PART ONE

The Scientific Foundation

1

Why Genetics?

THE OKLAHOMA SUNSHINE hit me right in the eyes, ending a long and tearful night. It was on odd way to celebrate the fourth of July, 1976, the bicentennial of the signing of the Declaration of Independence. I awoke in the farmhouse built by Robert Ross.* It was completed some thirty years before, when he was in his early thirties. The house and nearby barn were monuments to Robert's once considerable skill in carpentry—premortem tombstones for the man he had been before Alzheimer's disease destroyed his mind.

Robert's wife, Emma, and daughter, Ellie May, took me to see the corral he had tried to build at age forty. It was a rail fence, vastly simpler that the beautiful barn and farmhouse from which it extended. The fence was misshapen, boards wopperjawed and nails askew—an external embodiment of the decay in his cerebral cortex. Emma dated the onset of his illness to the day he came to her, tears of bitter frustration in his eyes, when he realized he did not know how to build the fence. He became aware he had lost the ability to think. It was the beginning of a devastating travail, destined to last twenty-five years. The awareness of dwindling capacities tortured him less as the years passed, because his capacity for any kind of awareness dissipated. The pain slowly shifted from Robert Ross to his family, and particularly to Emma, his partner for life.

*Names have been changed to protect the privacy of family members.

My first night in Oklahoma was absorbed in recapitulating the course of Robert's disease. It was a story of relentless loss. Over the first ten years, Robert slowly deteriorated, progressively losing his ability to do farm work, then to help with even the most menial chores. He became a ward of his wife and children, and by age fifty he could not even recognize them. Robert's blank stare was the most painful torture for his wife and children. These were the central characters of his life, its most cherished rewards. Now they were meaningless, literally beyond recognition, and they knew it even if Robert did not. Such a fire sears even those emotionally thick of skin, leaving permanent scars.

The Ross home country lay in the middle of the Oklahoma panhandle. This was land that at one time no state had wanted, and so it was here that North America's native populations were displaced in the face of the European migration westward. It later became a capital in the Dust Bowl. The Ross family weathered the Depression only to succumb to another catastrophe. Ten of Robert's thirteen siblings also developed Alzheimer's disease. The disease struck a half-dozen cousins. This toll of disease left the family reeling, attempting to deal with a remorseless foe they could not see, whose advance they could not stop.

I went to Oklahoma with Jeanie, a technician from a University of Colorado genetics laboratory, to meet the family and to collect blood and saliva samples. Our mission was to refine the already considerable mass of pedigree data we had assembled on this gifted but star-crossed family, and to bring samples back to Denver for analysis. I was twenty-three, fresh from my first year in medical school; Jeanie was in her early forties. I had taken one elective course in neurology, and had read seventy or eighty articles on Alzheimer's disease, the bulk of world literature at that time. (Contrast this with the more than four thousand items found in a search of the medical literature over a thirty month period in the early 1990s.) In those days, Alzheimer's disease was not yet a household word; it was a scientific backwater. A decade later, Alzheimer's disease got "hot" as a research topic, attaining that status in part because of a 1976 editorial by Robert Katzman in a leading neurology journal, which described its ravages on the population.[1]

I first met Robert Ross in the spring of 1976, at a Veterans Administration long-term care facility in Fort Lyon, in southeastern Colorado. The VA hospital was initially constructed to house those afflicted with tuberculosis, hence its isolated location on the arid plains. By sheer coincidence, my grandfather had been clinical director there in the 1930s, and my father, now a Denver physician, was born there. As hygiene and antibiotics conquered tuberculosis, the hospital was transformed into a mental health facility, accepting patients from a multistate area. It included a few special wards for long-staying patients. Robert had been there for several months when I first encountered him, flagged for further workup by James Austin, chairman of the department of neurology at the University of Colorado. Austin and other neurologists from the univer-

sity periodically visited the Fort Lyon facility. I came on this particular visit specifically to try to find cases of Alzheimer's disease that might have a genetic origin.

Robert was sixty when I met him, completely mute, with his arms contracted permanently into the fetal position. Robert's only responses to the outside world were myoclonic (involuntary) jerks provoked by loud noises or bright lights. Robert was the shell of a man who had been dearly loved. His story was typical of the clinical history of Alzheimer's disease, unusual only in having started at such a young age (forty) and having followed such a protracted course (ultimately twenty-five years). There were several notes in his files indicating that others in his family had Alzheimer's disease. Hence our interest. At the time, there was an active debate about whether Alzheimer's disease could be inherited. The most popular British textbook stated flatly, "It is not inherited,"[2] yet there were twenty or so papers in the literature, dating back to German research between the wars,[3–5] suggesting that some families carried an Alzheimer's gene.[6,7]

Austin was convinced that there was indeed an inherited form of the disease, and thought that a genetic research strategy was likely to be productive. The basic idea was to isolate a gene associated with the heritable form of Alzheimer's disease and then to determine the gene's function. It was a conceptually sound strategy that had been widely discussed for many diseases, although never successfully carried out at the time. The problem was that the tools to implement the strategy were primitive, and might well be inadequate for the task. The first step was to trace the disease through a large family, to see if there was a pattern suggesting inheritance of a single gene.

I met Emma Ross in Denver a few weeks after first seeing her husband, Robert. She had driven from Oklahoma to Denver, bringing a homemade pedigree with hundreds of individuals. The pedigree was written on pieces of blank school paper taped together. When unfolded, the collage covered a conference table. Some symbols had the wrong shapes, the data were incomplete, and the connections between some parts of the family were unclear, but the essentials were sound. The extended family tree represented years of diligent effort. The hundreds of hours that went into constructing the pedigree were far beyond what we had any right to expect. It was immediately obvious that the simplest explanation for who got Alzheimer's disease in the Ross family was a gene that caused Alzheimer's disease in a single dose; in other words, one copy of the bad gene from either parent was enough to trigger the disease. The Ross family might contain the information necessary to find the gene. Emma and I agreed that the next step was for me to meet the family.

The Ross pedigree spanned six generations, from Robert's great-grandfather to his grandchildren. The great-grandfather, of German extraction, immigrated from near the Volga River in Russia to the Midwestern plains of the United States in the decades following the American Civil War. The great-

grandfather reportedly suffered from severe confusion late in life. Robert's grandfather and father had similarly been said to have some mental illness, although it was not clear what it was. Their families had to take care of them, beginning in their late forties. Given the small population base and the tendency to marry within ethnic groups, it was possible that Robert's father and mother both carried the Alzheimer's gene, explaining why eleven of fourteen children were affected. (At the time, only nine of the eleven cases were clear-cut; the other two cases were confirmed years later.) It was also possible this family was just extraordinarily unlucky, having rolled the genetic dice and gotten bad rolls in all but three cases. The basis for inheriting Alzheimers's disease in the one group of brothers and sisters was not entirely clear, but the family was so large that it seemed likely to be a fruitful source of information. We needed to document the clinical histories and to obtain samples for biochemical and genetic analysis.

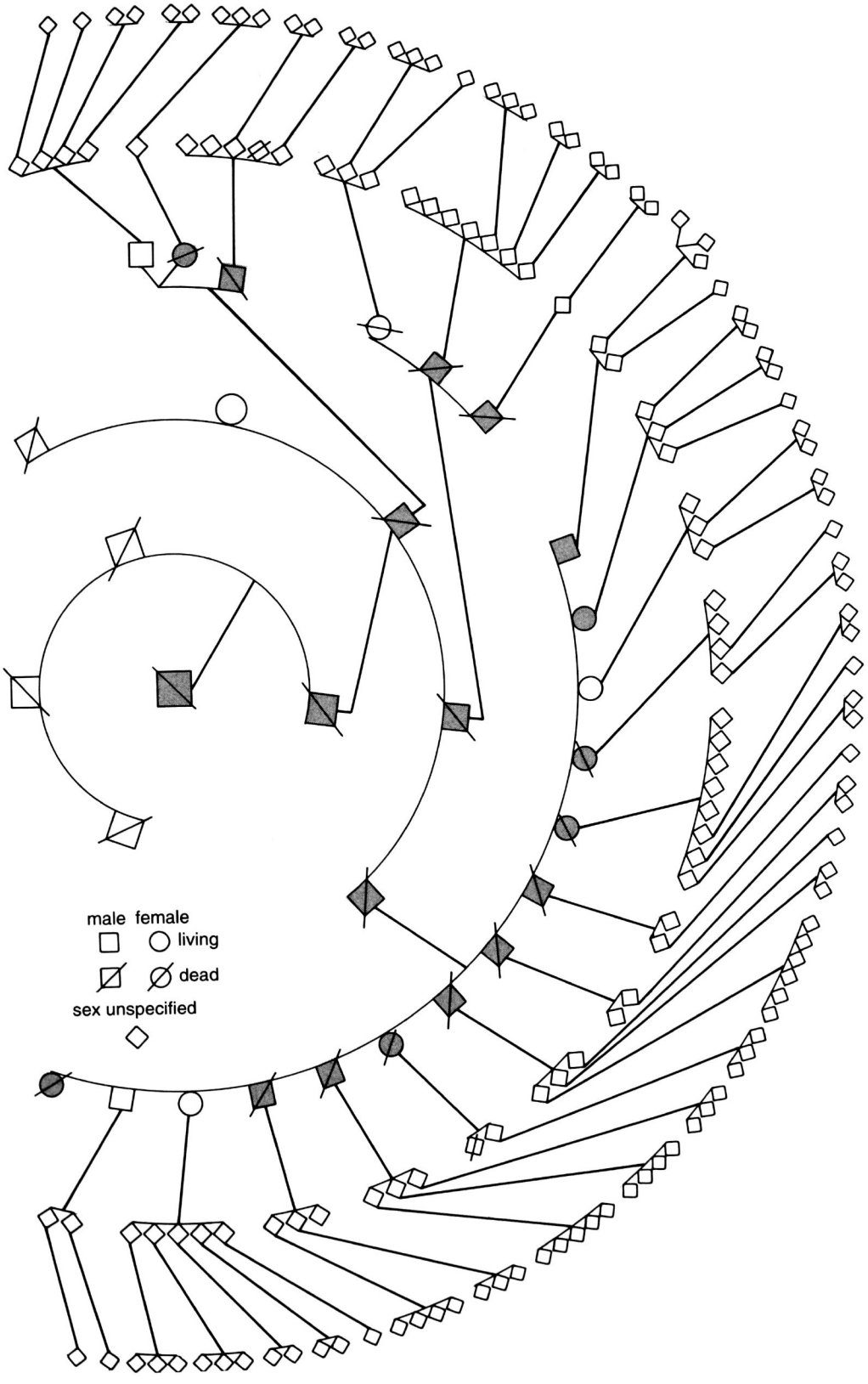
Jeanie and I met the Ross family at a schoolhouse reunion. It was a hot Oklahoma Saturday. Emma and Ellie May had contacted the branches and twigs of the family tree, and had arranged a family picnic to celebrate the Fourth of July. There were fifty to sixty people present. Jeanie and I already had a foretaste of the havoc Alzheimer's disease had wrought on this family. En route to Oklahoma, we stopped to see a cousin afflicted with the disease in nearby Kansas; before the meeting, we visited the family of a brother who had the disease but who had not yet admitted he was affected. The rest of the family already knew nonetheless. (Two years later, the symptoms were much worse, and the disease openly accepted, but this patient's immediate family remained estranged from the rest of the family.)

The Ross family grew up as Thomas Jefferson would have urged, as stalwart citizen farmers—sturdy stoics whose lives were centered on church, family, and farm. Moods swung with the weather, dependent on prospects for that year's crop. Members of the Ross family bore a heavy additional burden. Children of those who developed Alzheimer's disease stood an even chance of developing, or escaping from, the disease before they turned fifty.

Those born to an affected parent were constantly on watch for signs of early Alzheimer's disease in themselves. They had seen the wreckage of the

---

Pedigree of the Rosses shows the devastating impact Alzheimer's disease can have on a family. (Family members affected by the disease are indicated by gray symbols.) The disease was traced back four generations to Robert Ross's great-grandfather, who immigrated from Russia to the Oklahoma panhandle more than a century ago. The pedigree was initially constructed by Emma Ross (as she is known in this book). It was then corrected and amended as other family members were systematically contacted and clinical records and autopsy reports were checked on every person recorded as having Alzheimer's disease, on the brothers and sisters of those affected, and on those who married into the family. In this version, the pedigree has been altered somewhat to protect the confidentiality of family members. Nevertheless, the essentials have been retained, showing the four generations of affected family members, and many more at risk in the succeeding generations.


male female
living
dead
sex unspecified

disease, and loathed foisting it on their children and spouses. Those not directly at risk dreaded the day they would notice the first symptoms in loved brothers or sisters, uncles or aunts. Every car key forgotten, every light left on, every meeting missed became the focus of great distress. Was the disease beginning? This intensified the family dynamics, with denial, anger, and valiant acceptance constantly ripening on some branch of the family tree.

Emma was committed to shedding light on the disease. She assembled her pedigree, and kept records on deaths and births. Some members angrily resented Emma's meddling in their affairs. They protected their privacy against the incursions of scientists like me who sought to study the family in hopes of uncovering some clue to the causes of the dread disease. For some people at some times, the pain was just too overwhelming to let others near.

The stories of those who opened themselves to our inquiries were diverse, but they had a common tragic theme: the slow death of a mind in an otherwise healthy body. Each family described a period of grief that long preceded death and stretched on until death came, often years or even decades later. In many ways, death was a release. There was immense strife as once robust men and women became abject dependents. Every time a marriage was in prospect, there was a debate about when and how much to tell the prospective family member. The new spouse might someday bear responsibility for taking care of an Alzheimer's victim. Those who married into the family knowing the risks faced a difficult initiation. Others were not warned, and their resentment at having been kept in the dark haunted family gatherings.

As a twenty-three-year-old neophyte bearing the tools of science, I was eventually welcomed into the family as an intimate observer and archivist, privy to the most private family stories. I learned many details not known by others within the family—who had been adopted, who had artificial insemination, who was illegitimate. In the long tradition of medical research, with this intimacy I was handed the responsibility to guard the information. I was admitted to the inner sanctum because my art, molecular biology, was a source of hope—if not for the afflicted, then for their children.

In the schoolhouse, we ate a bounteous meal, replete with the Midwestern staples of beef, fresh corn, pie, and chocolate cake. I gave a short talk about Alzheimer's disease. The talk ended with an explanation of why Jeanie and I were there and how we hoped to locate the gene causing Alzheimer's disease by looking for other genes that might be near it, inherited along with it. Their large family and well-documented medical histories, I told them, would give us the best chance yet to get close to the Alzheimer's gene.

Genes are stretches of deoxyribonucleic acid (DNA) that contain the instructions to make a biological molecule. There are roughly six feet of DNA tightly coiled in each of the trillions of cells in the human body (with the exception of a few cells, like red blood cells, that lose their DNA as they mature). DNA is packaged with proteins into chromosomes, microscopic "colored bodies" in the cell's nucleus.

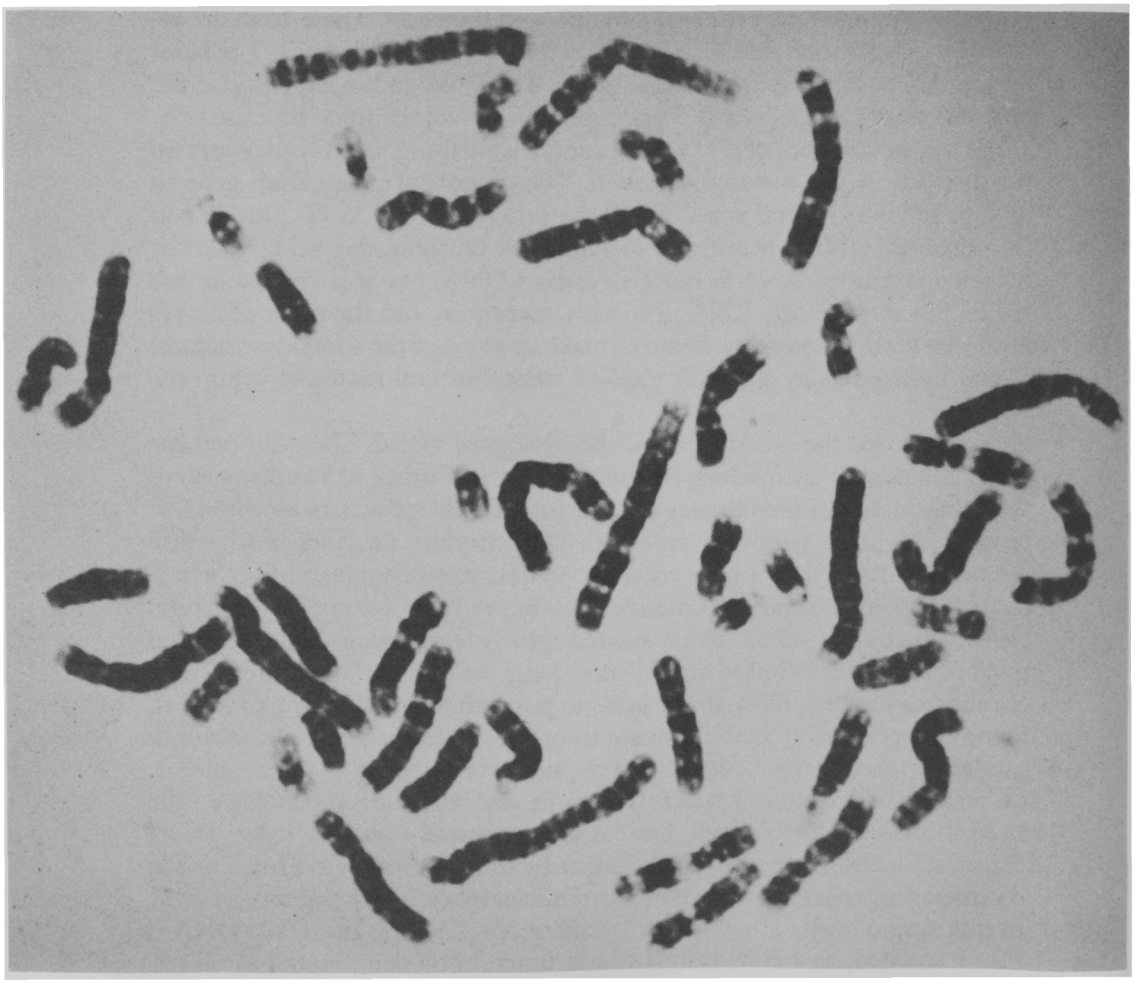
We were searching for a molecular handle on a mysterious disease. If we could find the gene's approximate location, its position among the chromosomes, we would take a step toward finding the gene itself, the actual DNA encoding some faulty molecule responsible for Alzheimer's disease. Through an extremely tedious but logical series of investigations, we might be able to find the molecule produced by the errant gene, and thus discover at least one molecular defect underlying the disease. We might even be lucky and find that the gene causing Alzheimer's disease was already known, but not yet associated with disease. There might well be other ways to develop Alzheimer's disease besides having a bad gene, but studying a clearly genetic form, in families such as the Rosses, was a logical scientific strategy. It was a relatively "clean" way to study a disease otherwise so immensely difficult to approach.

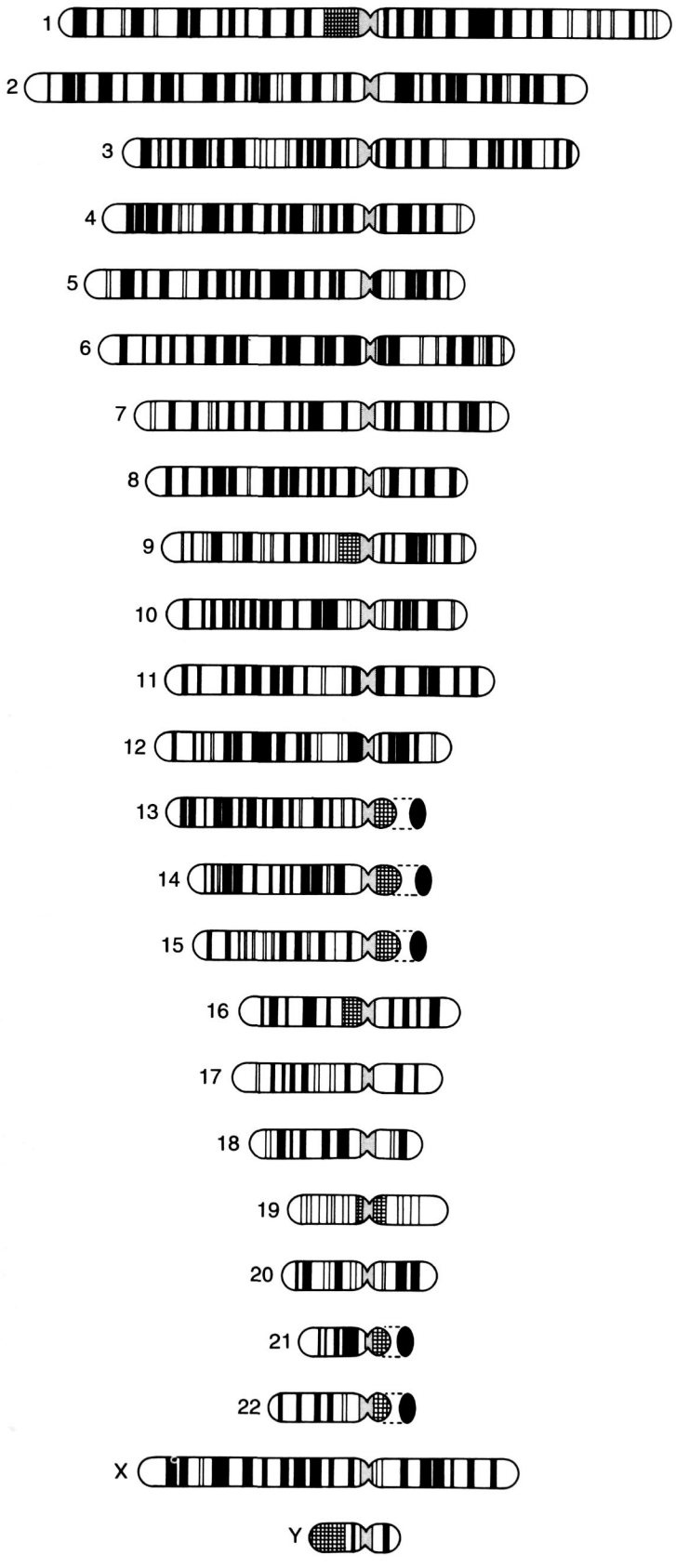
DNA, I explained, consists of long strings of chemical building blocks. There are four constituent chemicals, or nucleotide bases, abbreviated as A, C, G, and T (for adenine, cytosine, guanine, and thymine). These bases are attached to a backbone that is wound into the famous double helix. The bases form the steps in the spiral staircase of life. The DNA code is expressed in the order of the A's, C's, G's, and T's going up (or down) the spiral staircase.

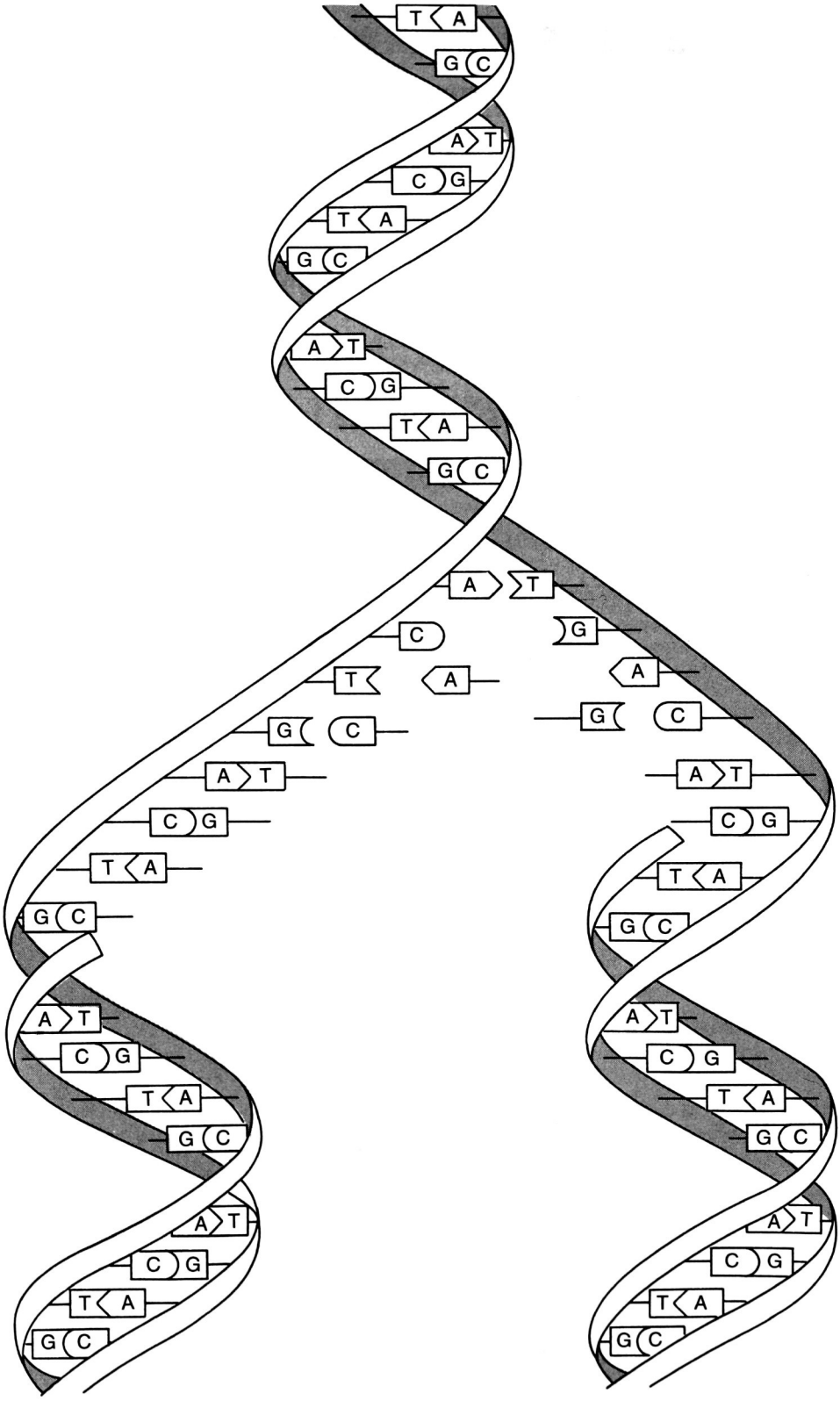
Genes are stretches of DNA that produce something. Usually they contain the instructions for making a protein. The process of going from gene to protein involves several steps. First, the order of A's, C's, G's, and T's in chromosomal DNA is transferred to a molecule of ribonucleic acid (RNA) by cellular machinery. RNA is quite similar to DNA, but it is chemically less stable. For most genes, RNA is, in turn, translated into the order of amino acids that make up proteins. Proteins make up many of the complex structures in and between cells, and they mediate most chemical reactions within the body.

Proteins are the workhorses of the biological world. They can become cellular structures themselves, or the precursors of other structures (such as the components of membranes that surround cells, or sugars that bond to proteins). A large family of proteins called enzymes catalyzes biochemical reactions within cells. Proteins are made up of strings of amino acids, of which there are twenty common varieties. All twenty have a common structural element that allows them to be knitted into protein strings. The structural backbone element is linked to a diverse range of chemical structures. These chemical differences allow amino acids to perform widely differing functions. Some amino acids are best at linking to others. Some are especially useful in catalyzing reactions of a particular type, such as removal of a water molecule or rupture of a chemical bond. Others fit smoothly into membranes. The twenty common amino acids present an enormously variable repertoire of chemical functions when strung together by the hundreds into proteins. The chemical properties of proteins are determined by the order and type of these twenty amino acids. The linear sequence of A's, C's, G's, and T's in DNA is thus translated, as a rule, into chemical function by determining the linear

Chromosomes, the repositories of genetic information, consist of extremely long DNA molecules bound to proteins and other cellular components and wound into the distinctive "supercoiled" shapes seen in the microscope. The micrograph below shows the full complement of 46 chromosomes in a cell of a normal human male during the metaphase stage of cell division. The chromosomes have been stained to reveal their characteristic banding patterns. The diagram at right shows the detailed banded structure of the 22 autosomes (or nonsex chromosomes, present in pairs in the micrograph) and the one pair of sex chromosomes (X and Y). The chromosomes are aligned along their centromeres, or constricted regions (gray areas). The total amount of DNA incorporated into a complete chromosome set of this type is the human genome. *Photo courtesy Department of Clinical Cytogentics, Addenbrookes Hospital, Cambridge, England / Science Photo Library / Photo Researchers, Inc.*




1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y


T A
G C
A T
C G
T A
G C
A T
C G
T A
G C
A T
C
T
G
A
A
G C
A T
C G
T A
G C
A T
C G
T A
G C
A T
C G
T A
G C
A T
C G
T A
G C
A T
C G
T A
G C

sequence of amino acids in proteins. A one-dimensional digital code in DNA is translated into a string of amino acids, which in turn folds into a three-dimensional functioning molecule.

Somewhere in the six feet of DNA was a gene that caused nerve cells to die prematurely, and Alzheimer's disease to develop in the Ross family. The smallest human chromosomes were estimated to contain about fifty million DNA bases strung together. The largest chromosome was roughly five times longer. One set of human chromosomes (one of each pair of chromosomes) contained an estimated three billion base pairs. Fishing the Alzheimer's gene out of this vast ocean of DNA was an awesome task. We needed a navigational chart—a map.

In 1976, there were only seventy or so "markers," genes whose location was known on the twenty-two pairs of nonsex human chromosomes.[8] These markers were the reference points by which to navigate on a genetic voyage in search of an unknown gene—an undiscovered island. Many chromosomes had only one or two markers each, so the signposts were few indeed. The tools of human molecular genetics were exceedingly imprecise; we as investigators were frustratingly impotent. But it was worth a shot. Any action was better than hopeless waiting. The seventy markers were useful, but far fewer than we needed to have a good chance of locating the gene.

