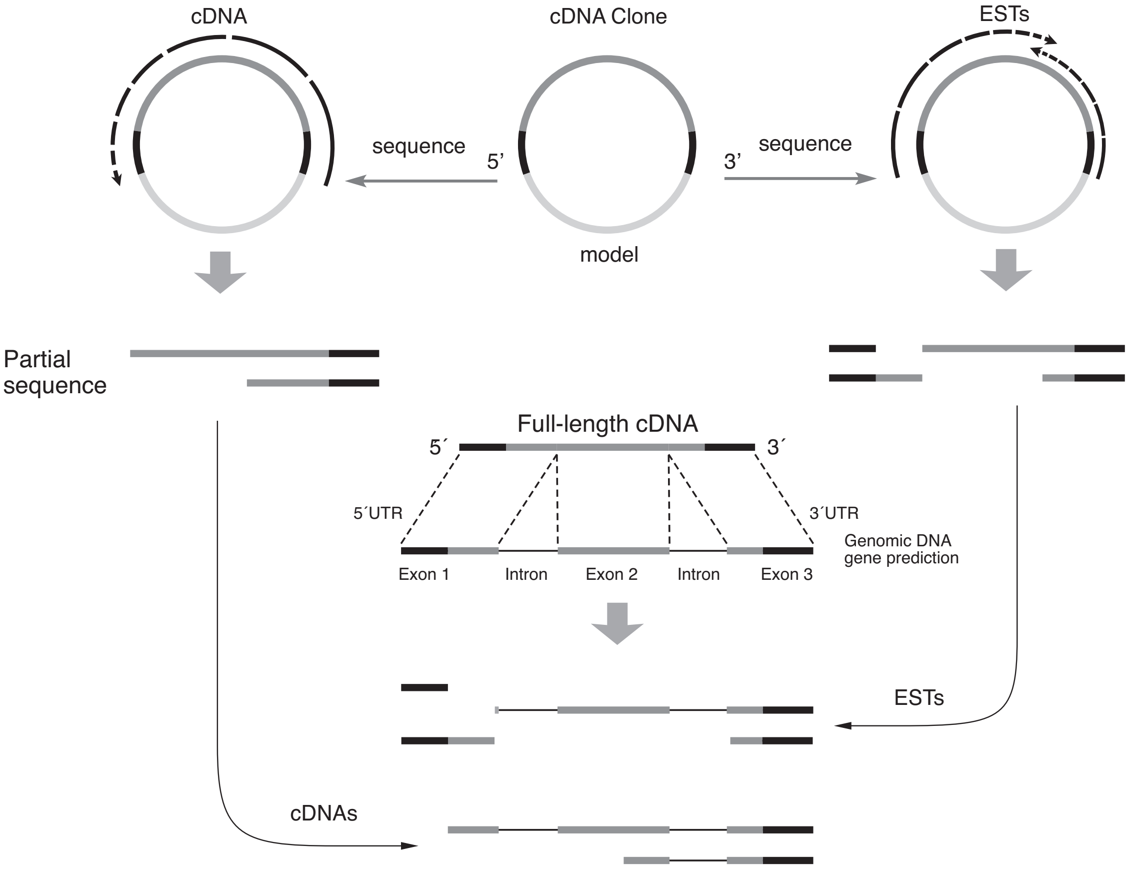
Much existing evidence for gene annotation has come from complementary DNA (cDNA) libraries, which represent the genes expressed by an organism or by a particular tissue or cell type. To make a cDNA library, mRNA is isolated from living cells, and the enzyme reverse transcriptase is used to convert mRNAs into cDNAs. These are copied into doubled-stranded DNAs, which are ligated into plasmid vectors and transformed into *E. coli*, creating a library of thousands of cDNA clones.

Full-length cDNA sequence is the best source of information about the 5′ and 3′ UTRs, which are not accurately predicted by computer programs. The 5′ UTR is the most difficult part of the gene to annotate. Reverse transcriptase extends from the 3′ end of the mRNA template, so incomplete extension typically produces many cDNAs (and cDNA clones) that are missing sequence at their 5′ ends. Expressed sequence tags (ESTs) are single-read DNA sequences obtained using primers at each end of the cDNA cloning vector. Because they can be generated quickly and inexpensively, ESTs are typically the most abundant biological evidence and are often available from related species. However, EST sequences are short (~500 nucleotides) and highly redundant.

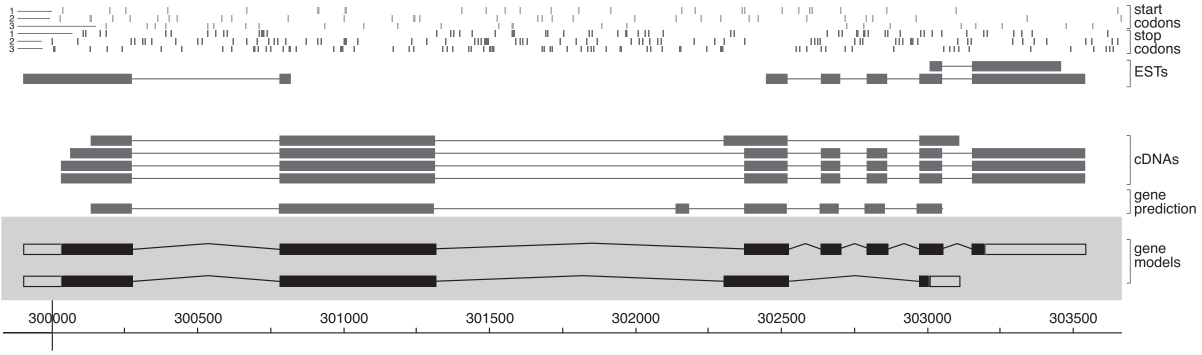

mRNA
Poly(T) primer
5'
3'
AAAAAAAAAA
TTTTTTTTTT
• Reverse-transcribe mRNA
mRNA
cDNA
5'
3'
3'
5'
AAAAAAAAAA
TTTTTTTTTT
• Degrade mRNA with RNase
cDNA
3'
5'
TTTTTTTTTT
• Synthesize second DNA strand with DNA polymerase
Double-stranded cDNA
5'
3'
3'
5'
AAAAAAAAAA
TTTTTTTTTT
• Ligate double-stranded cDNA into plasmid vector
• Transform different recombinant vectors into *E. coli*
• Grow transformed *E. coli* to produce more cDNA clones
cDNA Library
• Store cDNA clones in microtiter plates (with a potentially different cDNA clone in each well)

Making a cDNA library.

Annotation programs, such as Apollo, visually align gene models, cDNAs, and ESTs to genome sequence, providing evidence to extend 3′ and 5′ UTRs and confirm exon structure. ESTs and cDNAs with different arrangements of exons define alternatively spliced forms of mRNA that produce different proteins from the same genomic sequence.

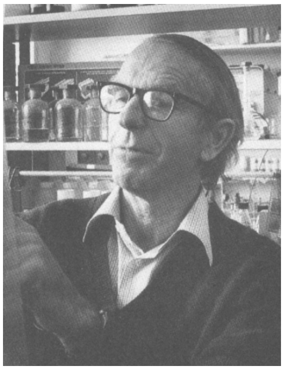


cDNA and EST evidence. A clone from a cDNA library is directly sequenced from the 3′ end to produce high-quality cDNA sequence (*left*). A full-length cDNA sequence offers the best evidence for confirming gene structure. Expressed sequence tags (ESTs) are short reads from the 5′ and 3′ ends of a cDNA clone *(right*). Because they are relatively easy to produce, ESTs typically offer abundant but short and highly redundant sequence evidence. Computer programs align cDNA and EST evidence with genome sequence to confirm the exon/intron structure of predicted genes.



Gene annotation and alternative splicing. This screen shot from the Apollo annotation program shows two alternatively spliced gene models based on a gene prediction plus EST and cDNA evidence.

## SANGER DIDEOXY DNA SEQUENCING



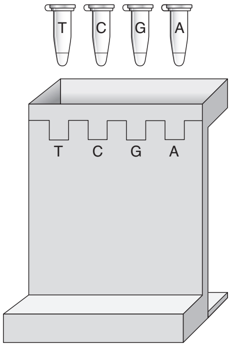
Fred Sanger

(Courtesy of Cold Spring Harbor Laboratory Archives.)

