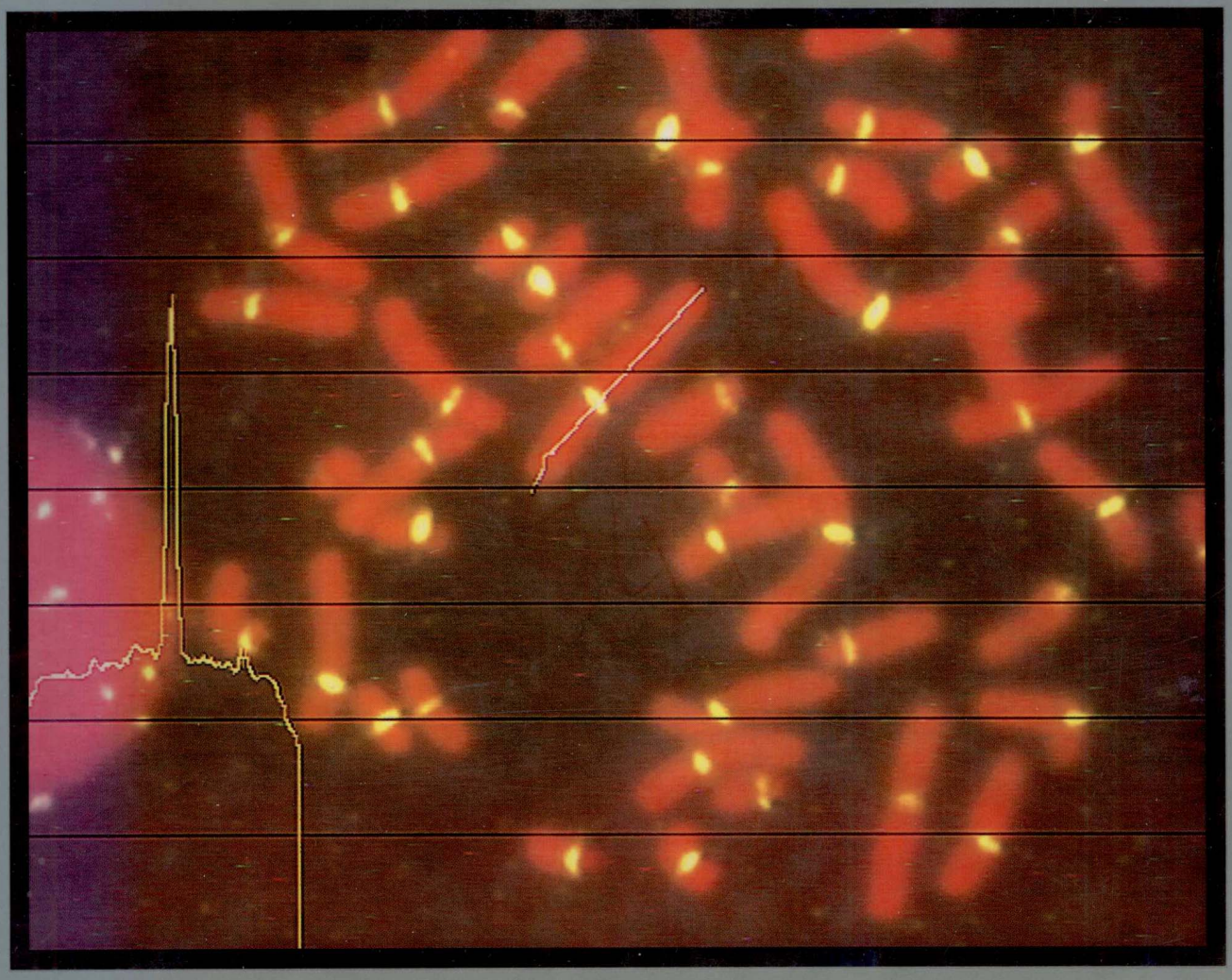
THE HUMAN GENOME PROJECT

DECIPHERING THE BLUEPRINT OF HEREDITY

Edited by Necia Grant Cooper

With a Foreword by PAUL BERG



THE HUMAN GENOME PROJECT

Deciphering the Blueprint of Heredity

Edited by NECIA GRANT COOPER

LOS ALAMOS NATIONAL LABORATORY

Foreword by PAUL BERG

BECKMAN CENTER FOR MOLECULAR AND GENETIC MEDICINE
STANFORD UNIVERSITY SCHOOL OF MEDICINE

UNIVERSITY SCIENCE BOOKS
Mill Valley, California

University Science Books
20 Edgehill Road
Mill Valley, CA 94941
Fax (415) 383-3167

Production manager: *Susanna Tadlock*
Science writers: *Steve Elder, Gerald Friedman, Richard Reichelt, Nancy K. Shera*
Editor's assistant: *Nadine Shea*
Technical illustration: *Andrea Kron*
Original art: *Jim Cruz, David R. Delano, Steve Elder, Mel Prueitt, Gloria Sharp*
Production: *Nadine Shea*
Photography: *John Flower*
Jacket and cover designer: *Robert Ishi*
Printer: *Palace Press*

