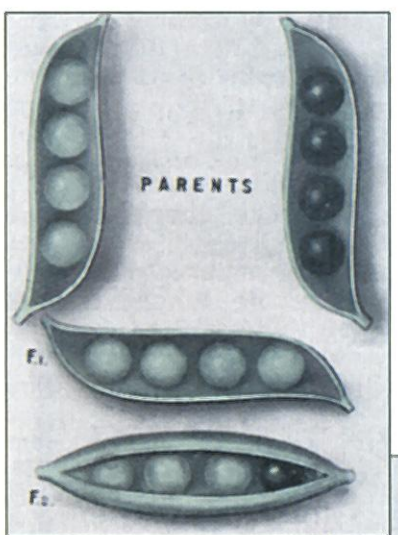
Genomics: Journey to the Center of Biology

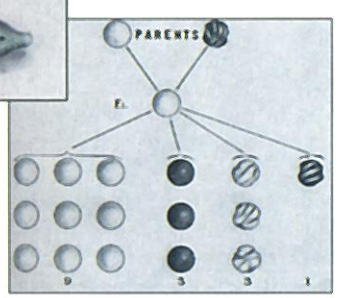
Eric S. Lander and Robert A. Weinberg

Without doubt, the greatest achievement in biology over the past millennium has been the elucidation of the mechanism of heredity. Heredity is surely the strangest of physiological processes: Organisms encapsulate instructions for creating a member of their species in their gametes, these instructions are passed on to a fertilized egg, and then they unfold spontaneously to give rise to offspring. The ancient Greeks puzzled over these remarkable phenomena. Hippocrates imagined that instructional particles were gathered together from throughout the adult body, having been shaped by experience, while Aristotle believed that the instructions were constant and inherent in the gametes. But philosophers could do no more than speculate for the ensuing 2000 years, because there was no way to probe the physical nature of these instructions.

How the nature of heredity came to be understood over the past 200 years is an extraordinary tale of scientific progress. In dizzying succession, biologists found that the heredity instructions followed specific rules of transmission, resided in the chromosomes contained in the nucleus, were embodied in the molecule DNA, were written in a precise genetic code, and could be read out in their entirety to specify organismic shape and function.

The solution to the problem of heredity turned out to have breathtaking elegance and generality. The instructions for assembling every organism on the planet—slugs and sequoias, peacocks and parasites, whales and wasps—are all specified in DNA sequences that can be translated into digital information and stored in a computer for analysis. As a consequence of this revolution, biology in the 21st century is rapidly becoming an information science. Hypotheses will arise as often in silico as in vitro. In this essay, we recount how this came to pass.





Pea wisdom. Early portraits of heredity laws: above, segregation of traits (second generation expresses dominant trait; recessive traits reappear in third) and at right, independent assortment of traits (in this case, seed texture and color).

BORDER ILLUSTRATIONS: MATT ZANG; CREDITS: T. H. MORGAN, *A CRITIQUE OF THE THEORY OF EVOLUTION* (PRINCETON UNIVERSITY PRESS, PRINCETON, NEW JERSEY, 1916)

Mendel's Revolution: Transmitting the Instructions

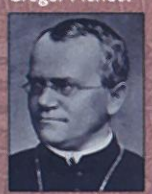
Heredity was the province of philosophers until Anton van Leeuwenhoek's invention of the simple microscope in the 17th century. Ironically, early microscopic studies diverted the field; observers convinced themselves that they could see tiny, preformed homunculi ensconced within individual spermatozoa. Preformation obviated the need to store and transmit instructions, but it raised perplexing philosophical questions, such as whether the entire human lineage resided, like nested Russian dolls, in Adam's sperm, and what role Eve played.

Scientific studies of heredity eventually emerged from a more practical quarter—economic forces driving improvements in agriculture. The Age of Discovery from the early 15th to the late 18th century brought thousands of new plant species to Europe, many of which were propagated and hybridized, improved cultivars being highly prized. The rapidly expanding international trade held great promise for increased economic returns on agricultural products.

One of the hotbeds of interest in agricultural improvement, including animal husbandry, was the city of Brünn (now Brno) in Moravia, center of the textile industry in the Austro-Hungarian Empire of the early 19th century. The high price commanded by imported Spanish wool spurred great interest in improving sheep breeds. Breeding programs, however, were conducted largely by trial and error with no

A Genomics Timeline

1865
Gregor Mendel



reports the results of his pea plant experiments, from which he discerned several fundamental laws of heredity. His results, which appeared in an obscure journal article in 1866, were ignored for 34 years.

1882
Walther Flemming publishes his observations of tiny threads—later known as chromosomes—inside salamander larvae cells that appear to be dividing.

1900
Hugo de Vries in the Netherlands, Erich Tschermak von Seysenegg in Austria, and Karl Correns in Germany simultaneously rediscover and verify Mendel's principles of heredity.

1902
Walter Sutton points out connection between chromosomes and Mendel's "factors," thereby expanding the science of genetics from the organismal level to the subcellular level.