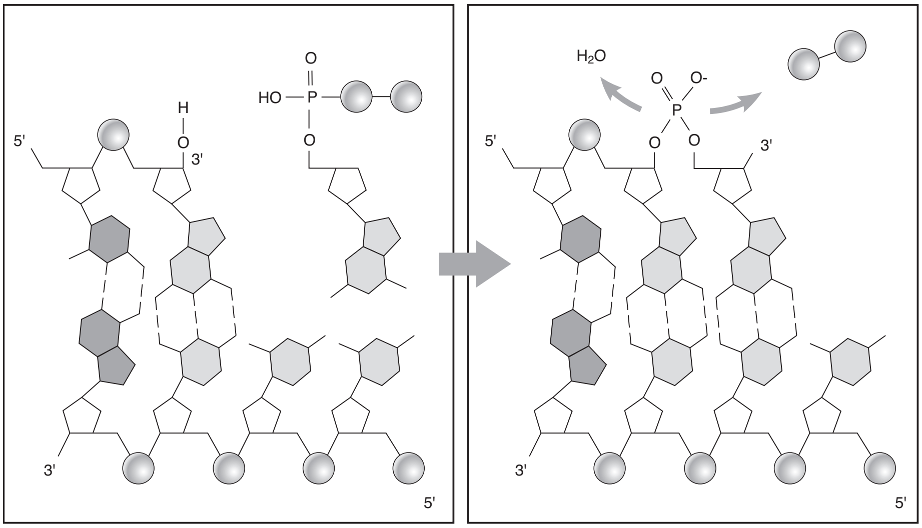
Modern genome sequencing was launched in 1977 when Fred Sanger at the Medical Research Council's Laboratory of Molecular Biology in Cambridge, England, developed a sequencing method based on enzymatic DNA synthesis. During primer extension by DNA polymerase, discovered in the 1960s, DNA synthesis is "primed" by short single-stranded primers that hybridize to each DNA strand. In the presence of the four deoxynucleotide triphosphates (dNTPs)—dATP, dTTP, dCTP, and dGTP—DNA polymerase will add nucleotides complementary to the single-stranded DNA template and extend the double-stranded region. Sanger found that if dideoxynucleotide triphosphates (ddNTPs) were included in the reaction, DNA elongation stopped when a ddNTP was incorporated. This occurs because ddNTPs lack a 3′ hydroxyl group (–OH), which is needed to form the phosphodiester linkage that joins adjacent nucleotides.

In the original dideoxy sequencing protocol, four reaction tubes (A, T, C, and G) are set up. Each of the reactions contains a DNA template, a primer sequence, DNA polymerase, and the four dNTPs (dATP, dTTP, dCTP, and dGTP), one of which is radioactively labeled. A single type of ddNTP is added to each of the four reactions—ddATP (to tube A), ddTTP (tube T), ddCTP (tube C), or ddGTP (tube G). Working from the primer, DNA polymerase randomly adds dNTPs or ddNTPs that are complementary to the DNA template. The ratio of dNTPs to ddNTPs in the reaction is adjusted so that a ddNTP is incorporated into the elongating DNA chain approximately once every 100 nucleotides. When a ddNTP is incorporated, synthesis stops, and a DNA strand of a discrete size is generated. After replication, there are millions of copies of the DNA sequence, each of which terminated at a different nucleotide position.

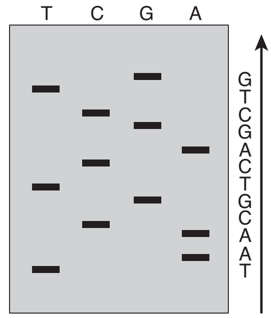
When the reactions are complete, the newly synthesized strands are denatured, and each reaction is loaded onto a different lane of a polyacrylamide gel. The synthesized fragments migrate through the gel according to size, and each lane eventually



Sanger's dideoxy DNA sequencing. Sequencing reactions are electrophoresed in a vertical polyacrylamide gel.



5′ to 3′ synthesis of DNA. (*Left*) An incoming nucleoside triphosphate aligns with the template strand. (*Right*) The 5′ carbon of the incoming nucleotide is joined to the 3′ carbon of the growing strand by way of a phosphodiester linkage. Pyrophosphate (two phosphate balls) and water are liberated.

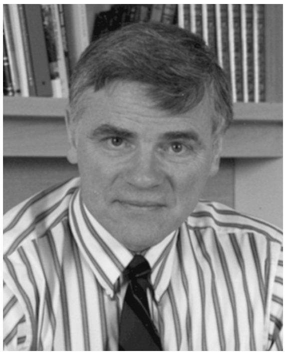


Autoradiogram of dideoxy sequencing gel. Bands are read from bottom to top of gel to generate sequence.

resolves to form a "ladder" of bands. Each band on the gel differs in length by a single nucleotide. Following electrophoresis, the gel is placed in contact with X-ray film. The radioactively labeled nucleotides expose the X-ray film, revealing the series of bands generated in the A, T, C, and G reactions. The gel is then "read" from bottom to top, beginning with the smallest DNA fragment and then scanning across the lanes to identify each successively larger fragment. Optical scanners became the first element of automation in DNA sequencing, producing computer files of the finished sequences. Using this approach, the following first small genomes were sequenced in the 1970s and 1980s:

- The bacterial virus ϕX174, 5386 nucleotides, by Frederick Sanger (1977).
- The mammalian virus SV40, 5224 nucleotides, by Walter Fiers, University of Ghent (1978).
- The human mitochondrion, 16,569 nucleotides, by Stephen Anderson, MRC Laboratory of Molecular Biology (1981).

## AUTOMATED DNA SEQUENCING


Leroy Hood
(Courtesy of Leroy Hood.)

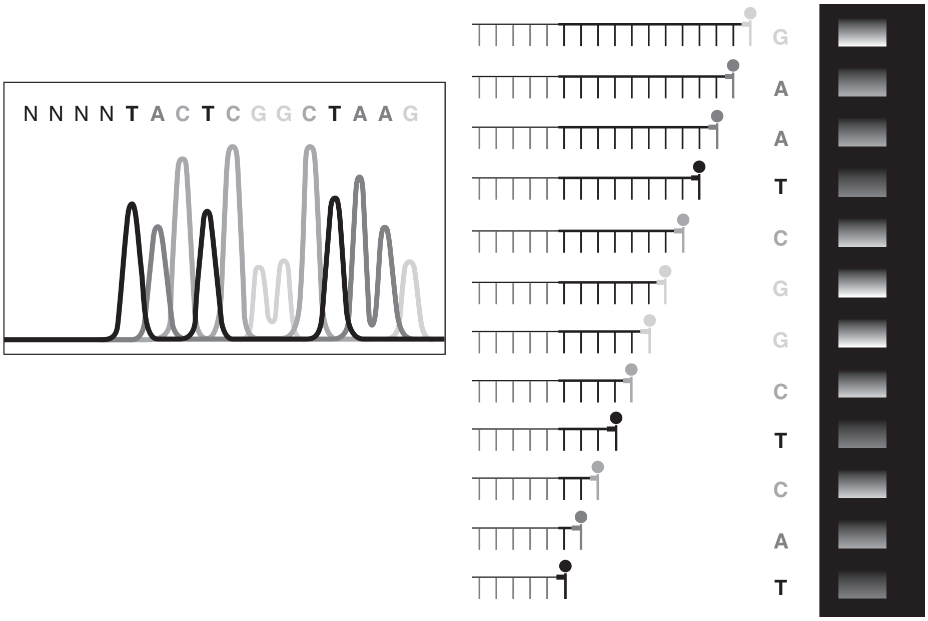
Automated sequencing was made possible by dye chemistry developed by Leroy Hood and Lloyd Smith at the California Institute of Technology. In 1986, they paired a different fluorescent dye with each of the four ddNTP reactions. The four sequencing reactions were loaded onto a single lane of a sequencing gel, and the fluorescent labels were detected as the terminated fragments passed an argon laser aimed at the bottom of the gel. When excited by the laser light, each fluorescent terminator emitted a colored light of a characteristic wavelength, which was then interpreted by computer software as an A (green), T (red), C (blue), or G (yellow) at that position.

Hood then collaborated with Mike Hunkapiller at Applied Biosystems, Inc. (ABI) to produce the first commercial instrument to read sequences from dye-labeled fragments. The ABI Model 370 DNA sequencer, first marketed in 1987, used a polyacrylamide slab gel to resolve ladders of DNA fragments labeled with fluorescent nucleotides. The sequencer incorporated a computer program that built a simulated gel image of fluorescent DNA bands as they were detected by a scanning laser at the bottom of the electrophoresis bed. The final output took the form of an electropherogram, showing colored peaks corresponding to each nucleotide position. The ABI Model 370 DNA sequencer, equipped with a 16-lane polyacrylamide gel, had the capacity to sequence as many as 20,000 nucleotides per day. Increasing the number of lanes to 32, 48, and, ultimately, 96 brought the daily output of each machine to 120,000 nucleotides or more.

