Genes, Peoples and Languages

The family tree relating human populations corresponds to another relating the languages of the world. Both trees imply a series of migrations; the biological evidence indicates a homeland in Africa

by Luigi Luca Cavalli-Sforza

More than 40 years ago, when I was studying bacterial genetics in the laboratory of Sir Ronald A. Fisher of the University of Cambridge, the place was saturated with mathematical theorizing. Thus, it is not surprising that I started thinking about a project so ambitious it seemed almost crazy: the reconstruction of where human populations originated and the paths by which they spread throughout the world. I reasoned that the task could be accomplished by measuring how closely living populations are related to one another and by deducing from this information a comprehensive family tree.

The goal is at hand. An exhaustive analysis of human genetic data gathered over the past 50 years, and of new data obtained with recently developed techniques, has enabled my colleagues and me to map the worldwide distribution of hundreds of genes. From this map, we have inferred the lines of descent of the populations of the world. Our tree agrees with another, smaller tree based on fundamentally different genetic data. Moreover, our reconstruction finds striking parallels in a recent classification of languages. Genes, peoples and languages have thus diverged in tandem, through a series of migrations that apparently began in Africa and spread through Asia to Europe, the New World and the Pacific.

The concept of a family tree is crucial to placing such events in their chronological sequence. If other factors are equal, the longer the time since two populations split apart, the greater the genetic difference—or distance—between them should be. Such an analysis could then be applied to a more complex history involving three or more populations.

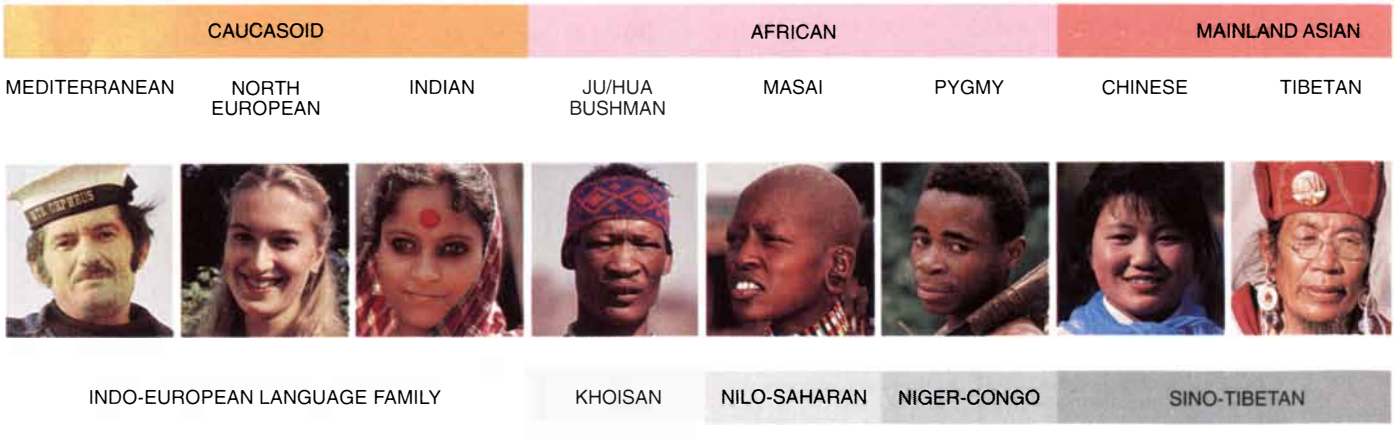
Human populations are sometimes known as ethnic groups, or "races," if one likes, although racist misuse of the term has made it rather odious. They are hard to define in a way that is both rigorous and useful because human beings group themselves in a bewildering array of sets, some of them overlapping, all of them in a state of flux. Languages, however, gave us a little help.

For much of its history, the human species was organized by tribes, or groups of fairly closely related people. Tribal affiliations continue to be of cardinal importance in traditional societies. In addition, there is often a one-to-one correspondence between language and tribe. Thus, languages offer a rough guide to tribes, and tribal membership, when available, provides a rough classification of populations.

Because the situation is far more complicated in metropolitan societies, we reduced our practical problems by focusing our study on aboriginal populations: those that occupied their present territories before the great migratory waves that followed the voyages of discovery in the Renaissance. Distances between these aboriginal groups cannot be abstracted from the presence or absence of a single inherited trait, or the gene that expresses it, because each group carries practically all the extant human genes. What does vary is the frequency with which the genes appear.

A good example is furnished by the vast set of data for the Rh factor, a human blood antigen that comes in two forms, positive and negative. The character is inherited in a simple way, and it has been studied in thousands of populations for reasons of public health. Physicians must identify pregnant women who are Rh negative and whose fetus is Rh positive and administer an immunologic treatment immediately after delivery. The treatment prevents the woman's body from making antibodies that might injure children she conceives later on. Rh-neg-

Ethnicity and Language



ative genes are frequent in Europe, infrequent in Africa and West Asia, and virtually absent in East Asia and among the aboriginal populations of America and Australia [*see map on next page*].

