CHAPTER 8

Applying DNA Science to Human Genetics and Evolution

WE STAND AT THE THRESHOLD OF A NEW CENTURY with the whole human genome stretched out before us. It is, at once, a public and intensely private record. Written in each person's DNA is a shared history of the evolution of our species and a personal portent of both the health and disabilities we may encounter as individuals.

High-throughput analyses using bioinformatic tools and DNA arrays are uncovering an exponentially increasing number of genes implicated in human diseases. Each gene identified in a known disease pathway automatically becomes a validated target for therapeutic development. Pharmaceutical companies are racing to convert this new knowledge into blockbuster drugs while the medical profession ponders how it will keep up with the flood of new diagnostic and treatment options. The implications extend far beyond medicine, as scientists seek genes behind the behaviors—and misbehaviors—that make us uniquely human.

The border between science fact and fiction will blur as we move further into the genome age. A tenfold increase in the information capacity of photolithographic DNA arrays could condense the entire human genome onto a chip the size of a postage stamp. Perhaps not everyone will be able to afford to carry a complete copy of their genome in a medical alert locket, but rapid scans of hundreds of medically and behaviorally important genes will almost certainly become part of standard medical care in developed countries.

What will it be like when we have a precise catalog of all the good, bad, and middling genes—and the wherewithal to determine who has which? On a personal level, will a genome-wide scan take on the aspect of genetic tarot, predicting the future course of our lives? In the face of such knowledge, will society continue to acquiesce to those who prefer to let nature take its course or will we gravitate toward a prescribed definition of the "right" genetic stuff?

This is not the first time that we have asked such questions. At the turn of the last century, science and society faced a similar rush to understand and exploit human genes. Eugenics was the name of the effort to apply principles of Mendelian genetics to improve the human species.

The eugenics movement began benignly in England with positive efforts by families to improve their own heredity. It took a negative turn in the United States, as well as in Scandinavia, where flawed data became the basis for laws to sterilize individuals and restrict immigration by ethnic groups deemed "unfit." These misguided attempts at eugenic social engineering formed part of the basis

Harry Laughlin and Charles Davenport Outside the Eugenics Record Office, 1912

(Courtesy of the Harry H. Laughlin Archives, Truman State University; http://www.eugenicsarchive.org.)

Field Worker Training Class, 1922

Students on a field trip to Kings Park mental hospital; Harry Laughlin on far right. (Courtesy of the Cold Spring Harbor Laboratory Archives; http://www.eugenicsarchive.org.)

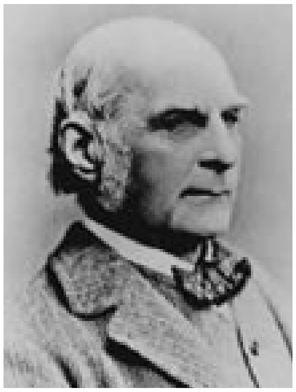
Eugenic and Health Exhibit, 1922

Fitter Families exhibit and examination building at the Kansas State Free Fair. (Courtesy of American Philosophical Society; http://www.eugenicsarchive.org.)

of the Nazi “final solution” to achieve racial purity, which resulted in the murder of more than 10 million Jews, gypsies, and other groups considered unfit.

We will thus begin the story of human genetics with the cautionary tale of its conjoined birth with eugenics, focusing on its development in the United States. Partly due to the stigma of association with the Holocaust and partly due the difficulty of conducting rigorous experimental studies, human genetics languished as a sort of scientific backwater after World War II, until molecular genetics provided new methods to track gene inheritance. After discussing the applications of genetics to the practical problems of human disease and identity, we will link these to the emerging story of how humans evolved and populated the earth.

CHARLES DAVENPORT AND THE EUGENICS RECORD OFFICE



Francis Galton
(Copyright The Galton Collection, University College London; http://www.eugenicsarchive.org.)

The term eugenics—meaning well born—was coined in 1883 by Francis Galton, a scientist at University College in London. Galton’s conception of eugenics arose from his earlier study, *Hereditary Genius* (1869), in which he concluded that superior intelligence and abilities were inherited with an efficiency of about 20% among primary relatives in noteworthy British families. Galton’s emphasis on the voluntary improvement of a family’s genetic endowment became known as “positive eugenics” and remained the focus of the British movement.

In Chapter 1, we saw how evolutionary thinking flowed into genetics by way of the ephemeral discipline of experimental evolution. In a similar way, evolutionary theory flowed into human genetics by way of eugenics. During the first decade of the 20th century, eugenics was organized as a scientific field by the confluence of ideas from evolutionary biology, Mendelian genetics, and experimental breeding. This synthesis was embodied by Charles Davenport.

Recall that, in 1898, Davenport became director of the Biological Laboratory at Cold Spring Harbor, New York, a field station to study evolution in the natural world. In 1904, on an adjacent property, he founded the Station for Experimental Evolution, whose researchers were among the very first adherents of Mendelian genetics. Davenport was among the first scientists to contribute to the genetic description of *Homo sapiens*. In 1907, he published the classical (although still incomplete) description of the inheritance of human eye color. He continued on to do early studies on the genetics of skin pigmentation, epilepsy, Huntington’s disease, and neurofibromatosis. Other researchers, including Archibald Garrod, described Mendelian inheritance in alkaptonuria, brachydactyly, hemophilia, and color blindness.

Davenport became interested in eugenics through his association with the American Breeders Association (ABA), the first scientific body in the United States to actively support eugenics research. The ABA members—including Luther Burbank, a pioneer of the American seed business—were literal in their aim to directly apply principles of agricultural breeding to human beings. This is aptly illustrated by Davenport’s book, *Eugenics: The Science of Human Improvement by Better Breeding*, as well as by the “Fitter Family” contests held at state fairs throughout the United States during the 1920s. These competitions judged families in the same context as the fastest racehorses, the fattest pigs, and the largest pumpkins.

In 1910, Davenport obtained funding to establish a Eugenics Record Office (ERO) on property adjacent to the Station for Experimental Evolution. A series of ERO bulletins, including Davenport's *Trait Book* and *How to Make a Eugenical Family Study*, helped to standardize methods and nomenclature for constructing pedigrees to track traits through successive generations. Constructing a pedigree entailed three important elements: (1) finding extended families that express the trait under study, (2) "scoring" each family member for the presence or absence of the trait, and (3) then attempting to discern one of three basic modes of Mendelian inheritance: dominant, recessive, or sex-limited (X-linked).

Eugenicists fared well on the first element, because large families were common in the first decades of the 20th century. However, scoring traits was a difficult problem, especially when eugenicists attempted to measure complex traits (such as intelligence or musical ability) and mental illnesses (such as schizophrenia or manic depression). In general, eugenicists were lax in defining the criteria for measuring many of the "traits" they studied. This led them to conclude that many real and imagined traits—including alcoholism, feeblemindedness, pauperism, social dependency, shiftlessness, nomadism, and lack of moral control—were single-gene defects inherited in a simple Mendelian fashion.

Much eugenical information was submitted voluntarily on questionnaires. Some families were proud to make known their pedigrees of intellectual or artistic achievement, whereas others sought advice on the eugenical fitness of proposed marriages. The circus performers on midways of nearby Coney Island offered eugenics researchers a trove of unusual physical trait differences, including giantism, dwarfism, polydactyly, and hypertrichosis. Notably, Davenport's correspondence with an albino circus family resulted in the first Mendelian study of albinism, published in the *Journal of Human Heredity*.

In addition to interviewing living family members, eugenics workers also used data from insane asylums, prisons, orphanages, and homes for the blind. Surveys filled out by superintendents were used to calculate the ethnic makeup of societal "dependents" and the costs of maintaining them in public institutions. With the mobilization for World War I, tens of thousands of men inducted for the draft provided a ready source of anthropometric and intelligence data. Notably, the Army Alpha and Beta Intelligence Tests, developed by Robert Yerkes of Harvard University, supposedly measured the innate intelligence of army recruits. African-American and foreign-born recruits were much more likely to do poorly on the Yerkes tests, because they mostly measured knowledge of white American culture and language.

THE CONSTRUCTION OF GENETIC BLAME

Whereas Francis Galton had focused on the positive aspects of human inheritance, the American movement increasingly focused on a "negative eugenics" program to prevent the contamination of the American germ plasm with supposedly unfit traits. The concept that some groups of people are genetically unfit dates back to Biblical references to the Amalekites. By about 1700, degeneracy theory supplied the "scientific" explanation that unfit people arose from bad environments which damaged heredity and perpetuated degenerate offspring.

Richard Dugdale, of the Executive Committee of the New York Prison Association, brought the concept of degenerate inheritance to eugenics in *The Jukes* (1877), a pedigree study of a clan of 700 petty criminals, prostitutes, and paupers living in the Hudson River Valley north of New York City.

Dugdale held the Lamarckian view that the environment induces heritable changes in human traits. He compassionately concluded the Jukes' situation could be corrected by providing them improved living conditions, schools, and job opportunities. However, this interpretation was discredited by American eugenicists, who embraced Mendel's genetics and Weismann's theory of the germ plasm. Together, these formed an interpretation that human traits are determined by genes which are passed from generation to generation without any interaction with the environment. Thus, when the ERO's field worker Arthur Estabrook reevaluated the Jukes in 1915, he found continued degeneration and placed the blame squarely on bad genes and the people who carried them.

Davenport's study of naval officers amusingly illustrates the extent to which eugenicists sought genetic explanations of human behavior to the exclusion of environmental influences. After analyzing the pedigrees of notable seamen—including Admiral Lord Nelson, John Paul Jones, and David Farragut—Davenport concluded that they shared several heritable traits. Among these was thalassophilia, "love of the sea," which he determined was a "sex limited" trait, because it was found only in men. Davenport failed to consider the equally likely explanations that sons of naval officers often grew up in environments dominated by boats and tales of the sea or that women were prohibited from seafaring occupations throughout the 19th and early 20th centuries.

Eugenic Social Engineering

Eugenics arose in the wake of the Industrial Revolution, when the fruits of science were improving public and private life. A growing middle class of professional managers believed that scientific progress offered the possibility of rational cures for social problems. Placing the blame for social ills on bad genes—and the people who carried them—raised the question: Why bother to build more insane asylums, poorhouses, and prisons when the problems that necessitated them could be eliminated at their source? Thus, negative eugenics seemed to offer a rational solution to age-old social problems. American eugenicists were largely successful in lobbying for social legislation on three fronts: to restrict European immigration, to prevent race mixing (miscegenation), and to sterilize the "genetically unfit."

As the eugenics movement was gathering strength, a phenomenal tide of immigration was rolling into the United States. During the first two decades of the 20th century, 600,000 to 1,250,000 immigrants per year entered the country through the facility on Ellis Island in New York Harbor (except during World War I). It is estimated that 100 million Americans alive today can trace their ancestry to an immigrant who arrived at Ellis Island. Also during this period, the nativity of the majority of immigrants shifted away from the northern and western European countries that had contributed most immigrants during the Colonial, Federal, and Victorian eras. Increasingly, the immigrant stream was dominated by southern and eastern Europeans, including large numbers of displaced Jews.

Doctor Examining Immigrants at Ellis Island, 1904
(Courtesy of National Park Service: Statue of Liberty Monument; http://www.eugenicsarchive.org.)

Immigrants Bound for New York on an Atlantic Liner

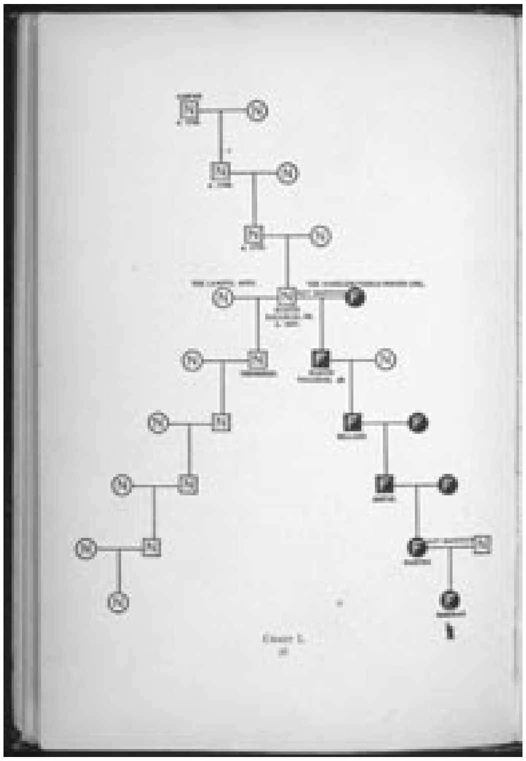
American eugenicists were overwhelmingly white, of northern and western European extraction, and members of the educated middle and upper classes. They looked with disdain on the new immigrants, many of whom settled in lower Manhattan. The plight of many of these immigrants—packed into tenements, plagued by tuberculosis and crime, and reduced to virtual serfdom in sweat shops—was sympathetically chronicled by Jacob Riis and the "muckraking" journalists. But to eugenicists, the immigrants' lot had little to do with poverty or lack of opportunity and had everything to do with their bad genes, which eugenicists feared would quickly "pollute" the national germ plasm. The eugenics movement provided a scientific rationale for growing anti-immigration sentiments in American society. Labor organizations fed on fears that working class Americans would be displaced from their jobs by an oversupply of cheap immigrant labor, while anti-Communist factions stirred up fears of the "red tide" entering the United States from Russia and eastern Europe.

As "expert agent" for the Committee on Immigration and Naturalization of the U.S. House of Representatives, ERO Superintendent Harry Laughlin became the anti-immigration movement's most persuasive lobbyist in the early 1920s. During three separate testimonies, he presented data that purported to show that southern and eastern European countries were "exporting" genetic defectives to the United States who had disproportionately high rates of mental illness, crime, and social dependency. The resulting Immigration Restriction Act of 1924 cut immigration to 165,000 per year and restricted immigrants from each country according to their proportion in the U.S. population in 1890—a time prior to the major waves of immigration from southern and eastern Europe. This had the desired effect of reducing southern and eastern European immigrants to less than 15,000 per year. Immigration did not regain prerestriction levels again until the late 1980s.

Of all the legislation enacted during the first four decades of the 20th century, sterilization laws adopted by 30 states most clearly bear the stamp of the eugenics lobby. Although the earliest sterilization law, passed in Indiana in 1907, was aimed at convicts and sex offenders, "feebleminded" persons became the major targets for eugenic sterilization. This owed much to Henry Goddard's influential book, *The Kallikaks* (1912). This effectively related study of the descendents of Martin Kallikak (the name is fictitious) was, in effect, a controlled experiment in positive and negative eugenics. Martin's marriage to a normal woman produced a normal lineage (from the Latin *kallos* for "goodness and beauty"). However, as a young militiaman in the Revolutionary War, he had an elicit union with an attractive but feebleminded barmaid, producing a second, "bad" lineage (*kakos*, for "bad"). Thus, the primary intent of eugenic sterilization was to curb the supposed promiscuous tendencies of the feebleminded, who threatened to perpetuate their kind and to contaminate good lineages, as surely as the case of Martin Kallikak.

Many of the early sterilization laws were legally flawed and did not meet the challenge of state court tests. To address this problem, Laughlin designed a model eugenics law that was reviewed by legal experts. Virginia's use of the model law was tested in *Buck v. Bell*, heard before the Supreme Court in 1927. Oliver Wendell Holmes, Jr., delivered the Court's decision upholding the legality of eugenic sterilization, which included the infamous phrase, "Three generations of imbeciles are enough!"

Carrie Buck, the subject of the case, had given birth to an illegitimate daughter and been institutionalized in the Virginia Colony for the Epileptic and the Feebleminded. Carrie was judged to be "feebleminded" and promiscuous. Arthur Estabrook examined Carrie's infant daughter Vivian and found her "not

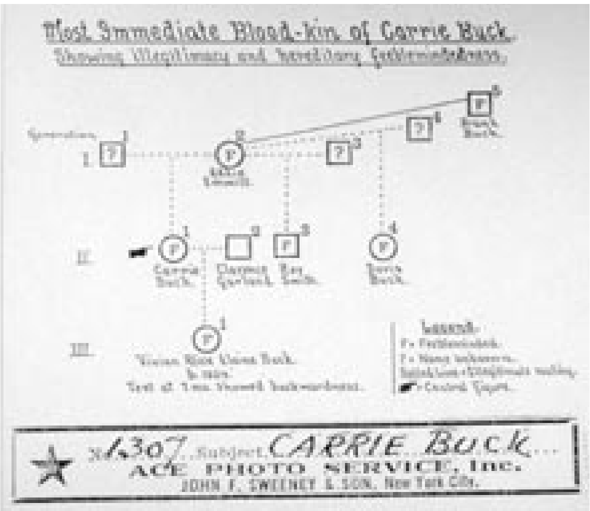


Henry Goddard's Pedigree of the Kallikak Family, 1912

The pedigree shows normal (N) and feebleminded (F) lines. (Courtesy of the University of Albany, SUNY; http://www.eugenicsarchive.org.)

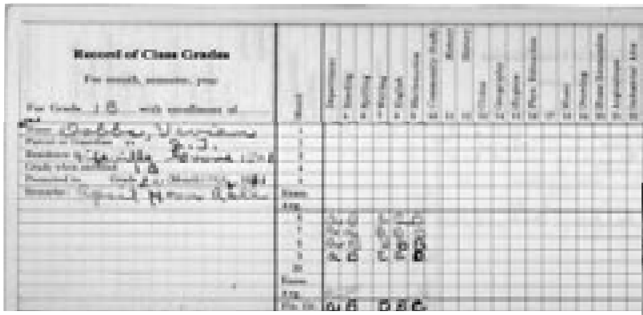
Carrie and Emma Buck, 1924

This evocative photo of daughter and mother on a bench at the Viriginia Colony for the Epileptic and the Feebleminded, in Lynchburg, was taken the day before the start of the Virginia trial that would lead all the way to the U.S. Supreme Court. (Courtesy of the University of Albany, State University of New York; http://www.eugenicsarchive.org.)


Most Immediate Blood-kin of Carrie Buck
Showing Illegitimacy and hereditary Feeblemindedness.
Generation
I
II
III
?
F
Addie Emmitt
Frank Buck
Carrie Buck
Clarence Garland
Roy Smith
Doris Buck
Vivian Alice Elaine Buck
b. 1924
Test at 7 mo. showed backwardness
Legend
F = Feebleminded
? = Nature unknown
Dotted Line = Illegitimate mating
= Central figure
No. 1307 Subject CARRIE BUCK
ACE PHOTO SERVICE, Inc.
JOHN F. SWEENEY & SON, New York City.

Pedigree of the Buck Family, 1924

This exhibit from the Virginia trial clearly illustrates three supposed generations of feebleminded females in the Buck family: Emma (Addie Emmitt), Carrie, and Vivian. Vivian's father, Clarence Garland, was the nephew of Carrie's foster parents, Mr. and Mrs. John Dobbs. Clarence had promised to marry Carrie, but disappeared by the time of her trial in 1924. (Courtesy of Paul Lombardo, Ph.D., J.D., and the American Philosophical Society; http://www.eugenicsarchive.org.)



Record of Class Grades

For month, semester, year

For Grade 1 B with enrollment of

Dobbs, Vivian

Vivian's First-grade Report, 1931

Listed under her adoptive surname, Dobbs, Vivian was a solid "B" student during her first-grade year at the Venable School in Charlottesville. She got straight "As" in deportment (conduct) and even made the honor role in April, 1931. She died a year later after a bout with measles. (Courtesy of Paul Lombardo, Ph.D., J.D.; http://www.eugenicsarchive.org.)

Carrie's Baby, Vivian, 1924.

This photo, taken the day before the Virginia trial, is believed to capture the "standard mental test" used by Arthur Estabrook to determine that Vivian Buck was feebleminded. Six-month old Vivian appears uninterested as foster mother Mrs. Dobbs attempts to catch her attention with a coin. (Courtesy of the University of Albany, State University of New York; http://www.eugenicsarchive.org.)

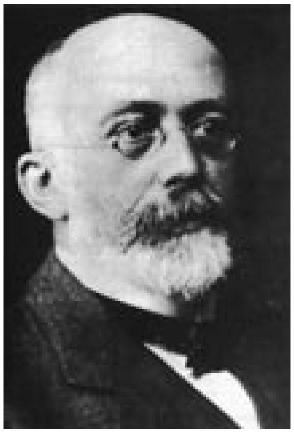
quite normal." It is impossible to judge whether Carrie was "feebleminded" by the standards of her time, but the child that Carrie bore out of wedlock was the result of her rape by the nephew of her foster parents. Clearly, Vivian was no imbecile. Later scholarship turned up Vivian's first-grade report, showing that she was a solid "B" student and received an "A" in deportment. Carrie was the first person sterilized under Virginia's law. *Buck v. Bell* was never overturned, and sterilization of the mentally ill continued into the 1970s, by which time about 60,000 Americans had been sterilized—most without their consent or the consent of a legal guardian.

Opposition and the End of Eugenics

Scientific opposition to eugenics came on many fronts and began even as it was being organized as a scientific discipline. In 1909, George Shull, at the Carnegie Station for Experimental Evolution, showed that the hybrid offspring of two inbred strains of corn are more vigorous than their inbred parents. The phenomenon of hybrid vigor also held true in mongrel animals, refuting eugenicists' notion that racial purity offers any biological advantage or that race mixing destroys "good" racial types.

Work by a number of scientists countered the eugenicists' simplistic assertions that complex behavioral traits are governed by single genes. Hermann Muller's survey of mutations in *Drosophila* and other organisms from 1914 to 1923 showed variation in the "gene-to-character" relation that defied simple Mendelian analysis. Many genes are highly variable in their expression, and a single gene may affect several characteristics (traits) at one time. Conversely, mutations in many different genes can affect the same trait in similar ways. Moreover, the expression of a gene can be altered significantly by the environment. Twin studies conducted by Horatio Hackett Newman also showed that identical twins raised apart after birth averaged a 15-point difference in I.Q. Lionel Penrose found that most cases at a state-run institution in Colchester, England, resulted from a combination of genetic, environmental, and pathological causes.

