Human Evolution and Its Relevance for Genetic Epidemiology

Luigi Luca Cavalli-Sforza

Genetics Department, Stanford University Medical School, Stanford, California 94305-5120; email: cavalli@stanford.edu

Annu. Rev. Genomics Hum. Genet. 2007. 8:1–15

First published online as a Review in Advance on April 4, 2007.

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

This article's doi:
10.1146/annurev.genom.8.080706.092403

1527-8204/07/0922-0001$20.00

Key Words

population genetics, races, medical genetics

Abstract

The invitation to write the prefatory article to this volume of the *Annual Review of Genomics and Human Genetics* inspired me to collect some thoughts, a few involving ideas that are not new, but perhaps worth resurrecting in light of recent observations made with the data emerging from the Human Genome Diversity Project (HGDP). Data from the many relevant studies based on the HGDP have been made public, as was originally the hope and plan of the project. Here I try to give a short summary of the evolution of modern humans, a unique species in many respects but of special interest to readers of this volume, and a few thoughts on the general rates of evolution that might be relevant to medical genetics and genetic epidemiology. I have made no attempt to give a general bibliography, not even of results from the HGDP, since most authors' conclusions are still unpublished. Citations are limited to very few general concepts and articles discussed in this preface.

GENOMES AND EVOLUTION

Until recently, the study of entire genomes appeared a great but remote dream. It has now begun in earnest and the number of species whose full genome has become publicly available is increasing rapidly. Comparison among species as distant as bacteria and eukaryotes including humans is making it possible to reconstruct with increasing accuracy the steps that led from lost ancestral genomes to those that survive. Genome analysis is clearly giving very strong support to Lamarck's original supposition, widely made known at the beginning of the nineteenth century, that all living beings had a single origin, even if the exact nature of this original organism will have to wait for more research. Science has had to fight skepticism and fear during all of its existence, and thus it is not too surprising that some do not accept evolution, not even as a hypothesis, and try to replace it with religiously based alternatives, at a time when genomics is turning the theory of evolution into one of humanity's most powerful theories.

Doubt and discussion of observations and their interpretations are, of course, the basic stimulus of scientific progress, but one must admit that the very high proportion of Americans who do not accept the evidence in favor of evolution is humiliating and scary, especially considering they are the inhabitants of the country that has been the cradle of much of the latest technology and science. Of course, we are aware that this error is anchored in the belief based on the famous statement implicit in Genesis, that the Earth and life has existed no more than six thousand years or so. Like all innovations, religions have costs and benefits. It is probably no coincidence that Lamarck's publication of his book *Philosophie zoologique* took place after the French revolution had freed the intellectual world of some of the brakes that retarded the development of science. Obviously, credulity and taste for legends about human origins are widespread among all human populations, and are not confined to early ages and cultures. Especially when powered by political interests, belief in such fairy tales can continue well into adulthood.

One of the problems of communication by language is that it is highly ambiguous, and subject to change in time and space. Words are often used metaphorically, and this may add to the confusion. The major bar to accepting evolution arises from the initial statement in Genesis that the Earth was created in six days, which is taken literally by a few fundamentalist churches. There are basic units of measurements of all important quantities like time, length, area, volume, and weight, but their meaning is subject to considerable variation in time and place, as are the words used for them. For obvious reasons, time is measured on the basis of cyclical intervals, of which there are several, the day being the sharpest, while month and year are less clear-cut, depending on geographic location. The "day" of the week of creation in Genesis is probably a poetical metaphor for a very long time period, like scientific eras. After all, what is a day for a timeless God? In another part of the Bible the word used for the years for patriarch's ages seems to have meant months. The use of words meaning time in the Bible is not the only case in which linguistic ambiguity has created major conflict.

THE BASIC IMPORTANCE OF DEMOGRAPHY FOR UNDERSTANDING AND MEASURING EVOLUTION

My own research interests have been largely centered on evolution, and the organisms about which I can claim some knowledge are humans and bacteria. It is well known that Darwin's attention was drawn to demographic thinking by Malthus' emphasis on the simple consequences of exponential growth of living organisms. All self-reproducing entities like living beings go through phases of exponential growth that soon reach limiting conditions, imposing serious constraints. One

of the great advantages of the study of human organisms is that demographic analysis is easier in our species than any other. Indeed, three of the four major factors of evolution are described quantitatively in demographic terms. Natural selection is a matter of differences among inherited types in probability of survival to reproduction age, and fertility (the number of children, extended to a full generation cycle). Fisher's fundamental theorem (8) uses these quantities to calculate what he calls the Darwinian fitness of different hereditary types. Special cases, like frequency-dependent selection and inclusive fitness, demand modifications of the theorem. Genetic drift is also a matter of the number of reproducing individuals in a population and of variation in the number of children. It is independent of natural selection. Migration is another straightforward demographic quantity, although for genetics it is important to distinguish two types of migration (2). Type 1 is individual migration, that is (or at least was historically) usually limited to short ranges, and due to the displacement of at least one of the two spouses made necessary by marriage. It typically decreases the variation among populations generated by drift, or by natural selection in different environments. Type 2 is group migration leading to colonization of new areas, which may have opposite effects, being likely to generate new drift through what is called "founder effect" (the usually small numbers of the founders of colonies). Founding generates further opportunities for drift because of the usually small numbers of individuals in generations following that of the founders. It may also give rise to new patterns of genetic variation because of natural selection in the new environments reached by the migrants.