One can estimate degrees of relatedness by subtracting the percentage of Rh-negative individuals among, say, the English (16 percent) from that among the Basques (25 percent) to find a difference of nine percentage points. But between the English and the East Asians it comes to 16 points—a greater distance that perhaps implies a more ancient separation. There is thus nothing formidable in the concept of genetic distance.

In reality, geneticists use formulas slightly more complicated than simple subtraction so that distances can tell as much as possible about evolutionary history. Should the fragments of a single population become utterly isolated from one another, for example, they will differentiate even in the absence of mutations and natural selection [see "The Genetics of Human Populations," by L. L. Cavalli-Sforza; SCIENTIFIC AMERICAN, September 1974]. Chance alone causes their respective gene frequencies to change, in a process called drift.

When other matters are equal, genetic distance increases simply and regularly over time. The longer two populations are separated, the greater their genetic distance should be. Distance might therefore serve as a clock by which to date evolutionary history. Statistical considerations show, however, that one cannot expect a single gene like Rh to provide an accurate chronology. It is essential to use averages of many genes in the calculus of genetic distances and, ideally, to retest conclusions with different sets of genes. Fortunately, thousands of genes are known, although only a small fraction has been tested in many populations.

There are many principles on which one can reconstruct trees from genetic distances. An example is furnished by a tree linking 15 populations that Anthony W. F. Edwards, now at Cambridge, and I published 27 years ago. The genealogy derives from distances calculated from genetic information then available, according to Edwards's formula for the "minimum genetic path." Essentially, this concept describes the tree having the smallest total branch length. When the tree is projected onto a map of the world so that its branching points match the populations' current homelands, the resulting pattern roughly coincides with anthropological reconstructions of ancient migrations [*see top illustration on page 107*].

Unfortunately, there is no strong evidence that the minimum genetic path provides the best way of fitting a tree to the data. Other tree-building methods may be more satisfactory for relating the length of branches to the passage of time and finding a datable "root" for the tree [*see bottom illustration on page 107*]. When possible, a root relates the populations to an outgroup—say, to chimpanzees, which are believed to have diverged from the line leading to humans between five and seven million years ago. If one assumes that the rate of evolutionary change is constant along all branches, one can equate their lengths to the time elapsed since they diverged. Such rooted trees may also be subject to biases, however, if some branches have undergone more rapid evolutionary change than others.

Mathematical techniques of population genetics can minimize biases by accurately predicting rates of evolution. The evolutionary model we used is the simplest. It predicts that the branches will evolve equally fast, provided that drift has been the major cause of change and that the various populations have been the same size, on average. Independent evidence confirms the former assumption; a judicious selection of populations makes the latter quite probable. Constant rates of evolution are likely when populations are large and live in territories spanning continents and for periods stretching back to original settlement.

With my colleagues Paolo Menozzi and Alberto Piazza of the universities of Parma and Turin, respectively, I designed a common analytical framework to study the history and geography of human genes. In a 12-year project we studied the body of genetic information that has accumulated over the past 50 years—more than 100 different inherited traits from about 3,000 samples taken from 1,800 populations. Most samples included hundreds or thousands of individuals. This set of data, which we call the classical set, is derived indirectly from the proteins that the genes express.

In addition, we recently developed an entirely new second set: molecular data studied directly in the coded sequences of DNA carried in the cell nucleus. Most molecular data we used were gath-

LUIGI LUCA CAVALLI-SFORZA has been professor of genetics at Stanford University since 1971. Born in Genoa in 1922, he earned an M.D. from the University of Pavia in 1944. He studied bacterial genetics in Italy and, from 1948 to 1950, at the laboratory of Sir Ronald A. Fisher of the University of Cambridge. He switched to human population genetics in 1952. Since then, he has studied consanguinity; genetic drift and the means of predicting it through demographic observations; the reciprocal relations between biological and cultural evolution; the cultural significance of names and surnames; and the reconstruction of human evolution. He has conducted fieldwork among African Pygmies and applied molecular techniques for the analysis of genes and for the permanent storage of genetic material taken from aboriginal populations.