Lloyd M. Smith
(Photo by Jim Dahlberg, courtesy of Lloyd M. Smith.)

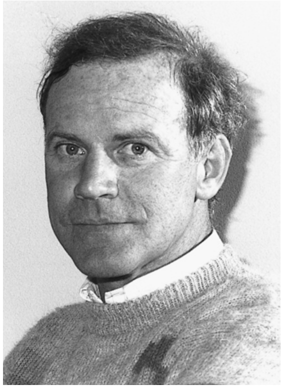
By allowing all four nucleotides to be analyzed in a single lane, Hood's fluorescent chemistry quadrupled the output of sequencing gels. Parallel improvements in DNA preparation further increased output. DuPont introduced "dye terminators," which attach a different fluorescent dye directly to each of the four terminator nucleotides (didATP, didTTP, didCTP, or didGTP). This allowed all four nucleotides to be labeled simultaneously in a single reaction. Polymerase chain reaction (PCR) was pressed into service to automate dye labeling, a hybrid method that became known as cycle sequencing.

The foundation PCR technology was discovered in 1985 by Kary Mullis at Cetus Corporation and uses enzymatic amplification to increase the copy number of a DNA



Sample output from an automated DNA sequencer. Electropherogram showing peaks corresponding to each nucleotide position (*left*). Representation of the dye-terminated sequences that correspond to each peak (*center*). Computer-generated representation of the sequencing gel (*right*).

Mike Hunkapiller
(Courtesy of Mike Hunkapiller.)

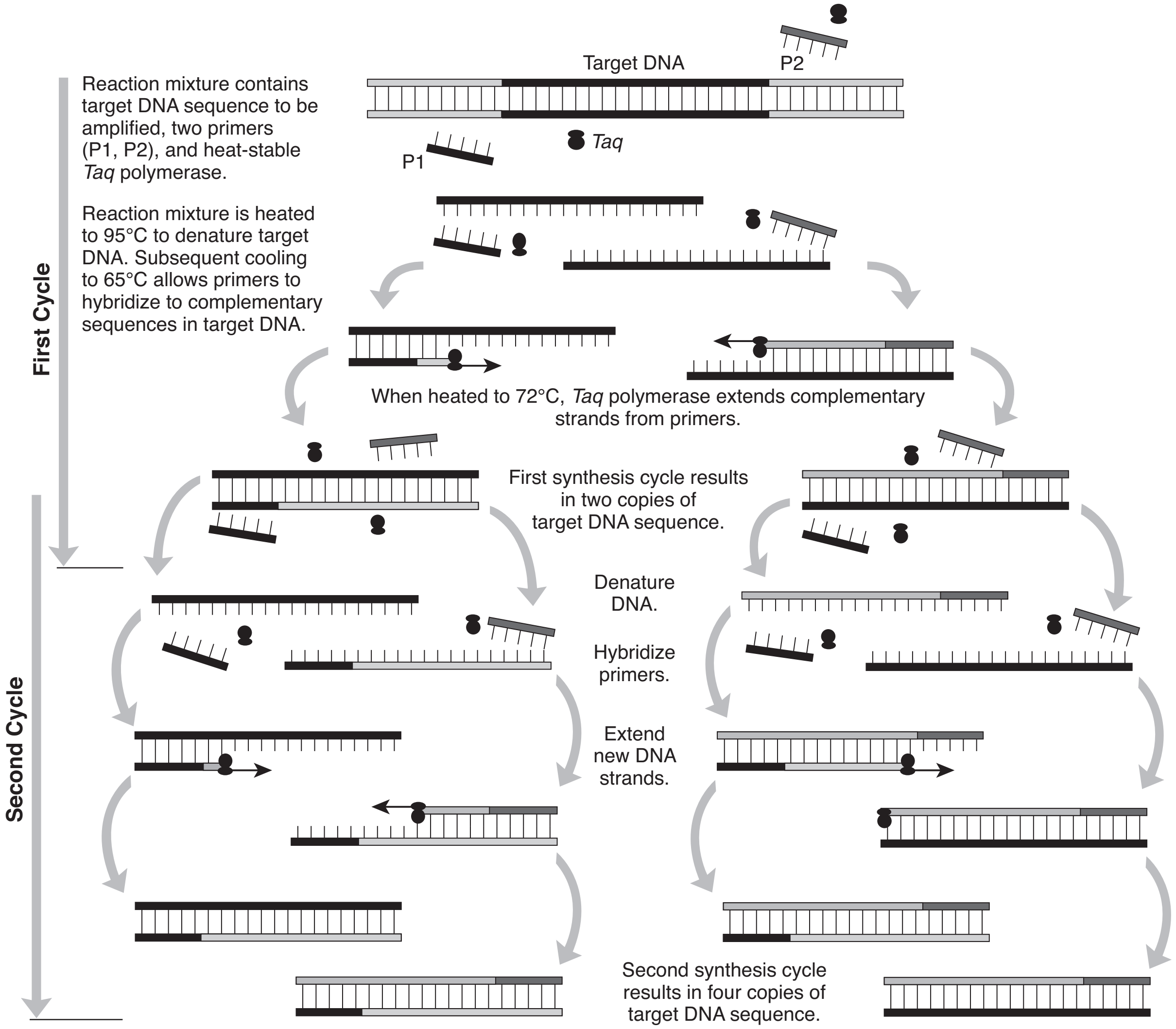


Kary Mullis
(Courtesy of Cold Spring Harbor Laboratory Archives.)

fragment. First, a pair of DNA oligonucleotide (oligo, meaning a few) primers of approximately 20 nucleotides in length are synthesized that bracket the "target" region to be amplified. The primers are designed to anneal to complementary DNA sequences at the 5′ end of each strand of the target region. The two primers are mixed in excess with a DNA sample containing the target sequence, a heat-stable polymerase, the cofactor magnesium ($Mg^{++}$), and the four deoxyribonucleoside triphosphates (dNTPs). *Taq* polymerase from *Thermus aquaticus*, a hot-spring dwelling bacterium, is commonly used.

A thermal cycler then takes the reaction mixture through multiple synthesis cycles, which typically comprise the following three steps:

- *Denaturing.* Heating to near boiling (94°C) denatures the target sequence and creates a set of single-stranded templates. Heating increases the kinetic energy of the DNA molecule to a point at which it is greater than the energy needed to maintain hydrogen bonds between base pairs, and the double-stranded DNA separates into single strands.
- *Annealing.* Cooling to approximately 65°C encourages oligonucleotide primers to anneal to their complementary sequences on the single-stranded templates. The optimum annealing temperature varies according to the proportion of A-T to G-C base pairs in the primer sequence. Because the primers are added in excess and are short, they will anneal to their long target sequences before the two original strands can come back together.
- *Extending.* Heating to 72°C provides the optimum temperature for the DNA polymerase to extend from the oligonucleotide primer. The polymerase synthesizes a second strand complementary to the original template.


Target DNA
P2
P1
Taq
First Cycle
Second Cycle
Reaction mixture contains target DNA sequence to be amplified, two primers (P1, P2), and heat-stable *Taq* polymerase.
Reaction mixture is heated to 95°C to denature target DNA. Subsequent cooling to 65°C allows primers to hybridize to complementary sequences in target DNA.
When heated to 72°C, *Taq* polymerase extends complementary strands from primers.
First synthesis cycle results in two copies of target DNA sequence.
Denature DNA.
Hybridize primers.
Extend new DNA strands.
Second synthesis cycle results in four copies of target DNA sequence.

DNA amplification using polymerase chain reaction.

During each synthesis cycle, the number of copies of the target DNA molecule is doubled. Twenty-five rounds of synthesis theoretically produce 1,000,000-fold amplification of the target sequence in as little as 20 minutes (in a two-temperature profile that eliminates a separate annealing temperature).

