ABRIDGED PAPERBACK EDITION

The History and Geography of Human Genes

WITH A NEW PREFACE BY THE AUTHORS

L. Luca Cavalli-Sforza, Paolo Menozzi, and Alberto Piazza

THE HISTORY AND GEOGRAPHY OF HUMAN GENES

The History and Geography of Human Genes

ABRIDGED PAPERBACK EDITION

L. Luca Cavalli-Sforza
Paolo Menozzi
Alberto Piazza

PRINCETON UNIVERSITY PRESS
PRINCETON, NEW JERSEY

Published by Princeton University Press, 41 William Street,
Princeton, New Jersey 08540
In the United Kingdom: Princeton University Press,
Chichester, West Sussex

Library of Congress Cataloging-in-Publication Data
Cavalli-Sforza, L. L. (Luigi Luca), 1922–
The history and geography of human genes /
Luigi Luca Cavalli-Sforza, Paolo Menozzi, Alberto Piazza.
p. cm.
Includes bibliographical references and index.
ISBN 0-691-08750-4 (unabridged cloth ed.)
ISBN 0-691-02905-9 (abridged pbk. ed.)
1. Human population genetics—History. 2. Human evolution.
3. Human geography. 4. Human population genetics—Research.
I. Menozzi, Paolo, 1946– . II. Piazza, Alberto, 1941–
III. Title.
QH431.C395 1993
573.21'5—dc20 93–19339

This book has been composed in Times Roman
Designed by Jan Lilly

The cover illustration is a map of the world showing four major ethnic regions. Africans are yellow, Australians red, and Caucasians green. Mongoloids show the greatest variation retaining some similarities with Europeans on one side (a light brown greenish tinge in middle Siberia) and with Australians on the other (a pinkish color in parts of America and on the way to it). The extensive gradients due to admixtures between Africans and Caucasoids in North Africa, and between Caucasoids and Mongoloids in Middle Asia, are clearly visible. (See chapter two.)

Princeton University Press books are printed on acid-free paper and meet the guidelines for permanence and durability of the Committee on Production Guidelines for Book Longevity of the Council on Library Resources

First printing of the abridged paperback edition, 1996

Printed in the United States of America

10 9 8
ISBN-13: 978-0-691-02905-4 (pbk.)

To our wives

Alba, Wallis, *and* Ada

CONTENTS

PREFACE TO THE PAPERBACK EDITION

THE FIRST edition of *The History and Geography of Human Genes* has been in print for less than two years and has received a favorable welcome by scientists and the public. In a friendly review that appeared in *Nature*, Jared Diamond compared it to a box of chocolates one keeps in the refrigerator as a source of precious morsels. He also noted that the volume is heavy to handle: the geographic maps of single genes occupy about half the volume and are, with their large format, responsible for its size. It seemed reasonable to print an abridged paperback edition, retaining only the unmodified text and its own bibliography and analytical index. The original edition, with its numerical tables and geographical maps, will still be available for those who prefer it, and for libraries where it can always be consulted.

Collecting data from the literature for the original book was begun in 1978 and ended in 1986, after which the genetic and statistical analysis was started, and the actual writing begun. During the long process of publication of a volume this size (the final proof correction took place in the summer of 1993, and publication in summer 1994), there were chances of updating the text. Inevitably, the need to avoid further delays restricted somewhat our taking advantage of these opportunities. We hope to put soon on the Internet the complete numerical data on which the book is based. They refer to almost 2,000 different populations and contain as many as 86,000 entries. The tables in the first edition of the book contain only a summary of this file.

In the last ten years new data have accumulated, enriching the information on "classical markers," and especially that on DNA markers. Quite a number of new experimental methods of molecular DNA analysis have been devised, generating new types of data, as well as new methods of evolutionary analysis. With the experience we accumulated, the existing organization and the available network of communications, updating the information and analysis should be much faster than the first collection and review of the existing data. Depending on the resources that will be available, old and new information could also be shared by interested readers much more promptly. The Human Genome Diversity Project aims at collecting a balanced and representative set of human samples across the world in order to have information on genetic differences of our species. In five to ten years it should accumulate an entirely new matrix of data, much more systematically and efficiently than has been possible so far. New evidence will probably resolve old problems and provide new clues and new problems worth considering.

For this paperback edition we still offer—to a wider audience, we hope—the same suggestions, proofs, and conclusions we have presented in the full volume. As we discuss more specifically in the Epilogue, which is reprinted from the first edition, many of our conclusions are tentative and are there to be tested, challenged, or confirmed. The reduced size of this volume does not reduce the importance we attribute to their message. What we have seen published since the writing of the manuscript has not suggested any dramatic change in our conclusions, but a new era is now beginning.

PREFACE

THIRTY years ago the first effort was made to reconstruct the history of human differentiation by employing the genetic divergence observed among human groups. The data base comprised gene frequencies, that is, frequencies of alleles at polymorphic loci known to be clearly inherited. Observed frequencies are very stable and seem to be rather insensitive to short-term environmental change. There are, however, very few if any data from the past, and stability in time is inferred from the stability in space, essentially the regularity of gene-frequency distributions and the very small differences usually observed among populations that live in widely different environments. Fortunately, in very recent times, new developments in molecular technology have generated the hope of obtaining substantial information from individuals or populations that have been dead for a long time.

Data from physical anthropology (including skin color, body build, and facial traits) had previously been the only source of information. Some of these data, especially measurements on bones, have the great advantage of being readable in the fossil material. Unfortunately, data available for the past have shown conspicuous changes in the last 200 years, as, for instance, the trend to increase in stature and changes in other measurements observed in Europe. It is difficult to ascribe these observations to genetic causes, and it is more likely that they represent responses to recent environmental changes. They are therefore less suitable for the study of genetic history. Even so, major differences observed in the fossil material have been important for reconstructing the general lines of evolution of the genus *Homo*. More detailed conclusions are still controversial because of the rarity of informative specimens and of dating difficulties in the time range of greater interest. Some of these limitations are slowly being removed.

Genetic data about extant populations useful for our purposes are extremely numerous. Two of the first polymorphic loci discovered, the ABO and RH blood groups, had considerable clinical importance and were tested very widely. Many other markers with no clinical interest were nevertheless investigated in many populations because of the anthropological information they can provide. Unfortunately, the existing data vary widely in number and geographic distribution. If they had been collected more systematically according to a rational plan, as occurred for important markers like those of the HLA system, we would have a much more informative body of data. Today molecular genetics is providing us with enormously more powerful technology, but the data base thus generated is still minimal, and we should better organize our future efforts.

There is another important reason for starting a major program in analyzing human diversity now. While our potential skills for analyzing human evolution are increasing, social changes taking place in developing countries are rapidly destroying the identities—if not the very existence—of the most important aboriginal populations. Thus, organized research efforts to save this precious information about our past have acquired a new urgency. Fortunately, recent technical developments make the prospects very exciting, so that this is a good time for taking stock of available knowledge and using it as a guide for planning future research.

This book was started with the desire to analyze the geography of human genes, using new techniques we have developed for the purpose of studying ancient human migrations. While the very demanding work of computerizing the enormous data base in existence was proceeding, it became clear that there was a need to analyze the same information with other techniques, developed by us and by others, which can lead to conclusions of historical interest. But the challenging task of reconstructing the history of human evolution can hardly be entirely satisfactory using only evidence provided by the genetic data. Information from historical, linguistic, anthropological, and archaeological sources is also useful, and it should be compared with the genetic evidence if we wish to reach fully satisfactory conclusions.

Needless to say, all these sources have their own limitations. Relevant data from history are infrequent, far from quantitative, and do not usually probe deep enough in time. Archaeology says very little about the physical populations it studies, but it gives dates and some, however vague, information on demography, especially on numbers of people, that are important for predicting the rates of genetic evolution. But archaeologists often find it difficult to distinguish between the migration of

people and the diffusion of artifacts or the techniques for making them. Linguistic change follows rules that are somewhat analogous to those of genetic evolution, except that it is much faster and the reconstruction of early stages is therefore especially difficult. Moreover, languages are sometimes replaced by others of totally different origin in a very short time, partially blurring the concordances. Physical anthropology can be misleading because certain physical traits observed in bones can sometimes change quickly with changing environmental conditions. Only genes almost always have the degree of permanence necessary for discussing fissions, fusions, and migrations of populations that took place during the history of our subspecies, which goes back for at least 100,000 years. A large fraction of the genetic variants we study appeared before that time. Their relative proportions have changed considerably since and can orient us in understanding population history.

Although population geneticists often summarize knowledge about the archaeology, history, and linguistics of the ethnic groups they have studied, there has been no comprehensive treatment or attempt at a global picture of our species from the points of view of general history that are relevant for genetics. We hope to fill this gap with the present volume. In the first chapter we give some general historical information on the subject, a discussion of the concept of race, its failure, and an elementary introduction to the major analytical techniques used for our purposes. We have tried to make the book readable to scientists of as many disciplines as possible, given that not only geneticists but also scholars from fields as diverse as archaeology, anthropology, history, geography, and linguistics have a potential interest in the subject. Most barriers to cross-disciplinary exchanges are the result of the specialized vocabularies of each field, and we have tried to counter this limitation as much as possible. This is tantamount to saying that lay readers could also understand this book, if they have the motivation necessary for going through a scientific analysis. Inevitably, the discussion is kept at an elementary level in each of these disciplines, and the language used is as simple as possible. Statistical methods and basic population genetics theory are explained in a qualitative way with economical use of scientific terms; all of which are defined at their first introduction.

The second chapter is dedicated to an analysis of the world data with the aim of understanding the general history of *Homo sapiens sapiens*. Trees of descent are reconstructed and compared with archaeological data and linguistic classifications. Other methods of analysis are applied to the global data for an evaluation of the genetic structure of the species as a whole.

The five chapters that follow are dedicated to the major geographic subdivisions of the inhabited Earth. We start with the continent where the genus and probably also the subspecies to which we belong have first developed, Africa, and then proceed with the other continents successively occupied, though not in the strict order of occupation: Asia, Europe, America, and Oceania. In each chapter we briefly discuss geography and ecology, and then history, starting with paleoanthropological and archaeological information. We pay special attention, when possible, to population numbers and densities, as well as to migrations that have special relevance for the evolutionary processes in which we are interested. Physical anthropology and linguistics follow. Then an analysis of the available genetic data is given for each continent in general and for its most important subsections. Geographic maps of genes for which there is enough information are given for the world and each continent at the end of the volume. "Synthetic" geographic maps derived from them are given in the text and show the major genetic patterns that can be abstracted from the total genetic "landscape" by suitable methods. There is not always enough genetic information to make full use of or to interpret all the historical and other nongenetic information given in the first sections of each chapter, nor is there enough of the latter to explain all details of the former, but we hope the unused information may be a stimulus for further research.

The last chapter is an epilogue that discusses generally our conclusions from a methodological point of view and the most urgent problems facing the continuation of research at this crucial time. We now have the tools for doing a much better job than has been done thus far at the level of both data collection and analysis. There is, of course, room for improvement in both, but the usefulness of living populations is being destroyed by a rapid increase in the rate at which human populations are vanishing. The mixing of formerly isolated groups is especially damaging for future research. This is a critical time for organizing our efforts before we lose a unique opportunity for understanding our genetic heritage.

As already mentioned, the second half of the book is dedicated to geographic maps for all genes for which the amount of data of aboriginal populations was deemed adequate. It is difficult to establish an objective criterion for deciding when data are sufficient for making a map, and the choice of alleles and continents represented in the maps was in part subjective. Gene frequencies from samples that were geographically too close had to be averaged before they were used in constructing maps. For different populations inhabiting the same region, we had to choose between discarding some of them or pooling them. In general, when there were both aboriginal populations and late settlers that could be easily distinguished, the former were chosen. The pooling of distinct populations living in the same narrow area generated local heterogeneities, which were systematically estimated and are shown on each map.

For satisfactory map construction, the regularity of the geographic distribution of the data is even more important than the total number of observations. Even for the most intensively studied genes, some areas are not well sampled. In order to give an idea of the strengths and weaknesses of each map from the point of view of the spatial distribution of data, we have indicated the locations at which data were available, as well as significant local heterogeneity, if any. No smoothing of the data could be perfect; we have therefore indicated where the calculated surface of gene frequencies departed significantly from the observations, and the direction of departure. A brief comment on the map of each gene is given in a special section of the appropriate chapter. These single-gene maps were used to generate the synthetic ones.

All gene frequencies obtained from the literature were used for building the geographic maps, but only a selection of populations tested for a greater number of genes was employed for tree analysis. The two methods, trees and geographic maps can be considered complementary descriptions of the same reality. The first stresses historical aspects, and the second the geographic ones. The historical interpretation of trees needs to be strengthened by tests of the validity of the hypotheses underlying them, which is sometimes possible. We usually find good agreement between genetic and nongenetic information, which encourages us further to believe in our conclusions. If nothing else, the presentation of clear hypotheses that can be tested is, we believe, a valuable contribution.

The gene frequencies of the population samples used for tree analysis in the various chapters and their sources are also given in the second part of the book (Appendixes 1 and 2). The bibliography of gene frequencies used for trees and maps is separate from that of works cited in the text (Appendix 3 and Appendix Bibliography). The largest part of gene-frequency data is also found in earlier tabulations that report the relevant sources, and we make specific reference to them population by population.

The task we set before ourselves was not an easy one, and we hope critical readers will recognize that the need to summarize a substantial amount of information of varied nature has inevitably generated the possibility of important omissions and errors. In particular, we apologize to authors who may feel their work has not been adequately considered. In many cases we have preferred to give our conclusions without comparing them with dissenting ones. Our excuse is that we wanted to present testable hypotheses and indicate the basis on which we have accepted or discarded them, without attempting to be fully comprehensive (a nearly impossible task). We are hopeful that our effort will help to spread knowledge and interest in human population genetics, and to recognize the usefulness of thinking in multidisciplinary terms. Much work is necessary for filling important gaps, for organizing future research more satisfactorily at an international level, and for making full use of the power of present techniques at this critical time, when crucial information is slipping out of our hands.