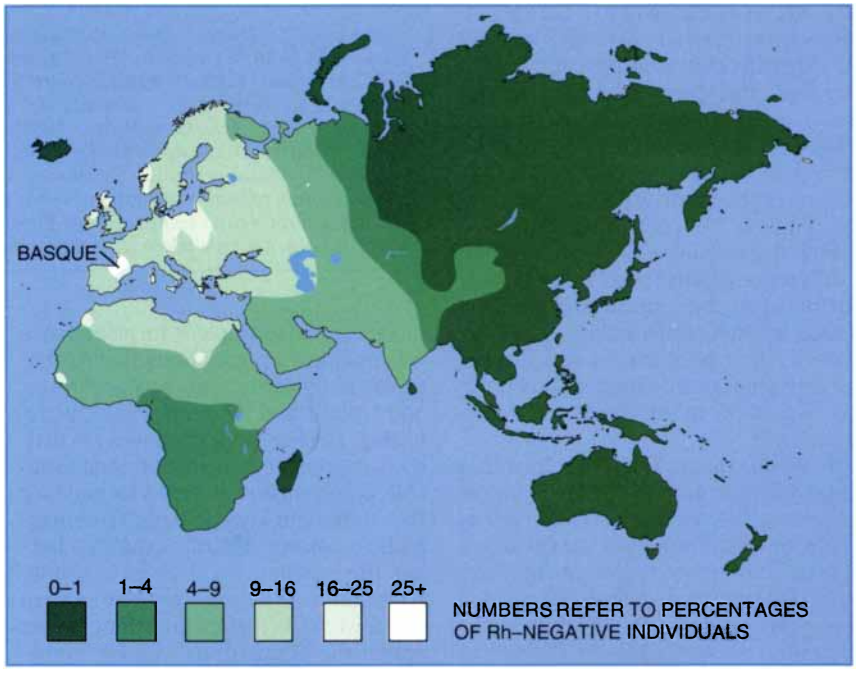
ered in a seven-year study undertaken in a collaborative effort between my laboratory at Stanford University and that of Kenneth K. and Judith R. Kidd of the genetics department at Yale University. Although such data are in many respects of higher quality than those based on gene products, so far they cover only one one-hundredth as many populations. In every comparison we have made thus far, however, the molecular data agree excellently with the classical data.

Our first result supports a conclusion that has emerged from studies of human physical and cultural remains: an African origin of our species. We found that the genetic distances between Africans and non-Africans exceed those found in other intercontinental comparisons. This result is exactly what one would expect if the African separation was the first and oldest in the human family tree.

The genetic distance between Africans and non-Africans is roughly twice that between Australians and Asians, and the latter is more than twice that between Europeans and Asians. The corresponding times of separation suggested by paleoanthropology are in similar ratios: 100,000 years for the separation between Africans and Asians, about 50,000 years for that between Asians and Australians, and 35,000 to 40,000 years for that between Asians and Europeans. In these cases, at least, our distances serve as a fair clock.

A different and quite elegant clock, it turns out, had been devised by other workers studying a kind of genetic data that differed fundamentally from our own. Their most interesting findings became available to us only when our analysis was nearing completion, but they confirmed our findings in all essential points. The set covers the relatively small number of genes encoded in the DNA of the mitochondria, cellular organelles that metabolize energy. We at Stanford had also initially studied such genes, but the late Allan C. Wilson and his colleagues at the University of California at Berkeley developed methodologies having higher resolution. (Here I can do but partial justice to Wilson's many contributions to molecular evolution. He died of acute leukemia in July, at the age of 56.)

Mitochondrial genes differ from those in the nucleus in fundamental ways. Nuclear genes derive about equal contributions from the father and the mother, but those in mitochondria are passed to offspring almost exclusively by the mother. This simple mode of inheritance makes mitochondrial genes very convenient for the estimation of genetic distances. They also have higher mutation rates than nuclear genes, so that one may in part alter the statistical determination of genetic distances, calculating them not from the frequencies of genes but from mutations in the genes themselves.