Cycle sequencing, like PCR, uses multiple rounds of denaturation, annealing, and extension, but uses only one primer and dye terminators. Using this sequencing technology, the Human Genome Project was initiated in 1988 as an international collaboration to determine the entire nucleotide sequence of the haploid human genome. The project ambled along until, in 1998, it received a psychological and technological challenge from J. Craig Venter, a former NIH researcher who had started a biotechnology company, Human Genome Sciences, and a nonprofit organization, The Institute for Genomic Research (TIGR). Venter announced that he had joined with ABI's Hunkapiller to start a new company, Celera, at which he intended to sequence the human genome in 3 years using a new capillary sequencing technology and Venter's shotgun genome assembly (discussed later).

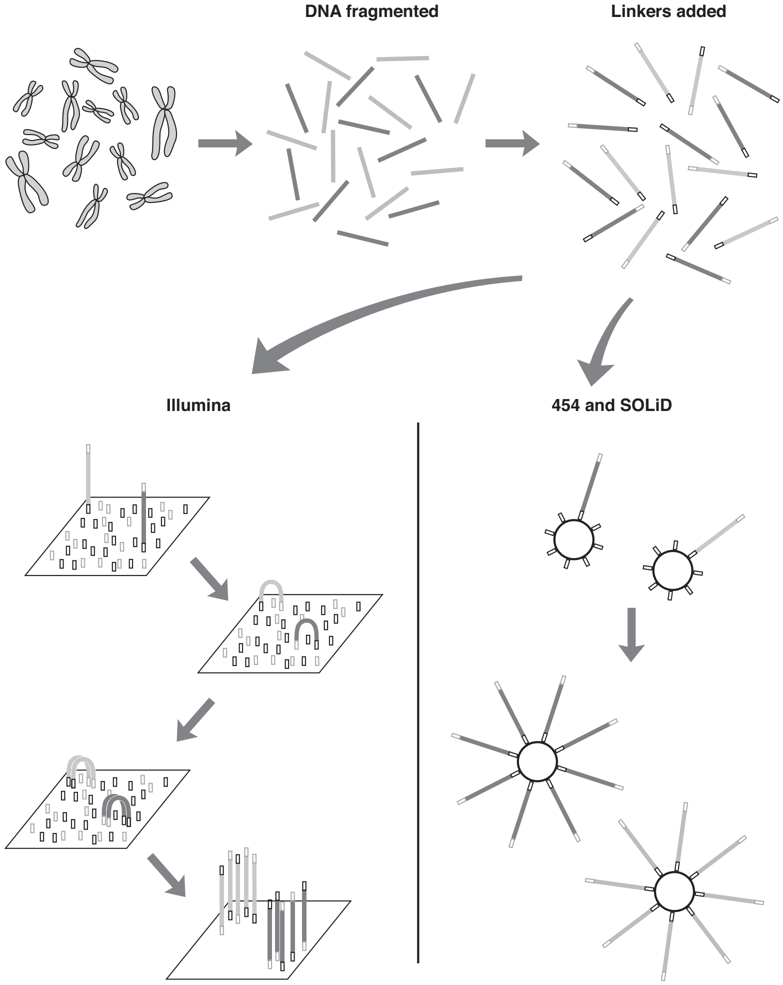
ABI's system replaced each lane of a slab gel with a silica capillary tube, each about the diameter of a human hair and filled with an electrophoresis resin. The capillary reduces heat generated during electrophoresis, allowing higher current and decreasing separation time. A 96-capillary array was linked to a robot mechanism capable of automatically reloading samples from 96-well microtiter plates up to 12 times per day. Throughput was further increased by 384-capillary instruments working from 384-well microtiter plates. This eliminated the time-consuming elements of pouring and loading sequencing gels, reducing human intervention to maintaining reagent levels and loading microtiter plates into the autoloader. Ultimately, Celera and the major centers of the international collaboration became sequencing factories outfitted with 30 or more capillary sequencers, each churning out up to 400,000 nucleotides of sequence per day. This was 400-fold faster than hand-sequencing methods available just prior to the start of the Human Genome Project. Increasing sequence reads to as high as 1000 nucleotides per capillary and limiting human intervention decreased sequencing costs from $1 to $0.01 per nucleotide and increased sequencing accuracy from 99% (one error per 100 nucleotides) to 99.9% (one error per 1000 nucleotides).

## NEXT-GENERATION DNA SEQUENCING

DNA sequencing always begins with fragmenting the genome under study using restriction enzymes, sonication (sound waves), nebulization with liquid nitrogen, or mechanical shearing (e.g., passing the DNA through a syringe). Each fragment must be enriched to provide enough DNA from which to detect each nucleotide in the fragment's sequence. In Sanger sequencing, the DNA enrichment is provided by bacterial cloning. Each fragment is ligated into a plasmid vector, which is, in turn, transformed into a bacterium. The bacterium replicates to create a colony of identical clones, each carrying multiple copies of the genome fragment. DNA is extracted from selected clones and forms the basis for individual sequencing reactions. Because some of these steps must be done by human technicians, the entire workflow cannot be fully automated.

Beginning in 2005, "next-generation" sequencers dramatically shortened the sequencing work flow by determining sequence directly from collections of genomic DNA fragments that are amplified by PCR. By 2012, next-generation sequencing had further decreased costs to only $0.10 per megabase (million base pairs). In most next-generation methods, short adapter molecules are first ligated to each end of the genomic DNA fragments. The adapters anchor the fragments to discrete locations on a substrate (a microbead or  plate surface) and then act as universal PCR primers to amplify each fragment in situ. Like bacterial colonies on a plate, the PCR colonies, or "polonies," are spatially isolated from one another. In most strategies, polonies are generated by emulsion PCR, in which the adapter-linked templates and water-soluble PCR reagents are emulsified in oil. This creates picoliter-sized reaction vessels in which each PCR is isolated in a tiny water droplet surrounded by oil.

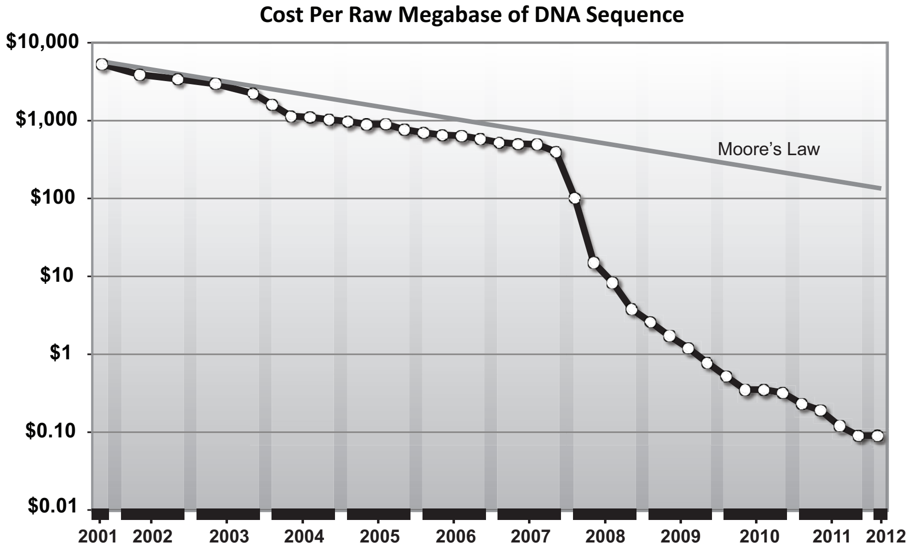
After PCR amplification, the genome fragments are denatured and the single-stranded molecules in each polony serve as templates for sequencing by DNA synthesis. Working from a universal primer within the adapter sequence, sequence is built up by adding nucleotides that are complementary to the genome templates. The reaction



Next-generation sequencing develops PCR colonies (polonies) that are physically separated from one another. Illumina sequencing isolates polonies at discrete locations on solid substrate; 454 and SOLiD sequencing isolate polonies on DNA-capture beads. Each polony, or feature, provides templates for sequencing by DNA synthesis.

vessel in which this occurs is termed a flow cell, because fresh reagents flow past the polony features during each synthesis cycle. The nucleotide added in each synthesis cycle is detected from each feature, building up millions of sequences in parallel. The repeated cycles of DNA synthesis across an array of polonies is referred to as cyclic array sequencing. Collecting sequence information requires a charge-coupled device (CCD) camera, a sensitive imager that can detect fluorescent or visible light “reports” from each tiny polony.