The fourth evolutionary factor, the source of all the evolutionary novelties, mutation in its umpteen forms, can also be described in demographic terms, although it is the one for which it is most difficult to get good estimates because it occurs rarely, and its precise measurement would demand greater numbers than we are usually prepared to handle. Genomic analysis, if extended to a large number of individuals and families, may provide precise measurements as the cost of testing decreases.

The kinetics of population changes under these four evolutionary factors is the main object of the mathematical theory of evolution developed especially by Fisher, Haldane, and Wright in the 1920s and 1930s. It rapidly became a stimulus for mathematical discoveries in stochastic processes, to which Fisher, Wright, and Kimura gave important contributions, and has continued to develop into a more and more complex mathematical theory. We are approaching a stage at which data will be available that allow these theories to produce estimates of the relative importance of these factors in a variety of real situations, and to further develop the general theory of evolution as well as of genetic epidemiology.

THE FREQUENCY OF POLYMORPHISMS AND HUMAN EVOLUTIONARY HISTORY

The analysis of individual genomic variation is clearly going to be one of the major next research steps, and it has already revealed a high frequency of polymorphisms. The main limitation is the cost of analysis of individual genomes. The recent availability of relatively cheap analysis has shown that many nucleotide sites of the order of millions are polymorphic in humans. Until the high cost of the full study of individual genomes comes down, we will have to be content with the polymorphisms now known, all of which are frequent enough that they could be detected on the basis of samples of a few hundred individuals. It will probably still remain true, however, that the most common polymorphisms involve single nucleotides, and those already known indicate that millions of nucleotide sites of our genome are polymorphic, roughly one per thousand of the nucleotide sites (3.14 billion) that form a human genome. The validity of current estimates of distributions of

polymorphisms is still rather approximate because of biases of ascertainment. Sites testable today have been found to be polymorphic on the basis of rather small samples of individuals, on which they were selected. The DNA chips now in existence test a somewhat arbitrarily preselected set of polymorphisms, and the criteria for choice are not always clear. With these limitations in mind, it remains exhilarating that we can now examine the polymorphisms at over a million nucleotide sites in the human genome. Naturally, when we are able to examine much larger numbers of individual genomes, the picture of polymorphisms may become much more complicated, but also much more instructive.

A remarkable finding is that the great majority of nucleotide polymorphisms are bi-allelic. A further constraint is that in mitochondrial DNA and Y chromosomes 90% or more of the sites are transitions, due to mutations A$\leftrightarrow$G, or T$\leftrightarrow$C, whereas in autosomes there may be a greater relative frequency of transversions. Bi-allelism is still rather frequent, whereas triallelism is relatively rare. This agrees with the simple hypothesis that transitions are more likely to occur than transversions as they involve less chemical difference, although the exact reason must involve a more precise biochemical explanation. The difference in the ratio of transitions to transversions between mtDNA and Y on one hand, and the autosomes on the other, is likely to reflect the differences between uni-parental and bi-parental transmission. DNA under uni-parental transmission has a fourfold higher evolutionary rate under drift, as the basic unit of drift: effective population size is four times smaller than for autosomes.

Another remarkable behavior of mutations that was first exemplified by the "African Eve" story is that all mitochondria of living humans descend from those of a single woman, not because there was a single woman in existence at a given time, but because mitochondria of all other women living at the same time were different, and distinguishable, from that of Eve, by one or more new mutations that occurred in Eve's mother (or close maternal ancestor that left no other descendants but Eve and Eve's descendants), and were not present in the other mtDNAs that existed at the time of Eve but are now extinct. In the famous paper of Allan Wilson's team (1), the "mitochondrial Eve"'s birth date was estimated roughly between 150,000 and 300,000 years ago. Neither the statement of Eve's time of origin nor her African origin had been rigorously proved at the time of their discovery, but much later research has given strong support to both, and the African origin of modern humankind has also received confirmation by archeology. The current estimate of the time of occurrence of Eve's distinguishing mutations is not far from the more recent extreme of the original estimate, 175,000, with a standard error around 10%.

The other uni-parentally transmitted chromosome, Y, has produced a (slightly more approximate) 50,000 years younger estimate. The difference between the birthdates of Adam and Eve is likely the consequence of polygyny, and in fact the ratio of the two dates is not far, as expected, from the ratio of the number of wives per man still observed in many parts of the world.

That all clones of these uni-parentally transmitted chromosomes existing today descend from one ancestral chromosome living some time ago, and all other clones that existed at the time of origin of the only ancestral surviving type are extinct, is a simple consequence of drift due to the variation of the number of children per individual. This expectation cannot be easily extrapolated to all other chromosomes, autosomes, and the X chromosome, or parts of them, because crossing over mixes up the DNAs of homologs and destroys the original genealogical sequence of nucleotides of the DNA strand. However, short DNA segments that had no crossing over still obey the law, except that it is not easy to distinguish the exact extremes of such segments. Most regular genes are too large to satisfy this requirement. When we are able to

discover the extremes of segments that have not undergone any crossing over since the single nucleotide mutations known in them originated, we may expect the time of origin of such short autosomal DNA segments to be four times as large (and three times for the X chromosome), on average, as that of uni-parentally transmitted, haploid chromosomes: i.e., 500,000 years (three times as large for the X chromosome).