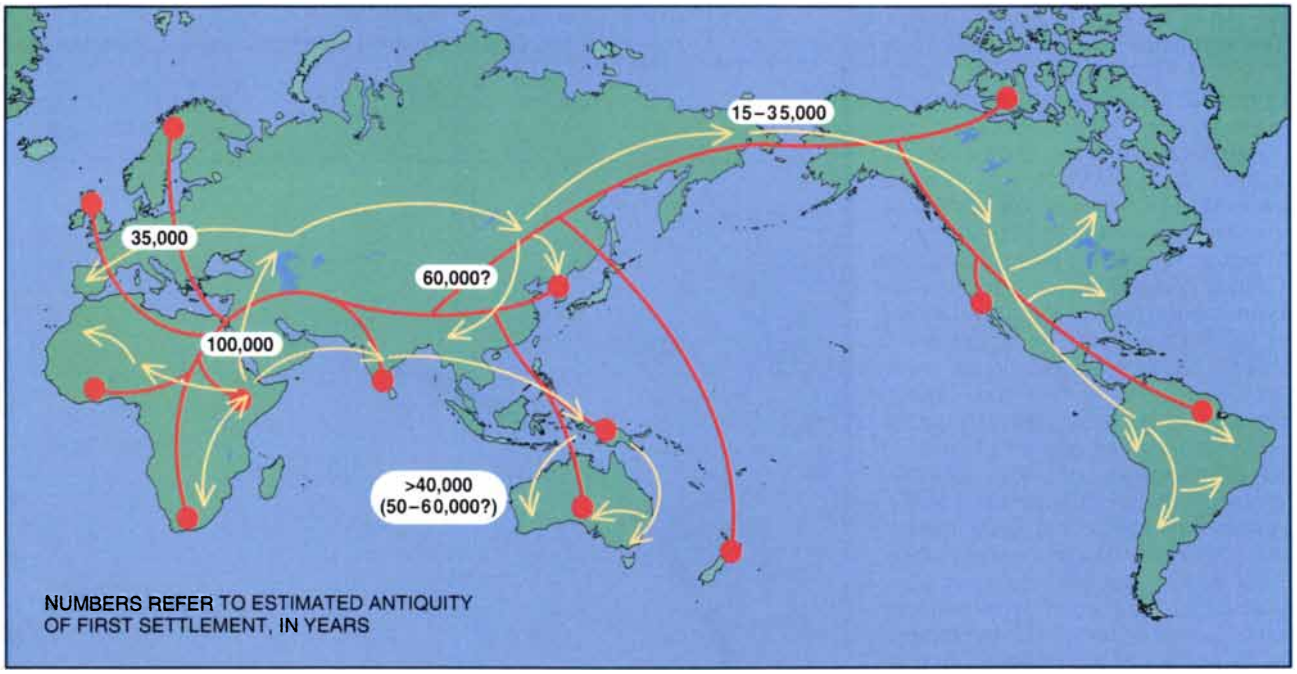
GENE MAP shows the Rh-negative factor to be most common among the Basques and less common further west. Such data suggest that the Basques preserve vestiges of an early European population that later mixed with newcomers from Asia.

The mitochondrial clock is based on the number of mutations that have accumulated rather than on the changes in gene frequencies. Whereas we hypothesized that our gene frequencies had drifted apart at constant rates, the Wilson group hypothesized that their mitochondrial genes had mutated at constant rates. Because of the nature of the data, it is easier to provide a root for a mitochondrial tree than for a nuclear one. One need merely compare one's tree with an external group—the Wilson group used chimpanzees—that is known to have diverged at a given date or range of dates.

From such distances, the Wilson group derived a tree of descent that showed more differentiation in Africa than anywhere else. That finding indicated that human mitochondrial DNA had been evolving for the longest time in Africa—that is, it can be traced to a single African woman. In addition, the workers were able to date the branching points of the tree by comparing DNA from humans and chimpanzees, whose lineages were known to have diverged about five million years ago. Their tree thus calibrated, the Wilson investigators were able to estimate the dates of its later branches. Most important, they were able to estimate that the African woman had lived 150,000 to 200,000 years ago. They have therefore confirmed our conclusions by completely independent means.

Recently the workers have brought their estimated date forward somewhat, but their African woman still precedes the date we assign to the divergence of African and Asian populations. In fact, she *should* be more ancient: the two dates refer to different events—the birth of an individual woman and the splitting apart of a population to which she belonged. The news media confused matters by giving wide circulation to the label "Eve" for this woman. In fact, we have no evidence that there ever was a time when only a single woman lived on the earth. Many other women might have lived at the same time, but their mitochondrial lineages simply went extinct.

Some of these conclusions remain controversial. Although paleoanthropologists agree that the genus *Homo* originated in Africa about 2.5 million years ago and that fossil evidence of anatomically modern *H. sapiens* appears only around 100,000 years ago, in Africa or near it, not all of them accept the "out of Africa" theory. One group maintains that modern humans emerged at a much earlier time and in many Old World populations all at once [see "The Emergence of Modern



GENES AND STONES tell similar stories. The earliest genetic tree (*red*) is projected onto a map; the ends have been placed in current homelands (*red dots*). More recent genetic work implies two routes of migration from Africa to Asia (*yellow*); details of the routes are speculative. Archaeological estimates of first-settlement dates appear beside migration routes.

Humans," by Christopher B. Stringer; SCIENTIFIC AMERICAN, December 1990].

Not only have we traced the earliest modern humans to Africa, we have also recovered evidence of a series of migratory waves. This pattern tells much about the origin of existing populations. Moreover, the interplay of our work with that of linguists and archaeologists promises to uncover still more detailed information.

In general, migratory processes reflect changes that may be viewed at once as pressures and opportunities. At many times, humans and their hominid progenitors have been able to increase in numbers greatly and thus to expand geographically. Such demographic success must have stemmed in the main from cultural developments, which for the prehistoric period must be inferred from the archaeological record. This record—bones and stone implements, for the most part—shows that Africa was indeed the original homeland of hominids. From there, migrations must have proceeded from Africa to Asia via the isthmus of Suez and, later, from Asia to Europe. These regions were settled by hominids by perhaps a million years ago.