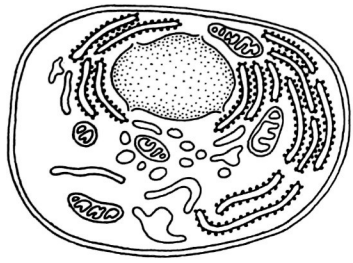
I returned to Oklahoma and Nebraska and Kansas and Texas several times over the next decade, always to a family reunion followed by an assembly line to gather more samples. We found other families in Colorado, in California, and in the Midwest and began to study them as well. Over the years, we could apply methods developed in the rapidly expanding area of human molecular genetics. Each year, we obtained more clinical records on family members. Each year, there was another seminar. At every family meeting, I could report progress, but nothing even vaguely resembling a major breakthrough. Twice there was a newly discovered victim whose misfortune cast a pall over the meeting, making our scientific progress seem paltry by comparison. Certainly our own work was inconclusive. It was the usual story of medical science—pushing inadequate analytical tools to the limit in search of some clue about how the body works. It was awful, but it seemed important to persist in the face of long odds.

A geneticist can work for years in a laboratory, never seeing an affected patient or commiserating with an afflicted family. The daily laboratory routine is relatively stable, if intense and demanding. Once the impact of a disease is directly experienced—the pain and devastation it causes for specific people—

DNA replicates itself by "unzipping" its two helical strands and incorporating new nucleotide building blocks from the surrounding medium in precisely the right order to form two identical copies of the original double helix. Each strand contains all the information needed to make the opposite strand, because the nucleotide bases represented by A's bind specifically to T's (and vice versa), while G's bind specifically to C's.

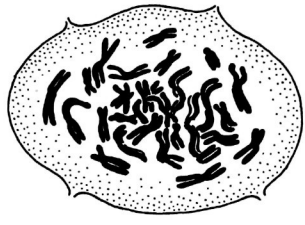
earth

cell

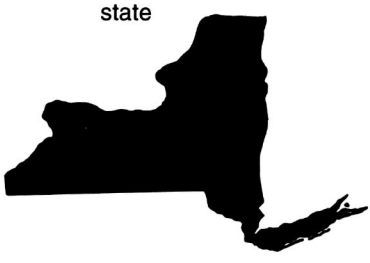


continent

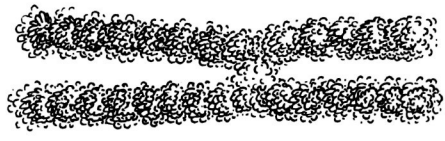
cell nucleus



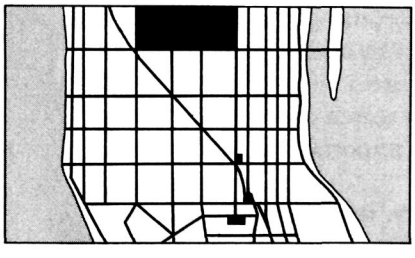
state



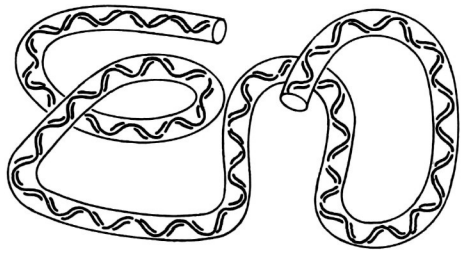
chromosomal DNA



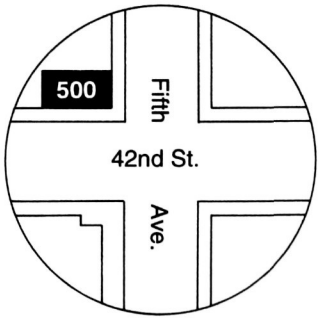
city



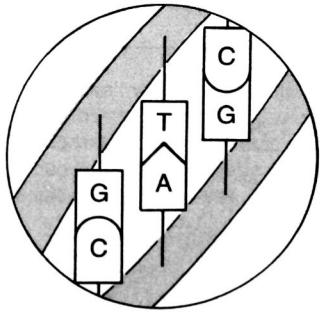
chromosomal DNA fragment



street address



codon



laboratory work acquires new meaning. It demands greater urgency. The stakes go up; the room for excuses and tolerance of delay go dramatically down. Laboratory manipulations become less an exercise in abstract problem-solving and more a holy crusade against a common enemy. Disease becomes evil; eradicating it a primary need. Medical research differs from other scientific fields in this respect. It is driven by this passion for life—the hunger to understand life in order to preserve it.

Robert Ross died in 1981, at age sixty-five, of pneumonia. Two hours later, I cut the spinal cord and lifted his brain out of the cranium. Robert's brain was now a pound of gelid mush in my hands. I weighed the brain on a scale. I will always remember this bizarre act, the culmination of years of work. I knew Emma and Robert's children. I had seen the farm he built. This was the brain that felt emotions and thought thoughts whose objects I had myself encountered. I knew that beloved wife, that farm, that family. The frustration of a mangled fence began here.

The ravages of Alzheimer's disease had reduced Robert's brain mass by a third. It was grossly abnormal, and the severity of the deterioration was apparent as soon as the brain tipped the scale. We sent the tissue off to D. Carleton Gajdusek's laboratory at the National Institutes of Health for analysis, along with tissue frozen from various organs. Under the microscope, Robert's brain was riddled with microscopic plaques, craters left by the bombs in his genes. His was roughly the sixtieth autopsy I performed during my internship year, but the most memorable by far. Studying Robert's brain was another small step toward ridding the world of Alzheimer's disease. If the tools of genetics had been more powerful at the time, it could have been a longer stride.

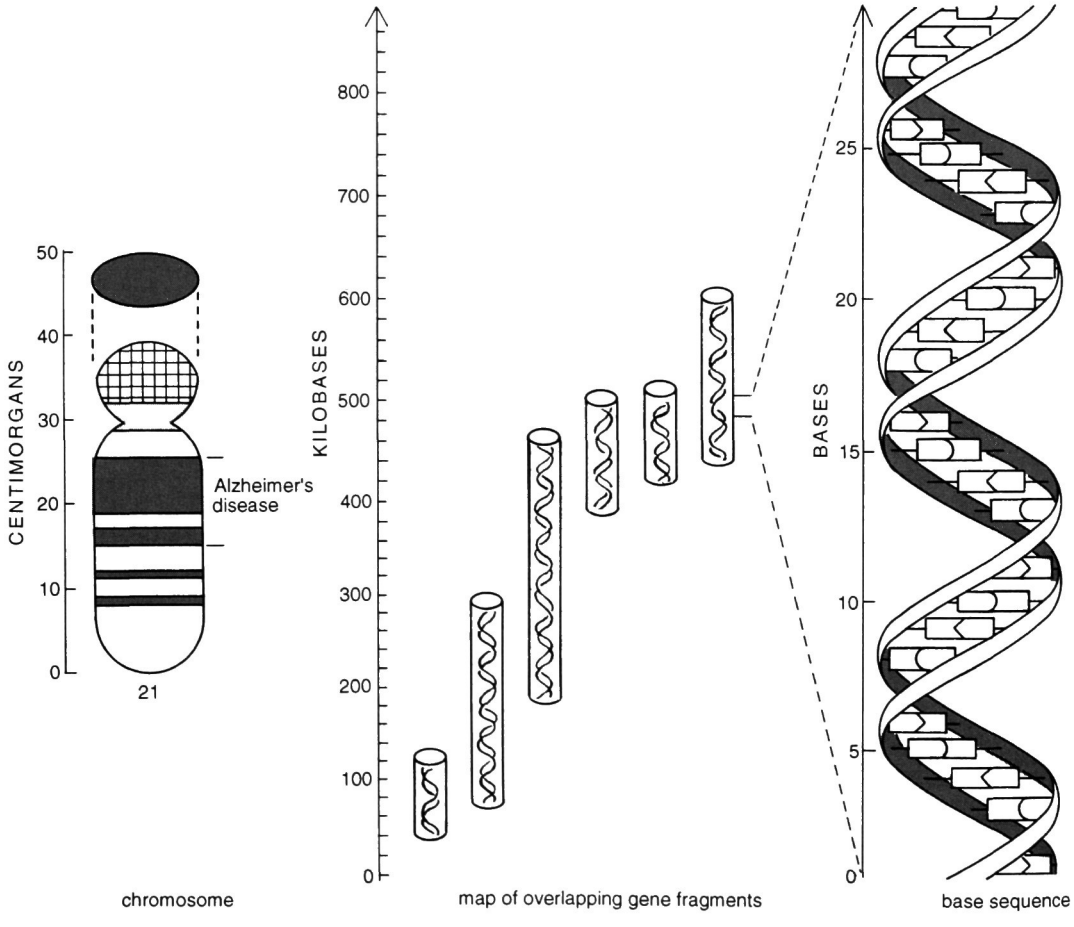
In 1987, two groups of investigators linked the inheritance of Alzheimer's disease to a region of chromosome 21.[9, 10] It appeared likely that somewhere in the DNA on that chromosome was a gene that caused the disease in some families, although apparently not the Rosses.[11–13] In 1991, several groups identified a mutation correlated with Alzheimer's disease in two families,[14; 15] but this mutation proved quite rare. It occurred in only a few of the families with Alzheimer's disease mapped to chromosome 21.[16] In other families, there is evidence of a gene on chromosome 19,[17] and in yet other families, on chromosome 14.[18] The chromosome 14 gene seemed likely to account for most cases of early-onset familial Alzheimer's disease. Alzheimer's disease has become one of the clearest examples of "genetic heterogeneity" in medicine—clinically similar disorders caused by different gene defects.[7]

The genetics of the Ross family are still obscure. The Rosses are among a

---

Geographic analogy gives a rough sense of the relative sizes of the subcellular entities involved in genome research. In the world of the cell, the information encoded in a triplet of nucleotide bases, or codon, corresponds loosely to the address of a single building.

Genetic maps differ in scale. The enlarged drawing of human chromosome 21 at left shows the region where the first linkage to Alzheimer's disease was identified in 1987. The scale shows the approximate length of this region in centimorgans, a measure of how often chromosomal segments are inherited together. (A distance of one centimorgan between genes indicates that they are likely to be separated only once in a hundred times in the process of meiosis, the kind of cell division that produces new sperm and egg cells.) The diagram at center is a physical map of a region of DNA, with an array of overlapping DNA fragments spanning the region. The scale is given in kilobases, or thousands of base pairs. Each fragment represents a length of DNA that has been cloned in yeast cells, so that large amounts can be copied and studied directly. By identifying clones containing adjacent chromosomal fragments, DNA from the region can be systematically scanned in search of genes. The diagram at right shows DNA mapping at its ultimate resolution, with the sequence of individual base pairs constituting the genetic information.



group now known as "Volga German" families. These families left Germany to live in the Volga River valley in prerevolutionary Russia. They suffered religious persecution, and many emigrated to the United States. Alzheimer's disease runs through a few of these families, which have been studied in the American Midwest. As noted above, the pattern of inheritance suggests a single gene, but none has yet been mapped.

This complex and confusing genetic story may well be a prototype of how researchers can use genetics to study neurological and psychiatric disorders. Diseases once thought to be coherent clinical syndromes may turn out to have several different causes. The clinical and anatomic similarities obscure underlying biological differences that the precision of molecular genetics can distinguish. In the case of Alzheimer's, the very tentative conclusion seems to be that there is at least one identified gene, the amyloid precursor protein gene, on chromosome 21, and perhaps another site on chromosome 21. There is strong evidence of another gene on chromosome 14, accounting for most familial cases with early onset, and a gene on chromosome 19 associated with later onset cases. There is yet another unmapped gene, or perhaps more than one, in the Volga German families. Four or more genes may cause what has until very recently been considered a single genetic form of Alzheimer's disease.

To add yet another complication, it remains unclear how many cases of Alzheimer's disease are genetic in origin, as opposed to other unknown causes (head trauma, viruses, environmental toxins, and other postulated agents). This is a matter of considerable controversy in neurology and genetics. The conventional wisdom is that only 10 to 15 percent of cases are associated with a single gene of major effect, but some researchers argue that the vast majority of cases may actually be due to genes whose effect is obscured because the disease begins so late in life that many affected people die before they develop symptoms.[19;20]

In many cases, some would say a majority, genes may be far less important, perhaps even irrelevant. In the absence of firm data, opinion runs rampant, and no one opinion is any better or worse than the others until it is proved right or wrong. The hope is that genetic studies will provide a tool to trace the causal path, leading to further progress and suggestions of other possible causes. The ultimate goals are prevention, treatment, and possibly even cure.

Molecular genetics is a short cut to understanding mechanism through structure. The great appeal of the genetic approach is its immense explanatory power. If a gene is part of a causal chain, then there is something concrete to study—how the gene turns on and off, what it produces, what the gene product might do. That is why it is so attractive to study Alzheimer's disease, cancer, arthritis, diabetes, and other major killing and disabling diseases through genetics. Genes do not do everything, and the genetic approach must be wedded to biochemistry and physiology to complete understanding of a causal chain, but molecular genetics has been advancing more rapidly than these other fields.

Technology emanating from molecular genetics will continue to shift the conceptual foundations of biology and medicine toward the study of DNA.

A familial form of Alzheimer's disease was known in the 1930s, but success in finding even an approximate chromosomal address for a gene causing it came only fifty years later. The discovery was made possible by new tools and techniques. Success awaited the construction of genetic maps, sets of markers on all the human chromosomes that could be used to trace the inheritance of regions of chromosomes through families. As the methods of molecular biology became more powerful, they were applied to problems of increasing scale and complexity. In the 1980s, a group of scientist-administrators independently spawned the idea of systematically mapping the human chromosomes and spelling out the molecular detail of the DNA they contain. The idea for a genetic map of the human chromosomes combined with technological innovations in other fields and eventually jelled into what became known as the Human Genome Project. The assault on genetic disease was but one of many historical roots of the genome project, but it was this root that lent the project special urgency. Alzheimer's disease was but one of hundreds of scourges at which molecular genetics took aim.

The Human Genome Project emerged as a unifying force, focusing the full intensity of molecular biology on the development of tools to crack open the diseases that eluded understanding. The tools were maps and methods; the genome project was a political package in which to present them to policymakers and the public.

2

Mapping Our Genes

BETWEEN THE TIME I first met Robert Ross and his death, a revolution began in human genetics. The manifesto of this revolution was a 1980 joint paper published in the *American Journal of Human Genetics*.[1] The paper, by David Botstein of the Massachusetts Institute of Technology (MIT), Raymond White and Mark Skolnick of the University of Utah, and Ronald Davis of Stanford University, proposed a systematic approach to finding and organizing markers on the human chromosomes. A map consisting of such markers spaced throughout the chromosomes could then be used to locate genes by correlating the inheritance of the markers with the inheritance of traits (including genetic diseases) in families. Human geneticists could trace the inheritance of small chromosomal fragments through families for the first time. The 1980 paper drew on techniques first developed for yeast genetics and extended ideas just then emerging in human genetics. The work leading up to this landmark 1980 paper began in 1978, when yeast geneticists Botstein and Davis were presented with a novel problem in human genetics.

David Botstein had a background eminently suitable for making this conceptual breakthrough. He had trained initially at the University of Michigan, one of the world's centers of human genetics after World War II, and had gone into yeast genetics at MIT. He was therefore imbued with human genetics, but worked on one of the most genetically tractable organisms, the source of many new molecular genetic techniques. Ronald Davis was long known to have a flair for devising new technologies in molecular genetics. Any time there was a new way to cut, insert, separate, purify, or otherwise manipulate DNA fragments, Davis's laboratory at Stanford was likely to be involved. Botstein and Davis crystallized a molecular marking strategy out of a complex brew of methods for DNA analysis.

The strategy of narrowing the region in which to search for genes was long used in experimental organisms, where controlled breeding and powerful genetic techniques simplified the task. The conceptual breakthrough in 1978 was to show how existing methods could be applied to the *human* genome. Botstein and Davis's technique was the conceptual engine that drove human genetics from the era of the horse-drawn carriage into the age of the automo-

bile. With these new maps, studying families like the Rosses changed from an improbable quest fueled by hope to a simple matter of persistence. To the extent that particular genes caused disease, their technique was a reliable way to find them.

Genetics became a major scientific field early in the twentieth century. As early as 1865, the Austrian monk Gregor Mendel had noted that the simplest way to explain the inheritance of certain characteristics of peas and other plants was to postulate "factors" donated by each parent. Exactly what physical structure conferred inheritance was not immediately clear, however. Chromosomes were first observed inside cells in 1877. Walter S. Sutton, a medical student working with Edmund B. Wilson at Columbia University, proposed in 1902 that chromosomes carried Mendel's hereditary factors.[2-4] Three years later, Nettie M. Stevens of Bryn Mawr College, also working with Wilson, explained how factors on the X and Y chromosomes could explain the inheritance of gender, independently corroborating her earlier work on insects.[5;6]

In 1906, the English scientist William Bateson, a champion of Mendelism, christened the study of inheritance "genetics."[5;7-9] In an independent coinage, Mendel's hereditary factors became "genes."[10] Thus, by 1910, the field had a name, and specific elementary objects to study.

Mendel's work, published in 1866, was largely ignored for thirty-five years, not because it was obscure or unavailable, but because its relevance to the dominant biological controversy of its day—evolution—was not immediately apparent.[5;9] His work was rediscovered independently in 1900 by three scientists in Holland, Germany, and Austria,[11;12] spawning the birth of Mendelian genetics. Genetic mechanisms to explain variation among generations immediately became the focal point in a protracted dispute about mechanisms of evolution. The controversy was ultimately resolved with the emergence of theoretical population genetics in the 1920s and 1930s. This new field combined statistical analysis of variations with the study of inheritance to explain how small genetic changes—mutations—could work with natural selection to explain evolution.[9]

Genetics grew rapidly with the study of the fruit fly *Drosophila melanogaster* and other species. Thomas Hunt Morgan, first at Columbia and then at the California Institute of Technology, blazed a path through the chromosomes of *Drosophila,* creating the paradigm for genetics in other organisms. The idea of looking for clusters of genetic traits, or characters, that were often inherited together—a phenomenon called genetic linkage—emerged from this group in a series of brilliant investigations.[12]

Applying the concepts of heredity emerging from the study of plants and other organisms, the British physician Archibald Garrod laid the foundation for medical genetics during the first decade of the twentieth century. At the Hospital for Sick Children in London, he studied the disease alcaptonuria. This condition caused a child's urine to turn dark, and later resulted in discolored cartilage and arthritis. Garrod studied the chemical abnormalities of the

affected children's urine, finding an excess of homogenistic acid, a metabolic by-product. This by-product accumulated like water behind a dam because the enzyme did not function. After conferring with Bateson, Garrod deduced that the inheritance of alcaptonuria could be explained by Mendel's hereditary factors; the enzyme defect must be determined by a gene.[13;14] Garrod further generalized his theory to the notion of "inborn errors of metabolism" for many other disorders.[15] Garrod thus established a firm link between genes and some human diseases.

Human gene mapping began in 1911, when Morgan's Columbia colleague Edmund B. Wilson deduced that the gene for color blindness must lie on the X chromosome because of its distinctive pattern of inheritance—fathers did not pass it on to sons, and it was rare among women.[16] The X chromosome could be distinguished by its size; females had two copies and males only a single copy. For five decades, study of the characteristic inheritance patterns of X-linked disease remained the most reliable gene-mapping method.

The first mapping of a human disease trait on another chromosome was published in 1968.[17] Genetic linkage to a human nonsex chromosome was first established in 1951,[18] but the nonsex chromosomes could not then be readily distinguished, so just which chromosome contained it was not known. It was possible to distinguish nonsex chromosomes in an occasional family when a particular chromosome had an unusual shape, or when chromosome fragments were rearranged and caused detectable clinical features, but vast regions of chromosomes other than X and Y resisted mapping. (At the time, geneticists erroneously believed there were forty-eight human chromosomes, further evidence of the technical limitations of the day.) In the late 1960s, two technical developments freed mapping from dependency on rare anomalies.

Somatic-cell hybridization mixed chromosomes from different organisms, fusing together cells from humans and other organisms. The mixed chromosomes fragmented and reorganized into metastable cell lines that retained various amounts of human DNA. It turned out that most rodent-human cell lines, after a few generations, kept mainly rodent DNA and only a small amount of human DNA, and were relatively stable over time.[19] If two genes were located near each other on the same chromosome, they would be expressed together in hybrid cell lines. By assembling large numbers of such cell lines, and devising ways to select only those cells containing genes of interest, it became possible to map genes by finding which genes were expressed together from different bits of chromosomes.[20] This was a laborious way to study the linkage of different known genes, suggesting their location near one another.

Linking genes to one another did not necessarily mean knowing which chromosomes contained them.[21] The largest and smallest chromosomes could be distinguished, but a large group of intermediate size generally could not. Geneticists needed a method to distinguish *all* of the chromosomes. Torbjörn

O. Caspersson of Copenhagen led the way, using fluorescent dyes and a microscope. Caspersson and others found that chromosomes could be distinguished by staining with DNA-binding fluorescent dyes, building on a line of work dating back to the 1930s.[22–25] These dyes did not uniformly stain the chromosomes, but they gave each chromosome pair a distinctive set of bands. Suddenly, geneticists could tell chromosomes apart. The same banding techniques detected deletions, rearrangements, and duplications of chromosomes.

Somatic-cell hybridization could thus look for expression of gene function, and chromosome banding allowed identification of each chromosome. Somatic-cell hybrids and chromosome banding launched human genetics on its quest for a complete gene map.[8]

The new techniques were considerable advances, but they still could not chart the chromosomes with sufficient precision to find individual genes, except in unusual cases. Changes in chromosome structure large enough to be visible under the microscope encompassed millions of base pairs of DNA, and such visible changes were rare. More subtle changes in DNA escaped detection. Only occasionally could a disease state be correlated with such detectable changes in chromosome structure. The techniques relied on gross structural changes and on expression of known gene products. They generally could not be used to locate genes of unknown function or to map new genes systematically.

The next advances grew out of the increasing power of molecular genetics. Molecular biology was largely a post–World War II phenomenon. Its two seminal events took place a decade apart. In 1943, Avery, MacLeod, and McCarty's discovered that DNA was the "transforming principle," conferring heritable traits from one bacterium to another.[26] This suggested strongly that DNA was the stuff of genes. In 1953, Watson and Crick revealed the double-helical structure of DNA.[27] This showed immediately how genes could be reliably passed from one cell to another by faithful copying of DNA. It also opened up an entirely new field devoted to understanding how genes guided cellular function.

The distinctive signature of molecular biology was to understand function through molecular structure. This "reductionist" conceptual strategy was borrowed from physics.[28] Early progress in molecular biology moved fastest in the study of bacteriophages, small viruses that infected bacterial cells. Beginning in the 1960s, however, molecular biology invaded field after field, applying its increasingly powerful tools to questions of greater complexity. In the middle to late 1970s, for example, molecular genetics was applied with astonishing success to the study of cancer, culminating in the discovery of oncogenes, or genes associated with cancer.

The first disease characterized at the molecular level was sickle-cell anemia. In 1949, genetic studies by James Neel at the University of Michigan showed

it was a recessive genetic disease;[29] biochemical studies by Nobel chemist Linus Pauling, at the California Institute of Technology, revealed structural changes in hemoglobin.[30] It seemed inevitable that Neel's gene would be causally linked to the protein defect. In the mid-1950s a change in one of the protein chains of hemoglobin was found,[31] suggesting a mutation in DNA encoding the protein.

Until the past few years, most of the tools of molecular biology followed this general outline, studying individual genes one at a time, starting with biochemical analysis of a gene product (the functioning protein). Applying molecular techniques to chromosome mapping pushed the techniques of molecular biology at both ends—mapping to the level of individual DNA base pairs at one end, and isolating and analyzing DNA fragments millions of base pairs in length at the other. The idea of a complete chromosome map also opened up the prospect of finding new proteins through the study of inheritance, rather than finding genes associated with known proteins, thus reversing the traditional gene-hunting strategy.

The prospects for finding unknown human genes began to brighten considerably in 1978, as molecular biology attained sufficient power to address problems in human genetics. Techniques of molecular genetics developed in the mid-1970s laid the foundation for a new kind of genetic map. In 1970, enzymes that cut DNA at specific base sequences were discovered. These quickly became precise tools to investigate DNA structure.[32;33] A highly reliable way to separate DNA fragments according to their length was another major innovation of the early 1970s.[34] Two groups of investigators independently discovered how to label short stretches of DNA with radioactive phosphorus to detect specific DNA sequences,[35;36] opening the way for a form of chromosome mapping.[37] In an early application of the technique, a fragment of DNA made from a hemoglobin gene distinguished just those fragments of DNA that included parts of the gene from among thousands of DNA fragments. A gene could thus be fished out of a sea of DNA.

The pieces were in place to construct a map of the human chromosomes, but learning how to combine the various techniques fruitfully required further insight. A series of papers on experiments with viruses and yeast showed how genetic linkage mapping might be accomplished in humans. DNA-cutting enzymes were first used to track genes in viruses.[38] DNA variations were used to show how a family of related yeast genes clustered together, most likely in a single chromosomal region.[39] In a series of experiments that presaged the idea of the human genome project, yeast DNA was cut into fragments using restriction enzymes, which recognized specific DNA sequences. The DNA fragments were then separated according to length, and probed with gene sequences to sort out which DNA fragments contained specific genes.[40] DNA fragments containing two normal genes were compared with a mutant version of the

same gene to find a change in a single DNA base pair. A gene was thus located by linkage to DNA marker variants, isolated by mapping the region in detail, and its mutant identified by looking at DNA sequence.

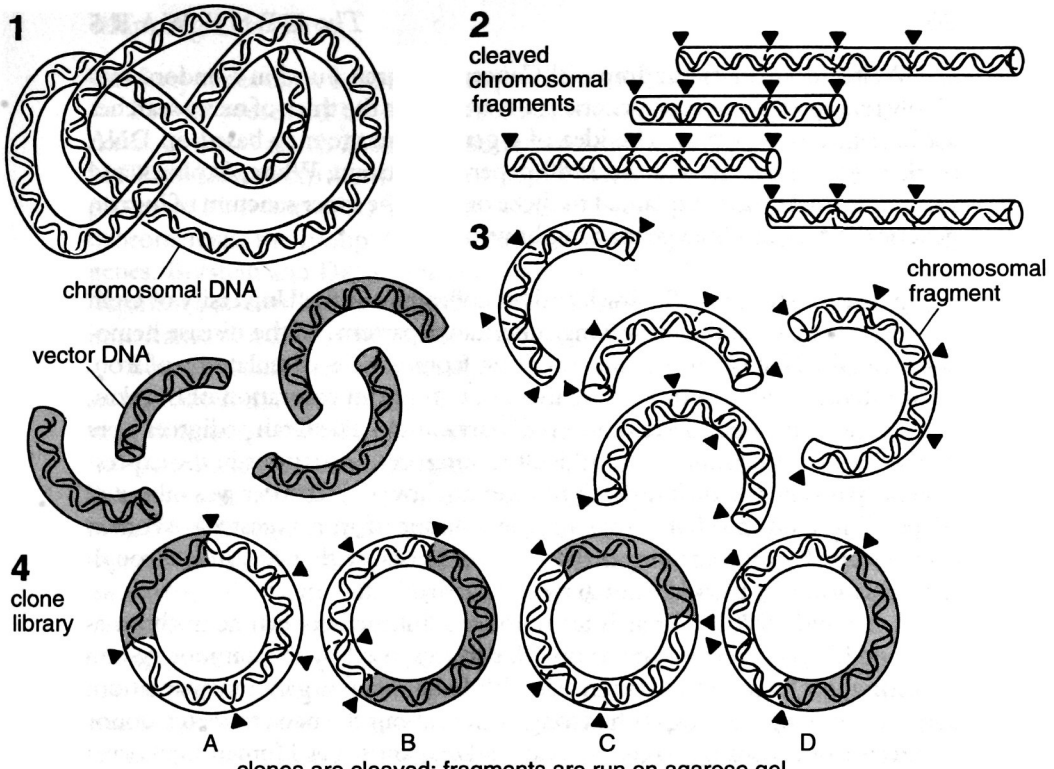
As the genetic studies of yeast and other organisms progressed, new techniques also fed a growth spurt in the emerging specialty of medical genetics. According to a leading authority in the field, the ability to distinguish chromosomes and to map at least some genes "gave the clinical geneticist his (or her) organ; just as the cardiologist had the heart, the neurologist the nervous system, the gastroenterologist the GI tract, the clinical geneticist had the genome."[41] Clinical genetics focused on lesions of the genome just as surgeons dealt with tumors or cardiologists dealt with damaged heart muscle. The first item on the agenda of clinical genetics was to define the lesions and characterize the diseases they caused.

In 1978, Yuet Wai Kan and A. M. Dozy at the University of California, San Francisco, found that a particular variant was commonly associated with the sickle-cell gene in families of North African origin. In 87 percent of cases, cutting DNA with a restriction enzyme generated a DNA fragment of a distinctive length, associated with the sickle-cell gene.[42] The sequence difference detected by the restriction enzyme was not in the gene itself, but could nonetheless mark the chromosome whence it came. In those families of Northern African descent, it was possible to tell whether a child inherited the chromosome associated with sickle-cell variant or the chromosome usually carrying a normal gene. If children had only the sickle-cell variant, then they were likely to have gotten the gene from both parents, and prone to develop sickle-cell disease. By tracking a DNA marker, one could indirectly track the gene in those families where the DNA sequence variants held. The marker variant, itself of no functional significance, was used to establish linkage with the mutant sickle-cell gene. Kan and Dozy noted that such markers could be quite useful for determining linkage with genes.