Mathematical models of population genetics provided evidence against the simplistic claim that degenerate families were increasing the societal load of dysgenic genes. The equilibrium model of Godfrey Hardy and Wilhelm Weinberg showed that, although the absolute number of dysgenic family members might increase over time, the frequency of any "negative" trait does not increase relative to the normal population. Feeblemindedness, thought to be a recessive dis-



Wilhelm Weinberg
(Reprinted, with permission, from Stern C. 1962. Wilhelm Weinberg, 1862–1937. *Genetics 47:* 1–5; (©Genetics Society of America.)

Godfrey Hardy
(Courtesy of Trinity College, Cambridge, England.)

order, presented a particular quandary. Although geneticists almost universally agreed that the feebleminded should be prevented from breeding, the Hardy-Weinberg equation showed that sterilization of affected individuals would never appreciably reduce the incidence of the disorder. Only a hideously massive program of sterilizing the vast reservoir of heterozygous carriers predicted by the equation would have any hope of significantly reducing the incidence of mental illness. Despite this, feeblemindedness was thought to be so rampant that many geneticists believed reproductive control could still prevent the birth of tens of thousands of affected individuals per generation.

Although he was a founding member of the board of the ERO, Thomas Hunt Morgan resigned after several years. He criticized the movement in the 1925 edition of his popular textbook, *Genetics and Evolution*, warning against the wholesale application of genetics to mental traits, and against comparing whole races as superior or inferior. He offered this advice: "...until we know how much the environment is responsible for, I am inclined to think that the student of human heredity will do well to recommend more enlightenment on the social causes of deficiencies...in the present deplorable state of our ignorance as to the causes of mental differences."

In 1928, Johns Hopkins geneticist Raymond Pearl charged that most eugenics preaching was "contrary to the best established facts of genetical science." A visiting committee of the Carnegie Institution in 1935 concluded that the body of work collected at the ERO was without scientific merit and recommended that it end its sponsorship of programs in sterilization, race betterment, and immigration restriction. Thus, the negative emphasis of American eugenics was completely discredited among scientists by the mid 1930s. Growing public knowledge of Germany's radical program of race hygiene led to a wholesale abandonment of popular eugenics. The ERO was closed in December 1939.

In the meantime, eugenics was gathering steam in Germany. Laughlin's model sterilization law was the basis for Nazis' own law in 1933, and his contributions to German eugenics were recognized by an honorary degree from the University of Heidelberg in 1936. Over the next several years, some 400,000 people—mainly in mental institutions—were sterilized. In 1939, euthanasia replaced sterilization as a solution for mental illness, and the lives of nearly 100,000 patients were ended "mercifully" with lethal gas. Overt euthanasia of mental patients ceased in 1941, when physicians with experience in euthanasia were reassigned to concentration camps in Poland, where they were needed to apply the "final solution" for Nazi racial purity.

PROBLEMS ON THE ROAD TO MODERN HUMAN GENETICS

Following the shocking revelations of euthanasia and human experimentation that took place in the Nazi concentration camps, it is not difficult to understand why human genetics research was largely avoided during the years following World War II. The fact also remained that prior to the advent of restriction enzymes and recombinant DNA, researchers simply did not have the tools to identify genes or to precisely locate them on chromosome maps.

Genetics had succeeded in *Drosophila*, for example, because a fly with a new mutation can be identified by visual inspection. The mutant fly can then be selectively mated with other flies of known characteristics to determine the mode of inheritance and chromosome position of the mutated gene. These analyses are simplified by the fly's rapid generation time and many offspring per generation. Furthermore, the lineage of each individual is known at the outset of a breeding experiment. Members of the experimental pool are most often physically and genetically identical, differing from one another only by one or, at most, several traits. This genetic homogeneity allows a specific trait or mutation to be observed against an essentially neutral background.

Human genetics differs from classical genetics in that the system under study cannot be easily manipulated. Although arranged marriages still take place in some cultures, people, for the most part, choose their own spouses and are generally opposed to being selectively mated. Thus, human geneticists must be content to work with the existing genetic makeup of related family members. In addition, they seldom have the luxury of following a single well-defined trait through successive generations; rather, they often must deal with a perplexing syndrome of variable traits. Most human populations tend to be outbred, meaning that they are physically and genetically heterogeneous. Thus, it is more difficult to identify genes—especially those with variable phenotypes—against this heterogeneous background.

Certain human populations, however, whose members are genetically isolated by geography or customs have a degree of genetic homogeneity. The relatively closed gene pools of the Icelandic people, the Old Order Amish, and the Mormons, combined with their habit of keeping meticulous genealogical records and having relatively large families, have made them amenable to genetic analysis. Customs prohibiting alcohol consumption among the Amish and Mormons make easier the analysis of mental and behavioral disorders, such as manic depression and schizophrenia, whose symptoms may be masked by alcohol or drug abuse.

The problems of human genetics were only solved as time eroded memories of Nazi eugenics, and when restriction enzymes and polymerase chain reaction (PCR) provided markers whose presence or absence can be scored with great certainty. It is worth remembering that during the entire reign of eugenics, DNA had not yet been shown to be the molecule of heredity, and nothing was known about the physical basis of mutation and gene variation.

With an understanding of gene variation has come a deeper understanding of disease complexity. Twin studies strongly indicate that genes have a dominant role in all aspects of human health and behavior. However, just as Hermann Muller observed in fruit flies, human diseases do not always exhibit a simple gene-to-character relation. An identical mutation may produce different physical symptoms (phenotypes) in different people. Conversely, different mutations may produce similar phenotypes in different people. As George W. Beadle and Edward Tatum found in *Neurospora*, mutations in any of several enzymes can have the same end effect—of altering or knocking out a biochemical pathway.

At one end of the spectrum are "simple," genetically homogeneous diseases, such as sickle cell anemia and cystic fibrosis, for which affected individuals share common mutations and highly similar symptoms. In the middle are diseases, such as β-thalassemia and neurofibromatosis 1, in which a variety of types of mutations in a single gene produce variable symptoms. At the other end of spectrum are "complex," genetically heterogeneous diseases, such as asthma and bipolar disorder, in which mutations in a number of genes—in combination with environmental factors—likely account for extremely variable symptoms.

DETERMINING THE CHROMOSOMAL BASIS OF HUMAN DISEASE

The development of new cytological methods, beginning in the 1950s, helped to bring human genetics out of its "dark ages." T.C. Hsu's treatment with hypotonic (low-salt) solution caused cells to swell, separating the chromosomes and making them easier to count. Wright stain, and later Giemsa and quinacrine stains, made it possible to identify each chromosome by size and distinctive staining patterns.

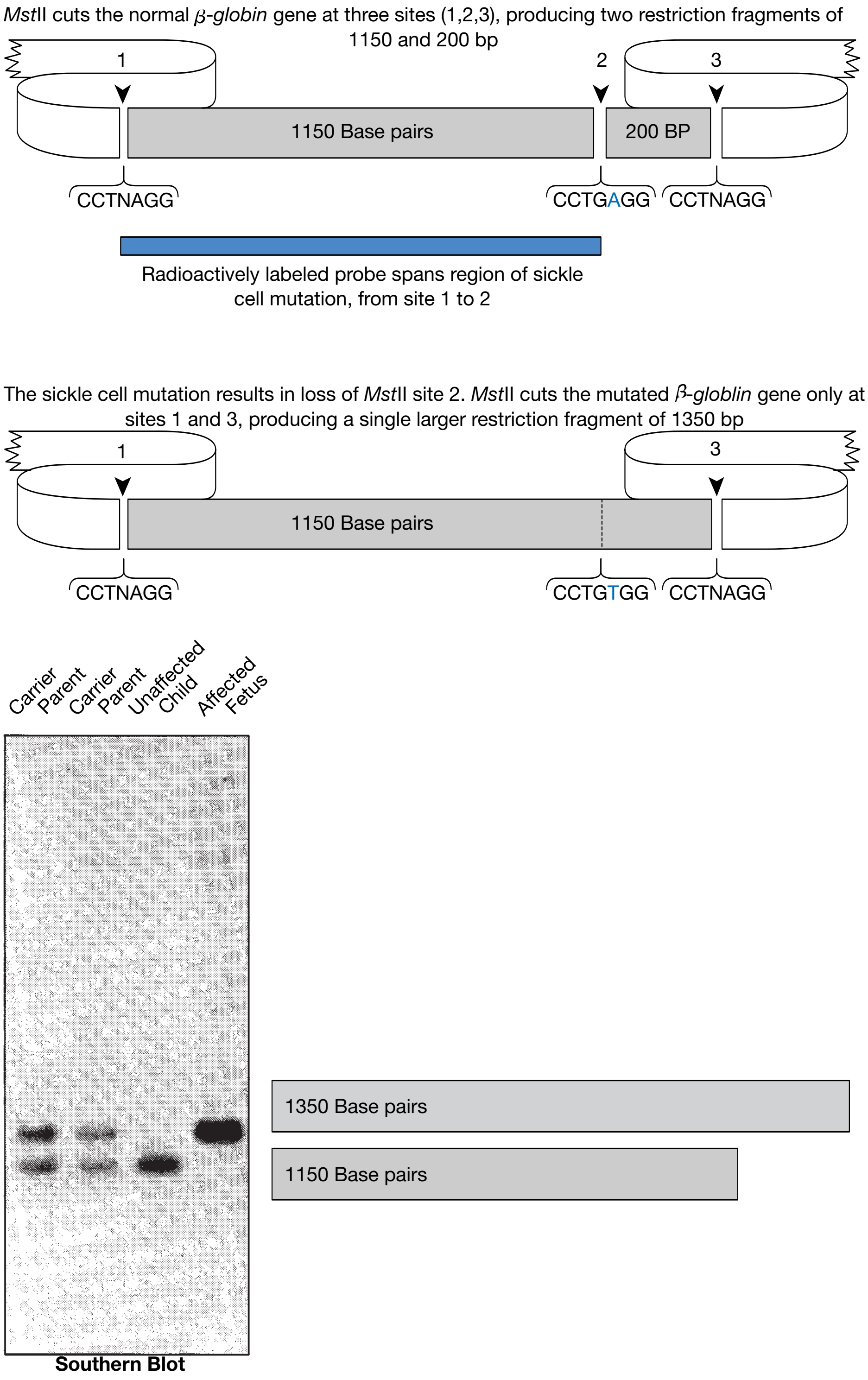
The state of human cytogenetics in the post-World War II years is best summed up by the fact that until 1956, humans were thought to have 48 chromosomes. This number was prejudiced by an earlier, and accurate, determination of 48 chromosomes in chimps. J.H. Tjio and A. Leven, of the National Institutes of Health, cleared up the matter when they showed conclusively that humans have 46 chromosomes. Within several years, cytologists established a direct relationship between human genetic disorders and abnormal chromosome number, or aneuploidy. Trisomy—an extra chromosome copy—was found in Down's syndrome (chromosome 21), Patau's syndrome (13), and Edward's syndrome (18). Abnormalities in sex chromosome number were also described for Turner's syndrome (X) and Klinefelter's syndrome (XXY). In the early 1960s, translocations were identified in some cases of Down's syndrome. Studies in the 1970s showed that chronic myeloid leukemia, Burkitt's lymphoma, and several other blood cancers are characterized by specific chromosome translocations (see Chapter 7).

Sickle Cell Brings Human Genetics into the Molecular Era

Recall Archibald Garrod's prophetic hypothesis, in 1908, that alkaptonuria and, by extension, other inherited disorders are caused by "inborn errors in metabolism." Beadle and Tatum proved that this is exactly the case in *Neurospora*. They showed that mutations in specific genes produce corresponding changes to enzymes, evidenced by heritable metabolic deficiencies. The elucidation of the molecular mechanism of sickle cell disease provided the first proof of this concept in humans and illustrates the accumulation of genetic knowledge over time.

Sickle cell disease was first described in 1910 by Chicago physician James Herrick, whose patient had anemia characterized by unusual sickle-shaped red cells. Over the years, evidence accumulated that it is a recessive disorder. In the mid 1940s, Irving Sherman, a medical student at Johns Hopkins School of Medicine, found that sickled blood transmits light differently than normal blood, suggesting structural differences in the hemoglobin molecule. William Castle, of Harvard Medical School, relayed this information to Linus Pauling at the California Institute of Technology, who had become interested in the molecular structure of hemoglobin. Castle supplied blood samples from sickle cell patients and healthy controls, from which Pauling and Harvey Itano isolated hemoglobin. When separated by electrophoresis, sickle-cell hemoglobin (Hb_s) migrates more slowly, showing it is less negatively charged than normal hemoglobin (Hb).

This was consistent with the work by Vernon Ingram, of the Cavendish Laboratory in Cambridge, England, showing that Hb contains more glutamic acid (a negatively charged amino acid) and Hb_s contains more valine (a neutral


*Mst*II cuts the normal *β-globin* gene at three sites (1,2,3), producing two restriction fragments of 1150 and 200 bp
1
2
3
1150 Base pairs
200 BP
CCTNAGG
CCTGAGG
CCTNAGG
Radioactively labeled probe spans region of sickle cell mutation, from site 1 to 2
The sickle cell mutation results in loss of *Mst*II site 2. *Mst*II cuts the mutated *β-globlin* gene only at sites 1 and 3, producing a single larger restriction fragment of 1350 bp
1
3
1150 Base pairs
CCTNAGG
CCTGTGG
CCTNAGG
Carrier Parent
Carrier Parent
Unaffected Child
Affected Fetus
1350 Base pairs
1150 Base pairs
Southern Blot

RFLP Diagnosis of Sickle Cell Anemia, 1982

The Southern blot shows the RFLP patterns of two carrier parents, an unaffected offspring, and amniotic fluid from an affected fetus. The carrier parents show a single copy of each RFLP: the 1350-bp fragment associated with the disease allele and the 1150-bp fragment associated with the normal allele (the 200-bp fragment is not detected by the probe). The unaffected child shows the 1150-bp fragment, and the affected fetus shows the 1350-bp fragment. Both offspring are homozygous and thus show a single relatively thick band, denoting two chromosomal copies of the normal (Hb) or mutated (Hb_s) gene. (Reprinted, with permission, from Chang J.C. and Kan Y.W. 1982. A sensitive new prenatal test for sickle-cell anemia. *N. Engl. J. Med. 307:* 30–32.)

amino acid). In 1956, Vernon Ingram and John Hunt independently sequenced the Hb and Hb_s proteins, finding that a glutamic acid at position 6 in Hb is replaced with valine in Hb_s. From this information, they used a genetic code table (showing that glutamic acid = GAG and valine = GTG) to predict that the A-T point mutation in the sixth codon is responsible for sickle cell disease.

The availability of protein and predicted DNA sequence facilitated the cloning of the α- and β-*globin* genes from a human genomic library in the early 1980s. (The methods used by Philip Leder to clone the β-*globin* gene are discussed in Chapter 5.) The combination of Southern blot and restriction enzyme analysis made DNA diagnoses possible for the causative lesions of many hemoglobinopathies.

In constructing early restriction maps of cloned human DNA, it became obvious that a point mutation can change a restriction enzyme recognition site, producing different-sized fragments, termed a restriction-fragment-length polymorphism (RFLP). These were the first DNA polymorphisms (poly for "many" and morph for "form") that could be readily detected. As discussed in Chapter 6, RFLPs also were the major type of marker employed in the early physical and linkage maps of the human chromosomes. Used in a local region of a chromosome, an RFLP also might detect the causative mutations of disease. Initially, RFLPs were detected by Southern blot analysis, using a radioactive probe that hybridizes to the polymorphic region.

The mutation responsible for sickle cell anemia was first detected by RFLP analysis in 1978 by Yuet Wai Kan and Andrea-Marie Dozy at the University of California, San Francisco. They used the restriction enzyme *Mst*II, which recognizes the sequence CCTNAGG (where N equals any nucleotide). The A-T mutation results in the loss of an *Mst*II recognition site that spans the region of sixth codon of the β-*globin* gene. Thus, the DNA from normal homozygous individuals, heterozygous carriers of the sickle cell trait, and homozygous sickle cell patients produces different restriction fragments when cut with *Mst*II.

Making Therapeutics from Cloned Genes

The isolation of insulin in 1921 by Frederick Banting and Charles Best of the University of Toronto, and their demonstration that it successfully corrects the metabolic defect of diabetes, paved the way for the treatment of other common metabolic disorders, notably hemophilia and pituitary dwarfism. Such deficiencies can be corrected by supplying the missing or underproduced protein: clotting factors VIII and IX for hemophilia, and human growth hormone (HGH) for dwarfism. However, ensuring adequate, contagion-free supplies of these therapeutic proteins proved difficult. Hormones are often produced in minute quantities in the body and, hence, are laborious and very expensive to isolate. In the case of human growth hormone, the number of patients who could be treated was limited by availability.

Approximately 8000 pints of blood were processed to yield enough clotting factor to treat a single hemophiliac for 1 year, and 7–10 pounds of pancreas from approximately 70 pigs or 14 cows were needed to purify enough insulin for 1 year's treatment of a single diabetic. The extraction of HGH was most onerous, requiring the pituitary glands from approximately 80 human cadavers to produce enough for a single year's therapy. The magnitude of the supply problem becomes obvious when one considers that patients suffering from these diseases require long-term treatment lasting a *minimum* of 5–10 years.

The risk of virus contamination is a most important consideration in any therapeutic product purified from mammalian cells. Simian virus 40 (SV40), which has proved so important in cancer research, was first isolated as a contaminant in poliovirus vaccine produced in monkey cells. Although there is no evidence of illness as a result of SV40-contaminated poliovirus vaccines, supplies of both human growth hormone and clotting factors have at one time or another been infected with life-threatening pathogens.

Prior to the identification of human immunodeficiency virus type 1 (HIV-1) and the development of virtually foolproof screening procedures, patients with hemophilia had a significant risk of contracting AIDS (acquired immune deficiency syndrome) from transfusion of contaminated clotting factors, as well as whole blood. During the window of time between the onset of the AIDS pandemic and the development of effective methods to screen for HIV and disable it in blood products, in 1983–1984, is it estimated that half of all hemophiliacs developed AIDS. According to 2001 statistics from the Centers for Disease Control, a total of 5234 American hemophiliacs have died of AIDS.

Therapeutic proteins isolated from animals, including porcine or bovine insulin, differ in amino acid makeup from the human protein they replace. The biological activity of an animal substitute may differ slightly from the native human protein, or it may elicit an immune response. Some diabetics had allergic reactions to porcine or bovine insulin, although this may have been due to impurities in the preparations and not necessarily differences in the amino acid sequence.

The development of new genetic tools made possible the cloning and production of a number of genes for medically important proteins, including insulin, clotting factors, tissue plasminogen activator, interleukin, interferon, erythropoietin, and colony stimulating factors. The Boyer-Cohen experiment (Chapter 4) showed that recombinant DNA methods can be used to transfer essentially any gene into *E. coli*, where the encoded protein may be expressed. This established a new paradigm of using cultured cells to produce therapeutic proteins to treat human metabolic disorders.

Producing therapeutic proteins from cloned human genes inside *Escherichia coli* hosts eliminates the risk of virus contamination and allergic sensitivity. Mammalian viruses cannot reproduce inside *E. coli* and hence cannot be co-isolated with the protein from a bacterial culture. The protein harvested from the bacterial culture has been expressed from a human coding region and is identical (or very nearly so) to the native protein. Thus, diabetics sensitive to bovine or porcine insulin do not have an adverse reaction to human insulin of recombinant origin.

Expressing Insulin and Growth Hormone in *E. coli*

The production of human insulin and growth hormone illustrate that expression of a human protein in *E. coli* typically requires detailed understanding of its biological synthesis and of the biochemical limitations of the bacterial cell. *E. coli* is incapable of processing eukaryotic pre-mRNAs or of performing the several posttranslational modifications needed to produce a biologically active form of insulin from its protein precursors.