The next stage is harder to recover because it depends on the time at which one imagines modern humans emerged from the hominid stock. In any case, it is clear that this emergence had already occurred when humans expanded from Asia to the Americas—an event that had to await a time when the Bering Strait was dry and the climate mild enough to permit passage by land. The settlement of Australia and the Pacific islands must also have been accomplished only recently, after the mastery of marine navigation.

Australia appears to have been settled by migrants from Southeast Asia at least 40,000 years ago and perhaps 10,000 to 20,000 years earlier than that. Archaeologists are divided, however, on the first entry into the Americas. So far the first fully convincing signs of humans in Alaska are dated to about 15,000 years ago. There seem to be ear-

CHANGE OVER TIME produces genetic differentiation, such as that reflected in this ethnic family tree (*left*). Drift, the mechanism of change, can be modeled by computer (*right*). When two halves of a population are first separated, they carry a gene at equal frequencies, but time and chance can eventually push them in opposite directions.

lier dates for sites in South America. The estimates therefore range from 15,000 to 35,000 years ago. Our nuclear genetic data suggest the settlement began around 30,000 years ago.

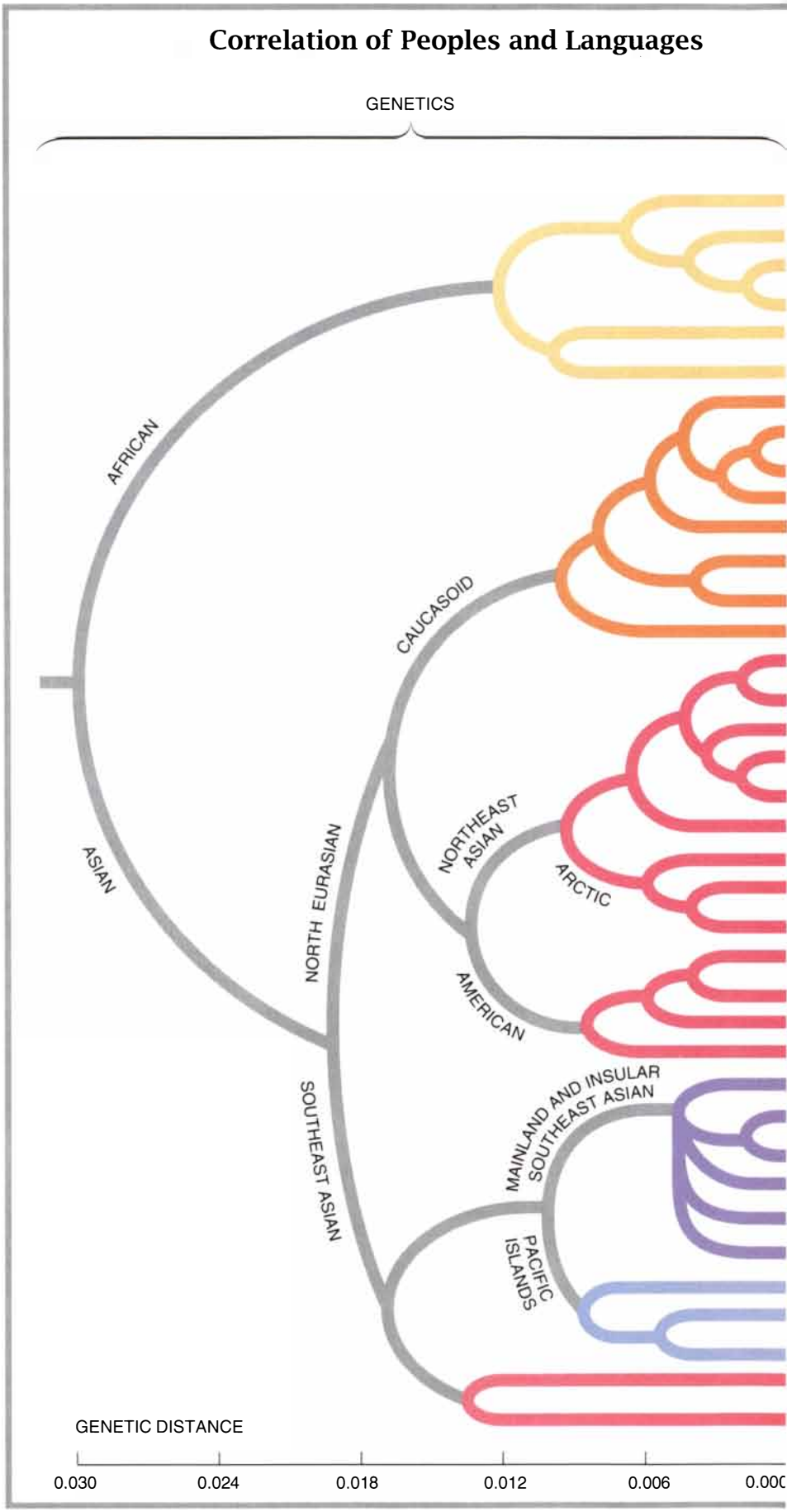
Europe has been swept by many migratory waves, but vestiges of the earliest settlement remain. A key to the puzzle was suggested in 1954 by Arthur E. Mourant, then at the Medical Research Council Population Genetics Laboratory in London, one of the first students of "gene geography." He hypothesized that the Basques (a population in northern Spain and southwestern France) were the oldest inhabitants of Europe and that they had conserved some of the pristine genetic constitution despite contacts with later immigrants. The theory is supported by data on the Rh-negative gene: the Basques have a higher frequency for the gene than any other population in the world. Work on many other genes confirms the hypothesis, as do studies of the profound linguistic differences between the Basque language and those of neighboring peoples.