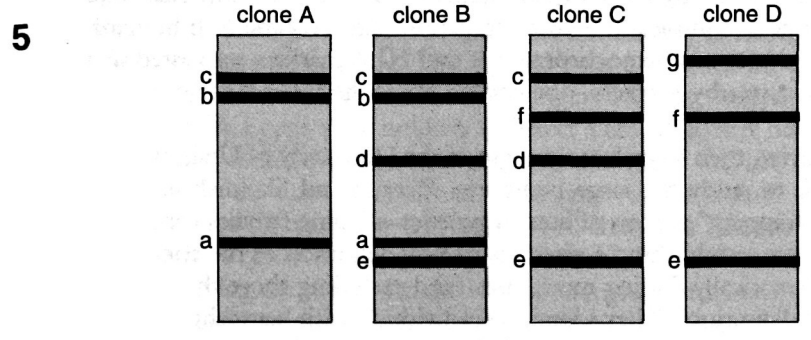
Two British groups also noted that normal variations among individuals could be used as markers to trace inheritance, and for linkage to genetic diseases and other traits.[43;44] Alec Jeffreys at the University of Leicester was most interested in studying variations among human populations; Ellen Solomon

---

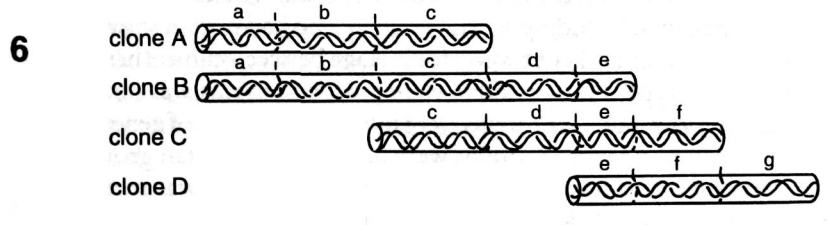
Physical mapping of chromosomal DNA is done in stages. First the DNA is cleaved into various lengths by special enzymes called restriction enzymes, which cut the DNA at specific sites (1). The resulting fragments are then combined with other fragments of DNA (vectors), typically forming circular loops of DNA that can be cloned, or copied in large numbers, in yeast or bacteria (2,3,4). The collection of cloned fragments, known as a clone library, provides a large enough supply of the original chromosomal DNA to analyze directly. The clones are next cleaved by enzymes and the fragments are separated according to size by running them through an agarose gel (5). By looking at the fragments common to different clones, researchers can piece together the original order of the fragments in the chromosomal DNA (6,7). A complete physical map can be assembled by correctly ordering the overlapping DNA fragments from one end of the chromosome to the other.

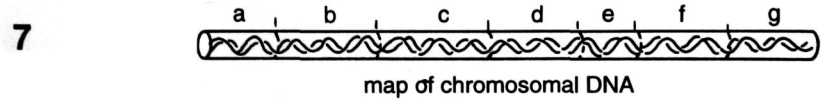


clones are cleaved; fragments are run on agarose gel



data are analyzed and order of genomic clones established





map of chromosomal DNA

and Walter Bodmer at the Imperial Cancer Research Fund in London, two highly regarded population geneticists, were also in the thick of many searches for human disease genes. The idea of a genetic linkage map based on DNA markers was in the air, and the 1980 paper by Botstein, White, Skolnick and Davis was the key that explained to those outside the inner sanctum of human genetics how such a linkage map might work.

In April 1978, Mark Skolnick and his colleagues at the University of Utah were trying to solve the mystifying inheritance patterns of the disease hemochromatosis. Hemochromatosis, resulting from toxic accumulations of iron, most often caused cirrhosis of the liver, orange-green coloration of the skin, diabetes, and insidious deterioration of heart muscle. The Utah pedigrees were meticulously documented, but difficult to interpret. Variations in the expression of symptoms made it difficult to ascertain how the disorder was inherited (depending among other things on the amount of iron ingested). Women often escaped symptoms longer than men did because they lost iron through menstruation (the blood lost each month carried iron with it).

In the mid-1970s, a French team noted a linkage between hemochromatosis and highly diverse proteins on cell surfaces, used by the immune system to distinguish "self" from foreign cells.[45–47] In studying organ transplantation, surgeons and immunologists had long known about the need to match donor and recipient according to cell surface markers, known as Human Leukocyte Antigen (HLA) types, to avoid organ rejection. HLA surface proteins were encoded by a gene complex known to reside on chromosome 6 in humans. The association between hemochromatosis and HLA markers suggested that by studying the nearby markers, one could dissect the inheritance of hemochromatosis.

Kerry Kravitz, then a graduate student at the University of Utah, worked with Skolnick to study the large pedigrees. Kravitz and Skolnick used the classic "boot-strapping" process of human genetics—finding families with many cases of a disease, establishing a rigorous clinical definition of the condition, and then systematically finding more cases (and excluding those that did not fit the clinical definition). They identified individuals with hemochromatosis, then tested iron levels in the patients' relatives, thus identifying new cases that had not been diagnosed. Better clinical information allowed another round of case finding, and so on. This had the added benefit of enabling treatment for the newly identified patients, including several women who had not yet shown symptoms. Kravitz and Skolnick kept testing for linkage between clinical hemochromatosis and HLA type. HLA markers were proteins, so this was not a case of direct DNA analysis, but the proteins were indirect indicators of genetic diversity. (The genes coding for the proteins were different.) The Utah group eventually accumulated enough cases to conclude that the disease was recessive—it required two copies of the gene, from both father and mother, to develop the disease.[48] Further analysis suggested ominously that the disease

was forty times more common than previously believed.[49]

Kravitz presented his initial results at the annual review of graduate students, at the Wasatch Mountain ski resort Alta in April 1978. Botstein from MIT and Davis from Stanford were invited as outside reviewers. After Kravitz presented his results of statistical associations between HLA variants and hemochromatosis, the group discussed how to use statistical associations to map genes. Botstein and Davis, both of whom were familiar with how restriction fragment patterns had been used in yeast, supported the notion of using coinheritance of a disease and some nearby marker as a mapping tool.

Botstein tends to think and talk excessively fast, and often at the same time. In one of his characteristic verbal explosions, he realized that correlating genetic differences with disease—the general approach used by Kravitz to track hemochromatosis—could be generalized and made much more powerful by direct analysis of DNA variations, using techniques Botstein and Davis were both familiar with in yeast.[50–55] If there were enough differences among individuals in families, the technique could be used to locate genes by dint of inheritance alone, with no knowledge of gene function and no particular candidate genes in hand.[56]

Botstein later recounted a vivid memory of looking up at Davis, both knowing that this was a conceptual breakthrough.[53–55] If one could only find enough genetic linkage markers spanning the chromosomes, a full-blown map should be possible. Genetic linkage in humans had been sporadically successful. The only markers available were generally genes, and these were usually not variable enough among individuals within a family to be able to trace their inheritance unambiguously. This limited their usefulness for genetic linkage analysis. The dearth of good markers hampered linkage analysis. Investigators searching for a gene were unlikely to detect a linked marker because the odds of finding a variable marker near the gene of interest was low. It was a crapshoot, with the odds stacked heavily in favor of the house. Moreover, the markers were distant from one another, so that their relative order and the distance between them were hard to determine. This complicated the process of determining the size of the chromosomal region containing a gene, and the gene's orientation with respect to different markers, even if a nearby marker was linked to the gene. The 1980 paper noted that "no method of systematically mapping human genes has been devised, largely because of the paucity" of markers that varied frequently among individuals.[1] Finding a large collection of such markers dispersed throughout the genome would be an enormous task, but it should theoretically work. Once these markers were ordered relative to one another, they could anchor a map, and be used to search for genes expeditiously, *even if one knew nothing more than the pattern of inheritance in a family.*

The importance of large families was the reason Skolnick was in Utah in the first place. Mormon families trace their pedigrees in great detail for reli-

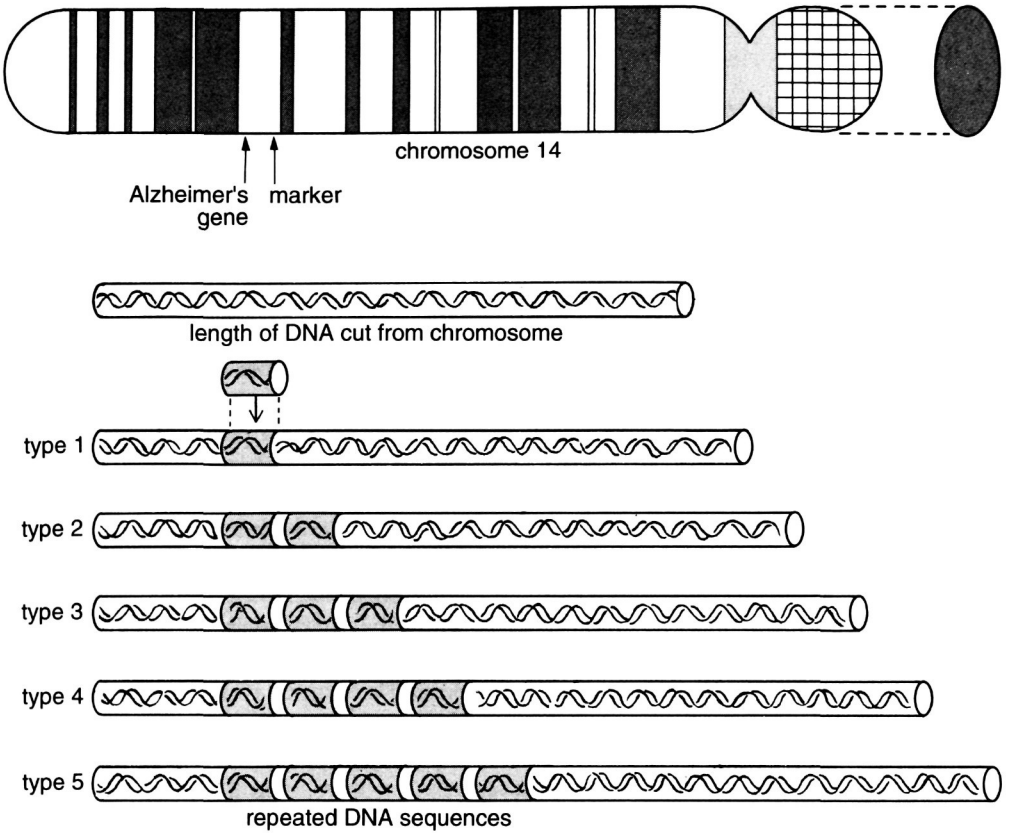
gious reasons, searching for distant relatives who can be guided to salvation. The Church of Latter-Day Saints supports an elaborate research center for genealogy. Members of the church tend to have large families, again to increase the number of those redeemed by the faith. The scrupulous attention to family history has created a gold mine for human genetics. Indeed, something like mining is involved, as some of the genealogical records are carefully preserved in a mountain vault, to prevent their destruction in the event of war.

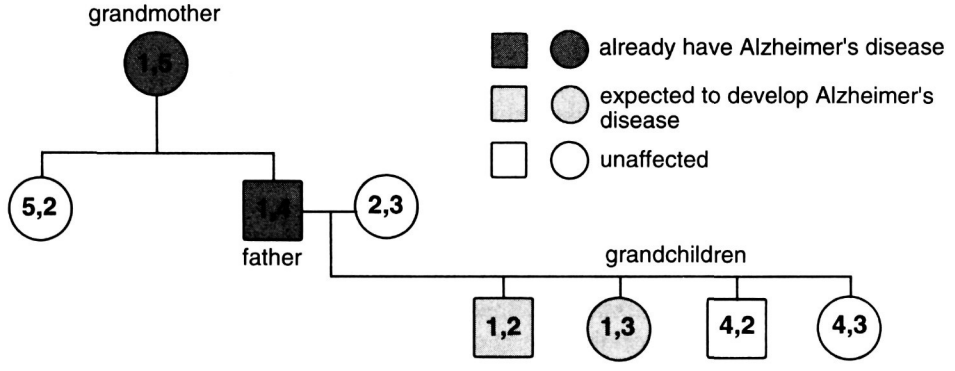
Skolnick set out to computerize large pedigrees for genetic analysis. He and others devised computer algorithms to do the tedious computations of probabilities for genetic linkage. He recognized the importance of Botstein and Davis's discussion, but he was not a molecular biologist. Botstein and Davis were intrigued, but they had many other projects more directly related to their past interests. The idea of a concerted effort to construct a genetic linkage map of DNA markers awaited another investigator's initiative.

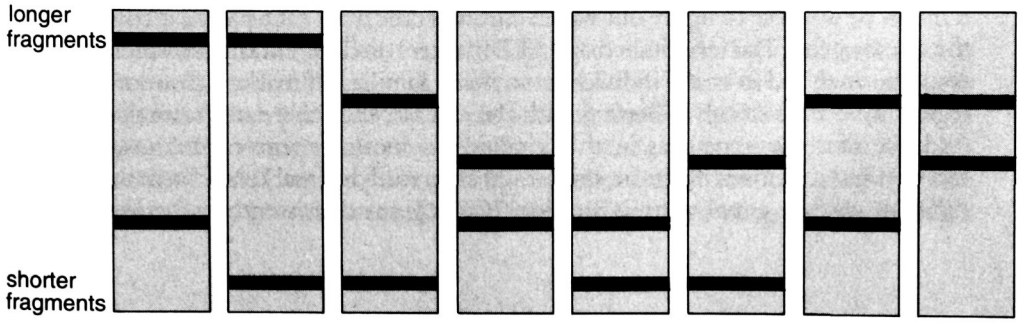
Enter Raymond L. White. White, who was then at the University of Massachusetts at Worcester, came to the project that would establish his scientific reputation through a triangular MIT connection. White had been a graduate student with Maurice Fox at MIT. Soon after the fateful 1978 meeting at Alta, Fox bumped into Skolnick at a National Institutes of Health (NIH) meeting in Bethesda, Maryland, on breast cancer. Skolnick conveyed Botstein's idea to Fox, who then spoke with Botstein when he got back to MIT. Fox called White to tell him about Botstein's idea. White, becoming claustrophobic about prospects for a career studying the genetics of blowflies,

---

DNA Marker near the gene for Alzheimer's disease on chromosome 14 can be used to trace the course of the disease through a family. As the drawing at the top of the opposite page shows, the marker is not coincident with the Alzheimer's gene, but is close to it and thus is highly likely to be inherited with it. In this hypothetical family, there are five variations of the DNA marker. Repeated DNA sequences have been inserted one, two, three, four, and five times into DNA fragments cut out of each family member's DNA, differentially increasing the length of the respective fragments. Since each person inherits two copies of chromosome 14, he or she can have makers of two different lengths (known technically as restriction fragment length polymorphisms, or RFLPs). In this case, the grandmother exhibits RFLP pattern 1,5 and has Alzheimer's disease. Her son shows pattern 1,4 and is also affected. This indicates two things: (1) that he inherited a copy of chromosome 14 with pattern 1 from his mother and pattern 4 from his father; (2) that Alzheimer's disease is associated with pattern 1, since this son must have gotten the disease along with pattern 1. The unaffected sister in this generation inherited the other copy of chromosome 14, bearing pattern 5 from her mother and pattern 2 from her father. With this information, it is possible to predict that two siblings in the next generation are likely to carry the Alzheimer's gene, and thus to develop the disease (if they live long enough), because they also inherited pattern 1, while the two other siblings will be spared. The use of genetic markers such as RFLP's to track which copy of a chromosome was inherited by different family members depends on finding a marker that differs among family members. It was by repeating this process in many different families, and in different cases in the same family, that investigators were able to build up the evidence that a specific region of chromosome 14 is associated with Alzheimer's disease. The precise location of the gene and the nature of the gene defect can be determined only when the gene is found in the region.


chromosome 14
Alzheimer's gene
marker
length of DNA cut from chromosome
type 1
type 2
type 3
type 4
type 5
repeated DNA sequences


grandmother
1,5
5,2
father
2,3
already have Alzheimer's disease
expected to develop Alzheimer's disease
unaffected
grandchildren
1,2
1,3
4,2
4,3


longer fragments
shorter fragments

was looking for an exciting new project to pursue. Botstein had severe space constraints at MIT, and could not immediately embark on a new venture.[57] White called Botstein, with whom he had worked for a year while Fox was on sabbatical; White had found his project. They began collaborating on a research proposal to seek grant funds from NIH.[52–54; 58; 59]

White took the lead drafting a proposal to seek NIH grant funds. The application was sent to NIH on February 27, 1979. Its first goal was "to develop a new set of genetic markers for the human genome based on DNA restriction fragment length polymorphisms (RFLPs)."[60] These RFLPs were markers of human genetic variation that could be used to trace the inheritance of chromosome regions through family pedigrees. The grant proposal mentioned Skolnick's HLA linkage work, and recounted the prior use of RFLPs in analysis of viruses and yeast. The proposal then laid out the great advantages for studying human inheritance if "one does not have to 'isolate the gene' in order to do mapping." By tracing the inheritance of naturally occurring variants, one could look for associations with disease genes, without knowing which protein the disease gene produced. An RFLP map would be just the tool needed to solve the biggest problem confronting most genetic diseases—to find unknown genes. If enough RFLPs could be mapped to the chromosomes, they would "provide a new horizon in genetics. . . . If successful, this endeavor will transform research with a low probability of success—familial linkage studies—into a legitimate endeavor. . . . the new tool will also [permit analysis of] syndromes which seem to 'run in families' but are too difficult to characterize genetically."[60]

White got his funding to start work at Worcester, and by November 1979, White and his colleague Arlene Wyman had isolated their first human RFLP. The probe from the first RFLP was named pAW101 (for "plasmid, Arlene Wyman 101").[61] The RFLP showed almost as much genetic variation as the HLA locus, making it likely that the two chromosomes in any given person would often be distinguishable, and that parents would have different markers as well. This heterogeneity made it a wonderful tool to find a gene near it on the tip of chromosome 14, where Wyman eventually mapped it.

If RFLP differences similar to the one associated with the sickle-cell gene could be found throughout the chromosomes, then the inheritance of chromosomal regions could be similarly traced through families and correlated with the presence or absence of disease. If an unknown disease gene, say the one causing cystic fibrosis, was located near the site detected by an RFLP, then it might be possible to figure out which chromosome from each parent carried the disease gene. The inheritance of RFLP markers on different chromosomes could be analyzed in many individuals in many families. If markers from one region were consistently inherited with the disease, then there was statistical evidence that the gene was in that region. To locate a gene of unknown location and unknown function, one would keep studying markers from many different chromosomal regions and look for regions consistently associated

with the disease. Given enough markers, one would eventually find one near the gene. Having found the approximate location on the chromosomes, DNA from that region could be isolated and studied in greater detail in hopes of finding the mutation itself—the DNA change that caused a genetic disease. The concept was elegant and powerful, and the first step was to develop enough RFLP markers.

White soon moved to Utah, to take advantage of the Mormon pedigrees and because the Howard Hughes Medical Institute was willing to give him substantial funding to construct a map. Over the next few years, the Utah group systematically searched for RFLP markers, refining their techniques. They made DNA from members of more than forty families available for genetic typing by other groups, and contributed more markers to the genetic map of humans than any other group.

Another team dedicated its efforts to finding genetic linkage markers and to constructing systematically a complete genetic linkage map. Helen Donis-Keller led a group at Collaborative Research, Inc., a private firm located near Boston. During 1979, 1980, and 1981, Botstein had several conversations with NIH staff about assembling a complete RFLP map. He finally despaired of enticing NIH into the ring, but he believed an RFLP map was centrally important. He and Davis were both on Collaborative Research's scientific advisory board. White and, for an even shorter period, Skolnick were also initially consultants to the company. The possibility of RFLP mapping was much discussed among the company's scientific advisers. Collaborative Research was casting about for new markets and technical avenues, wanting to create a limited research and development partnership (a tax-favored investment tool in vogue until it was dismembered in the 1986 tax-reform law).

Donis-Keller had come to gene mapping through molecular biology. She worked as a graduate student with Nobelist Walter Gilbert at Harvard. She then moved to Harvard Medical School to work on viral molecular biology with Bernard Fields. Gilbert called her in 1981 to see if she would like to work for the newly forming biotechnology company Biogen, and she was hired as its third U.S. employee. Biogen grew rapidly, and working there became chaotic. Donis-Keller left Biogen for Collaborative Research in the spring of 1983. Nobelist David Baltimore headed the scientific advisory group to Collaborative Research. He, Botstein, and others prevailed on Donis-Keller to join Collaborative Research, to do strategic planning. Genetic linkage mapping with RFLP's became one of several projects under discussion as priorities for the company.

Donis-Keller applied for an NIH Small Business Innovation Research award, but was told the work was "not sufficiently innovative." She and James Wimbush then went to Wall Street, looking for $50 million. They came close to securing venture capital under a limited-partnership arrangement, but ultimately failed. They did presentations for Johnson & Johnson, Union Carbide, and other large companies. Donis-Keller laid out the strategy of completing a

genetic linkage map and using it to locate genes, thereby making available a new method to detect genetic conditions. Having the tools for genetic linkage would not only spin off diagnostic tests, but would also give the company an edge in finding genes more quickly. Other methods could be used to find the gene itself, providing a target for drug development and gene therapy. Donis-Keller and the upper management at Collaborative Research were repeatedly rebuffed on Wall Street. She was appalled at the lack of vision among American corporations. They simply did not believe that genetics would be critical to cancer, heart disease, or other major health problems that bred diagnostic and therapeutic markets.

In 1984, Tom Oesterling became president of Collaborative Research, and the company decided to go ahead with RFLPs, using internal funds. The genetic linkage mapping team at Collaborative Research grew from four in April 1983 to twenty-four by the end of 1984. Work progressed steadily for three years. By the summer of 1987, things began "to come together."[62] The Collaborative Research team decided to push for a complete genetic linkage map in time for the human gene mapping conference in Paris that September. Rumors that the Utah team was planning to publish such a map were rampant, and they spurred Collaborative Research's efforts. Jean-Marc Lalouel, from the Utah Group, told one of the company's researchers at the Paris meeting that Utah had "won the war," and the Boston group feared it might be true.[63] The Utah group did distribute a pamphlet at the Paris conference, but the Donis-Keller team had submitted their publication to the journal *Cell* just before leaving for Paris.

Collaborative Research's announcement of a genetic linkage map caused quite a stir in the fall of 1987.[64–66] The corporate office decided to hold a press conference at the American Society of Human Genetics meeting in October, and announced the impending publication of a genetic linkage map containing markers from the public domain and from the company's own collection.[62;67] Following the press conference, other groups opened fire. The map was said to be incomplete and the spacing between some markers was indeterminate, but the group at Collaborative Research calculated that their map was sufficient to locate 95 percent of any new genes and markers. White reported that the Utah group would publish maps one chromosome at a time when there were no gaps. Some of the controversy derived from the fact that about a fourth of the markers in the map had been discovered elsewhere, and the family resources necessary to do the mapping were contributed by a variety of other groups. Someone had to publish the first map, however, and no one group could ever claim full credit for the pooled resources. Ambivalence about a for-profit company's sponsorship of work in the *Cell* paper further complicated professional rivalries that were already intense. Despite the professional tensions, or perhaps abetted by them, the genetic linkage map beginning to coalesce around the efforts of the groups at Utah, Collaborative Research, and elsewhere became an enormously powerful tool.

The number, pace, and scale of hunts for human disease genes increased dramatically in the late 1980s. RFLP mapping reached a fever pitch, often flashing a sharp, competitive edge. In musing on the history of their field, geneticists James Crow and William Dove compared the 1987 "map flap" with publication of the first genetic linkage map, Alfred Sturtevant's 1913 paper on *Drosophila:* "These quiet beginnings stand in abrupt contrast to the current hubbub over the human linkage map and the proper definition of a map. With its rival factions and the glare of publicity, the mapping race is almost a genetic Olympics."[68] Crow and Dove betrayed a tinge of nostalgia, even tacit disapproval, of the style among the new upstarts. But the world had changed.

With maps in hand, gene hunts became highly competitive races. Teams led by Francis Collins and Ray White crossed the line to the neurofibromatosis (type I) gene in a dead heat, and leveraged favors out of editors eager to publish hot new findings. Their articles came out the same day in *Science* and *Cell*[69] The glare of publicity made genetic linkage mapping and gene hunting a high-stakes game, a national sport with guaranteed coverage in the *New York Times* and other major newspapers. Within a decade, genetic linkage mapping had gone from stepchild to celebrity, a scientific Cinderella story. But Cinderella needed an escort to protect her from her stepsisters.

The glue holding the various genetic linkage efforts together, despite the rivalries and tensions, was the Centre d'Etude du Polymorphisme Humain (CEPH), a Paris organization founded by Nobelist Jean Dausset with funds from a scientific award and gifts from a private French donor.[70;71] CEPH was formed with the express purpose of enabling groups to pool their efforts in constructing a complete genetic linkage map. The first meeting to piece together the coalition took place in November 1984, when major mapping groups from Europe and North America descended on Paris, or rather ascended to a hallowed terrain where common good transcended rivalry. Cell cultures from members of large reference families around the world were collected by CEPH. Twenty-seven families came from the Utah Mormon pedigrees. One family came from the thoroughly studied Amish pedigrees of Pennsylvania Dutch country. Two families were small branches of an enormous Huntington's disease cluster in Venezuela, and ten French families were contributed by Dausset. Together they constituted forty families well suited for genetic linkage analysis. The idea was to test members of the CEPH family panel and to use the resulting RFLP marker data to make linkage maps of all the chromosomes. Each group that participated in CEPH agreed to genotype all informative families in the panel (to use their markers on the DNA taken from individuals in those families) and to send the data back to a shared database.

While many of the CEPH families were initially found by studying pedigrees for specific diseases (Huntington's, bipolar disorder, and other conditions), gene-hunting was *not* the thrust of the CEPH collaboration. It was instead to orient DNA markers through systematic analysis of *reference* fami-

lies. The families were not used because their pedigrees showed the inheritance of genetic diseases, but because there were sufficient numbers of living members in well-defined pedigrees from whom DNA was readily available. Once the markers were mapped in the reference families, then the same markers could be used to hunt for genes in pedigrees containing any variety of genetic diseases or other traits. Government funding agencies appeared oblivious to this distinction between gene-hunting and systematic mapping. Map construction would take a massive effort to find informative markers, map them, and order them relative to one another. Gene hunts might eventually produce a map, but the bedrock of gene hunts was family pedigrees constructed with scrupulous attention to the accuracy of diagnosis. Those hunting for specific genes continued to be supported, and many groups hunted by finding clusters of genetic linkage markers, but those attempting to produce global genetic linkage maps got little help from the government.

The world of science was the beneficiary of the competition among Donis-Keller, White, and the other genetic linkage mappers. Between them, the Utah and Collaborative Research groups performed a great service. Wyman and White's first marker was found under a grant from the National Institute of General Medical Sciences, but for the complex and vast effort needed to find sufficient markers and to orient them on a map of the human genome, funding came almost entirely from the Howard Hughes Medical Institute and corporate sources. This government policy failure played a major role in later debates about the need for a systematic genome project. NIH's rejection of overtures to support the construction of genetic linkage maps was mentioned repeatedly in debates about whether it would not similarly spurn other mapping ventures.

Mapping by genetic linkage harked back to a style of mathematical genetics developed late in the last century and early in this century, some of which preceded the term "genetics." The approach was fundamentally classical genetics—the study of the inheritance of observable differences among individuals—supplemented by clinical observation to define the genetic characters under study, as augmented by the modern tools of molecular marking. The process relied on the mathematics of probabilities to make correlations. Those who studied evolutionary biology and population genetics immediately understood the significance of genetic linkage mapping. They were joined by a few medical geneticists comfortable with the statistical techniques of linkage.

When RFLP markers helped locate the genes responsible for Huntington's disease[72] and Duchenne muscular dystrophy in 1983,[73] human geneticists took notice. The Duchenne gene was already known to reside on the X chromosome, by dint of its inheritance pattern, but the location of the Huntington's gene was a complete mystery, and there was little prospect of finding it by traditional methods. (The Huntington's story is told in greater detail in Chapter 16.) The first spectacular successes were followed in short order by polycystic kidney disease,[74] retinoblastoma,[75] and cystic fibrosis[76–79] in 1985. These

were each major prizes, and genetic linkage mapping with RFLPs took center stage.

According to a *Newsweek* feature in August 1987, a disease a week was being mapped by genetic linkage.[80] Technical advances further extended the ability to work backward from approximate gene location, determined by linkage to a marker, to find the gene itself and identify its product (in most cases, a protein). The first successful search for a gene of unknown function, starting from chromosomal location, ended in 1987 with the cloning of a gene causing chronic granulomatous disease.[81] This was soon followed by the genes for Duchenne muscular dystrophy[82] and retinoblastoma.[75;83;84] In each of these cases, however, the gene's approximate location (on the X chromosome for chronic granulomatous disease and Duchenne; on chromosome 13 for retinoblastoma) was already known from patterns of inheritance, human-hamster cell hybrids, and the study of patients with small chromosomal deletions. RFLPs played a role in narrowing the search, but the thinness of the RFLP map and limits inherent in using only pedigree studies to locate genes required additional strategies.

The sea change came with discovery of the gene for cystic fibrosis (CF).[85–87] The CF story propelled molecular genetics to the fore. Huntington's disease was the first mapped by linkage to an RFLP marker,[72] the first great triumph of RFLP linkage. But knowing the gene's location did not lead to the Huntington's gene itself, which remained elusive until 1993. In contrast, CF was mapped by RFLP in 1985, and the gene found in 1989.

CF was one of the most common seriously disabling single-gene diseases in Europe and North America. It became the first disease for which a gene of completely unknown location was mapped by genetic linkage, and then the regional DNA studied until a gene was found and the protein product identified. The CF gene was first located on chromosome 7 by Lap-Chee Tsui of the University of Toronto, collaborating with Collaborative Research.[76;88;89] Rumors of the chromosome 7 location early in 1985 induced other groups to look intensively for other markers nearby. The Utah group and Robert Williamson's group at St. Mary's Hospital in London both found linked markers even more tightly linked to CF (meaning their markers were closer to the gene).[77;78;88;89] The CF linkage studies were highly competitive and were covered closely by the scientific press.[88;90] Competition produced quick results. By locating the CF gene in a region of chromosome 7, RFLP mapping solved one of the highest priority problems, and one of the knottiest in human genetics. The race did not stop with locating the gene, but continued with sustained intensity for four more years until the gene was isolated. Indeed, it did not end even then, as there were still prospects of gene therapy and targeted drug development to pursue.