ACKNOWLEDGMENTS

OUR LIST of acknowledgments is long, and we hope we are not leaving out anyone. Many people helped us with computerizing our data and in many other ways. In particular, we want to thank Juliana Hwang for computer programming, and Mariangela Trentadue, Sharon Feingold, and the late Michelle Leo for data input. Most text figures were computer-generated by Megan Betz, and in part by Kim Ha. In addition to this effort, which was rather heroic considering the computer programs that were used at the time, they took part in the final preparation of the bibliography. Much of the statistical analysis was skillfully performed by Nazario Cappello, Eric Minch, Joanna Mountain, and Sabina Rendine. To all of them we express our heartfelt thanks.

The production of the geographic maps was an eight-year endeavor undertaken at the University of Parma under the auspices of its computer center, which made available unlimited amounts of computer time and untold miles of plotter paper and ink. We could not have carried out our project without the unlimited support and assistance of the computer center staff. Words cannot express our gratitude for Enzo Siri's wizard capacity for solving graphics programming problems and his biblical patience in producing the maps that appear in this book.

The consistently creative help of Eleonora Olivetti and Sabina Rendine, always over and above the demands of duty, in preparing the plots included in each map and the color synthetic maps played a critical role in completing the graphic part of our work.

Many people kindly read parts of the manuscript and suggested changes: Peter Bellwood, Jaume Bertranpetit, Anne Bowcock, Giacomo Giacobini, Barry Hewlett, Eric Minch, James V. Neel, Colin Renfrew, Philip Rightmire, S. M. Sirajuddin, Robert Sokal, Chris Stringer, Ken Weiss. Merritt Ruhlen read the entire manuscript. Mike Crawford helped greatly with data from Siberia. Frank Livingstone, Arthur Mourant, and Don Tills provided us with proofs of their tabulations in advance of publication of their books. Franco Scudo made us aware of the Darwin citation used in chapter 8. Rosalba Guglielmino Matessi contributed in processing and interpreting the hemoglobin data included in this book. None of these helpers must be blamed for our mistakes, but we are grateful for their help and encouragement. Finally, Ed Tenner and Judy May of Princeton University Press provided the initial encouragement to extend our 1978 geographic analysis of Europe to the rest of the world, and the entire Press (in particular, Emily Wilkinson) have been remarkably helpful and patient as we missed deadline upon deadline for presentation of our final manuscript. Copy editing, an endless job, fell upon the shoulders of Teresa Carson, who contributed gracefully to improving the final product. Last but not least, Marilyn Anderson lightened our load by providing secretarial help with sensitivity and good spirit.

Grants-in-aid should be acknowledged. Without the National Institutes of Health (NIH) grants GM20467 and GM10452, this research would not have been possible. Financial help also came from the NIH National Library of Medicine (grant LM04106), the Lucille P. Markey Charitable Trust, the Wenner-Gren Foundation, and the National Science Foundation (Anthropology Division). On the other side of the Atlantic, the Italian CNR (Consiglio Nazionale delle Ricerche) Projects "Biotechnology and Bioinstrumentation," "Genetic Engineering," "Biological Archive," and MURST (Ministero Universita' Ricerca Scientifica Tecnologica) funds of 40% and 60% are gratefully acknowledged. Computer time for graphics (plots and color maps) was supported by grants and facilities of CSI-Piemonte (Consorzio per il Sistema Informativo, Torino, Italy). The help of their staff was especially appreciated.

THE HISTORY AND GEOGRAPHY OF HUMAN GENES

1 INTRODUCTION TO CONCEPTS, DATA, AND METHODS

1.1. INTRODUCTION

For some time, geneticists had been aware of a certain amount of genetic variation among the individuals forming a species, but the remarkable extent of this variation was not appreciated until about 25 years ago. Conspicuous human traits like hair and eye color clearly vary from one individual to the other in many populations; these differences are easily perceived by the layman, as are variation in height, weight, body build, and facial traits, which are also genetically determined to some extent. Their hereditary transmission, however, is complex, and these traits contribute little to our understanding of the extent of variation. The first example of clear-cut genetic variation, that of ABO blood groups, was described at the beginning of the century (Landsteiner 1901). Dissimilarities between individuals regarding ABO blood-group variation are due to small chemical differences between molecules found at the surface of red blood cells.

These studies were soon extended to other blood-group systems, and a body of data began to accumulate showing that different human populations have different proportions of blood groups. However, the first glimpse of the staggering magnitude of genetic variation came later—beginning in the 1950s and coming to full development in the 1960s—when individual differences for proteins could be systematically studied. A *protein* is a large molecule made of a linear sequence of components called *amino acids*; different proteins vary considerably in their amino-acid composition and serve very different functions. The relationship between structure and function has been demonstrated for many proteins. The same protein may show small, strictly inherited differences between individuals. The first example was observed in the protein hemoglobin, in which the replacement of a specific amino acid by another was shown to determine a hereditary disease known as sickle-cell anemia. This first case of "molecular pathology" was detected by subjecting the protein to an electric field with a procedure called electrophoresis (Pauling et al. 1949; Ingram 1957). The amino-acid replacement involved in sickle-cell anemia causes a change in the electric charge of the hemoglobin molecule, which allows the separation of normal and sickle-cell hemoglobins. Electrophoretic analysis has since been further developed and has helped detect a great deal of variation in proteins. It is now known that the majority of the tens of thousands of different proteins found in an organism exist in more than one form, so that some individuals may have one form of the protein, whereas others may have another form.

Protein variation is still the tip of the iceberg. Only when the analysis could be carried out at the level of the hereditary material itself, deoxyribonucleic acid (DNA), could the full extent of individual genetic variation begin to emerge. This technique became widely available only in the 1980s, and although comparisons of segments of DNA in different individuals are still rare, they are becoming more common. They are, however, adequate to convince us that there is much more variation at the DNA

level than was suspected when only proteins and blood groups could be analyzed.

Techniques of DNA analysis are still being developed rapidly, and the future will undoubtedly see more and more attention being paid to individual variation at the DNA level. Meanwhile, an enormous wealth of information has accumulated and keeps accumulating on individual variation studied with immunological techniques (as the blood groups are) or with electrophoresis of proteins.

If we know that there exist different genetic types of a specific protein or other strictly inherited character, we can count individuals carrying one type or the other and establish the proportions of that type in the population being examined. These proportions vary from one population to another because they change over time in each population in a relatively unpredictable manner. The change in proportions of these types over time is the *evolutionary process* itself. It proceeds slowly but incessantly over generations. The analysis of populations living today in different places gives us a cross section in time of this continuing process, which is inevitably diverse in the various parts of the inhabited Earth.

Our primary interest is in understanding this evolutionary process. The first task is to describe the existing variation, using a variety of techniques that lend themselves to this work and allow us to test the relevant evolutionary models. We restrict our interest to aboriginal populations, which we define as those already living in the area of study in A.D. 1492. After this time, geographic discoveries stimulated the expansion and migrations of the economically more advanced populations all over the planet. Some movement took place before A.D. 1492, but at a smaller scale. Ordinarily, populations that migrated after that date have mixed only partially with earlier residents and are easily recognizable on the basis of physical appearance and historical and social knowledge. They, and some populations that are highly isolated and/or have had a complex history—such as Samaritans, Jews, Gypsies, and several others—need special study and are not considered in this book. Samaritans, as well as many Jewish populations, have been the object of analysis by Batsheva Bonné-Tamir (1980; Bonné-Tamir et al. 1992). Several general articles and books have been dedicated to Jews (e.g., Mourant 1978; Carmelli and Cavalli-Sforza 1979; Karlin et al. 1979; Morton et al. 1982; Livshits et al. 1991).

One way of studying living populations is a geographic representation of the data. For this purpose we first consider each *gene* (a segment of DNA endowed with a specific function) by itself, and for each gene we separately analyze the different forms that we can recognize, the *alleles* of that gene. The proportion of a given allele in different populations is the raw material of this approach. It is well established that the proportion of an allele varies considerably from place to place, but usually there is little difference between neighboring populations so that the greatest variation is observed at large distances. It is thus possible to prepare geographic maps representing these proportions for a particular allele (also called *allele frequencies*, or simply, *gene frequencies*) when a sufficient number of populations have been tested. The standard procedure is to draw *isogenic curves* or lines connecting points of equal gene frequency.

Geographic maps of an allele are useful for understanding facts specific to that allele, including its evolutionary history and the effects of evolutionary factors like mutation and natural selection. The geographic distribution of a particular allele may give information on the place of origin of the genetic change (*mutation*) that generated it. Correlations of the distributions of gene frequencies with environmental parameters at the geographic level have been instrumental in the discovery of specific genetic adaptations. The sickle-cell anemia gene was the first example, because its geographic distribution showed a correlation with that of malaria (Haldane 1949). The hypothesis that this gene may confer resistance to malaria was later confirmed by more direct tests.

For a long time anthropologists tried to reconstruct evolutionary relationships and history on the basis of a single character or gene. A favorite for over 100 years was the cephalic index (the percentage of skull breadth to length) introduced shortly before the middle of the last century. However, with a single trait, two populations of different origin could well turn out to be more or less identical. Anthropometric traits of this kind also have another very serious drawback: there is no guarantee that the character is completely under the control of biological inheritance and the variations observed could be due to short-term response to environmental changes. This was shown by Boas (1940) at the beginning of the century, but this lesson was, and still is, usually forgotten. The main advantage offered by such traits, namely the availability of data from fossil bones, was therefore minimized because of the uncertain nature of the observed differences.

After the first blood-group system (ABO) was discovered, ABO gene frequencies soon became a favorite for classifying populations. The information thus obtained, however, is also inadequate, even if it escapes to a large extent the limitation of possible short-term changes under direct environmental effect. Every gene frequency varies over time in ways that can be considered, at least superficially, nearly random. Therefore, it is not surprising that populations having clearly different evolutionary histories may show similar gene frequencies. This drawback can be avoided if one cumulates the information from more than one gene. As one increases the number of genes considered simultaneously, the probability that a similar confusion takes place becomes more and more remote. In 1963 it was shown that even with as few as 20 alleles from five genes one could successfully attempt

a reconstruction of human evolution (Cavalli-Sforza and Edwards 1964). Later experience proved that a larger number is desirable or even necessary.

Several methods allow us to combine the information from many genes into appropriate statistical indices. They are usually called *multivariate* to distinguish them from those using single traits or genes (*univariate*).

Multivariate analysis is especially useful for understanding evolutionary forces that tend to operate in a parallel fashion on all genes: migration and random genetic drift (the random fluctuation of gene frequencies in time, to be further explained later). These and other methods are applied to the existing data with the aim of extracting information of genetic and evolutionary interest.

The reconstruction of human evolution, including the fissions, the major migrations, and the understanding of the roles of mutation, drift, and natural selection is often difficult and challenging. There is clearly little hope of an experimental approach to our species, in which the evolutionary process could be repeated and interfered with in known ways. This, as well as the present almost total lack of fossil data on genetic variation (from populations living at earlier times), generates a strong desire for external evidence that can support the conclusions of genetic analysis. Fortunately, information from other sources can supply some clarification. The credibility of our conclusions can be greatly strengthened if these conclusions can be confirmed in the light of an interdisciplinary approach. Results from genetic data should be compared with relevant knowledge from other fields, in particular, paleoanthropology, prehistory, history, the geographic and ecological setting, and the cultural evidence that comes indirectly from linguistic studies. We have considered such feedback an essential part of our analysis, and we have designed our book in order to satisfy this requirement. The remainder of this chapter is dedicated to an introduction to specific concepts, data, models, and methods.

1.2. Genetic definitions

The purpose of this section is to provide some elementary definitions for readers who have no background in genetics. Genetic information is present in every cell of an organism in the form of chromosomes, of which there exist 23 pairs per cell in humans. Each pair is made up of one member of paternal and one of maternal origin that are morphologically indistinguishable but show subtle differences detectable at the chemical level. The main constituent of chromosomes, and the carrier of genetic information, is deoxyribonucleic acid (DNA), a long thread consisting of a linear sequence of relatively small molecules called *nucleotides*, or simply *bases*. Each nucleotide is chosen from four different ones, indicated by the symbols A,C,G,T. A short segment of DNA may look like a superficially random sequence of nucleotides, for example, TAACATGCCAT. . . .

The order of the nucleotides is actually responsible for the specific actions of DNA and is copied almost without error at reproduction of cells and individuals. Thus, the progeny contains DNA with a sequence essentially identical to that of the parent, and this is the mechanism that ensures the maintenance of the properties of living organisms.

The DNA thread is most probably continuous along a chromosome and is extremely long, since the average number of nucleotides per chromosome is over 100 million. In spite of its continuity, one can recognize shorter segments in the DNA that have specific functions and are called *genes*. The genes we recognize most easily are those that direct the structure and shape (and thus also the function) of proteins, complicated biological molecules that perform a great variety of specific activities in the cells. A chromosome may contain, on the average, many thousands of genes, each made of thousands of bases.

At cell division, DNA is replicated so that each of the two daughter cells generated by the division of one cell contains DNA that is practically identical to that of the parent cell, with very few errors in replication. Such errors are transmitted to the progeny because the new DNA is the master from which all future copies are made. Transmission error in the reproduction of DNA is called *mutation*. It may be the replacement of a nucleotide by another of the four, or the addition or deletion of nucleotides. Mutation may have trivial or serious consequences for the whole organism, depending on the alteration in the function of the specific gene to which the altered DNA belongs. Mutation in the dividing cell of an organism made up of many cells, like humans, may lead to alteration of part of the organism, but it is not transmitted to descendants unless it occurs in *germinal* cells, or *gametes*. Gametes are dedicated to the production of individuals of the next generation, and mutations occurring in them can be passed on to progeny and thus have evolutionary consequences. The male (sperm) and female (egg) gametes contain only 23 chromosomes. Reduction in number takes place by a very precise mechanism of random assortment of chromosomes so that each gamete receives only one member of each chromosome pair. The union of a sperm and an egg generates a new cell, a *zygote,* which again has 46 chromosomes, that is, 23 pairs in each of which one member is of paternal and the other of maternal origin (fig. 1.2.1).