Library of Congress Catalog Number: 93-085290
ISBN 0-935702-29-6

Printed in Hong Kong

10 9 8 7 6 5 4 3 2 1

THE HUMAN GENOME PROJECT

Deciphering the Blueprint of Heredity

Foreword

Paul Berg

THE HUMAN GENOME PROJECT is the first internationally coordinated effort in the history of biological research. It aims to determine the complete sequence of the nearly 3 billion base pairs that constitute the human genome, and, in its course, to identify the 100 thousand or so genes that define the human species. A subsidiary but essential goal is to learn the base-pair sequences of several model experimental organisms: the bacterium *Escherichia coli* (3 million base pairs), the yeast *Saccharomyces cerevisiae* (14 million base pairs), the nematode *Caenorhabditis elegans* (80 million base pairs), the invertebrate *Drosophila melanogaster* or fruit fly (165 million base pairs), and the mammal *Mus musculus* or house mouse (3 billion base pairs). The purpose of this unprecedented effort is to learn the DNA sequences that determine each organism's phenotypic characteristics and guide its development. This information is critical for expanding and refining our understanding of cellular and organismal functions.

In the course of these analyses, we are sure to identify genes governing a variety of human diseases and, thereby, to develop new strategies for their diagnosis, prevention, and therapy—and, we hope, for their cure. Thus, completing the Human Genome Project will not be the endpoint of our effort. Rather, it will provide us with a new beginning in our quest for understanding life on our planet: its origins; its evolution; the wonders of its development, capabilities, and maintenance; and its vulnerability to the stresses and challenges of the environment.

This highly readable volume summarizes the strategies for proceeding with the project, and it describes the tools that have been developed, and those that still need to be developed, in order to accomplish this formidable task. Thus, the ways in which genetic-linkage, physical, and sequence maps are being constructed are clearly explained. Especially interesting are the discussions by the founders of the discipline and its leading participants concerning the project's implications with regard to human health and the ethics of dealing with the diagnostic and predictive capabilities that will follow.

Paul Berg

Director, Beckman Center
for Molecular and Genetic Medicine,
Stanford University

A GUIDE TO THIS BOOK

Studying the Human Genome is one of the great enterprises of 20th Century science—and one that will carry modern molecular biology well into the 21st Century.

The Human Genome Project rests firmly on the foundation of a half-century of research into the fundamental nature of the gene. This epoch-making progress in molecular biology is chronicled in the first 64 pages of this book. Here the gene is progressively defined—from Mendel's invisible particle of inheritance to Watson and Crick's informational DNA molecule, with its Morse code-like message for the synthesis of enzymes and other proteins. A series of well-illustrated diagrams lays out the Genetic Code and the mechanism for the retrieval and use of the information stored in DNA to manufacture various proteins (pages 45–48).

The section of the book that establishes the DNA → RNA → protein dialectic of classical molecular biology is followed by a discussion of the new and powerful recombinant DNA techniques upon which the Human Genome Project rests. These enable one to dissect the huge cellular chromosomes, with their thousands of adjacent genes, into smaller DNA segments whose individual genes can then be analyzed. Here Restriction Enzymes are introduced as the tools for cleaving large DNA molecules at specfic nucleotide sequences, along with the electrophoretic technology for separating individual gene-size DNA pieces according to size (pages 52–58). The introductory chapter concludes by explaining how the isolated DNA segments can be individually transplanted ("cloned") into self-duplicating minichromosomes in order to maintain them in living bacterial cells. Collectively these cells form a "library" of the human genetic information, with each bacterium corresponding to one volume of the library. Ultimately, individual bacterial cells can be grown and processed to generate the large amounts of purified human DNA necessary for analysis at the molecular level (pages 58–64).

A series of Nobel Prizes recognizes the importance of this pioneering work—

- To Watson, Crick, and Wilkins in 1962 for their discovery of the structure of DNA.
- To Nirenberg, Khorana, and Holley in 1968 for their elucidation of the Genetic Code.
- To Arber, Smith, and Nathans in 1978 for their discovery of Restriction Enzymes.
- To Sanger and Gilbert in 1980 for their development of methods for determining the nucleotide sequence of DNA molecules.
- To Berg in 1980 for developing the first methods for cloning genes.
- To Bishop and Varmus in 1989 for the use of recombinant DNA techniques to identify several of the genes involved in cancer.
- To Sharp and Roberts in 1993 for their discovery of RNA Splicing and Split Genes.
- To Mullis and Smith in 1993 for the Polymerase Chain Reaction, which allows one to selectively amplify chosen DNA molecules, and for methods that allow the production of biologically useful mutant DNA molecules.

The analysis of human genes in which mutation has led to disease is a primary objective of the Human Genome Project.

In recent years hundreds of human diseases have been identified that have a genetic basis. Some are as mild as red-green color blindness, while others are as severe as the defects in lipid metabolism that lead to cardio-vascular disease. The recombinant DNA analysis of mutant genes from affected individuals has shown that these diseases generally result from simple, interpretable changes of one or a few of the three billion base pairs that constitute the human genome. An immediate practical consequence of identifying such genetic lesions is that it becomes possible to devise prenatal DNA tests to determine whether an individual has inherited two recessive defective genes—or one dominant gene—and is therefore at risk for the disease.