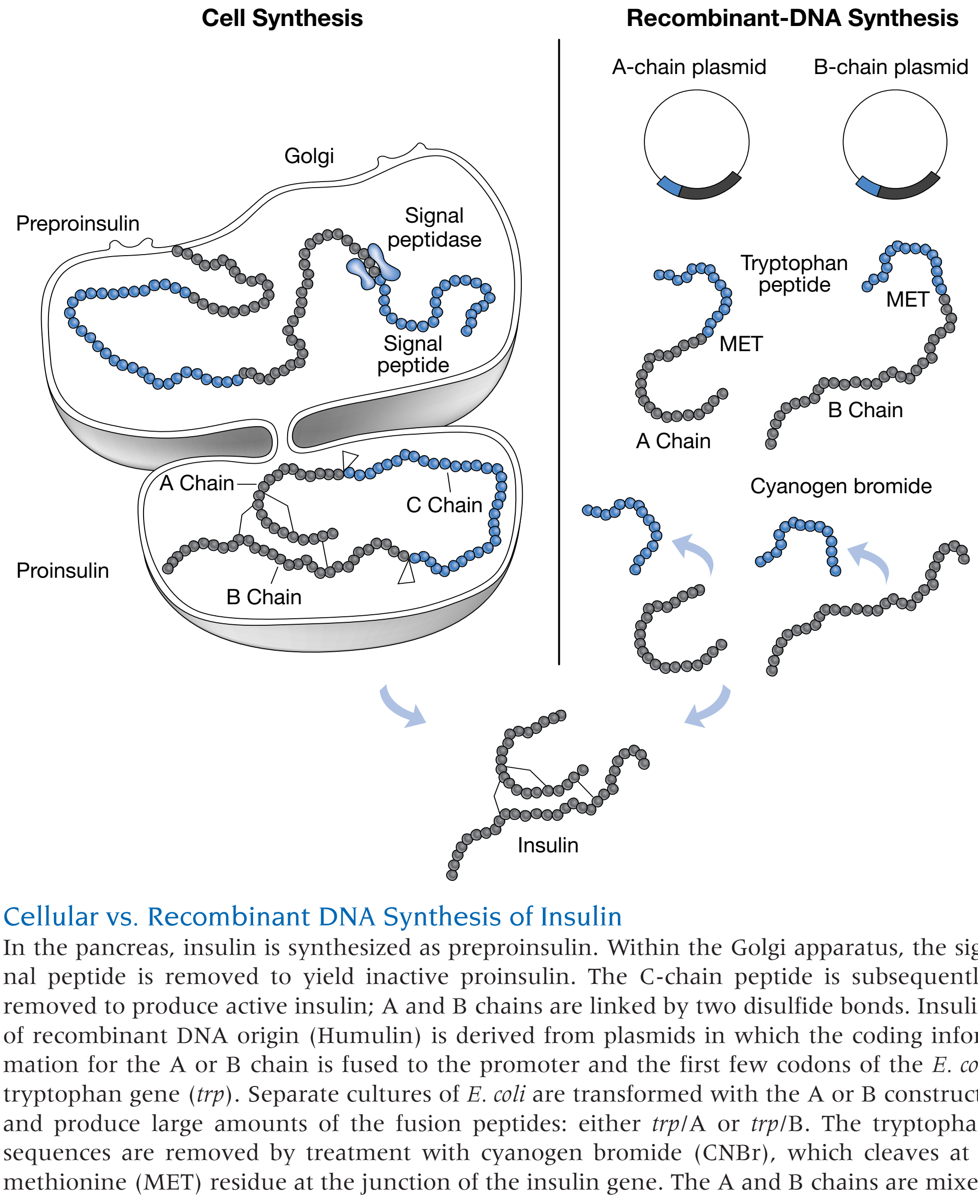
Some Approved Drugs Produced from Cloned Genes

Product	Generic name/Company	Year of first U.S. Approval	Approved for
Recombinant human insulin	Humulin/Eli Lilly & Co.	1982	Diabetes mellitus
Recombinant human growth hormone	Protropin/Genentech, Inc.	1985	Growth hormone deficiency in children
Recombinant interferon-α	Intron A/Scherin-Plough	1986 1988 1988 1991 1992	Hairy cell leukemia Genital warts Kaposi's sarcoma Hepatitis C Hepatitis B
Recombinant hepatitis B vaccine	Recombivax HB/Merck & Co.	1986	Hepatitis B prevention
Epoetin alfa	EPOGEN/Amgen Ltd.	1989	Anemia of chronic renal failure
Recombinant interferon-γ	Acctimune/Genentech, Inc.	1990	Chronic granulomatous disease
Colony-stimulating factor (CSF)	Leukine/Immunex Corp.	1991	Bone marrow transplantation
Recombinant anti-hemophiliac factor	Recombinate rAHF/Baxter Healthcare	1992	Hemophilia A
Recombinant DNase I	Pulmozyme/Genentech, Inc.	1993	Cystic fibrosis
Recombinant coagulation factor IX	AlphaNine SD/Alpha Therapeutic Corp.	1996	Christmas Disease Hemophilia B

The mature insulin molecule consists of two polypeptide chains—an A chain of 30 amino acids and a B chain of 21 amino acids—which are held together by disulfide linkages. However, this active insulin results from sequential modifications in two precursor molecules: preproinsulin and proinsulin. The gene for insulin consists of two coding exons separated by a single intron. Following splicing of the pre-mRNA, a functional transcript is translated into a large polypeptide called preproinsulin. The molecule includes a 24-amino-acid signal peptide at its amino terminus, a feature of many secreted proteins needed for their proper transport through the cytoplasm. The signal peptide, which is the first part of the preproinsulin molecule produced, anchors the free-floating ribosome to the endoplasmic reticulum (ER) and is subsequently clipped off as the molecule passes through the ER membrane.

The result is a molecule of 84 amino acids called proinsulin, whose looped shape is maintained by cross-linking disulfide bonds. Proinsulin makes its way to the Golgi apparatus, where a converting enzyme removes 33 amino acids from the middle of the connecting loop (the C chain), leaving the remaining A and B chains held together by the disulfide linkages. This yields active insulin, which is stored in a secretory granule for eventual release into the bloodstream.

The human genomic sequence could not be used directly to produce active insulin in *E. coli*, because the bacterium lacks the enzyme systems needed (1) to splice out the intron sequence to produce a mature mRNA, (2) to remove the signal sequence from preproinsulin, and (3) to remove the C chain from proinsulin. Although other methods have been used, the strategy first employed in 1979 by Eli Lilly & Co. to produce recombinant human insulin neatly sidesteps these constraints by simply omitting the above sequences. The nucleotide sequences coding for the A and B chains of active insulin were chemically synthesized and cloned into separate expression plasmids. A bacterial strain containing each expression vector produces a fusion bacterial/human polypeptide, which is harvested and subsequently treated with cyanogen bromide to remove the bacterial amino acids. Cyanogen bromide cleaves the fusion polypeptide at the methionine residue that begins the human sequence. This treatment, which also alters tryptophan, is only useful because, by happenstance, neither the A nor B insulin chain includes tryptophan or additional methionine residues.



Cellular vs. Recombinant DNA Synthesis of Insulin

In the pancreas, insulin is synthesized as preproinsulin. Within the Golgi apparatus, the signal peptide is removed to yield inactive proinsulin. The C-chain peptide is subsequently removed to produce active insulin; A and B chains are linked by two disulfide bonds. Insulin of recombinant DNA origin (Humulin) is derived from plasmids in which the coding information for the A or B chain is fused to the promoter and the first few codons of the *E. coli* tryptophan gene (*trp*). Separate cultures of *E. coli* are transformed with the A or B constructs and produce large amounts of the fusion peptides: either *trp*/A or *trp*/B. The tryptophan sequences are removed by treatment with cyanogen bromide (CNBr), which cleaves at a methionine (MET) residue at the junction of the insulin gene. The A and B chains are mixed together, and disulfide bonds are formed by a chemical process.

Purified A and B chains are then mixed in equal portions and incubated under conditions that form the disulfide linkages.

Human growth hormone, a polypeptide of 191 amino acids, is also produced in recombinant *E. coli*. The coding sequence for the first 24 amino acids of the expressed gene is synthesized chemically, whereas amino acids 25 through 191 are derived from a cDNA copy of *HGH* mRNA isolated from pituitary cells. The recombinant HGH differs by one amino acid from normal HGH due to the fact that *E. coli* is unable to remove the initiator methionine residue that is removed posttranslationally in human cells.

Expressing t-PA, Erythropoietin, and Interferons in Mammalian Systems

Both insulin and HGH are relatively simple proteins that do not undergo glycosylation or other posttranslational modifications. Bacteria do not possess the enzymatic machinery for making posttranslational modifications to proteins, so genes encoding extensively modified proteins must be cloned into eukaryotic expression systems. Chinese hamster ovary (CHO) cells have proven to be the most popular mammalian system for expressing human therapeutic proteins. However, expressing proteins in most eukaryotic cells is much more costly than in bacteria. Whereas a cloned gene can be engineered to produce up to 40% of total protein production of *E. coli*, a cloned product may account for only 8% of total protein output in eukaryotic cells.

Tissue plasminogen activator (t-PA) is an example of a therapeutic protein that must be expressed in mammalian cells. t-PA is a protease that attacks fibrin, a major protein involved in forming blood clots. Patients that demonstrate early signs of heart attack or stroke are administered t-PA, which acts by destroying small blood clots that can potentially form blockages in arteries. Until the overexpression of cloned t-PA was attained in eukaryotic cells, the major clot destroyer in use was streptokinase, a protease isolated from *Streptococcus*. As a foreign protein, streptokinase may cause immune reactions, hemorrhaging, and other side effects. For these reasons, t-PA was initially hailed as a major improvement in the treatment of heart attack. However, retrospective studies have shown little difference in the recovery of patients treated with t-PA versus those treated with streptokinase.

Erythropoietin (EPO), which stimulates production of red blood cells from stem cells in the bone marrow, is useful for treating anemia, AIDS, and patients undergoing chemotherapy and bone marrow transplants. Both EPO and HGH have less reputably been used as performance boosters among athletes. Whereas HGH increases strength by increasing muscle mass, EPO increases endurance by increasing the oxygen-carrying capacity of the blood.

Discovered in 1957, interferon gained a reputation as having wondrous antiviral, anticancer, and immune modulatory effects. However, quantities were too limited to prove its usefulness in laboratory and clinical trials. When cloned interferon made significant quantities available, it was found that interferon modulates the function of several types of immune cells—macrophages, cytotoxic lymphoctyes, and B cells—increasing expression of immunoglobulins and human leukocyte antigens (HLAs). In 1992, James Darnell, at The Rockefeller

University, elucidated how interferon initiates a signal transduction pathway through which immune cells are primed to recognize and degrade the RNA of infecting viruses. However, interferon proved to be relatively toxic at pharmacological doses and generally failed to live up to its wonder drug hype. Although interferon has not proven to be the "magic bullet" that some had hoped for, its several forms have proven broadly useful in treating a number of diseases. Interferon-α is the most effective treatment available for hepatitis B and C, which infect hundreds of millions people worldwide, and is also used in treating leukemias. Interferon-β is the most effective treatment for multiple sclerosis, although it is still uncertain how it functions in controlling this disease. Interferon-γ is used to treat osteopetrosis and chronic granulomatous disease.

THE IMPORTANCE OF DNA POLYMORPHISMS

The obvious candidates for medications from cloned genes were soon virtually exhausted. Biologists had to come to grips with the harder work of cloning a disease gene in the absence of knowledge about its protein product. To do this, a disease gene first must be mapped to a chromosome position by linkage to known loci. After a precise location is identified, the region containing the gene is identified in a genomic library and then subcloned and examined bit by bit. Comparing DNA from affected versus unaffected individuals in a family can then potentially turn up obvious mutations that confirm the identity of the disease gene among a number of nearby candidates. This method would come to be known as positional cloning.

Unfortunately, there simply were not enough genes on the human chromosome maps to support the sort of linkage studies needed to find diseases. In 1911, Edmund B. Wilson of Columbia University had, by default, mapped the first human gene to the X chromosome, when he discovered that the inheritance of color blindness is "sex-limited." However, human chromosome mapping progressed slowly over the ensuing decades. The presence of a testis determining factor was inferred in 1959, making it the first Y-linked "gene."

By 1980, only 120 human genes had been assigned chromosome locations. A similar number of genes (135) had been mapped in *Drosophila* chromosomes. However, the lag in human gene mapping becomes clear when one compares the average number of genes mapped per chromosome (4 pairs for *Drosophila*, 23 pairs for humans). Excluding the gene-poor chromosome that determines maleness in each species, that represented an average of 34 genes per chromosome for *Drosophila*—a gene map almost seven times denser than the human map, which had only five genes per chromosome. Moreover, the majority of *Drosophila* genes had been given precise locations by linkage analysis, as well as in situ hybridization, to banded chromosomes. Few human genes had been precisely located.

This impasse was solved by a simple proposition made by David Botstein, Ronald Davis, and Mark Skolnick at a scientific meeting at Alta, Utah in 1978. They proposed that the human chromosome map could be populated with physical variations in the DNA molecule itself. These DNA polymorphisms, which are assayed by gel or capillary electrophoresis, would substitute for phenotypic or biochemical variants used in classical linkage analysis.

DNA Polymorphisms and Human Identity

At this point, it makes sense to leave our story of finding human disease genes to discuss the parallel development of DNA polymorphisms in establishing human identity. We will return to gene cloning a little later, armed with a more detailed understanding of the evolution of DNA markers. During the 1980s and 1990s, molecular analysis would become increasingly powerful as the RFLP polymorphisms, usually having only two alleles, were supplanted by repeat polymorphisms—VNTRs (variable number of tandem repeats) and STRs (short tandem repeats)—with increasing numbers of alleles.

Although fingerprints and thumbprints have been used as personal identifiers since ancient times, only in the 20th century did they come into use in criminal cases. Francis Galton, the father of the eugenics movement, studied the fingerprints of thousands of schoolchildren from around the British Isles. His 1892 treatise, *Finger Prints*, showed how to analyze the various patterns of whorls and loops, noted their relative frequencies, and suggested how fingerprints could be rigorously used in criminal cases. This book formed the basis of forensic DNA fingerprinting still in use today.

Francis Galton's Analysis of Fingerprints, 1891

Many of Galton's examples came from schoolboys. These fingerprints are from the Hanway Street School, London. (Copyright The Galton Collection, University College London.)

The term “DNA fingerprinting” was coined to allude to the traditional use of fingerprints as a unique means of human identification. Whereas classic fingerprinting analyzes a phenotypic trait, DNA typing directly analyzes genotypic information. When properly conducted, DNA-based testing can provide positive evidence of a person’s identity. In contrast, the phenotypes detected by blood grouping and leukocyte antigen testing are shared by sufficiently large numbers of individuals that they are not, in the strictest sense, tests of identity. Rather, they are exclusionary tests that can only prove that forensic evidence does not match a suspect or that persons are not related.

All that is required for DNA fingerprinting is a small tissue sample from which DNA can be extracted. This can be blood or cheek cell samples in a paternity case, a semen sample from a rape victim, dried blood from fabric, skin fragments from under the fingernails of a victim after a struggle, or even several hairs (with the attached roots) combed from a crime scene. Ted Kaczynski, the “Unabomber,” was definitively linked to the case when his DNA type matched the one obtained from cells left when he licked a stamp used on a letter. Using the best available techniques, a DNA type can be obtained from cells in *fingerprints* on a glass or other hard surface. The time is approaching when a criminal will not be able to afford to leave even a single cell at a crime scene.

Variable Number of Tandem Repeats

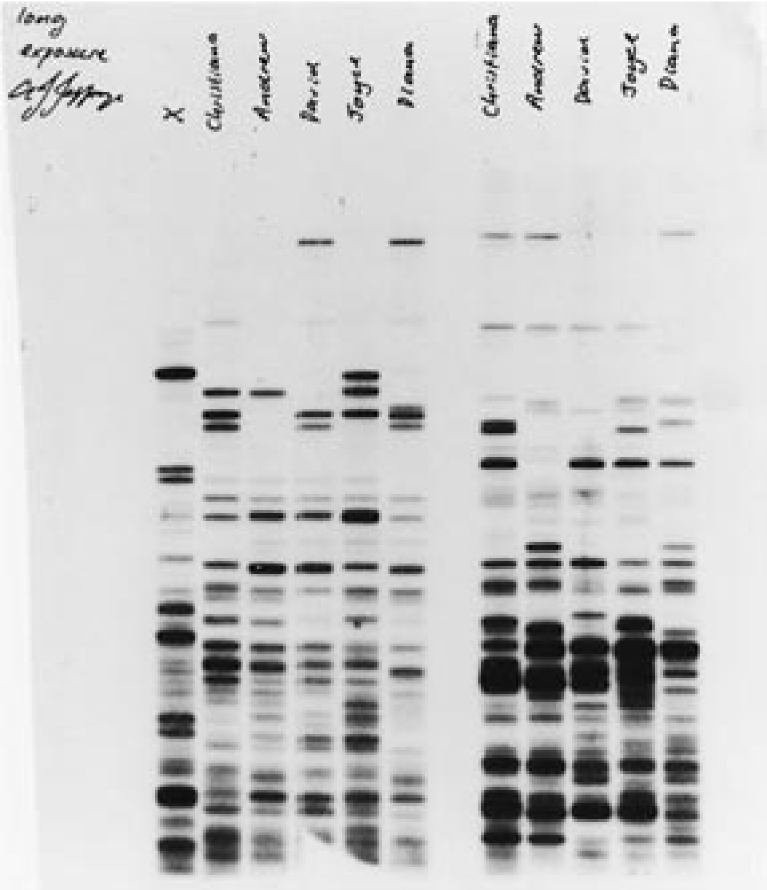
British researcher Alec Jeffreys was the first to realize that DNA polymorphisms can be used to establish human identity. He coined the term DNA fingerprinting and was the first to use DNA polymorphisms in paternity, immigration, and murder cases. The discovery, in 1984, of the so-called “Jeffreys’ probes” arose from the investigation of the “minisatellite” fraction of highly repetitive DNA in the human genome. (Recall Roy Britten’s experiments from Chapter 6.) Minisatellites, composed of short repeated DNA sequences that “hover” in a chromosome region, were first described in 1980 by Arlene Wyman and Ray White at the University of Utah. Each minisatellite proved to be composed of

Alec Jeffreys, 1989
(Courtesy of A. Jeffreys, University of Leicester.)

tandem repeated units ranging in size from 9 to 80 bp. The number of repeats at a particular locus was variable between homologous chromosomes, hence the acronym VNTR (variable number of tandem repeats).

Working at the University of Leicester, Jeffreys found two "core" sequences that are common to a set of VNTRs associated with the myoglobin gene locus. Assaying for them by Southern blotting produced a DNA fingerprint that was a composite of VNTRs at multiple loci, leading to the term "multilocus probes." In his analysis, radioactive probes hybridize to restriction fragments that have partial homology with the core sequence, typically detecting 20–30 interpretable bands. These distinctive banding patterns are inherited in a Mendelian fashion, with half of the bands derived from the mother and half from the father.

The "Ghana Immigration Case" (1985) provided the first practical test of DNA fingerprinting. The case involved Christiana Sarbah and her teenage son Andrew,



Use of Multilocus Probes in the Ghana Immigration Case, 1985

This Southern blot shows the first use of DNA fingerprints as evidence in a court of law. The match between many bands in their DNA fingerprints proved a family relationship between a Ghanian boy, Andrew, who wished to remain in England with his mother (Christiana) and his siblings (David, Joyce, and Diana). Bands not shared by a child and the mother were inherited from the father. The DNA fingerprint of an unrelated person is shown in lane X. The evidence was prepared by Alec Jeffreys, who originated the use of multilocus probes and coined the term "DNA fingerprint." (Courtesy of A. Jeffreys, University of Leicester.)

who immigrated to England after living for some time with his father in Ghana. Although depositions and other information showed that Christiana and Andrew were almost certainly related, the British Home Office ordered that Andrew be deported in the absence definitive evidence to prove Christiana's parentage. Jeffreys agreed to assist with the appeal case, believing it would be an ideal test of the DNA fingerprint technology he had recently developed. He used his myoglobin VNTR probes to produce DNA profiles from blood samples from Christiana, Andrew, and three siblings—David, Joyce, and Diana. Because of the lack of the father's blood sample, Jeffreys reconstructed the father's fingerprint from bands present in the three undisputed children, but absent in Christiana. About half of Andrew's bands matched bands in the father's compilation and the remaining bands were all present in Christiana's fingerprint. The possibility of this happening by chance is greater than one in a trillion. The Home Office accepted the DNA fingerprint evidence and allowed Andrew to stay in England.

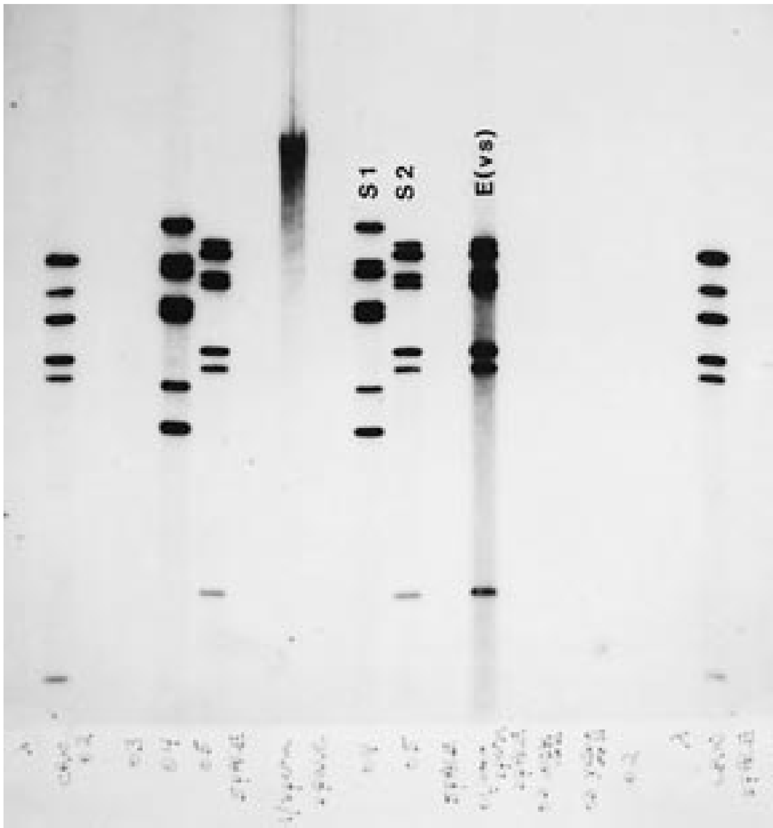
Jeffreys' probes essentially analyzed a number of VNTR polymorphisms simultaneously. The multiple bands created by the multilocus system proved difficult to analyze and standardize. The system was prone to produce artifact bands whenever a restriction enzyme failed to cut entirely, and it could be difficult to determine whether a sample had digested to completion. Furthermore, the number and frequency of alleles were never rigorously worked out, making it impossible to accurately determine the relative rarity of one fingerprint over another.

Jeffreys' multilocus probes were supplanted by single-locus probes which identify a polymorphism that occurs at a single location on one chromosome. The majority of RFLPs that had been discovered through the mid 1980s were point mutations that destroy or create a restriction enzyme recognition site. This type of RFLP has only two alleles and three genotypes (++, +–, and ––). Thus, gene mappers and forensic biologists alike sought out more variable polymorphisms as they became increasingly available in the late 1980s.

Beginning in 1987, Yusuke Nakamura, Ray White, and others at the University of Utah began a systematic search for single-locus VNTRs, ultimately providing more than 100 useful polymorphic loci scattered throughout the genome. Probes for these VNTRs became widely used in gene mapping and DNA fingerprinting. Each probe hybridizes to a unique hypervariable region of the genome and generates a pattern consisting of one or two bands from an individual's DNA, depending on whether they are homozygous or heterozygous at that locus. Used alone, a single-locus probe only detects one or two differences; however, "cocktails" of several probes came into use for forensic purposes.