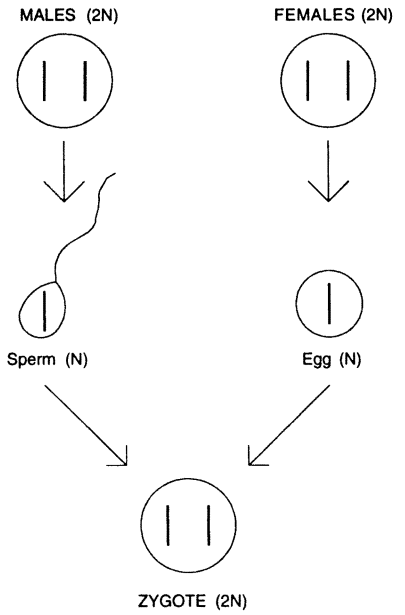

Fig. 1.2.1 The mechanism of reduction of chromosomes by which gamete cells are formed. From the gamete cells, a fertilized egg cell (a zygote) develops. For simplicity, only one chromosome pair is represented. In humans, N, the number of chromosomes per cell, is 23.

Mutation results in a new gene that is slightly different from the old one; the two types are called *alleles* of that gene. In the first generation after mutation, there is only one individual in a given population carrying the new "mutant" allele. If the first individual carrying the mutation reaches the adult stage and has several offspring, there is a higher chance that the new allele originated by mutation will be found in later generations; however, many new mutations are lost in the first few generations. It is also possible that under the influence of evolutionary forces described later a new allele will become more and more frequent, on the average, in succeeding generations; and after many generations no copies of the old allele may be left in any of the individuals forming the population. This replacement of an old allele by a new one (the "fixation" of a new mutant) is the elementary process of evolution. It may take a great number of generations, on the average, tens or hundreds of thousands.

Even if mutation is rare, there will be at least some dozens of different mutations in any gamete, considering all the many different genes it contains. The total number of nucleotides in a gamete is very large (3 billion), and the mutation rate per generation may be of the order of 1 in 200 million nucleotides. Most mutations will happen in different genes, but over time the same gene may be hit again by another mutation. In this way, many alleles of the same gene can arise and coexist in a population. When we detect the presence of two or more alleles of a gene in a population, we call the gene *polymorphic*. Polymorphisms produce the genetic markers used in all sorts of genetic studies, including evolution. The usual types of polymorphisms analyzed are summarized in the next section.

Because all the cells of every individual have one (and the same) gene of paternal and one of maternal origin for each type of gene, some individuals may have received different alleles of a polymorphic gene from their parents. They are called *heterozygotes*, whereas individuals that received the same alleles from both parents are called *homozygotes*. The percentage of individuals that are heterozygous at one gene is the *heterozygosity* of the gene, which is the simplest measure of the degree of polymorphism for that gene. It is equal to zero if the gene is not polymorphic.

Assuming that there are only two alleles, say M and m for a gene, an individual can be MM, mm (homozygous), or Mm (heterozygous). These are the three possible *genotypes*, and if the three can all be distinguished by direct observation or by laboratory tests, then it is easy to determine the gene frequency of M or m. In fact, it is enough to count alleles. If there are, for instance,

5 MM individuals,
6 Mm,
3 mm

on a total of 14 individuals then there are $2 \times 5 = 10$ M genes in MM homozygotes and 6 in heterozygotes for a total of $10 + 6 = 16$ M genes. There also are 6 m genes from heterozygotes and $2 \times 3 = 6$ m genes in mm homozygotes, for a total of $6 + 6 = 12$ m genes. Altogether there are $16 + 12 = 28$ genes, or twice the number of individuals counted. The *gene frequency* of M is $16/28 = 0.57$ (57%) and that of m is $12/28 = 0.43$ (43%). The sum of frequencies of all alleles of a gene is 1 (100%).

It may be impossible to count genes directly. If MM individuals are indistinguishable from Mm, but both are different from mm, one cannot separately count individuals that carry two M or only one M. This phenomenon is known as *dominance*, and the types that we can distinguish are called *phenotypes*, meaning distinguishable by appearance. Even if dominance makes it impossible to count MM and Mm individuals separately, under certain conditions one can still determine gene frequencies. The major assumption is that the choice of mates is random for that particular gene. Let us call p the gene frequency of M; that of m is $(1 - p)$ if there are only two alleles. Homozygotes for a given allele are expected to be present in a randomly mating population with frequency equal to the square of the gene frequency of that allele, that is, p^2 for MM, and $(1 - p)^2$ for mm. Heterozygotes are expected to be twice the product of

the gene frequencies of the two alleles of which they are formed, in this case $2p(1-p)$. Thus, the three genotypes *MM, Mm, mm* should have frequencies p^2, $2p(1-p)$, $(1-p)^2$. This rule is named after its discoverers, Hardy-Weinberg. It is easily extended to more than two alleles. We often must have recourse to this rule, but its validity is restricted to populations and genes for which mating is random (as discussed in sec. 1.7). In this book we do not give frequencies of genotypes or phenotypes, but only gene frequencies calculated directly from them by gene counting or by application of the Hardy-Weinberg rule. For traits that are conspicuous—like hair, skin and eye color, or height—mating is not random, and hence this rule would not apply. In any case, their genetic determination is usually unclear or complex. For further reading on these topics see Cavalli-Sforza and Bodmer 1971a; Bodmer and Cavalli-Sforza 1976a; Christiansen and Feldman 1986.

Our current evolutionary thinking is mostly in terms of gene frequencies and their changes. Genes, unlike phenotypes, are essentially stable because mutation is rare; their frequencies are stable in time except for the evolutionary factors we consider later: mutation, migration, selection, and random drift. Thus, the genetic study of evolution is essentially an analysis of the role played by these factors in observed changes.

The study of gene frequencies restricts our analysis to the behavior of single points or very short segments of DNA. As we extend our knowledge, we will direct our attention more and more frequently to the structure of longer and longer segments. The longer a segment, the more polymorphisms for single points it can accommodate. The information will increase but so will its complexity. Long DNA sequences do not behave rigidly in evolution but can exchange segments. The ease with which the sequencing of DNA has recently become possible will make the direct study of long DNA segments more and more commonplace, but most of our present data allow us little more than a point-by-point approach.

1.3. Techniques for Detection of Polymorphic Markers

Polymorphism refers to the presence of more than one allele of a gene in a population. Because we usually observe relatively small samples in terms of numbers of individuals, we tend to call a gene polymorphic if the rarer allele is not too rare—say, not less than 1%—so that there is a good chance of observing a polymorphism in a sample of 100 individuals or more because this is the order of magnitude of most sample sizes. Examined at the highest resolution (the DNA level), almost all genes are highly polymorphic, but this expectation is unfortunately based on few data. It is well known that some DNA regions are much more polymorphic than others. It is more difficult to estimate the average individual variation at the DNA level. Very roughly, 1 in every 500 nucleotides, on the average, differs in two chromosomes taken at random from a population, and therefore also in the two members of a chromosome pair of a random individual. Because genes are usually made up of many thousands of nucleotides, every gene is likely to be polymorphic if fully analyzed. We know, however, that some DNA segments—in particular, those coding for proteins—are much more highly conserved in evolution than others, and therefore we can expect them to show much less individual variation than others.

Polymorphic genes are caught in the middle of a transition from the first appearance of a mutation to its likely final event, fixation or extinction. This is a long process, and most of the time we cannot say which of two alleles is the older and which the newer (the "mutant," by definition). In the human species we can obtain some information on this point by looking at the presence of one or the other allele in the nearest species, chimpanzees or gorillas. Whether or not we know the remote history of polymorphisms, however, they provide us with pointers to the variation of specific chromosome segments. In this sense they are *genetic markers* and thus our door to the understanding and measurement of genetic variation.

The markers we analyze are conveniently classified by the technique used for detecting them and the tissues to which they apply. Polymorphisms can be found in almost any cell or biological fluid, but blood is by far the favorite because it is most easily obtained and gives the greatest opportunities for detecting them. The list of markers analyzed on a geographic basis sufficiently wide for our purpose is given in table 1.3.1. The most important categories of genetic markers are the following.

Blood groups. Blood groups are detected in red blood cells by immunological techniques. Substances at the surface of red cells act as *antigens*; that is, they determine the production of *antibodies* when injected in other individuals of the same or a different species. Antibodies are proteins (immunoglobulins), potentially produced by every individual, but only in large amounts when the organism is stimulated with the specific antigen. Antibodies also react specifically with antigens in test tubes in ways that can be easily made visible. The first system of "blood groups" discovered were called ABO, A and B being the antigens on the surface of red cells with which

Table 1.3.1. Genetic Markers Selected for Use in This Book because Adequate Data Are Available for Many Populations

Name of Locus	*Symbol*	*Chromosome Location*	*Alleles Used*
ABO blood group	*ABO*	9q34.1-34.2	*A, B, O, A1, A2*
Acid phosphatase 1	*ACP1*	2p25	*A, B, C*
Adenosine deaminase	*ADA*	20q13.11	*1*
Adenylate kinase 1	*AK1*	9q34.1-34.2	*1, 2*
Alkaline phosphatase placental	*ALPP*	2q37	*S1, F1*
α-1-antitrypsin	*PI*	14q32.1	*M, F*
β lipoprotein, Ag system	*AG*		*X*
β lipoprotein, Lp system	*LPA*	6q26-27	*Lp(a+)*
Ceruloplasmin	*CP*	3q23-25	*A*
Cholinesterase 1	*CHE1*	3q26-qter	*U*
Cholinesterase 2	*CHE2*	2q	*+*
Complement component 3	*C3*	19p13.3-13.2	*S, F*
Diego blood group	*DI*		*A*
Duffy blood group	*FY*	1q21-25	*A, B, O*
Esterase D	*ESD*	13q14.1-14.2	*1*
Glucose-6-phosphate dehydrogenase	*G6PD*	Xq28	*A-, B-, def*
Glutamic-pyruvate transaminase	*GPT*	8q24.2-qter	*1*
Glycine-rich β-glycoprotein; factor B	*BF*	6p21.3	*S, F, F1, S0.7*
Glyoxalase I	*GLO1*	6p21.3-21.1	*1*
Group-specific component	*GC*	4q12-13	*1, 1F, 1S, 2*
Haptoglobin	*HP*	16q22.1	*1, 1F, 1S, 2*
Hemoglobin, α	*HBA*	16p13.3	
Hemoglobin, β	*HBB*	11p15.5	
Hemoglobin, δ	*HBD*	11p15.5	
Hemoglobin, γ	*HBG*	11p15.5	
HLA-A histocompatibility type	*HLAA*	6p21.3	*1, 2, 3, 9, 10, 11, 28, 29, 30, 31, 32, 33*
HLA-B histocompatibility type	*HLAB*	6p21.3	*5, 7, 8, 12, 13, 14, 15, 16, 17, 18, 21, 22, 27, 35, 37, 40, 41*
Immunoglobulin* GM1; GM3	*IGHG1G3*	14q32.33	*za;g* *zax;g* *f;b0b1b3b4b5* *za;b0b1b3b4b5* *za;b0b1c3c5* *za;b0b1c3b4b5* *za;b0stb3b5* *fa;b0b1b3b4b5* *zx;g* *za;b0sb3b5*
Immunoglobulin KM (Inv)	*IGKC, KM*	2p12	*1&1,2*
Kell blood group	*KEL*		*K, k, Kpa, Jsa*
Kidd blood group	*JK*	18q11-12	*A, B, O*
Lactate dehydrogenase	*LDH*		*A & B variants*
Lewis blood group	*LE*	19	*Le, le, Le(a+)*
Lutheran blood group	*LU*	19q12-13	*A*
Malate dehydrogenase	*MDH1*	2p23	*1*
MNS blood group	*MNS*	4q28-31	*M, N, S, s, S^u, MS, Ms, NS, Ns, He*
P	*P1*	22q11.2-qter	*1*
Peptidase A	*PEPA*	18q23	*1*
Peptidase B	*PEPB*	12q21	*1*
Peptidase C	*PEPC*	1q42 or 1q25	*1*
Phenylthiocarbamide tasting	*PTC*		*T*
Phosphoglucomutase 1	*PGM1*	1p22.1	*1*
Phosphoglucomutase 2	*PGM2*	4p14-q12	*1*
Phosphogluconate dehydrogenase	*PGD*	1p36.2-36.13	*A, C*
Phosphoglycerate kinase 1	*PGK1*	Xq13	*1, 2*

(continued)

Table 1.3.1 *(continued)*

Rhesus blood group	*RH*	1p36.2-34	*D, C, E, D^u, C^w, CDE, CDe, CdE, Cde, cDE, cDe, cdE, cde, V*
Secretor	*FUT2(SE)*	19q	*Se*
Superoxide dismutase 1	*SOD1*	21q 22.1	*1*
Transferrin	*TF*	3q21	*C, D*

* The GM markers also appear in the numeric notation. The correspondence between it and the alphanumeric notation we use will be found in Steinberg and Cook (1981).

anti-A and anti-B antibodies combine, respectively. An individual may carry the A antigen, the B antigen, both, or neither. Thus, four groups of individuals are defined and can be unequivocally tallied (fig. 1.3.1). Individuals of the same ABO blood group show the same reactions to the testing reagents employed and can also exchange red cells by transfusion without adverse consequences. The detection of the ABO system and its inheritance predates World War I. Many other blood-group systems were detected after ABO, of which only a few (especially RH; Landsteiner and Wiener 1940; Levine and Stetson 1939) are important in clinical practice. RH (formerly Rh) has a large number of alleles and is probably a family of adjacent genes. In addition to ABO and RH, MN blood groups and a few others are very widely studied.