One of the first steps in the Human Genome Project is to develop a rough map of the whole human genome, which can then, by ever expanding effort, be increased in resolution until finally every base pair is assigned its own particular location. As a case study for exploring

human chromosomes and their genes, the middle section of the book focuses on the molecular analysis of Chromosome 16 carried out at the Los Alamos National Laboratory. Chromosome 16—some 100,000,000 base pairs in length and about 3% of the human genome—contains a number of important genes, including several involved in disease. For example, the chromosome contains the metallothionein gene family that gives rise to a set of proteins that bind to heavy metals such as mercury, and thus play a role in the body's defense against these toxic environ-mental agents. Among the other important genes on Chromosome 16 are the hemoglobin gene, where mutations are associated with severe anemia, a gene involved in kidney disease, one associated with a neurodegenerative disorder known as Batten's disease, and a gene where DNA rearrangements lead to a form of leukemia (pages 184–185).

The chapter on "The Mapping of Chromosome 16" explains the general strategy by which the Los Alamos National Laboratory has defined the molecular architecture of this chromosome. The discussion traces the purification of Chromosome 16 by a process called Flow Sorting, in which the chromosome to be isolated is first marked by the binding of a fluorescent targeting molecule (pages 182, 236–246). The purified chromosome then becomes the source of material for thousands of specific DNA fragments generated by restriction enzymes. Collectively these restriction fragments represent the individual gene areas of the chromosome. Step by step the Los Alamos researchers explain how the various restriction fragments are ordered into a linear array—a series of adjacencies determined by virtue of overlapping DNA information. In the end a continuous DNA map is constructed in much the same way as the word Biochemistry can be constructed from the overlapping terms biochem and chemistry.

An interesting aspect of the analysis of Chromosome 16 is the discussion of the chromosome's various classes of repeated DNA elements. These multiply-reiterated short sequences not only serve as important landmarks along the length of the chromosome, aiding in the mapping of other genes, but are often interesting in their own right. For instance, one such repeated element, Alu, is a defective version of the gene that codes for an RNA molecule important in the transfer of enzymes and other proteins between the cytoplasm and other cellular compartments (page 186).

The pioneering methods used to map Chromosome 16 are already being extended to the construction of molecular maps for other human chromosomes, en route to a full understanding of the three billion base pair human genome.

Engineers will be fascinated to see the extent to which modern molecular biology has been automated.

The massive effort required to identify, clone, and sequence the entire human genome is not merely a matter of organizing many technicians to work repetitively until finally the three billion base pairs are put together. In fact, techniques available only in the last 5–10 years and ones currently being developed, especially in computer-controlled automation and data analysis, are essential to the Human Genome Project.

One of the most innovative techniques to have been developed and mechanized in the last few years is called the Polymerase Chain Reaction (PCR). This is a way to obtain large amounts of a specific DNA segment by repeatedly amplifying a very small amount of starting material. The technique is so sensitive that a single cell containing two copies of a particular sequence can, after amplification, yield sufficient DNA to allow all of the standard molecular analyses, such as cloning, restriction mapping, and sequencing.