Because each feature can be as small as 1 μm, millions of features can be included in a single array of modest size. This makes it possible to obtain hundreds of millions of nucleotide sequences in a single run on a next-generation sequencer. In fact, next-generation sequencers can generate a genome’s worth of sequence in several sequencing runs. The dramatic labor and reagent savings become apparent when one considers that a single run of a cyclic array sequencer replaces more than a million individual sequencing reactions—and the accompanying bacterial cloning, culturing, and DNA preparation. The microliter volumes of each of millions of Sanger sequenc-



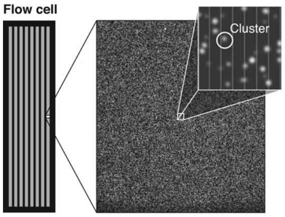
Exponentially declining cost of DNA sequencing. Before the advent of next-generation methods, which develop sequence directly from collections of amplified DNA molecules, the decreasing cost of DNA sequencing generally followed Moore's Law. Named for Intel Corporation founder Gordon Moore, this law predicts that the number of transistors integrated on a computer chip will double approximately every 18 months—with a corresponding decrease in cost. However, since 2008, DNA sequencing costs have decreased 1000 times faster than predicted by Moore's Law.

ing reactions are replaced with a single reaction that effectively uses picoliter or femtoliter amounts for each feature.

454 Life Sciences released the first cyclic array sequencer in 2005. In this method, adapter-terminated genome fragments are hybridized to complementary adapter sequences immobilized on DNA-capture beads, such that each bead contains a single library fragment. After emulsion PCR, the beads with attached polonies are arrayed on a picotiter plate—a microfabricated chip containing wells 28 μm in diameter. Each well accommodates a single DNA-capture bead, along with smaller beads containing immobilized enzymes for pyrosequencing (light sequencing). Bst polymerase from *Bacillus stearothermophilus* generates pyrophosphate with each nucleotide addition. The pyrophosphate then reacts with ATP sulfurylase, luciferase, and luciferin (the light-emitting pigment from fireflies) to produce bioluminescence.

During sequencing, a single nucleotide species is added to the array, and any well in which that nucleotide is added to a template produces a burst of light. A CCD camera records light events in each well (channel) during each synthesis cycle. This is an asynchronous system. During each round of synthesis, different features may or may not incorporate the selected nucleotide, and at any point in time, sequences of different lengths are generated from different polonies. Because there is nothing to terminate a sequence, repeated nucleotides in a sequence (such as A-A-A-A) are added during a single synthesis cycle. Hence, the number of repeats must be inferred by a proportional increase in luminescence, with an A-A-A-A repeat producing a light burst about four times the amplitude of a single A nucleotide. The 454 machine produces the longest reads of any of the next-generation synthesizers (700 nucleotides).

Using this technology, the entire diploid genome (six billion nucleotides) of Nobel Prize–winner and DNA structure discoverer James D. Watson was sequenced in 2007. Hoping to downplay the risks of this sort of detailed genetic knowledge, Watson allowed his entire sequence to be made available online (http://jimwatsonsequence.



Polony features from Illumina sequencer. Next-generation sequencing produces an array of polonies (PCR colonies). Nucleotide additions in each synthesis round produce light emissions at individual polonies, which are recorded by a charge-coupled device (CCD) camera. Each lane of an Illumina flow cell is imaged as ~100 image "tiles," each of which captures activity from as many as 200,000 features. Each dot in this Illumina random array is an individual polony or cluster of identical templates.

cshl.edu/cgi-perl/gbrowse/jwsequence/), except for the *ApoE* gene, which predisposes a person to Alzheimer's disease. Watson's sequence was completed by a handful of 454 scientists in 4 months at a cost of about $1.5 million, compared to the first haploid human genome (three billion nucleotides) completed in 15 years by thousands of scientists worldwide at a cost of about $3 billion.

The Illumina Genome Analyzer is based on technology developed by Gerardo Turcatti and colleagues. In this method, adapter-flanked DNA fragments are amplified by bridge, or cluster, PCR. The adapter sequences tether the library fragments to complementary sequences attached to a solid substrate in a flow cell. During PCR, the template forms a bridge between anchored pairs of adapters (primers) so that every PCR product remains attached to the gel. Successive cycles of PCR can be likened to the motion of a "Slinky," in which the motion always proceeds from one anchored point to another. The bridging limits the distance between each amplicon to the length of the library fragment. Thus, after PCR, the ~1000 amplicons of a single genomic template are clustered in a discrete polony.

As with automated Sanger sequencing, the Illumina system uses dye terminators that halt DNA synthesis upon incorporation into an elongating DNA molecule. Whereas traditional dye sequencing uses fluorescent dyes that are irreversibly attached to dideoxy terminators, Illumina sequencing exploits a removable dye. The fluorescent label replaces the 3′ hydroxyl on each nucleotide, blocking further addition and ensuring that only one nucleotide is added per cycle.

In each synthesis cycle, all four fluorescently labeled nucleotides (A, T, C, and G) are added to the flow cell. This is a synchronous sequencing system, because one of the four nucleotides is added to the same nucleotide position, at the same time, in each polony feature. The incorporated nucleotide is excited by a laser, and a CCD camera records the color emitted from each polony. After imaging, the fluorescent dye is chemically removed, regenerating the 3′ hydroxyl and preparing the template for the next cycle of synthesis and imaging. Because each synthesis cycle is a discrete event, the Genome Analyzer accurately detects each nucleotide in a repeated element (such as A-A-A-A). This technology produces paired-end reads of 100 nucleotides per feature.

The Applied Biosystems SOLiD system exploits synthesis by ligation. Like the 454 system, this system generates polonies by emulsion PCR on DNA-capture beads (1 µm). The SOLiD system uses a degenerate collection of short oligonucleotides (eight or nine nucleotides) in which all possible sequence combinations are represented. Each oligonucleotide is labeled with one of four removable fluorescent dyes, which corresponds to the nucleotide at the fifth position. The dye also functions as a blocker, allowing only one oligonucleotide to be ligated in each synthesis cycle.

During the first round of synthesis, a universal primer is annealed to the adapter sequences on each polony feature. An oligonucleotide with a sequence complementary to the template sequence is then ligated to the end of the universal primer. After ligation, the polonies are imaged in four channels to determine the nucleotide in the fifth position. After imaging, the oligonucleotide is cleaved between the fifth and sixth position, releasing the dye and creating a free end for the next ligation. After each successive round of ligation, sequence information is collected on every fifth nucleotide in the template (positions 5, 10, 15, 20, etc.). The universal primer and ligated sequence are then denatured from the template. A new universal primer that has one less nucleotide at the end compared to the first universal primer is then annealed to the template to generate an offset that will allow a different set of positions to be sequenced with the ligated oligonucleotides, e.g., positions 4, 9, 14, 19, etc. After five rounds of synthesis with different universal primers, each with one fewer nucleotide than the previous primer, a complete sequence is generated for each feature. The SOLiD system produces paired-end reads of 60 nucleotides per feature.