In a special issue of *Science* published on October 12, 1990, or eleven days after the genome project officially began, scientists took stock of their accom-

plishments to date. The hugeness of the task before human geneticists was starkly apparent. There were inconsistencies in nomenclature, the genetic distances measured for the same regions differed markedly, depending on analytical assumptions, and little of the genome had been mapped in detail.[91;92] The human genome was indeed still "Terra Incognita."[93] In another special issue of *Science* just two years later, the tone was far more upbeat. A news piece noted that the genome project "hit its stride even sooner than its most ardent enthusiasts had predicted. Data are pouring out of the genome centers, new technologies are coming on line, and perhaps most notably, the first two high-resolution maps of human chromosomes are now complete."[94] *Science* also published a linkage map of the human genome incorporating more than sixteen hundred markers, including many markers of far greater usefulness than those in published in the 1987 Collaborative Research map.[95] The map's authorship also changed in interesting ways from its 1987 counterpart. The authors were referred to as the NIH / CEPH Collaborative Mapping Group, with many collaborators listed for each chromosome, and Helen Donis-Keller served as overall coordinating editor. Four weeks later, *Nature* published another second-generation linkage map, arising from the prodigious efforts of a French collaboration. This map contained 814 markers spanning an estimated 90 percent of the genome, and most of the markers were far more useful for tracing inheritance than their 1987 counterparts.[96;97]

In retrospect, the systematic search for chromosomal markers and the construction of linkage maps were among the most significant accomplishments of human genetics in the 1980s and into the 1990s. Their full impact was not to be felt for several years. Maps not only made gene-hunting easier but also opened entirely new possibilities for tracing the inheritance of multiple genes and the study of how genes in one region influenced those in another. The groups in Utah, at Collaborative Research, at CEPH, in the high-technology French collaboration, and at many other laboratories throughout the world forged a genetic linkage map of a human being, a practical tool never seen before.

Genetic linkage mappers blazed a trail for those who would use human genetics to crack the tough nuts of human disease—diseases that had resisted assault by traditional research methods. The scientific strategy to answer the question "What gene is at fault?" was beautifully laid out by science writer Maya Pines, in a report prepared for the Howard Hughes Medical Institute.[98] The construction of genetic linkage maps, ironically, got little aid at first from the government agencies charged with supporting the overall biomedical research effort. This was partly because private institutions, most notably the Howard Hughes Medical Institute and CEPH, were quicker to respond and partly because the visionaries took matters into their own hands when they encountered obstacles to government support.

The formation of CEPH proved a watershed in human genetics. Dausset's idea, the commitment of the Donis-Keller, White, and other laboratories to

share data, and the agreement to make DNA from common families generally available represented a considerable commitment from each participating group. Each family pedigree took immense effort to construct and to check. Agreeing to share access to these pedigrees entailed a degree of cooperation in map construction often overlooked in the race to find individual disease-associated genes by *using* the map. CEPH promised a coherent approach to map the human genome, and it was called a "human genome mapping project" as early as 1985.[71] The collaborative arrangement had its weaknesses and tensions, but it outlived the public clashes to unify the efforts. When NIH joined the effort in 1990, and with the emergence of a high-technology French collaboration at more or less the same time, progress was even more rapid. Despite being recognized as a genome mapping project, however, genetic linkage mapping efforts did not grow into "the" Human Genome Project. Those constructing genetic linkage maps could stake a legitimate claim on the Human Genome Project, but the bureaucratic edifice bearing that title grew from different sources, from three independent proposals to determine a reference DNA sequence of the entire genome.

Genetic linkage mapping was eventually folded into the genome project as it evolved, but it was at best an afterthought in the earliest genome project proposals. The history of the genome project would have been a more logical progression had human genetics spawned it. The focus on DNA sequencing that gave rise to the genome project was more than just a matter of emphasis—the sequencing proposals formulated in 1985 and 1986 came from a different group of individuals not directly engaged in RFLP mapping. There was some overlap of interests, but the impetus for DNA sequencing arose from those who contemplated the structural study of DNA, not from classical genetics and the study of inherited characters. The confluence of structural and classical genetics was delayed, but it was inevitable.

It was a long way from determining chromosomal location by RFLP mapping to isolating a gene. The work was tedious, methods were unreliable, and they often proved inadequate to the task. Finding genes required better methods to study DNA from a given chromosomal region. To move expeditiously from approximate chromosomal location to actual gene required a different kind of map—a physical map. Physical mapping bridged the gap between genetic linkage and DNA sequence, an important intermediate step. The techniques for making physical maps were first developed in other organisms, and only later applied to humans.

3

Of Yeasts and Worms

From the 1950s onward, and even into the 1990s, molecular biology focused its increasingly sophisticated analytical techniques on more complex organisms. Methods developed and tested on viruses, bacteria, and yeasts were exported into the study of other organisms such as mammals, including humans. The process was by no means as simple as starting on the smallest organisms, completing work, and packing up to move on to the next larger one. Molecular work on humans, mice, and other complex organisms all began soon after molecular biology was founded, but the central thrust of early molecular biology was on understanding the basic relationships among DNA, RNA, and proteins. The organisms principally used to illuminate these processes were viruses and bacteria. The philosophy was to focus on systems that could be understood mechanistically. The strategy was overtly premised on reducing life processes to molecular mechanisms.[1] Over the course of three decades, the tools of molecular biology were used more and more to dissect the biology of progressively larger genomes: from viruses to bacteria to yeasts to multicellular organisms.

The study of structure began with proteins and genes. Since proteins ultimately were encoded by genes, and because some of the most powerful new techniques involved the manipulation of gene fragments, the study of DNA grew in importance. Papers in a widening circle of fields used DNA at some stage of experimentation, and recombinant DNA techniques opened up entirely new ways to ask nature questions biologists had yearned to address for decades.

The logic of reductionism was pushed to its extreme in the study of several microbes. The systematic description of entire genomes began in the 1970s and accelerated into the 1980s. The structure of DNA in small viruses was described first, then bacteria were mapped. As the 1980s progressed, there was much talk about how to map the genomes of ever larger organisms, especially man.

Genetic linkage mapping was a great advance. It could determine the approximate location of a gene. Finding the gene itself, however, was a vexing

problem. The first step in this process was to fragment DNA from the region in question, make copies of the DNA segments, and reorient them so the DNA could be studied directly. The method to do this was cloning. Cloning is a way to make thousands or millions or even billions of copies of a stretch of DNA. The process starts by inserting a bit of DNA into a virus that infects bacteria. When used to clone DNA, the modified virus, shorn and sculpted to perform its copying function, is called a "vector." The vector, containing the DNA insert under study, is then put into its bacterial host, where it is copied. Bacteria grow fast, and some vectors proliferate inside each bacterium. Billions of copies of a DNA insert can be made with ease.

A Caltech-Harvard team led by Tom Maniatis of Harvard hit upon the idea of cloning *all* the DNA in the genome of an animal, by breaking the DNA into small fragments that could be individually cloned.[2] Because bacteria could be handled by the millions, it was possible to take the DNA from an organism and grind it up into fragments small enough to be copied in bacteria. Millions of different bacterial colonies would contain different DNA inserts fifteen thousand to twenty thousand base pairs in length. Collections of such colonies were called "libraries." If the process of breaking the chromosomal DNA into fragments and cloning it were nearly random, then any given short stretch of chromosomal DNA would be represented in four or five different clones in the million-clone library (for a human genome, or one of similar size). The great advantage of having such clones was that the DNA was readily available for further analysis, since it could be copied in the bacteria. By making a DNA fragment library of the human genome available, the Maniatis group advanced biomedical research in many fields a considerable distance.

With all of the DNA available in one fragment or another, how could one then find a specific gene? The problem of finding just those clones containing a gene of interest was straightforward when something was already known about the gene. Maniatis and his group, for example, had DNA probes to detect rabbit hemoglobin genes. They used these probes to "light up" the colonies that contained clones with parts of the hemoglobin protein genes.[2] They then fished out the few clones containing bits of the gene. By analyzing these fragments in detail, they could tell how the DNA fragments overlapped and could construct a map. This map was not a genetic linkage map, however, but a map that showed how DNA in the chromosome lined up—a physical map. The main difference was that the measure in a genetic map was how often two genes or DNA fragments were separated during *inheritance*. This was in turn a measure of how cells copied and passed on DNA in the production of egg and sperm cells. Distance on a physical map, by contrast, was measured in base pairs—*how far apart* physically two genes or fragments were.

The *order* of genes or other landmarks was always the same on both linkage and physical maps, but the relative distances could be quite different. Chromosomal regions varied five- to tenfold in how frequently DNA was ex-

changed while producing sperm and egg cells. Genes only a few thousand base pairs apart in one region could be separated as often as those hundreds of thousands of base pairs apart in another region, and it differed between males and females for most regions as well. On average, a 1 percent change of recombination translated to a million base pairs (in humans). Both kinds of maps were important, as they served different functions.

Genetic linkage maps provide a bridge from studying how a feature was inherited in an organism to locating the genes for that feature; physical maps are ways to directly catalog DNA by region. The problem of how to make a genetic linkage map in humans had, in principle, been solved by RFLP markers. A parallel problem was how to make a physical map of the human genome and other genomes of interest. As discussion of the Human Genome Project began in 1985, physical mapping of large genomes was just beginning through work on yeasts and nematode worms.

Two groups began independently to apply the cloning and ordering strategy to make maps of this pair of model organisms. Maynard Olson's laboratory at Washington University began to map the chromosomes of *Saccharomyces cerevisiae,* or baker's yeast, which is used not only in baking but also in wine and beer fermentation. John Sulston and Alan Coulson at the Medical Research Council laboratory in Cambridge, England—later joined by Robert Waterston at Washington University in St. Louis—worked toward a physical map of *Caenorhabditis elegans,* a soil-dwelling nematode about a millimeter long. Genome-scale physical mapping of both organisms began in the early 1980s and showed promising results by 1986.[3;4] Both projects focused on organisms central to biological understanding.

Yeast was emerging as the core model for eukaryotic genetics, that is, the genetics of organisms whose cells have a separate nucleus containing the chromosomes. (Bacteria and many other lower organisms do not sequester their chromosomes in a separate compartment, or nucleus; they are called prokaryotes, meaning "before nucleus.") The 12.5 million base pairs in the genome of *S. cerevisiae* were a logical early target for physical mapping. Having a set of ordered genomic DNA clones would be an extremely powerful addition to the already formidable armamentarium assembled to conquer yeast genetics. Botstein, together with Gerald Fink of the Whitehead Institute for Biomedical Research, noted several features marking yeast as a model, most important "the facility with which the relation between gene structure and protein function can be established."[5] The wealth of data on yeast mutants from classical genetics, usually bred by selecting for those yeast cells that could survive under stressful conditions, combined with the immense power of DNA exchange within yeast cells, made it possible to introduce mutations into known genes. The effects of these mutations could be quickly assessed because of the short generation time. By introducing mutations, it was generally possible to snare

genes in the genome. By studying what happened to the protein or RNA gene product when a mutation occurred and correlating it to how the organism's biology changed, it was possible to draw inferences from gene structure to protein function, and thence to physiology. The speed and precision of yeast biology expedited efforts to link the triad of gene, protein, and function, so that "proteins first discovered elsewhere but present in yeast may best be studied first in yeast."[5]

The cooperative sociology of the yeast research community was another important factor noted by Botstein and Fink: "newcomers find themselves in an atmosphere that encourages cooperation. . . . not only are the published strains and mutants generally made available, but many (if not quite all) laboratories in the field routinely exchange strains, protocols, and ideas long before publication." Yeast genetics was ripe for a structural approach, and stocks of ordered DNA clones representing the entire genome would be an immensely useful tool. Olson's proposal to make a physical map from bacterial clones like those pioneered by Maniatis was thus enthusiastically greeted by his peers, and his grant was approved.

Olson planned to start by making a library, but then to try to put the books in order. The ordering strategy was to find DNA clones that overlapped one another. By finding next neighbors, then the next, and so on, eventually the order of the cloned DNA fragments would be established and they could be assigned to their chromosomal region of origin.

Yeasts were wonderful experimental models for many aspects of eukaryotic genetics, but these single-celled organisms were unsuitable for studying the complex interactions found in more complex organisms. Yeasts do not have brains or adrenal glands, for example, and so they do not fabricate intricate connections between brain cells or develop specialized hormone-secreting cells to communicate to other organs widely separated in the body. Yeasts are far from simple, but large animals are immensely more complex, with trillions of cells somehow coordinated into a whole.

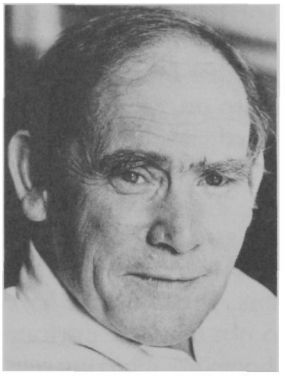
The ideal organism to address such questions, it turned out, was the nematode *Caenorhabditis elegans*. This worm was an unlikely candidate to win a beauty contest, but its three-day generation time and penchant for self-fertilization, thus automatically establishing pairs of identical chromosomes, made it an excellent choice for genetic inquiry.[6] Like many other basic lines of inquiry in molecular biology, the foundation for *C. elegans* biology was laid predominantly at the Medical Research Council (MRC) laboratory in Cambridge, England.

Sydney Brenner of the MRC selected *C. elegans* in the early 1960s as a model to study multicellular phenomena, especially the nervous system.[7–9] Brenner wanted the smallest animal possible that was nonetheless complicated enough "to study the effects of mutations in single genes . . . to isolate mutants

affecting the behavior of an animal and see what changes have been produced in the nervous system."[7] Brenner outlined his first ideas about how a research program might go forward in an October 1963 proposal to the MRC:

> The *new major problem* in molecular biology is the genetics and biochemistry of control mechanisms in cellular development. . . . Part of the success of molecular genetics is due to the use of extremely simple organisms which could be handled in large numbers: bacteria and bacterial viruses. . . . We should like to attack the problem of cellular development in a similar fashion, choosing the simplest possible differentiated organism and subjecting it to the analytical methods of microbial genetics. Thus we want a multicellular organism which has a short life cycle, is easily cultivated, and is small enough to be handled in large numbers, like a microorganism. It should have relatively few cells, so that exhaustive studies of lineage and patterns can be made, and should be amenable to genetic analysis.[9]

Brenner went on to tout the virtues of *Caenorhabditis briggsiae,* which was the initial focus until supplanted by another species, *C. elegans.* Brenner's logic



Sydney Brenner, a researcher at the Medical Research Council's molecular biology laboratory in Cambridge, England, was an influential early champion of genome mapping. His experimental animal of choice, beginning in the 1960s, was the nematode worm *Caenorhabditis elegans. Courtesy Sydney Brenner*

followed that being pursued by Seymour Benzer at Caltech to understand neural function and other complex phenomena in *Drosophila,* the fruit fly. To carry out Brenner's agenda, it was necessary to find mutant nematodes, with observable differences, and also to assemble an awesome mass of structural information—the lines of descent of all cells in the worm's body, and the connections between them. Several laboratories at the MRC laboratory in Cambridge dedicated themselves to doing just that.

John Sulston, through a monumental effort, traced the development of the more than nine hundred somatic cells in the nematode's body, by watching the worms develop under a microscope with special optics that enabled him to observe every cell in the transparent body of the nematode.[6;7;10–13] Sulston made meticulous records of which cells produced which, and thus created a "pedigree" of all 959 nongerm cells. This was an incredible feat, producing the

basic information necessary to trace what happened when the development of specific cells was disrupted. With this set of lineages, it was possible to observe directly the effects of killing a particular cell and to observe how its death affected the nematode's behavior and ability to survive.

In another *tour de force,* John White, Eileen Southgate, and Nichol Thompson of the MRC group reconstructed the "wiring diagram" of the worm's nervous system, analyzing twenty thousand photographs taken by an electron microscope.[6; 10; 14] They traced the main connections linking all 302 nerve cells. These two efforts are mind-boggling in their detail. The nematode was a reductionist's delight. It was conceivable that with these basic tools it would become possible to understand the entire organism's biology in all its mechanistic detail. The notion was not that all the biology would be explained by structural details, but that the structure was the best scientific strategy to try to get at both genetic and environmental factors influencing development.

The cell lineages and connectivity maps were the starting points to study what happened when something disrupted normal structure. Structural genetics was the crucial missing element. Even as the work began on *C. elegans,* the purpose was to correlate behavior with structure; DNA was the conceptual starting point. Sulston and Brenner estimated the size of the genome early on, to see what they were up against. They estimated the genome size by grinding up nematodes and seeing how much DNA their cells contained, and also by observing how long it took for separated DNA strands to reassemble into double helices. (The more complex the genome, the longer this took.) These methods suggested that the genome consisted of eighty million base pairs, giving it "the smallest value of any animal."[15] (This estimate was increased to 100 million base pairs in the late 1980s. Corrections came because the physical map then nearing completion gave more accurate data, and it became clear that the genome of the bacterium *Escherichia coli,* whose size had been a scaling factor for the *C. elegans* calculation, was larger than originally thought.)[16] The genome was segmented into six chromosomes containing seven hundred genes that were mapped by 1988.[6] A physical map of the worm's genome was the critical next step. Sulston and Coulson took it.

Extending the length of DNA in each clone simplified the physical mapping process. Dozens of laboratories helped improve cloning vectors that could consistently contain DNA inserts thirty thousand to forty thousand base pairs long, and these quickly became standard fare. This reduced the number of DNA fragments that had to be sorted through and ordered to make a library. With these advances, it became possible to take DNA from chromosomes, clone it, and reconstruct the order of cloned DNA fragments. Eventually a complete map of the *C. elegans* genome could be assembled. This kind of physical map had an enormous advantage—the chromosomal DNA would be not only mapped but also cloned and stored in the freezer for further analysis. If one wanted to study DNA from a region known to contain a gene, for

example, one could go the freezer and pull it out, or call the scientist who had it in his or her freezer.

Coulson and Sulston labored for several years to make collections of *C. elegans* DNA fragment clones and then "fingerprint" the fragments.[4] DNA fragment patterns were stored in a computer, which looked for other clones that might exhibit a similar pattern, thus indicating overlap. Coulson and Sulston had over 80 percent of the *C. elegans* genome covered by sixteen thousand clones in 1986. When ordered into overlapping clusters, they fell into seven hundred groups. This was a great boon to the close-knit nematode research community, but the problem remained of how to close the gaps. Those seven hundred clusters should eventually resolve into the six chromosomes. David Burke, Georges Carle, and Maynard Olson of Washington University in St. Louis helped solve the problem with a gift from yeast.

Burke, a student in Olson's laboratory, became interested in the goings-on in chromosomal structure. Burke and Carle constructed cloning vectors that yeast cells would recognize as chromosomes. Because they were treated as chromosomes, the fragments of DNA being copied could be enormously large. The idea was to co-opt the normal cellular machinery that copied and distributed DNA to new cells. Burke succeeded in making artificial chromosomes containing DNA fragments more than ten times larger than could be cloned in bacteria. The products of Burke and Carle's artifice became known as yeast artificial chromosomes, or YACs for short.[17]

The length of DNA fragments cloned in YACs helped solve several serious problems at once. First, far fewer clones were needed to span a chromosomal region. Second, the problems in cloning some genes in bacteria might be overcome in yeast, whose biology was more similar to that of higher organisms.[18] Third, the longer fragments improved prospects for detecting overlap, making it easier to find next-neighbor clones. The fraction of the clone needed to detect overlap was smaller, dramatically improving the speed of making a complete physical map.[19] Using YACs, the MRC and Washington University groups were able to span many gaps, reducing the number of contiguously mapped regions (nicknamed contigs) from 700 to 346 in seven months.[20] By late 1989, the number of gaps in the *C. elegans* map was down to 190;[21] by 1992, more than ninety million base pairs of the genome were physically mapped, with only forty remaining gaps.[22] A new refrain was heard in nematode laboratories: "The gaps in maps are filled mainly with the YACs."

Physical mapping was also greatly assisted by the ability to separate much longer DNA fragments in the laboratory. In 1984, David Schwartz and Charles Cantor, then at Columbia University, developed the DNA separation technique known as pulsed-field electrophoresis. The method was an elegant, if somewhat slow, way to distinguish DNA fragments millions of base pairs in length. Dealing with such enormous fragments entailed special handling to minimize inadvertent fragmentation and applying electric fields that changed direction periodically.[23] Many embellishments of this technique soon fol-

lowed, and it became possible to make maps of large chromosomal regions by this technique relatively quickly. Although initially the technique did not result in sets of *clones* of DNA from the region for further study, it did make mapping much faster, and it could help establish landmarks for subsequent analytical steps. With YAC-sized clones, pulsed-field electrophoresis was necessary for separation. The method became integral to physical mapping involving very large DNA fragment clones, until supplanted by faster techniques.

The physical maps of yeasts and nematodes became extremely powerful tools for genetics and for general understanding of biology. They eliminated the need for each researcher to develop a clone library laboriously and to screen it independently. The nematode physical map was put in a computer database available to all laboratories. Those who discovered genes or mutant organisms fed into the system, thus linking physical and genetic maps in a living network of cooperating laboratories. This well-stocked toolbox—containing genetic linkage maps, cell lineage maps, physical maps, and mutant strains with known characteristics—enabled biologists to approach an understanding of *C. elegans* unequaled by understanding of any other organism of comparable complexity.[24] The wealth of structural detail, the quality of the researchers, and the persistent pursuit of the newest technologies proved the prescience of Brenner's insight.

The problem of physically mapping the human chromosomes was greatly underestimated. It had been blithely assumed that yeast and nematode maps would be completed quickly and that the same techniques would translate quickly and readily to the far larger and more complex human genome. The estimates of complete human chromosome maps within two or three years, proffered in 1987 and 1988, proved too optimistic, but only by a half decade or so. Despite misgivings as the genome project was launched in 1990, the first physical maps of human chromosomes were published in 1992. David Page and his group at the Whitehead Institute produced a map of the Y chromosome,[25;26] and a team led by Daniel Cohen of France reported a map of chromosome 21.[27;28]

The road to physical maps of yeast and nematode genomes was not yet at an end. With physical maps nearing completion, the next step was to determine the entire DNA sequence of the yeast and *C. elegans* genomes. A consortium of European laboratories began to sequence yeast chromosomes in 1988, joined by a group at Stanford University in 1990. The Washington University and MRC groups began a transatlantic joint project to sequence the *C. elegans* genome in 1990.[22] The road of a billion nucleotides began with a single base.

4

Sequence upon Sequence

THE FAR-REACHING SIGNIFICANCE of the discovery of DNA's double-helical structure was immediately apparent to many scientists. The physicist George Gamow, for example, was one of the first to realize that the information stored in DNA must be in the form of a four-letter digital code. Writing ten months after Watson and Crick first described the structure of DNA, he noted:

> The hereditary properties of any given organism could be characterized by a long number written in a four-digital system. . . . the enzymes (proteins), the composition of which must be completely determined by the deoxyribonucleic acid molecule, are long peptide chains formed by about twenty different kinds of amino acids, and can be considered as long "words" based on a twenty-letter alphabet."[1]

Like the order of 0's and 1's, the two-letter digital code in computer software, the order of the chemical subunits of DNA contained the instructions not only for the assembly of proteins but also for their biological function. The software was useless without the hardware to translate it, but a great deal could be learned by looking at the software code. If one was trying to understand what a computer was doing or the nature of an error that disrupted it, then the software was a good place to start.

Determining the order of bases in the entire genome is the core idea that started the genome project. Each nucleated cell of the human body contains forty-six chromosomes: twenty-two pairs of nonsex chromosomes and a pair of sex chromosomes (two X's for females, and X and a Y for males). The idea, as initially posed, was to determine a reference sequence for each chromosome. Each chromosome is an extraordinarily long DNA molecule, from fifty million to hundreds of millions of base pairs in length, bundled with proteins and RNA. The total number of base pairs in a complete reference sequence of the human genome was estimated at more than three billion, necessarily a rough estimate until physical maps are completed.

Given that no more than a few hundred thousand base pairs had ever been sequenced in a contiguous region in the mid-1980s, sequencing the genome was a largely impractical idea when first posed. The idea nonetheless initiated a debate within science that broadened the definition of the Human Genome

Project; DNA sequence information remained a central, if no longer exclusive, objective. Having the map and sequence information would not impart all knowledge about biology, because most interesting questions are about function rather than structure. DNA sequence information was, however, enormously useful for several reasons.

The DNA code is shared among organisms. Genes with similar functions are often historically related, having sprung from a common ancestral gene. Their DNA sequences are similar, and hunting for such similarities is a powerful way to get hints about the function of a new gene. If a newly discovered disease gene is similar in sequence to a yeast gene that codes for a cell surface receptor, for example, experiments to look for receptor proteins are in order. A cancer-associated gene whose sequence is similar to a growth-regulating molecular switch in yeast is a clue to the origins of cancer.

DNA sequence is, in this sense, the *lingua franca* of biology, because all organisms speak it. Most genes are conserved through evolution or built from bits and pieces of existing genes, as tweaked and tuned by evolutionary history. Examining sequence similarities discloses the historical relationships between genes, and hence the relatedness of the proteins they produce. Structural similarity suggests (but does not establish) functional similarity. The sequence of amino acids in a protein can provide the same kind of information about relatedness, but it requires having sufficient amounts of pure protein to analyze. It is much easier to examine sequence at the level of the gene, because DNA can be spliced and copied, providing enough material to sequence. By using DNA sequence to suggest the function of the protein it produces, scientists can shortcut the long and tedious process of purifying the protein.

DNA sequence is also a natural way to catalog genetic information. What better way to keep track of genetic information than by storing its digital code? The cataloging process can lead to surprises. The first regions selected in the project to sequence the *C. elegans* genome, for example, revealed far more genes than expected. The region was selected, in part, because it was known to be gene-rich, but the number of genes was twice as large as projected. The project also turned up several genes of known function that had not been previously found.[2]

Sequence data also promised to serve as a starting point for biology in a way that had not until then been systematically pursued. In early 1992, a massive European collaboration succeeded in sequencing chromosome 3 of baker's yeast, the first chromosome of a nucleated cell to be sequenced. The sequence was achieved by thirty-five laboratories coordinated through a European Community project, culminating a complex three-year collaboration. An even larger and more complex collaboration than the *C. elegans* sequencing effort, it was a major stride forward. The publication in *Nature* had 147 authors from thirty-seven institutions.[3]

The complete sequence contained 182 apparent genes, only thirty-four of which had ever been mapped; this was somewhat higher than a theoretical

estimate of 160 genes made by a Japanese group.[4] It showed the promise of sequencing for discovering genes, as "even in a genome as small and as intensively studied as that of yeast, only a minor fraction of the genes has been identified by classical means."[3] The authors of the *Nature* article concluded that systematic sequencing projects could "reveal new functions that have been missed by more traditional approaches and also illuminate the mechanisms of genome evolution."[3] With the sequence in hand, an obvious goal was to determine the function of these 182 genes. This was much more feasible in yeast than most other organisms because, in the authors' words, "the functional analysis of novel genes discovered from the sequencing is facilitated by the easy methods for gene disruption and replacement . . . available in yeast."[3] It turned out that only 20 percent of the newly sequenced genes were similar to those found in various databases, and that "yeast molecular geneticists are working on only a small subset of the problems presented by their organism."[3] Once this functional catalog expanded, it would serve as the reference book for studying function in other organisms. The yeast sequencing project was a major boost to European genetics. The battle was joined to sequence the remainder of the yeast genome, with groups from Canada, the UK, Japan, and the United States mounting projects on one or more chromosomes, in hopes of having a complete reference sequence by year 2002.[5] The *C. elegans,* yeast, and other large sequencing projects began the slow process of turning biology on its head—starting from DNA sequence information and working toward function rather than the other way around.

As noted in Chapter 1, DNA is transcribed into RNA, usually on the way to protein. RNA serves several functions, some of it involved in editing and splicing the genetic code. RNA can be a messenger—the vehicle to translate from DNA code to strings of amino acids that become proteins. Some RNA becomes part of the cellular machinery and is never translated into protein. Proteins, however, make up most cellular structures and mediate the vast bulk of chemical reactions within cells.

The flow of information is usually from DNA through RNA to protein—what Crick called the "central dogma," as elaborated in the late 1950s and early 1960s.[6] We now know that information can also go from RNA to DNA when cells are infected with some viruses, reversing the flow, as occurs with AIDS infection. Some cellular RNA is occasionally copied into DNA and then inserted into chromosomes as well, but these counterexamples are small eddy currents in the torrential outflow of information that begins with DNA.

Chromosomal DNA is the terrain to be mapped by the genome project, and DNA sequencing provides the map with the highest possible resolution. The order of base pairs in DNA is the raw information. Getting at that order is consequently of central importance. In finding the specific defect for cystic fibrosis, for example, several laboratories engaged in massive sequencing efforts. Indeed, the Du Pont company, which developed an automated DNA

sequencing instrument and sold it for a few years, crowed that its technology had uncovered the gene.[7] A few humans were also involved, but the importance of technology was nonetheless worth noting. Two methods for determining DNA sequence were developed independently at Cambridges separated by the Atlantic Ocean.