Protein electrophoresis. Proteins, the main product of genes, move in an electric field with mobility that depends on their surface electric charge, which in turn depends on their chemical structure. The main polymorphisms studied are those of proteins present in the liquid part of blood (serum or plasma) or in red blood cells. The first detection of a blood protein polymorphism was that of hemoglobin (Pauling et al. 1949), showing the molecular nature of the mutation leading to sickle-cell anemia. Abundant serum proteins (e.g., haptoglobin) were found to be polymorphic soon thereafter. Proteins active as specific catalysts in biochemical reactions (enzymes) are usually present in the blood in small concentrations. When it was learned how to uncover enzymes through very sensitive and specific staining reactions, enzymes provided the first statistical evidence that polymorphisms are much more common than earlier believed. This was discovered simultaneously in humans and in *Drosophila* in 1966 (Harris 1966; Lewontin and Hubby 1966). An example of a two-allele variation detected by electrophoresis is given in figure 1.3.2.

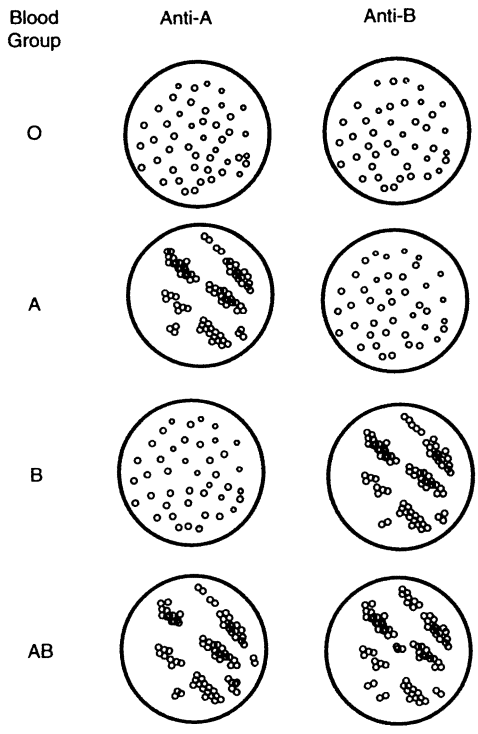

Fig. 1.3.1 Reactions of red blood cells of O, A, B, and AB individuals to anti-A and anti-B reagents.

Human lymphocyte antigens. A precious addition to the arsenal of genetic tools has been that of human lymphocyte antigens, HLA (in vertebrates known as MHC, major histocompatibility complex). These are proteins located on the surface of white blood cells, which participate in the formation of antibodies and have practical importance for organ transplants. Work started in the

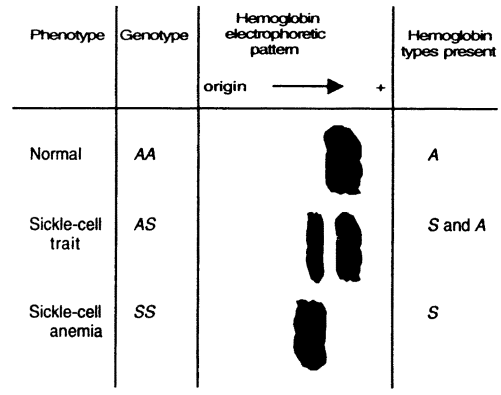

Fig. 1.3.2 Electrophoresis of a protein (hemoglobin) in which two alleles are detected as bands running at different speeds in the electric field. A homozygote (*AA* or *SS*) has only one band; the heterozygote (*AS*) has both.

early 1960s on this superfamily of genes has shown it to be almost as polymorphic as all the rest of non-DNA markers together and has generated the most informative single genetic system known today.

Immunoglobulins. Variants of the immunoglobulins (*GM, KM, AM,* etc., formerly, *Gm, Km, Am,* etc.) found in plasma or serum are tested by special immunological techniques and provide a rich source of genetic variation. Immunoglobulins are the usual "antibodies." Other protein variants are also tested by similar immunological techniques.

Other polymorphisms. There are a few other polymorphisms detected phenotypically by specific techniques (immunodiffusion, often done for lipoproteins, immunoelectrophoresis, autoradiography, etc.). A widely studied one, tested by tasting a substance called phenylthiocarbamide (PTC) has long been suspected to have irregular inheritance. A recent reanalysis (Reddy and Rao 1989) indicates penetrance of the heterozygote is incomplete and there might be a small contribution by polygenes. Gene frequencies of this widely studied marker may be slightly incorrect, therefore, but it would have been difficult to exclude it at this stage; we believe the approximation arising from its inclusion is likely to be negligible.

All the above markers reveal variation at the level of proteins or protein products. They were discovered some time ago and are the only ones for which information is abundant. In general, the longer the time since discovery, the more data at the geographic level are available (fig. 1.3.3), with exceptions tied to the practical difficulties of testing some of them.

DNA polymorphisms. In the future one can expect a sharp rise of information on variation of polymorphisms detected by direct study of DNA. At the moment, data on DNA markers are minimal for many world populations. (They are reviewed briefly in sec. 2.4 in chap. 2.) A method used for DNA study at the population level is restriction analysis. Restriction enzymes cut DNA at specific sites, short segments defined by sequences of usually four, six, or (rarely) more nucleotides ("restriction sites"). A mutation in these sequences will prevent cutting; other mutations may generate new restriction sites. DNA fragments resulting from restriction of the DNA of an individual are electrophoresed and thus separated according to size. Those belonging to a region under study are revealed specifically by binding to DNA "probes," human DNA segments from the chromosome region of interest, cultivated in bacteria and especially prepared by attaching radioactive or other labels that allow their detection. A polymorphism appears as variation of labeled fragment sizes in different individuals (fig. 1.3.4) and is called RFLP, or restriction fragment length polymorphism. By this technique, hundreds of thousands, and perhaps millions, of polymorphisms can be detected; to

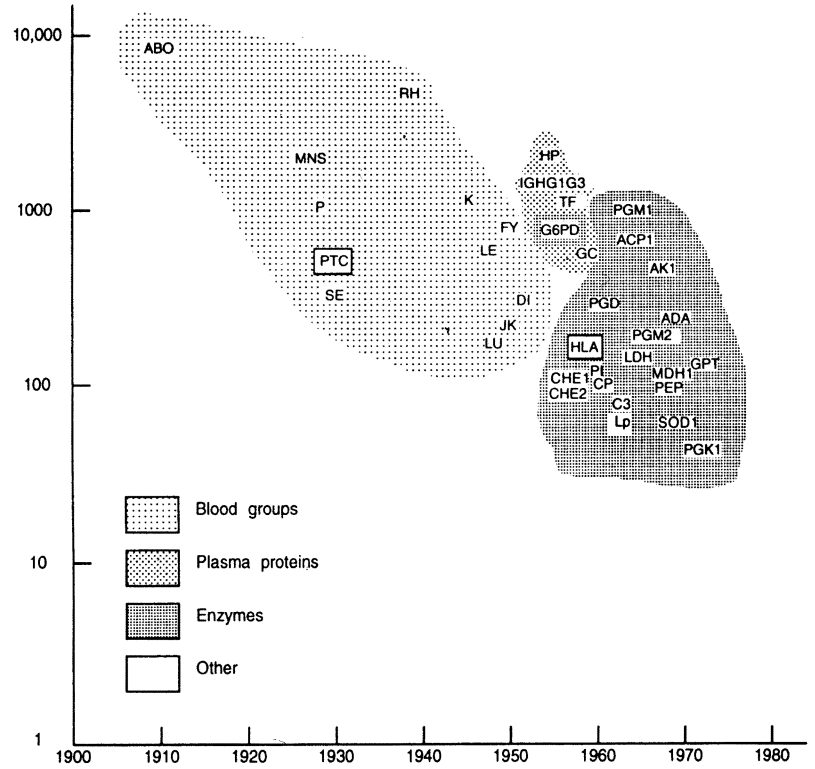

Fig. 1.3.3 Year of discovery of major genetic markers (abscissa) and number of observations in our data base.

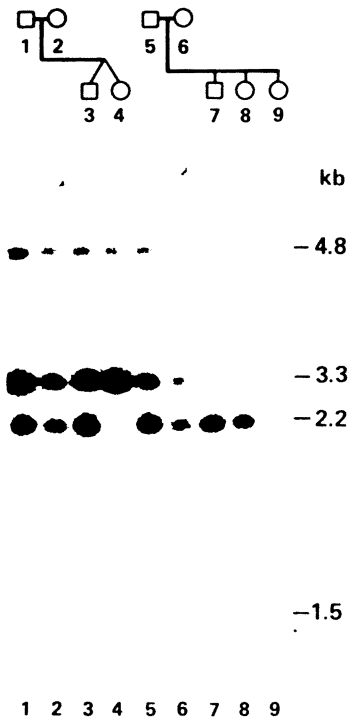

Fig. 1.3.4 DNA polymorphisms using restriction enzyme *DraI* of the *CD8* gene detected by restriction-fragment-length analysis. Alleles of 3.3 and 2.2 Kb of gene *CD8* behave as codominant Mendelian traits. DNA was digested with the restriction enzyme *DraI*, fractionated by agarose-gel electrophoresis, transferred to a nylon support by Southern's method (1975), and hybridized with a radiolabeled cDNA probe that detects the *CD8* gene.

date, more than 2,000 probes showing polymorphism are available.

DNA sequencing is the ultimate method of analysis, but existing sequences are almost always limited to one or few individuals. Powerful techniques (in particular, PCR, the polymerase chain reaction) (Erlich 1989) make the tests of known polymorphic nucleotide replacements and the sequencing of DNA segments much faster and more sensitive. PCR uses the capacity of DNA to be multiplied indefinitely by DNA polymerase, and a single DNA molecule can be amplified at will, usually with little error. One of the major novelties made possible by PCR is the analysis of very old samples in which small amounts of DNA are still left and are not too badly damaged. Some encouraging results have been described (Pääbo et al. 1989). Old samples have also been tested a number of times for non-DNA polymorphisms, in particular ABO; results are usually unsatisfactory, mostly because many individuals give uncertain reactions.

There are textbooks on blood-group polymorphisms (Race and Sanger 1975) and on protein and enzyme polymorphisms (Giblett 1969; Brock and Mayo 1978; Harris 1980). They were published a few years ago, but work on these lines has been slower in recent times. Summaries on HLA are found in a collection of symposia and workshops called *Histocompatibility Testing*. A list of all available DNA polymorphisms appears yearly in the publications of the Human Gene Mapping workshops.

1.4. THE EVOLUTION OF GENE FREQUENCIES

In this section we give a very elementary introduction to population genetics for readers who have no background in genetics. More complete introductions can be found elsewhere (Crow and Kimura 1970; Cavalli-Sforza and Bodmer 1971a; Bodmer and Cavalli-Sforza 1976a; Christiansen and Feldman 1986; Nei 1987; Hartl and Clark 1989; Weir 1989).

Gene frequencies change over time. Mutations supply the raw material by generating new alleles and even new genes, when whole regions are duplicated. Thus, *mutation* is a key ingredient of evolution. Without it, evolution would soon come to a standstill. But a specific mutation very rarely reoccurs in an individual other than the first, and thus the rate of recurrence of the same mutation has little effect on the overall rate of evolution of a particular mutation. Rather, its fate depends on the other three evolutionary forces, *migration, natural selection,* and *random genetic drift,* all of which can affect the gene frequency of an allele present in a population. The first two drive gene frequencies in specific, and to some extent predictable, directions; of the two, natural selection has special importance in determining the future of a species. It is the only evolutionary factor that has direct *adaptive* consequences, because it is the automatic process sorting out and favoring useful mutations while eliminating deleterious ones. It thus makes the functional improvement of living organisms possible. Drift is nondirectional because it is the effect of random sampling of a gamete at each generation, and does not have any simple adaptive consequences. Like drift, mutation is random but may have different probabilities in different directions.

Natural selection is the automatic choice of "fitter" types, which can eventually make an initially very rare type, a single mutant, the most common in a population, provided it is advantageous to the individuals carrying it. The complex adaptations we observe in living organisms would have essentially zero probability of spreading to

whole populations and species by mere chance. Natural selection is responsible for these extraordinary functional adaptations and the complex mechanisms responsible for them. Before Darwin, and after him for people who have not really grasped the power of natural selection, these adaptations have often appeared, understandably, as the product of design, and hence of intelligent creation. Under closer scrutiny, biological adaptations are wonderful but clumsy, like the result of "tinkering" (Jacob 1977), the accumulation of useful mechanisms not by design, but by trial and error, in a historical process, dictated by the chances of spontaneous mutations happening at particular times and places. When mutations offer acceptable solutions to the needs of organisms, they are adopted via natural selection. But they inevitably set later constraints on the further evolutionary process (see, e.g., Crick 1988).

Seen at the most elementary level, natural selection is simply the automatic enrichment of populations in genetic types that produce more descendants, and impoverishment in those that produce fewer. The rate of change under natural selection can be predicted on the basis of the numbers of descendants of each genetic type, strictly speaking, the number of children reaching sexual maturity. This number is called *Darwinian fitness* and is based on demographic parameters like survival and fertility. It is usually expressed in a relative scale, comparing two or more phenotypes or genotypes in the same population. On the basis of Darwinian fitness of two genetic types, one can predict which type, if any, will prevail in the end, and the rate of the process of change in gene frequencies, provided fitnesses do not change over time.