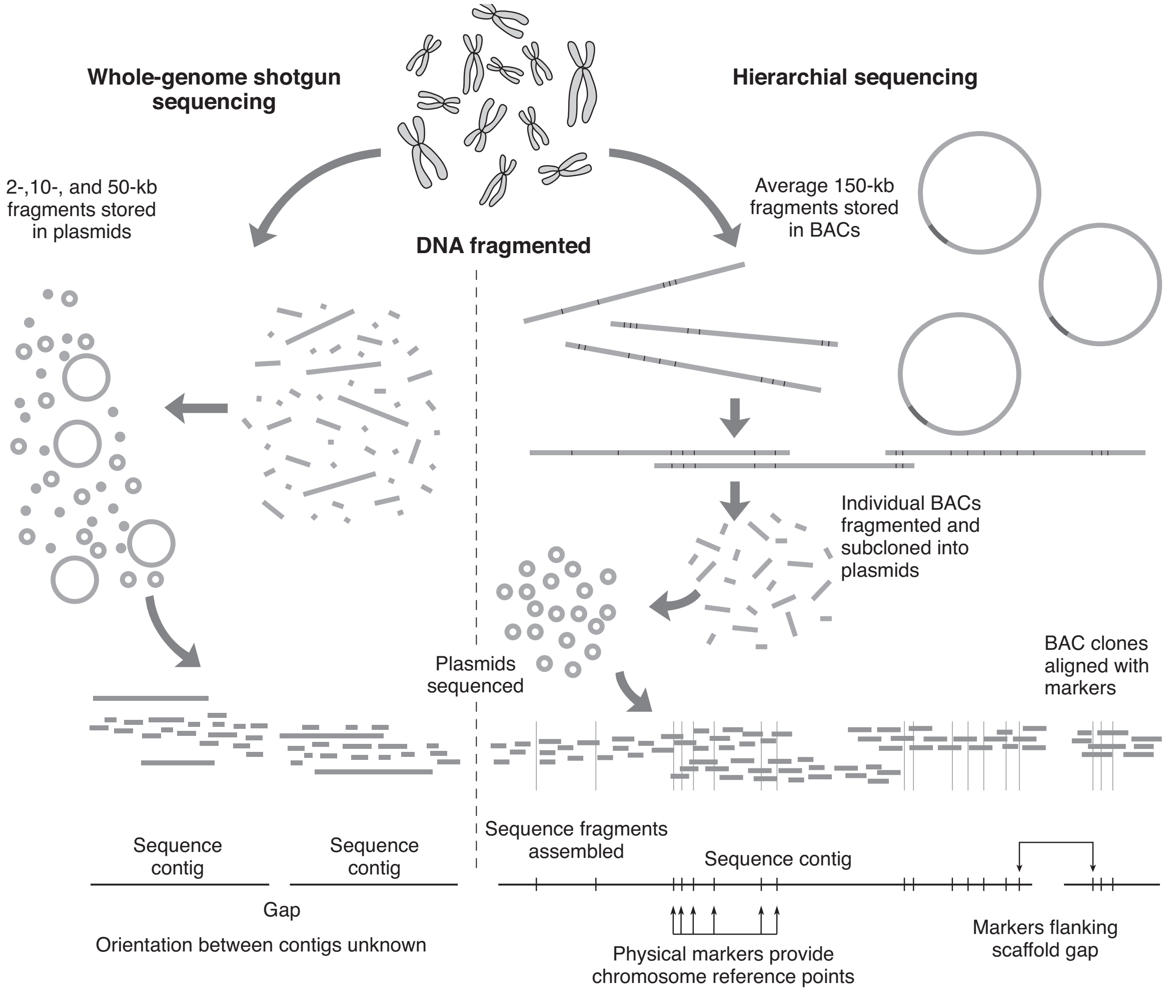
After founding 454 Life Sciences, Jonathan Rothberg developed the Ion Torrent. Released in 2010, this instrument uses an entirely different detection system than other next-generation sequencers. Rather than using an optical sensor to detect light emitted from fluorescent dyes, Ion Torrent uses a semiconductor sensor to detect pH changes during DNA synthesis. The semiconductor chip used in the instrument has up to 660 million microwells, each containing a different single-stranded DNA template from the genome under study. During each synthesis cycle, a new deoxynucleotide triphosphate (A, T, C, or G) is flowed through the chip. The addition of the defined nucleotide to its complementary partner on the template strand by DNA polymerase is accompanied by the release of pyrophosphate (P2) and a hydrogen ion ($H^+$). An ion-sensitive field-effect transistor (ISFET) detects the pH change in the microwell, and an electrical signal from the transistor is directly interpreted by the base-calling software. As with the 454 instrument, Ion Torrent infers the number of nucleotides in a homopolymer tract (such as A-A-A-A) by the relative strength of the signal for that nucleotide addition. Ion Torrent is the fastest of the commercially available sequencers, completing up to 200 bases of sequence per microwell in a 2-hour run.

## GENOME ASSEMBLY STRATEGIES

Whether 75 nucleotides are obtained per polony feature or 800 nucleotides are obtained per capillary, each channel of a sequencing instrument represents only a minute fraction of a genome. After sequencing comes the task of assembling millions of sequenced fragments into large contiguous sequences and, ultimately, into whole chromosomes. Two major strategies are used in assembling information obtained from DNA sequencing: whole-genome shotgun and hierarchical cloning. Whole-genome shotgun is a bottom-up method and is the fastest and most economical means to sequence a genome. The name alludes to fragmenting the genome into tiny bits, as with a close-range shotgun blast, and then sequencing the resulting short fragments. The genome under study is typically fragmented by enzyme cleavage, sonication, or mechanical shearing.

Because next-generation sequencing begins with short fragments generated in this way, it is, in fact, the latest evolutionary step in whole-genome shotgun sequencing methods. However, as originally conceived, shotgun sequencing involves ligating genome fragments into bacterial plasmids and transforming the resultant recombinant molecules into *E. coli* bacteria. Each transformed bacterium is cultured in a separate well of a 384-well plate and grows to produce clones of identical bacteria, all of which carry the same insert of genomic DNA. Thus, each of millions of plasmid clones can be identified by a specific position on a master plate. Clones are randomly selected, and 600–800 nucleotides of sequence are generated from each end in a single sequencing run (the middle part of the cloned fragment is generally not sequenced).

In a typical plasmid library, each nucleotide in the genome under study is represented on eight to ten different cloned fragments; this is called 8x–10x coverage. Next-generation sequencing typically provides at least 20x coverage. This level of redundancy is needed for the assembly step, in which computer algorithms sift through the masses


Whole-genome shotgun sequencing
Hierarchial sequencing
2-,10-, and 50-kb fragments stored in plasmids
DNA fragmented
Average 150-kb fragments stored in BACs
Individual BACs fragmented and subcloned into plasmids
Plasmids sequenced
BAC clones aligned with markers
Sequence contig
Sequence contig
Sequence fragments assembled
Sequence contig
Gap
Orientation between contigs unknown
Physical markers provide chromosome reference points
Markers flanking scaffold gap

Whole-genome shotgun and hierarchical clone sequencing strategies. Hierarchical sequencing has the advantage of using physical markers to anchor assembled sequences (contigs) to exact chromosome locations.

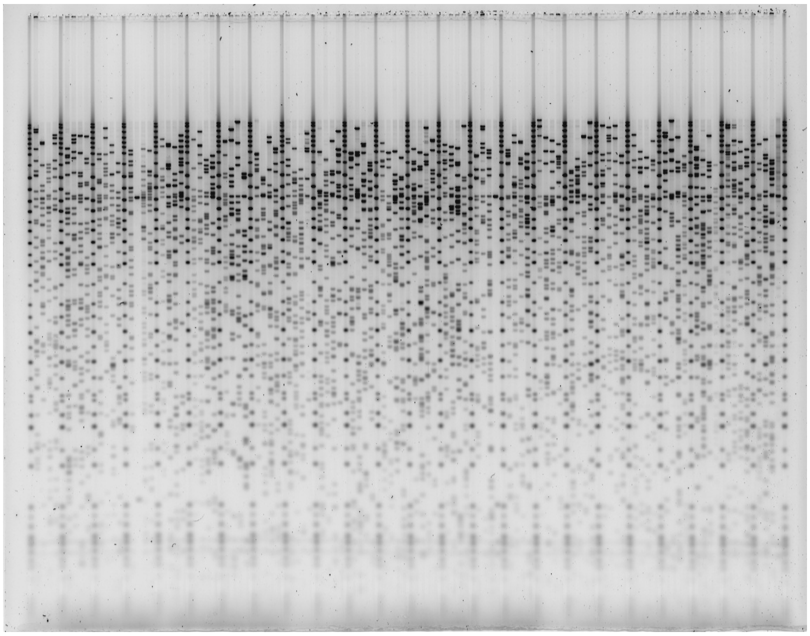
of DNA data to align the short sequences according to shared regions of overlap. This ultimately produces stretches of contiguous sequence, or contigs.

Assembly of a whole-genome shotgun, however, is confounded by any repetitive DNA whose repeat sequence is longer than the average length of sequence reads: 800 nucleotides in capillary sequencing, 700 nucleotides in 454 sequencing, and 60–200 nucleotides for other next-generation sequencing methods. In addition, many genome regions will not be represented in the library used to generate the sequencing plasmids or features. In the absence of a physical map or existing reference assembly, whole-genome shotgun sequencing produces numerous relatively short contigs whose orientations to one another or to precise chromosome locations are virtually impossible to ascertain. In many cases, assembly of a whole-genome shotgun sequence can be guided by a complete assembly of a close evolutionary ancestor, because chromosome regions display a high degree of synteny (conserved gene order) among related species.

Ultimately, variations of a more laborious method—hierarchical cloning—have been used to generate nearly complete reference genomes for a number of key organisms. Hierarchical cloning is a top-down strategy that creates an ordered library of large DNA molecules, maintained as individual bacterial clones and mapped to specific chromosome locations. Bacterial artificial chromosomes (BACs), developed in the mid 1990s by Mel Simon at the California Institute of Technology, have proven to be the most useful cloning vehicles for large DNA fragments. These circular chromosomes, which can contain up to 250 kb of inserted DNA, are extremely stable and amenable to large-scale automation. Plasmids, in contrast, are much smaller and can hold up to only 2–50 kb of inserted DNA.