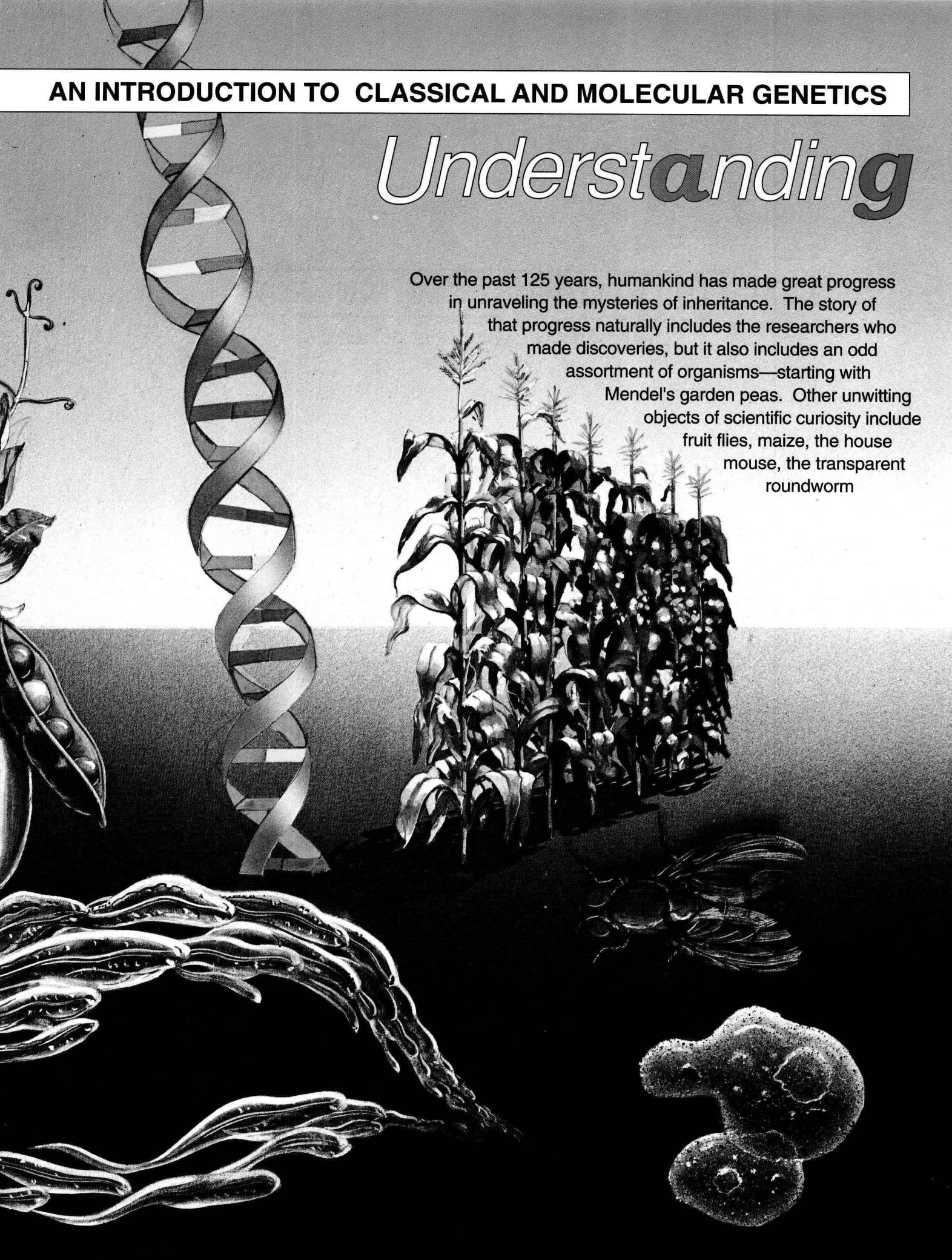
Ultimately, the order of nucleotides in the pieces of DNA that have been mapped onto the human genome must be determined. As with many techniques, the DNA sequencing procedure (pages 151–159) has undergone a continuing revolution. When Greg Sutcliffe in Gilbert's lab sequenced the first plasmid, pBR322—about 40 nucleotides at a time—it took three years and over $100,000 to determine the 4,361 base pair sequence. Modern techniques have been so significantly improved by the introduction of DNA sequencing machines that all of pBR322 could now be sequenced in a few days for under $1000. Indeed, researchers are rapidly generating hundreds of thousands of base pairs of sequence information monthly, which are read directly into computers, where the data can be accessed by other researchers and subjected to detailed comparative analysis using sophisticated computer programs (pages 250–295).

This unusual book integrates the basic science, the specific research goals, and the public policy considerations that are involved in the Human Genome Project. The understanding gained in this project will have profound effects both in our lifetime and beyond. For the information that will be collected about the human genome will permanently reshape our understanding of human disease and indeed of human diversity.