1910
Thomas Hunt Morgan and coworkers in the "fly lab" show that some genetically determined traits are sex linked. They also confirm that some trait-determining genes are located on specific chromosomes.

underlying rationale. Possessed with remarkable foresight, Brünn's civic leaders organized societies to promote scientific research, citing the importance of discoveries such as those of Copernicus and Newton and expressing hope that the world would someday be similarly indebted to a son of Brünn.

This extravagant hope was indeed to be fulfilled. C. F. Napp, head of the Pomological and Oenological Society of Brünn and abbot of the Augustinian monastery, kept his eye out for scientifically trained young men to join his remarkable monastery. The best of these recruits was Gregor Mendel, who had studied physics in Vienna before joining the abbey. The revolution in the understanding of heredity that followed was not triggered by a monk working in isolation who accidentally stumbled upon the laws of genetics. Rather, Mendel worked in an incubator focused on promoting scientific progress in what today would be called agricultural biotechnology.

Mendel's pea breeding allowed him to observe genetic dominance and the segregation of traits. In fact, these phenomena had been described qualitatively decades earlier. But Mendel took a quantitative approach, using his physics training and his breeding data to formulate a theory providing, for the first time, a mechanistic description of the laws of heredity.

Mendel proposed that heredity information was passed from parent to offspring in discrete packets, which he called "factors." Different factors were responsible for distinct aspects of a pea plant's appearance, such as seed shape or flower color. His key insight was that the factors occurred in pairs, with one member of the pair being passed on from each parent. The two factors governing a trait might carry conflicting instructions, in which case the voice of one might dominate in determining the appearance of the individual. Nonetheless, the other factor would persist in latent form, and its effects could reappear in later generations in predictable ratios.

Mendel's 1865 report *(1)* in the *Journal of the Brünn Society of Natural Science* fell on deaf ears. He worked at the periphery of the scientific community, and he published in an obscure journal. But the real problem was that Mendel's formalisms were mathematical and his factors were abstractions. Mendel's laws would only gain a wide audience long after his death, when they could be related to biological realities—visible cellular structures.

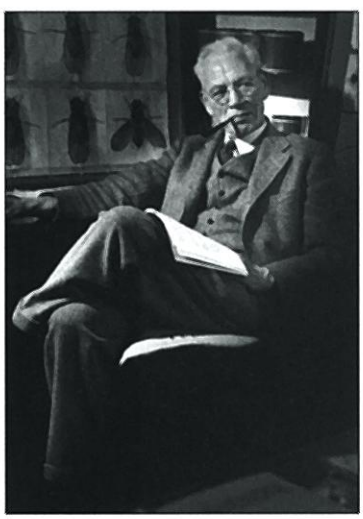
Chromosomes: The Cellular Basis of the Instructions

By the mid-1880s, biologists began to recognize that the physical seat of heredity must lie in the cell's nucleus. Microscopists found that recently fertilized eggs carried two equally sized "pronuclei," which later fused. These pronuclei derived from the sperm and the unfertilized egg, which seemed to contribute equally to heredity. Moreover, close examination of a spermatozoon indicated that this cell was little more than a nucleus with a tail.

The most obvious components of the nucleus were its chromosomes, whose behavior could now be studied with precision through greatly improved staining and microscopy techniques. Like the entities that harbored heredity instructions, chromosomes appeared to duplicate with each cycle of cell growth and division. Still, researchers remained unsure about the relation between chromosomes and heredity. Some theories, for example, held that each chromosome carried a complete set of the heredity instructions. The ensuing ferment revived interest in understanding the laws of heredity via breeding experiments.

In the early months of 1900, three researchers—Hugo de Vries, Erich Tschermak von Seysenegg, and Karl Correns—independently reported rediscovering Mendel's work and laws *(2)*. Their work revealed little more than what Mendel had found 35 years earlier, but the scientific community was now primed to listen. The papers sparked the genetics revolution that continued unabated throughout the 20th century.

An initial challenge was to prove the connection between genes and chromosomes. The most important advances would come from the study of the fruit fly *Drosophila* in the laboratory of Thomas Hunt Morgan at Columbia University. Arguably, the greatest insights came from Alfred Sturtevant, who was performing undergraduate research in Morgan's lab. Sturtevant analyzed a large body of experimental results describing the frequency with which pairs of genes were cotransmitted when passed from parent to offspring. He realized that these data could be explained by a simple model in which the genes were arrayed along a linear "linkage map," with nearby genes being cotransmitted more often than gene pairs located far from one another along his maps. Sturtevant realized that these maps showing the positions of genes must correspond to the thread-like chromosomes *(3)*. Gene mapping rapidly became a powerful tool of genetics, although the definitive proof of the connection between linkage maps and chromosomes came later in the 1930s with studies by Barbara McClintock on maize chromosomes *(4)*.



Father of gene mapping. Alfred Sturtevant realized early last century that genes were linked in linear arrays.