Natural selection acts directly only on phenotypes, but it acts on genotypes in an indirect fashion, depending on the extent to which the phenotype is determined by the genotype. Its genetic effect is thus dictated by the correspondence between the genotype and the phenotype. Phenotypes are chosen or discarded by natural selection, but the phenotypes on which selection acts are not necessarily the same that seem superficially "fitter" to us. Moreover, natural selection may differ in different environments. As an example, there are three phenotypes corresponding to the three genotypes for the sickle-cell gene: *AA*, *AS*, *SS*. The last of the three is severely sick and often dies from sickle-cell anemia. One would expect the gene *S* to be rapidly eliminated (in fact, it usually is, but not under all conditions). The situation is quite different in the presence of malaria, the great killer in tropical and subtropical environments.

By electrophoresis of hemoglobin or by DNA analysis, we can distinguish all three sickle-cell genotypes, whereas simpler laboratory tests separate only the first from the other two. In addition to ensuring the poor survival of the *SS* genotype, natural selection also acts at another level: in certain malarial environments, *AA* individuals survive less well than *AS*, whereas in nonmalarial environments no selective difference is noted between these two genotypes. This classic example shows that natural selection may well change in different environments. It also shows a peculiar evolutionary behavior in that, under environmental conditions favoring the spread of a certain type of malarial parasite (*Plasmodium falciparum*), the heterozygote is at a selective advantage over both homozygotes. The evolutionary outcome of natural selection in favor of the heterozygote is the stabilization of the frequencies of both alleles *A* and *S* at an equilibrium value usually near 90% *A* and 10% *S*; thus, neither allele becomes fixed, but the allele with higher fitness as a homozygote retains a correspondingly higher frequency. In the case of sickle cell, the anemia observed in the *SS* homozygote is devastating and is especially bad in the presence of malaria. Under such conditions, the *SS* genotype is about 10 times less viable than the *AA* genotype, and the allele frequencies at equilibrium in the presence of malaria tend to stay close to a similar ratio, 10 *A* versus 1 *S*. Thus, a heterozygous advantage determines a balanced or stable polymorphism. It is likely, however, that many other polymorphisms we observe are not stable, but their frequencies are slowly changing in time.

In the absence of heterozygous advantage, the allele favored by natural selection will eventually prevail and the other or others will be eliminated. The time taken by the selective process to reach this goal depends on the strength of selection, as measured by relative Darwinian fitnesses of the competing genotypes. These are often expressed as selective coefficients, s, which are percentage differences between the Darwinian fitness of a given genotype (or phenotype) and the Darwinian fitness of a genotype (or phenotype) taken as reference. If the heterozygote has a Darwinian fitness exactly between that of the two homozygotes, the formula for calculating the time t in generations necessary for an advantageous gene to increase from a gene frequency q_0 to $(1 - q_0)$ is especially simple:

$$t = \log[q_0/(1 - q_0)]/s,$$

where log is the natural logarithm. This formula assumes that selection remains at the same level all the time. Naturally, one can think of many other models in which selection varies in space or time, and in many other ways; the present model is the simplest.

Taking, for example, $q_0 = 0.01$ such that t is the time required to go from $q = 1\%$ to $q = 99\%$, and expressing the time in years (with 25 years per generation), the formula predicts that the spread to the population of the advantageous gene will take the following number of

years, given the selection coefficient s: for $s = 0.1\%$, 115,000 years; 0.3%, 38,300 years; 1.0%, 11,500 years; 3.0%, 3800 years; 10.0%, 1150 years.

Unfortunately we know very little about the selection coefficients that prevail in human evolution, because it is difficult to measure them. It would be necessary to observe an impossibly large number of individuals, especially if s is small. A selection coefficient of 10% is very large, and values of this order of magnitude have been measured in the case of the advantage of the heterozygote for sickle-cell anemia versus the normal homozygote in the presence of malaria, as well as for other similar cases of genetic resistance to malaria, such as thalassemia. Another gene for which a selection coefficient of the advantageous type has been estimated is that for lactose tolerance, the capacity of an adult to digest the milk sugar, lactose (practically all young individuals digest lactose until 3–4 years of age). Its frequency varies considerably between populations, being very low where adults do not use fresh milk, and 50%–100% where they use large amounts. The evaluation of this selection coefficient is based on the length of time since bovines and ovines were domesticated and fresh milk thus became available for consumption.

In a previous estimation that did not take into account the interaction between cultural transmission of the custom of drinking milk as an adult and genetic transmission, the selection coefficient for lactose tolerance was estimated at minimally $1\frac{1}{2}\%$–3% (Bodmer and Cavalli-Sforza 1976a). Considering the cultural transmission of milk consumption, the selection coefficient must be at least 10% (Feldman and Cavalli-Sforza 1989).

Other genes also showing great differences between populations are those that control immunoglobulin types, perhaps because they involve differential resistance to infectious diseases, the incidence of which varies widely in different parts of the world. There is no direct estimate of selection coefficients for these genes. One can venture a guess since these genes are among those showing the greatest differences in frequencies between human populations, and these differences must have evolved over the last 50,000–100,000 years. From the data given above, it is likely that the selection intensities involved were expressed by s values very rarely greater than 1%. These are the highest selection coefficients likely to be affecting favorable genes in human populations. There is, however, very strong selection against deleterious genes determining serious diseases and early deaths.

Random genetic drift (also called *drift*) is the fluctuation of gene frequencies from one generation to another as a result of random sampling of gametes (sperm and egg cells). The transition from one generation to another is effected by gametes. The adults participating in the production of offspring, and the gametes made by them and used for this purpose completely determine the gene frequency of the next generation (Cavalli-Sforza and Bodmer 1971a). When a population is small, the total number of gametes involved in forming the next generation is also small and subject to a large *sampling deviation* that depends on the total number of parents contributing to the next generation. Qualitatively, in a small population, the gene frequency may fluctuate wildly from one generation to the other; whereas in a large one, it will be more stable, and stability increases with population size. The effects of drift also accumulate in time, because the gene frequency of a generation is determined entirely by the gene frequency of the preceding generation, without memory of earlier gene frequencies. As a result, deviations caused by sampling increase over time, and the gene frequency of a population may eventually become 0% (extinction of the allele) or 100% (fixation). In fact, fixation or extinction are the inevitable fate of an allele if drift continues long enough, irrespective of population size; but the process takes much longer in large populations, on the average, in proportion to population size. Figure 1.4.1 shows computer simulations with three different numbers of individuals and three different initial gene frequencies. In the two cases with the smallest population sizes and in one with an intermediate one, the rarer allele was lost after a few generations.

The census size of a human population cannot be taken directly as an estimate of the N value in the simplified model used for figure 1.4.1, which, as in virtually all theoretical treatments, is that of a population reproducing synchronously. What matters in practice is the number of active parents, and because about only one-third of the individuals in a real population are of parenting age, the "effective population size" N_e, corresponding to that of a synchronously dividing population is approximately one-third of the census size. More accurate estimates may be made but are usually not necessary.

It is interesting to compare the consequences and to consider the interactions of drift with the other evolutionary factors. The effectiveness of drift depends on demographic factors: we have seen the importance of population size in figure 1.4.1. Migration also strongly affects and in general reduces the consequences of drift, and we consider it later. It is worth stressing that drift affects all genes in a quantitatively similar way, though it affects each randomly and independently of other genes, in the sense that a small population will have high drift fluctuations for *all* genes, whereas a large one will have very little drift, but again for *all* genes. By contrast, natural selection affects each gene in a unique way; we may anticipate, however, that many of the genes we study seem totally unaffected by selection, that is, they are *selectively neutral*. All genes are subject to the effect of drift, and even a mutation favored by

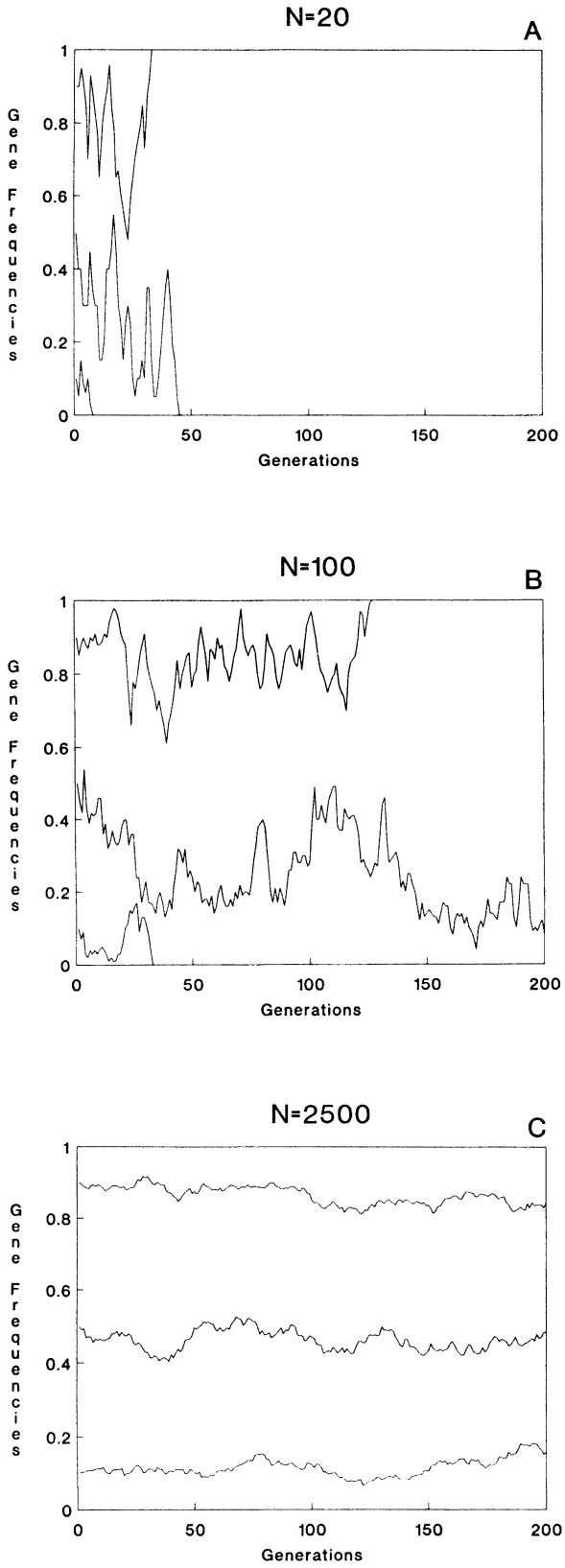

Fig. 1.4.1 Effect of population size on drift. Simulation experiments at three different initial gene frequencies (90%, 50%, 10%) with: A, N = 20 individuals; B, N = 100; C, N = 2500.

selection may be lost because of drift. This can happen with a fairly high probability in the first generations after its appearance by mutation while it is still rare; it is then more strongly exposed to chance effects, almost as strongly as a gene unaffected by selection. However, after an initial period, as soon as the favorable alleles have reached a higher frequency, drift will have little effect on the selective process of genes that are at an advantage.

A mutation that has just occurred is exposed to a fate that depends largely on its effect on the carrier. If this is unfavorable, it will be soon eliminated, together with the carrier or carriers; if favorable, it will increase and may ultimately become fixed, although at the beginning there may be some uncertainty on account of drift, as we have seen. If the mutation is selectively neutral, only drift will matter, and its final fate is either extinction or fixation. Extinction is much more likely, and there is a great probability that a new mutation will disappear in a few generations. But there is a small probability that it will survive and may eventually be fixed under drift alone, fixation taking, on the average, a long time that increases with increasing population size. One might superficially deduce that the pressure of new mutations will have very little effect, per se, on the process of evolutionary change. However, this conclusion, which was popular for some time, has been shown to be wrong. A population is made up of many individuals, all of which are exposed to mutation, and this counterbalances the rarity of mutations. In fact, it has been proved theoretically that the rate of neutral evolution—that is, under mutation and drift alone, without selection—equals the mutation rate (Kimura 1968, 1983).

The relative number of amino-acid differences of the same protein in two different species increases with the time of evolutionary separation between the two species (suggested by geological data). From such numbers one can calculate the rate of molecular evolution. For a given protein it is approximately constant, and there are reasonable explanations for the rate differences observed between proteins. Similar considerations can be extended to the evolutionary rates of DNA. Kimura noted that evolutionary rates thus calculated are comparable in order of magnitude to mutation rates, and this was one of the arguments for proposing that changes observed in molecular evolution are mostly or almost exclusively due to selectively neutral mutations. It is clear that disadvantageous mutations are not uncommon and are fairly rapidly eliminated, but they can be disregarded for this purpose at least at a first approximation; the real question is the relative importance of advantageous mutations versus neutral ones, because these are the only two types of mutations that are fixed in evolution. There must be favorable mutations, or adaptation would not occur, but

the analysis suggests that they are indeed rare with respect to essentially neutral ones.

Kimura's theory that molecular evolution is mostly due to neutral mutations was initially met with strong skepticism by many, but evidence in its favor has accumulated (Kimura 1983). In DNA and protein regions of vital importance for function, one finds perfect—or almost perfect—conservation; that is, variation does not occur between individuals and occurs only rarely between species. This indicates strong selective control against changes that would be deleterious; it also shows that evolutionary improvement in this region is rare or absent. However, variation is quite common in chromosome regions that are not of vital importance. A clear example is the variation observed, for instance, in pseudogenes, which are derived from duplications of active genes but are completely inactive. The gene function is maintained by the active gene, but the inactive copy is not directly exposed to the action of natural selection. If we compare the variation of a pseudogene with that of the corresponding active gene, we find a great difference: unlike the active gene, which is under selective control, the pseudogene can freely accumulate all the variation that can be produced and fixed in evolution when there is no control by natural selection. Thus pseudogenes are under the influence of mutation and drift alone and can be observed to be about 5 times more variable between species than a functioning gene (Li and Graur 1991). Clearly, the active gene is evolutionarily more stable because natural selection weeds out all unfavorable mutations.