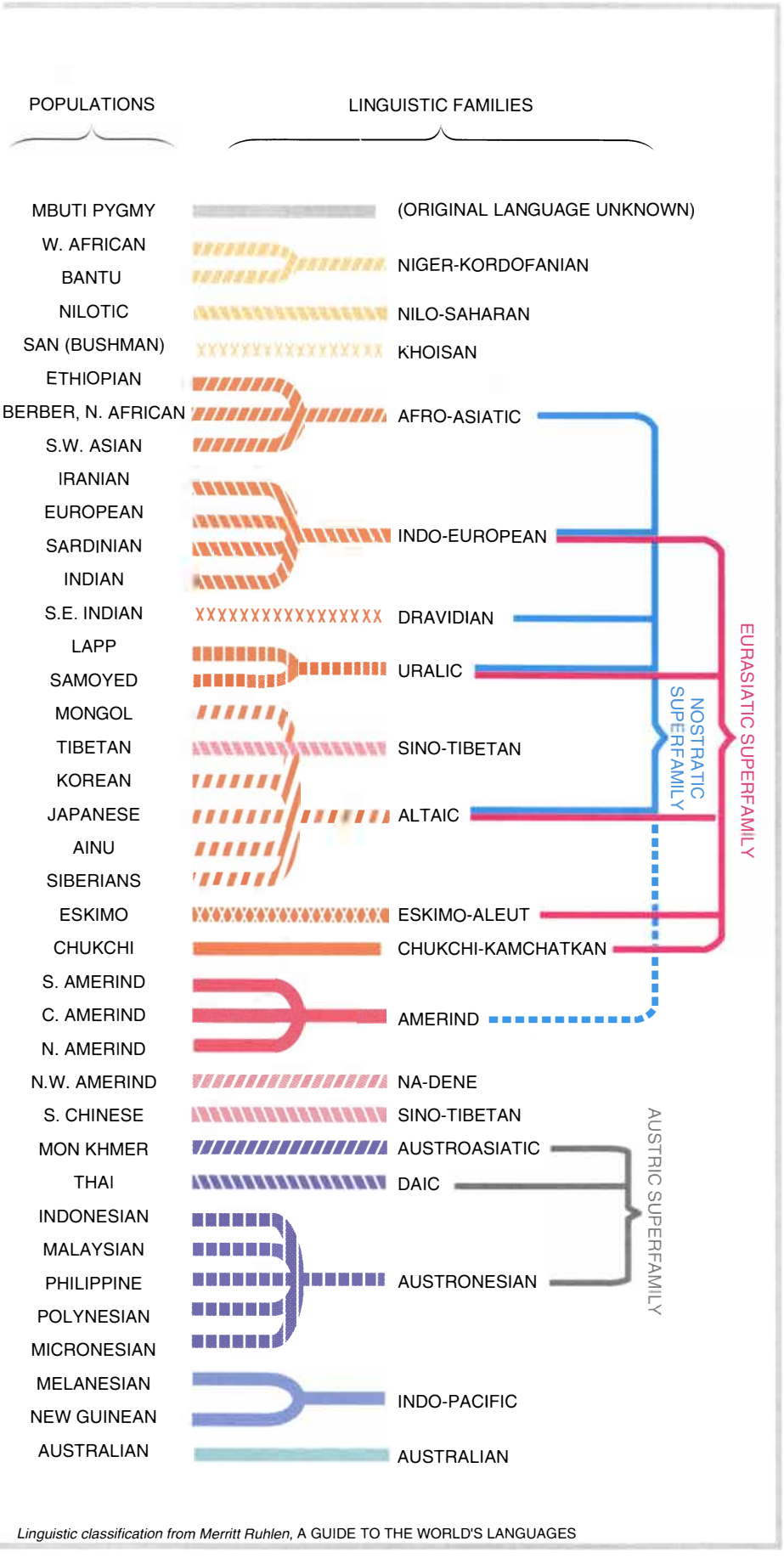
Indeed, a recent analysis of how genes vary from one part of Europe to the next suggested a model of how all of Europe was settled. This model holds that early Neolithic farmers brought their genes, culture and Indo-European languages from the Middle East to Europe in a process of slow expansion [see "The Origins of Indo-European Languages," by Colin Renfrew; SCIENTIFIC AMERICAN, October 1989]. Because the Basques' forebears lived at the far end of that migratory path, they presumably underwent the least genetic admixture with the farmers.