The average time of these nonrecombined short DNA segments in bi-parentally transmitted chromosomes is likely older than 500,000 years on average, because an unknown, but, as we discuss later, potentially large proportion of them must be under selective heterozygous advantage. If this does not belong to a specific DNA segment, some heterozygous advantage will be borrowed from that of neighboring DNA: "associative overdominance through linkage disequilibrium" (13). In fact, in HLA there are polymorphisms that might be 15 million years old, and examples that have a longer evolutionary age have begun to appear in the literature. There is, of course, a large variation in individual evolutionary times under drift alone (10), but a two-million-years age would be a relatively rare occurrence with an expectation of 500,000 years, and there are other reasons for suspecting heterozygous advantage in many polymorphisms.

A few anthropologists still support an old anthropological hypothesis, called multiregional theory, that the current subdivision into four or five different races arose at least 1.7 million years ago (mya), at the time of the first expansion of *Homo erectus*, and that subsequent events have not erased the evidence of this ancient subdivision. These anthropologists hope to save the multiregional theory by showing that some of the genes currently polymorphic in modern humans originated long ago, away from Africa (a very difficult job), but this approach will not likely be of great help. It is perfectly possible, of course, that mutations originating in Asia that took place before the more recent expansion may have traveled back to Africa a long time ago, and there are already examples of them, although clear ones are clearly posterior to the latest African expansion.

HUMANS AS A COSMOPOLITAN INVASIVE SPECIES

Other complications arise in the application of population genetics to our evolutionary history because our species has gone through major bottlenecks in the total number of individuals, followed by later major expansions. They are mostly effects of major cultural innovations, permitted by the continuous increase of our capacities to communicate and to invent, that have enormously increased the power of cultural evolution compared with other animals, even our nearest cousins. This is reflected in the archeologically observed increase of the volume of our brain, which is now about four times the size of the ancestor common to our nearest Primate cousin, chimpanzees. By contrast, the brain of chimpanzees did not change much. The separation from chimps occurred at least 5–6 mya, somewhere in Africa, and a large number of species originated from that separation especially, if not only, in the human line, of which only one, ours, survives, and it is genetically very homogenous (*H. sapiens sapiens*). Therefore, it must have had a very recent evolutionary origin, as archeologists also show (approximately around 150,000 years ago, in East Africa).

The new species that originated in the first half of the interval of time separating us from chimps are classified into a genus different from ours, *Australopithecus* (so named because it was found predominantly in South Africa). It has been decreed that our genus, *Homo*, descended from it between 2.5 and 3 mya with the species called *habilis*, because it showed the first evidence of making stone tools, however rough. Around 1.7 mya tools had increased in number and complexity, allowing the species now called *H. erectus* to spread from Africa to the whole Old World (Eurasia). Very probably fire (first archeological evidence 1.6 mya)

also helped to generate this early expansion, being used for a variety of jobs, including defense against cold, and against wild animals, as well as for cooking food and making tools. Many species were formed at that time by allopatric speciation, but all Australopithecines and species of the genus Homo other than ours have been extinct for quite some time. There have been recently changes of these species names that are not entirely stabilized.

The next great expansion, this time to the whole world, was much later, and was started by a small population—genetic data indicate it may have been made of only 1000 individuals—living in East Africa. Genetic evolutionary time estimates indicate there may have been a slow beginning of growth and expansion perhaps 100,000 years ago, but there was a much more impressive, archeologically proved, successful beginning 50–60,000 years ago of an expansion to Asia and beyond, that continued until the whole world was settled. By that time, language was fully mature and was one of the key ingredients of a sophisticated communication, very useful for spreading and accumulating cultural knowledge. Language is certainly the chief difference between humans and animals. Tribes are basically (or were originally) social groups speaking the same language and sharing cultural identity. The language spoken by the East African "tribe" that started growing demographically, and continued to do so after it began to expand to the whole world, is most probably the one from which all 6000 languages spoken in the world originate. Linguists find difficulties in accepting statements that there was a single original language, from which all those existing today originated, because linguistic evolution is very fast and has produced profound differences between present languages. But important evidence in its favor is that any modern living human, apart from some rare genetically disabled individuals, can learn equally well any of the languages now in existence. Other major innovations that favored this expansion were inventions like navigation, that allowed far islands like Oceania and New Guinea to be reached over 40,000 years ago, and a new set of tools of increasing complexity and usefulness.

At about or shortly after 13,000 years ago the last great glaciation ended, and ice started retiring from Europe. It was the last of a series of glaciations that had started, with interruptions, some 80,000 years ago. It had transformed the world, and the climate change now favored different types of grass that affected the fauna, and also humans. The number of individuals on Earth is estimated to have been, at this time, between 1 and 15 million. This is, in terms of orders of magnitude, a thousand times less than now, but a thousand times larger than before the expansion. The demographic expansion that had started with modern humans and caused a great geographic expansion gave rise to a number of new innovations, such as the bow and arrow, the use of bamboo where it grew (in East Asia), the use of clothes in cold climate. Especially at the end of the last glaciation there were also improvements in stone tools, which became smaller and more varied (called mesolithic, especially in Europe), but the economy was still one of hunting and gathering (and fishing) or, as it is called, in short, foraging, and population density was growing in several more favorable areas, with temperate climates. In these areas archeologists have noted the apparently independent starts of a major innovation, food production: agriculture, and animal breeding, replacing food collection from nature. Both are obviously applications of the discovery of major secrets of life: its modes of reproduction, and the possibility of genetically modifying plants and animals to suit our needs.