All considerations made thus far apply to *closed populations* that do not receive migrants. The species is, by definition, a population closed to migration from other species, but sections of a species can undergo cross-migration. In fact, practically all populations exchange individuals with other populations of the same species and such "migration" usually shows strong dependence on distance: the shorter the geographic distance, the higher the migration rate. Physical obstacles like mountains and rivers can further reduce migrational flow, and routes and means of transportation increase it. Because intercontinental travel began around A.D. 1500, we limited our study to *aboriginal* populations, those that were in place before that date; intercontinental transportation has subverted the earlier patterns of migrational flow.

It is useful to consider two different types of migration: (1) the relocation of individuals or small clusters, like families, leaving one group of individuals—a village, a town, a city—and moving to another, and (2) the relocation of groups, usually larger, moving to new, often uninhabited, territory (Cavalli-Sforza 1973). The first is very frequent and takes place mostly at short distances, generating a continuous internal mixing of populations. The second is rare, but can sometimes cover large distances. The second type was responsible for the occupation of large regions and whole continents. We call the first type *individual migration* and the second, *mass migration* or *colonization*, when the distinction is necessary.

The classical results of drift referred to above are given for closed *isolated* populations that receive no immigrants. The migratory exchange of individuals between these populations tends to buffer the effects of drift, the more strongly, the greater the fraction of immigrants received by a population per generation, m. Knowing effective population size N_e and m, assumed constant per generation, one can evaluate the reduction of drift caused by migration. The quantity $N_e m$ is an index of the degree of relative genetic isolation of a population made up of N_e individuals and receiving a proportion m of immigrants. It allows one to take into account the joint effects of drift and migration. The larger $N_e m$, the smaller the fluctuations of gene frequencies expected over time or over space, that is, between neighboring populations examined at the same time. Simple as it seems, this quantity is not always easy to estimate, especially for larger populations with a complex structure.

In the colonization of a new territory the "founders"—that is, the first colonizers—are sometimes few. They may therefore have an important effect on the subsequent history of the population, and the same will be true of every other demographic bottleneck that may occur at a later time. The phenomenon is especially prominent in the colonization of small islands, but is also observed in *isolates*, populations that do not mix with neighbors for social and/or geographic reasons. Extreme cases have been well documented in a few religious isolates like the Amish (Bachman 1961) and Hutterites (Hostetler and Huntington 1980) in the United States or the Samaritans in Israel (Bonné-Tamir, in prep.). Some investigators prefer to distinguish between drift and "founders' effect," as if they were two different phenomena (Bellwood 1979), but founders' effect is clearly only an episode of drift. Given the slow rate of human reproduction, an initially small population will remain relatively small for a certain number of generations thereafter, increasing the total drift effect over and above that resulting from the small number of founders. Many rare alleles are usually eliminated in these processes because they were either absent among the founders or were lost in later generations. If founders were few or there were strong bottlenecks, rare alleles that happen to be present at the beginning and survive until later will eventually be found at higher frequencies in comparison with other populations. It is especially easy under these conditions to find that one or more genetic diseases have become very frequent in the island or "isolate," whereas others have completely disappeared.

The colonization of large new territories, including those of uninhabited or scarcely inhabited continents or large regions, are of special importance for the history of human evolution. When conditions are favorable to the growth of the migrants and exchanges with the parental population are limited or absent, these colonizations are major examples of *fissions*. In this process a sample from the parental population generates, by a sort of budding, a new population that expands to occupy the space available to the degree of saturation compatible with the new environment and the available food technologies. When aborigines were present in sufficient numbers in the areas to which migrants were directed, there could be impingement, intermingling, or the formation of new population boundaries between earlier settlers and newcomers on account of linguistic, geographic, and ecological isolation. Often relics of more or less unmixed aborigines may remain for as long a time as isolates in refugia.

Clearly, mass migration offered many chances for the formation of new groups that became separated from the original population and had the opportunity to begin diverging from it. Often this has brought in contact groups showing substantial genetic differences, which may have maintained their individuality but may also have exchanged immigrants in one direction or both. When migration takes place prevalently in one direction (from one group to another) it is often referred to as *gene flow*.

It is characteristic of both individual and mass migration that all genes are equally interested in the exchange. It is also true of drift that all genes are affected with equal intensity, but each in an unpredictable direction, whereas in migratory exchanges all genes are affected in parallel, predictable ways. One can evaluate the extent of gene flow into a population when the gene frequencies of the parental populations are known or can be estimated with reasonable assurance.

Technological innovations have frequently determined local population growth. Sometimes the innovations (e.g., new food technologies or easily produced weapons) may have rapidly diffused to neighbors, affecting them in a similar way, and little if any genetic variation was determined. But if such innovations were complex and depended on specific social structures, their diffusion under exclusively cultural contacts may have been slow or impossible. Once the growth of the new population oversteps the limits of local population density compatible with the new conditions, it usually determines outside migration, often to the nearest possible place. Cycles of local demographic increase followed by migration can, in the long run, cause major geographic expansions of an initially small population. Such processes generate *population expansions* comparable to those that can take place in the occupation of uninhabited, or less densely inhabited, continents or large areas. All these processes have had an important role in the construction of the present genetic picture of populations and will be discussed in more detail below (sec. 2.7). Under these different mechanisms of evolution, gene frequencies have varied from population to population, and one of our tasks is to use modern geographic and ethnic variation for the purpose of reconstructing various aspects of this history. Using terms common in anthropology, the problem is that of inferring diachronic variation on the basis of the synchronic variation. Obviously, there are limits to the extent to which this is possible in the absence of fossil data. Continuous cross-checking with independent sources of evidence is the best insurance that conclusions are correct.

1.5. Classical attempts to distinguish human "races"

The study of human "races" dates to antiquity. The existence of conspicuous differences between humans of diverse geographic origins must have been a familiar sight to the first long-distance travelers. The father of Greek history, Herodotus (fifth century B.C.), gives the name, geographic location, customs, and physical appearance of a great number of people, mostly around the Mediterranean. He is the father not only of history, but also of anthropology (Myres 1953). Sometimes the information he gives us is clearly legendary or mixed with tales and superstitions, but at other times he has been vindicated by modern archaeology. When he lists the ethnic groups contacted by the Greek traveler Aristeas at the extremes of the Central Asian steppes, one is tempted to recognize proto-Mongolian nomads in some of them.

Egyptians and Phoenicians were certainly aware of the existence of sub-Saharan African populations. The Roman empire was in contact with Africans, Indians, and indirectly, by trade, with East Asians. The Roman naturalist Pliny the Elder (first century A.D.) had a naive explanation of the physical differences between Africans and Europeans, which he thought were a direct consequence of the climate. The Roman poet Lucretius (first century B.C.) had a more subtle approach to evolution that anticipated the idea of natural selection. But for Pliny, Africans are "burnt by the heat of the heavenly body near them, and are born with a scorched appearance, with curly beard and hair"; while in the north, being far from the sun, "the races have white frosty skins, with yellow hair that hangs straight" (Rackham 1979).

Although taxonomic ideas and examples, biological or not, go back mostly to Aristotle (fourth century B.C.), serious attempts at a classification of human races had to wait for substantial geographic knowledge. This became

common only in the eighteenth century when interest in the classification of animals and plants was already flourishing. A definition of the *species* of living beings was given by John Ray (1627–1705) and is basically the same that we follow, namely, a group of individuals that can interbreed.

One of the first naturalists who discussed human variation, the Frenchman George Leclerc comte de Buffon (1707–1788), was a pioneer evolutionist whose work is said to have influenced Lamarck. Buffon used a definition of species very similar to Ray's, but he probably reached it independently. He clearly stated his conviction that humans are a single species and,

> after multiplying and spreading over the whole surface of the earth, they have undergone various changes by the influence of climate, food, mode of living, epidemic diseases, and the mixture of dissimilar individuals. At first these changes were not so conspicuous, and produced only individual varieties; these varieties became afterwards specific, because they were rendered more general, more strongly marked, and more permanent by the continual action of the same causes; they are transmitted from generation to generation, as deformities or diseases pass from parents to children.

Many citations like this one and the ones below are found in Count (1950).

Lists of races, or varieties as they were called by Linnaeus (1707–1778), appear in the eighteenth century with Linnaeus and with Kant (1724–1804), who also made various hypotheses on their mechanism of origin. Kant's hypotheses are unconvincing today, but the philosopher acknowledged their futility in the lack of adequate knowledge. J. F. Blumenbach (1752–1840), considered the father of physical anthropology, exercised great influence with his doctor of medicine thesis from the University of Göttingen (Blumenbach 1775). He stated that the human species is one, with five varieties: Caucasian (he might have been the first to use this term), Mongolian, Ethiopian (including all Africans), American, and Malay (including the islands of Southeast Asia and the part of Oceania then known). At that time skin color, the most conspicuous of all traits, had the dominant role it still has in the layman's mind. He defined Caucasians as we define Caucasoid today, including Europeans, North Africans, and people from the Near East and India. He did not, however, include Lapps and Finns, whom he assigned to Mongols. He stated that he chose the name of this variety from the Mount Caucasus, on the basis of what one might call a poetical motivation, because of the widespread belief that this region harbors the most beautiful people, like the Georgians who live in the southern part of the Caucasus. He also considered this area the likely origin of modern humans and followed Buffon in regarding white as the original color of the human species.

Trying to get away from the ever-present and obviously unsatisfactory criterion of skin color, the Swedish anatomist Anders Retzius (1796–1860) showed it was possible to generate a classification of races using craniometric criteria. Retzius invented the *cephalic index*; the ratio of the width to the length of the skull. This measurement enjoyed tremendous success in physical anthropology for a century, until the advent of multivariate analysis and genetic markers after World War II. Its popularity was tied to the simplicity of measurement both in living individuals and in skulls, including fossil ones, and to the superficial impression of precision it conveys. Ideas of biometry were introduced at that time by the Belgian scientist, Adolphe Quetelet (1796–1874). After World War II, interest in the cephalic index essentially disappeared because its heritability is probably low and because the index is sensitive to short-term environmental effects.

In the early nineteenth century, other means of distinguishing human races were suggested, and some contributors also argued against complete interfertility, challenging the idea that there is only one species. The summary of Charles Darwin (1809–1882) in *The Descent of Man, and Selection in Relation to Sex* (Darwin 1871) is especially illuminating. He noted arguments given by others for and against full interfertility in humans (we have no doubt today that there are no limitations in humans to interfertility). In the face of contrasting evidence existing at the time, Darwin concluded nevertheless that the species is likely to be one because all the races "graduate into each other"; moreover, "the races of man are not sufficiently distinct to inhabit the same country without fusion; and the absence of fusion affords the usual and best test of specific distinctness." He also specified that, however conspicuous the differences between races, they are mostly unimportant, because for most important traits, including mental ones, there is much similarity. In spite of the external difference between American aborigines, Negroes, and Europeans he was "incessantly struck . . . with the many little traits of character showing how similar their minds are to ours." Concerning classification problems, Darwin cited 12 authors who all disagreed on the numbers of races, giving numbers that vary from 2 to 63; he cited this disagreement as further evidence that "it is hardly possible to discover clear distinctive characters" between races, because they "graduate into each other."

As to the origin of variation, Darwin believed that "the external characteristic differences between the races of man cannot be accounted for in a satisfactory manner by the direct action of the conditions of life; the differences between the races of man, as in color, hairiness, form of features, etc., are of a kind which might have been expected to come under the influence of sexual selection." It is noteworthy and unfortunate that very little research has been done in humans on the evolutionary consequences of the choice of mates.

The American anthropologist Franz Boas (1858–1942) was among the first to throw considerable doubt on the

evolutionary stability of quantitative phenotypic variation like stature, limb measurements, and in general most anthropometric traits. In a classic work (Boas 1940) that compared physical characteristics of children of immigrants to the United States with those of relatives who did not migrate, he showed the magnitude of short-term environmental effects. As was almost inevitable at that time, his work was statistically weak. In any case, confidence in anthropometry remained unshaken for a long time and may still be strong in the most conservative quarters. The magnitude of short-term environmental effects is well documented, and there also exist slow environmental changes, the physiological effects of which are difficult to test, but which throw considerable doubt on genetic interpretations of phenomena like the recent secular trend for an increase in stature in Europe and other parts of the world. Nonmetric variation of bones has recently become popular, but evidence that it is determined by genotype and is insensitive to short-term environmental change is still far from adequate.

We believe that the major breakthrough in the study of human variation has been the introduction of genetic markers, which are strictly inherited and basically immune to the problem of rapid changes induced by the environment. One should not expect, of course, that in the long term they show complete stability; otherwise, there would be no evolution. The nature and the dynamics of the major forces that mold the frequencies of genetic markers are well understood: natural selection (including also sexual selection), mutation, migration, and chance. Chance is effective in two ways: (1) because mutations are rare and random, the occurrence of a specific mutation at a particular point of time and space can be considered a chance event; (2) random genetic drift is another strictly indeterministic process.

The pioneers of this approach, Hirszfeld and Hirszfeld (1919), showed (fig. 1.5.1) the differences in frequencies of A and B blood antigens in different ethnic groups, sampled from the armies of World War I. They proposed a biochemical index to differentiate populations on the basis of the two antigens. Starting in the 1930s, American immunologist W. Boyd used information on gene frequencies of ABO and the other blood groups then known (MN and P; RH became known in 1940) for reconstructing the evolutionary history of humans and the differentiation of races (Boyd 1950). Boyd and others also started looking for ABO antigens in mummies, research that met with criticism because of the possible contamination with related bacterial antigens and destruction by specific bacterial enzymes.

The theoretical contributions of R. A. Fisher (1890–1962) to the understanding of the structure of RH (see Race and Sanger 1975) and his interest in evolutionary applications stirred considerable activity on blood-group studies in Great Britain. The person who emerged as the major student of human evolution through genetic markers was Arthur Mourant, who was instrumental in improving the quality of research on populations, thanks to his expertise in genetic hematology, his typing of many interesting ethnic groups, and his publication of the first modern tabulation of gene-frequency data with an evolutionary interpretation (Mourant 1954). This book was followed by three more volumes, which have been a major source of information (see sec. 1.9). A tabulation of gene frequencies of immunoglobulin polymorphisms by Steinberg and Cook (1981) was a useful complement. In addition, a recent tabulation of genetic data from a selected number of human populations, with some geographic displays, has been published by Nei and Roychoudhury (1988). We closed our data collection in the summer of 1986 and were therefore unable to use new information from this book.