David Dressler
HARVARD MEDICAL SCHOOL

THE HUMAN GENOME PROJECT

Deciphering the Blueprint of Heredity

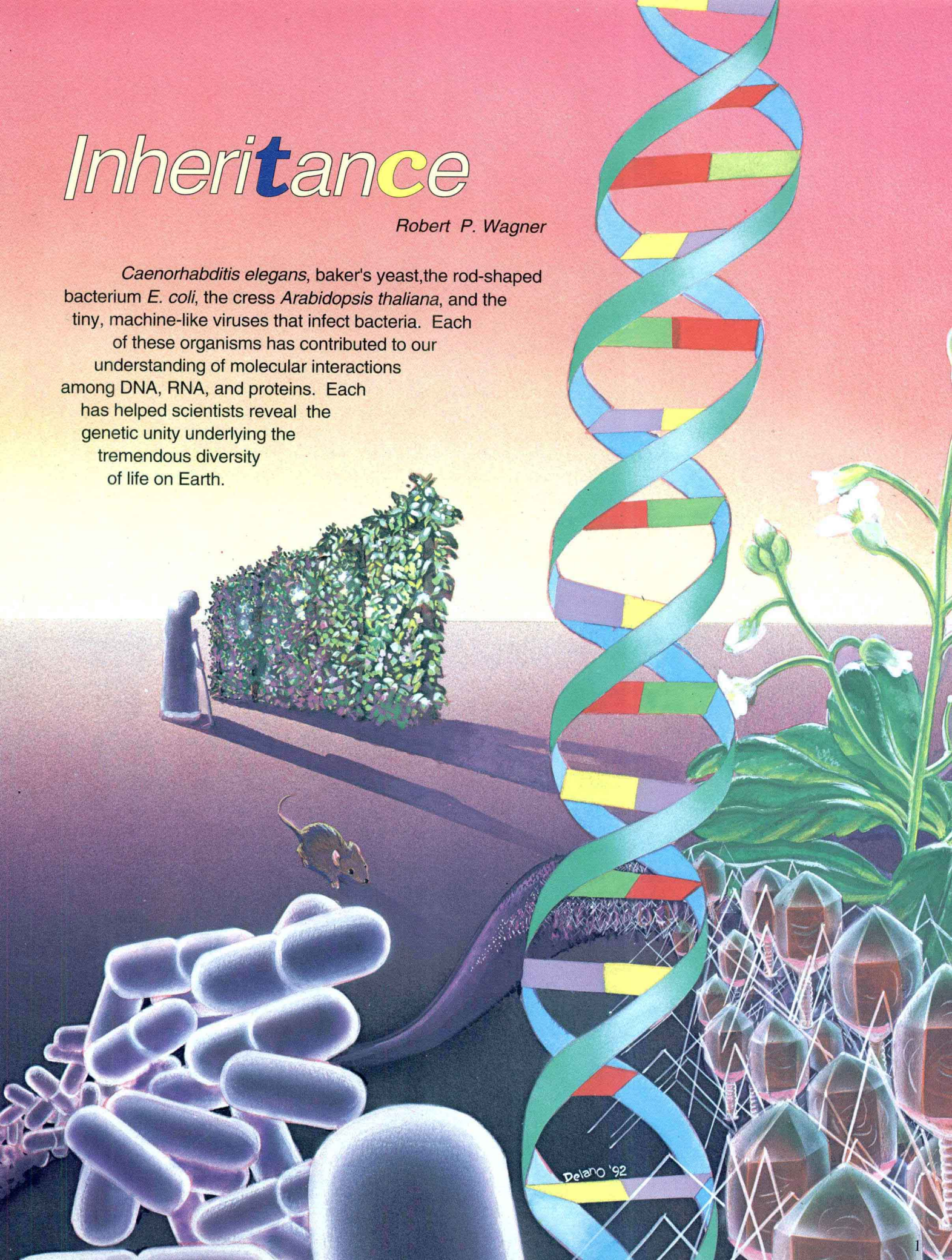

AN INTRODUCTION TO CLASSICAL AND MOLECULAR GENETICS
Understanding
Over the past 125 years, humankind has made great progress in unraveling the mysteries of inheritance. The story of that progress naturally includes the researchers who made discoveries, but it also includes an odd assortment of organisms—starting with Mendel's garden peas. Other unwitting objects of scientific curiosity include fruit flies, maize, the house mouse, the transparent roundworm

Contents

Inheritance

Robert P. Wagner

Caenorhabditis elegans, baker's yeast,the rod-shaped bacterium *E. coli*, the cress *Arabidopsis thaliana*, and the tiny, machine-like viruses that infect bacteria. Each of these organisms has contributed to our understanding of molecular interactions among DNA, RNA, and proteins. Each has helped scientists reveal the genetic unity underlying the tremendous diversity of life on Earth.



That like begets like—that what is now called a species begets offspring of the same species—must have been evident to the earliest humans. Recognition of the inheritance of variations within a species must also have come early, since domestication of animals undoubtedly involved elimination of individuals with undesirable characteristics (a penchant for human flesh, for example). The first animals to be domesticated may well have been members of the dog family, which were used as food, and domestication of canines may have started even before the advent of *Homo sapiens*. The remains of an old hominid relative of ours, *Homo erectus* (also known as Java or Peking man), have been found associated with those of a dog-like animal in 500,000-year-old fossils. The earliest canine remains associated with our own species are a mere 12,000 years old. The domestication of food plants probably began between 8000 and 9000 years ago, although some authorities contend that the domestication of cereals preceded that of most animals.

Humans must also have very early related mating between "male" and "female" animals, including humans, with the subsequent issuance of offspring. Sexual reproduction in plants was probably recognized much later—many plants, after all, are discreetly bisexual—but at least 4000 years ago, as evidenced by the Babylonians' selective breeding, through controlled pollination, of the date palm (*Phoenix dactylifera*), which occurs as separate male and female trees. (The dates borne by a female tree result from fertilization of its eggs by sperm-containing pollen from male trees.)

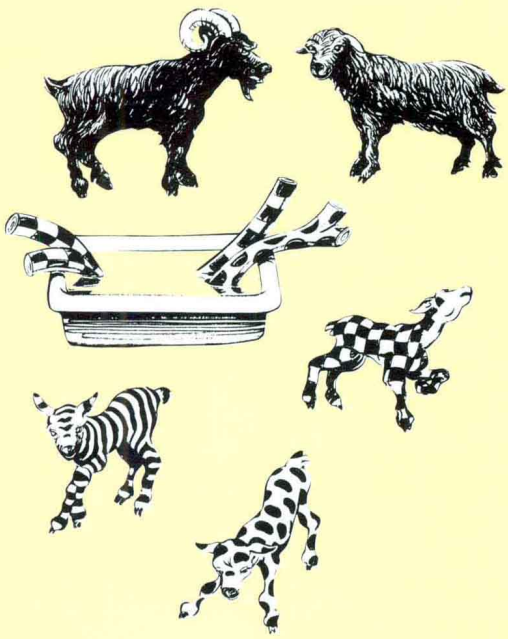
The oldest recorded thoughts about heredity appear in the religious writings of the ancient Hindus and Jews, which reveal recognition of the heritability of disease, health, and mental and physical characteristics. The caste system of the Hindus, the hereditary priesthood among the Jews of the tribe of Levi, and later, in Homer's time, the inheritance of the gift of prophecy are a few reflections of ancient thinking about the link between successive generations of humans. Some of those ideas, which of necessity were based primarily on philosophical outlook rather than scientific fact, are discussed briefly in "Early Ideas about Heredity."