Because each probe identifies a discrete locus, the frequency of each allele can be determined in population studies and the Hardy-Weinberg equation used to calculate the occurrence probability of each genotype. The VNTR loci that became most useful in forensics were those with 10 or more alleles and with a high degree of heterozygosity in many human populations, thus maximizing the ability to discriminate between two individuals. Consider the case of D1S80, a VNTR on chromosome 1 in which a 16-nucleotide unit is repeated from 14 to more than 41 times, creating 29 different alleles.

By studying the occurrence of alleles in human populations, one can calculate the probability of an individual's DNA fingerprint having this or that combi-



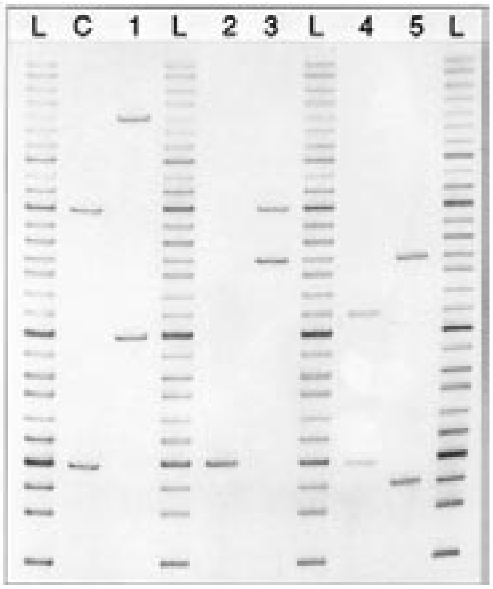
Use of Single-locus Probes in a Criminal Case, Palatka, Florida, 1988

This Southern blot is from a case in which two friends, Randall Jones (S2) and Chris Reesh (S1), were accused in the double rape-murder of a Florida woman and her boyfriend. A cocktail of single-locus probes showed an exact match between the DNA fingerprint of semen obtained from the female victim, E(vs), and the DNA fingerprint from Jones' blood sample. Jones received the death penalty—the first time this sentence was handed down in the United States on the strength of DNA fingerprint evidence. Jones is currently on Death Row at the Union Correctional Institute in Raiford, Florida. Reesh served 9 months of an 8-year sentence as accessory. (Courtesy of Cellmark Diagnostics.)

nation of DNA bands, or the probability of two DNA samples matching each other. Adding a second VNTR increases the ability to distinguish between individuals. Different VNTRs used in identity testing have been chosen on different chromosomes. This way, one can be assured that each VNTR is unlinked from the others—that each VNTR is inherited independently. If they are unlinked, then the probability of any two bands being inherited together is the product of their individual occurrences. By the mid 1990s, most forensic laboratories were producing types with five to eight unlinked markers.

Although it was initially challenged in the courts, due to its extreme sensitivity and potential for contamination, PCR eventually supplanted Southern blotting in forensic analysis. PCR made possible extremely rapid protocols that required very small amounts of template and obviated the use of radioactivity. D1S80 was among the first to be adapted for use in a forensic PCR kit. Briefly, a small sample of blood or other cells is lysed by boiling, the cell debris is removed by centrifugation, and the PCR reagents are added directly to the crude extract.

1	2	3	4	5	6	7	8	9
				Hardy-		D1S80	Locus 2	Combined
	Type	Allele	Frequency	Weinberg	Calculation	Probability	Probability	Probability
C	18/31	18	0.263	2pq	2 (0.263 x 0.058)	0.0305	0.0050	0.000153
		31	0.058					
1	24/37	24	0.318	2pq	2 (0.318 x 0.003)	0.0002	0.0035	0.0000007
		37	0.003					
2	18/18	18	0.263	p^2	(0.263 x 0.263)	0.0692	0.0075	0.000519
		18	0.263					
3	28/31	28	0.050	2pq	2 (0.050 x 0.058)	0.0058	0.0025	0.0000145
		31	0.058					
4	18/25	18	0.263	2pq	2 (0.263 x 0.055)	0.0289	0.0045	0.000013
		25	0.055					
5	17/24	17	0.013	2pq	2 (0.013 x 0.318)	0.0008	0.0065	0.0000052
		24	0.318					



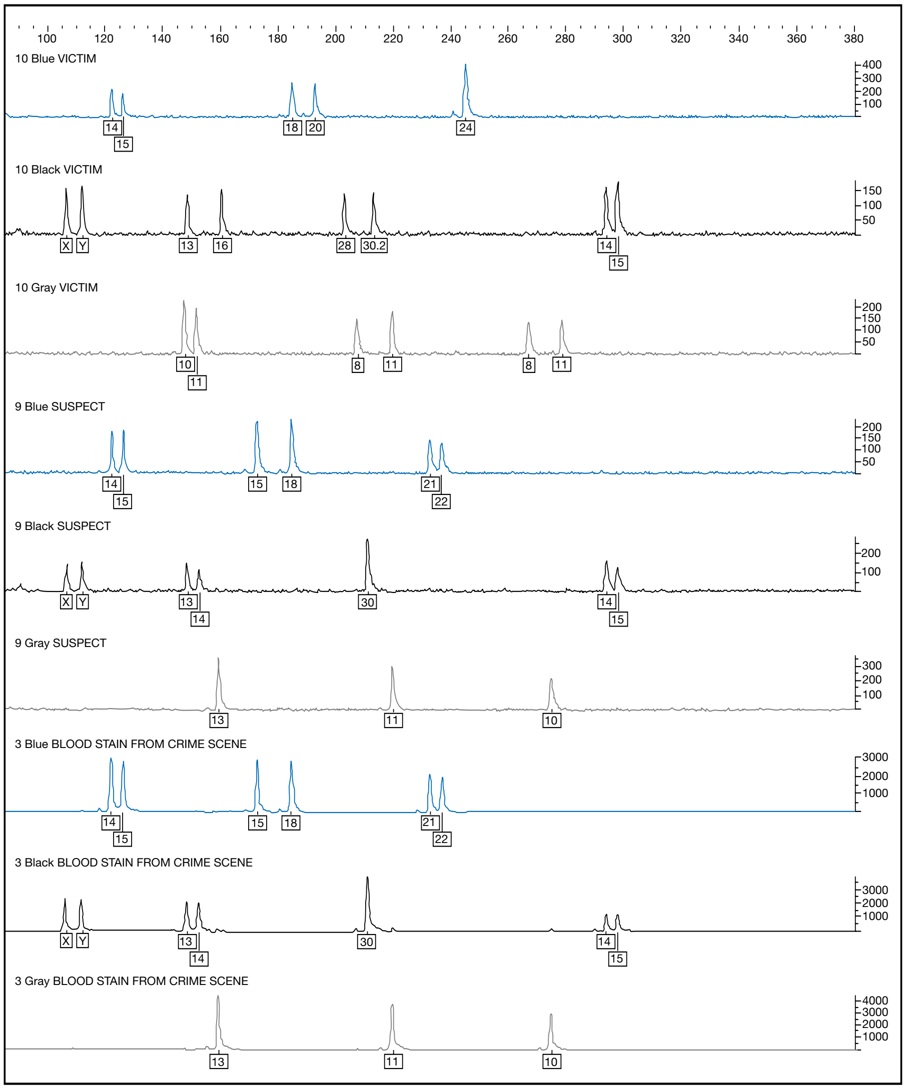
Use of PCR to Amplify the D1S80 Locus, 1991

Shown are a control (C) and five types (1–5). Type lanes show the major alleles of the system, composed of 14–41 repeats of a 16-nucleotide unit (L lanes show DNA size marker ladder). The table shows how to “score” the alleles (Columns 1–3), with allele frequencies for a Hispanic-American population (Column 4). Assuming Hardy-Weinberg equilibrium, the frequency of each DNA type is calculated in Columns 5–7. Hypothetical frequencies for DNA types at a second locus are given in Column 8. Provided the two loci are unlinked, then the combined frequency of their coinheritance is the product of their individual occurrences (Column 9). With each additional marker, the DNA types become increasingly diversified, with some types being orders of magnitude rarer than others. (Courtesy of Applied Biosystems/Perkin Elmer.)

Following the appropriate number of synthesis cycles, the amplified DNA is separated by electrophoresis in a polyacrylamide gel, stained with ethidium bromide or silver, and visualized directly.

Short Tandem Repeats

The most recent stage in the evolution of DNA polymorphisms has been the employment of “microsatellites,” with repeat units of two to five nucleotides. Their potential use in forensic DNA science was first suggested in 1992 by Thomas Caskey of the Baylor College of Medicine. The short repeat unit of STRs (short tandem repeats) creates smaller alleles, providing a greater chance of “rescuing” an STR polymorphism from degraded DNA samples than a longer VNTR. Because they came into popular use later than other polymorphic systems, STR analysis was developed primarily using PCR technology. Although STR alleles can be separated in various electrophoresis systems, their small size allows automated analysis using DNA sequencers.



Alleles Detected

Sample	Amelogenin	D3S1358	vWA	FGA	D8S1179	D21S11	D18S51
Victim	XY	14, 15	18, 20	24	13, 16	28, 30.2	14, 15
Suspect	XY	14, 15	15, 18	21, 22	13, 14	30	14, 15
Blood Stain From Crime Scene	XY	14, 15	15, 18	21, 22	13, 14	30	14, 15

Sample	D5S818	D13S317	D7S820	D16S539	THO1	TPOX	CSF1PO
Victim	10,11	8, 11	8, 11	9, 11	7, 9	9, 11	10, 12
Suspect	13	11	10	9, 12	6, 9	8, 11	9, 12
Blood Stain From Crime Scene	13	11	10	9, 12	6, 9	8, 11	9, 12

Use of STRs in a Criminal Case, Suffolk County, New York, 2000

A typical criminal case from a 1997 homicide in which a blood stain from the crime scene was tested against blood from the victim and from the suspect, who was wounded during a struggle. Shown are sequencing results for nine STR loci, plus an XY marker, using three color channels. Four additional loci were run in another channel, which is not shown. The frequency of the suspect/blood stain type in different populations is: Caucasian 0.000000000000000372 (3.73 x 10^{-16}); African-American 0.000000000000000103 (1.03 x 10^{-16}); Hispanic 0.0000000000000000267 (2.67 x 10^{-17}). (Courtesy of J. Galdi, Suffolk County Crime Laboratory.)

STR alleles are perfectly suited to fluorescent detection on a DNA sequencer, allowing forensic scientists to make use of the four dye labels (red, green, blue, and yellow) as a separate "channel." Since each STR polymorphism typically produces a tight range of common alleles, three to four STRs with differing ranges in allele size can be labeled with the same dye and detected in a single channel. With this step, "multiplex" and "megaplex" polymorphism analyses gained a mechanization and reproducibility akin to hospital metabolite testing.

In 1997, the Federal Bureau of Investigation (FBI) recommended that a 13-marker panel of STRs, plus an XY marker, become the standard in criminal investigations. With this number of independently inherited polymorphisms, the probability of even the most common combination is in the tens of billions. Thus, modern DNA testing has the capability of uniquely identifying each and every person alive today. As of June 2002, the FBI's Combined DNA Index System (CODIS) contained 1,013,746 DNA profiles, including 977,895 profiles of convicted offenders.

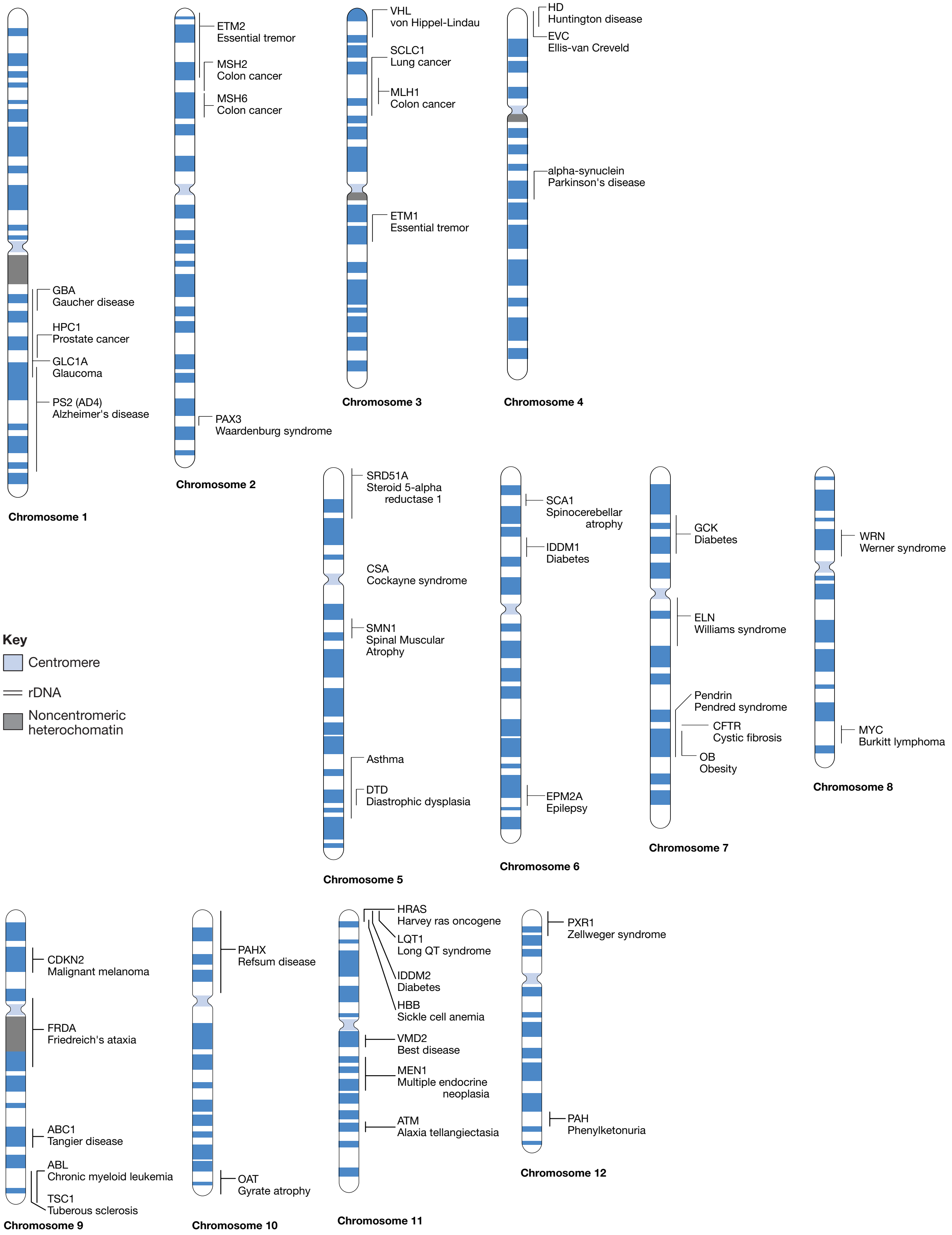
GENE CLONING: FROM LINKAGE TO DNA DIAGNOSIS

During meiosis, paired chromosomes align and homologous regions are exchanged when chromatids "cross over" with one another. Usually, large DNA fragments, on the order of tens of millions of nucleotides, are moved between chromatids. The further apart two chromosome loci (locations), the greater the possibility that they will become separated during a crossover event. As discussed in Chapter 1, the frequency of recombination is a measure of the genetic distance between two sites on a chromosome. If two loci have a recombination frequency of 1%, they will become separated once in 100 meiotic recombinations. A recombination frequency of 1% is referred to as 1 centiMorgan (cM).

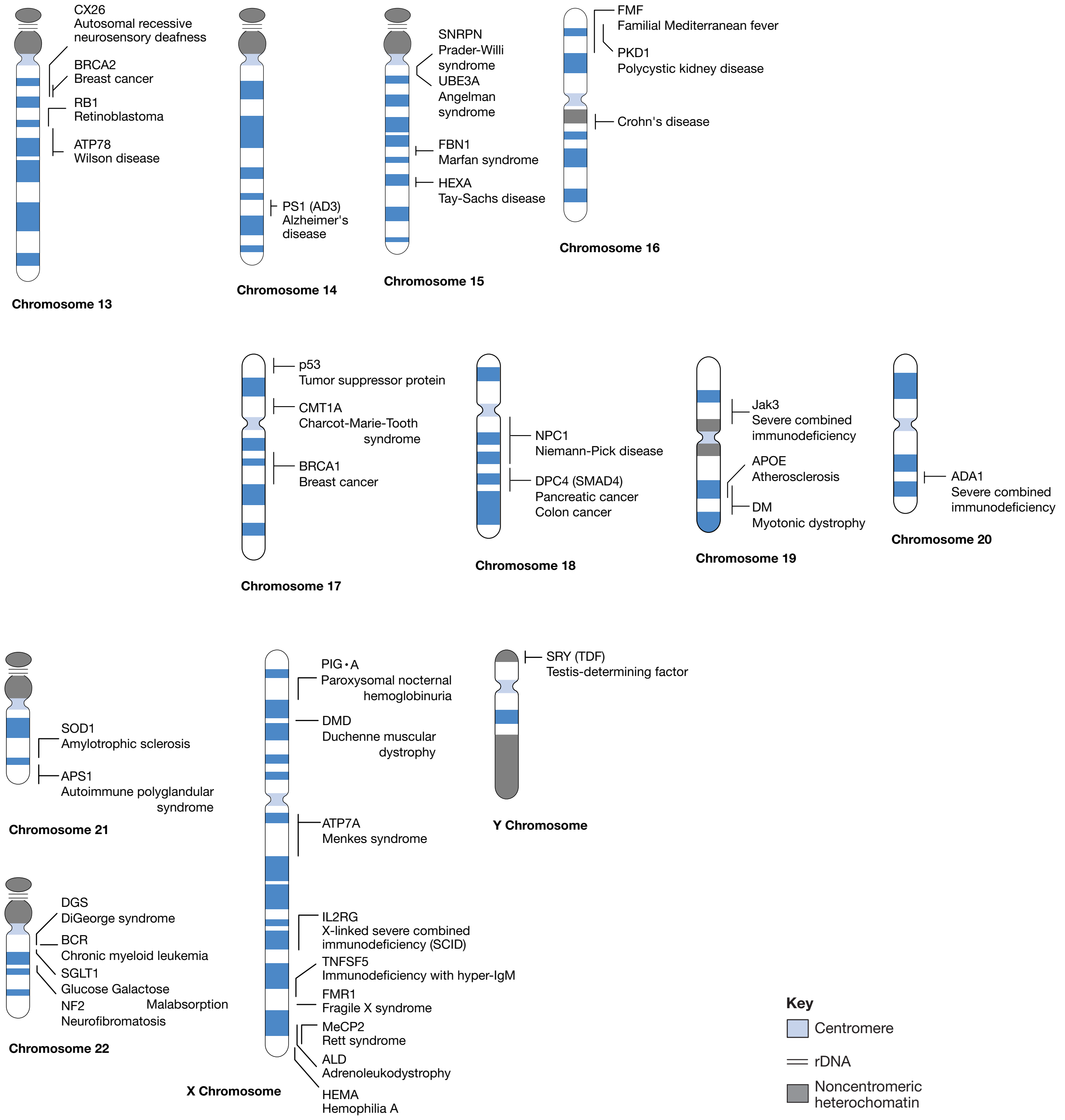
For any two chromosome loci, there is an equilibrium between two states: crossover and linkage. For distant loci, the equilibrium shifts toward a high recombination frequency. At a distance of 50 cM, recombination reaches a maximum of 50%. Loci separated by 50 cM or more are said to be unlinked, with an equal chance that they stay together or become separated during meiosis. Linkage increases as distance and recombination frequency decrease, reaching 0% for loci very close to one another.

At 0 cM distance, there is only one possible state—linkage—so equilibrium is formally impossible. Thus, when crossover never occurs between loci, they are said to be in linkage disequilibrium. The loci, and the entire region between them, are inherited as a single unit. Since the extent of linkage disequilibrium varies from region to region on the chromosome, defining blocks of linkage disequilibrium on human chromosomes is an important ongoing effort.

To pass the threshold for linkage, a marker must lie within 20 cM of a disease gene locus. This means that the two loci will be separated in 20 of 100 meiotic recombinations or, conversely, that the marker will be present in 80% of patients screened. Assuming that the probes are generated randomly and that they are evenly distributed throughout the genome, there is a 1 in 135 probability of any probe being linked to the disease (2700 cM per genome/20 cM linkage distance).

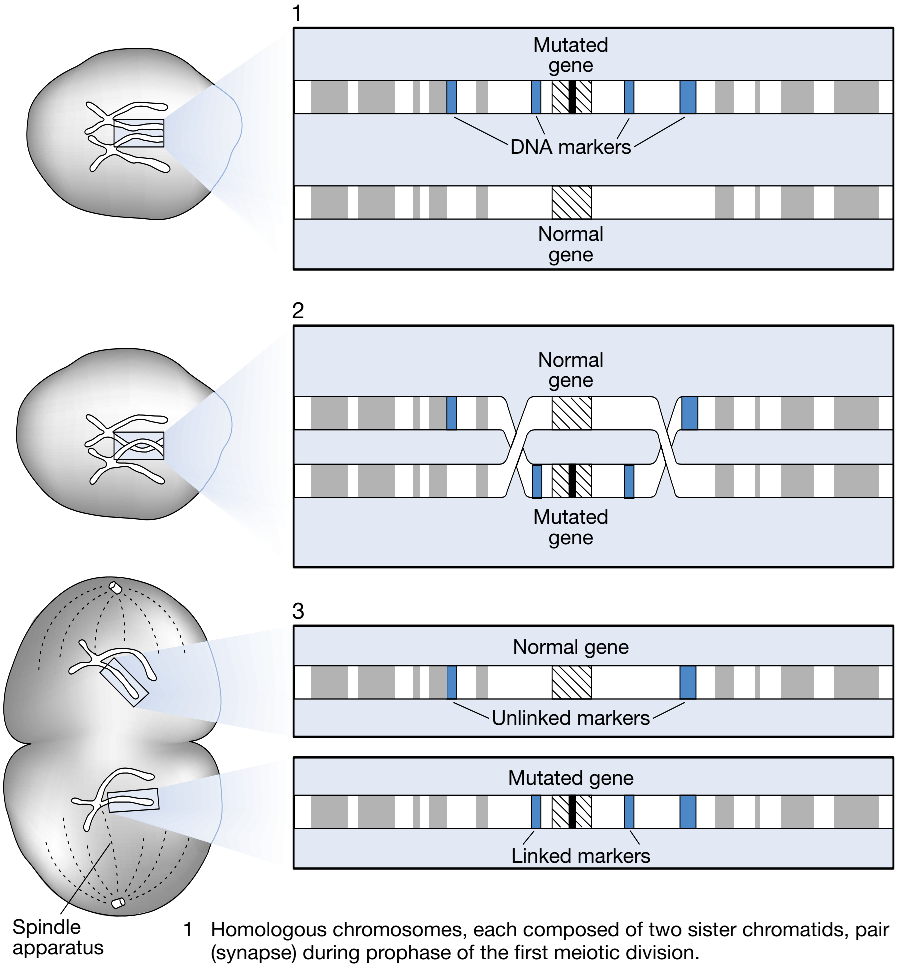


Cloned Genes Involved in Genetic Diseases and Cancer



Cloned Genes Involved in Genetic Diseases and Cancer (continued)

(Adapted, with permission, from the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland.)