DNA: The Biochemical Basis of the Instructions

The early 20th century witnessed the birth of another experimental science: biochemistry. This marriage of biology and chemistry sought to understand life by isolating molecules and reconstituting living processes in the nonliving extracts prepared from cells. The biochemists had a clear agenda: to systematically dismantle the notion of vitalism, which held that ineffable "life forces" were responsible for the complex attributes of living cells and tissues. By 1925, they had triumphantly shown that many biochemical reactions could be reproduced in the test tube using the organic catalysts called enzymes. The science of genetics, however, was disconnected from this forward rush of biochemical progress. Genes seemed hopelessly inaccessible: How could one possibly purify heredity in a test tube? Indeed, could heredity ever be understood through biochemistry and the increasing number of molecular species being unearthed in living cells?

A first, tentative foray toward the molecular embodiment of genes came from the 1927 work of Hermann Muller, then in Texas, who demonstrated that x-rays could mutate the genes of *Drosophila (5)*. This provided geneticists with a powerful tool. They no longer needed to rely on the spontaneous randomness of nature to generate variants of the "wild-

CREDITS: TOP TO BOTTOM) STOCK MONTAGE; INSTITUTE ARCHIVES/CALTECH; KATHARINE SUTLIFF

type" genes normally found in flies. Conceptually, Muller's discovery was even more far-reaching, showing that genes were physical entities susceptible to being damaged just like other molecules in the cell. But still, the central question remained: What kind of molecules explained heredity?

A year later, some steps were taken toward an answer. Fred Griffith in England made the serendipitous observation that an extract prepared from virulent, disease-causing *Pneumococcus* bacteria could transmit the trait of virulence to a benign strain. Once the benign bacteria had acquired these instructions, they and their descendants in turn showed all the traits of virulence. The instructions for inducing disease, whatever their nature, persisted in the virulent bacteria long after their death by heat treatment *(6)*.

By the mid-1930s, Oswald Avery, Colin McLeod, and Maclyn McCarty, working at the Rockefeller Institute in New York, took on the daunting task of purifying the elusive substance that conferred virulence. By 1944 they had the answer. Deoxyribonucleic acid (DNA) molecules extracted from virulent bacteria sufficed to transfer the genetic instructions for virulence. Destruction of the DNA resulted in loss of these instructions, while destruction of bacterial proteins seemed to have no effect on the information transfer *(7)*.

Their conclusion was controversial. DNA molecules were widely regarded as boring, monotonous chains composed of four nucleotides—ostensibly structural scaffolds of the chromosomes. Protein molecules were far more interesting. They were biochemically and structurally more complex, and for this reason seemed to offer far more possibilities for harboring genetic information. But DNA survived the skeptics. When purified of all but 0.02% of contaminating protein, DNA continued to be potent in transmitting genetic information. Most compelling were the 1952 experiments of Alfred Hershey and Martha Chase, who showed that when bacterial viruses inject their genetic information into host cells, DNA enters the cell, while the protein coat remains on the outside *(8)*.

Still missing was an understanding of how DNA—or any molecule—could store and encode heredity instructions. This intellectual puzzle had attracted the interest of some eclectic physicists, including Niels Bohr and his student Max Delbrück. They struggled to explain the long-term stability of genes in terms of molecules residing in deep potential wells and even suggested that new laws of physics might be needed to explain life. These issues were distilled by Erwin Schrödinger in a brilliant and popular 1945 book, *What is Life? (9)*.

Schrödinger proposed that genes must be "aperiodic crystals" consisting of a succession of a small number of isomeric elements whose precise sequence constitutes the heredity code in the manner of the Morse code. Although these ideas did nothing to identify the responsible molecular structures, they did attract many newcomers to the field—including a youthful James Watson, who set off to Cambridge, England, determined to work on the nature of the gene. There, he teamed up with former physicist Francis Crick.

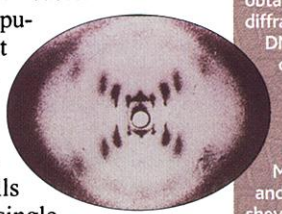
Watson and Crick's revelation of the double-helical structure of DNA struck like a thunderbolt in April 1953 *(10)*. Just as Schrödinger had predicted, DNA was an aperiodic crystal, being composed of four nucleotide bases along its strands. The double helix, by pairing A with T and G with C on opposite strands, explained how genetic information could be copied (the complementarity of nucleotides meant that each strand could serve as a template for the assembly of a complete double helix) and how mutations could arise (occasionally the copying process might go awry). In one stroke, Watson and Crick had explained the key problems of genetics.

The Genetic Code Through the Recombinant DNA Revolution: Deciphering the Instructions

The Watson-Crick model made clear that the instructions must be encoded by the sequence of the bases in the strands of the DNA double helix. But how, specifically, were these instructions read out to build the components of a living organism? By 1964, the outlines of the solution had been worked out. The DNA segment corresponding to each gene is copied into a messenger RNA molecule, whose base sequence is then used to direct the synthesis of a specific protein from amino acid building blocks. Marshall Nirenberg used synthetic RNAs to crack the genetic code by which triplets of bases (nucleotides) constitute genetic "words" specifying particular amino acids *(11)*. In principle, the secret of life had been laid bare.