To construct a BAC library, genomic DNA is randomly cleaved to produce fragments averaging 150,000 nucleotides in length. Partial digestion can be achieved by using a low concentration of a restriction enzyme, sometimes in combination with a methylase that protects a portion of the cutting sites. Alternately, rare-cutting restriction enzymes such as MluI, NruI, and PvuI, which have recognition sequences that occur infrequently in most eukaryotic genomes, can be used. The resulting restriction fragments are ligated into separate BACs, and these recombinant molecules are transformed into *E. coli*. Each of hundreds of thousands of BAC clones can be identified by a specific position on a master 384-well plate. A typical BAC library achieves 8x–10x coverage of the whole genome.

The BAC libraries are analyzed in several ways to identify sets of BAC clones with shared sequences. For each experiment, samples of the BAC clones are transferred from the master library plates onto replica plates for analysis. In this way, the master library is maintained for future reference. BAC fingerprints are generated by digesting BAC clones with several restriction enzymes and electrophoresing the restriction fragments through a gel. BACs that share a common banding pattern, or fingerprint, must share an overlapping region of sequence. BAC-end sequences are generated by running a sequencing reaction on either end of each BAC clone, and matches are determined by sequence alignment. Finally, genetic markers and "overgo probes" repre-



Gel photo of BAC fingerprint clone panel, 2000.
(Courtesy of John McPherson, School of Medicine, Washington University, St. Louis.)

senting known genes are hybridized to the BAC clones. Information from BAC fingerprints, BAC-end sequences, genetic markers, and overgo probes is used to sort BAC clones into "bins" according to their shared sequence information. The genetic markers assign bins to physical locations on chromosomes, whereas the overgo probes provide the relative positions of known genes.

Next, a "tiling path" is selected—an economical set of BACs to sequence that cover the majority of the genome. Each BAC clone along the tiling path is shotgun sequenced. First, each BAC is sheared by forcing the DNA solution through a syringe. This produces fragments several thousand nucleotides in length. The sheared DNA is then separated on a gel, and the fragments are subcloned into plasmids to fill several 384-well plates. "Paired-end reads" are obtained by sequencing 600–800 base pairs of sequence from each end of each subcloned fragment. Finally, a computer aligns overlapping reads to provide the entire sequence of each BAC clone.

The finished BAC clones are then aligned to provide a contiguous sequence, or contig. An assembler program then strings together sequences from local contigs to produce a nearly continuous chromosome sequence. In many cases, unlinked contigs can be oriented with respect to one another using BAC ends, plasmid ends, known mRNAs, and physical/genetic map information to achieve the highest order of large-scale integration called "scaffolds." As the name implies, a scaffold uses map features to affix contigs to chromosome maps, showing the linear relationship of unmerged contigs. Finally, the sequence is improved, or finished, to fill in gaps and extend the contigs. Sequencing primers are used to extend 3′ sequence, whereas 5′ gaps are filled by sequencing PCR products.

## DNA MICROARRAYS

DNA sequencing is complemented by DNA microarrays (or DNA arrays), which provided the first means to analyze large numbers of DNA sequences in parallel. Conceived in the mid 1990s by Pat Brown at Stanford University, DNA arrays based on cDNA libraries were originally developed to analyze the expression patterns of thousands of genes at a time. In his method, different cDNAs were spotted at discrete positions on a glass slide coated with polylysine. The spotting can be done with a set of needle-like pins or even with an inkjet printer! The negatively charged DNA molecules form ionic bonds to the positively charged polylysine substrate, holding them firmly in position in the microarray. The finished microarray thus contains immobilized probes representing thousands of genes from a single cell type or from a single species.

Patrick O. Brown
(Courtesy of Pat O. Brown.)

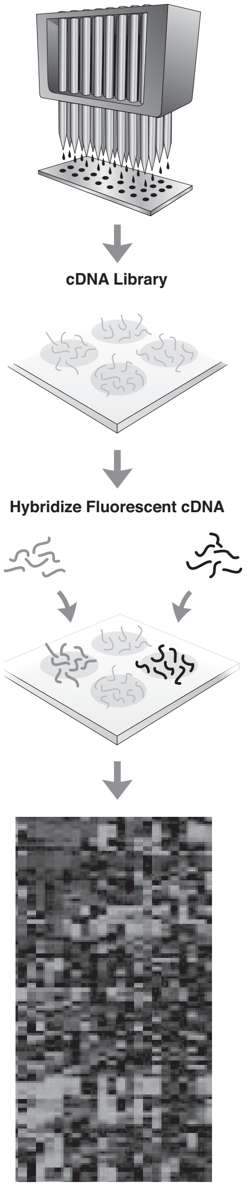
A typical experiment is based on a microarray spotted with cDNAs representing all genes in the genome of an organism. mRNA is isolated from experimental and control cells to be compared, e.g., tumor versus normal cells, mitotic versus quiescent cells, or cells from different tissues. cDNAs made from the mRNA samples from each cell type are then labeled with either a green or red fluorescent dye; the dyes are similar to those used for automated DNA sequencing. The labeled cDNAs are incubated with the microarray, where they hybridize to positions containing complementary sequences. Unbound cDNAs are washed away, and the microarray is imaged under a fluorescence microscope. Red or green signals indicate genes that are differentially expressed in the two populations of cells, and the intensity of the signal indicates the level of expression. In one early experiment, a microarray containing 17,856 cDNAs expressed by the lymph nodes was used to compare diffuse large B-cell lymphomas from different cancer patients. Two different expression patterns, which correlated with different clinical outcomes, were found in tumors that could not be distinguished by microscopic examination.

The company Affymetrix, founded by Steven Fodor, took a different approach to the construction of DNA microarrays. Combining microphotolithography borrowed from computer chip manufacture and combinatorial chemistry from the pharmaceutical industry, Affymetrix patented an industrial method to produce high-quality DNA microarrays. Rather than attaching cDNA probes to the array, oligonucleotides are built anew at individual positions on a quartz wafer using light-directed chemical synthesis. Each wafer may yield 50–400 GeneChips, depending on the number of probes in the microarray.

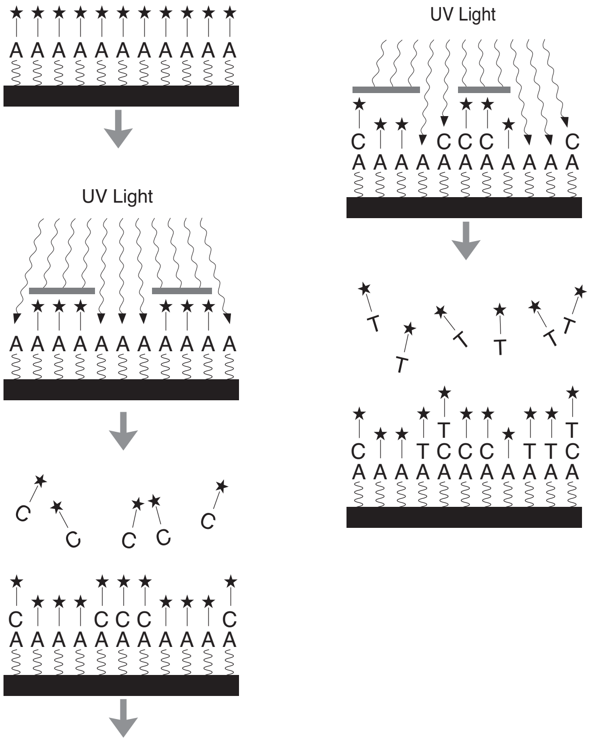
To make a GeneChip, the wafer is first coated with a linker molecule. Each linker molecule is attached to a single nucleotide with a protecting group that blocks polymerization, in the same manner that a dideoxynucleotide terminates a growing nucleotide chain. The protector group is sensitive to light (photolabile) and is released on exposure to ultraviolet (UV) light. A filter mask is placed between the wafer and the UV light source so that only specific positions are exposed to the light and become deprotected. A new nucleotide is then added to the chain at these deprotected positions. A new protecting group is added to these positions at the end of each synthesis step and the process starts over. A computer program controls the process, and a wafer containing a wide variety of oligonucleotides can be built up in 50–100 synthesis steps.