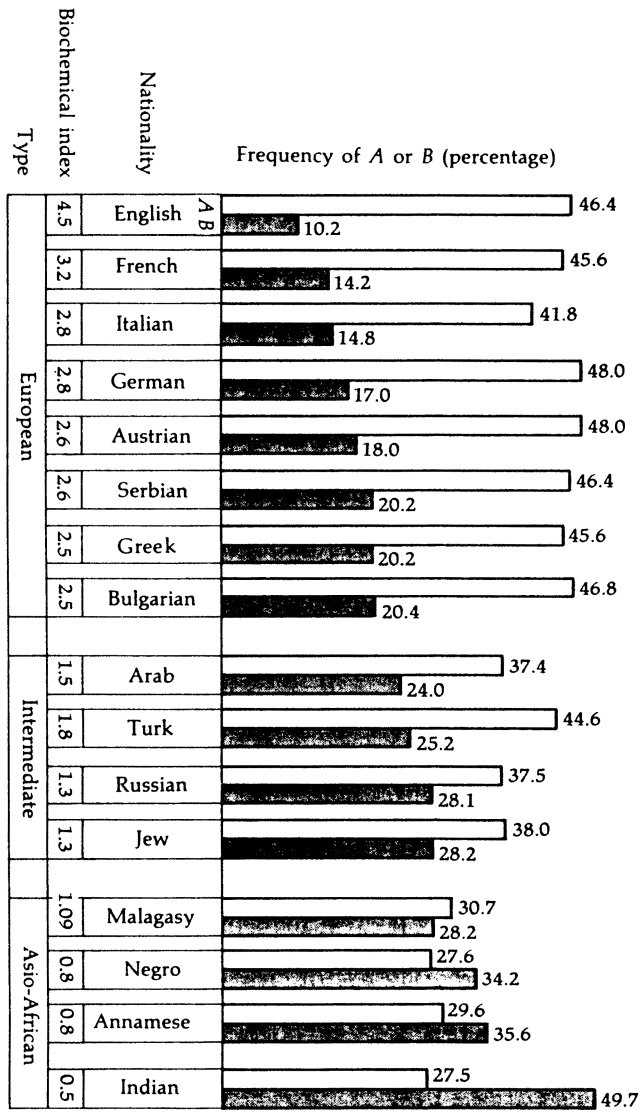

Fig. 1.5.1 This graph showing ethnic differences in *ABO* gene frequencies was the first use of genetic markers to study racial differences. The figures given are percentages of positive reactions with anti-A and anti-B reagents. The "biochemical index" is the ratio of *A* to *B*. (Taken from Hirszfeld and Hirszfeld [1919, pp. 505–537] by Bodmer and Cavalli-Sforza [1976, p. 576].)

1.6. Scientific Failure of the Concept of Human Races

The classification into races has proved to be a futile exercise for reasons that were already clear to Darwin. Human races are still extremely unstable entities in the hands of modern taxonomists, who define from 3 to 60 or more races (Garn 1971). To some extent, this latitude depends on the personal preference of taxonomists, who may choose to be "lumpers" or "splitters." Although there is no doubt that there is only one human species, there are clearly no objective reasons for stopping at any particular level of taxonomic splitting. In fact, the analysis we carry out in chapter 2 for purposes of evolutionary study shows that the level at which we stop our classification is completely arbitrary. Explanations are statistical, geographic, and historical. Statistically, genetic variation within clusters is large compared with that between clusters (Lewontin 1972; Nei and Roychoudhury 1974). All populations or population clusters overlap when single genes are considered, and in almost all populations, all alleles are present but in different frequencies. No single gene is therefore sufficient for classifying human populations into systematic categories.

As one goes down the scale of the taxonomic hierarchy toward the lower and lower partitions, the boundaries between clusters become even less clear. The evolutionary explanation is simple. There is great genetic variation in all populations, even in small ones. This individual variation has accumulated over very long periods, because most polymorphisms observed in humans antedate the separation into continents, and perhaps even the origin of the species, less than half a million years ago. The same polymorphisms are found in most populations, but at different frequencies in each, because the geographic differentiation of humans is recent, having taken perhaps one-third or less of the time the species has been in existence. There has therefore been too little time for the accumulation of a substantial divergence. The difference between groups is therefore small when compared with that within the major groups, or even within a single population. In addition, our species and its immediate predecessor, *Homo erectus*, showed considerable migratory activity in all directions, some of which are likely to have resulted in admixtures between branches that had separated a long time before. Whatever genetic boundaries may have developed, given the strong mobility of human individuals and populations, there probably never were any sharp ones, or if there were, they were blurred by later movements. There may still exist weak genetic boundaries in some regions, but they only mean that there has been less local admixture across certain barriers. For instance, Barbujani and Sokal (1990; Sokal et al. 1988) have found a number of weak genetic boundaries in Europe linked with geographic, ecological, and linguistic differences (see chap. 5).

From a scientific point of view, the concept of race has failed to obtain any consensus; none is likely, given the gradual variation in existence. It may be objected that the racial stereotypes have a consistency that allows even the layman to classify individuals. However, the major stereotypes, all based on skin color, hair color and form, and facial traits, reflect superficial differences that are not confirmed by deeper analysis with more reliable genetic traits and whose origin dates from recent evolution mostly under the effect of climate and perhaps sexual selection. By means of painstaking multivariate analysis, we can identify "clusters" of populations and order them in a hierarchy that we believe represents the history of fissions in the expansion to the whole world of anatomically modern humans. At no level can clusters be identified with races, since every level of clustering would determine a different partition and there is no biological reason to prefer a particular one. The successive levels of clustering follow each other in a regular sequence, and there is no discontinuity that might tempt us to consider a certain level as a reasonable, though arbitrary, threshold for race distinction. Minor changes in the genes or methods used shift some populations from one cluster to the other. Only "core" populations, selected because they presumably underwent less admixture, confer greater compactness to the clusters and stability to the classification tree. Although the hope of producing a good taxonomy is a lost cause—a minor scientific loss—that of reconstructing evolutionary history retains full strength and has the advantage that hypotheses can be tested on the basis of other, independent sources of data. Greater confidence in the conclusions must come from agreement with external sources of relevant evidence rather than from internal analysis.

The word "race" is coupled in many parts of the world and strata of society with considerable prejudice, misunderstanding, and social problems. Xenophobia, political convenience, and a variety of motives totally unconnected with science are the basis of racism, the belief that some races are biologically superior to the others and that they have therefore an inherent right to dominate. Racism has existed from time immemorial but only in the nineteenth century were there attempts to justify it on the basis of scientific arguments. Among these, social Darwinism, mostly the brainchild of Herbert Spencer (1820–1903), was an unsuccessful attempt to justify unchecked social competition, class stratification, and even Anglo-Saxon imperialism. Not surprisingly, racism is often coupled with caste prejudice and has been invoked as motivation for condoning slavery, or even genocide. There is no scientific basis to the belief of genetically determined "superiority" of one population over another. None of the genes that we consider

has any accepted connection with behavioral traits, the genetic determination of which is extremely difficult to study and presently based on soft evidence. The claims of a genetic basis for a general superiority of one population over another are not supported by any of our findings. Superiority is a political and socioeconomic concept, tied to events of recent political, military, and economic history and to cultural traditions of countries or groups. This superiority is rapidly transient, as history shows, whereas the average genotype does not change rapidly. But racial prejudice has an old tradition of its own and is not easy to eradicate.

1.7. Identifying Population Units

Although, in principle, the individual can be the unit of evolutionary study, this requires testing every individual for a large number of genes, and the amount of information generated soon becomes prohibitive. This approach has been attempted thus far only in studies of mitochondrial DNA (Brown 1983; Johnson et al. 1983) (discussed in sec. 2.4). The development of simple sequencing techniques has permitted great expansion of this approach, but a considerable increase in the efficiency of present methods of statistical analysis is still required. Moreover, trees developed using this method give a different sort of information from the population trees that we examine in this book. They supply mutational histories and refer to events that are not the same as the populational events that we follow here.

A rigorous analysis should start with a definition of the "population" to be sampled; in practice, one deals with samples that have already been collected and tested so that one is limited to deciding whether a sample is acceptable. A "Mendelian population" is one formed by individuals who mate randomly (see, e.g., Cavalli-Sforza and Bodmer 1971a). The "Hardy-Weinberg" rule (HW), briefly outlined in a preceding section, predicts the distribution of observed phenotypes in a diploid organism like the human one, given random mating, gene frequencies, and dominance rules, if any. It is well known that HW is extremely robust and the great majority of observed samples satisfy it. Deviations usually arise from one of four sources:

1. The genetic model (postulated alleles and dominance relationships) is incorrect.
2. Laboratory procedures or testing reagents are not always adequate.
3. Natural selection eliminates some phenotypes preferentially (an example is the effect of malaria mortality when testing the distribution of phenotypes for sickle-cell anemia among adults).
4. The population sample is heterogeneous, being made of socioeconomic and geographic strata that do not mate randomly with each other and differ in their gene frequencies.

The test of deviation from HW is made by the statistical index of goodness of fit, χ^2. When this is greater than a predetermined amount, the deviation from HW is considered statistically significant, but the numerical value of χ^2 is proportional to the number of individuals forming the sample. If there is a true deviation, the larger the sample, the more likely the deviation is significant. Large samples are thus inevitably more likely to show deviations, and many smaller samples, which apparently satisfy HW, may be actually as unsatisfactory as the larger ones, but have a smaller chance that their deviation from HW is statistically significant. Very large samples seem less satisfactory from the point of view of fitting HW expectations because they are almost always made up of blood donors, large numbers of which have been tested for blood groups important for transfusion. These large samples are inevitably drawn from a large area and thus come from a population that is more likely to be heterogeneous. An evaluation of the heterogeneity between samples of the same geographic or ethnic origin can alert us to the existence of such problems. For this reason, we have in certain cases avoided the use of sample sizes as weights when calculating mean gene frequencies. Further information on the application of Hardy-Weinberg can be found elsewhere (Cavalli-Sforza and Bodmer 1971b; Bodmer and Cavalli-Sforza 1976b, c).

A deviation from HW can determine a systematic error in the estimation of gene frequencies in the case of dominant genes, for example, in the *ABO* and *RH* systems. This error may be relatively important when comparing populations that are genetically close but often becomes trivial when comparisons are between very different populations, as is practically always the case in our work. However, it is difficult to set a nonarbitrary, objective threshold at which to include or exclude a gene-frequency estimate, based on the relevant HW χ^2. We have therefore made only rare exclusions on this basis.

The definition of Mendelian populations based on *panmixia* (general random mating within a population)—and hence the test of HW equilibrium—are, in practice, of limited use for our purposes. If applied to whole populations rather than to population samples, the HW test would probably prove that many populations are heterogeneous when they come from towns of large size. It is difficult to guess at what level this would be perceptible, but one cannot, of course, set a size threshold for this phenomenon, since the chance of significant heterogeneity will increase continuously with town size. The

redundancy internal to any population or sample that includes several members of the same family is one cause of increase of χ^2 above random expectation and is not easy to take rigorously into account. In addition, the tendency of marriages to occur at a short distance between places of residence or of birth tends to generate some geographic heterogeneity unless the area investigated is extremely small. Socioeconomic heterogeneity and other strong barriers to free interchange can also create stratifications almost within each population studied. Thus, validity of the HW test is not crucial to investigations making very broad comparisons. Most populations have some internal genetic heterogeneity, even if it is not detectable in every case, and it is of little importance for our purposes. In the geographic maps that we present, heterogeneity between populations occupying the same geographic locality is indicated and evaluated at an opportune significance level. (All the symbols used on the geographic maps are explained in table 1.14.1.) Local heterogeneity is tested under the hypothesis of binomial random sampling, and this test is very sensitive with large samples. It thus may respond to minor differences that are most frequently trivial compared with the genetic distances between populations that we are interested in studying. The explanation of these heterogeneities is usually sought in the coexistence, in the same small area, of different ethnic groups that do not undergo frequent genetic exchange.

These difficulties are sometimes not easily understood by people unfamiliar with human populations, their demography, and population genetics (Bateman et al. 1990a; Cavalli-Sforza et al. 1990). The term *deme* has been commonly employed to indicate a population unit that is panmictic and receives specified proportions of migrants from other specified populations. Except for a very few human populations, one cannot give an operationally useful definition of a deme, for reasons similar to those that make it difficult or impossible to define races. Demes are of course much smaller in size than races, but the continuity of the variation of gene frequencies and of mating distances at the geographic scale of demes is even more extreme than for races (Cavalli-Sforza 1958, 1963, 1986a, b).

It may be worth mentioning some of the reasons that make it difficult to define human demes. Candidates could be ethnographic units (e.g., tribes) or geographically defined clusters of people (villages, towns, cities). They are all usually endogamous to some degree and may come closer to the definition of a deme, but there are always many possible, embarrassing choices. Many tribes have undergone extraordinary demographic expansions (e.g., in Nigeria) and are subdivided in complex ways. In tribal as in modern society, the choice of mates is largely dictated by geographic, socioeconomic, religious, ethnic, and other constraints. The gene pool is therefore subdivided and stratified in very complex ways. Moreover, especially in Africa, most villages are made of several ethnic groups, and even most smaller tribes are spread over many villages (for a rare demographic investigation of a farming population in Africa, the Ngbakas of the Central African Republic, see Thomas 1963). In modern society most of the farming population may be very sparse (e.g., in the Po Valley of northern Italy; also in agricultural regions of the United States), with many people living in isolated houses near the farms. The farming population, once 90%–95% of the population, has decreased enormously (to 10%–15% or less) with the modernization of the economy. At present, the population tends to conglomerate in big cities with extremely complex patterns of residential segregation. The effect of geographic distance on the probability of marriage is well known and applies, though in different degrees, to all populations but is usually an incomplete description of the patterns of population distribution. Changes in means of transportation and labor opportunities have of course altered profoundly marriage customs: Dahlberg used the term "breakdown of isolates" to indicate this phenomenon, which includes a fall in consanguineous marriages and an increase in geographic distance between birth places of mates (see also Cavalli-Sforza 1957, 1958, 1963, 1986b; Cavalli-Sforza and Bodmer 1971; Bodmer and Cavalli-Sforza 1976b; Cavalli-Sforza and Hewlett 1982; Wijsmann and Cavalli-Sforza 1984).