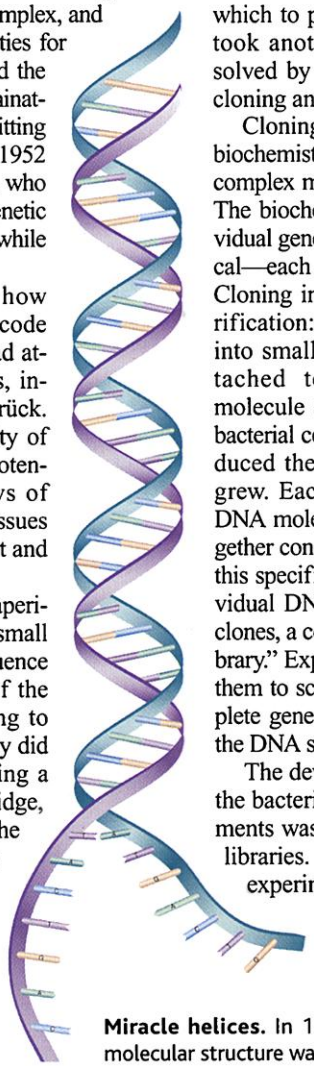
In practice, there was a catch. Although biologists had deciphered the code for translating DNA information into proteins, they could not yet read any natural DNA sequences—not even the sequence of a single gene out of the thousands present within a cell. They lacked the text on which to practice their newfound deciphering skills. It took another 15 years for this problem to be fully solved by the two recombinant DNA technologies of cloning and sequencing.

Cloning circumvented the limitations of traditional biochemistry, which relies on isolating molecules from a complex mixture based on their chemical idiosyncrasies. The biochemical approach is useless for purifying individual genes because chemically, they are virtually identical—each is simply a stretch of DNA bases. Cloning introduced a new twist on purification: Large genomes were cut into small segments; each was attached to a special "vector" molecule and then introduced into bacterial cells, which faithfully reproduced the foreign DNA as the cells grew. Each bacterium received a single DNA molecule, ensuring that descendant cells would together constitute a "clone," all harboring exact replicas of this specific DNA segment. Scientists thus purified individual DNA segments by propagating them in distinct clones, a collection of which came to be called a "gene library." Experimenters then devised clever steps to enable them to screen the millions of separate clones in a complete gene library and pick out the rare clones carrying the DNA segment and thus gene of interest.



The development of the vectors capable of directing the bacterial cell to reproduce the individual DNA segments was a key technical step in the creation of these libraries. Here, biologists exploited highly successful experiments of nature such as viruses and plasmids, which are cellular parasites known to co-opt cells into making hundreds or thousands of copies of viral and plasmid DNA molecules *(12)*.

DNA sequencing technology formed



Miracle helices. In 1953, DNA's now iconic molecular structure was first revealed.

CREDITS: (LEFT TO RIGHT) CARIN CAIN; PHOTO RESEARCHERS

1927
Working with fruit flies, Hermann Muller determines that x-rays can cause genetic mutations.

1928
Fred Griffith discovers the phenomenon of transformation, in which some unknown "principle" transforms a harmless strain of bacteria into a virulent one.

1944
Oswald Avery, Colin MacLeod, and Maclyn McCarty prove that DNA, not protein, embodies the heredity material in most living organisms.

Late 1940s
Erwin Chargaff discovers one-to-one correspondence between adenine and thymine and between cytosine and guanine—a key piece of information for determining the structure of DNA.

1952
Rosalind Franklin obtains x-ray diffraction data of DNA, which become central to the elucidation of DNA's molecular structure. Martha Chase and Alfred Hershey report experiments with bacteriophages that help prove DNA is the molecule of heredity.

1953
James Watson and Francis Crick announce their discovery of the double-helix structure of DNA. They write in a 958-word *Nature* article: "It has not escaped our notice that the specific pairings we have postulated immediately suggest a possible copying mechanism for the genetic material."

Mid-1960s
Marshall Nirenberg, H. Gobind Khorana, and others crack the triplet code that maps messenger RNA codons to specific amino acids.

1969
A team at Harvard Medical School led by Jonathan Beckwith isolates the first gene, specifically, a bacterial gene whose protein product is involved in sugar metabolism.

1970
A team at the University of Wisconsin, led by H. Gobind Khorana, synthesizes a gene from scratch, beginning what might be called chemical genetics.

1972
Using restriction enzymes from Herbert Boyer's research group, Paul Berg and colleagues produce the first recombinant DNA molecules.

1973
The era of genetic engineering begins when Stanley Cohen, Herbert Boyer, and co-workers insert a gene from an African clawed toad into bacterial DNA.