It should be noted that one can only hope to recover a pattern of settlement that reflects successful migrations. There might have been failures, too. In the Americas, to take a very recent example, the Vikings established short-lived settlements, but the genetic contribution they might have made to the local gene pool is unknown.

Our third major finding was that the distribution of genes correlates surprisingly well with that of languages. We concluded that, in certain cases, a language or family of languages can serve to identify a genetic population. A striking example is provided by the nearly 400 languages in the Bantu family of central and southern Africa, which are related to one another and correspond closely to tribal boundaries and the genetic affiliations among tribes. The reason this should be so had been advanced on linguistic grounds in the 1950s by Joseph H. Greenberg of Stanford.

Correlation of Peoples and Languages





Linguistic classification from Merritt Ruhlen, A GUIDE TO THE WORLD'S LANGUAGES

Greenberg's hypothesis, which has since won wide support, was that the Bantu tongues are descended from one tongue or a handful of closely related dialects spoken by early farmers in eastern Nigeria and Cameroon. As the farmers expanded into central and southern Africa beginning at least 3,000 years ago, their languages diverged but not so much as to obscure their common origin. Because the explanation applies to the genes of these populations, Bantu—originally a linguistic category—can now be extended to designate a group of populations having both a linguistic and a genetic basis.

In 1988 my colleagues and I published a genetic tree of evolutionary origins of 42 world populations, together with their respective linguistic affiliations. The tree demonstrates that the genetic clustering of world populations closely matches that of languages. With very few exceptions, the linguistic families seem to have a relatively recent origin in our genetic tree. Moreover, recent attempts by two groups of linguists to generate higher linguistic categories ("superfamilies") gave results that were also in line with the totally independent genetic evidence. It was exciting to discover that we had confirmed a conjecture made by no less a pioneer than Charles Darwin, who stated, in chapter 14 of *On the Origin of Species by Means of Natural Selection,* that if the tree of genetic evolution were known, it would enable scholars to predict that of linguistic evolution.

Why should genetic and linguistic evolution correspond so closely? The answer lies not in genetic determinism but in history: genes do not control language; rather the circumstances of birth determine the languages to which one is exposed. Linguistic differences may generate or reinforce genetic barriers between populations, but they are unlikely to be the leading cause of the correlation. Human evolution is punctuated by the splitting of populations into parts, some of which settle elsewhere. Each fragment evolves linguistic and genetic patterns that bear the marks of shared branching points. Hence, some correlation is inevitable.

One may object that complete separations, such as those established when a splinter group migrates to a new continent, must be rare. But one does not need oceans or mountain ranges to separate populations: simple distance will do the job, as genetic studies in many species prove. Because migratory interchange is normally greater at short distances, one expects and finds a higher degree of genetic difference the farther apart two subgroups are placed. It

is just the same for languages. When there are no special barriers, both genetic and linguistic variations tend to be continuous, and discontinuity tends to appear in both when there are some barriers to free intermigration.

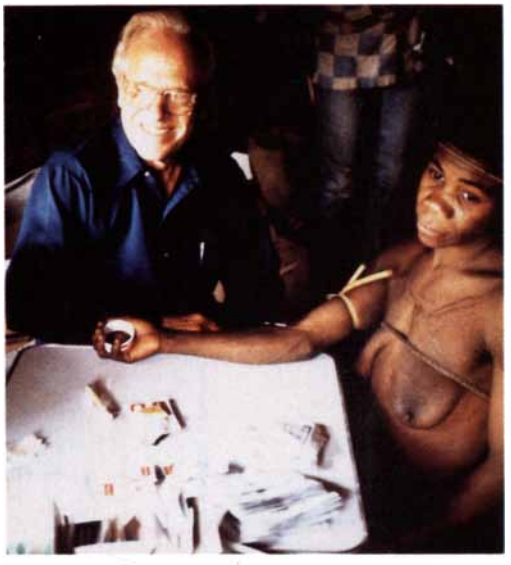
Two kinds of exceptions should be noted to our rule of the correspondence between genes and languages: those in which there is a replacement of language and those in which there is a replacement of genes. The former occurs when people give up their ancestral language for a new one, perhaps that of immigrants, conquerors or a newly risen cultural elite. Such replacements do not always occur, however, and they are less likely when the new language derives from a different family. Basque is an extreme case of a relic language that has evidently survived through thousands of years of continuous linguistic turnover in neighboring regions.