Food production began independently in various areas, at least in the sense that the plants and animals that were domesticated were the local ones already used as food by hunters and gatherers (fish was domesticated only very recently). The major areas were the Middle East and Turkey: the oldest, which is near the border between Turkey and Syria, was an already mixed agro-pastoral site

(Abu Hureyra, 11,500 years ago). Domestications of local plants and animals began somewhat later, in China (independently in north and south), in the Sahara, which was not dry at the time, in New Guinea, in the Mexican plateau, and in West Africa. From the places of origin, food production spread slowly at rates of 1–2 km per year, both carried by farmers that looked for new fields at the periphery of the expansion, and imitated by the local hunter-gatherers, who learned from their neighbors and sometimes intermarried with them.

Agriculture changed the diet considerably. The first crops developed in very different parts of the world (wheat, maize, rice) are still today the most important in the whole world. Agriculture also increased sedentariness and decreased the nomadism associated with foraging, but a later development of relatively pure pastoralism in semiarid areas encouraged some special types of nomadism.

Sedentariness and increase of population density caused wide social stratification, job differentiation, the rise of governments, of chances for big property and power that were essentially impossible with foraging. Still later, new expansions were favored by new inventions: metals. The first was bronze (around 5000 years ago) and some 1500 years later iron developed in a similar area, mostly at the boundary between Europe and Asia above the Caucasus. In the same area the horse was domesticated. Together, metals and the horse generated the beginnings of major wars between neighboring governments, and men, originally hunters, now became warriors. Large numbers of war prisoners were turned into slaves. It became possible for large tribes to resettle in new favorable areas, their families traveling by cattle- or horse-driven carts. Horse-mounted Mongols began major conquests in Asia. The camel was domesticated later. With horse and camel, Arabs rapidly conquered Turkey and North Africa, Sicily, and Spain, beginning in the seventh century AD. Arab merchants also spread to East and South Africa. Some single inventions and innovations proved extremely powerful, and their influence on our life, including our genetics, keeps increasing at a growing rate. Cultural history has thus begun to dominate the genetic picture of our species, and to affect the fate of many other species as a consequence of our actions on the environment [comparisons of archeological and genetic observations of modern human evolution appear in Cavalli-Sforza & Feldman (5) and in Cavalli-Sforza et al. (6, 7)].

THE CONTRIBUTIONS OF MUTATION AND SELECTION TO THE POLYMORPHIC PICTURE

Of the four factors of evolution, mutation is responsible for the origin of polymorphisms, and natural selection for the change of their frequencies in response to the environment. In diploid species selection in the form of heterosis increases the stability of polymorphisms. By contrast, drift tends both in the short term and in the long run to destroy polymorphisms, and to differentiate populations one from the other, whereas individual migration mostly tends to redistribute the existing polymorphisms within the species. Kimura (10) has shown that most mutations are selectively neutral in molecular evolution, so that mutation and drift are the main causes of most differences among species. But what about the millions of polymorphisms that show no major frequency differences in our species, at least between the four populations so far tested by the (National Institutes of Health (NIH) Genome Institute) International HapMap Project? There are, of course, some genes, or more precisely DNA segments, that show strong differences among these populations, signaling differential selection that extends to neighboring short DNA lengths because of linkage disequilibrium, but the great majority of polymorphisms show few differences among the HapMap populations, in agreement with the overall homogeneity

across our species compared with most other mammals.

A question of interest is: Why are there so many polymorphisms? The following situations come to mind:

1) The polymorphisms were already in the species before its origin, under differential selection that drove the two alleles to their present relative frequency, but may have ceased to operate, so that they are now neutral and fluctuate under drift alone, in equilibrium with migration.
2) Polymorphisms were driven to their present average frequency by linkage with neighboring genes that are or have been under selection, and the present differences among populations are due to drift, in equilibrium with migration.
3) Polymorphic frequencies may be the result of equilibrium between opposing mutation rates in the absence of other pressures, but gene frequency changes under mutation pressure alone are extremely slow. Moreover, the bi-allelism of almost all polymorphic nucleotide sites shows that we are far from an equilibrium of mutation rates. We would expect that at many sites more than two nucleotides would be present at equilibrium under mutation pressures. In fact, the time it takes for equilibrium to be established under mutation at rate μ, in a very large population, is of the order of $1/\mu$ generations. We have little knowledge of mutation rates, and mostly for very few genes; estimates of average values of mutation rates are mostly derived from evolutionary analysis, but we also know that some individual mutation rates may be very different from the average. The order of magnitude of changes due to mutation rates alone would demand millions of generations. There are also major bars to the observation of equilibria determined by neutral mutation rates alone: They are accidents causing strong drift by founder episodes, like those that generate new species, and differential selection effects in different environments.
4) Selective advantage of heterozygotes will generate stable polymorphic equilibria. Even if one of the alleles is lethal or almost so, heterozygote frequencies near 20% or 30% have been observed at equilibrium, for instance for sickle cell anemia. Heterozygous advantage is also spread to neighboring, closely linked neutral genes. The rate at which such polymorphic equilibria are reached is much higher than under mutation pressure alone. In the first example of heterotic equilibrium studied, sickle cell anemia, the selection coefficient involved was of the order of 10%; that is, the fitness of the normal homozygote was 10% smaller than that of the heterozygote, and was enough to give a 20% frequency of heterozygotes at equilibrium, even though the other homozygote was basically lethal. We have few other heterotic examples as clear as that of sickle cell anemia (or thalassemia, which is very similar). With selection of this strength about 100 generations are necessary to take a population of 10,000 individuals from the first appearance of a mutation to near equilibrium. In the simplest conditions, the time in generations necessary for replacement by a new mutant is proportional to the reciprocal of the selection coefficient. Thus, in the course of the latest modern human expansion (50,000 years is ca. 2000 generations), a new mutation that appeared at the beginning and showed heterozygous advantage might be close to equilibrium today, even with selection coefficients as small as 0.5%. If the equilibrium due to heterozygous advantage existed before the start of the expansion, it may have been due to smaller selection coefficients, even as small as 1/10,000, that would of course take a proportionately longer time to establish [a simple numerical