1976
Genentech, the first genetic engineering company, is founded in South San Francisco.

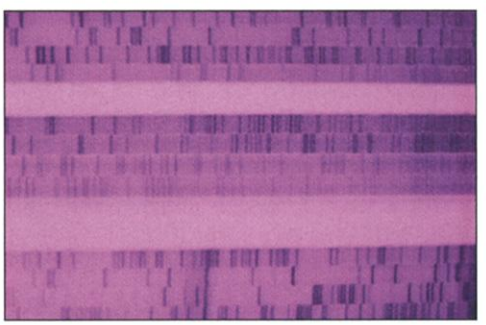
1983
James Gusella and co-workers locate a genetic marker for Huntington's disease on chromosome 4. This leads to scientists having the ability to screen people for a disease without being able to cure it.

the other half of the recombinant DNA revolution of the 1970s. Two strategies—one pioneered by Fred Sanger, the other by Walter Gilbert—made it possible to determine with relatively high accuracy the sequences of DNA fragments a few hundred bases long *(13)*. Soon, individual genes cloned from large cellular genomes became objects of study.

Sequencing technology advanced rapidly, driven by an unquenchable thirst for sequence information. In the late 1970s, an entire doctoral thesis might be devoted to reporting the sequence of a gene of several thousand DNA bases. By century's end, technologists had developed automated sequencing machines capable of cranking out up to a half-million bases per day.

The Genomics Revolution

DNA sequencing soon produced surprises by revealing connections between genes that previously had seemed unrelated. Two early examples involved cancer-causing genes: the oncogenes *sis* and *erbB*. One research team cloned these genes and determined their DNA sequences. Meanwhile, unrelated groups of researchers who were more biochemically inclined isolated two proteins—platelet-derived growth factor (PDGF) and the receptor for epidermal growth factor (EGF)—and determined the amino acid sequences of both. To everyone's surprise, the DNA sequences of the oncogenes corresponded nearly perfectly to the amino acid sequences of the well-studied growth-controlling proteins *(14)*. These identifications revealed instantaneously how the *sis* and *erbB* oncogenes transform normal cells into cancer cells.



Life's bar code. Raw genetic sequence data, like this autoradiograph, began changing biological research in the 1970s.

Such connections were only the beginning. Comparisons of gene sequences revealed that strikingly similar proteins were encoded in the genomes of organisms as distantly related as yeast and mammals. For example, the proteins governing the progression of a yeast cell through its cycle of growth and division were found in very similar form in the cells of humans. These cross connections soon numbered in the thousands, then tens of thousands. It became clear that the evolution of life on this planet was stunningly conservative. Once nucleated cells evolved more than 1.5 billion years ago, the great majority of the proteins invented at the time were perpetuated in myriad descendant cells—sometimes with only minor changes. Often the genes encoding these early proteins multiplied and diversified a billion years later, spawning large families of related genes and proteins having diverse, sometimes totally novel functions.

Recognition of gene families produced enormous synergy in research, as the function of one member could often be deduced from that of its known relatives. When the gene responsible for cystic fibrosis was cloned, sequence analysis immediately suggested that it belonged to a family of proteins that transport ions through membranes—a fact that was then readily tested and confirmed in the laboratory *(15)*.

The connections across vast phylogenetic distances also drove biologists to reconceptualize their research. Those researching organisms such as worms, flies, and yeast began portraying their work as opening windows on the universal rules of life on this planet, not just on the idiosyncrasies of the arcane organisms they studied. Researchers working on sea urchin and frog development found themselves catapulted into cancer research meetings, using a common vocabulary with cancer researchers to describe proteins that play equally important roles in early embryogenesis and in the development of human malignancies.

Sequence analysis also revolutionized the study of evolution, by making it possible to draw phylogenetic trees relating organisms on the basis of similarities in their genes rather than shared physical characteristics. By the 1980s, the availability of vast quantities of sequence data and sophisticated computer-based analytic tools led to a wholesale redrawing of the branches and twigs of the tree of life.

The studies of individual genes represented stunning achievements, but these successes soon promoted an even grander vision: the systematic study of complete genomes, soon referred to as genomics. The first foray into genomics was a proposal to use DNA technology to extend Sturtevant's original concept of genetic mapping to the human being. Instead of tracing the inheritance of visible mutations as had been done in fruit flies, David Botstein and colleagues proposed in 1980 that one could construct a complete genetic map of the human chromosomes by following the inheritance of common DNA sequence variations, termed DNA polymorphisms *(16)*. Each polymorphism could be used to plant a sequence marker at a specific site on a genetic map of a chromosome. One could then localize genes causing specific human diseases by matching their inheritance patterns with those of the signposts on these genetic maps.

The first success using this strategy came in 1983, when the gene causing Huntington's disease was shown to map to the tip of the short arm of human chromosome 4 *(17)*. The first comprehensive human genetic map with 400 signposts was constructed by 1987, and much denser maps with more than 10,000 such markers were available a decade later. Medical genetics was revolutionized as the genes causing more than 1000 human diseases were soon mapped to specific chromosomal sites.

An even more expansive vision was expounded in 1985: The entire human genome would be sequenced, providing a complete catalog of every human gene. On its face, the proposal seemed quixotic, a logistic impossibility. The human genome encompasses 3 billion bases of DNA; then-current sequencing technology could only read out lengths of about 300 bases in each analysis. Decades of work by vast hordes of technicians would surely be required to complete the task.