1 Homologous chromosomes, each composed of two sister chromatids, pair (synapse) during prophase of the first meiotic division.

2 Recombination occurs when chromatids cross over, exchanging DNA fragments. Linked markers remain with the original chromatid, whereas unlinked markers become separated from it.

3 During anaphase, the recombined chromatids separate into two different daughter cells (in a subsequent meiotic division, the sister chromatids will segregate into separate haploid sex cells).

Fate of Linked and Unlinked Markers during Meiotic Recombination

Accurate DNA diagnosis becomes feasible once a marker has been located within 5 cM of the disease gene. At this distance, the recombination frequency (and probability that the marker and gene become unlinked) is 5%. Conversely, the coinheritance of the marker and the disease gene, as well as the accuracy of diagnosis, is 95%. The addition of a "flanking" marker within 5 cM on the other side of the gene theoretically increases the accuracy of predictions to 99.75%. (The chance of both markers becoming unlinked is 0.05 x 0.05 = 0.0025.)

DNA diagnosis relies on linking one allele of a polymorphic marker to the inheritance of a disease phenotype. Because the alleles present at the polymorphic locus may differ from family to family, it is often necessary to follow a linked polymorphism through the pedigree of the family under study. This establishes which particular polymorphic allele is associated with the disease state in that particular family. It is also necessary to identify heterozygous carriers of the disease gene in whom one polymorphic allele segregates with the disease gene and a different polymorphism segregates with the normal gene.

In general, the closer the marker to the disease locus, the more accurate the diagnosis. A marker and a gene in extremely close proximity may be in linkage disequilibrium, in which case, they are always coinherited. Such markers, including those actually located within the disease gene, can provide accurate diagnosis *without* a family history. (The sickle cell polymorphism discussed earlier falls into this category.) However, markers located within a very large disease locus may not even be in linkage disequilibrium with the gene itself. For example, markers located at the 5´ end of the *dystrophin* gene have a recombination frequency of 5%—an apparent distance of 5 cM, or about 6 million nucleotides!

Although many DNA diagnoses originally relied on Southern blotting, most new diagnostic tests rely almost exclusively on PCR. Southern analysis typically takes 24 hours or more to complete, whereas a PCR analysis can be completed in several hours.

The Triumph and Frustration of Cloning the Huntington's Disease Gene

In 1983, Huntington's disease (HD) was the first major disease locus mapped by RFLP/linkage analysis and illustrates the method's power—and difficulties. HD is a degenerative nervous system disorder that invariably leads to loss of motor function, mental incapacitation, and early death. It is a rare example of an autosomal dominant lethal. HD has been perpetuated in the human gene pool because the onset of symptoms usually occurs well after the affected individual is capable of reproducing.

The linkage analysis performed by James Gusella's group, at the Massachusetts General Hospital, was based on studies carried out simultaneously by Nancy Wexler, of Columbia University and the Hereditary Disease Foundation, on the inheritance of HD in a group of patients living at Lake Maracaibo, Venezuela. This population made an ideal case study, because it is an extended family composed of 9000 members. In addition to establishing a pedigree showing the inheritance of the disease, Wexler and her co-workers collected blood samples from more than 2000 affected persons and family members. These samples were returned to the laboratory and used to establish lymphoblastoid cell lines that can be contin-

Nancy Wexler and the "Pedigree Wall" at Columbia University, 1987

The pedigree traces the inheritance of Huntington's disease through extended families living at Lake Maracaibo, Venezuela. (Courtesy of S. Uzzell.)

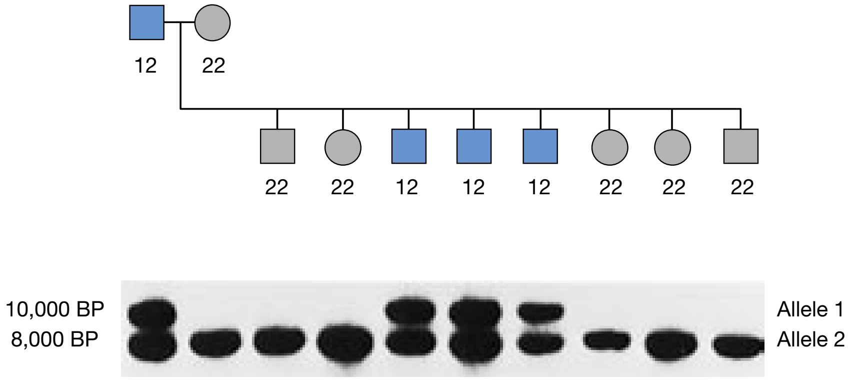
uously cultured. This is accomplished by fusing white cells in the blood with immortal cancer cells or by transforming them with an immortalizing oncogene from a tumor virus. In either case, the cultured white blood cells provide an easy source of DNA needed for the next, and most laborious, step.

James Gusella radioactively labeled at random a number of cloned DNA fragments from a human genomic library. These were then used to probe Southern blots of DNA from HD patients and unaffected family members. Gusella was looking for a probe that identifies a polymorphism whose appearance parallels the pattern of inheritance of the disease. The assumption was that all members of the extended Lake Maracaibo family inherited the disease from a single common ancestor. Thus, if a polymorphic marker is tightly linked to the disease gene, it should be coinherited by all of the affected individuals. Luck was with Gusella: A linked marker was identified with the 12th probe tested. Using this probe, he demonstrated that the HD gene is located near the telomere of the short arm of chromosome 4. Gusella's initial good fortune did not continue, and it took almost 4 years to identify a tightly linked marker, this one also located on the centromere side of the disease locus. Theoretically, he should have been able to find other linked markers located some distance from these original markers. The intervening distance would be occupied by the disease gene, thus providing flanking markers on either side. This ultimately proved impossible, because the HD gene lies very near the telomere.

HD patients turned out to have different profiles of genetic markers, indicating independent origins of the disease. Using a complex analysis of haplotype linkage disequilibrium between markers in different pedigrees, the Huntington's Disease Collaborative Research Group finally cloned the HD gene in 1993, 10 years after its initial mapping. Affected members in all 75 HD families used in the study showed a length polymorphism in the coding region of the *huntingtin*

Triplet Repeat Disorders

Disease	Protein	Number of triplet repeats (normal/disease)
Huntington's disease	huntingtin	6–35/36–180
Fragile X mental retardation	FMR1	30/60–200
Spinocerebellar ataxia		
SCA1	Ataxin-1	6–39/40–88
SCA2	Ataxin-2	14–32/33–77
SCA3	Ataxin-3	12–40/55–86
SCA6	Ataxin-P/Q Ca^{++} channel	4–18/21–31
SCA7	Ataxin-7	7–17/34–200
SCA12	PPP2R2B	7–32/55–93
Spinobulbar muscular atrophy (SBMA)	Androgen receptor	9–36/38–65
Dentatorubral and pallidolyusian atrophy (DRPLA)	Atrophin-1	3–36/49–88
Ataxia with intellectual deterioration	TATA-binding protein	25–42/45–63
Schizophrenia	KCNN3	12–28/Long alleles over-represented
Male infertility	POLG1	10/0



DNA Diagnosis of Huntington's Disease, 1987

This pedigree shows the coinheritance of a polymorphic allele and Huntington's disease. Allele 1 is linked to the Huntington's disease gene, and allele 2 is linked to the normal gene. The affected father and three affected sons (*blue*) all carry one copy of each allele. The unaffected mother and offspring (*gray*) have copies of the normal allele. (Courtesy of T.C. Gilliam, Columbia University.)

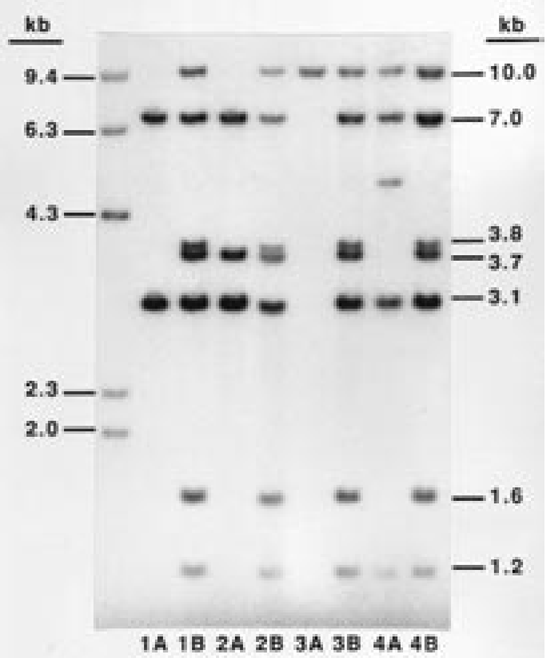
gene caused by a CAG repeat. The expansion of nucleotide triplets had been previously identified in Fragile X mental retardation and has a role in several other disorders.

The normal *huntingtin* gene has 6–35 CAG repeats; the mutated version in HD patients has 36–180 repeats. The number of repeats correlates with age when symptoms appear: Individuals with 36–41 repeats may never have symptoms, whereas those with more than 50 repeats develop symptoms before age 20. Since the repeat occurs in a coding exon, each additional repeat adds another unit of the amino acid glutamine to the expressed huntingtin protein. This alters the three-dimensional structure of the huntingtin protein, changing its interactions with other cell proteins.

Cloning the Duchenne Muscular Dystrophy Gene

Several important genes were cloned during the interval between the mapping of the HD locus and the eventual cloning of *huntingtin*. The X-linked disease Duchenne muscular dystrophy (DMD) was among the first disease loci to actually be cloned in the absence of knowledge of its protein product. The early success in identifying this gene relied on evidence showing that a number of DMD patients have large deletions clustered in a region of the X chromosome known as Xp21. In addition, females having the disease were found to have a break in their active X chromosome at position Xp21. This suggested that the deletions are associated with pathology, causing a loss of part of the normal gene at this locus. A combination of strategies, including RFLP analysis, was used to isolate the disease gene.

The *dystrophin* gene, as it has become known, is one of the largest and most complex genes yet discovered. Encompassing more than 2,000,000 bp and possessing more than 60 exons, it produces a 14,000-bp mRNA that codes for a protein containing 4000 amino acids. The *dystrophin* gene appears to be prone to



DNA Diagnosis of Duchenne Muscular Dystrophy, 1986

This early Southern blot analysis shows the DNA-banding patterns of four sets of brothers (1–4). Unaffected boys (in *B* lanes) show seven DNA bands, which are protein-coding exons of the *dystrophin* gene on the X chromosome. Brothers with muscular dystrophy (in *A* lanes) have deletions that eliminate one or more exons. (Courtesy of Jan Witkowski, Banbury Center, Cold Spring Harbor Laboratory.)

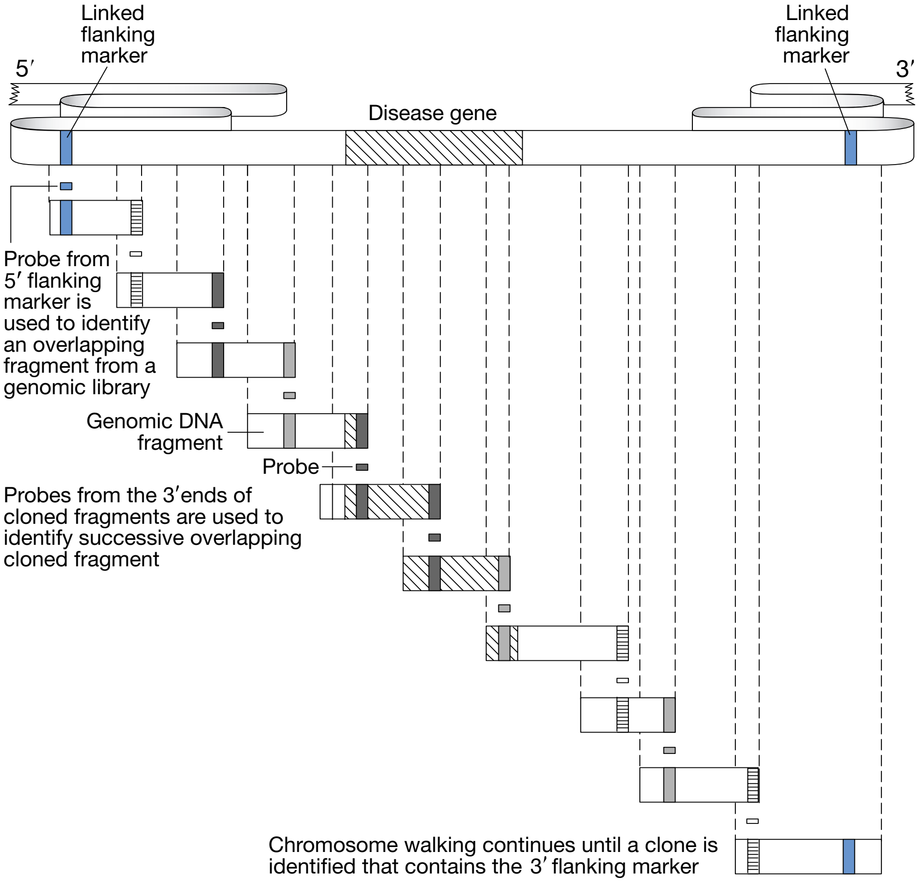
damage. DMD patients show many different deletions of exons of the gene, which effectively knock out production of any functional dystrophin. In the milder form of Becker muscular dystrophy, deletions in the *dystrophin* gene produce a semifunctional dystrophin protein. Diagnosis by Southern blot analysis, using a cDNA of the *dystrophin* mRNA, could detect various exon deletions. This was replaced by multiplex PCR, where multiple sets of primers are used to amplify ten of the commonly deleted exons in a single PCR experiment.

Cloning the Cystic Fibrosis Gene

The 1989 isolation and analysis of the causative gene for cystic fibrosis (CF) on chromosome 7 was a case study of the practice and power of modern molecular genetics. First, it resulted from an interdisciplinary collaboration among more than 25 scientists at the Hospital for Sick Children and University of Toronto, the University of Michigan, and the University of Pittsburgh. Second, it was the first disease gene identified *entirely* using the methods of positional cloning. Unlike DMD, CF is not characterized by large-scale deletions or rearrangements that could be used to map the gene to its chromosomal location.

Once a linked marker was identified within 1 cM of the disease locus, the researchers used the strategy of "chromosome walking" to clone the gene responsible for CF. Generally, chromosome walking works as follows: First, a genomic library of large DNA fragments (20,000–40,000 bp) is constructed that encompasses the disease locus. The closest linked marker is used as a probe to isolate its corresponding genomic clone. Following restriction mapping of the clone, a restriction fragment is isolated from the end of the clone closest to the disease locus. This fragment is used to reprobe the library to identify an overlapping clone. The endmost fragment of this clone is then used to reprobe the library, and another overlapping clone is isolated. Through such a succession of overlapping clones, one "walks" along the chromosome region spanning the disease locus, eventually reaching the flanking marker on the other side. The overlapping fragments are then assembled to produce a map of the disease locus.

Researchers screened ten genomic libraries and isolated the *cystic fibrosis transmembrane conductance regulator* (*CFTR*) gene, which spans approximately 250,000 nucleotides. The coding exons of *CFTR* predict a protein of 1480 amino acids. The CFTR protein is involved in the transport of sodium chloride and water in and out of the epithelial cells that line the lungs and the digestive system. As in sickle cell anemia, the primary genetic lesion in CF is a specific mutation affecting a single amino acid. Approximately 70% of CF patients show a 3-bp deletion, named deltaF508, that results in loss of a single phenylalanine residue at amino acid position 508 of the CFTR polypeptide. With this mutation, the cell excretes out too much salt, and too little water, resulting in a sticky mucus that clogs the lungs and extra salt in the patient's sweat.


Linked flanking marker
5′
Linked flanking marker
3′
Disease gene
Probe from 5′ flanking marker is used to identify an overlapping fragment from a genomic library
Genomic DNA fragment
Probe
Probes from the 3′ends of cloned fragments are used to identify successive overlapping cloned fragment
Chromosome walking continues until a clone is identified that contains the 3′ flanking marker

Using Chromosome Walking to Clone a Disease Gene

PHARMACOGENOMICS

Throughout the second half of the 20th century, major pharmaceutical companies amassed "libraries" containing hundreds of thousands of chemical compounds. These numbers have increased dramatically with the advent of combinatorial chemistry, which builds up compounds from simple chemical components—analogous to the way DNA probes are built up on a photolithographic DNA chip (Chapter 6). Each of these compounds is a potential pharmaceutical that can fight disease by altering the activity of a gene or its corresponding protein. For example, a number of compounds have been developed against histamines and other molecules involved in allergic reactions.

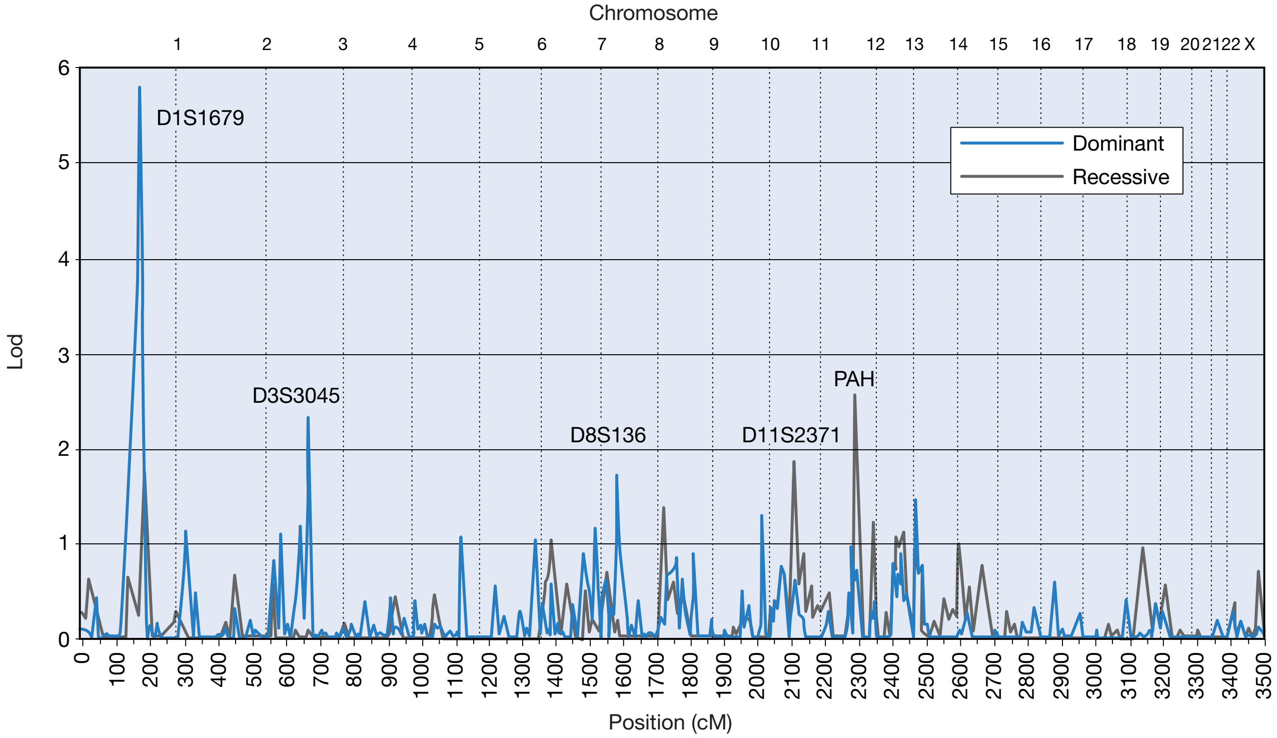
Pharmaceutical development, however, has been hampered by a relative lack of metabolic "targets" against which companies can test their huge compound libraries. The availability of the human genome sequence promises to solve this problem by presenting drug developers with a trove of new targets. Using the human genome sequence to inform drug discovery is termed pharmacogenomics. Each gene that is definitively linked to a disease becomes a validated target for drug discovery. Knowledge of mutations in that gene, and the corresponding changes in the three-dimensional structures of the encoded protein, allows one to develop strategies for screening compound libraries. Rational drug design carries this concept a step further by using the target protein's own structure to predict the properties of small molecules that can bind to an active site or otherwise modulate the protein's activity. This was the triumph of Gleevec, the first anticancer drug developed using detailed knowledge of protein kinase receptors (Chapter 7).

As more genes, and therefore proteins, are identified in a disease pathway, treatments increasingly can be tailored to specific defects in a metabolic or signal transduction pathway. Defects in different proteins in the same pathway may cause the same disease or symptoms. (Recall Beadle and Tatum's experiment, where mutations in different genes produced the same metabolic phenotype.) Thus, the same apparent disease may present different drug targets, depending on which gene in a pathway is mutated. For example, a mutation in the cytoplasmic domain of the EGF (epidermal growth factor) receptor is blocked by Gleevec; however, a different drug would be needed to counter a mutation in the EGF extracellular domain.