introduction to this type of analysis is found in Cavalli-Sforza & Bodmer (4)].

The safest way to test for the existence of heterozygous advantage is the demonstration of deviation from Hardy Weinberg equilibrium among adults in favor of heterozygotes, but it took of the order of a thousand individuals to prove this in the case of sickle cell anemia. With smaller selection coefficients the number of individuals required is prohibitive, given the size of samples available today. However, an analysis of 91 genes studied at the genome level, in which selection had major effects, showed that 22, i.e. 24%, of them included heterosis (17). It is likely that heterosis involves an even higher fraction of polymorphisms, considering that the stabilization of a polymorphism also extends to neighboring neutral genes because of linkage disequilibrium. Naturally, this stabilizing effect due to linkage disequilibrium with selected variants is smaller the weaker the linkage (13). It is possible that the similarity of frequencies of neighboring polymorphic sites shown by HapMap populations is due to a near ubiquity of heterosis, including heterosis borrowed from neighboring sites by linkage disequilibrium, but a real test will demand adequate numbers of individuals.

DEMOGRAPHIC BOTTLENECKS FOLLOWED BY DEMOGRAPHIC GROWTH

Particularly important in establishing the genetic picture of a population are demographic bottlenecks, especially if they are followed by major demographic growth. We know that our species had such a history at the time of its origin in the last 100,000 years, with spread limited almost only to Africa in the first 50,000 years of this period, and followed by expansion to the whole world in the second half. Later, there were some major expansion events at the time of the beginning of agriculture, beginning around 10,000 years ago. These generated more local spreadings, which were at least partially responsible for the major four or five clusters that correspond approximately to continents (see below), but there were many further later ones that were much more limited geographically. Transoceanic navigation in post-Columbian times led particularly to the spread of Europeans, but this cultural innovation was also responsible for the spread of Indians and of Chinese to many economically attractive parts of the world.

Social customs have tended to keep most societies fairly endogamous after their migrations away from the mother country, although there have also been examples of partial melting pots, especially in the United States and even more in Brazil. But in Europe and elsewhere, groups of colonizers kept some of their original cultural identity, including their language, for periods of remarkable length. There are also examples of extraordinary conservation of languages. The Basque language may have descended, with considerable transformation, from a family of languages that were spread to Eurasia and to North America more than 20,000 years ago, perhaps with the first arrivals from Africa. However, most language families spread more recently with agricultural developments, i.e., in postglacial times, and largely replaced earlier linguistic family groups.

Still more recently there was a multiplication of minor expansions, beginning with the metal ages and later. The Jewish diasporas brought Jews to various continents 2500 and 1900 years ago, and again in more recent times, and Jews grew in numbers almost everywhere. Less than 2000 founders, mostly Dutch, joined by some German and French protestants settled in Capetown, grew by a factor of almost 1000-fold in 350 years, and now form about half of the white population of South Africa. French Canadians of Quebec originated at about the same time, descendants of approximately 1000 French women who were invited by King Louis XIV to become "Daughters of the King," receiving a dowry if they accepted to marry French

trappers who were active in Quebec. Here also a 1000-fold increase took place in 350 years. These populations have all been well studied genetically and examples of genetic rarities or diseases present in one or a few founders, sometimes identified from the genealogies, are now relatively frequent in these and similar populations.

A word of caution: one may wonder if some numerical values related to evolutionary times, including many evolutionary time values estimated by simulation of evolutionary processes, are entirely correct, because they are usually calculated assuming exponential population growth. It is very likely that the population growth of our species during its geographic expansion from East Africa was far from exponential, and rather linear with time, or at most quadratic. The best model for demographic-geographic expansions (we use the short "demic" for their combination) is most likely to follow the extension to population expansions of the model introduced by Fisher (9) for advantageous mutations, called "the wave of advance," which shows that expansion under constant population density results in a constant rate of population growth, i.e., is linear and not exponential in both time and space.