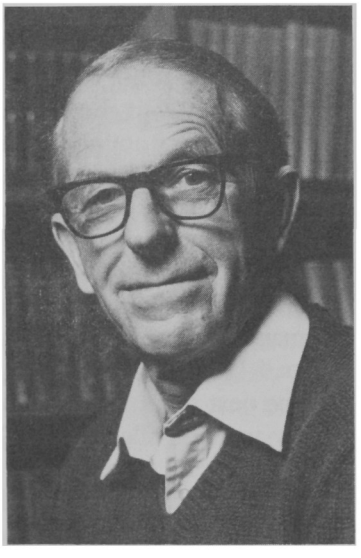
Methods to sequence DNA were conceptual extensions of work on sequencing proteins. In 1945, Frederick Sanger in Cambridge, England, set out to determine the order of amino acids in insulin, the protein hormone used to treat diabetes.[8] The sequence of insulin was a landmark in protein chemistry and earned Sanger a Nobel Prize in 1958. Protein sequencing was given a major boost in 1950, when Pehr Edman of the University of Lund in Sweden discovered a way to chop one amino acid at a time from the end of a protein.[9] By removing one amino acid at a time, the sequence was directly determined. Previous methods had required breaking proteins into small fragments, analyzing the order of amino acids in the fragments by a variety of methods, and then reconstructing the overall order of the original molecule. Edman's method was not only easier to understand but also proved well suited to automation. By 1967, instruments to determine amino acid sequence for proteins were on the market. These evolved into rapid and reliable protein sequencing instruments.[10;11]

The next step up from protein sequencing was RNA. It took Robert W. Holley and his colleagues at Cornell University seven years to determine the order of seventy-seven bases in one form of RNA. Like the first protein-sequencing methods, they broke the RNA molecule into small fragments and reconstructed the order.[12;13] For several years, DNA sequencing was done by transcribing it to RNA and then deducing the RNA sequence of short segments. This was the strategy used to determine the first DNA sequence, published in 1971: the short "sticky ends" of bacteriophage lambda (λ).[8;14] The methods were too slow and tedious to be scaled up for large DNA molecules. Sanger recognized this, and the importance of more efficient DNA sequencing.

Sanger's DNA sequencing method took advantage of how cells make linear strings of DNA. Cell enzymes start from DNA in a chromosome and use it as the template to make copies, preserving the order of base pairs. The chemical backbone of the base-pair unit (nucleotide) in the DNA molecule is the same, a sugar and a phosphate group. The only variable is which base is inserted. The outer backbone is thus a monotonous repeat of sugar-phosphate-sugar-phosphate . . . The information is contained in the order of bases in the "rungs" of the DNA ladder. Each base has only one complementary base. A binds to T and C binds to G. DNA in chromosomes is stored as two strands of DNA bound to each other, with one strand exactly complementary to the other. In making copies of the chromosomes, the strands are unzipped and new copies made of each strand. A new pair of identical double-stranded DNAs results.

Sanger's idea was to adapt the cell's natural machinery, but to introduce

chemical tricks to produce the DNA sequence. Sanger's initial method was to supply all of the components, but to "starve" the reaction of one of the four bases needed to make DNA. The cell would copy strands of DNA, but would run out of one of the bases. All the chains would run out just before the same base, starved of that one component. In the sequence ACGTCGGTGC, for example, starving for T would produce ACGTCGG(blank) and ACG(blank). The T in the longer fragment was produced before it ran out of T precursors;



Frederick Sanger pioneered work on the sequencing of DNA, RNA, and proteins at the MRC's Cambridge laboratory. His emphasis on understanding biological function through molecular structure has guided British participation in genome research from the beginning. *Courtesy Frederick Sanger*

the shorter fragment, perhaps because the reaction started just a bit later, ran out before it got there. By separating the resulting molecules by length, it would be obvious that the eighth and fourth positions should be T, because the chains of seven and three bases (one less than the "T" positions) were present. Starving for G would produce AC(end), ACGTC(end), ACGTCG(end), and ACGTCGGT(end), meaning G was in positions 3, 6, 7, and 9. The whole sequence would follow directly from starving for each of the four nucleotide precursors.

Sanger's next trick was to find chemicals that were inserted in place of A's, C's, T's, and G's, but caused a growing DNA chain to end. In the above example, the fragments for the "false" T (T*) would be ACGT* and ACGTCGGT*. With one terminator for each base type, the sequence could again be read directly by just measuring how long they were. Sanger presented his first partial DNA sequence to an awestruck audience in May 1975[15–17] and published the simpler chain-terminator method in 1977.[18]

Sanger noted in a marvelous autobiographical article reviewing his career, "Sequences, Sequences, and Sequences," that "of the three main activities

involved in scientific research, thinking, talking, and doing, I much prefer the last and am probably best at it."[8] He also gave some insight into why the MRC laboratory in Cambridge played such a central role in the development of modern molecular biology.

> I was in the fortunate position of having a permanent research appointment with the (British) Medical Research Council, and was not under the usual obligation of having to produce a regular output of publishable material, with the result that I could afford to attack problems that were more "way out" and longer-term: in fact, as few others could adopt this approach, I felt under some obligation to do so. . . . I like the idea of doing something that nobody else is doing rather than racing to be the first to complete a project.[8]

Several thousand miles away, in the other Cambridge, Allan Maxam and Walter Gilbert of Harvard developed an entirely different sequencing method, based on chemical disruption of DNA. They were studying the regulation of a bacterial gene, a region of DNA that served as a "switch" to turn the gene on and off. In the off state, no RNA was transcribed; when it was on, RNA was produced copiously. A protein that stuck directly to DNA appeared to throw the switch. When the protein, dubbed the repressor, bound to its DNA target, it blocked RNA transcription of genes nearby. Maxam and Gilbert wanted to study the specific DNA site recognized by the repressor protein. They isolated a DNA fragment from the region and showed that a short stretch of DNA was "protected" from degradation when the repressor was present. When bound to DNA, the repressor protein protected the DNA from digestive enzymes. They made RNA from this region and spent two years laboriously determining the sequence of a twenty-four-base-pair region of DNA, using the fracture-and-reconstruct methods.[19]

A Soviet scientist, Andrei Mirzabekov, visited the laboratory twice during this period. His first visit, late in 1974 or early in 1975, was brief. He was finding ways to break DNA at specific base pairs by selectively adding methyl groups to specific DNA bases. Mirzabekov found that dimethyl sulfate destabilized the DNA, leading to breakage specifically at adenine (A) and guanine (G) bases. Mirzabekov, Gilbert, Maxam, and graduate student Jay Gralla discussed the possible use of such DNA-fragmenting reactions to study the repressor-binding region. They already knew the DNA sequence from the region, but wanted to know precisely where the protein bound. The idea was that protein binding would block not only enzymes but also chemical methylation—the addition of methyl (CH_3) groups to the bases in DNA. If they compared DNA fragments fractured in the presence and absence of repressor, there would theoretically be a stretch of DNA that would break without repressor, but would be protected with it.

The first experiment was a total mess, but the second showed the expected pattern. The results were reported at a Danish symposium in the summer of

1975.[20] This was independent verification of the digestion experiments, but direct chemical fragmentation of DNA gave a much more specific binding profile.

The new method had another extremely appealing feature. If reaction conditions could be found to fracture DNA selectively at each of the four bases constituting any DNA, it would be possible to read the DNA sequence directly by separating fragments according to length. Maxam adjusted reaction conditions until he could fragment DNA at G only, or at both A and G.[21] If a fragment appeared in both reactions, it was a G; if it appeared in the A + G reaction but not the G, it was an A. Maxam then found a similar method to break DNA at cytosine (C) and thymine (T). The base occupying each position in DNA could thus be inferred from the fragmentation in four separate chemical reactions. A DNA sequencing method was born.

Late in that summer of 1975, Maxam gave a talk at the annual New Hampshire Gordon Conference on nucleic acids, while Gilbert was in the Soviet Union. Maxam distributed a protocol for Maxam-Gilbert sequencing at that meeting.[22] Their method of DNA sequencing was published in 1977.[23]

Sanger later confided, "I cannot pretend that I was altogether overjoyed by the appearance of a competitive method [for DNA sequencing],"[8] although the two methods proved to have complementary strengths. The preferred approach varied according to what was sequenced and how the DNA was prepared.

Thus, between 1974 and 1976, two independent techniques for sequencing DNA were developed, each an elegant solution to a central methodological problem. The capacity to sequence DNA opened up an enormous range of experiments and complemented the other major technical triumph of molecular biology during this period—artificially recombinant DNA. Embellishments of these techniques were used to determine the sequence of progressively larger fragments, and eventually whole genomes. The first sequence of twelve base pairs in 1968 grew to the 5,386-base-pair genome of the bacterial virus phi-X, achieved in 1977 with the new Sanger method. The DNA sequence of the small chromosome within a human mitochondrion, the cell's energy pack, was determined in 1981. It consists of more than sixteen thousand base pairs.

Genome of a virus, one of the first—and smallest—genomes ever to be decoded, was worked out initially in 1977 by Frederick Sanger and his colleagues at the Medical Research Council Laboratory of Molecular Biology in Cambridge, England. The genome is represented here by a sequence of more than 5,000 letters, corresponding to the four nucleotides (adenine, thymine, cytosine, and guanine) of the viral DNA molecule, which has only a single circular strand during part of its life cycle. Starting from the top, the sequence runs from left to right on odd-numbered lines, and from right to left on even-numbered lines. (In its circular form, the two ends of the molecule are connected.) The particular sequence shown, representing the entire genetic inheritance of the extremely small bacterial virus designated phi-X174, is divided into nine genes, which code in turn for the amino acid sequences of nine different proteins. More than 500,000 such pages would be needed to similarly display the human genome.

GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTCGGATATTTCTGATGAGTCGAAAAATTATCTTGATAAACGAGGAATTACTACTGCTTGTTTA
TCAACTACCGCTTTCCAGCGTTTCATTCTCGAAGAGCTCGACGCGTTCCTATCCAGCTTAAAAGAGTAAAAGGCGGTCGTCAGGTGAAGCTAAATTAAGC
AACGATTCTGTCAAAAACTGACGCGTTGGATGAGGAGAAGTGGCTTAATATGCTTGGCACGTTCGTCAAGGACTGGTTTAGATATGACTCACATTTTGTT
TGAACTGAGTACTAAAGAATGGATAATCACCAACTTGTCGTAGCCTGAGTCTATCATTAGGTGCGAGAAAATTTTACAGTTGTTCTCTTAGAGATGGTAC
TACTGAACAATCCGTACGTTTCCAGACCGCTTTGGCCTCTATTAAGCTCATTCAGGCTTCTGCCGTTTTGGATTTAACCGAAGATGATTTCGATTTTCTG
TCGCTCCCATAGGATGTTTCAGGTCGCATGGTATTTGCGTTCGGAGTTGCGTCGCTGCTCGTGCTCTCGCCAGTCATCGTTAGGTTTGAAACAATGAGCA
TTCCTGCTCCTGTTGAGTTTATTGCTGCCGTCATTGCTTATTATGTTCATCCCGTCAACATTCAAACGGCCTGTCTCATCATGGAAGGCGCTGAATTTAC
CAGTCATTCTTGCAGTCACAAAGGACGCGCATGTGCGTTCCATTTGCGCTTGTTAAGTCGCCGAAATTGGCCTGCGAGCTGCGGTAATTATTACAAAAGG
GCAGAAGAAAACGTGCGTCAAAAATTACGTGCGGAAGGAGTGATGTAATGTCTAAAGGTAAAAAACGTTCTGGCGCTCGCCCTGGTCGTCCGCAGCCGTT
TTAAATAGGAGTTCATTCCCCGGCTTCGGGGACGTTATTTTTAACAACTGGTGGATGTATGGTTTCTGCTGCCGGAAATGCGAACGGAAATCATGGAGCG
ATGTCTAATATTCAAACTGGCGCCGAGCGTATGCCGCATGACCTTTCCCATCTTGGCTTCCTTGCTGGTCAGATTGGTCGTCTTATTACCATTTCAACTA
TACAGATGTCATCTCAGTTATCGTTCCGGTGCTGCGTTACCTCTTTCTGCCTCTCGCGGTTGCCGCAGGTAGAGCTTCCTCAGCGGTCGCTATTGGCCTC
TTTTACTTTTTATGTCCCTCATCGTCACGTTTATGGTGAACAGTGGATTAAGTTCATGAAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACCAA
GATATCTATAGTTTATTGGGACTTTGTTTACGAATCCCTAAAATAACCATAGTCCCAATTAGCACGGTTCTTTTCGCCGTACCAGTTATATTGGTCATCA
CGTATTTTAAAGCGCCGTGG...ATGCCTGACCGTACCGAGGCTAACCCTAATGAGCTTAATCAAGATGATGCTCGTTATGGTTTCCGTTGCTGCCATCT
TCGTCGAACGTCTGGGTATTACAGTTATCTACACCATCTTCAGCAGTAAACCGCTCTTTCGAGTCAGAGTCCTCCTTCGCCTCGTCAGGTTTACAAAAAC
TATGCTAATTTGCATACTGACCAAGAACGTGATTACTTCATGCAGCGTTACCATGA.GTTATTTCTTCATTTGGAGGTAAAACCTCATATGACGCTGACA
ACTTGTGCTGGTCTTTTGACCGGATTGCTGCAAACCAGTCAAGGTAGTTGTAGTATCGGTCTACGGGTCTCTAATCTCGCGTACTGTTCATTTCCTGCCA
ACAGACCTATAAACATTCTGTGCCGCGTTTCTTTGTTCCTGAGCATGGCACTATGTTTACTCTTGCGCTGGTTCGTTTTCCGCCTACTGCGACTAAAGAG
GGAAGTATCTTTAAAGTGCGCCGCCGTTCAACGGTATGTTTTGTCCCAGCGGTCGTTATAGCCATATTCAGTTTCGTGGAAATCGCAATTCCATGACTTA
ATGTTTTCCGTTCTGGTGATTCGTCTAAGAAGTTTAAGATTGCTGAGGGTCAGTGGTATCGTTATGCGCCTTCGTATGTTTCTCCTGCTTATCACCTTCT
TTGACTTGCTGACTTTGTGACCAGTATTAGTACCAACGCTTATTCATGCGCAAGAACCTTTAGTGGTCTTCCGCCAAGGACTTACTTACCCTTCGGAAGT
GTTGCAGTGGATAGTCTTACCTCATGTGACGTTTATCGCAATCTGCCGACCACTCGCGATTCAATCATGACTTCGTGATAAAAGATTGAGTGTGAGGTTA
TTTCAGACTTTGTACTAATTTGAGGATTCGTCTTTTGGATGGCGCGAAGCGAACCAGTTGGGGAGTCGCCGTTTTTAATTTTAAAAATGGCGAACGCAAT
TATTTCTCGCCACAATTCAAACTTTTTTTCTGATAAGCTGGTTCTCACTTCTGTTACTCCAGCTTCTTCGGCACCTGTTTTACAGACACCTAAAGCTACA
GGACTAATCGCCGCAACTGTCTACATAGGTAGACTTACGTTACTTCTTTTGGTGGTAATGGTCGTATTTGGCAGTTTGATAGTTTTATATTGCAACTGCT
TTGTTTCAGTTGGTGCTGATATTGCTTTTGATGCCGACCCTAAATTTTTTGCCTGTTTGGTTCGCTTTGAGTCTTCTTCGGTTCCGACTACCCTCCCGAC
TAACGGCCCGCAAGCCCCTTCCTGCAGTTATCAGTGTGTCAGGAACTGCCATATTATTGGTGGTAGTACCGCTGGTAGGTTTCCTATTTGTAGTATCCGT
AACGTCTACGTTGGTTTCATGGTTTGGTCTAACTTTACCGCTACTAAATGCCGCGGATTGGTTTCGCTGAATCAGGTTATTAAAGAGATTATTTGTCTCC
AACTGGCGGAGGTTTGTTAAATCTGTACCGCGGTGGTCGTTCTCGTCTTCGTTATGGCGGTCGTTATCGTGGTTTGTATTTAGTGGAGTGAATTCACCGA
AAAGCCGCCTCCGGTGGCATTCAAGGTGATGTGCTTGCTACCGATAACAATACTGTAGGCATGGGTGATGCTGGTATTAAATCTGCCATTCAAGGCTCTA
CTTCACGGTCGGACGTTGCATGGAAGTTCTTCAGGAAATGGTCGAAATCGTTATCGTGTGCTTTGTTTTGATCCCCGCCGGAGTAGTCCCAATCCTTGTA
TGCCGTTTCTGATAAGTTGCTTGATTTGGTTGGACTTGGTGGCAAGTCTGCCGCTGATAAAGGAAAGGATACTCGTGATTATCTTGCTGCTGCATTTCCT
GGTCAACGTAAAATCATTCGAGAAAAACTAAGAGTTTAGGCCGCAGTTGGTATGGTCGTCTCCTTCGTAGTCGTGGTCGTGCGAGGGTTCGTAATTCGAG
ACAATCAGAAAGAGATTGCCGAGATGCAAAATGAGACTCAAAAAGAGATTGCTGGCATTCAGTCGGCGACTTCACGCCAGAATACGAAAGACCAGGTATA
CCTTTGGACGACAACGAACCTTTCTAACCACAAAAGGTATTATCTGCGTTGCGTCGTCATCTGAGGAAGACA.CTTATTCGTTCGTAGAGTAAAACACGT
GAGATTATGCGCCAAATGCTTACTCAAGCTCAAACGGCTGGTCAGTATTTTACCAATGACCAAATCAAAGAAATGACTCGCAAGGTTAGTGCTGAGGTTG
GGTCTTCGTCGTAGTCACTGCTGTAATCTTTATAGGAAACGTCATCGCGGTTATACTCTTCTCGGTATGGCGACTAAGACGCAAACGACTACTTGATTCA
TGTGGTTGATATTTTTCATGGTATTGATAAAGCTGTTGCCGATACTTGGAACAATTTCTGGAAAGACGGTAAAGCTGATGGTATTGGCTCTAATTTGTCT
TAAGTCTTCCCATTATTCTTGCTTGGTATTTTTTCGGAGGTTCTAAACCTCCGTACTTTTGTATGTTAACCCTCCCACAGTTAGGACTGCCAATAAAGGA
GTCACGCTGATTATTTTGACTTTGAGCGTATCGAGGCTCTTAAACCTGCTATTGAGGCTTGTGGCATTTCTACTCTTTCTCAATCCCCAATGCTTGGCTT
CGGCAGTTGTATGTATAGTGGTAATAGCTTGAGTTGCGGGACGTATGCTTTTCTGTCTTAGAGAAGGTTCTCGAACTACGACAATAGGTAGACGAATACC
CATAAGGCTGCTTCTGACGTTCGTGATGAGTTTGTATCTGTTACTGAGAAGTTAATGGATGAATTGGCACAATGCTACAATGTGCTCCCCCAACTTGATA
CCCGCAAGTCGTCGGTCGAACGTTTTGACGCATTGGCAGAAGAGCAAGAGATTTTTGGTAAAAAGCAGGGGAAGCCCCGCCACCAGATATCACAATAATT
TCTTAAGGATATTCGCGATGAGTATAATTACCCCAAAAAGAAAGGTATTAAGGATGAGTGTTCAAGATTGCTGGAGGCCTCCACTAAGATATCGCGTAGA
GATTAGCCAGCAGTCGGTTGCACTCTCACAGTTTTTGCTATTTGGTTGGTAGTCGTACTCGGACAGCGTAACGTAAGTAGTTTGCGACTTATCGTTTCGG
AGGCGTTTTATGATAATCCCAATGGTTTGCGTGACTATTTTCGTGATATTGGTCGTATGGTTCTTGCTGCCGAGGGTCGCAAGGCTAATGATTCACACGC
CGATGGACATCCTTCACAGGCGTATTTCACGTGGCGTACCTTTACTTCTGCCGGTAATCGACATGGTATGAGTCCGTGTGTTTTTATGACTATCGTCAGC
GTTGACCCTAATTTTGGTTGTCGGGTACGCAATCGCCGCCAGTTAAATAGCTTGCAAAATACGTGGCCTTATGGTTACAGTATGCCCATCGCAGTTCGCT
TGTATCTTTGGTTGTCGGTATATTGACCATCGAAATTCGCCGAGTGGAAATCGTAGTTGTCCGGTGTTGGTTGGTCTTGCACTTTTTCGCAGGACGCACA
GGCGAAATACGTTAACAAAAAGTCAGATATGGACCTTGCTGCTAAAGGTCTAGGAGCTAAAGAATGGAACAACTCACTAAAAACCAAGCTGTCGCTACTT
CGAACCATTCAACCTAATTCGTGAGGCACCTGTCTAAACAGTAACACTCGTAAAAGTAGGGCTTCAACGCCGAGTAAGACTAAGACTTGTCGAAGAACCC
TGGGTTACGACGCGACGCCGTTCAACCAGATATTGAAGCAGAACGCAAAAAGAGAGATGAGATTGAGGCTGGGAAAAGTTACTGTAGCCGACGTTTTGGC
ACGTCCAACCTATGCGGTTAGTAAAAATAGCTTCGCGCGTATTTAAACTCGTCTAAACAGCAGTCTCCAACGCGG

In 1984, the MRC Cambridge group sequenced the 172,000-base-pair genome of the Epstein-Barr virus, the cause of infectious mononucleosis and other conditions. The final landmark of the 1980s was the human cytomegalovirus, whose genome has more than 229,000 base pairs.[24] These achievements were accomplished without much automation. Computers were used to assemble the data into a coherent whole, but the vast bulk of the effort was done by human hands, eyes, and minds.

The next major step for DNA sequencing was to make it faster, cheaper, and more accurate. With enormous stretches of DNA whose sequence was yet to be determined, and with sequencing broken down to a series of standard procedures in repetitive steps, DNA sequencing was a natural target for automation.

Automation of biochemical reactions entailed mixing diverse reagents in small volumes, generally running a reaction inside a single small droplet. Molecular biologists and biochemists frequently ran dozens of reactions at once, each in its own test tube. Getting machines to do some of these mindless tasks would free postdoctoral fellows and graduate students to do more creative things, to move their minds from their hands to their science. By augmenting the duration, reliability, and speed of laboratory work, automation made possible experiments too large and complex to do manually. In the words of one molecular biologist, a robot was the "ultimate postdoc," an endorsement of the new instrumentation but also an acknowledgment of the tedious tasks routinely relegated to young and eager minds in molecular biology.

Devising instruments to automate processes used in molecular biology was pursued at only a few universities. Most of the work was concentrated in companies that sold analytical instruments to laboratories. A few companies were formed to develop instruments, usually growing out of academic centers to fill market niches left vacant by larger companies. Procedures to determine the order of amino acids in proteins were first automated in the late 1960s, followed by instruments to synthesize short proteins from amino acids. Analysis of DNA was next in line. Serious efforts to *synthesize* short segments of DNA, essential to developing highly sensitive probes for analyzing genetic experiments, began in the late 1970s and proved successful by the early 1980s.

Automation of DNA sequence determination began around this time in both Japan and the United States. In the United States, the first successful efforts leading to the current generation of fluorescence-based DNA sequencers began in the late 1970s at Caltech, one of the few academic centers interested in both molecular biology and instrument development.

Leroy Hood was at the center of efforts to marry high-technology instrumentation to molecular biology. Hood grew up near Shelby, Montana, a high school quarterback who directed his team to successive state championships. He went to Caltech as an undergraduate, and upon graduation joined an accelerated M.D. program at Johns Hopkins. He returned to Caltech to pur-

sue a Ph.D., working in the laboratory of William J. Dreyer, whose interests centered on protein structure. In 1967, Dreyer's group was involved in automating the Edman reaction for protein sequence determination. The Caltech group worked with the Beckman Corporation to make an instrument. It became the first in a long line of triumphs, establishing Caltech as the capital of biotechnology instrumentation. Over the next two decades, Hood and his group at Caltech emerged as preeminent innovators.

Hood's main contribution to instrumentation was to create a laboratory environment with a breadth of expertise ranging from engineering through organic chemistry to molecular biology. His contributions to biology were broader than instrumentation. He was one of the leaders in molecular immunology, aimed at understanding the fundamental mechanisms at play in the body's principal defense system. He commanded one of the largest molecular biology groups in the world,[25;26] with a legion of young faculty, postdoctoral fellows, graduate students, technicians, and others that generally hovered around sixty-five, and edged over one hundred in the summers.[27]

Developing instruments was a respectable enterprise at Caltech.[28] An accomplished fabrication shop helped build prototype machines, and the Hood group promulgated a philosophy that tied instrument development to solution of pressing problems in molecular biology, especially in immunology but also in several other areas. The group stayed near the forefront of molecular biology and defined the cutting edge of instrumentation development.

The Caltech group discussed ways to automate DNA sequencing procedures as soon as the Sanger and Maxam-Gilbert procedures became known, but there were several difficulties to be overcome. Henry Huang labored for several years to automate the standard DNA sequencing methods, using funds donated to Hood's group by Monsanto and money obtained from the Caltech president's fund.[29] Huang worked to build an apparatus that could detect DNA fragments, but the problem of detecting a very small signal from the DNA amid a very noisy signal, compounded by the primitive computers of the day, doomed the enterprise. Huang's efforts sparked a continuing interest in DNA sequencing, however, and it was picked up by Lloyd Smith when he joined the laboratory in April 1982.

Smith came to the Hood group from Stanford, where he had applied lasers and fluorescent methods to the study of biological questions.[30;31] While he did not initially go to Caltech to automate DNA sequencing, Smith became intrigued by the problem. Hood had several times urged Huang to try fluorescent labeling instead of ultraviolet absorption, but Huang pointed out the difficulties.[29] Huang also considered measuring molecular mass directly to detect DNA, but again the technical problems were daunting. The Caltech group lacked the expertise in organic chemistry needed to attach fluorescent dyes to DNA. Smith filled that critical gap.

Huang left Caltech in September 1982, and work on DNA sequencing moved to Smith. A complex enterprise grew up, involving the talents of a team

of people at Caltech and the biotechnology instrumentation firm Applied Biosystems. Smith's expertise combined organic chemistry and instrumentation, including the use of fluorescence and lasers. Smith worked to find appropriate dyes, then worked on the tedious and artful task of finding ways to attach them to DNA without perturbing the detection of DNA fragment size. He and the rest of the Caltech team also had to devise a practical way to activate the dyes by laser and to detect fluorescent emissions. Over two years, the Caltech team pushed inexorably forward.[32;33]

Hood secured the necessary funding from the Weingart Institute and also used funds from corporate donations to the Caltech group from Baxter-Travenol, Monsanto, and Upjohn. Hood's ability to cultivate enthusiasm went well beyond academic circles, undergirding his prodigious fund-raising abilities. Priming the money pump was necessary to lay the foundation supporting a legion of young investigators who fiddled with expensive hardware.

While a DNA-sequencing prototype was being assembled at Caltech, a parallel effort moved forward at the nascent Applied Biosystems. Kip Connell from Hewlett-Packard directed the team there. Steve Fung worked on different ways to link fluorescent dyes to DNA. From 1983 through 1985, there was much give and take between Applied Biosystems, which focused on detection and slab gel techniques, and Caltech, where the emphasis was on fluorescent dye chemistry. The Caltech group announced its prototype in a June 1986 presentation.[32] By May 1986, Applied Biosystems had a commercial instrument ready for testing. Its DNA sequencer hit the market in 1987, and was soon joined by a rival fluorescence-based machine manufactured by Du Pont[34] and another machine based on detecting radioactive phosphorus.[35;36]

Both the Caltech and Applied Biosystems projects got a big boost from two brothers and a fateful car ride. One day late in 1982, Tim Hunkapiller was driving his older brother Mike to the airport. Both brothers worked in Hood's group at Caltech. The Hunkapiller brothers grew up in Oklahoma and attended university there. Mike made his way to Caltech, where he studied physical chemical methods to understand enzyme-mediated reactions. In 1976, two years after getting his Ph.D., Mike planned to return to Oklahoma. Hood talked him into staying at Caltech, and Mike got involved in designing an instrument to sequence proteins in much smaller amounts than possible with the instruments developed previously at Caltech and elsewhere. Mike devised a new method that was far more sensitive and could be done on much smaller samples.[10] The result was a prototype instrument to do protein sequencing that could be used for a wide array of problems unapproachable with the previous generation of instruments.

The Caltech group probed several companies for interest in manufacturing the protein sequencer. They approached Becton-Dickinson, Du Pont, and Beckman Instruments. Beckman middle managers saw little prospect for expansion of the protein sequencing market they already dominated. They seemed

to believe a new instrument might compete with their existing one. Du Pont nibbled but did not bite.

Applied Biosystems was founded to build instruments for the new biotechnology. Marvin Carruthers and Hood were on the board of AmGen, a biotechnology company. Carruthers, from the University of Colorado, worked with Hood's group to devise chemistry for a different machine—one that made short stretches of DNA with specified base sequences. The Carruthers collaboration with Hood's group at Caltech culminated in a machine brought to market by Applied Biosystems in 1982.

The group of venture capitalists who helped found AmGen spoke with Hood and Carruthers, who talked enthusiastically about an emerging instrumentation market for molecular biology. The venture capital sponsors were willing to try to tap this market, so they provided capital to start the company that became Applied Biosystems.

A group from Applied Biosystems visited Caltech in the spring of 1981, just after the first gas-phase protein sequencer had been developed. Applied Biosystems later picked up the license for the protein sequencer and evinced interest in the protein synthesizer, DNA synthesis machine, DNA sequencer, and other instruments under development at Caltech. With this suite of four instruments, a laboratory could break down proteins and DNA into their component sequences or build up a specified protein or DNA sequence from scratch. These were essential steps in a wide range of molecular biology experiments. Instruments to sequence and synthesize proteins and DNA formed the technological quartet for a new approach to biological research.

Applied Biosystems became a successful startup company, beating its larger rivals in the instrumentation business. It first marketed the Caltech-inspired protein sequencer in February 1982, and followed in December with the DNA synthesizer based on improved chemistry developed in Caruthers's laboratory. (Beckman had a sublicense to this chemistry as well.) Arnold Beckman, whose foundation was a generous funder of research at Caltech, hit the roof. His managers might not have been interested, but, he felt, Hood should have notified him directly that Beckman Instrument's middle managers had tunnel vision. While Beckman himself no longer had line authority, his name was still on the company, and he believed he could have taken remedial action. Beckman's fury intensified while Applied Biosystems displaced Beckman Instruments as the dominant force in protein sequencing within a year of introducing its instrument. Hood learned that selling a new idea required approaches to top management and also convincing middle management and corporate technical experts. Individuals at many levels could block a new idea; progress required all the gates to be open. Tensions between the Hood group and Beckman were ultimately healed over, after a few years' delay in securing the donation for the new Beckman Institute.