FINDING GENES BEHIND COMPLEX DISORDERS

Recall that most products made from cloned genes treat uncommon or tightly defined disorders. With the notable exception of statins used to treat hypercholesterolemia, the DNA revolution has not been successful in offering new or improved treatments for common disorders, such as asthma and noninsulin-dependent diabetes. Furthermore, molecular genetics and genomic biology have not yet produced a rational drug for the treatment of major behavioral disorders, notably schizophrenia and bipolar disorder (manic depression).

Asthma, diabetes, schizophrenia, and bipolar disorder are all examples of complex, or heterogeneous, disorders. Each appears to involve multiple genes whose expression is further modified by environmental factors, for example, air quality in asthma, diet in diabetes, and drug or alcohol abuse in schizophrenia and



LOD Scores for a Genome-wide Scan for Schizophrenia, 2000

Affected and unaffected individuals in 22 families with schizophrenia were genotyped at 381 marker loci throughout the genome producing significant linkage on one long arm of chromosome 1. (Reprinted, with permission, from Brzustowicz L.M., Hodgkinson K.A., Chow E.W.C., Honer W.G., and Bassett A.S. 2000. Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21-q22. *Science 288:* 678–682.)

bipolar disorder. To further complicate matters, each disorder has a range of severity and expression that may make it difficult to standardize diagnosis to the point that all researchers would evaluate the same pedigree in an identical manner.

The goal of population studies, like family studies, is to link particular DNA polymorphisms to a disease phenotype. The LOD (logarithm of the odds) score is the key statistical method used to establish linkage in family and population studies. On the basis of an observed recombination frequency between a marker and a putative disease locus, the LOD score is a ratio of the probability (odds) of a pedigree occurring at that linkage value divided by the probability (odds) of no linkage. A LOD score of 3, which is generally considered the threshold for possible linkage in a complex disorder, means that linkage is 1000 times more likely than no linkage. As a logarithmic function, like the Richter earthquake scale, each LOD score increases by a factor of 10. Thus, a LOD score of 3 represents a 10 times closer association between a marker and a locus than does a LOD score of 2.

The LOD score is extremely sensitive to changes in data analysis and laboratory errors. The change in diagnosis of a single person in an extended pedigree may be enough to lessen the association between the marker and a phenotype and, thus, weaken statistical linkage. Despite these problems, a number of genome scans and family studies during the past 10 years have reported linkage for schizophrenia with loci on chromosomes 1, 6, 8, 10, 13, 15, and 22. The case is similar for bipolar disorder, where linkage has been reported on chromosomes 4, 12, 13, 18, 21, and 22. Although the linkage reported in any single study is modest, the fact that the same regions have turned up again and again in different pedigrees is consistent with the hypothesis that multiple susceptibility genes contribute to an individual's overall risk of schizophrenia and bipolar disorder.

Many researchers believe that isolated populations offer great promise in the search for genes behind complex disorders. Since they preserve only a fraction of human diversity, isolated populations present a relatively homogeneous genetic background against which it may be easier to identify genes for common and heterogeneous disorders. Many behavioral studies have focused on the Amish, among whom alcohol and drug abuse is rarely a confounding problem. The islanders of Tristan de Cuhna, in the middle of the Atlantic Ocean, are interesting for their extremely high rates of asthma.

A polymorphism that is "informative" (coinherited with the disease locus) in one population may not be informative in a different population. Thus, as we saw in the case of the *Huntingtin* gene, individual linked markers may fail to establish linkage if a disease has arisen separately in different populations. The causative lesion of many disorders also varies, with some groups having unique, or "private," mutations not seen in other groups.

The Old Order Amish and Mennonites provide an object study in founder effect. Members of these two religious sects, known as the "plain people," live in agricultural communities where they eschew most modern technology and dress in simple clothing without adornment or buttons. To this day, they disdain motor vehicles in favor of horse and buggy. Lancaster County, Pennsylvania, remains a homeland for both groups, each of which is predominately derived from fewer than 100 individuals who settled there in the 1700s. Like many groups with small founding populations, they have concentrated mutations for some otherwise rare metabolic disorders. For example, maple syrup urine disease (MSUD) affects about 1/250,000 children in the general population but strikes about 1/400 Amish and Mennonite children.

Methylcrotonyl-CoA carboxylase (MCC) deficiency, a related but usually less severe disorder of the breakdown of the amino acid leucine, provides an example of a "private mutation." This disease has an overall frequency of about 1/50,000 in Caucasian populations, but it reaches a frequency of about 1/1500 among the Old Order Amish and Mennonites of Lancaster County. Affected Amish children have a G-to-C missense mutation at position 295 of the β-subunit of the *MCC* gene, whereas Mennonites have a frameshift mutation caused by a T insertion at position 518. Thus, although they live nearby and share a closely related religion, lifestyle, and ethnic background, each of these groups has inherited a different point mutation responsible for a rare disorder.

Single-nucleotide Polymorphisms

Imagine the complexity of searching for any of several potential genes involved in a heterogeneous disorder in which different genes or gene combinations may produce similar phenotypes in different populations. The various genes are likely to be associated with different markers in different population groups. Association studies using single-nucleotide polymorphisms (SNPs) hold potential in solving the problem of linking markers to the genes involved complex disorders.

Although the term SNP burst on the scene in the late 1990s, they are nothing but point mutations. To put this into perspective, there is about 1 nucleotide difference per 1200 nucleotides in two comparable chromosomes. This translates into about 3 million single-base differences and 100,000 amino acids differences between any two people. However, most single-nucleotide mutations are rare in a population; to be generally useful in gene scans, an SNP must have a population frequency of at least 1%.

Because SNPs are the most frequent type of polymorphism, there are potentially hundreds of useful SNP markers in a region of linkage disequilibrium that is associated with a disease gene. A region of linkage disequilibrium is termed a haploblock, because it is inherited, without recombination, like the haploid mitochondrial DNA (mtDNA) or the Y chromosome. A set of SNPs, or other markers, within the haploblock are inherited together as a haplotype.

Different populations have accumulated different SNPs within the haploblock. Thus, affected individuals from different populations may share certain markers within the haploblock, whereas other markers will be unique in certain populations. Just as one may find a consensus sequence for promoter regions and intron/exon splice junctions, haplotypes can be identified that represent a consensus of SNPs that are coinherited with the disease gene across many populations. Although no individual SNP is likely to have great predictive value, the combinatorial effect of an SNP haplotype can be a powerful tool in linkage studies. Thus, in 2002, the National Human Genome Research Institute announced a 100-million-dollar project to establish a haplotype map of the human genome, potentially containing 200,000 haploblocks.

Many researchers are confident that once the human genome map is heavily populated with SNPs, disease genes can be identified in heterogeneous populations of unrelated individuals. For example, a sample could be drawn from a database of all individuals who suffer from severe asthma, irrespective of their population group. An equivalent control sample is then drawn of healthy people. Each patient and control are SNP typed across his or her whole genome or across a specific candidate region. A haplotype is constructed for each person, using SNPs that occur in haploblock regions of the genome. Then, computer algorithms search for a consensus haplotype that is associated with the disease locus. Since haplotypes may encompass tens or hundreds of SNPs, this type of association analysis is much more complex for determining linkage with one marker at a time. It is not clear how saturated with SNPs the genome map must be before pure association analysis of this type will become possible, but it may be as few as 1 million SNPs.

Pharmacogenetics

Everyone at one time must have taken pause at the paradox of a physician asking us if we are allergic to a particular drug. After all, the doctor should be the one to inform us of a potential problem. Unfortunately, trial and error is the only way to determine a response to most drugs—it takes an allergic reaction to know if we are allergic! SNPs offer the potential of predicting a negative response *before* a drug is taken. Thus, the endgame of genetic medicine is pharmacogenetics, predicting drug response and tailoring treatment to each person's genetic makeup. However, before we enter this era of personalized medicine, experts today believe we must pass through a period of "population medicine," where drugs are targeted according to a generalized profile of the population group that most closely matches the patient.

Although it is very much in vogue today, the term "pharmacogenetics" was first coined in 1959 by Freidrich Vogel. This was based on earlier evidence that drug responses are inherited and vary between population groups. Notably, African American soldiers serving in Italy during World War II suffered adverse effects, including hemolysis, from the antimalaria drug primiquine. This was correlated with glucose-6-phosphate dehydrogenase (G6PD) deficiency, which, ironically, provides some protection against malaria.

Drug response is largely mediated by so-called metabolic enzymes in the liver—the cytochrome P450 monooxidases (CPY450s)—which detoxify compounds and metabolize many drugs into their bioactive forms. People who are "extensive metabolizers" efficiently convert a given drug to its active form and/or metabolize it at a rate that provides the desired therapeutic effect. "Poor metabolizers" fail to convert enough of the drug to its active form or metabolize it at a rate that fails to produce a therapeutic effect. "Toxic metabolizers" convert the drug into a toxic product or metabolize it so slowly that it accumulates to toxic levels.

In the late 1970s, Robert Smith of St. Mary's Hospital, London, noticed an unusually high incidence of side effects, including an unusual fainting response, among patients prescribed the antihypertension drug debrisoquine. He found that about 8% of Caucasians (but less than 2% of Black and Asian populations) are poor metabolizers, handling debrisoquine 10–200 times less efficiently than extensive metabolizers. Michel Eichelbaum, of the University of Bonn, found similar disparities in the metabolism of sparteine, an anti-arrythymic. This led to the realization that both drugs are metabolized by the CPY2D6 enzyme and that poor metabolizers inherit a defective CPY2D6 enzyme. Subsequent work revealed that CPY2D6 is involved in deficient responses to at least 40 common drugs, including codeine, dextromethorphan, beta-blockers, monoamine oxidase inhibitors, tricyclic antidepressants, antipsychotics, neuroleptics, and fluoxetine (Prozac). Cloning and sequencing of the *CPY2D6* gene in 1988 showed that poor metabolizers have polymorphisms that produce splicing errors or amino acid substitutions. Recent research showed that several SNP haplotype pairs in *CPY2D6* predict response to the anti-asthma drug albuterol, with striking differences in haplotype distribution between population groups.

Screening for relevant *CPY450* polymorphisms would be a perfect application for gene chip and would be a logical first step in the development of pharmacogenetics.

THINKING ABOUT HUMAN HISTORY AND POPULATIONS

Each person's unique disease susceptibilities and responses to drugs are, in large part, the balance between our uniqueness as individuals and the similarities we share with others in historical population groups. Written in each person's DNA is a record of our shared ancestry and our species' struggle to populate the earth. Our ancient ancestors moved around and eventually out of Africa. They moved in small groups, following river valleys and coastlines, reaching Asia and Europe. Land bridges that appeared during recurring Ice Ages allowed them to reach Australia and the Americas.

As these early people wandered, their DNA accumulated mutations. Some provided advantages that allowed these pioneers to adapt to new homes and ways of living. Most were nonessential. Mutations are the grist of evolution, producing gene and protein variations that have allowed humans to adapt to a variety of environments—and to become the most far-ranging mammal on the planet. The same mutational processes that generated human diversity—point mutations, insertions/deletions, transpositions, and chromosome rearrangements—also generated disease.

It may be hard to see from our current vantage point, but the entire industrial revolution has occupied only about 0.1% of our 150,000-year history as a species. The cradles of western civilization—classical Greece and Rome—take us back into only 2% of our history. The earliest city-states of Mesopotamia, Babylonia, Assyria, and China take us back only 4% of way into our past. At 6%, we reach the watershed of agriculture, which changed forever the way humans live and work. After language, the domestication of plants and animals is the single greatest civilizing factor in human history. Increased production and performance of domesticated organisms made possible urbanization and task specialization in human society. Thus, the labor of fewer and fewer farmers produced enough food and clothing materials to satisfy the needs of growing numbers of nonfarmers—artisans, engineers, scribes, and merchants—freeing them to develop other elements of culture. Reaching back the remaining 93% of our history, to the dawn of the human species, we lived only as hunter-gatherers.

The fastest evolving part of our genome, the mitochondrial control region, accumulates about one new mutation every 20,000 years. Mutations are five- to tenfold less frequent in most regions of the nuclear chromosomes. Thus, virtually every gene in our genome is, on average, only one event away from our hunter-gatherer heritage. This leads to two far-reaching conclusions that substantially broaden our understanding of evolutionary processes and the origin of human disease:

- Throughout most of human history, the hunter-gatherer group was a basic population unit upon which evolution acted.
- Our basic anatomy, physiology, and many aspects of behavior are essentially identical to the hunter-gatherers who ranged through the ancient landscapes of Africa, Europe, Asia, Australia, and the Americas.

It may be difficult for many people today to conceive of what is meant, in a genetic sense, by a human population. This is because, over the past quarter century, people have become extremely mobile. Airplanes and four-wheel vehicles have made it possible to travel virtually anywhere in the inhabited world within a day or two. Major urban centers have become cosmopolitan, with mixes of people representing many races and cultures. Even so, today there are still regions of the world where people are born, reproduce, and die all in the same village. This essentially defines the "classical" definition of a human population: a group of people who, by reason of geography, language, or culture, preferentially mate with another.

Unique human populations—for example, the Saami of Finland, the Ainu of Japan, the Nanuit of Alaska, the Yanomami of Brazil, the Pygmies of Central Africa, and the Bushmen of Southern Africa—have preserved unique cultures and languages. Their genomes preserve the genetic residue of a time when all human beings lived in smaller and more cohesive groups. Small populations are subject to the founder effect, “inbreeding,” and genetic drift (a random fluctuation of nonessential alleles). Over millennia, these effects join with selection to concentrate particular gene variations within different population groups. Gene variations come into equilibrium when a population grows to several thousand individuals.

THE BIOLOGICAL CONCEPT OF RACE

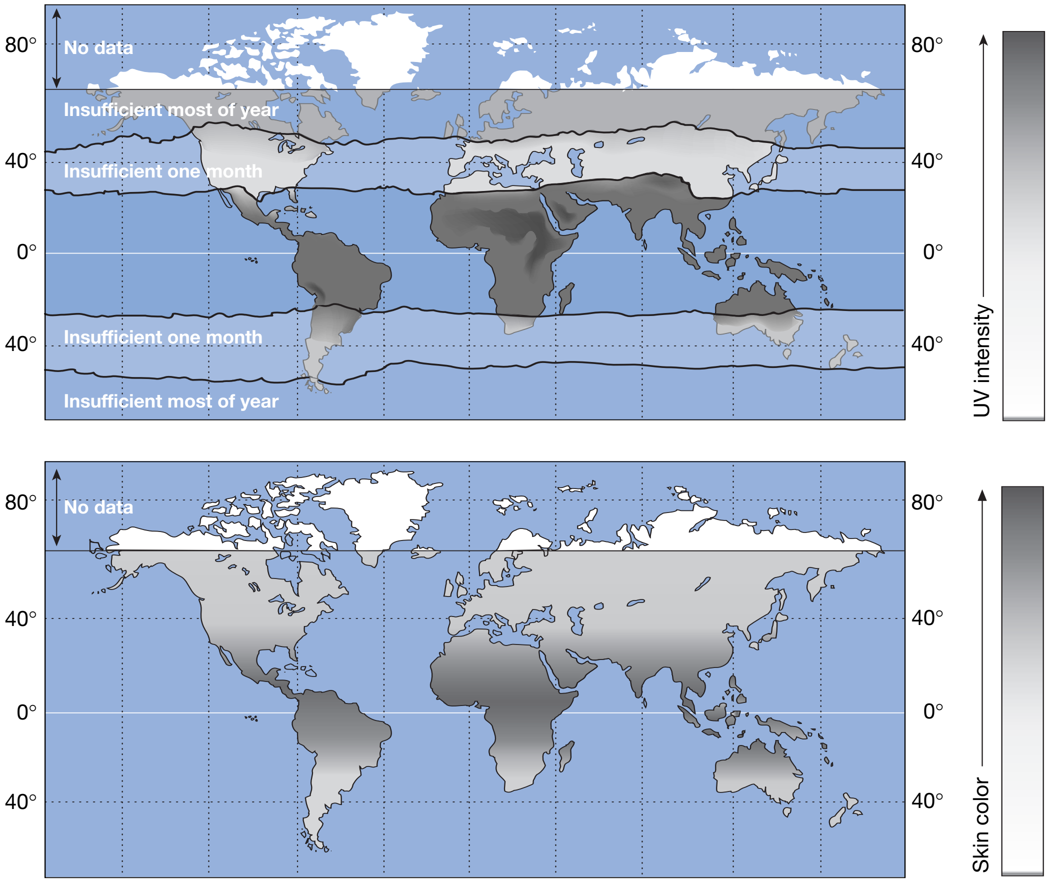
Most people can readily define characteristics that make them different from others. The most obvious difference between people is the color of their skin, followed by hair and eye color, hair texture, and shapes of body and facial features. These physical characteristics, in combination with cultural and religious practices, have been generalized into the related concepts of race and ethnicity. Unfortunately, racial and ethnic prejudices have fueled many of the worst events in human history.

The physical characteristics we associate with race and ethnicity likely are controlled by a mere handful of the 30,000 to 50,000 genes in the human genome. Variation in human skin color is determined by levels of two different forms of a pigment produced by melanocytes in the dermis layer of the skin. Eumelanin is brown-black and pheomelanin is red-yellow. However, the genetic basis of pigmentation is not well understood, and only one gene involved in human pigment variation has been located. But why did different population groups develop different skin colors in the first place?

Biologists assume that early human ancestors had light skin covered by dense hair—like our near primate relative, the chimpanzee. Australopithecines and other early human ancestors probably looked and acted like tall chimpanzees, but with the ability to walk upright for longer periods of time. *Homo erectus*, with its striding gate, spent more time tracking prey and foraging in the open African savannas. Increased activity in the open sun, and a larger brain to be protected from overheating, necessitated efficient evaporative cooling. This would have selected for individuals with larger numbers of sweat glands. (Chimps have very few sweat glands.) However, wet hair hinders evaporation, so a trend toward evaporative cooling also favored a reduction in body hair. This is a plausible explanation of how humans came to have nearly hairless bodies.

In the absence of protective hair, it is generally assumed that dark-pigmented skin developed in hominid populations in Africa as protection against the damaging effects of UV radiation. Most skin cancers develop later in life, well after reproductive age. Thus, it seems unlikely that melanin’s anti-cancer effect, alone, could have provided enough selective advantage for dark skin.

In 1967, W. Farnsworth Loomis, of Brandeis University, offered an explanation of why lighter skin evolved among populations living at higher latitudes. Short-wavelength UV (UVB) radiation in sunlight triggers a reaction in the skin to produce vitamin D, which is important in skeletal formation and immune function. Vitamin D synthesis by the skin is especially important for people without diets rich in this vitamin, as would have been the case for most early hominids. Thus, Loomis hypothesized that lighter skin offered a selective advantage as people migrated out of Africa, allowing them to absorb more of the reduced UV light that penetrates the atmosphere in the higher latitudes.



Ultraviolet Light and Vitamin D Production, and Skin Color

(*Top map*) Populations that live in the tropics near the equator receive enough UV light from the sun to synthesize vitamin D all year long. In temperate zones, people lack sufficient UV to make vitamin D at least one month of the year. Those nearer the poles do not get enough UV light most months for vitamin D synthesis. (*Bottom map*) Shown are predicted skin colors for humans based on UV light levels. In the Old World, the skin color of indigenous peoples closely matches predictions. In the New World, however, the skin color of long-term residents is generally lighter than expected, probably because of their recent migration and factors such as diet. (Adapted, with permission, from Jablonski N.G. and Chaplin G. 2002. Skin deep. *Sci. Am.* *287:* 74–81.)

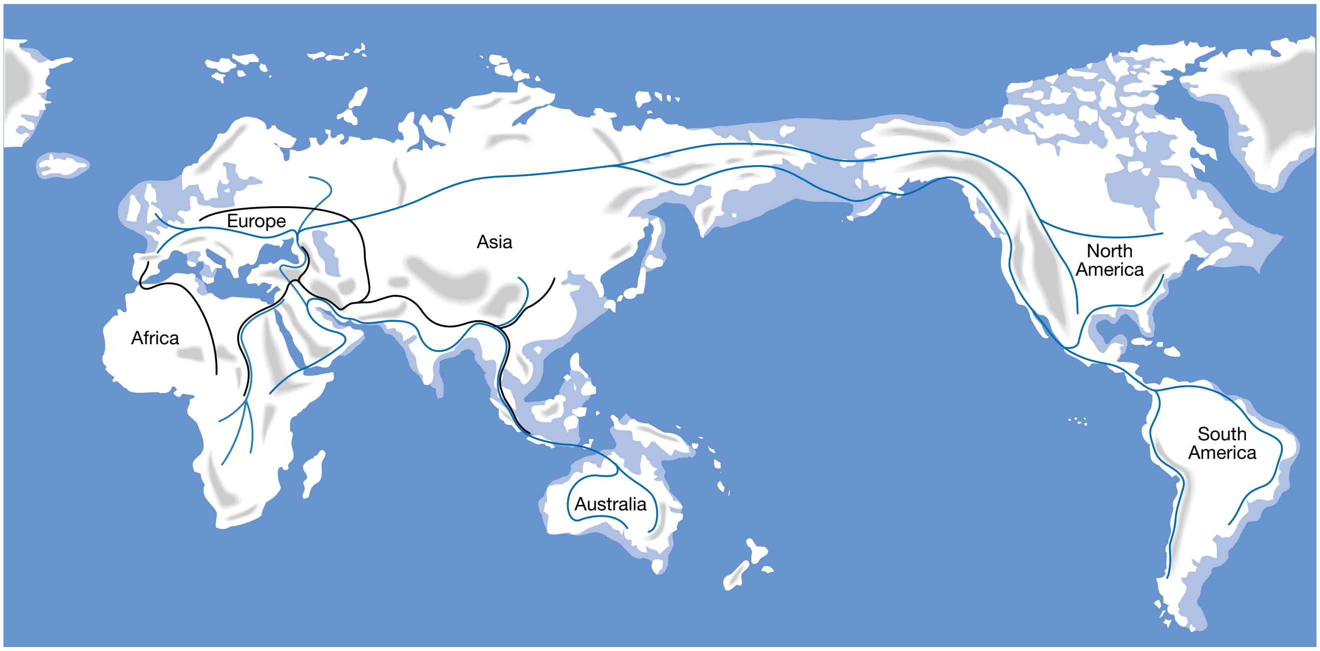
Recent modeling of worldwide UVB radiation, based on satellite mapping of the earth's ozone layer, shows a correlation between levels of UVB that are sufficient for vitamin synthesis and skin pigmentation. Thus, dark-pigmented skin is found in the tropics where there is sufficient UVB to synthesize vitamin D year-round. Lighter skin, but with the ability to tan, is found in the subtropical and temperate regions, which have at least 1 month of insufficient UVB radiation. Very light skin that burns easily is found north of 45 degrees, where there is insufficient UVB year-round.