Moreover, some argued that sequencing the human genome was a fool's errand because the vast majority of it—perhaps 95%—does not encode proteins or regulatory information. These sequences were derogatorily labeled "junk DNA." Why, some asked, expend enormous effort to acquire detailed sequence information about DNA that had slim prospect of ever yielding insight into biological function?

But the proposal prevailed. Several years of debate restructured the initial plan into a series of staged subprojects. The

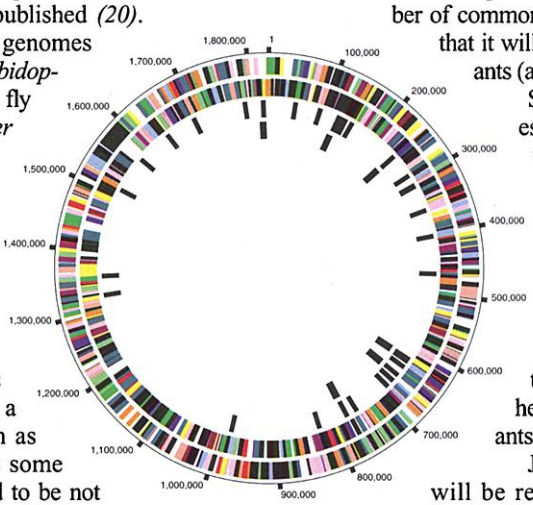
CREDITS: (TOP TO BOTTOM) VISUALS UNLIMITED; GENENTECH

relatively small genomes of important experimental organisms—bacteria, yeast, flies, and worms—would be attacked first, before turning to that of the human. Biologically interesting in their own right, these genomes would serve as pilot projects designed to refine the tools for automated sequencing and computational analysis of genomic information. These efforts were organized in 1990 as the international Human Genome Project, biology's first attempt to create a large-scale infrastructure for studying life.

The first project to get under way was the sequencing of the 12-million-base (Mb) genome of the yeast *Saccharomyces cerevisiae*, with sequences of individual chromosomes pouring out between 1992 and 1996 in an international collaboration involving dozens of labs *(18)*. In 1995, the first complete bacterial genome was produced—the 1.8-Mb *Haemophilus influenzae*. It was all generated in a single laboratory using a "shotgun" technique in which the whole genome is randomly fragmented, the fragments are sequenced, and the results are merged and reassembled into one coherent genome-length sequence *(19)*.

The results transformed cell biology. For the first time, biologists could enumerate the full complement of genes and proteins required to run a living cell. Included here were the essential hardware components of nucleated (eukaryotic) cells and those of the simpler nonnucleated (prokaryotic) cells.

By 1998, the genome of the first multicellular organism—the 97-Mb DNA sequence of the roundworm *Caenorhabditis elegans*—was published *(20)*. And sequencing of the genomes of the mustard weed *Arabidopsis thaliana* and the fruit fly *Drosophila melanogaster* was already nearing completion as this past century drew to a close. One truly astounding result, long suspected, was definitively confirmed by these sequencing efforts: The number of distinct genes required to template a complex organism such as the fruit fly (which has some 13,000 genes) was found to be not much greater than the ~6000 carried in the genome of the single-celled baker's yeast.



Circle of flu. Genome map of the first organism (*Haemophilus influenzae*) to be fully sequenced.

The pace of sequencing has only quickened. The sequence of the human genome is expected to be completed in rough form this year and in finished form not long thereafter. Biologists have begun to think of the complete sequence of an organism's genome as a necessary starting point for serious research.

The Future: Global Views of Biology

The availability of the complete parts lists of organisms, that is, catalogs of all their genes and thus all their proteins, has been redirecting biologists toward a global perspective on life processes—to study the role of all genes or all proteins at once. Twentieth century biology triumphed because of its focus on intensive analysis of the individual components of complex biological systems. The 21st century discipline will focus increasingly on the study of entire biological systems, by attempting to understand how component parts collaborate to create a whole. For the first time in a century, reductionists have yielded ground to those trying to gain a holistic view of cells and tissues.

This new approach promises stunning breadth of perspective. At the same time, it threatens to inundate scientists in a flood of data that will overwhelm their ability to interpret it. Powerful new types of bioinformatics will clearly be required to assimilate and interpret the data that will issue from various types of genomics research. We focus here on a few of these new global perspectives whose outlines are already clear.

Human genetics will be affected in profound ways. Once the human genome is sequenced, a follow-on task will be to understand the spectrum of genetic variation in the human gene pool and its relation to disease. This turns out, surprisingly, to be a tractable problem because of the relatively recent vintage of our species. The current world population of 6 billion people descends from a few tens of thousands of progenitors who inhabited Africa some 150,000 to 200,000 years ago. Such small populations can maintain only a limited degree of genetic diversity—typically, only a few common variants in the coding sequences of each gene in their genome. Moreover, the few thousand generations of subsequent exponential population expansion have been too few on an evolutionary time scale to alter the spectrum of common variation substantially. As a result, the modern human population has much less intraspecies genetic variation than, for example, chimpanzees. Recent experimental studies have confirmed the limited number of common variants in typical genes, raising the prospect that it will be possible to catalog all of the common variants (alleles) of all human genes.