In 1982, Mike Hunkapiller was consulting with Applied Biosystems. Mike followed Applied Biosystems' early history with great interest, but was more

interested in several biology projects at Caltech. Within a year, he changed his mind. He left Caltech to join the firm in July 1983, later becoming vice president for science and technology.[37] Applied Biosystems ultimately succeeded in developing the full set of four instruments for DNA and protein sequencing and synthesis and then began to work on yet other projects, such as the development of robots able to perform diverse biochemical reactions. Mike Hunkapiller was head of the development team.

Tim Hunkapiller had an entirely different scientific background and personal style. He was studying evolutionary biology and doing fieldwork in Oklahoma when he applied to Caltech for graduate school. He applied to Caltech in part because the application was free. He was rejected, but with the suggestion that he might be accepted if he took more physical chemistry. He agreed to take a special tutorial and was accepted at Caltech.

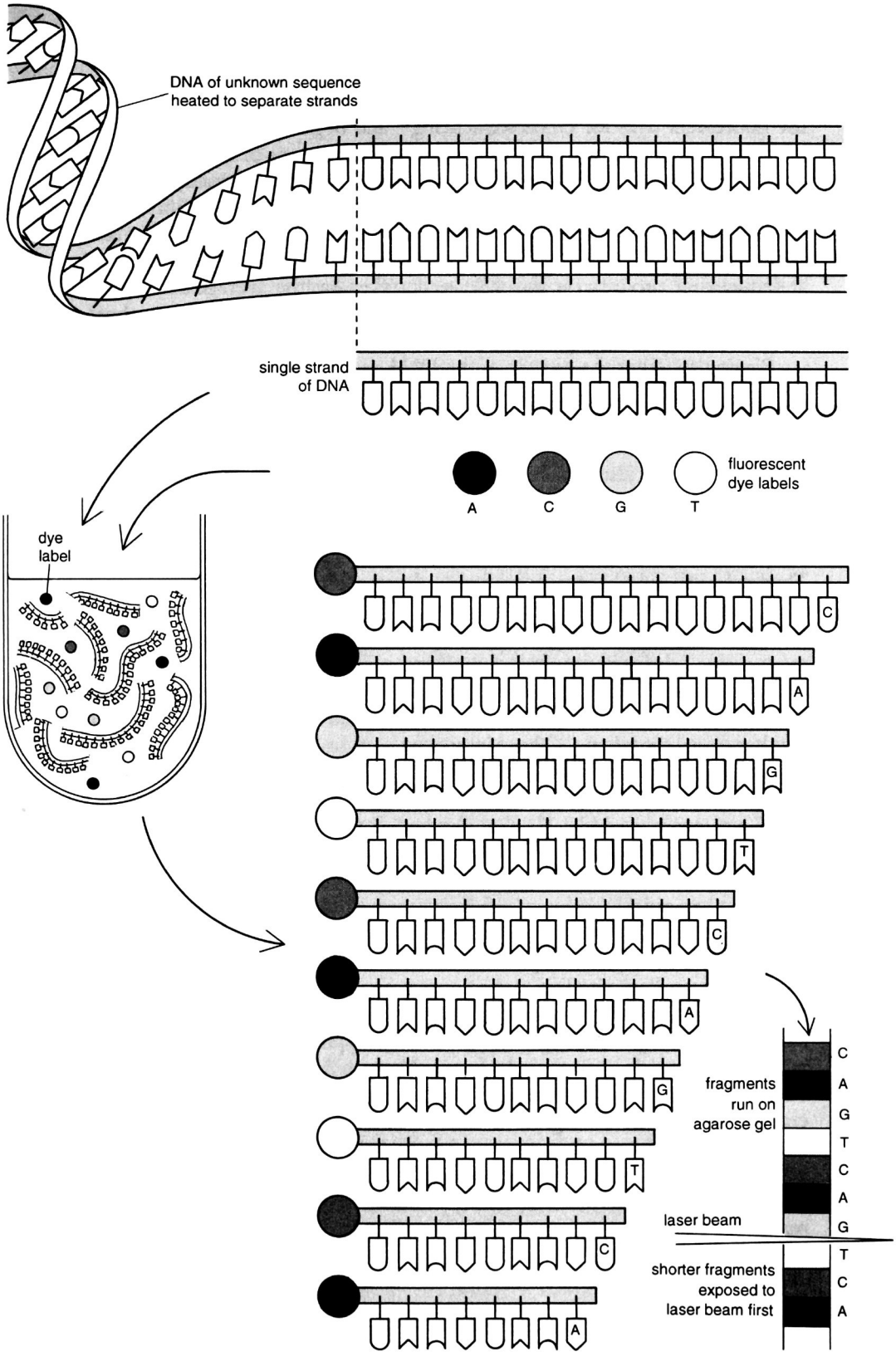
The Hunkapiller brothers' different styles led to a fruitful scientific collaboration. The DNA-sequencing project was one beneficiary. Tim's eye was caught by an article about separating molecules in gels within very thin capillary tubes, lodged in silicon wafers like those used for computer chips. Tim and Mike were speculating about how capillary separation techniques might be used to determine DNA sequences when Tim brought up the possibility of using four different color dyes, one for each nucleotide of DNA. This resuscitated Huang's idea in a new form. They discussed the idea with Smith, who was receptive. Smith pointed out that color dyes would not work because they would not be sufficiently sensitive to reveal the minuscule amounts of DNA in a gel. He suggested fluorescent dyes. With the Hunkapillers and Smith all enthusiastic, the balance tipped. Mike's support was especially critical, since he was Hood's trusted lieutenant.

It was a very long road, however, from idea to instrument. Smith donned the yoke at Caltech, and Applied Biosystems mounted its separate but interwoven effort. Tim Hunkapiller did not work directly in the group that Smith forged to make the DNA sequencer. Tim instead went on to focus on the use of computers to analyze information derived from molecular biology and became a prominent national figure in discussions on the subject.

While Applied Biosystems was hard at work in California, several other companies were also developing DNA sequencing machines. In 1987, the

---

Automated DNA sequencing technique is used to determine the order of the A's, C's, G's, and T's in a cloned sample of DNA. The DNA is first heated to separate the strands. The single strands are then cut at various points, and the resulting fragments are added to a mixture of fluorescent dyes in which strands that end in a specific letter are all tagged with the same fluorescent dye. The mixture is next run through an agarose gel, which in turn is passed through a laser beam. The shorter fragments, which travel farther in the gel, are exposed to the beam first, followed by successively longer fragments. Each base (A,C,G, or T) emits a different fluorescent color corresponding to its position in the DNA sequence. The DNA sequence can be inferred from the sequence of fluorescing colors emitted as the fragments in the gel go past the laser beam.


DNA of unknown sequence
heated to separate strands
single strand
of DNA
A
C
G
T
fluorescent
dye labels
dye
label
fragments
run on
agarose gel
laser beam
shorter fragments
exposed to
laser beam first

chemical and pharmaceutical giant Du Pont announced a different method to use fluorescent dyes for DNA sequencing[34] and began to market its Genesis 2000 instrument within a year. EG&G Biomolecular, one of scores of EG&G companies that originally spun off from MIT in the postwar period, marketed another machine that was based on detecting radioactive phosphorus. The EG&G machine cost considerably less than the fluorescent instruments and was based on methods quite familiar to molecular geneticists.[35] It was aimed at sequencing projects on the scale undertaken by the average laboratory, however, not on the mega-sequencing scale envisioned by genome enthusiasts.[36]

Piecing together the history of the DNA sequencer revealed the tensions between science and industry in a highly competitive environment. Such tensions over priority and who had which idea first permeate science, and money intensified the conflict. The DNA sequencer became a source of great pride in collaboration, but also of some friction between Applied Biosystems and Caltech and among the major collaborators who made the first successful machines. Control over the underlying patents and royalties hung in the balance, in addition to scientific credit for originating an important new technology. The final truth was that no individual could take full credit.

Just as Nobel selection committees were perpetually unfair in conferring a prize on "winners" in science—ignoring the way science had changed so that most major advances required the efforts of hundreds, not one or two—choosing among the contributors to a technological development was not just perilous, it was nonsense. Lee Hood clearly created the working environment that harnessed the talents of those interested in technology and biology. Henry Huang kept the idea of automated DNA sequencing alive long enough to pass the baton. Lloyd Smith directed the team that fabricated the Caltech prototype DNA sequencer, as Mike Hunkapiller had done before him on protein sequencing. Tim Hunkapiller linked the analytical instruments to their computer interpretation. An Applied Biosystems team built its own prototype instrument that was quickly and successfully commercialized and spread throughout the world. No villains misappropriated the property of others, but competition for priority—and perhaps money—loosened the bonds of trust. Success shattered the collegium of science.

The commercial overlay of technology development also provoked international tensions. In this regard, the Caltech–Applied Biosystems group was regarded collectively as a national treasure. Norman Anderson, who had worked at several national laboratories over the years and had himself developed several instruments for biological research, chatted with me about U.S.–Japan trade tensions at a human genome meeting in July 1986. Referring to Japanese economic competition, he saluted Hood by saying, "Thank God he's on our side."[38] The United States was hardly alone in developing DNA sequencing machines and other instrumentation for molecular biology. There was move-

ment aplenty on the other sides. The United States had a lead, but Europe and Japan were in the contest.

In Japan, the Science and Technology Agency (STA) began in 1981 to support a project to automate DNA sequencing. This program was the brainchild of Akiyoshi Wada, who was handed a mantle from STA to improve the analysis of DNA. He chose to focus on instrumentation and to automate well-established techniques, rather than simultaneously to develop new methods and automation technologies. Wada enticed several corporate sponsors (Fuji Photo, Seiko, and Matsui Knowledge Industries) into the project, which was eventually housed at the RIKEN Institute in Tsukuba Science City.[39–46] An independent automation effort at Hitachi culminated in a DNA sequencer that in 1989 was marketed only in Japan. (The Japanese genome program is discussed in detail in Chapter 15.)

The automation effort at the European Molecular Biology Laboratory in Heidelberg began in the early 1980s, supported by several European governments. Wilhelm Ansorge directed a team that produced an instrument that used fluorescent-dye detection. The EMBL instrument employed a scheme of fluor-labeled bases somewhat different from both the Caltech and the Du Pont designs.[47] The EMBL prototype served as the basis for the ALF DNA sequencing system marketed by LKB-Pharmacia, a Swedish company, beginning in 1989. The ALF system had complementary strengths to the Applied Biosystems approach, and large-scale sequencing projects used both.[2]

The ability to synthesize and sequence proteins and DNA revolutionized molecular biology; automating these tasks promised to consolidate the revolution. Hood and others saw automation as enabling assaults on larger and harder problems. He, Mike Hunkapiller, and Lloyd Smith preached that sequencing would continue to accelerate and that dedication to massive sequencing initiatives should await the development of better technology and better strategies for employing it. Automation would make currently impractical goals more practical.[48–50] Hood projected impacts in the practice of diagnosis and treatment, well beyond basic research applications.[51]

As the Human Genome Project shifted from a topic of debate to an ongoing research program in the early 1990s, it was becoming clear that large-scale DNA sequencing efforts would prove to be more than mere quantitative extensions of existing technology. After several years of research a second generation of automated "sequenators" appeared on the verge of development into instruments. New techniques seemed capable of increasing the speed of sequencing, reducing the amount of DNA needed to derive a sequence, and extending the number of bases that could be sequenced at a time.[52] Improvements seemed likely to be ten times or so more efficient at deriving sequence information from DNA prepared for sequencing. The steps for preparing DNA and the computer algorithms to piece sequences together began to loom as the main obstacles. Dedicated sequencing on a grand scale would entail

much greater attention to the accuracy of data, quality control of laboratory practices, quantitative assessments of the validity of base pair assignments, sophisticated algorithms to weave long stretches of sequence data together, mathematical and computer methods to make comparisons of sequence data, and generally a more systematic production mode of operation.[53]

DNA sequencing caught the fancy of those who saw a new way to do biology. A simple but revolutionary new technique for producing enormous amounts of short stretches of DNA, called the polymerase chain reaction (PCR), promised yet another revolution.

Kary Mullis invented the PCR technique while working for Cetus Corporation. Mullis was prone to flights of fancy and emotive explosions. Of the three qualities needed to excel in corporate research and development—creativity, willingness to work hard, and an ability to work with others—he was off the charts on the first, passable on the second, and encountered difficulties with number three. He conceived of PCR in 1983, and it eventually worked. Mullis's Ph.D. was in chemistry, and he was hired to work on DNA synthesis at the University of California, San Francisco Medical Center, during the late 1970s. In 1979, he was hired by Cetus, a company named for the whale in its logo, in Emeryville across the Bay. At Cetus, Mullis eventually headed the group that made short stretches of DNA for experiments. Synthesizing DNA and using it to detect specific sequences led him to think about ways to copy DNA without having to clone it in bacteria, yeast, or other organisms. PCR was a powerful and simple technique to do just that.

PCR was an enormously powerful technique. By 1989, its revolutionary implications had begun to emerge. In genome mapping, the rapidity and simplicity of the technique enabled genome researchers to contemplate using the common language—short stretches of defined DNA sequence taken from PCR reactions—to merge genetic linkage maps and a variety of physical maps.[54] C. Thomas Caskey, a genome research director from Baylor University, introduced Mullis at a 1990 DNA sequencing conference, asserting that the genome project itself had become practical only in the wake of Mullis's discovery.[55] PCR was a godsend.

While heaven-sent, PCR did not arrive by direct descent. It came by a more circuitous route, namely Route 128 overlooking the Anderson Valley in Mendocino County, a "moonlit mountain road into northern California's redwood country."[56] Mullis struck upon the idea while driving along it with his coworker one Friday evening, for a weekend respite at his cabin.[56;57] Later, back at Cetus, "I ran my favorite kind of experiment: one involving a single test tube and producing a yes or no answer. Would the PCR amplify the DNA sequence I had selected? The answer was yes."[56]

Mullis presented his idea as a scientific poster at Cetus's scientific retreat in June 1984. It was a troubled period in his personal life, and Mullis blew his fuse. He got into a late-night shouting match with another Cetus scientist,

and was so combative that he kept many awake until past three in the morning with phone calls and bouts of yelling. Someone finally called a private security officer to walk Mullis along the beach until he calmed down. His tenure at Cetus was seriously in question: Mullis was creating havoc—shouting, threatening a coworker who was going out with his erstwhile girlfriend, arguing with Cetus's evening guards when he didn't have his badge to enter the building after hours. Rather than fire Mullis, Cetus ended his duties as a head of the DNA synthesis laboratory and let him pursue the PCR idea full-time.[58]

By June 1985, Mullis and many other Cetus scientists and technicians had made it work. PCR had gone from an idea to a demonstrated technique of great promise. It was beginning to be used throughout the company. The question of when and whether to publish it was discussed at several meetings. Mullis was to write up the basic idea promptly, and another team was to work on a paper using PCR to detect the sickle-cell mutation, as the first application to a practical problem. Mullis was slow to write up the basic method, becoming obsessed instead with producing fractal images on the company computer, and Cetus decided to let the sickle-cell paper go ahead, with Randall Saiki as the first author, because he had generated the data. Saiki presented the results in October. The paper was published in the December 20 issue of *Science*.[59] Mullis's paper, however, was rejected—first by *Nature* and then by *Science*. The importance of the technique and the numerous variations not covered in the Saiki paper were apparently lost on the reviewers. Like art critics in the time of Cézanne, they missed the point.

Mullis was not listed as senior author of the work cited as the standard first paper on PCR, an unfortunate consequence of his wanting to include additional experiments, his procrastination, and the short-sighted reviews at *Science* and *Nature*. Mullis had not been eager to publish the technique in the first place, and now he was cut out of the traditional standard of due credit. Most molecular biologists first heard of his technique only when he presented it at the June 1986 symposium, "The Molecular Biology of *Homo sapiens*" at Cold Spring Harbor Laboratory on Long Island, where his talk duly impressed the cognoscenti.[60] This presentation was arranged when a Cetus scientist called James Watson, who was organizing the meeting, to alert him to this important new technique. In his Cold Spring Harbor paper, Mullis noted the process of copying DNA repeatedly "seems not a little boring until the realization occurs that this procedure is catalyzing a doubling with each cycle in the amount of the fragment defined by the positions" of the short synthetic DNA stretches inserted into the reaction.[60] With the laws of exponential arithmetic, a series of doublings quickly amounted to a lot of DNA copies. The audience included many of the most esteemed molecular biologists in the world, and Mullis got center stage.

The original description of the method, along the lines of the paper rejected by *Science* and *Nature*, was finally published in a specialized journal only in late 1987.[61] In 1989, *Science* hailed PCR and its molecular accouterments as the

first exemplars in an odd new annual ritual of declaring a "molecule of the year."[62;63] *Science* thus belatedly recognized the fundamental importance of the technique, while neglecting to mention its earlier editorial mistake.

The publication experience caused the simmering tempers at Cetus to boil over. Mullis left Cetus in the summer of 1986, within months of his triumphant talk at Cold Spring Harbor.[58]

Copying DNA became possible in the 1970s with the advent of DNA cloning, and was a major advance, but PCR avoided the limitations and extra steps involved in molecular cloning. If what you wanted was information from a short stretch of DNA, rather than a long piece of DNA to study, PCR was the answer. PCR made it possible to amplify DNA fragments from much smaller samples of DNA, even down to the theoretical limit, a single molecule. PCR also made the copying process faster and cheaper and avoided the rearrangements and cell-culture manipulations in a bacterium, yeast, or other cell.

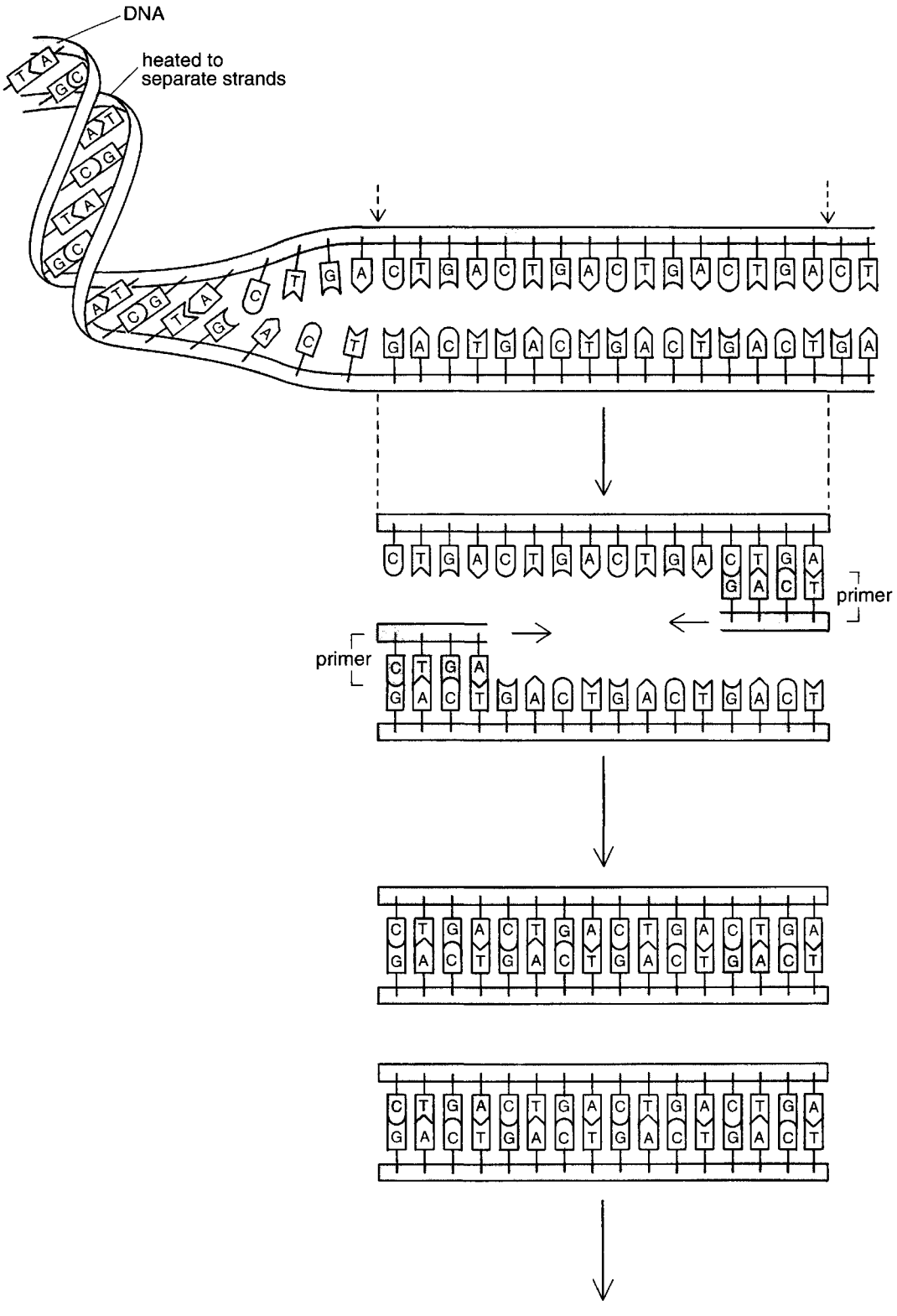
The PCR reaction was carried out in a test tube, using well-defined ingredients. These included the DNA to be analyzed, the nucleotide precursors to make new DNA, and a DNA polymerase enzyme. The first DNA polymerase was discovered by Arthur Kornberg at Stanford in 1955, in work that earned him a Nobel Prize,[64] and the intervening years had turned up a host of details about the process.[65] The PCR reaction built on what was known about how enzymes synthesized new DNA strands from existing ones.

The critical element for PCR was a set of DNA primers, short stretches of DNA of defined sequence at the ends of the DNA region to be copied. When bound to sample DNA, the primers created short regions of double-stranded DNA. The polymerase enzymes could not start making DNA strands from anywhere, but had to start from one end of a double-stranded region. Given such a starting point, as provided by primer that bound to native DNA, the enzyme could proceed to make a second strand complementary to the first. Twin primers pointing in opposite directions bracketed the DNA to be copied, thus defining the DNA region to be copied by PCR.

Starting with a very small amount of sample DNA and appropriate primers, a fragment could be copied (amplified) thousands, millions, or even hundreds of billions of times. The PCR reaction entailed heating the reaction mixture in cycles, to separate the DNA strands from each other. The original reaction

---

Polymerase chain reaction is used to rapidly amplify a specific DNA region. The region to be amplified is first bracketed by a pair of DNA primers, which set the starting points for making new strands of DNA, in opposite directions. The new strands of DNA are then separated, and copies are made again with a new pair of primers having the same sequence, so that only the same region of DNA is consistently copied. Each cycle leads to a near-doubling of DNA from that region. After dozens of repetitions, the process can yield billions of copies of the original DNA. The inventor of the PCR technique, Kary B. Mullis, was awarded the 1993 Nobel Prize in chemistry for this major contribution to genetic research.


DNA
heated to
separate strands
primer
primer

used a conventional DNA polymerase that had to be replaced with each cycle, making the process more complex and more expensive.[59] By mid-1986, the Cetus group was using a DNA polymerase enzyme isolated from a bacterium, *Thermus aquaticus,* found in hot springs, where bacteria routinely copied DNA at temperatures near the boiling point, as used in the PCR process.[66;67] This sped the process, obviated the need to replace enzyme, and had other technical advantages, such as facilitating automation.

PCR involved no esoteric instruments, only a reliable means to heat and cool the reaction mix and the use of an enzyme and chemical reagents. PCR was wondrously flexible. It made DNA analysis simple enough that laboratories formerly hesitant to do work at the DNA level could do so. It was used to trace human origins to Africa (although this proved more controversial than initially imagined), to detect infections rapidly and with exquisite sensitivity, to diagnose genetic disease, to study the complex evolution of the immune system, and for a plethora of other applications.[57;68] As PCR was being discovered, for example, the AIDS epidemic was just becoming known. During these same years, the virus causing AIDS was discovered. The problem of detecting infection was an urgent scientific problem, and detection of AIDS became one of the early applications for PCR. The first publication using PCR involved the detection of the sickle-cell gene,[58] demonstrating its usefulness in diagnosis of genetic disease. It was quickly applied to the diagnosis of infections,[69] to the study of immune function, and to the study of cancer. The generality and simplicity of the method created an entirely new market for reagents, heating-and-cooling instruments (called thermal cyclers), primers, primer synthesizers, and enzymes.[70] It spawned a mini-industry.

PCR became the subject of a patent battle when chemical giant Du Pont dug up a series of 1971–1974 papers from the laboratory of H. Gobind Khorana, an MIT Nobel laureate. Du Pont began marketing products for PCR amplification, challenging the Cetus patent and claiming the Khorana papers had placed the idea in the public domain. Cetus took Du Pont to court. In the first legal skirmish, Cetus emerged the winner as the U.S. District Court in San Francisco sustained Cetus's patent claims.[57;71–75] Chiron, a nearby biotechnology company, announced it intended to buy Cetus a few months later. PCR rights were sold to the Swiss pharmaceutical conglomerate Hoffmann–La Roche as part of the deal. Roche trumpeted its acquisition of PCR rights on December 11, 1991, and formed a new unit, Roche Molecular Systems. The deal gave Roche research and diagnostic rights, while Perkin-Elmer retained rights for reagents, instruments, and nondiagnostic applications.[76] The sale provoked *Nature* to query, "Is Cetus Selling the Family Silver?"[77] Cetus had licensed research and diagnostic uses of PCR to Perkin-Elmer and Roche several years earlier, so *Nature*'s question was too late and misdirected. Cetus licensed PCR to Hoffmann–La Roche because it simply did not have sufficient marketing and distribution capacity for medical diagnostics, and it was seeking a company that would more aggressively pursue PCR applications than the

previous licensee (Kodak). In February 1992, Roche announced a relaxation of licensing arrangements, eliminating up-front fees for academic and nonprofit institutions and setting a maximum royalty rate of 9 percent for them.[78] While the top brass at Cetus may have neglected PCR for too long, many down in the research and development trenches were well aware of its enormous potential and eager to move toward applications.

Mullis's insight came at a time when molecular genetic techniques had advanced sufficiently to make use of it. Techniques to make the short stretches of DNA used as primers artificially were laborious until automated instruments were developed early in the 1980s. To make useful primers, at least some sequence information on the target DNA was usually needed, so the practical application of PCR awaited facile sequencing techniques. Once it was described in 1985, PCR exploded through the molecular biology community. A bibliography compiled by Perkin Elmer Cetus listed three publications based on the technique in 1985, twenty in 1986, seventy-five in 1987, 280 in 1988, and 860 in 1989.[79] In 1990, Cetus stopped publishing the bibliography because it was growing too fast; by 1992, there were five hundred articles a month published using PCR. The point was proved.

Technological developments surged forward in genetic linkage mapping, physical mapping, and DNA sequencing in the period 1980–1986. These technical developments set the stage for a science policy debate that culminated in ideas for a concerted genome project. Before the project could emerge from the primordial technological soup, energetic people with vision and persistence had to champion new ideas and create institutions to sustain them. The technological groundwork was in place, but the Human Genome Project also required that the new technology be harnessed to a scientific project by securing a budget and establishing a bureaucratic structure. Several individuals independently brought forth their ideas for an audacious new biological enterprise in 1985.

PART TWO

Origins of the Genome Project

5

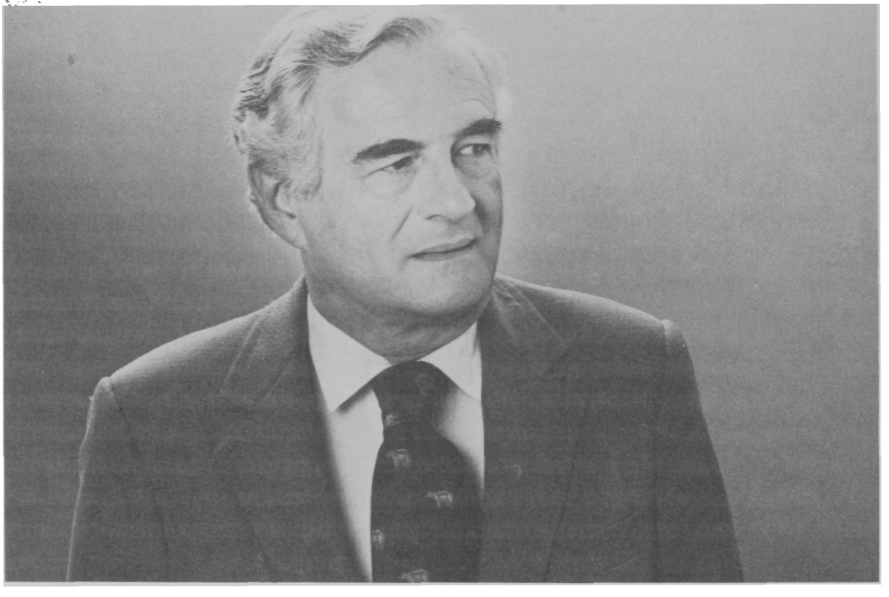
Putting Santa Cruz on the Map

THE FIRST MEETING focused specifically on sequencing the human genome was convened in 1985 by Robert Sinsheimer of the University of California at Santa Cruz. While the genome project did not grow out of the meeting, or even emerge as a topic of discussion, the 1985 Santa Cruz gathering did plant the seed.

Planning for Sinsheimer's May 1985 meeting at Santa Cruz began the previous October, when Sinsheimer called several faculty biologists—Robert Edgar, Harry Noller, and Robert Ludwig—into his office. Sinsheimer was then chancellor at UCSC. As such, he had been a participant in several major science planning efforts. These included relations with the three national laboratories managed for the Department of Energy (DOE) by the University of California (Los Alamos, Lawrence Berkeley, and Lawrence Livermore national laboratories), discussions of the California state proposal to house the Superconducting Super Collider, and, most directly, the Lick Observatory. The UCSC faculty in astronomy had an international reputation. As a biologist, Sinsheimer wanted biology to achieve similar stature. He wanted, he said, to "put Santa Cruz on the map."[1]

Others had previously conceived of large, concerted mapping projects and technology development, but these did not grow into the genome project.