In 2000, Nina Jablonski and George Chaplin, of the California Academy of Sciences, offered a more complete explanation for the evolution of dark-pigmented skin among early hominids in Africa. They proposed that melanin protects the body's stores of the B vitamin folate, which is essential for reproduction and embryonic development. This conclusion came from the synthesis of several lines of research: (1) Exposure to sunlight rapidly reduces folate levels in the blood. (2) Treating male rodents with folate inhibitors impairs sperm development and induces infertility. (3) Folate deficiency during pregnancy, including reduction apparently induced by overuse of tanning beds, increases risk of neural cord defects in infants. Thus, as early hominids spent more time hunting and gathering on the open savanna, those with darker skin would have had greater reproductive success and produced more healthy offspring.

WHAT THE FOSSIL RECORD TELLS US ABOUT HUMAN EVOLUTION

Thoughts about the alleged differences between the races pale when one considers that evolutionary theory, as well as popular genealogy, demands that all human beings alive today share a common ancestor at some point in the distant past.

The fossil record shows that the human species arose in Africa, and all people alive today share a common ancestor there. Anthropologists estimate that the human lineage diverged from other primates about 6–7 million years ago, with chimps being our closest living relative. Among the most primitive human ancestors were members of the genus *Australopithecus*, which lived about 3 million years ago. Remains of Australopithecines have been discovered primarily



“Out of Africa” Theory of Human Evolution

Ancient humans of the species *Homo erectus* left Africa 1.8 million years ago, reaching Europe and Asia (*black lines*). Groups of *Homo sapiens,* from whom all modern humans are descended left Africa about 70,000 years ago (*blue lines*). These groups replaced any remaining ancient populations, reaching Asia and Australia about 60,000 years ago and entering Europe about 45,000 years ago.

in the Rift Valley of Africa. Early members of our own genus, *Homo erectus*, arose in the same region about 2.5 million years ago. These "archaic" hominids migrated out of Africa approximately 1.8 million years ago to found populations in Europe, the Middle East, and southern Asia.

The earliest fossils of modern humans, or artifacts made by them, have been found in southern and eastern Africa, dating to about 140,000 years ago. Remains of modern humans dating to 100,000 years ago have been found in the Middle East, to 60,000 years in Asia and Australia, and to 45,000 years in Europe. By modern humans, we mean members of our own species *Homo sapiens*, who share with us important anatomical features (skull shape and size) and behavioral attributes (use of blades, bone tools, pigments, burial goods, representational art, long-distance trade, and varied environmental resources). These humans subsequently spread to Micronesia, Polynesia, and the "New World" (North and South America).

How modern humans emerged is a matter of debate between proponents of two opposing theories. Supporters of the *multiregional theory* contend that modern human populations developed independently from archaic hominid (*Homo erectus*) populations in Africa, Europe, and Asia. Early modern groups evolved in parallel with one another and exchanged members to give rise to modern population groups.

Supporters of the *displacement theory*, commonly known as "out of Africa," contend that all modern human populations are derived from one or several modern population groups that left Africa beginning about 70,000 years ago. These founding groups migrated throughout the Old World, displacing any surviving archaic hominids. Thus, scientists all agree that our earliest hominid relatives arose in Africa, but disagree on when the direct ancestors of living humans left Africa to populate the globe.

THE DNA MOLECULAR CLOCK

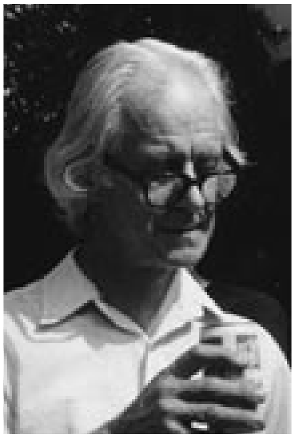
At first thought, there does not seem to be any way in which DNA could provide us information about the origin of modern humans. The oldest hominid DNA isolated thus far dates back only about 60,000 years, well after our emergence as a species. In fact, we can study our evolutionary past by looking at the DNA variation of humans and primates alive today. Although this might seem to be a contradiction, remember that the DNA of any individual bears the accumulated genetic history of its species.

When two groups split off from a common ancestor, each accumulates a unique set of random DNA mutations. Provided mutations accumulate at a constant rate, and occur sequentially (one at a time), then the number of mutations is proportional to the length of time that two groups have been separated. This relatively constant accumulation of mutations in the DNA molecule over time is called the "molecular clock." An event that has been independently established by anatomical, anthropological, or geochronological data is used to attach a time scale to the clock. For example, the human molecular clock is typically set using fossil and anatomical evidence suggesting that humans and chimps diverged 6–7 million years ago.

Because of its high mutation rate, the mitochondrial control region evolves more quickly than other chromosome regions—it has a faster molecular clock. The fast mutation rate means that lineages diversify rapidly, amplifying differences between populations. However, rapid mutation also introduces the confounding problem of "back mutation" where the same nucleotide mutates more than once, returning it to its original state. Multiple mutations at the same position also cause an underestimation of the total number of mutation events. Thus, the number of observed differences between human and chimp sequences are less than one would expect to have occurred in the 6–7 million years since the lineages diverged. The chance of back or multiple mutations is much smaller over the period during which modern humans have arisen. So, the number of observed mutations among living humans is very close to the actual number that has accumulated since we arose as a species.

Mitochondrial DNA (mtDNA) offers another important advantage in reconstructing human evolution: With very few documented exceptions, the mitochondrial chromosome is inherited exclusively from the mother. This is because mitochondria are inherited from the cytoplasm of the mother's large egg cell. Any paternal mitochondria that may enter the ovum at the moment of conception are identified by different ubiquitin proteins expressed on their surface and destroyed. The lack of paternal chromosomes with which to recombine greatly simplifies the analysis of mitochondrial inheritance. The mitochondrial genome is inherited intact over thousands of generations, without the confounding effect of crossover with a paternal chromosome. Because the mitochondrial genome is haploid, having only a contribution from the mother, mtDNA types are termed haplotypes ("half-types").

WHAT DNA TELLS US ABOUT HUMAN EVOLUTION



Allan Wilson

(Courtesy of Cold Spring Harbor Laboratory Archives.)

Throughout the 20th century, fossils provided the only tools for reconstructing human origins and the field remained virtually the sole province of anthropologists. In 1987, Allan Wilson and co-workers at the University of California at Berkeley moved anthropology into the molecular age when they made mtDNA haplotypes for 145 living humans. Using a molecular clock like that described above, they constructed a tree that extrapolated back to a common ancestor who lived about 200,000 years ago.

It is important to note that the tips of the branches of Wilson's evolutionary tree were the 145 individual humans whose mtDNA types he had determined. Although most individuals generally came out on branches with others from their regional population group, some individuals fell on branches with other groups. This illustrated a high degree of mixing between human population groups. Importantly, Africans turned up on several non-African branches, but only African individuals were found on the branch closest to the root of the tree. Thus, Wilson concluded that the so-called mitochondrial "Eve" most likely lived in Africa, providing DNA evidence for the recent dispersion of modern humans "out of Africa."

During the next decade, molecular reconstructions of human lineages were conducted with autosomal and Y chromosome polymorphisms. Unlike mitchondrial SNPs, which have a high rate of back mutation, each Y chromosome SNP is believed to represent a unique mutation event that occurred once in evolu-

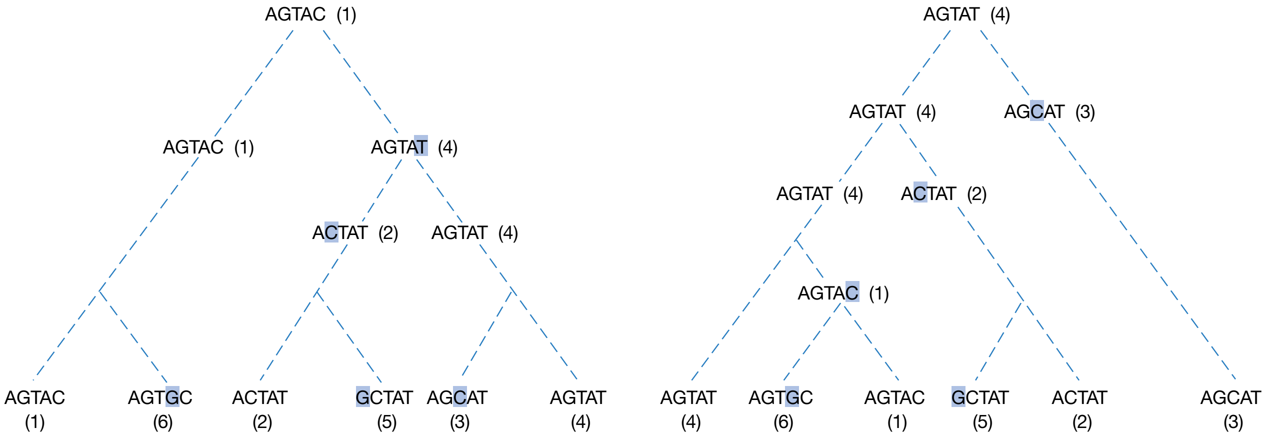
A. Six allelic sequences

Sequence 1) agctggctgaatgctatctgcgtcgcgcgaaataacgtcagcaattcgttacat**c**tctctagggc
Sequence 2) agctggctgaatgctatctgc**c**tcgcgcgaaataacgtcagcaattcgttacatttctctagggc
Sequence 3) agctggctgaatgctatctgcgtcgcgcgaaa**c**aacgtcagcaattcgttacatttctctagggc
Sequence 4) agctggctgaatgctatctgcgtcgcgcgaaataacgtcagcaattcgttacatttctctagggc
Sequence 5) agctggctga**g**tgctatctgc**c**tcgcgcgaaataacgtcagcaattcgttacatttctctagggc
Sequence 6) agctggctgaatgctatctgcgtcgcgcgaaataacgtcagca**g**ttcgttacat**c**tctctagggc

B. The six haplotypes for the five SNPs in the allelic sequences above.

```
Position                11          22          33          44          55
Sequence 1)  ----------a----------g----------t----------a----------c----------
Sequence 2)  ----------a----------c----------t----------a----------t----------
Sequence 3)  ----------a----------g----------c----------a----------t----------
Sequence 4)  ----------a----------g----------t----------a----------t----------
Sequence 5)  ----------g----------c----------t----------a----------t----------
Sequence 6)  ----------a----------g----------t----------g----------c----------
```

C. Two possible trees for the evolutionary relationships among the haplotypes above, assuming that mutations occur sequentially.



Using Haplotypes to Create Trees Showing Evolutionary Relationships

tionary history. The accumulated DNA evidence has confirmed Wilson's original story. Several lines of evidence point to the emergence of modern humans in Africa about 150,000 years ago.

- The greatest amount of DNA variation occurs in Africa, suggesting that African populations have been accumulating mutations for the longest period of time. Europeans have only about half the variation of African groups, suggesting that they are only about half as old.
- Most Asian and European variations are a subset of variations found in African populations, suggesting that Asian and European populations are derived from an African source.
- The deepest roots of a tree diagram of human variation contain only Africans. Ancient alleles have not been found in non-African populations.

Interestingly, comparisons of mitochondrial and Y chromosome polymorphisms suggest that men and women have had different roles in the peopling of

A

Greece	GGTACCACCCAAGTATTGACTCACC	25
Japan	GGTACCACCCAAGTATTGACTCACC	
Greece	CATCAACAACCGCTATGTATTTCGT	50
Japan	CATCAACAACCGCTATGTATTTCGT	
Greece	ACATTACTGCCAGCCACCATGAATA	75
Japan	ACATTACTGCCAGCCACCATGAATA	
Greece	TTGTACGGTACCATAAATACTTGAC	100
Japan	TTGTACGGTACCATAAATACTTGAC	
Greece	CACCTGTAGTACATAAAAACCCAAT	125
Japan	CACCTGTAGTACATAAAAACCCAAT	
Greece	CCACATCAAAACCCCC**T**CCCC**AT**GC	150
Japan	CCACATCAAAACCCCC**C**CCCC**GC**GC	
Greece	TTACAAGCAAGTACAGCAATCAACC	175
Japan	TTACAAGCAAGTACAGCAATCAACC	
Greece	**C**TCA**A**CTATCACACATCAACTGCAA	200
Japan	**T**TCA**G**CTATCACACATCAACTGCAA	
Greece	CTCCAAAGCCACCCCTCACCCACTA	225
Japan	CTCCAAAGCCACCCCTCACCCACTA	
Greece	GGATA**C**CAACAAACCTACCCACCCT	250
Japan	GGATA**T**CAACAAACCTACCCACCCT	
Greece	TAACAGTACATAGTACATAAAGCCA	275
Japan	TAACAGTACATAGTACATAAAGCCA	

B

Greece	CCAAGTATTGACTCACCCATCA**A**CA	25
Neandertal	CCAAGTATTGACTCACCCATCA**G**CA	
Greece	ACCGCTATGTAT**T**TCGTACATTACT	50
Neandertal	ACCGCTATGTAT**C**TCGTACATTACT	
Greece	G**CC**AG**CC**ACCATGAATATTGTAC**G**G	75
Neandertal	G**TT**AG**TT**ACCATGAATATTGTAC**A**G	
Greece	TACCATAA**A**TACTTGAC**C**ACCTG**T**A	100
Neandertal	TACCATAA**T**TACTTGAC**T**ACCTG**C**A	
Greece	GTACATAAAAACC**C**AATCCACATCA	125
Neandertal	GTACATAAAAACC**T**AATCCACATCA	
Greece	AA**A**CCCCC**T**CCCCATGCTTACAAGC	150
Neandertal	AA**C**CCCCC**C**CCCCATGCTTACAAGC	
Greece	AAG**T**ACAGCAATCAACC**C**TCAACT**A**	175
Neandertal	AAG**C**ACAGCAATCAACC**T**TCAACT**G**	
Greece	TCA**C**ACATCAACT**G**CAACTCCAAAG	200
Neandertal	TCA**T**ACATCAACT**A**CAACTCCAAAG	
Greece	**C**C**A**CCC**C**T**–**CACCCACTAGGATA**C**C	225
Neandertal	**A**C**G**CCC**T**T**A**CACCCACTAGGATA**T**C	
Greece	AACAAACCTACCCACCCTT**A**ACAGT	250
Neandertal	AACAAACCTACCCACCCTT**G**ACAGT	
Greece	ACATAG**T**ACATAAAG**C**CATTTACCG	275
Neandertal	ACATAG**C**ACATAAAG**T**CATTTACCG	

C

Chimpanzee	TTCTTTCATGGGGAAGCA**A**ATTT**AA**	25
Greece	TTCTTTCATGGGGAAGCA**G**ATTT**GG**	
Chimpanzee	GTACCACC**T**AAGTA**C**TG**G**CTCA**TT**C	50
Greece	GTACCACC**C**AAGTA**T**TG**A**CTCA**CC**C	
Chimpanzee	AT**T**A**–**CAACCGCTATGTATTTCGTA	75
Greece	AT**C**A**A**CAACCGCTATGTATTTCGTA	
Chimpanzee	CATTACTGCCAGCCACCATGAATAT	100
Greece	CATTACTGCCAGCCACCATGAATAT	
Chimpanzee	TGTAC**A**GTACCATAA**TC**AC**CCA**ACC	125
Greece	TGTAC**G**GTACCATAA**AT**AC**TTG**ACC	
Chimpanzee	ACCT**A**TAG**C**ACATAAAA**T**CC**ACC**TC	150
Greece	ACCT**G**TAG**T**ACATAAAA**A**CC**CAA**TC	
Chimpanzee	**–**ACAT**T**AAAACC**TT**C**A**CCCCATGCT	175
Greece	**C**ACAT**C**AAAACC**CC**C**T**CCCCATGCT	
Chimpanzee	TACAAGCA**C**G**C**ACA**A**CAATCAACCC	200
Greece	TACAAGCA**A**G**T**ACA**G**CAATCAACCC	
Chimpanzee	**C**CAACTATC**GA**ACAT**A**AA**ACA**CAAC	225
Greece	**T**CAACTATC**AC**ACAT**C**AA**CTG**CAAC	
Chimpanzee	TCCAA**C**G**A**CAC**TT**CTC**C**CCCAC**CCT**	250
Greece	TCCAA**A**G**C**CAC**CC**CTC**A**CCCAC**TAG**	
Chimpanzee	**A**ATACCAACAAACCTACCC**T**CCCTT	275
Greece	**G**ATACCAACAAACCTACCC**A**CCCTT	

D

Neandertal	CCAAGTATTGACTCACCCATCAGCA	25
Neandertal	CCAAGTATTGACTCACCCATCAGCA	
Neandertal	ACCGCTATGTAT**C**TCGTACATTACT	50
Neandertal	ACCGCTATGTAT**T**TCGTACATTACT	
Neandertal	G**TT**AG**TT**ACCATGAATATTGTACAG	75
Neandertal	G**CC**AG**CC**ACCATGAATATTGTACAG	
Neandertal	TACCATAATTACTTGACTACCTGCA	100
Neandertal	TACCATAATTACTTGACTACCTGCA	
Neandertal	GTACATAAAAACCTAATCCACATCA	125
Neandertal	GTACATAAAAACCTAATCCACATCA	
Neandertal	A**A**CCCCCCCCCCCATGCTTACAAGC	150
Neandertal	A**C**CCCCCCCCCCCATGCTTACAAGC	
Neandertal	AAGCACAGCAATCAACCTTCAACTG	175
Neandertal	AAGCACAGCAATCAACCTTCAACTG	
Neandertal	TCATACATCAACTACAACTCCAAAG	200
Neandertal	TCATACATCAACTACAACTCCAAAG	
Neandertal	ACGCCCTTACACCCACTAGGATATC	225
Neandertal	ACGCCCTTACACCCACTAGGATATC	
Neandertal	AACAAACCTACCCACCCTTGACAGT	250
Neandertal	AACAAACCTACCCACCCTTGACAGT	
Neandertal	ACATAGCACATAAAGTCATTTACCG	275
Neandertal	ACATAGCACATAAAGTCATTTACCG	

Was Neandertal Our Direct Ancestor?

Representative two-way comparisons of mitochondrial control region sequences used to determine whether Neandertal was the direct ancestor of modern humans. Sequence differences are highlighted in blue. (*A*) Comparison of modern Greek and Japanese humans shows 6 differences over 275 nucleotides (less than the 379 nucleotides analyzed by Svante Pääbo). (*B*) Modern Greek and Neandertal comparison shows 26 differences. (*C*) Modern Greek and a chimpanzee comparison shows 48 differences. (*D*) Comparison between two Neandertals shows 6 differences.

the planet and in the mixing of genes among population groups. Generally, mtDNA types show a gradation (or cline) of allele frequencies from one geographic region to another. This is the signature of "gene flow," the slow and steady exchange of genes between adjacent populations, which occurs in many cultures when women leave their families to live in their husbands' villages. Y polymorphisms, on the other hand, show discontinuities between adjacent regions, suggesting that men have not moved freely between local groups. However, related Y chromosome types do leave the signature of migrations and war campaigns that abruptly transplant genes over long distances. Thus, members of the Black South African tribe, the Lemba, have the telltale signature of Cohanin Jewry displaced from the Middle East. The most common Y chromosomes in cosmopolitan southern Japan clearly were transplanted from Korea in the last several hundred years, but the Ainu of the northern islands of Japan have an ancient affinity with Tibetans.

Was Neandertal Our Direct Ancestor?

Since their initial discovery in the Neander Valley of Germany in 1856, the heavy-set bones of Neandertal have fascinated scientists, as well as the general public. Neandertal had a brain capacity within the range of modern humans and was certainly an archaic member of the genus *Homo*. Neandertal ranged through Europe, the Middle East, and Western Russia beginning about 300,000 years ago and became extinct about 30,000 years ago. Clearly, during part of its span on earth, Neandertal shared its habitat with modern humans. Thus, there was a longstanding controversy about whether Neandertal was the direct ancestor of modern humans.

According to the multiregional model, modern humans developed concurrently from several distinct archaic populations living in different parts of the world. Under this model, Neandertal must be the intermediate ancestor of modern Europeans. Other archaic hominid fossils, Java and Peking man (*Homo erectus*), were the ancestors of modern Asians. According to the "Out of Africa" model, Neandertal was displaced by modern *Homo sapiens* who arrived in Europe about 40,000 years ago.

Svante Pääbo, 1997
(Photo by Margot Bennett, Cold Spring Harbor Laboratory.)