Such common variants attract enormous interest, because it is suspected that they may hold the key to inborn susceptibility to common diseases. A few cases in point are already known, such as common variants in the apolipoprotein E gene that predispose carriers to Alzheimer's disease or in the clotting factor V gene that predispose carriers to deep venous thrombosis *(21)*. Some human geneticists believe that such examples represent only the tip of an enormous iceberg. The challenge here will be to identify the full collection of variants and then test their correlation with diseases.

Just as comparison within the human species will be revealing, so too will comparisons between species. Evolution is a grand experiment in which myriad sequence changes within a gene are tried and tested in the crucible of selection. Evolutionary comparison among organisms illuminates those sequences that play important functional roles in protein structure or gene regulation, and thus have been retained unaltered over extended periods of evolutionary time. Identifying evolution's successful experiments will reveal key functional features of important genes and proteins, obviating years of painstaking laboratory experimentation.

Evolutionary sequence comparison should allow us to identify genes that were crucial to the creation of new species; such genes are likely to have undergone strong selection and more rapid sequence evolution. One putative example of such a gene has been proposed in the fruit fly *(22)*. It would be fascinating to find candidates for the genes and genetic changes that triggered the speciation between our progenitors and those of chimpanzees.

Global approaches are also proving central to efforts to understand the physiology of cells and organisms. Key here will be our ability to survey which genes within a given cell

1983 cont.
Meanwhile, Kary Mullis conceives of the polymerase chain reaction, a chemical DNA replication process that will greatly quicken the pace of genetic science and technology development.

1984
Alec Jeffreys develops "genetic fingerprinting," a molecular biological analog of traditional fingerprinting for identifying individuals by analyzing polymorphic (variable) sequences in their DNA.

1986
The Human Genome Initiative, later called the Human Genome Project, is announced. The goal is to sequence the entire human genome and provide a complete catalog of every human gene.

1987
A large, collaborative effort yields the first comprehensive human genetic map with 400 signposts.

1988
The National Center for Human Genome Research is created, with the goal of mapping and sequencing all human DNA by 2005.

1990
The Human Genome Project is formally launched, with a completion date set for 2005.

W. French Anderson performs the first gene therapy procedure on a 4-year-old girl with an immune disorder known as ADA deficiency. (It didn't work.)

1990 cont.
Mary-Claire King

finds evidence that a gene on chromosome 17 causes an inherited form of breast cancer and increases the risk of ovarian cancer.

1992
An international collaboration produces a rough map of genetic polymorphism: the variable genetic regions along all 23 pairs of human chromosomes that govern person-to-person biological variation.

1995
The Institute for Genomic Research reports the first complete DNA sequence of the genome of a free-living organism—the bacterium *Haemophilus influenzae*.

1996
The first complete sequence of the genome of a eukaryote (the yeast *Saccharomyces cerevisiae*) is reported by an international effort involving some 600 scientists in Europe, North America, and Japan.

1998
The first genome of a multicellular organism—the 97-megabase DNA sequence of the roundworm *Caenorhabditis elegans*—is published by the *C. elegans* Sequencing Consortium.

2000
Projected completion of the first draft of the sequence of the entire human genome.

Each month, Britannica.com enhances the Pathways of Discovery essay with links to relevant items within and without *Encyclopedia Britannica*'s vast stores of information. To access this month's Pathways essay and all previous ones, go to www.britannica.com and click on the "Science" channel.

are being read out (expressed) and which ones are silent. A starting point will come from successful monitoring of how the level of each expressed RNA and protein differs among the cells of different tissues in response to different physiologic signals, or in various disease states. Already, microarray detectors exist that allow researchers to measure the RNA levels corresponding to each of the 10,000 or so known genes (still only 10% of the total), and various approaches are being developed for studying complex mixtures of expressed proteins—the new science of proteomics.

Because the spectrum of expressed proteins within a cell determines its biology, such comprehensive descriptions will provide the basis for understanding precisely why, for example, brain cells differ from kidney cells. They will identify biological markers characteristic of disease states, leading to techniques for early identification. They will help classify cancers into distinct subtypes, making it possible to know a tumor's lineage, the nature of the genetic mutations that led to its appearance, and, in the long run, whether it will respond to a particular therapy. And they will reveal the strategies of attack that a pathogen launches against its host and the defenses mounted by the host to defeat the invading pathogen.

The big picture. With gene chips like this one, researchers can study thousands of genes at once.

Proteins that interact physically invariably communicate with one another. So other techniques, such as the "two-hybrid screen" and its relatives, are being developed to identify these interactions. Maps of these associations will, in turn, shed much light on the design of the channels that send and process signals within living cells.

The long-term goal is to use this information to reconstruct the complex molecular circuitry that operates within the cell—to map out the network of interacting proteins that determines the underlying logic of various cellular biological functions including cell proliferation, responses to physiologic stresses, and acquisition and maintenance of tissue-specific differentiation functions. A longer term goal, whose feasibility remains unclear, is to create mathematical models of these biological circuits and thereby predict these various types of cell biological behavior.