THE SERIAL FOUNDER EFFECT IN THE EXPANSION OF MODERN HUMANS

The examination by Rosenberg et al. (16) of the set of 1064 human DNA samples provided by the Human Genome Diversity Project (HGDP) with ca. 400 microsatellite loci generated a cascade of further research. Prugnolle et al. (14) showed that the average heterozygosity of the 52 populations decreased in linear fashion by about 16% as a function of geographic distance from the place of origin. We confirmed their result with twice as many microsatellites, using 400 additional microsatellites that were made public later, and offered (15) an interpretation by a simulation that predicted well the observed fall of genetic variation, measured by heterozygosity. The simulation used a model of "serial founder effect" that can be summarized as follows: The original East African population grew from an initial small size to saturation density, and then generated a colony at some distance, which also grew to saturation. The first colony then generated a second colony that moved further away; it grew and the process continued until the furthest point inhabited on Earth was reached. The origin of each colony was thus an episode of founder effect that was repeated serially at every new colonization. The simulation showed a linear fall of average heterozygosity at a rate that corresponded closely with that observed for the microsatellite data, using numbers for tribes and growth models that were anthropologically reasonable. According to this model, 100 steps (each step being a successive colonization event) could, in 40,000 years, cover the 25,000 kms separating the origin, put arbitrarily in Addis Ababa, from the terminal colony in the southern tip of South America. We allowed for obligatory waypoints that avoided the direct crossing of oceans. A very similar exercise by Liu et al. (12) suggested 300 steps.

These models can be improved, but at least they show that one can explain the remarkable linearity of the process that was observed, on time and space scales compatible with the real conditions. A simulation using Fisher's wave of advance model (9) also fits the observations and is probably preferable in that it would correspond more closely to anthropological reality, but would demand, for a detailed fit, more specific anthropological knowledge than available for hunter-gatherers' societies.

THE CORRELATION OF GEOGRAPHIC AND GENETIC DISTANCES

The Ramachandran et al. analysis (15) of the Rosenberg et al. data (16) also estimated the correlation between genetic and geographic distances calculated for all the

possible pairs of the HGDP populations. It showed an extremely high linear correlation between the two distances ($r = 0.89$), one of the highest correlations ever observed in biology. It should be emphasized that the 52 HGDP populations are all indigenous, i.e., are today located in the approximate places where they were before the great migrations that took place in post-Columbian times (so that in America the collection includes only Native Americans, etc.). It is therefore obvious that all populations belonging to the same continent are more similar to each other than those from different continents, but this description, which is characteristic of many simple race models, is not sufficient. In fact, populations from different continents that are geographically close are also more genetically similar than expected on the simple hypothesis that they are part of their respective continents, and geographic distance between them is unimportant. Thus, for instance, Arabs and Ethiopians, Spaniards and Moroccans, and Turks and Greeks can be expected to be more similar to each other than two random African or two European or two Asian populations. The few comparisons that deviate from the linearity of the correlation of geographic and genetic distance involve some African populations that are particularly ancient in human evolutionary history, and have had more time to diverge genetically. Also slightly out of place in the linear correlation are populations that are known to be partial admixtures of very different groups, as is also expected.

The simplest interpretation of these results, also taking account of the serial founder effect, is that the human genetic differentiation observed for microsatellites is to a very large extent due to the equilibrium of drift and migration. Major effects of natural selection are not detected, but preliminary observations with other types of mutants or by other approaches also show evidence for natural selection of the directional type, as demonstrated by major differences among different HapMap populations. This summary can cover only the major aspects of these findings, but the original paper shows that drift can account for at least 78% of the variation observed for microsatellites. However, it leaves uncertain what role, if any, heterosis had in this process.

GENETIC VARIATION BETWEEN AND WITHIN POPULATIONS, AND THE RACE PROBLEM

In the early 1980s, Lewontin (11) showed that when genetic variation for protein markers is estimated by comparing two or more random individuals from the same populations, or two or more individuals from the whole world, the former is 85% as large as the latter. This means that the variation between populations is the residual 15%, and hence relatively trivial. Later research carried out on a limited number of populations and mostly, though not only, on protein markers has confirmed this analysis. The Rosenberg et al. data actually bring down Lewontin's estimate to 5%, or even less. Therefore, the variation between populations is even smaller than the original 15%, and we also know that the exact value depends on the choice of populations and markers. But the between-population variation, even if it is very small is certainly enough to reconstruct the genetic history of populations—that is their evolution—but is it enough for distinguishing races in some useful way? The comparison with other mammals shows that humans are almost at the lower extreme of the scale of between-population variation. Even so, subtle statistical methods let us assign individuals to the populations of origin, even distinguishing populations from the same continent, if we use enough genetic markers. But is this enough for distinguishing races? Darwin already had an answer. He gave two reasons for doubting the usefulness of races: (1) most characters show a clear geographic continuity, and (2) taxonomists generated a great variety of race classifications. Darwin lists the numbers of races estimated by his contemporaries, which varied from 2 to 63 races.