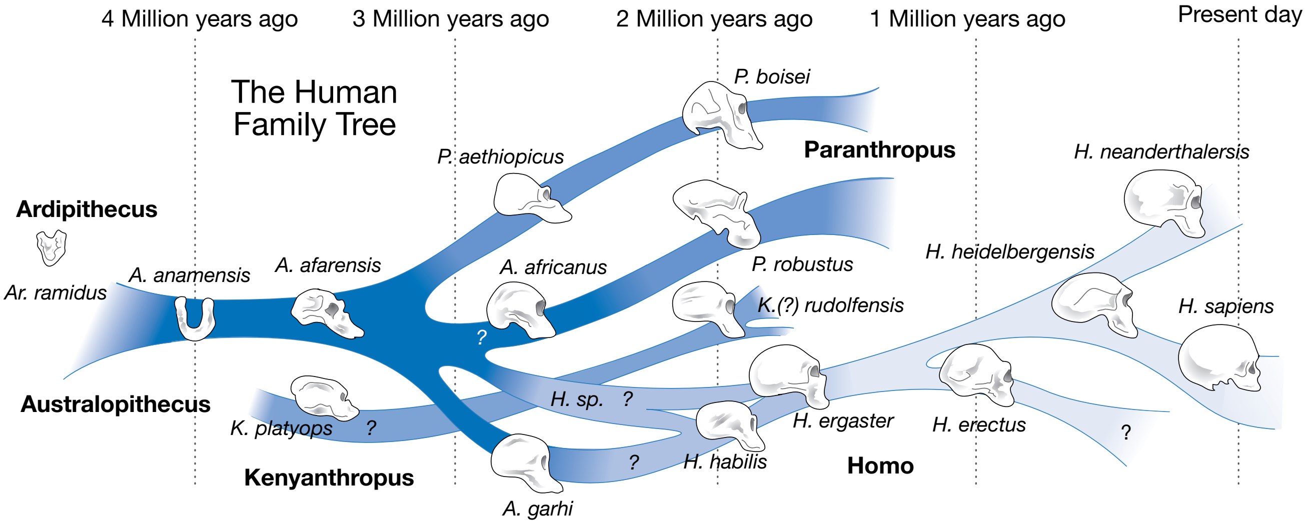
In 1997, at the University of Munich, Svante Pääbo, a student of Allan Wilson, further revolutionized human molecular anthropology when he added a 40,000-year-old DNA sample to the reconstructions of hominid evolution. Pääbo extracted DNA from the humerus of the original Neandertal-type specimen, amplified the sample by PCR, and cloned the resulting products in *E. coli*. The cloned fragments were then used to reconstruct a 379-bp stretch of the mitochondrial control region. Pääbo drew about 1000 human mitochondrial control region sequences from the Genbank database and compared them in pairs. He found an average of 7 mutations between these pairs of modern humans, representing the average variation accumulated since the divergence from a common ancestor. However, he found an average of 27 mutations when he compared each of the 1000 modern human sequences against the reconstructed Neandertal sequence. This put Neandertal outside the range of variation of modern humans.

If one takes the mutation rate of the mitochondrial control region to be about 1/20,000 years, then 7 mutations equals 140,000 years to a convergence to a common ancestor, in accordance with the earliest fossil record of modern humans. The 27-mutation difference between living humans and Neandertal suggests that our lineage converges on (or diverges from) a common ancestor about 550,000 years ago. This weighed still more DNA data in favor of the "Out of Africa" model.

"Bushy" Evolution

Until about 30 years ago, the "single-species hypothesis" dictated that only a single human ancestor lived on the earth at any point in prehistory. Each new Australopithecine or hominid-like fossil was immediately considered one of a direct succession of ancestors of modern humans. This created the image of an evolutionary "tree" with a long, straight trunk and essentially no branches until one reached racially and ethnically diverse modern humans. Thus, when racial segregation was still a way of life in the United States and elsewhere around the world, "straight-line" evolution provided scientifically minded people the comfort of believing in a shared evolutionary past, but put some distance between modern human groups who considered themselves different from each other.

This followed from a concept of gradual evolution, in which features that define a new stage are essentially properties of the preceding stage. In other words, there was a certain preordained "sense" in the way evolution proceeded. However, as more and more distinctive fossils emerged, it became impossible to fit them all into one straight path toward modern humans. Some had to be evolutionary dead ends and not a part of our direct ancestry. The concept of a "bushy" human pedigree, with short branches due to frequent extinctions, was



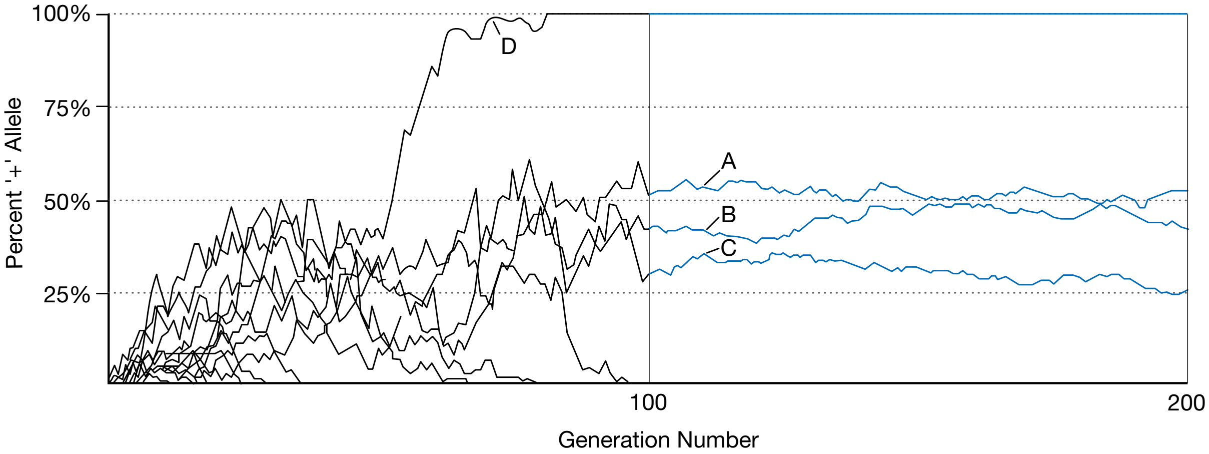
Bushy Evolution

Current theories of hominid evolution show a bushy lineage, rather than a straight line. This evolutionary scheme is based on the work of Donald Johanson. (Redrawn, with permission, from the Institute of Human Origins, Arizona State University.)

predicted by the model of punctuated evolution articulated in the early 1970s by Niles Eldridge of the American Museum of Natural History and Stephen Jay Gould of Harvard University. According to this model, evolution proceeds in abrupt fits and starts, as adaptive changes arise in small, dispersed populations.

Svante Pääbo's demonstration that the mtDNA variation of Neandertal lies outside the variation of modern humans provided DNA evidence to support the bushy tree concept. It makes better sense of the fossil evidence, but also makes better sense of the fact that most genetic variation accumulated during a time in human history when hunter-gatherer populations averaged perhaps 50 persons. A bushy evolutionary tree is exactly what one would expect from selection acting upon "private" sets of mutations that arise in small, relatively segregated groups of hunter-gatherers. It is most useful to consider that mutation and selection are separate events, often widely separated by time. Thus, neutral mutations anticipate future circumstances—environmental or behavioral—when they may provide a selective advantage.

Working from the premise that most mutations did not confer an immediate selective advantage to the hunter-gatherer, then most new mutations would survive or be extinguished in the group according to the whims of genetic drift. In this way, different hunter-gatherer groups accumulated different sets of mutations. On occasions, one or more of the accumulated mutations would provide a selective advantage to one population or another—and groups exchanged alleles through intermarriage. However, as climate, circumstances, or the luck of genetic drift changed, different lineages would become extinct, leaving behind a fossil record of its distinctive features.



Genetic Drift in Hunter-Gatherer Groups

This simulation illustrates the fate of a new, neutral mutation that occurs in 100 hunter-gatherer groups. Each group maintains a population of 50 individuals who mate randomly, with respect to the new allele (+). The + allele is lost from the majority of populations within 10 generations and from 96% of populations after 100 generations. However, the + allele rises to frequencies of 30–50% in three populations (*A,B,C*) and reaches 100% in a single population (*D*). At the end of 100 generations, each population expands to 2000 individuals, and the + allele frequency stabilizes in populations *A*, *B*, and *C* over the next 100 generations. This represents the sort of population growth enabled by the advent of agriculture (although it would not have occurred over a single generation). Populations *A*, *B*, *C*, or *D* would thrive under a new environmental challenge to which the mutation confers a selective advantage. Thus, the drift of a neutral allele may favor certain populations in the future.

Climate Changes and Population Bottlenecks

The bushy nature of the tree of human evolution tells us that many hominid lineages have become extinct in the past. Several lines of evidence suggest that the lineage shared by all humans alive today also came perilously close to extinction at one or more points in the last 100,000 years.

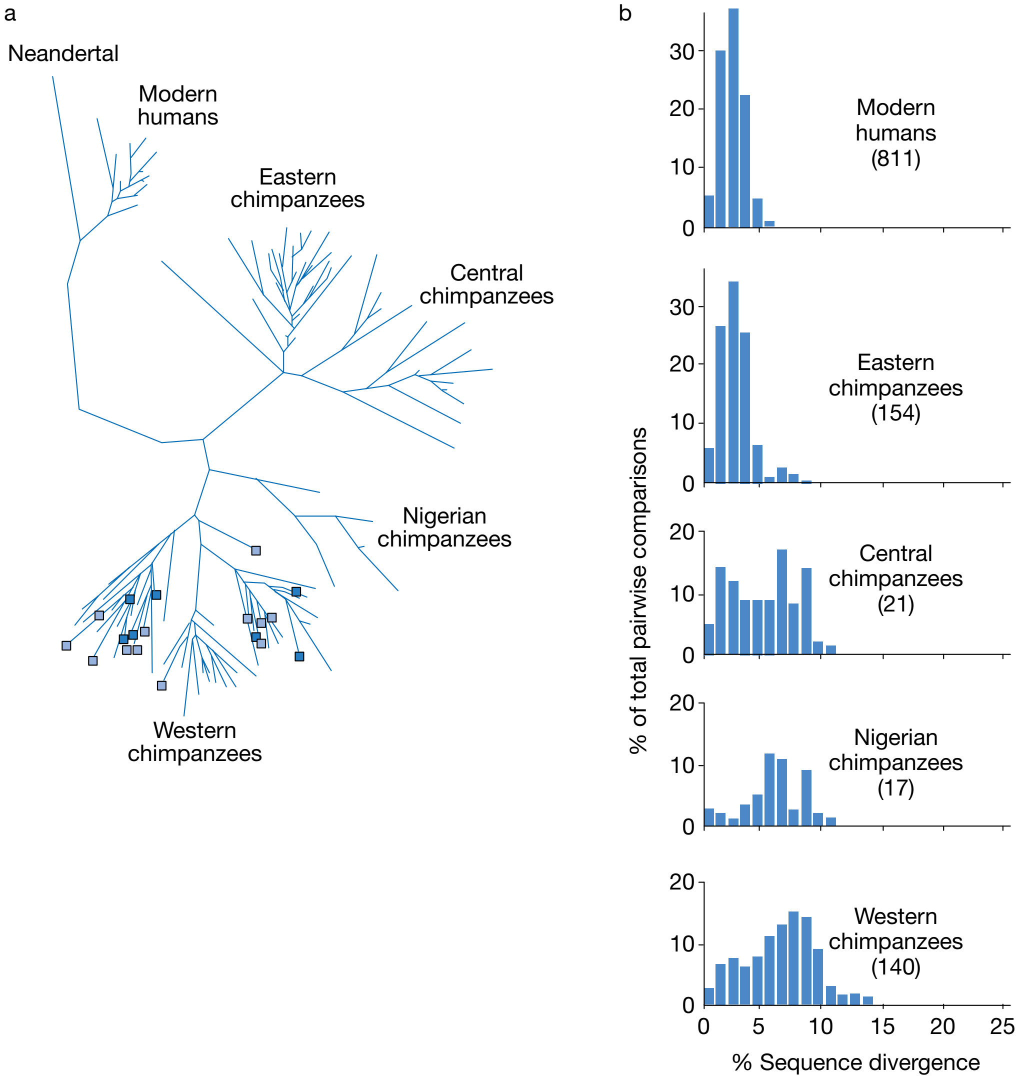
On the surface, humans appear very diverse. Different populations have acquired distinctive morphological adaptations, including skin color, body shape, and pulmonary capacity, that allowed humans to inhabit virtually every biogeographical region of the earth. Despite these morphological differences, the human species as a whole has surprisingly little genetic diversity. Differences between populations account for only about 10% of human genetic variance, and there is no evidence for separate human subspecies.

This contrasts sharply with chimpanzees, which are restricted to similar habitats in equatorial Africa and have few morphological differences. Despite this seeming homogeneity, scientists recognize as many as four distinct subspecies of chimps living in eastern, central, and western Africa and Nigeria. There is substantial genetic diversity between, and within, the chimp subspecies. Notably, there is a greater diversity of mtDNA types among members of a single troop of western chimps than among all humans alive today. This striking lack of genetic diversity in the human species supports the contention that we are a young species that has gone through several "bottlenecks" that drastically reduced the human population—and genetic diversity. During these periods, the entire human population may have shrunk to as few as 1000 individuals, clinging to life in scattered refuges.

The fossil record shows that modern humans first left Africa about 100,000 years ago, traveling via the Sinai Peninsula into the Middle East. However, this group seems to have stalled, never reaching Asia or Europe. The track of modern human migrations grows cold worldwide until about 60,000 years ago, when we find evidence of a second movement out of Africa, hugging the coast of the Sinai Peninsula and across the narrow neck of the Red Sea into Asia and Australia. There is no sign of *Homo sapiens* in Europe until about 45,000 years ago, when they appear to burst on the scene at a number of sites almost simultaneously.

Stanley Ambrose, of the University of Illinois, has provided a plausible explanation for the aborted first venture of *Homo sapiens* out of Africa. This was the volcanic eruption of Mt. Toba in Sumatra, about 71,000 years ago. Toba's 30 x 100-km caldera (more than 30 times the size of Crater Lake in Oregon) was produced by the largest known eruption of the Quaternary Period, the most recent 1.8 million years of the earth's history. The eruption produced as much as 2000 km^3 of ash, leaving ash beds across the Indian Ocean and into mainland India. By comparison, the 1984 eruption of Mt. St. Helens in Oregon produced 0.2 km^3 of ash.

Although the airborne ash would have settled after several months, Mt. Toba also injected huge amounts of sulfur into the atmosphere, where it combined with water vapor to form sulfuric acid. Greenland ice cores show heavy sulfur deposition for six years following the eruption, indicating a lingering, sun-obscuring haze worldwide. The reduced solar radiation reaching the earth's surface would have lowered sea surface temperatures by about 3ºC for several



Comparison of Genetic Diversity in Humans and Chimpanzees

(*a*) The family tree compares mitochondrial control region sequences from 811 humans and 332 chimps. Note the extensive branching of chimp groups, especially the western subspecies, indicating a high degree of diversity. Blue boxes indicate chimps from two social groups, each of which exhibits greater diversity than the entire human population. (*b*) The bar graphs show pairwise sequence differences. Note the tight clustering of humans with less than 5% sequence divergence, compared to the broader distribution and greater sequence divergence among central, Nigerian, and western chimps. (Reprinted, with permission, from Robinson R., ed. 2003. *Genetics*. MacMillan Reference USA, New York; Figure created by Dr. Stanley H. Ambrose, Department of Anthropology, University of Illinois, Urbana.)

years. Pollen records suggest that much of Southeast Asia was deforested following the eruption, and significant changes are also recorded in the pollen profile of Grand Pile, in France. Greenland ice cores show that the eruption of Toba was followed by 1000 years of the lowest oxygen isotope ratios of the last glacial period, indicating the lowest temperatures of the last 100,000 years. Thus, it is not difficult to believe that the eruption of Toba produced several years of volcanic winter, followed by 1000 years of unrelenting cold. This surely would have decimated human populations outside of the scattered refuges in tropical Africa.

60 Miles
18 Miles

Toba Caldera, Sumatra, Indonesia
(Adapted from the Landsat Pathfinder Project at: http://edcdaac.usgs.gov/pathfinder/pathpage.html.)

The fossil record shows that the modern humans who first reached the Middle East were replaced after the Toba eruption by cold-tolerant Neandertals, illustrating that adaptations are relative to environmental factors. Interestingly, Neandertal seem to have gone through similar population bottlenecks during its 250,000 or so years on earth. Mitochondrial control region sequences have been obtained from two additional Neandertal specimens from Croatia and the Caucasus. Added to the German sample, these represent about half the range of Neandertal. Comparisons of these samples suggest that Neandertal had only about the same (limited) level of diversity as modern human populations, despite the fact that this species existed nearly twice as long as currently has *Homo sapiens*.

The Hunter-Gatherer Remains

Before the advent of RFLP and SNP data, disease and protein polymorphisms of the blood system, including ABO and Rh groups, human leukocyte antigens, and globin variants, provided the only means to study human population variation quantitatively. Using these data, Luigi Luca Cavalli-Sforza, Paolo Menozzi, and Alberto Piazza were among the first to attempt to reconstruct the genetic history of Europe. They identified gradients, or clines, in gene frequencies across Europe using principal component analysis. The first principal component identified a gradient emanating from the Middle East and diminishing through northwestern Europe. They interpreted this east-west cline as genes that origi-

Ötzi the Iceman

(Reprinted, with permission from Schiermeir Q. and Stehle K. 2000. Frozen body offers chance to travel back in time. *Nature 407:* 550.)

nated in the Fertile Crescent and spread westward with agriculture, at a rate of about 1 km per year. Farming arose about 10,000 years ago and roughly marks the boundary between the Paleolithic (Old Stone Age) and Neolithic (New Stone Age). Because the east-west component accounted for the greatest variance across many genes, they concluded that Neolithic genes had essentially replaced Paleolithic genes. In their quest for new agricultural lands, farmers moved inexorably northwest through Europe, mixing with and eventually displacing the hunter-gatherer populations they encountered.

The advent of mtDNA typing of ancient remains and living Europeans challenged Cavalli-Sforza, Menozzi, and Piazza's "demic diffusion" model of the wave-like expansion of Neolithic farmers. The first data came from "Ötzi the Iceman," a 5000-year-old mummy found frozen in the Tyrolean Alps in 1991. An international team obtained mitochondrial control region sequences from tissue samples in 1994. Ötzi's mtDNA type, and others differing from his by only a single nucleotide, proved common among people alive today of European ancestry. Ötzi's DNA looks modern, because he is fully modern. Five thousand years is less than a single mutation from the present, even by the fast mitochondrial clock.

Bryan Sykes of Oxford University extended the human mitochondrial lineage directly back to the Paleolithic, when he analyzed DNA from a human tooth from the Cheddar Gorge in the south of England. Excavated from a limestone cave, the tooth was carbon dated to 12,000 years ago, approximately 6,000 years before farming reached England. Even so, like Ötzi the Iceman, Cheddar Man's identical DNA type and many others differing from his by only a single nucleotide are common among Europeans alive today. Clearly, Cheddar Man was a hunter-gatherer, yet his DNA has survived into the present time. So there could not have been a virtually complete replacement of Paleolithic hunter-gatherers by Neolithic farmers.

Analysis of mtDNA from living Europeans by Sykes and Antonio Torroni (of the University of Rome) identified seven major mitochondrial haplogroups.

"Seven Daughters of Eve"

"Daughters"	Age	Origin	% of Modern Europeans
Ursula	45,000	Greece	11
Xenia	25,000	Southern Russia	6
Helena	20,000	Southern France	46
Velda	17,000	Northern Spain	5
Tara	17,000	Central Italy	9
Katrine	15,000	Northeastern Italy	6
Jasmine	10,000	Middle East	17

The fictitious names given by Bryan Sykes to the founders of the seven major European mitochondrial haplogroups are based on the alphabetic classification system of Antonio Torroni.

Sykes popularized the founders of these lineages as "the seven daughters of Eve." Six of these lineages, representing about 80% of Europeans alive today, are derived from Paleolithic stocks dating back before the advent of agriculture. Only about 20% of European haplotypes are young enough to represent the new genes of Neolithic farmers. This is not far different from the 28% of variance described by Cavalli-Sforza's first principal component.

mtDNA has thus shown an unbroken continuity of inheritance from the Paleolithic hunter-gatherers clear through to suit-clad urban humans alive today. Although agriculture did eventually reach England and the rest of Europe, it came largely on its own. The culture of farming diffused, rather than the farmers themselves. This will almost certainly hold true for the diffusion of agriculture from other ancient centers in Africa, Asia, and the Americas. "Modern" Neolithic farmers did not outpace the Paleolithic hunters, as they had replaced the Neandertals before them. The Paleolithic hunter-gatherers lacked nothing—genetically, physiologically, or behaviorally—that they needed to move into the modern age. These hunter-gatherers became farmers—and they became us.

FOR FURTHER READING

Carlson E.A. 2001. *The unfit: The history of a bad idea*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Cavalli-Sforza L.L., Menozzi P., and Piazza A. 1994. *The history and geography of human genes*. Princeton University Press, Princeton, New Jersey.

Gould S.J. 1985. Carrie Buck's daughter. In *The flamingo's smile: Reflections in natural history*, pp. 306–318. W.W. Norton, New York.

Jablonski N.G. and Chaplin G. 2002. Skin deep. *Sci. Am.* **287:** 74–81.

Kevles D.J. 1995. *In the name of eugenics: Genetics and the uses of human heredity*. Harvard University Press, Cambridge, Massachusetts.

Micklos D., ed. 1999. DNA from the beginning (http://www.dnaftb.org). Dolan DNA Learning Center, Cold Spring Harbor, New York.

Micklos D., ed. 2000. Genetic Origins (http://www.geneticorigins.org). Dolan DNA Learning Center, Cold Spring Harbor, New York.

Micklos D., ed. 2000. *Image archive on the American eugenics movement*. Dolan DNA Learning Center, Cold Spring Harbor, New York. At: http://www.eugenicsarchive.org

Micklos D., ed. 2002. Your Genes/Your Health (http://www.ygyh.org). Dolan DNA Learning Center, Cold Spring Harbor, New York.

Stanley S.M. 1996. *Children of the ice age: How a global catastrophe allowed humans to evolve*. Harmony Books, New York.

Sykes B. 2001. *The seven daughters of Eve: The science that reveals our genetic ancestry*. W.W. Norton, New York.