The European Molecular Biology Laboratory had in 1980 seriously contemplated sequencing the entire 4,700,000-base-pair genome of the bacterium *Escherichia coli*,[2;3] but that project was judged technically premature. Norman Anderson, who had worked at several DOE-funded national laboratories during two decades, had a track record of devising instruments for molecular biology, including high-pressure liquid chromatography, two-dimensional protein electrophoresis, and zonal centrifugation.[4] He and his son Leigh lobbied during the late 1970s for a national effort to catalog genes and blood



Robert Sinsheimer, as chancellor of the University of California at Santa Cruz, convened the first meeting on sequencing the human genome in May 1985. Although the institute for human genome sequencing that he envisioned for the UC Santa Cruz campus never materialized, the impetus for such a project remained. *Don Fukuda photo, courtesy University of California, Santa Cruz*

proteins,[5] and Senator Alan Cranston pushed for a dedicated $350 million program in the early 1980s. Even then, there was talk of the need to collect DNA sequence data.[3] Father and son continued to urge adoption of their program in the national laboratory system and at DOE. Their efforts were known by DOE administrators, and may indeed have helped set the stage for the genome project, but they had not crystallized into a dedicated science program.

The inspiration for Sinsheimer's DNA sequencing proposal was a telescope.[6;7] A group of University of California astronomers wanted to build the biggest telescope in the world. The venture was ultimately successful, produc-

ing the Keck Telescope on Mauna Kea in Hawaii, which saw first light on November 24, 1990. This success came only after clearing several high hurdles.

In 1984, the costs of enlarging the giant telescope on Mount Palomar or constructing a facility of similar size were estimated in the range of $500 million, a large fraction of the expense associated with manufacturing an enormous mirror. Jerry Nelson of the Lawrence Berkeley Laboratory hit upon the idea of using thirty-six hexagonal mirrors to replace a single large one, reducing cost estimates eightfold. By computer adjustments of the hexagonal array, the complex of smaller and cheaper mirrors could provide the same resolving power. A piece about the telescope appeared in the *San Jose Mercury*. Soon after the article appeared, a Mr. Kane called the laboratory.[8] He thought he might know a donor interested in funding the telescope, the widow of Max Hoffman. Hoffman had made a fortune as the U.S. importer of Volkswagen and BMW automobiles, and had left an estate of several tens of millions of dollars, the Hoffman Foundation, whose trustees were his widow and two others. Mrs. Hoffman signed most of the papers for a $36 million donation for the Hoffman Telescope project the day before she died. It was the largest single gift in the history of the University of California, but it had to be returned. That $36 million return was the event that stimulated the DNA-sequencing idea.

The $36 million donation, generous as it was, fell $30 to $40 million short of what was needed to build the telescope. Further donors were needed, and the University of California was having trouble finding them. Since the telescope was already named for Max Hoffman, it was more difficult to entice further large donations. The University of California finally sought help from Caltech, a private university. The University of California got more than it bargained for. After finding several smaller donations, Caltech got an agreement from the Keck Foundation, built with Superior Oil money, to fund the entire telescope if the name was changed to the Keck Telescope. The Hoffman Foundation, having lost the glory of being the major donor and having lost its most interested trustee, was not interested in helping build a smaller sister telescope or in using its funds for other suggested alternatives.

Sinsheimer wondered if an attractive proposal in biology could recapture the interest of the Hoffman Foundation. He pondered whether there were opportunities missed in biology because of biologists' proclivity to think small, in contrast to their colleagues in astronomy and high-energy physics. Sinsheimer's laboratory had purified, characterized, and genetically mapped a bacterial virus, phi-X-174.[7] Its 5,386-base-pair genome was the first of any organism's to be sequenced, by Frederick Sanger in 1978.[9] Sinsheimer followed the progression of DNA sequencing to larger and larger organisms. As he thought about targets for a large biology project, Sinsheimer struck upon sequencing the human genome, fully a million times larger than the viral genome and ten thousand times larger than the biggest sequencing project to date. He sought

counsel of his colleagues at UCSC about establishing an institute to sequence the human genome, and in October 1984, he called the meeting with Noller, Edgar, and Ludwig.[10]

Edgar, Ludwig, and Noller were at first stunned by Sinsheimer's audacity, but as they began to think through the scientific approach that would lead to sequencing the entire genome, they decided that it would be a useful goal and would generate equally useful results along the way. In particular, the process of sequencing would entail physical mapping, a valuable enterprise in its own right. Edgar and Noller prepared a position paper for Sinsheimer on Halloween 1984, which became the basis for Sinsheimer's letter to University of California president David Gardner on November 19.[11] The Santa Cruz scientists proposed that the DNA sequencing institute could be

> a noble and inspiring enterprise. It some respects, like the journeys to the moon, it is simply a "tour de force"; it is not at all clear that knowledge of the nucleotide sequence of the human genome will, initially, provide deep insights into the physical nature of man. Nevertheless, we are confident that this project will provide an integrating focus for all efforts to use DNA cloning techniques in the study of human genetics. The ordered library of cloned DNA that must be produced to allow the genome to be sequenced will itself be of great value to all human genetics researchers. The project will also provide an impetus for improvements in techniques . . . that have already revolutionized the nature of biological research. . . .[12]

Sinsheimer urged Gardner to approach the Hoffman trustees with his new idea, asserting:

> It is a an opportunity to play a major role in a historically unique event—the sequencing of the human genome. . . . It can be done. We would need a building in which to house the Institute formed to carry out the project (cost of approximately $25 million), and we would need an operating budget of some $5 million per year (in current dollars). Not at all extraordinary. . . . It will be done, once and for all time, providing a permanent and priceless addition to our knowledge.[11]

Sinsheimer also discussed the idea with James Wyngaarden, director of the National Institutes of Health, in March 1985. Sinsheimer noted that Wyngaarden was "attracted by the idea," and he urged Sinsheimer to approach the National Institute of General Medical Sciences if the May meeting reached consensus on the project's feasibility.[13] Sinsheimer concluded he would have to find a source of funds. To do so, he would need the blessing of some internationally recognized scientists to lend the project credence.

The next phase was to call a meeting of experts from around the world. Noller wrote to Sanger, with whom he had worked several years earlier. Sanger's reply was encouraging: "It seems to me to be the ultimate in sequencing and will probably need to be done eventually, so why not start on it now? It's difficult to be certain, but I think the time is ripe."[14] Edgar, Noller, and Robert Ludwig convened the meeting on May 24 and 25, 1985, bringing together an eclectic mix of DNA experts. Bart Barrell was Sanger's successor as head of

large-scale sequencing at the MRC Cambridge laboratory. Walter Gilbert represented the Maxam-Gilbert approach to DNA sequencing. Lee Hood and George Church were Americans pushing sequencing technology, Hood through automation and Church (who had done his graduate work with Gilbert) through clever ways to extract more sequence data from each experiment. Those familiar with genetic linkage mapping were also invited, including David Botstein, Ronald Davis, and Helen Donis-Keller. John Sulston and Robert Waterston were invited to report on their efforts toward constructing a physical map of *C. elegans*. Leonard Lerman was a technologically oriented biologist from Boston, and David Schwartz had pioneered the techniques for handling and separating DNA fragments millions of base pairs in length. Finally, Michael Waterman of the University of Southern California was brought for his expertise in mathematics, DNA sequence analysis, and databases.

Over the course of an evening and a day, the group decided that it made sense systematically to develop a genetic linkage map, a physical map of ordered clones, and the capacity for large-scale DNA sequencing.[7] The first sequencing efforts should focus on automation and development of faster and cheaper techniques.[15] The workshop concluded, significantly, that a complete genome sequence was not feasible, as such an undertaking would require large leaps in technology. "In the meantime, one should concentrate on the sequencing of regions of expected interest (polymorphisms, functional genes, etc.). The first few percent should be of great interest."[15]

The idea of sequencing the human genome was out in the open. A later account of the meeting captured its modest aspirations as "Genesis, the Sequel."[6] Sinsheimer sent letters and a summary of the meeting to several potential funding sources, including the Howard Hughes Medical Institute (HHMI) and the Arnold and Mabel Beckman Foundation, but there were no takers.[16–18] Contacts with the Hoffman Foundation, while the initial impetus for the meeting, were not permissible. The University of California president's office now handled the foundation, so the Santa Cruz campus could make no direct approach. The NIH route was blocked by the need to ask for a facility in which to do the work and the large budget required. A major construction effort entailed approval from the UC system. NIH might be approached to fund the project, but not the facility in which to do the work, and not until the facility was built. These were formidable obstacles. Sinsheimer concluded the only solution was to find a private donor for the building first, but his access to large sources of private money also had to go through the UC president's office. The Hoffman funds were never recouped by the University of California.

Sinsheimer later reflected:

> I was certain of the value of the proposal. The human genome surely would someday be sequenced, once and for all time. The achievement would be a landmark in human history and the knowledge would be the basis for all human biology and medicine of the future. Why not now?[7]

Sinsheimer contemplated going directly to Congress. He discussed the institute idea with Leon Panetta, his congressman. Panetta was supportive, but indicated his awareness that proposals of such magnitude would have to go through the UC president's office.[19] Sinsheimer was frustrated in his attempts to cultivate interest in Gardner's office. As Sinsheimer neared retirement, prospects for a human genome sequencing institute at UC Santa Cruz quietly died. While he did not get his institute, the Sinsheimer Laboratory for biology was dedicated by UC president Gardner, Senator Mello of California, and Assemblyman Farr, with a public lecture by Charles Cantor, in February 1990.[20] The idea of sequencing the human genome moved on to other pastures, having acquired a life of its own.

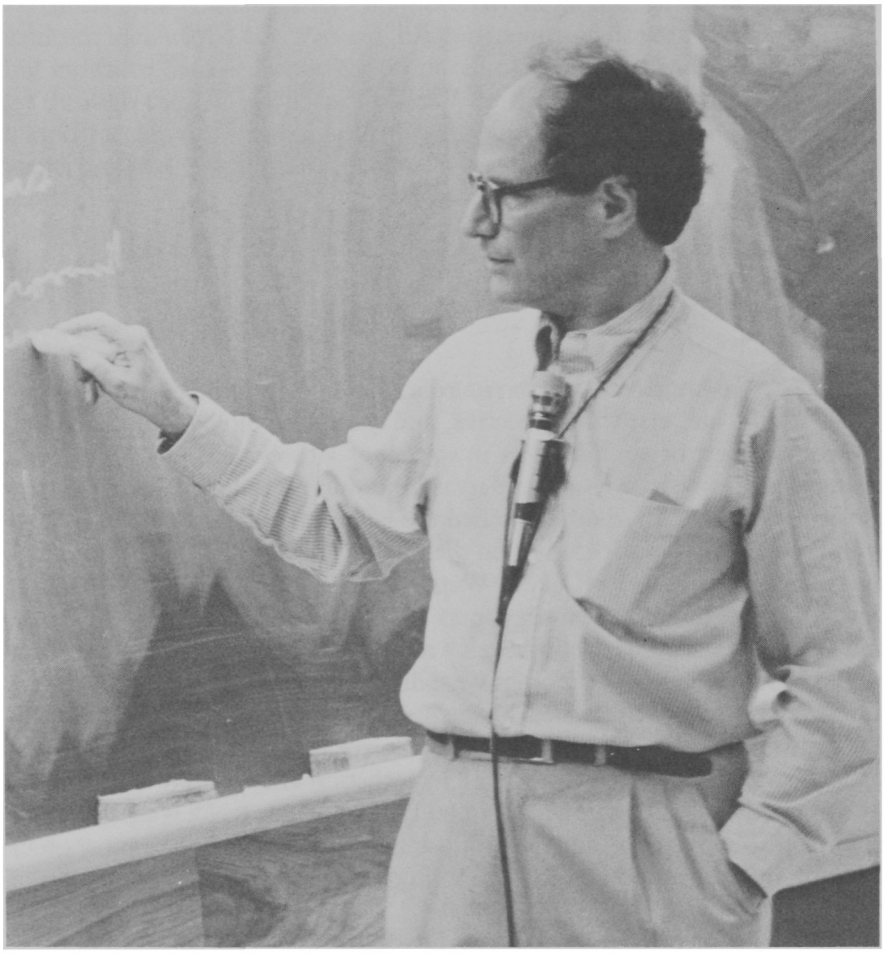
6

Gilbert and the Holy Grail

SINSHEIMER HANDED THE TORCH TO Walter Gilbert—Nobel laureate, erstwhile executive, and molecular biologist of legendary prowess. Gilbert began his career in science as a theoretical physicist. As an assistant professor in physics at Harvard, he wanted to learn about the new molecular biology. In 1960, he joined the laboratory of James Watson, who with François Gros was then hot on the track of messenger RNA. In a videotape taken of a meeting to celebrate Watson's sixtieth birthday in 1988, Gilbert described how he was given six papers to read when he first joined Watson and Gros, in contrast to the hundreds a new postdoctoral or graduate student would be handed today.[1] ("Things were different then.") Watson would hold a stopwatch while Gros sloshed a large flask of bacteria and Gilbert poured in ten to twenty millicuries of radioactive phosphate, to label the RNA in the bacteria. Messenger RNA was then a hypothetical entity, postulated to exist by some, but not yet a known commodity. Messenger RNA was, of course, eventually found to exist, and the group at Harvard joined those at the Pasteur Institute in Paris and the MRC Cambridge laboratory in the front ranks of molecular biology.

RNA is copied from stretches of DNA, then spliced, and finally transported out of the cell nucleus to serve as the code to assemble amino acids into proteins. Gilbert's career in molecular biology started with an extremely important problem. His reputation built even more on work that began in 1965 to find the repressor protein, an on-off switch for the gene that produced a bacterial protein. This was one of the most hotly contested races of its day in molecular biology. Gilbert commented on this phase of his work: "By the time the repressors were actually isolated, which was late in 1966, they had become a—Holy Grail?"[2] The mythic theme would return two decades later, by which time Gilbert was among the most respected thinkers in molecular biology.

Gilbert searched for the repressor protein with Benno Muller-Hill of Germany. The *lac* genes, involved in digesting sugars, were turned on and off in response to the presence or absence of sugars in the growth medium surrounding bacterial cells. The simplicity of the *lac* operon system made it a central target of molecular biology. Gilbert and Muller-Hill found the *repressor* protein



Walter Gilbert jots down his estimate of the cost and time it would take to sequence the entire human genome at a rump session of a symposium on the molecular biology of *Homo sapiens*, held at the Cold Spring Harbor Laboratory in June 1986. Gilbert left Harvard University in 1982 to become chief executive officer of the biotechnology firm Biogen; he returned to Harvard two years later and has been there ever since. *Victor McKusick photo, courtesy Cold Spring Harbor Laboratory Library*

that flipped this genetic switch in 1966.[3] It was a period of intense rivalry and cooperation with Mark Ptashne, who worked on a similar problem in a laboratory just down the hall.[4] Ptashne had come to Harvard to work under Watson and was trying to find a different repressor protein, one that turned genes on and off in the bacteriophage, or bacterial virus, named phage lambda.[5] Gilbert and Muller-Hill found their repressor just a few months before Ptashne found his. The next step was to study how the switch was thrown.

In the late 1960s and early 1970s, Gilbert isolated the DNA region that controlled the *lac* genes, called the operon, or genetic-switch region. This was the first segment of DNA isolated.[6] He chose to study the dynamics of the system by analyzing the structure of DNA in the region. This was the work that led to DNA sequencing, described in Chapter 4. Gilbert was thus a part of several landmark developments in molecular biology: the discovery of messenger RNA, the isolation of the *lac* repressor, and the technical miracle of DNA sequencing. It was not the end.

Gilbert joined a three-way race to isolate, study, and express the gene for insulin, one of the most studied proteins in all biology. Since its discovery in the 1920s, insulin had been used in treatment of diabetes. It was the first protein sequenced (by Sanger), and because of its therapeutic use, it was an obvious candidate protein to make using recombinant DNA technology as soon those methods were discovered in the mid-1970s. Gilbert threw his hat into the insulin ring in 1976. This and his past work took him on a short digression into commerce. Gilbert was among the founders of the Swiss-American biotechnology firm Biogen, created in 1978 while Gilbert's laboratory was working to clone insulin. Gilbert was enticed into involvement by a venture capital group hoping to establish the new company. At the scientific end, Gilbert's group at Harvard was the first to trick bacteria into producing the insulin protein, only the second mammalian protein ever so produced.[8]

Gilbert's star rose higher in 1980, when he shared the Nobel Prize for chemistry with Paul Berg of Stanford and Sanger. This was a special year for the Nobel, as these three scientists have a reputation as truly exceptional molecular biologists, even compared to other Nobel laureates. Each has not only left a significant personal legacy of science, but also left a trail of scientists trained in their laboratories and likely to travel to Stockholm themselves someday.

In 1982, Gilbert became chief executive officer at Biogen. Harvard forced him to choose between keeping his professorship and running a biotechnology company. He shook the academic world when he left his American Cancer Society chair at Harvard to direct Biogen. Biogen, however, did not fare well; it lost $11.6 million in 1983 and $13 million in 1984.[8] Gilbert resigned as CEO in December 1984 and returned to Harvard, where he became chairman of the department of biology. (Biogen continued to lose money after Gilbert left the helm.) In 1988, Gilbert was named Loeb University Professor at Harvard.

After leaving Biogen, Gilbert traveled to the South Pacific. The group organizing the Santa Cruz meeting sought him out, failing to locate him for many weeks. Robert Edgar finally reached Gilbert with a letter in March,[9] and Gilbert agreed to come. His addition was significant. After attending the Santa Cruz meeting, Gilbert became the principal spokesman for the Human Genome Project for the better part of a critical year.

Gilbert proved an articulate visionary, transmitting excitement to other

molecular biologists and to the general public. He translated the ideas at Santa Cruz into specific operating plans in a memo back to Edgar two days after the workshop. In it he offered a strategy for Sinsheimer's institute, although privately he was not convinced that it should be located in Santa Cruz:

> . . . In the early years the institute may want to be a sequencing resource—taking genes and probes from outside and returning sequences, cosmids [clones], and probes to the outside. . . . I expect that the most rewarding information scientifically will be in the first 1 percent of total sequence, if the work is focused, that most of the information, in the sense of interesting differences, will be in the next 10 percent, and the last 90 percent—of intron and intergenic regions—will be the least informative, but the increase in speed of sequencing should make each of these three phases take roughly equal times—or possibly make the last faster than the first.[10]

In this letter, he returned to a familiar motif, noting, "The total human sequence is the grail of human genetics—all possible information about the human structure is revealed (but not understood). It would be an incomparable tool for the investigation of every aspect of human function." Gilbert's Holy Grail proved an enduring rhetorical contribution to the genome debate. Indeed, it captured more than perhaps he intended. The Grail myth conjured up an apt image; each of the Knights of the Round Table set off in quest of an object whose shape was indeterminate, whose history was obscure, and whose function was controversial—except that it related somehow to restoring health and virility to the Fisher King, and hence to his kingdom. Each knight took a different path and found a different adventure.

Gilbert carried the ideas from Santa Cruz into the mainstream of molecular biology. He gave informal presentations on sequencing the genome at a Gordon Conference in the summer of 1985, and at the first international conference on genes and computers in August 1986.[11] Gilbert was extremely well connected, and he infected several of his colleagues with enthusiasm, including James Watson.

Gilbert gave the genome project much greater notice than it would otherwise have achieved. His role was featured in the *U.S. News & World Report, Newsweek, Boston* magazine, *Business Week, Insight,* and the *New York Times Magazine.*[12–17] He joined Watson, Hood, Bodmer, and others as the star of video documentaries on the genome project.[18] Gilbert and Hood wrote supporting articles for a special section in *Issues in Science and Technology* published by the National Academy of Sciences.[19;20] Gilbert and Bodmer promoted the genome project in editorials for *The Scientist.*[21;22] Gilbert thus stoked the genome engine, preserving the spirit of Santa Cruz.

Gilbert provoked a major controversy, however, when he decided to try to take the genome project private. He began thinking about establishing a genome institute himself in 1986. In January 1987, Michael Witunski, president of the James S. McDonnell Foundation, approached Gilbert with the idea of

foundation support to help create such an institute. This idea died when the foundation funded a study to assess the genome project at the National Research Council of the National Academy of Sciences. Gilbert participated in a spate of meetings convened to debate the genome project during late 1986, and he became a member of the NRC committee. In spring 1987, he decided to take the commercial plunge. He resigned from the NRC committee and announced plans to form Genome Corporation.

Gilbert's idea for Genome Corp. was to construct a physical map, do systematic sequencing, and establish a database.[6] The business objectives included selling clones from the map, serving as a sequencing service, and charging user fees for access to the database. The market would be academic laboratories and industrial firms, such as pharmaceutical companies, that would purchase materials and services from Genome Corp. The purpose was not so much to do things that others could not do at all, but rather to do them more efficiently, so that outside laboratories could purchase services more economically than they could perform the services themselves. In Gilbert's words, "Twenty years ago, every graduate student working on DNA had to learn to purify restriction enzymes. By 1976 no graduate student knew how to purify restriction enzymes; they purchased them. Historically, if you were a chemist, you blew your own glassware. Today, people simply buy plastic."[23] Genome Corp. could free biologists to focus on biology instead of wasting time making the things used in their experiments. These precedents fueled Gilbert's quest for funding from venture capitalists over the course of 1987 and into 1988. By late 1987, however, Wall Street's enthusiasm for biotechnology had turned to skepticism, and the stock market crash in October made capitalizing Genome Corp. all but impossible. The highly publicized efforts to start a genome project in the federal government made prospective investors leery of competing with the public domain. Genome Corp. could succeed only if Gilbert stayed so far ahead of academic competition that others would come to him for services, rather than waiting for the information and materials to be made freely available.

Gilbert was unabashed after the demise of Genome Corp. He remained a highly visible spokesman for a vigorous and aggressive genome project. He was consistently at the high end when making projections of what could be done in the way of mapping and sequencing. He was a technological optimist. Younger scientists balked at his enthusiasm for targeted, production-mode work and feared that he was publicly proclaiming goals too ambitious to attain. They loathed his almost monomaniacal focus on production-style DNA sequencing and bristled at his image of genome research as factory work. They complained bitterly that they would be held accountable for achieving impossible objectives set by policymakers listening to Gilbert; they felt they were being asked to climb Mount Everest after having only strolled a few miles along the Appalachian Trail.

If Gilbert was to blame for setting the sights too high, however, he would

at least be there on the firing line with the rest of genome researchers. Gilbert did not indulge in mere rhetoric, but committed his laboratory to be among the pioneers of large-scale DNA sequencing. In 1990, he proposed to sequence the genome of the smallest free-living organism, *Mycoplasma capricolum,* a small bacterium of goats.[24] This project was among the handful of sequencing projects intended to move sequencing from a theoretical possibility to a new way of understanding life. The genetics of the organism were not nearly so thoroughly studied as those of many other bacteria. Gilbert proposed to determine the DNA sequence of the bacterium's 800,000 base pairs, thought to contain five hundred or so genes. He hoped to reconstruct the biology of the organism by starting from its DNA sequence. The idea was not that sequencing would address all the questions of biological interest, but that starting from sequence would answer them faster.

Gilbert's project on *M. capricolum* joined other pilot sequencing projects on model organisms. These were among the grants given out in the first year's operation of the National Center for Human Genome Research at NIH.[24] A European consortium began a multicenter sequencing effort directed at yeast chromosomes. Botstein and Davis also proposed to start sequencing the yeast genome at Stanford (working from the physical map of yeast made by Maynard Olson). The groups working on the nematode *C. elegans* began systematic large-scale sequencing, in a transatlantic collaboration between John Sulston and Alan Coulson in England and Robert Waterston at Washington University in St. Louis.

Gilbert was not content to contribute only to the sequencing effort. His natural talents tended toward more theoretical generalizations. He was among the first to postulate an explanation of why genes were broken into different regions of DNA—with islands of base sequence to be translated into protein separated by long stretches of other sequences. In an article titled "Why Genes in Pieces?" he suggested that the role of fragmentation was to promote the shuffling of useful protein modules throughout the genome, enabling them to be used in different contexts.[25] Indeed, it was his terminology for DNA regions—"exons" for the parts that coded for protein and "introns" for the segments that separated exons—that eventually caught hold. Gilbert and, independently, Russell Doolittle postulated that the exon modules in DNA encoded protein substructures; these could be mixed and matched to serve similar functions in different proteins. They could be moved about in the genome over many generations, and the long intron sequences between the exons made this more feasible physically. Gilbert pushed the idea further a decade later, asserting in a controversial paper that nature had in fact settled on a relatively small set of structures to play with, several thousand or so, and built up the full complexity of existing organisms from a small fraction of the possible permutations.[26]

Gilbert also conveyed an ever enlarging vision of the role of molecular

genetics in biology, and the genome project in particular. He foresaw what science historian Thomas Kuhn had termed a "paradigm shift" in biology, with the science becoming driven more by theory. Molecular biologists would do experiments to test ideas first arising from the analysis of masses of information stored in computers. The cloning and sequencing that preoccupied the time of so many graduate students and postdoctoral fellows would be relegated to robots or specialized commercial services. "To use this flood of knowledge, which will pour across the computer networks of the world, biologists not only must become computer-literate, but also change their approach to the problem of understanding life. . . . The view that the genome project is breaking the rice bowl of the individual biologist confuses the pattern of experiments done today with the essential questions of the science. Many of those who complain about the genome project are really manifesting fears of technological unemployment."[27]

A genome program robust enough to sustain such a vision required a bureaucratic structure. The process of erecting this structure was at least as arduous as the science itself. At the beginning of 1987, as Gilbert formulated plans for Genome Corp., there was no center to support these and similar efforts in genome mapping and sequencing. Genome Corp. died, or rather was stillborn. While Gilbert despaired of federal leadership for the genome project, it was eventually two federal agencies that defined it. By the end of 1990, both the Department of Energy and the National Institutes of Health had genome programs with budgets totaling almost $84 million, and there were dedicated genome programs in the United Kingdom, Italy, the Soviet Union, Japan, France, and the European Communities. This remarkable bureaucratic transformation began late in 1985.

7

Genes and the Bomb

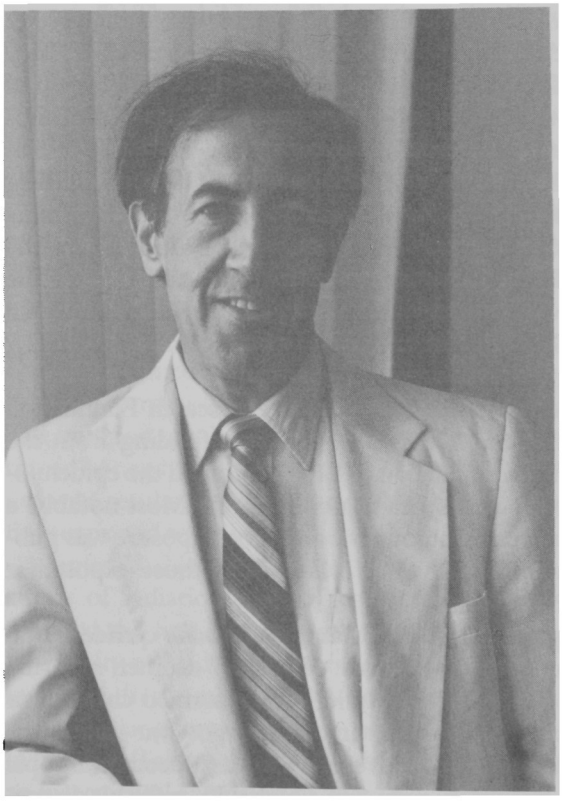
By proposing a Human Genome Initiative in the Department of Energy in 1985, Charles DeLisi thrust the Human Genome Project onto the public policy agenda. In so doing, he forced the ponderous bureaucracies at the Department of Energy (DOE) and the National Institutes of Health (NIH) into action. Several roots of DeLisi's genome research program can be traced back to the Manhattan District Project to build an atomic bomb. Some led through studies of the biological effects of dropping the bombs at Hiroshima and Nagasaki. Others led through the mathematicians who helped create the initial atomic bomb and, after World War II was over, the hydrogen fusion bomb.

In spring 1985, DeLisi became director of the Office of Health and Environmental Research (OHER) at DOE, the division responsible for funding the bulk of life sciences and environmental research for the department. The Nobel laureate physicist Arthur Holly Compton started the first biology project related to nuclear fission in 1942, at the University of Chicago, site of the first nuclear chain reaction.[1] He was aware of the dangers of radiation to workers, based on early experiences with X-rays and radium. Compton became one of the most important advisers to the federal government in the postwar period, chairing the Committee on the Military Value of Atomic Energy.[2]

Over the years, the mandate of the biological research program broadened considerably to include many biological effects of energy production, in addition to radiation biology. The bureaucracy underwent several reorganizations, from the Manhattan Project to the postwar Atomic Energy Commission (Public Law 79-585) to the Energy Research and Development Administration (Public Law 93-438). Jimmy Carter made a promise to create a Department of Energy in his 1976 campaign for President. The promise was made good in 1977 (Public Law 95-91), carrying with it the biology program that DeLisi later inherited.

In the period immediately after World War II, the Atomic Energy Commission (AEC) was a major supporter of genetics research. The AEC had a relatively large research budget at a time when the National Science Foundation was just coming into existence and the National Institutes of Health were quite small. Even the small fraction of the AEC budget devoted to genetics

dwarfed other genetics programs, and the national laboratories funded by AEC grew into centers on the forefront of research. This picture changed as the NIH budget increased steadily for three decades, leaving DOE in the dust. The National Institute of General Medical Sciences (NIGMS) became the principal funding source for basic genetics. Molecular biologists trained in the 1970s and 1980s were accustomed to thinking of NIGMS as the wellspring of genetics; older geneticists who might remember the AEC's role were smaller in number and generally separate from those who founded molecular biology.