Oligonucleotide arrays, constructed either by photolithography or by attaching synthesized oligonucleotides to glass, have largely replaced spotted cDNA arrays. Because they control for the copy number and size of each gene probe (typically 50–70 nucleotides), synthetic oligonucleotide arrays are easier to calibrate, and they provide more consistent results in high-throughput applications. Synthetic arrays have another major advantage in that they can be based on annotated genes and complete genome sequences available in public databases. For these reasons, synthetic DNA arrays have supplanted printed arrays for a range of analyses including the following:

- Whole-genome arrays containing probes developed from every annotated gene in a genome sequence are used to study differential gene expression.
- So-called SNPchips containing hundreds of thousands of single-nucleotide polymorphisms (SNPs) are used for generating personal genetic screens or for large-scale population studies. The same is true for copy-number variations (CNVs).
- With multiple probes for each exon of a gene, exon arrays support detailed studies of gene transcription and alternative splicing.
- Tiling arrays investigate 25-nucleotide "tiles" spaced at 10-nucleotide intervals to scan a genome for novel mRNA transcripts or transcription-factor-binding sites.
- High-density resequencing arrays that interrogate each position of an entire genome are used to sequence annotated genomes from multiple individuals to provide information about population variation.

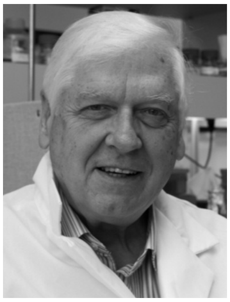


Typical DNA array experiment. A cDNA library is spotted onto a polylysine-coated slide. Differentially labeled cDNA probes are added from two different cell types or experimental treatments, one labeled with a red fluorescent dye (black) and the other labeled with a green fluorescent dye (gray). The probes hybridize to corresponding cDNAs, showing which genes are active under each of the two conditions.



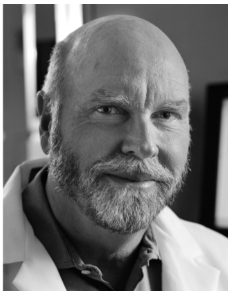
Making a GeneChip. Adenine molecules with photolabile protecting groups (stars) are attached to a quartz wafer. A blocking mask (gray) allows UV light to expose specific regions of the wafer, which removes the protecting group from a population of adenines. A second nucleotide (C) is added to unprotected adenines. Another mask is added, deprotecting a different set of nucleotides, and a new nucleotide (T) is added. The process is repeated to build up olignucleotides with defined sequences at known positions on the wafer.

## MINIMAL AND SYNTHETIC GENOMES



Hamilton Smith

(Courtesy of the J. Craig Venter Institute.)



J. Craig Venter

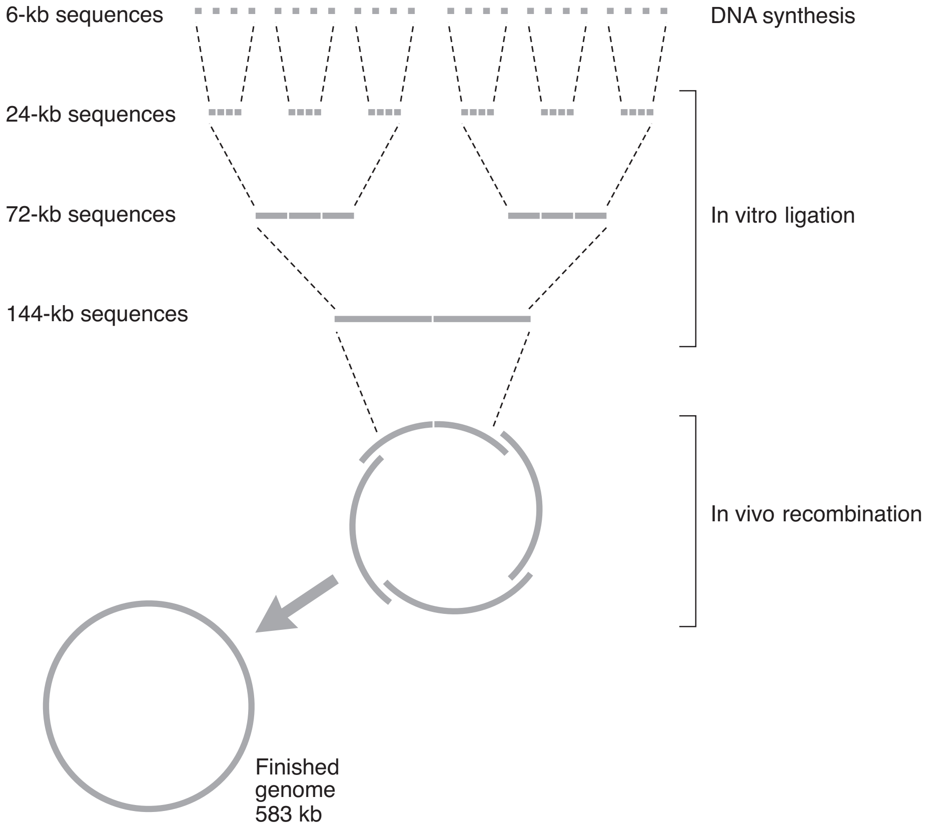
(Courtesy of the J. Craig Venter Institute.)

This consideration of genome structure, function, and sequencing leaves us with several interesting questions. What is the smallest genome—the minimal set of genes—required for life? If a genome is merely biological information encoded in the sequence of a DNA molecule, can scientists synthesize a minimal genome to recreate life in vitro (meaning literally "in glass," but, practically, in a test tube)?

Recently, these questions have been most vigorously pursued by a research team headed by Hamilton Smith and Craig Venter at the J. Craig Venter Institute in Rockville, Maryland. (Smith shared the 1978 Nobel Prize for the discovery of restriction enzymes.) Har Gobind Khorana and coworkers at the National Institutes of Health chemically synthesized the first gene in 1978: the 207-bp gene for tyrosine suppressor tRNA. The advent of automated DNA synthesis in the 1980s, and later the Internet, made it possible for researchers to order small DNA molecules (notably PCR and sequencing primers) online for overnight delivery.

In 2004, Smith and Venter took DNA synthesis into a new realm when they manufactured a biologically active viral genome from scratch. Working from the published sequence of the bacterial virus ϕX174, they designed single-stranded oligonucleotides (42 mers) such that the ends of complementary strand sequences were staggered. When annealed, these formed "sticky" ends that provided templates to align adjacent sequences during a subsequent ligation reaction. Double-stranded ligation products averaging 700 nucleotides were then joined by polymerase cycling assembly that incorporated overlapping templates on each cycle to produce full-length genomes of 5386 bp. The extended products were then electroporated into *E. coli*, where the synthetic molecules directed the production of encapsulated, fully infective viruses. The bacterium *Mycoplasma genitalium* has the smallest genome yet found for any free-living organism. Although it carries 485 protein-coding genes, approximately 100 of these are nonessential when disrupted individually. To determine the smallest gene set that is simultaneously required for life, Smith and Venter proposed to synthesize various reduced genomes and test them inside bacterial cells. As a step toward this goal, in 2008 the team synthesized a correct copy of the 582,970-bp *M. genitalium* genome. Synthetic DNA sequences of 5–7 kb went through several rounds of in vitro ligation to produce overlapping constructs equaling about one-fourth of the *Mycoplasma* genome. These fragments were then joined—by homologous recombination within yeast cells—to create a synthetic *Mycobacterium* chromosome that was stably propagated in yeast.

In 2010, the team produced the first living cell under the control of a synthetic genome. For this experiment, they assembled the entire 1.08 million–bp genome of the fast-growing *Mycoplasma mycoides*. Synthetic 1-kb cassettes went through three rounds of recombination in yeast to produce 10- and 100-kb assemblies—and finally a complete genome, which included several "watermark sequences" that can be decoded into quotes from English literature! Next, they inserted the synthetic genome into a recipient cell of the related species *Mycoplasma capricolom*. These "synthetic cells" reproduced normally and had the expected phenotype of *M. mycoides*. PCR of DNA from synthetic cells identified the watermark sequences, and whole-genome sequencing



Synthesizing the *Mycoplasma genitalium* genome.

detected several mutations that occurred during assembly but none from the recipient *M. capricolom*.

No vital force is required to "reboot" a living cell from the information inherent in a synthetic DNA molecule. Although organisms have evolved chromosomes to perpetuate their genetic legacy, this information can simply be stored in and retrieved from a digital file. This puts a point on the notion that DNA is information.