The Dawn

The first significant advances toward our current understanding of inheritance came in the late Renaissance with the work of the English physician William Harvey (1578–1657) and the invention of the microscope (circa 1600). Harvey is best known for his discovery of the dynamics of the circulation of the blood, but he also propounded a new view about the relative importance of the contributions of male and female animals to the creation of offspring. Previously, the female contribution, the egg, had been regarded as mere matter, matter that assumes a form dictated entirely by the male's semen. But Harvey proposed that both egg and semen guide the development of an offspring. His observation of the eggs of many species led him to conclude (in *De generatione animalium*, 1651) that "*ex ovo omnia*." That everything arises from an egg was meant to apply to humans also, even though Harvey had never seen the eggs of humans or any other live-bearing creature.

EARLY IDEAS ABOUT HEREDITY

Ancient beliefs about heredity included the idea that inborn characteristics are inherited from parents, as well as the idea that they could be affected by external influences on the parents at conception or during pregnancy. The biblical story of Jacob's wages (Genesis, chapter 30) combines both. Jacob had agreed to tend the flock of his uncle and father-in-law, Laban, if he could take when he left all the unusually colored animals: the sheep with dark wool and the goats with white streaks or speckles. But Laban, a deceitful and greedy man, took his few such animals three days' journey away. The remaining stock he assumed would not produce offspring of the colorations Jacob had named. However, Jacob peeled tree



branches to make them striped and spotted and stood them in the watering troughs when the stronger goats were mating nearby. The kids from those matings, unlike their parents, had the markings that made them his, and they were more vigorous than the offspring of the weaker goats. He herded the sheep so they faced Laban's dark-colored goats; they then bore dark-colored lambs. Today the appearance in offspring of characteristics different from those of either parent can be attributed to the combined effects of the genetic contributions of each parent (see "Mendelian Genetics").

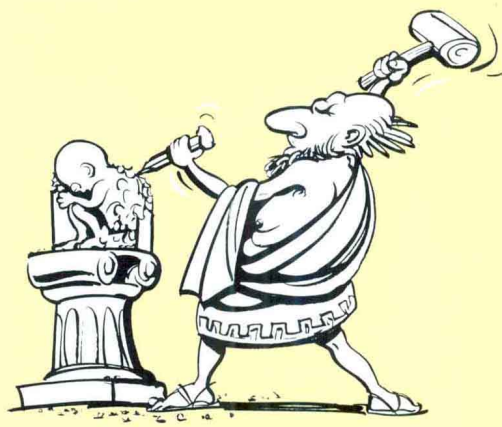
The ancient Greeks gave considerable attention to human inheritance in their writings. Plato, for example, made cogent statements about human traits being determined by both parents. He emphasized that people are not completely equal in physical and mental characteristics and that each person inherits a nature suited to fulfilling only certain societal functions. Also prominent in the thinking of the early Greeks was the inheritance of acquired characteristics. Aristotle, for example, wrote that

> children are born resembling their parents in their whole body and their individual parts. Moreover this resemblance is true not only of inherited but also of acquired characters. For it has happened that the children of parents who bore scars are also scarred in just the same way in just the same place. In Chalcedon, for example, a man who had been branded on the arm had a child who showed the same brand letter, though it was not so distinctly marked and had become blurred.

The idea that external influences play a role in heredity persisted even until the early part of the twentieth century. We now know that the idea contains some truth. For example, ionizing radiation, many chemicals, and infection by some viruses can cause heritable changes, or mutations, but generally those changes are entirely random and cannot be directed toward specific outcomes.

One of the more remarkable theories about human inheritance, pangenesis, was developed in about the fifth century B.C. and espoused by Hippocrates and his followers. According to that theory, semen was formed in every part of the male body and traveled through the blood vessels to the testicles, which were merely repositories. Variations of the theory lasted well into the ninteenth century A.D. and were even accepted by Charles Darwin. Pangenesis was for some reason dominant in the thinking of the philosophers and theologians of the Middle Ages. Albertus Magnus (1193–1280), his pupil Thomas Aquinas (1225–1274), and the naturalist Roger Bacon (circa 1220–1294) all accepted pangenesis as a fact. One variant of the theory was the idea that both male and female produced semen. According to Paracelsus (1493–1541), semen was an extract of the human body containing all the human organs in an ideal form and was thus a physical link between successive generations.