Charles Delisi, as director of the Office of Health and Environmental Research in the Department of Energy, set aside the first funding for human genome research at DOE in 1985, in effect putting the genome project on the public policy agenda for the first time. *Courtesy Boston University*

DeLisi's idea for a DOE genome project spun off from an effort to study changes in DNA wrought in the cells of the atomic bomb survivors known in Japanese as the ***hibakusha*** ("those affected by the bomb"). They had been exposed to one of the most cataclysmic events of all time, but it was just the beginning of their collective nightmare.

The history of the genome project is linked to an attempt to determine if there would be a final, genetic wave of effects from bomb exposure. Specifically, investigators wanted to assess the frequency of inherited mutations caused by exposure to the atomic bombings. Those exposed to the bombings suffered

through many phases of radiation effects. Many people were vaporized, burned to death, or otherwise killed immediately by the bomb blast. Among those who survived the first hours, many died of radiation sickness that killed off cells in the immune system, skin, and intestinal lining. Fetuses *in utero* at the time of the bombing had an increased risk of microcephaly (small head and brain associated with mental retardation). Among burn victims, large deforming keloid scars formed in the months after exposure. A few years later, a wave of leukemias passed through the *hibakusha*. After a decade, they began to show somewhat increased rates of cancer in the breast, thyroid, gastrointestinal tract, bone marrow, and other tissues.

The *hibakusha* were severely stigmatized in the postwar period.[3;4] They were intensively monitored for decades with exhaustive medical follow-up of their health status, in one of the largest, most complex, and longest epidemiological studies ever attempted. In 1947, the U.S. National Academy of Sciences established the Atomic Bomb Casualty Commission (ABCC), with funding from the Atomic Energy Commission, to study the effects of the Hiroshima and Nagasaki bombs. The ABCC used legions of researchers to interview the *hibakusha,* eliciting details related to radiation exposure and health effects. The purpose of the ABCC was to gather information—not to provide treatment, a fact that aroused considerable resentment among the *hibakusha*.[3–5] Eventually, the Japanese government set up special health programs.

In 1975, the ABCC became the Radiation Effects Research Foundation (RERF), based in Hiroshima and Nagasaki, with joint funding from the governments of the United States and Japan. RERF continued the epidemiological investigations and conducted other related research. Most notably, a major reassessment of the nature and amount of radiation exposure was published in 1987, substantially changing dose estimates of those exposed at Hiroshima.[6]

One of the sources of stigma was a belief that the *hibakusha* carried mutations caused by the radiation they experienced. *Hibakusha* women reported they were rejected as mates because they would have deformed children or would pass on mutations and genetic disease. In the early postwar period, the extent of mutational damage to atomic bomb survivors was indeed a hot topic of controversy. H. J. Muller, fresh from receiving a Nobel Prize for his discovery that radiation could induce mutations, used his new fame to sound the alarms. Speaking of the *hibakusha,* he observed that "if they could foresee the results 1,000 years from now . . . they might consider themselves more fortunate if the bomb had killed them."[7] Alfred Sturtevant was even more apocalyptic about radiation exposure: in a letter to *Science,* he warned that atomic bombs already exploded "will ultimately result in the production of numerous defective individuals—if the human species itself survives for many generations."[8]

Such dire predictions were made by some of the most expert geneticists of the day. They fed a growing public fear of radiation that long predated atomic

bombs, but was greatly intensified by the mystery surrounding the Manhattan Project and its awesomely powerful products.[9] Nonetheless, the fears were products more of speculation than of observation. The speculations were not purely fabricated; they were based on animal studies, but in this case projections from other organisms proved errant, with distressing effects on the *hibakusha* and their children. The findings from extensive monitoring for three decades were contradictory: according to one expert, "the overwhelming impression that one gains from the analyses of the genetic data . . . is that there is not compelling evidence of genetic change in the offspring of exposed parents."[10] The children failed to show significantly higher rates of cancer or other disease, including birth defects and genetic disorders. If bomb exposure to their parents had produced inherited mutations, they were subtle and hard to detect among the DNA changes that normally occur between generations.[11]

The data were too sparse to drive choices among policies. While radiation clearly increased mutations, no one could say how many or what were the consequences in humans. Historian Susan Lindee concluded that "flexibility in the quantitative side of the argument contributed to flexibility in the 'acceptable' parameter."[5] One group of scientists noted that the species was unlikely to go extinct as a consequence of radioactive fallout, but this was small consolation to a public more interested in intermediate endpoints—the generations destined to live in the meantime.

An enormous range of interpretations was compatible with limited data. The question of whether the *hibakusha* suffered from heritable mutations continued to nag human geneticists. The ABCC studies were expected to produce negative results all along, an odd instance of a major commitment to a project fully expected to be inconclusive.[12]

James V. Neel and others devoted their careers to careful study of the effects of radiation on the genes of the *hibakusha* and their children. Neel founded the first department of human genetics in the United States, at the University of Michigan, based in part on funds to study the genetic effects of radiation. In the mid-1980s, a group sought to apply the emerging techniques of molecular genetics to the quantitative measurement of heritable mutations in humans. Taking the analysis down to the level of DNA sequence was merely an incremental extension of decades of work.

RERF convened a genetics study conference on March 4 and 5, 1984, in Hiroshima. Conferees recommended that cell lines be created from the *hibakusha,* and that "methods for direct examination of DNA should be introduced with all deliberate speed."[13] This recommendation could be interpreted any number of ways, and the International Commission for Protection Against Environmental Mutagens and Carcinogens elected to hold a meeting focused specifically on new DNA techniques. The Department of Energy funded the meeting. Mortimer Mendelsohn of Lawrence Livermore National Laboratory asked Ray White to organize the meeting.

White selected Alta, Utah, as the meeting site. At the same venue where

Botstein and Davis struck upon the idea of systematic RFLP mapping six years before, the masters of technology convened to discuss direct analysis of DNA. White invited an extraordinary mix of molecular and human geneticists to the meeting. The meeting, which lasted from December 9 to 13, 1984, took place in a blizzard. The skiing was memorable; the science was even better.

The 1984 Alta meeting planted the seeds for George Church's embellishments of the Maxam-Gilbert sequencing methods. Many of the young molecular biologists had never met Neel; indeed, some had never heard of him. Maynard Olson, destined to figure prominently in the genome story, was deeply impressed by Neel's commitment.[14] Olson was just beginning to get results on his physical mapping project of yeast. Charles Cantor presented some of the first data using the method he and David Schwartz described for separating million-base-pair fragments of DNA for mapping. The genetic linkage mappers, White foremost among them, had already found their first few RFLPs. Most of the participants had never met one another; as discussion heated up, the meeting became a boiling cauldron of ideas. The roiling broth within contrasted with the blizzard outside, isolating the participants from the world and lending intensity to the discussion.[15]

The conclusion of the meeting was, ironically, that the methods of direct DNA analysis were inadequate to detect the expected increase in mutation frequency from radiation exposure at Hiroshima and Nagasaki.[15–18] In attaining its specific end, the conference was a disappointment, but it brought together a welter of related ideas that would grow into the DOE genome project. The links were a congressional report and Charles DeLisi, a new face at DOE.

The congressional Office of Technology Assessment (OTA) was then doing a report on technologies to measure heritable mutations in man. Exposure to Agent Orange, environmental toxins, and radiation were coming before congressional committees as public policy problems.[19] Mike Gough, then an OTA project director, was present at the Alta meeting and discussed the various technologies in a draft report sent to the Department of Energy for review. The report was published in 1986 as *Technologies for Detecting Heritable Mutations in Human Beings*.

DeLisi had the idea for a project dedicated to DNA sequencing, structural genetics, and computational biology while reading the October 1985 preliminary draft of the OTA report.[20–23] DeLisi was then the newly appointed head of the Office of Health and Environmental Research at DOE. In a scene typical of Washington, he reflected on programs under his direction by reading about them in a report prepared by outsiders.

Once he had the idea, DeLisi moved quickly. He and David Smith, a scientist-administrator also working at DOE headquarters, barraged one another with notes and memos about how to plan this major new initiative. While most of the offices in and around Washington eased into the Christmas

lull, Smith and DeLisi were busy crafting a new science initiative. Smith and DeLisi asked the biology group at Los Alamos National Laboratory for comments on DeLisi's idea. The Los Alamos group replied with a dense, scattered, but wildly enthusiastic five-page memo just before Christmas, prepared by physician Mark Bitensky and others.[24] The memo bubbled over with enthusiasm about the potential technical and human health benefits that a structural approach to genetics would open up. The discussion centered on DNA sequencing and barely mentioned physical or genetic mapping. The Los Alamos group found another appealing argument for a concerted research program, arguing that such a project could become a "DNA-centered mechanism for international cooperation and reduction in tension."[24]

The memo saw the national laboratories emerging from the shadow of the atomic bomb. In Bitensky's words, "[J. Robert] Oppenheimer's statement 'I am become death, the Destroyer of Worlds' gives way to 'the National Laboratories are become the ultimate advocates for the understanding of human life.' "[24] He referred to Oppenheimer's quote from the *Bhagavad Gita,* uttered upon the explosion of the atomic fission bomb test at Alamogordo, New Mexico.[25–27] Los Alamos even checked with Frank Ruddle of Yale, to ensure that he would be willing to testify before Congress if called. With this initial encouragement, Smith and DeLisi began to pull the bureaucratic levers in Washington.

DeLisi outlined the political strategy to garner support from the scientific community, from their superiors at DOE, and from Congress.[28] Smith responded with a note about rumors of previous discussions, at a Gordon Conference and at the University of California the previous summer, but he did not know what had come of these.[29] Smith cautioned that criticisms would plague the DOE proposal for some time to come: it was not science but technical drudgery, directed research was less efficient than letting small groups decide what was important, and efforts should be concentrated on genes of interest rather than global sequencing. DeLisi bounced back: "Regarding the grind, grind, grind argument . . . there will be some grind; what we are discussing is whether the grinding should be spread out over thirty years or compressed into ten." He estimated that "we are talking about $100–150 million per year spread out over somewhat more than a decade," and he asserted that such a project certainly would rate as more important than the lower 1 percent of biology grants that funding of this magnitude would displace. The political effort, he argued, should focus not on whether it would displace other work, but instead on how to gain support for new funding.[30]

In order to reach out to the scientific community, DeLisi and Smith asked Los Alamos to convene a workshop: (1) to find out if there was consensus that the project was feasible and should be started; (2) to delineate medical and scientific benefits and to outline a scientific strategy; and (3) to discuss international cooperation, especially with the Soviet Union. A planning group at Los Alamos got together on January 6 to begin planning the workshop.[31] The

meeting was shaped in a series of notes and calls back and forth between DOE headquarters and Los Alamos.[32]

The workshop was held in Santa Fe on March 3 and 4, 1986, with "a rare and impassioned esprit."[33] Frank Ruddle chaired the meeting. Discussion at the Santa Fe workshop added an emphasis on integrating genetic linkage and physical maps and the process of making physical maps.[34;35] Participants agreed on the importance of the new venture and on part of what it should entail, but opinions failed to converge on how to organize the effort. Nobelist Hamilton O. Smith of Johns Hopkins University found that "perhaps the most impressive feature of the meeting was the unanimous consensus that sequencing the entire human genome is doable . . . [although] how to implement such a heroic and costly undertaking is less clear."[36] Anthony Carrano and Elbert Branscomb of Lawrence Livermore National Laboratory stressed the importance of clone maps and warned that "a program whose announced purpose was simply to 'sequence the human genome' might unnecessarily and incorrectly arouse fears of territorial and financial usurpation in the biomedical research community."[37] Events proved their political acumen; fears of a massive mindless sequencing operation became the major threat to scientists' support of the human genome project.

David Comings, a human geneticist from the City of Hope Medical Center in southern California, was further from the mark when he asserted that the whole physical mapping component might be funded "without any stirring up of any congressmen or other related creatures."[38] Those awful creatures proved altogether too alert and intrusive.

Beyond the first rationale, the study of heritable mutations, DOE had a second reason to mount a genome project. DOE managers wanted to capitalize on the resources of the national laboratories, with their ready access to exotic high technology, the best complex of supercomputers in the world, and multidisciplinary teams of scientists.

> The Genome Project also fit naturally within a broader DOE mission, and that is the utilization of the Labs to solve nationally important problems in areas that required their unique capabilities. In the case of Genome, the uniqueness was experience with large multidisciplinary projects, and a history of breakthroughs in applying engineering to the medical sciences (nuclear medicine being the paradigm). To the extent that large portions of the project could not be comfortably accommodated at most universities, this second rationale ultimately became as important as the first.[39]

This justification was liable to seem self-serving, however; the arguments sounded like a typical bureaucracy's merely expressing its proclivity for self-perpetuation. And so it was. David Botstein showed his knack for subtle understatement, calling the DOE genome initiative "DOE's program for unemployed bomb-makers."[40] Lee Hood was more diplomatic, noting:

> The argument they had enormous technological resources that could be focused on this problem was utterly irrelevant, unless they had the key individuals that could

integrate those in a focused and productive way, to take advantage of biology as well as the technology. So on all of those counts, I think DOE had not convinced the world in 1985 that they had the wherewithal to take on the Human Genome Initiative.[41]

The future of the national laboratories proved crucial to the DOE's genome effort. Mutation detection was the intellectual origin, but it was too weak a foundation on which to build a major new program. A new direction for the national laboratories, to channel their ample intellectual and technological energies, became a much more powerful drive once engaged. The laboratories were a natural political base with a well-developed support structure. Scientists at several of the national laboratories were enthusiastic about the idea and were already doing related research. DeLisi's idea started from a narrow base, mutation detection, but then grew to encompass a much larger political goal, the salvation of the national laboratories.

DeLisi discussed the possibility of a genome project with his immediate superior, Alvin Trivelpiece, who supported it and charged the DOE life sciences advisory committee (the Health and Environmental Research Advisory Committee, or HERAC) to report back to him about it. Trivelpiece and DeLisi had discussed why DOE did not have the same high stature in biology that it had in high-energy physics, and they aspired to lift DOE to the forefront of biology on the wings of a genome project. Trivelpiece, as director of the Office of Energy Research, reported directly to the Secretary of Energy (then John Herrington), who in turn reported directly to the President.

On May 6, 1986, six months after his initial idea, DeLisi produced an internal planning memo to request a new line-item budget. This went to Trivelpiece and up through the DOE bureaucracy. DeLisi argued for a two-phase program. Phase I had three components. The first, physical mapping of the human chromosomes, the central element, would take five or six years. The other two components were development of mapping and sequencing technologies and renewed attention to how computer analysis could assist molecular genetics (especially sequence analysis). As physical mapping progressed, parallel efforts would proceed, to prepare for Phase II, the sequencing of the entire genome. High-speed automated DNA sequencing and enhanced computer analysis of sequence information were both essential to making the transition from Phase I to Phase II. DeLisi's background in computational biology, his previous experience in interpreting DNA sequence information at the National Cancer Institute, came to the fore here. Phase II, contingent on success in all three parts of Phase I, was to sequence the banks of DNA clones that constituted the physical map.

DeLisi spoke of a project analogous to a space program, except that it would entail the efforts of many agencies and a more distributed work structure, with "one agency playing the lead, managerial role. . . . DOE is a natural organization to play the lead."[42] A six-year budget of $5, $10, $19, $22, and $22 million was proposed for fiscal years 1987–1991.[43] Plans survived the

internal DOE review, and a series of meetings was scheduled, beginning in July 1986, with Judy Bostock, the DOE life sciences budget officer in the presidential Office of Management and Budget (OMB), and with her boss, Thomas Palmieri.

OMB perches atop the federal bureaucracy, with responsibility to oversee management and prepare the President's budget request to Congress each year. Mention of OMB sends shivers of fear down the spines of most who work for the federal government. OMB is the dank home of malicious obstructionists and ax-toting budget officers. The genome project charged into the dark castle—the New Executive Office Building a block from the White House—to face the naysayers and dream-stealers. As the exception that proves the rule, the genome project got a major boost from OMB.

DeLisi's genome meetings with Bostock were focused on planning for fiscal years 1988 and beyond. Bostock was an erstwhile physicist from MIT, intrigued by prospects of improving the speed and efficiency of biological research, who believed that better instrumentation could improve the quality of biology.[44;45] She saw molecular biology as an extremely inefficient process with postdoctoral and graduate students doing mindless manual work that would be better done by robots or automated instruments. DeLisi was proposing a program to analyze DNA faster and with less human effort, a laudable goal that capitalized on the resources of national laboratories. Bostock bought DeLisi's plans, clearing a major obstacle from the road to Congress.[46]

DeLisi succeeded in his dealing with the DOE and OMB bureaucracies, but he also needed an endorsement from scientists. The OHER advisory committee, the Health and Environmental Research Advisory Committee (HERAC), endorsed the plan for a DOE genome initiative in a report from its special *ad hoc* subcommittee. The subcommittee was a blue-ribbon scientific group chaired by Ignacio Tinoco, a highly respected chemist from the University of California at Berkeley, then on a sabbatical year at the University of Colorado. The HERAC report urged a budget of $200 million per year and made a case for DOE leadership of the effort. The introduction to the report laid out the rationale:

> It may seem audacious to ask DOE to spearhead such a biological revolution, but scientists of many persuasions on the subcommittee and on HERAC agree that DOE alone has the background, structure, and style necessary to coordinate this enormous, highly technical task. When done properly, the effort will be interagency and international in scope; but it must have strong central control, a base akin to the National Laboratories, and flexible ways to access a huge array of university and industrial partners. We believe this can and should be done, and that DOE is the one to do it.[47]

Budget projections made by the committee were not directly coupled to the multiyear DOE-OMB budget agreement. The HERAC report was issued in April 1987, at least seven months after DeLisi began to reprogram funds,

and four months after the budget agreement with OMB.[48-51] The process of formulating a budget began with DeLisi's notes to David Smith in December 1985 and continued more broadly at a genome conference hosted by the Los Alamos National Laboratory in Santa Fe, New Mexico, in March 1986. In letters sent to the organizers after that meeting, budget estimates covered a wide range and generally focused on only one or two components. By the second Santa Fe conference in January 1987, planning had become more systematic. Several of the participants met over lunch at that conference to discuss what the budget should be. David Padwa, who had previously been involved with founding an agricultural biotechnology company, Agrigenetics, noted some political constraints on the budget. It had to be large enough to command congressional attention, so it would have to be at least $50 million to $100 million per year, but it could not be so large it threatened other research interests. The discussion continued at a meeting of the HERAC subcommittee at the Denver Stouffer's Hotel, February 5 and 6, 1987, a month before their report was to be considered by the full HERAC. Generating cost estimates was delegated to Lee Hood. The second day's meeting started at nine in the morning, and Hood's plane was delayed, so the group began to discuss what could be done within the range of budgets thought to be reasonable for OHER to request. There was discussion of how much physical mapping and sequencing could be done with $20 to $40 million, the maximum thought politically feasible.

Hood entered the meeting at ten o'clock, armed with some handwritten notes, including a menu of technologies and attendant costs. The proposal included technology development, physical mapping, mapping and sequencing of model organisms (yeast and bacteria), and regional sequencing of interesting chromosomal regions (e.g., those packed with genes). His estimates were $200 to $300 million per year for a full program. Someone asked if that was at all possible, since it was a full order of magnitude higher than earlier discussions. Hood did not wait for an answer, and asked passionately whether the budget would drive the vision or the vision would drive the budget. With this, the group deliberated over some technical details of how to make the projections and settled on a figure of $200 million. This brought the budget projection into the range judged politically attractive over the course of previous discussions.

The HERAC subcommittee did not discuss which agency should lead the Human Genome Project at its final meeting to draft its report. This was pointed out to HERAC when it met to consider the subcommittee report in March 1987. By April, when the report was released, Tinoco as subcommittee chairman and Mort Mendelsohn, a member of the subcommittee and chairman of HERAC, had canvassed the members. They wrote the language favoring DOE leadership. Later interviews with members of that subcommittee revealed that at least seven of the fourteen had reservations about giving DOE a blank check, but agreed to the suggested language because they feared inac-

tion on the part of NIH; it was more important to them that the project proceed than that NIH direct it.

Despite the go-ahead from his superiors at DOE, from OMB, and from the scientific community as represented by HERAC, DeLisi's job was still not complete. There was a two-step process in each house of Congress. Before a federal agency can fully implement a major new initiative, Congress has to authorize it and separately appropriate funds for it. These twin processes are interdependent but distinct.

Appropriations committees in the two houses are parallel. They allocate funds according to the executive department expending the funds and follow a relatively stable annual routine. The President's budget proposal is prepared, first by each department and then by OMB. In January the President's budget goes to Congress, where it is referred to the appropriations committees. Except in unusual circumstances (as occurred once during the Reagan years, violating the spirit, and probably also the letter, of the Constitution), the House takes action first, and the Senate works from the House figures. The appropriations committees cannot authorize new programs, but can only fund activities authorized by other committees. The interpretation of these distinctions can be tight or loose, depending on the circumstances. (One of the nation's first large science agencies, the U.S. Geological Survey, for example, was created and operated for years under a rider to an appropriations bill, without an authorization statute.)[52;53]

To get the genome program started, DeLisi took $5.5 million in funds from the preexisting fiscal year 1987 budget and reallocated them to the genome effort. Such limited "reprogramming" was standard fare, permitted by the appropriation and authorization committees within reasonable limits. For 1988 and later budgets, however, DOE needed support from its authorization committees. DeLisi noted the need for congressional action in his first personal note to David Smith,[28] and he began to hold meetings with congressional staff in 1986. This was unfamiliar territory for DeLisi, who was given to shyness and new to defending a program on Capitol Hill. There was little problem in the Senate, as DOE could in all likelihood count on strong support from Senator Pete Domenici and tacit approval of Senator Wendell Ford, the key figures on the authorization committee. Domenici also sat on the appropriations and budget committees. The problem was in the House.

Staff of the relevant DOE authorization subcommittee in the House were getting mixed signals about the DOE genome initiative. Congressman James Scheuer chaired the subcommittee with jurisdiction over DeLisi's program. Scheuer's staff read the generally negative response to DOE's plans in *Science* magazine; phone calls to biologists elicited both support and opposition. Eileen Lee, the biologist on staff, was uncertain what tack to take. She called on OTA staff, including me, to help plan a hearing, in hopes of penetrating the network of scientists concerned with the genome project.

DeLisi's problem was complicated by the politics of his other programs. Scheuer's staff was generally supportive of DOE staff initiatives, but DeLisi had problematic relations with Eric Erdheim, staff for Claudine Schneider, the ranking minority (Republican) member on the subcommittee. It was unclear to Scheuer's staff whether they should expend the political capital to defend DeLisi against Erdheim on the genome project. Claudine Schneider was generally suspicious of DOE's record on research into environmental health hazards, although she eventually decided DeLisi's program was good. As the hearing approached, the genome project became the battleground for a skirmish between Democrats and Republicans on the subcommittee staff.

About a week before the hearing, I was invited to meet with subcommittee staff from both parties. I could sense the tension in the room, but was blithely unaware of its origin,despite the fact that my wife, Kathryn, worked in Claudine Schneider's office at the time. As we were drifting apart after the meeting, Eileen Lee whispered to me that she thought Erdheim had asked James Watson to testify against the DOE genome program. A few minutes later, as I was preparing to leave the subcommittee's rabbit warren of offices, Erdheim took me aside to tell me he was thinking about calling the Delegation for Basic Biomedical Research to seek testimony from Watson or David Baltimore. Erdheim "had problems with what DeLisi was doing in his programs,"[54] and he was skeptical of the genome proposal. What did I think of that? I suggested that he had better find out what Watson or anyone else from the delegation would say before he invited him.

Eileen Lee arranged for Leroy Hood to testify before the committee. Hood agreed, oblivious to the political maelstrom swirling around him. At the March 19 hearing, he delivered an impassioned plea for the genome project.[55] Hood asserted a role should be found for DOE, NIH, and NSF. He thus deftly if unwittingly ducked the troublesome question of which agency should hold the reins. Scheuer's staff had agonized about the possibility of a Hood-versus-Watson contretemps, but Watson did not show. (Watson later said he was never asked to testify.)[56]

Rep. Schneider's latent distrust broke the surface in a series of questions about forthcoming DOE reports on health effects of radiation among submarine workers, radiation effects among the *hibakusha,* health effects in nuclear plant workers, and "least cost" energy. (DeLisi later noted that Schneider praised these reports when she got them.)[39] Despite the dramatic warning signals, the genome program coasted through the hearings unscathed. The DOE program was probably more vulnerable at this hearing than at any other point in its evolution. DeLisi, unaware of the backroom shenanigans, had cleared his highest hurdle.

The appropriations process was less troublesome than authorization and presented no major obstacles once the genome project had OMB approval. The DOE budget process for fiscal years 1988 and 1989 held true to the initial agreement with OMB—seeking $12 million and $18 million, respectively. It

began to exceed the initial agreement only in 1990, when it sought $28 million instead of the original $22 million.

After the March authorization hearing before Scheuer's subcommittee, I escorted Hood (who did not know me then) to the elevators and out to catch a taxi, through the labyrinthine Rayburn House Office Building. He asked, "Is that it?" I asked what he meant. He replied, "Do we get the money?" I was struck, not for the first time, by how much of the process that went into federal research funding was unknown to even the most sophisticated of its recipients. I said something about this being just the first of many steps toward DOE's budget. It was far from a done deal. Hood dashed into a cab and headed for National Airport. He was a long way from home.

DeLisi's ideas found fertile soil in the U.S. Senate, but for reasons different from his own. Senator Pete Domenici was a staunch supporter of the national laboratories in his home state of New Mexico, although he believed that they produced far less long-term benefit for the local economy than they should. He convened a panel of influential policymakers to discuss the future of the national laboratories one Saturday morning, May 2, 1987, in the U.S. Capitol. The meeting featured Barber Conable, a former New York congressman and head of the World Bank; Donald Fredrickson, former director of the National Institutes of Health; Ed Zschau, former California congressman and successful entrepreneur; Jack McConnell, director of advanced technologies for Johnson & Johnson; and the directors of several national laboratories.

In the midst of the meeting, Domenici asked, "What happens if peace breaks out?"[57–60] The bulk of the work supported at the two laboratories in New Mexico was focused on nuclear-weapon production and defense-related research and development. Domenici wanted to know how the immense research resources of the national laboratories could be better integrated into the national economy.[61] He also sought a new mission for national laboratories that did not depend on Cold War rhetoric and that might move them into the growth areas of science, including biology. Domenici knew that sooner or later the Reagan defense spending juggernaut would lose steam.

Donald Fredrickson, then president of the Howard Hughes Medical Institute, asked if the national laboratories might play a role in the human genome project. After the meeting, Jack McConnell helped draft legislation that resulted in Senate bill 1480. By that time, Los Alamos was already beginning its genome program, a year and a half after DeLisi's initial idea. This show of strong support from the Senate nonetheless helped secure the DOE program's future at a time of potential vulnerability.

DeLisi and Smith anticipated many of the arguments that would be made for and against the genome project. But what was missing from their thoughts proved just as important—competition with NIH and acceptance among molecular biologists and human geneticists proved even more important than they might have thought. DeLisi remarked later that "moving unilaterally was

not my preference, nor did I consider it optimal."[62] He had a strong potential ally in Vincent DeVita, director of the National Cancer Institute, where DeLisi had worked before. DeVita's power was waning, however, and he was soon to leave the NCI directorship. NIGMS was the NIH institute responsible for funding most basic genetics, but DeLisi's relations with NIGMS were more distant and there was a much greater difference in styles.

DeLisi saw a hole, put his head down, and ran. He put the genome project on the public agenda, but it was not a clean run for the end zone.

> The well-known NIGMS response was that if it were to be done, they should do it, but it should not be done. . . . One of my choices was to use the NIH style of cautious consensus building. At times, perhaps most of the time, that is the best procedure; but in my judgment, this was not such a time. I made a deliberate decision to move vigorously forward with the best scientific advice we could muster (HERAC). I am quite willing to take the criticism, rational or not, that such movement provokes. . . . I would have been far more timid about subjecting myself to . . . criticisms . . . if I saw my future career path confined to government.[62]

DeLisi decided to risk attack and push forward. His relations with NIGMS director Ruth Kirschstein, director of the most relevant scientific program at NIH, were intermittent and distant. Those of us observing the process could readily see that the two principal figures in genome politics at DOE and NIH were ill at ease with each other. DeLisi and Kirschstein were both, however, consummate professionals. They avoided direct conflict while encouraging staff exchanges and cooperation. Both later glossed over this period during which their objectives were at cross purposes and their roles inherently cast them in opposition, attributing the perception of conflict to science reporters covering genome politics. The reporters were telling the truth. The tension between DOE and NIGMS was real. The amazing feature of the genome project is that the conflict was contained. It never broke into destructive distrust or resulted in NIH and DOE taking positions that would force them into direct confrontation before Congress. Staff members on Capitol Hill were well aware of the potential for open conflict between NIH and DOE. Some even eagerly awaited the public theater it would provide. Had the battle lines been drawn, the genome project as a whole would almost certainly have been delayed or destroyed.

Several technical elements are remarkable by their absence from early consideration at DOE. There was very little discussion of genetic linkage mapping—the first and arguably the most important step toward making the project useful to the research community—and scant attention to the study of nonhuman organisms as either pilot projects or even scientifically important subjects to study. DeLisi explained these gaps as resulting from a *presumption* that RFLP mapping and work in other organisms would proceed apace, and that the genome program would merely augment the ongoing efforts in these related but distinct areas.[63] A memo from George Cahill corroborates that

DeLisi stressed the importance of comparative genome mapping in man, mouse, and other organisms at the initial meeting of the HERAC subcommittee.[51]

Genetic linkage maps and work on other organisms were, however, clearly subsidiary to the main goals of the initial DOE program: DNA sequencing technology, computation, and physical mapping. By 1990, the genome project was redefined so that genetic linkage maps and physical maps of model organisms and humans were accorded first priority, with sequencing to follow when (and if) it became affordable and sufficiently rapid. In the reoriented genome project, DNA sequencing was subtly removed from the top spot and subordinated to other goals.

The seeming neglect of genetic linkage mapping and nonhuman genetics drove a wedge between DOE and much of the biomedical research community. The enthusiasm driving the DOE human genome proposal proved sufficient to keep it going, but it was a rough ride.