Gene replacement, usually partial, occurs when one population mixes with another. The mixing can be perfectly gradual, affecting the relative frequency of all genes in like proportion. This gradualism sharply distinguishes genes from languages, which in principle are either replaced or not. A language retains its ancestral integrity even if it adopts vast numbers of words from another linguistic family or subfamily. Linguists agree, for example, that English remains a member of the Germanic subfamily despite its borrowings from the French, the Greek and the Latin. What matters is that the structure and basic vocabulary retain family traits.

This difference means that when a tiny minority imposes its language on a conquered majority, language replacement is close to complete, but gene replacement is in proportion to the demographic ratio. Hungarians, for example, speak a language from the Urals (which divide Europe from Asia) imposed by the Magyar conquerors of the Middle Ages but carry a European genetic pattern. Only with some difficulty can one detect any traces of the Magyar genes in the modern population.

Large-scale gene replacement is perhaps rarer. But at least one likely example can be seen in our pair of complementary trees: the Lapps, or Saame, of northern Scandinavia. Their language also belongs to the Uralic family, but their genetic pattern suggests a mixture between Mongoloid peoples of Siberia and Scandinavians, who are responsible for the majority of their genes. Genetic admixture is evident also in the Lapps' hair and skin, which vary from extreme light to dark. A situation not unlike that of Lapps applies for Ethiopians, a genetic mixture of Africans and Caucasoids from Arabia, with a predominance of the former.

Even a modest trickle of genes can produce great effects if it continues long enough. A classic example is that of African Americans, who today derive on average 30 percent of their gene pool from people of European ancestry. This is the mixture that would have resulted had 5 percent of all black unions been with Europeans in each generation since the institution of American slavery and had all the progeny been classified as black. Another 1,000 years of such flow would leave but little of the original African genome.



GENETIC SAMPLE is taken from a member of the Aka tribe of African Pygmies by the author.

It is perhaps surprising that so much of the expected correlation between languages and genes remains, despite the blurring caused by gene or language replacements. In part, this may reflect our concentration on aboriginal populations. In any case, other analyses now confirm the existence of this correlation at a microgeographic level, sometimes in a dramatic way. Perhaps the most striking example is furnished by the close agreement between our analysis of genetic patterns in Native Americans and Greenberg's recent classification of New World languages into three major families. The two studies proceeded independently, using quite different kinds of data, yet each strongly implies that there were a handful of discrete migrations into the Americas.

The ultimate explanation of this correlation of genes and culture must be sought in the two mechanisms of transmission: horizontal and vertical. Genes, always transmitted from parents to children, describe a vertical path through the generations. Culture can also pass vertically from generation to generation, but unlike genes, it can also be transmitted horizontally, between unrelated individuals. High fashion, for example, is generally transmitted from Paris to the rest of the world every season (although Italy now appears to be taking the lead). In the modern world horizontal transmission is becoming increasingly important. But traditional societies are so called precisely because they retain their cultures—and usually their languages—from one generation to the next. Their predominantly vertical transmission of culture most probably makes them much more conservative.

Gene and language replacements are more than annoying exceptions to our rule. Each exception operates according to rules of its own, which should explain much about the evolution of populations and languages and hence of the development of human culture. Studies of such replacements could thus complement the work we have done. Anthropological fieldwork must catch up with such tools and with the rapidly vanishing data. Priceless evidence is slipping through our fingers as aboriginal populations lose their identity. Growing interest in the Human Genome Project may, however, stimulate workers to gather evidence of human genetic diversity before it disappears.

FURTHER READING

CULTURAL TRANSMISSION AND EVOLUTION: A QUANTITATIVE APPROACH. L. L. Cavalli-Sforza and Marc W. Feldman. Princeton Univeristy Press, 1981.

DRIFT, ADMIXTURE AND SELECTION IN HUMAN EVOLUTION: A STUDY WITH DNA POLYMORPHISMS. A. M. Bowcock, J. R. Kidd, J. L. Mountain, J. M. Hebert, L. Carotenuto, K. K. Kidd and L. L. Cavalli-Sforza in *Proceedings of the National Academy of Sciences,* Vol. 88, No. 3, pages 839–843; February 1, 1991.

RECONSTRUCTION OF HUMAN EVOLUTION: BRINGING TOGETHER GENETIC, ARCHAEOLOGICAL AND LINGUISTIC DATA. L. Cavalli-Sforza, A. Piazza, P. Menozzi and J. L. Mountain in *Proceedings of the National Academy of Sciences,* Vol. 85, No. 16, pages 6002–6006; August 1988.

HISTORY AND GEOGRAPHY OF HUMAN GENES. L. L. Cavalli-Sforza, P. Menozzi and A. Piazza (in press).