Also prevalent during the Middle Ages was the concept of entelechy, the Aristotelian idea that the way an individual develops is determined by a vital, inner force. The determining force is provided by the male and transmitted in his semen. The female provides no semen but only, so to speak, raw material. Aristotle compared the roles of male and female in the creation of an offspring with the roles of sculptor and stone in the creation of a sculpture.



Other forms of vitalism continued to be popular even up to the beginning of the twentieth century primarily because people lacked knowledge about the nature of the physical connection between generations of animals and plants.

With his naked eye Harvey could see no form in a newly laid, fertilized chicken egg. But he assumed the form that did appear later arose epigenetically from matter that has some sort of inherent, though invisible, organization. The theory of epigenesis—that an organism arises from structural elaboration of formless matter rather than by enlargement of a preformed entity—dates back to Aristotle, but Harvey differed from Aristotle in seriously doubting that the living can arise from the nonliving. Experimental justification for his doubt came about a century later.

Thoughts about heredity would probably not have advanced beyond Harvey's had it not been for the compound microscope, an invention credited sometimes to Zaccharias Janssen and sometimes to Galileo. Other Renaissance men noted for their discoveries with the microscope and improvements to its design are regarded as the founders of microscopy: Nehemiah Grew (1641–1712), Robert Hooke (1635–1703), Antoni van Leeuwenhoek (1632–1723), Marcello Malpighi (1628–1694), and Jan Swammerdam (1637–1680). Their observations—among which were sperms in semen and structural elements, dubbed cells by Hooke, in plant and animal tissues—formed the foundations of the science now called cell biology.

Users of the early, low-resolution microscopes could (and did) let their imaginations run wild. Some thought they saw miniature humans, homunculi, preformed in human sperms; others saw tiny animals, animalcula, preformed in animal eggs. Those apparitions led to resurrection of the theory of preformation originally propounded by Democritus and other Greeks. In the eighteenth century the preformation theory developed into the encapsulation theory, which stated that, at the time of creation, all future generations were packaged, one inside the other, within the primordial egg or sperm. Logically, all life would come to an end when the last homunculus or animalculum was born. The encapsulation theory died—because it was ridiculous—although many eminent biologists were its fierce advocates up to the beginning of the nineteenth century.

The higher-resolution microscopes of the later half of the eighteenth century allowed Caspar Friedrich Wolff (1734–1794) to observe the development of chicken embryos. His work clearly showed that the components of a new organism are not preformed but, as stated two millenia before by Aristotle and a century before by Harvey, arise from the undifferentiated matter of the fertilized egg.

The Great Awakening

Modern biology may be said to have been born in the nineteenth century, several hundred years after the beginnings of modern chemistry and physics. Earlier biologists were either physicians or naturalists (what we now call botanists and zoologists), and their work focused on structure, physiology, and classification. But the nineteenth century brought several developments that were basic to emergence of the newer branches of biology, including cell biology and genetics.

The Rise of Cell Biology. During the first half of the nineteenth century, evidence accumulated for the so-called cell theory, which states that the cell is the structural and functional unit of all organisms. The diversity of cell shapes and sizes was noted (see "The Variety of Cells"), and various intracellular structures were observed (see "Components of Eukaryotic Cells"). Of particular importance to genetics is the membrane-bound intracellular structure called the nucleus, which was found to be a common feature of the cells of all organisms more complex than bacteria and blue-green algae. Organisms possessing a nucleus were classified as eukaryotes, and organisms lacking a nucleus were classified as prokaryotes.

Later, during the early 1850s, came the momentous finding, embraced in the aphorism *omnis cellula e cellula*, that cells divide to form new cells. A leading proponent of the idea that all cells come from cells was the German physician Rudolph Virchow (1821–1902). A cancer specialist, among other things, Virchow asserted that cancer cells arise from cells pre-existing in the body and do not, as earlier physicians had thought, arise by spontaneous generation from unorganized matter.

Another development was the realization that gametes (sperms and eggs) are also cells, in particular cells specialized for transmitting information from one generation of a sexually reproducing organism to the next. The remarkable difference in size between sperms and eggs was found to be due to cell components other than their nuclei, and that observation, coupled with the belief that sperms and eggs contain the same amount of hereditary information, indicated that hereditary information resides in the nuclei of gametes. The nucleus was found to be the site also of the information transmitted from one cellular generation to the next.