Rosenberg et al. (16 and later work) analyzed the relative statistical power of the most efficient subdivisions of the data with a number of clusters varying from 2 to 6, and showed that five clusters have a reasonable statistical power. Note that this result is certainly influenced by the populations chosen for the analysis. The five clusters are not very different from those of a few partitions that had already existed in the literature for some time, and the clusters are: (*a*) a sub-Saharan African cluster, (*b*) North Africa–Europe plus a part of western Asia that is approximately bounded eastward by the central Asian desert and mountains, (*c*) the eastern rest of Asia, (*d*) Oceania, and (*e*) the Americas. But what good is this partition? The Ramachandran et al. (15) analysis of the same data provides a very close prediction of the genetic differences between the same populations by the simplest geographic tool: the geographic distance between the two populations, and two populations from the same continent are on average geographically closer than two from different ones. However, the Rosenberg et al. analysis (16) adds the important conclusion that the standard classification into classical continents must be modified to replace continental boundaries with the real geographic barriers: major oceans, or deserts like the Sahara, or other deserts and major mountains like those of central Asia. These barriers have certainly decreased, but they have not entirely suppressed genetic exchanges across them. Thus, the Rosenberg et al. analysis confirms a pattern of variation based on pseudocontinents that does not eliminate the basic geographic continuity of genetic variation. In fact, the extension by Ramachandran et al. of the original Rosenberg et al. analysis showed that populations that are geographically close have an overwhelming genetic similarity, well beyond that suggested by continental or pseudocontinental partitions. Genetic variation is thus very nearly continuous, as Darwin said, but behind this continuity is another discontinuity that has not yet been analyzed. It is due to barriers among small social groups, based on religious separations and socioeconomic stratifications. These also create differences in the environment of individuals and small groups, through diet and economic or other, self-imposed limitations. This analysis suggests that it is important to recognize and study groups much smaller than the traditional major "races." To clarify what groups we want to distinguish, we first have to very clearly answer the question: What purpose do we have in mind when we try to distinguish races?

The number of groups to be distinguished depends on the differences among populations that are really useful for some valid purpose, and we still have not decided on criteria to choose populations that we want to distinguish. Therefore, the real question remains: What do we want races for? Let us start by agreeing on what could be the most important reason for defining "races." Incidentally, I am inclined to dismiss the word "race" because of its connection with the odious episodes of racism with which we are continuously confronted. The word "populations" is useful in statistics for defining the group from which we draw samples, but is in practice used arbitrarily, and perhaps the most neutral term could be that used by Rosenberg et al. (16): "clusters." We want to define useful genetic clusters. But more than a term to be used, what I am looking for here is a general agreement on a good reason for doing research aiming to define a useful genetic stratification of populations, and it seems to me we can really find it in research that can be of help for medicine, that is for diagnosis and therapy.

The expression "ethnic groups" is also useful, but especially in situations in which it is not clear if the basic difference is of genetic or cultural (including socioeconomic) nature, or both.

SHALL WE TAILOR DIAGNOSIS AND THERAPY TO THE INDIVIDUAL GENOME?

Today it might also be objected to that the long-term answer is actually another one. We

do not need any types of races at all, even for medicine. The individual genome is going to be the best information for diagnosis and therapy, especially considering the strong individual variation in response to many drugs. But how long will we need to wait before this is true? The most important diseases (the "complex" diseases) do not behave, except in rare cases, in a simple Mendelian way. They are complex in two ways. One is that many of them are likely polygenic—that is, there may be a great variety of genes that contribute to the disease. Specific Mendelian diseases are often caused by different genes and often different variants within each gene causing different patterns, and in polygenic diseases almost every individual case or pedigree might have a somewhat different polygenic system of its own. The other way is that the environment causes additional variation, and the individual environmental effects are usually poorly known.

It is clear that genomic knowledge will be very useful for rapidly extending our knowledge of Mendelian diseases, and of individual responses to drugs that may be available for a specific disease, but when we come to the most difficult diseases of all, the complex diseases, the environment is also very important. It is clear that we must be prepared to deal with the importance of environmental stratification within the genetic cluster, much of which may be due to socioeconomic, rather than genetic, differences.

We obviously hope that the genome will also be very useful for understanding more of the genetic background of complex diseases (provided environmental contributions are accurately studied at the same time). Here, however, we may have to wait for an entirely new development beyond the genome formed by the DNA of germinal cells, which accounts for inheritance. Information of what happens in somatic cells, i.e., epigenetics, will also be necessary. Thus, for genome knowledge to give a response to all or most individual queries much work remains to be done, and perhaps the only sensible answer to the question "how long will it take?" is that we can hope that there will be important progress in this century.

Even so, for many of these investigations it may still be necessary or useful for a while to study genetic clusters of individuals to solve specific problems, but they will have to be smaller populations than simply one per continent or pseudocontinent. Genetic epidemiology shows that many Mendelian diseases are concentrated in some, usually small, social or ethnic groups, especially for the rarer diseases. The reason is simple for recessive diseases: The recent demographic expansions that have occurred in so many parts of the world in the past few centuries or millennia have created clusters of cases of specific genetic diseases that originated from a single or a few mutations that occurred during the demographic growth of specific populations. These social groups are still meaningful today (for example, the demographic bottlenecks followed by major episodes of demographic growths, mentioned above). It is these genetic clusters that will likely be very useful, perhaps also for studying complex diseases, but the conclusions may remain valid only or mostly for the clusters in which they were obtained.

The number of cases of a well characterized disease may give a rough indication of the time of origin of the relevant mutations and populations. But some diseases can be caused by a great variety of mutations in the same gene, and then there is often considerable clinical variety associated with the various mutations, as is clear, for instance, with cystic fibrosis. The size and importance of the gene contribute to this complexity. The demographic history of a population may help to predict which groups are likely to show more genetic similarity in their disease patterns. The remarkable differences in genetic pathologies found among Ashkenazim and Sephardim (Jews of northern-European and of Mediterranean origin) populations indicate that genetic clusters of individuals of medical interest may have to be small to be really useful; even a relatively small group like Jews has

to be split into subclusters for medical genetics research. If the size of these groups should be an example of the size of useful genetic clusters from a medical point of view, then one would need of the order of a thousand genetic clusters for the whole species.