More powerful tools will be needed to realize these goals, at the level of both instrumentation and the computerized processing of biological information, which has now become the cottage industry of bioinformatics. Biologists will require gene-specific reagents to disrupt the function of each component in the cell and study the rippling effects of each such disruption on other genes and proteins within the cell. Various techniques, such as antisense reagents complementary to individual genes and small-molecule screening, are currently being explored with the aim of finding a general technique for disrupting intracellular circuits in a specific, highly targeted fashion. The challenge of disrupting every gene in a human cell is daunting but perhaps not insurmountable—after all, human genes number a mere 100,000 or so, a figure that seems less formidable with the passage of time.

Biology enters this century in possession, for the first time, of the mysterious instruction book first postulated by Hippocrates and Aristotle. How far will this take us in explaining the vast complexity of the biological world? Will we ever be able to draw a protozoan or a peacock, knowing only its DNA sequence? We are hard pressed to provide an answer at the beginning of the century. Still, we proceed with great optimism: The solutions to many problems long resistant to attack are now within reach. The prospects of 21st century biology are surely breathtaking.

At the same time, we must confront this new world soberly and with some trepidation. The genetic diagnostics that can empower patients to seek personalized medical attention may also fuel genetic discrimination. The understanding of the human genetic circuitry that will provide cures for countless diseases may also lead some to conclude that humans are but machines designed to play out DNA cassettes supplied at birth—that the human spirit and human potential are shackled by double-helical chains. So the most serious impact of genomics may well be on how we choose to view ourselves and each another. Meeting these challenges, some quite insidious, will require our constant vigilance, lest we lose sight of why we are here, who we are, and what we wish to become.

References and Notes

1. G. Mendel, *Verh. Naturforsch. Ver. Brünn* **4**, 3 (1866). (The original paper was published in *Verh. Naturforsch. Ver. Brünn* 1865, which appeared in 1866.)
2. K. G. Correns, *Ber. dtsch. bot. Ges.* **18**, 158 (1900); E. Tschermak von Seysenegg, *Ber. dtsch. bot. Ges.* **18**, 232 (1900); H. de Vries, *Ber. dtsch. bot. Ges.* **18**, 83 (1900).
3. A. H. Sturtevant, *A History of Genetics* (Harper & Rowe, New York, 1965); A. H. Sturtevant, *J. Exp. Zool.* **14**, 43 (1913).
4. H. B. Creighton and B. McClintock, *Science* **17**, 492 (1931).
5. H. J. Muller, *Science* **46**, 84 (1927).
6. F. Griffith, *J. Hyg.* **27**, 113 (1928).
7. O. T. Avery, C. M. MacLeod, M. McCarty, *J. Exp. Med.* **79**, 137 (1944).
8. A. D. Hershey and M. Chase, *J. Gen. Physiol.* **36**, 39 (1952).
9. E. Schrödinger, *What is Life?* (Cambridge University Press, New York, 1945).
10. J. D. Watson and F. H. C. Crick, *Nature* **171**, 737 (1953).
11. M. W. Nirenberg and J. H. Matthaei, *Proc. Natl. Acad. Sci. U.S.A.* **47**, 1588 (1961); M. Nirenberg and P. Leder, *Science* **145**, 1399 (1964).
12. J. D. Watson, M. Gilman, J. Witkowski, M. Zoller, *Recombinant DNA* (Freeman, 2nd ed., New York, 1992).
13. A. M. Maxam and W. Gilbert, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 560 (1977); F. Sanger, *Science* **214**, 1205 (1981).
14. J. Downward *et al.*, *Nature* **307**, 521 (1984); R. F. Doolittle *et al.*, *Science* **221**, 275 (1983); M. D. Waterfield *et al.*, *Nature* **304**, 35 (1983).
15. J. R. Riordan *et al.*, *Science* **245**, 1066 (1989).
16. D. Botstein, R. L. White, M. Skolnick, R. W. Davis, *Am. J. Hum. Genet.* **32**, 314 (1980).
17. J. F. Gusella *et al.*, *Nature* **306**, 234 (1983).
18. R. A. Clayton, O. White, K. A. Ketchum, J. C. Venter, *Nature* **387**, 459 (1997).
19. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995).
20. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
21. W. J. Strittmatter *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 1977 (1993); R. M. Bertina *et al.*, *Nature* **369**, 64 (1994).
22. C.-T. Ting, S.-C. Tsaur, M.-L. Wu, C.-I. Wu, *Science* **282**, 1501 (1998).

23. Related work of E.S.L. is supported by a grant from the National Institutes of Health/National Human Genome Research Institute. The authors would like to thank Maureen T. Murray for her comments on the text.

Eric S. Lander and Robert A. Weinberg are members of the Whitehead Institute for Biomedical Research and professors in the Department of Biology at the Massachusetts Institute of Technology, Cambridge, MA 02142. R.A.W. is a Daniel K. Ludwig Professor and an American Cancer Society Research Professor.

CREDITS: (LEFT TO RIGHT) UNIVERSITY OF WASHINGTON; AFFYMETRIX GENECHIP