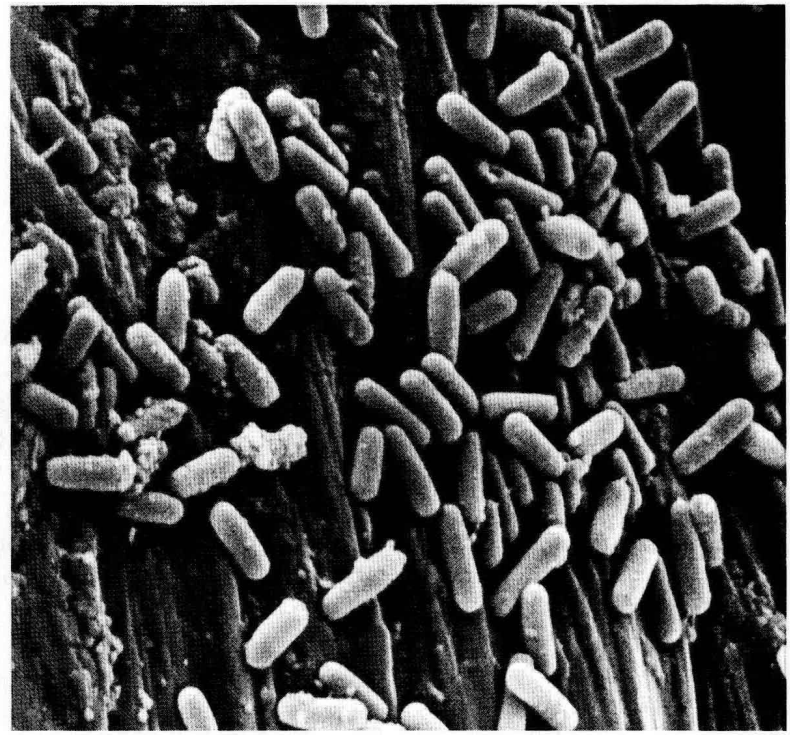
The above developments led to formulation of the law of genetic continuity, which succinctly summarizes what was probably the most important advance toward the understanding of living systems up to that time: Life comes only from life through the medium of cells.

By the late 1880s hereditary information had been localized further to intranuclear elements that can be seen with the microscope during the mitotic phase of the cell cycle, the phase that culminates in cell division (see "The Eukaryotic Cell Cycle"). The elements, which were named chromosomes because they can be stained (selectively colored) with certain dyes, are most easily observed during the portion of the mitotic phase called metaphase. (We now know that each "metaphase chromosome" consists of two duplicates of a single chromosome bound together along a more or less central region.)

Facts accumulated about chromosomes (see "Chromosomes: The Sites of Hereditary Information"). All the somatic cells (cells other than gametes) of a sexually reproducing organism have the same even number of chromosomes, the so-called diploid number, whereas all its gametes have the same so-called haploid number of chromosomes, which is exactly one-half the diploid number. Furthermore, the diploid and

THE VARIETY OF CELLS

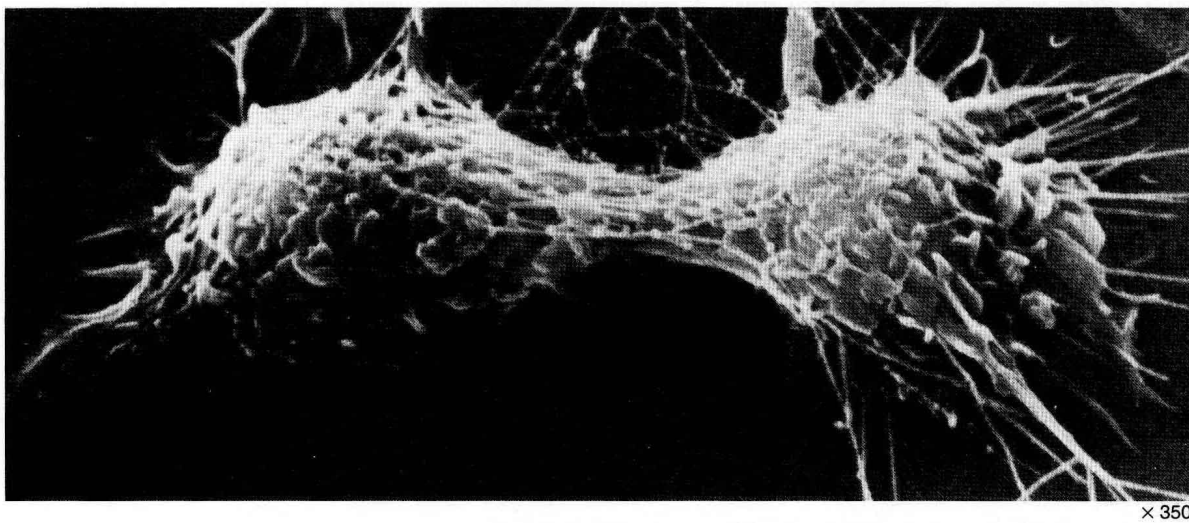
Cells vary in shape from the most simple to the indescribably complex. Shown here are electron micrographs of a few examples from nature's cornucopia.



× 4500

***Escherichia coli,* the most studied of all bacteria**

From *Molecular Biology of the Cell,* second edition, by Bruce Alberts et al. Copyright 1989 by Garland Publishing, Inc. Reprinted with permission. Courtesy of Tony Brain and the Science Photo Library.



× 3500

Mouse fibroblast during the final stage of cell division

From *Molecular Biology of the Cell,* second edition, by Bruce Alberts et al. Copyright 1989 by Garland Publishing, Inc. Reprinted with permission. Courtesy of Guenter Albrecht-Buehler.