It is probably legitimate to state that, as genetic predisposition to many diseases is important, even for the response to some endemic infections, recognizing and studying useful genetic clusters may help us toward the aims of diagnosis and therapy. In practice, we may be going in the direction of studying smaller and smaller clusters and end up with the individuals, but the use of information from an individual's genome may need many more years to become truly adequate. Until then, it seems to me we need to enlarge and improve on collections of the HGDP type, my next subject, and also try new, different approaches.

THE HGDP PROJECT AND ITS FUTURE

Nature Reviews Genetics published a short history of HGDP-Centre d'Etude du Polymorphisme Humain (CEPH) (3). The present collection is made of 1064 cultures of B-lymphocytes from 52 populations that were already in existence. All continents except Australia are represented, but inevitably not in the way that was originally desired, with populations spread at fairly regular distances. There is some geographic clustering of the populations because some countries (China, Pakistan, Israel) have been especially active and generous in providing cell lines. The project was made possible by the kindness of research workers who donated cell lines they had collected, and by the fact that CEPH already had all the necessary equipment for growing cell lines, producing DNA, and preparing vials for distribution to laboratories, and made it available for the HGDP. The equipment had been assembled earlier for producing the human linkage maps that used the Utah and a few other pedigrees for linkage studies. Thus, HGDP-CEPH required very little money.

The original idea of the HGDP project was to collect 10,000 cell lines of 25 individuals from each of 400 populations. The Genographic Project is planning to collect 100 individuals from each of 1000 populations but just DNA from saliva and, in a small fraction of cases, blood samples. Unfortunately, it will not generate cell lines. With present technology, the analysis of 1000 individuals is still somewhat demanding for a single laboratory unless it is limited to a gene or to a sample of special markers. There is an inevitable trade-off between the number of individuals and that of markers that will be examined, mostly for economic reasons. However, we are now entering a new era, in which we can become more ambitious, but also need greater capitals.

Cultures of B lymphocytes are still the best material to collect and store. DNA will remain for some time the most important product of genome collections, but cells also offer important avenues of research into epigenetics. Their RNA and proteins can give information on the epigenetic processes, and on epigenetic evolution taking place during the history of the individual, at least for all genes relating to general maintenance and reproduction of the cell, and for specific immunological functions that occur in the B lymphocyte. Therefore, they are also informative on the immunologically significant environment of the populations.

It is important to maintain and extend the present HGDP collection. So far, the HGDP has been developed with minimum investment. There are hopes that it will be expanded, even if governmental support for science has unfortunately shrunk almost everywhere. Clearly, countries where genome studies are more likely to be fruitful in the short term are those for which it is reasonable to invest proportionately more at this stage. It is practically inevitable that the two biggest clusters of populations, one including Europe, North Africa, and west Asia

including Pakistan and India, and the other East Asia, will be studied preferentially, given that these two areas have the strongest economic development, and that they also intergrade genetically into each other. But it will also be important to dedicate a reasonable part of the effort to the indigenous parts of the rest of the world—Africa, Oceania, and the Americas—otherwise recent immigrants to the more highly developed part of the world will not have access to the biomedical information available to other residents. Not only is this inequality undesirable but our understanding of human evolution will remain incomplete without adequate consideration of the other continents.

LITERATURE CITED

1. Cann RL, Stoneking M, Wilson AC. 1987. Mitochondrial DNA and human evolution. *Nature* 325:31–36
2. Cavalli-Sforza LL. 1973. Some current problems in human population genetics. *Am. J. Hum. Genet.* 25:82–104
3. Cavalli-Sforza LL. 2005. The Human Genome Diversity Project: Past, present and future. *Nat. Rev. Genet.* 6(4):333–40
4. Cavalli-Sforza LL, Bodmer WF. 1999. *The Genetics of Human Populations*. New York: Dover Press
5. Cavalli-Sforza LL, Feldman MW. 2003. Biology as history: population genetic approaches to modern human evolution. *Nat. Genet.* 33:266–75
6. Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. *The History and Geography of Human Genes*. Princeton, NJ: Princeton Univ. Press
7. Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J. 1988. Reconstruction of human evolution: bringing together genetic, archeological and linguistic data. *Proc. Natl. Acad. Sci. USA* 85:6002–6
8. Fisher RA. 1930. *The Genetical Theory of Natural Selection*. New York: Dover Press
9. Fisher RA. 1937. The wave of advance of advantageous genes. *Ann. Eugen.* 7:355–69
10. Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624–26
11. Lewontin RC. 1985. Population genetics. *Annu. Rev. Genet.* 19:81–102
12. Liu H, Prugnolle F, Manica A, Balloux F. 2006. A geographically explicit model of worldwide Human-Settlement theory. *Am. J. Hum. Genet.* 79:230–37
13. Ohta T, Kimura M. 1970. Development of associative overdominance through linkage disequilibrium in finite populations. *Genet. Res.* 16:165–77
14. Prugnolle F, Manica A, Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* 15:1022–27
15. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support for a serial founder effect originating in Africa, using the relationship between genetic and geographic distance in human populations. *Proc. Natl. Acad. Sci. USA* 102:15492–97
16. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd HM, et al. 2002. Genetic structure of human populations. *Science* 298:2381–85
17. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. 2006. Positive natural selection in the human lineage. *Science* 312:1614–20

Annual Review of Genomics and Human Genetics

Volume 8, 2007

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* chapters may be found at http://genom.annualreviews.org/