From a practical point of view, a definition that makes it possible, if necessary, to obtain another sample from the same population is the major requirement for statistical validity. The indication of the population sampled by an earlier research worker is usually sufficient for this purpose. Even in the case of the American Indian populations that were most thoroughly investigated, the Yanomama and the Makiritare, one cannot expect a new sample to satisfy conditions of a good statistical sample, that is, to be within binomial sampling error. The genetic heterogeneity between villages is very high (Ward and Neel 1970) and their genetic composition is unstable in time and space. The source of much of this extreme heterogeneity is that internal population rearrangements take place by partitions that tend to follow lines of kinship. In general, because of the internal genetic heterogeneity of every population, one might find some variation between different samples, but most of the time it will not be embarrassingly large as with the Makiritare or Yanomama.

The details of the geographic or ethnic origin of the published population data given in the original papers varies considerably and may pose some practical difficulties. Generally, a gene frequency refers to a population occupying a geographic area that is sometimes poorly defined. A number of published samples, fortunately small, is given with very little detail (for instance, the population of origin may be defined simply as Australian aborigines or Africans). When other, better-

defined samples exist for the same genes, the poorly defined ones have been omitted. Even the indication of nationality—often the only information available—is not satisfactory, since many countries have substantial internal ethnic variation. When it was necessary to include any such poorly specified sample, it referred to the capital.

Populations defined as mixed, without giving details, were systematically excluded. When the original paper gave no information on this point, strong internal evidence of admixture with other widely different ethnic groups coming from the distribution of markers, was considered sufficient reason for discarding the population (for instance, the presence of nontrivial numbers of A blood-group individuals or of sickle-cell anemia in Central and South American Indian populations). We considered it important to be conservative in this respect, as one might easily bias the estimates if one were too extreme in applying this criterion. We retained populations that had up to 10% admixture, rarely more, but unfortunately there are few markers that allow a reliable diagnosis of admixtures from internal evidence.

Multivariate analysis posed more difficult problems. Here there is a need for finding the maximum possible number of genes for each population, and many authors tested only a small set of markers. It was therefore necessary to provide a reliable method of pooling populations. There was a smaller chance of being wrong when the same population name was used in different studies, even if some confounding is certainly introduced by pooling different samples of the same population collected by different authors. Especially with small tribes, however, there was frequently a necessity of using somewhat wider definitions of populations, allowing the pooling of tribes, in order to reach a satisfactory number of markers. Few populations have been tested systematically for a large number of markers.

Our database contains 76,676 gene frequencies. They correspond to 6633 samples with different geographic locations and to 1915 different population names. These numbers posed a gigantic problem of classification and there would be little to be gained from physical anthropology classifications if any existed.

Our main criterion in pooling populations for generating higher categories was geographic, but it was clear that, especially for populations from the developing world, the geographic criterion had to be supplemented with general anthropological information of some kind because populations of widely different origins occasionally live only short distances from one another. We decided to resort to linguistics when other criteria failed since it is increasingly clear that there is a certain amount of parallelism between the linguistic and genetic evolution of populations. This parallelism is certainly incomplete, however, and there are many well known exceptions. On the one hand, the use of a linguistic code of classification of our populations offered the possibility of giving us a chance to test further the genetic-linguistic parallelism and the deviations from it, a problem of interest per se. On the other hand, the pooling of populations on the basis of linguistic association offered an additional criterion of grouping, which we used within groups defined by geographic and other classical ethnic criteria. It is important that linguistic classifications also usually follow geographic criteria, so that the two go hand-in-hand, but the linguistic criterion is usually finer and more often attuned to ethnic differences, especially in developing countries. In fact, tribal names are very often the same as those of languages. The linguistic code is discussed in the following section.

By this process of pooling, we reduced the initial 6633 samples, most of which had been tested for very few genes (alleles), to exactly 491 populations at the lowest level of clustering. This reduction involved both culling and pooling. These populations are listed in Appendixes 2 and 3, their gene frequencies and ethnic composition are given, and a bibliography is provided. Many of the 491 "populations" appear in the analyses of each region in chapters 3–7, which are dedicated to single continents. A further selection from the 491 populations was carried out for analyses at the level of whole continents, always excluding populations with the smaller number of genes. For the purpose of further increasing the average number of genes in the analysis of the whole world, a final reduction of the number of populations was made by culling populations that had too few genes and no affine groups with which they could be pooled, and pooling those that showed high affinity. With this second cycle of culling and pooling, the number of populations decreased to 42 (with 120 independent alleles), but still with a nonnegligible number of gaps. Although some sacrifices had to be made, the sample of 42 populations was still reasonably compatible with the desire to adequately represent the whole world. Their analysis is described in chapter 2. The gene-frequency data of the 42 populations appear in Appendix 1.

1.8. Linguistic Classification

The most recent linguistic classification lists 4736 languages (Ruhlen 1987). This indicates a million speakers per language, on the average, but a handful of languages are spoken by hundreds of millions of people, the great majority by tens of thousands of individuals, and many by only a few hundred or less than one hundred. Lan-

guages in the last category are likely to become extinct in a few generations. The same fate has already befallen a number of others, some of which were studied before their disappearance. Except for the very few widely spoken languages, there tends to be a one-to-one correspondence of tribal names to language names. Thus, except in the case of large modern nations in which the identity of original tribes is usually—though not entirely—lost, languages offer a powerful ethnic guidebook, which is essentially complete, unlike strictly ethnographic information. Moreover, there exist phylogenetic classifications of languages, which in linguistics jargon are called "genetic," and they supply a partial taxonomic hierarchy from which we could build a linguistic numerical code. Naturally such a code should not be used automatically for a biological classification, but it was a convenient point of departure that was modified on the basis of other information.

An important consideration is that modern linguistic classifications recognize major groups or *phyla* (also called families). Leaving aside a few isolates, which, although reasonably well studied, cannot be classified in the existing taxonomy, the few unclassified languages, the recent hybrids (pidgins and creoles), and the invented languages like Esperanto, there are 17 major taxonomic groups listed by Ruhlen. A list of these 17 phyla is given in the next chapter. Thanks to Dr. Ruhlen, we had access to his classification while it was still unpublished. Books like *Classification and Index of the World's Languages*, and the *Ethnologue* (Voegelin and Voegelin 1977; Grimes 1984) were also very useful for synonymies and geographic information. Grimes was especially useful for demographic census data. Some information was also obtained from the *Encyclopaedia Britannica* (1974) as well as government publications for the USSR and China (*The USSR in Figures for 1986* and *China Handbook Editorial Committee* 1985).

The code we eventually adopted for classifying our populations is geographic–anthropological (physical)–linguistic–ethnographic, the order of the four words reflecting the average importance of each criterion in making decisions in uncertain cases. On the basis of this code, we classified all gene-frequency data into a four-tier hierarchy: the continent, major classes within the continent, and two lower hierarchical levels. The third tier included clusters of very unequal numerical importance and was designed to recognize major groups as well as special populations likely to deserve separate analysis. The fourth tier consisted of the 491 populations initially selected for multivariate analysis and consisted of tribes, countries, or regions of countries, depending on areas. Our code was therefore inspired by pragmatic considerations and does not necessarily correspond to subdivisions that eventually turned out to be entirely meaningful from a phylogenetic point of view, but it made it possible to easily reconstruct linguistic phyla and their major subdivisions.

Tribes do not necessarily correspond to Mendelian populations as defined above; the discrepancy between the two is especially important when tribes are numerically large. In fact, in large tribes, social stratification and geographic differentiation may be very pronounced and even relatively small tribes, when carefully analyzed, have shown internal heterogeneity. This is especially marked, for instance, in the case of the Yanomama and the Makiritare (Ward and Neel 1970). Supratribal, national aggregates and their geographic subdivisions are even worse in this respect. Apart from the heterogeneities that may arise in various circumstances, however, tribes, when not too large numerically, are a reasonable approximation to a population unit for the purpose of genetic analysis; in any case, there is usually no better choice. Because the ideal of a Mendelian population is difficult to attain, we consider the smallest subdivision available as a genetic pool of individuals who mate randomly or nearly so for most practical purposes of genetic analysis, recognizing that this unit may be larger than a Mendelian population but still offers the best practical compromise.

It is reassuring to note that the patterns of linguistic variation in space parallel those of genetic and/or geographic variation, as shown in a number of detailed studies on small regions (Sardinia, Piazza et al., in press; Micronesia, Cavalli-Sforza and Wang 1986), as well as on a wider scale (North America, Spuhler 1979; Europe, Sokal et al. 1988; Barbujani and Sokal 1990; Central America, Barrantes et al. 1990). There are in fact good a priori reasons why cultural and genetic pools have close similarities: both genetic and cultural contacts take place by the same routes; they respond to the same geographic and ecological barriers; and they also can influence each other, in the sense of mutual reinforcement. For example, take a tribe as an imperfect example of a cultural pool; a tribe is frequently endogamous, a property that fits to some extent the idea of a genetic pool. At a more general level, the constitution of a genetic pool is determined by geographic factors, socioeconomic distance, and a variety of cultural factors (religious, linguistic, etc.), all of which also operate on cultural pools and affect them in a parallel way. Although investigations of the joint effects of all these variables would be very interesting, there do not seem to be any. It seems likely that two individuals have a higher probability of marrying if their distance in any of these scales is shorter.

Important correlations are thus created between genetic pools on one side and sociocultural pools on the other. There are limitations, however, to the parallelism of linguistic and genetic evolution. Languages evolve much faster than genes; two languages may become mutually unintelligible in a thousand years or less because of progressive differentiation. Formally, this is similar

to the origin of two different species in biology. Speciation involves the loss of interfertility, in some measure the genetic equivalent of the loss of communication, but speciation takes on the order of a million years. Moreover, a language can be replaced by an entirely different one in as little as three generations as a result of political events leading to domination by a new people. By contrast, the genetic changes accompanying the replacement of a language, usually by invasion followed by imposition of the language of the new masters, may be difficult to detect genetically because the new masters are often numerically a small fraction of the whole population they dominate. It is also possible that extensive gene replacement has occurred through prolonged contact and gene flow from neigbors, without language change. One thus expects, and finds, minor and major inconsistencies in the comparison of genes and languages.

Linguistic analysis does not cease to be useful in the analysis of large national and supranational aggregates in which a common language is spoken. The distinction, however, may have to be drawn at the level of dialects and is inevitably more subtle. The simplest method of measuring the similarity of dialects, or of languages that are not too widely separated, consists of evaluating the proportion of words that clearly have a common origin, even though they may have undergone some phonological or semantic change. Such words are called *cognates*. The percentage of words that are cognate in two languages (or dialects) is a measure of the languages' similarity and also of the linguistic affinity of the corresponding populations. Attempts at correlating the similarity (or its converse, the distance) between two languages with their time separation (glottochronology) have been only partially successful.

Dialects of a language should show, on the average, smaller reciprocal distances than languages. The transition from dialect to language is, however, continuous and there is a gray area in which designating two forms of speech dialects or languages is arbitrary.

1.9. Nature and Sources of the Data

We have confined our analysis to aboriginal populations that were in their present location at the end of the fifteenth century when the great European migrations began. We have thus excluded Black Americans and all the recent colonizations of Caucasoid, Chinese, and Indian origins. We have also excluded all manifestly mixed populations; those stated to have 25% or more external admixture; all populations from Israel, for reasons already stated in section 1.2; and various isolates, including migrant groups like Gypsies. We have also usually excluded populations classified as living "abroad," for which there is only a vague indication of origin.

All data from the major compilations cited in the Appendix that satisfy the above definitions have been included in the data bank, irrespective of sample size (excluding sizes below 50 individuals, with very few exceptions for geographic areas where data were extremely rare) and of Hardy-Weinberg χ^2. As already discussed, the rationale behind this last decision is that a usually small heterogeneity is expected in the majority of the samples, but is not usually found because the great majority of samples are of small size, and hence unlikely to show heterogeneity. The error in estimating the frequencies of dominant genes because of this heterogeneity is likely to be small and unbiased compared with the differences between populations that we study. It would be difficult to set an arbitrary general limit to χ^2 above which data are ignored.

Gene frequencies from populations with the same geographic coordinates were averaged (weighting by sample size), and heterogeneity χ^2's were calculated in each case. The indications of local heterogeneity that appear in geographic maps are based on these estimates.

Our joint work began as geographic analysis of genetic European data in 1977. After publication of our first paper on what we called synthetic geographic maps of Europe (Menozzi et al. 1978a), we decided to extend the work to the rest of the world. At that time, the extensive tabulation by Mourant et al. (1976a) had appeared, but it carried only data published up to 1972. An update was therefore necessary, and it was begun by a computer search using common key words and available retrieval software. In this way, we also identified journals most frequently employed for publishing articles of interest. These journals were systematically searched by hand.

When our work was fairly advanced, we became aware that Tills, Kopeć, and Tills were updating the tabulations by Mourant et al. (1976) from 1972 onward. These authors summarized the next 9–10 years of data, and eventually published it (Tills et al. 1983) in a format very similar to that used earlier by Mourant et al. (1976a), using the same numbers for tables of the same genes. These numbers are used in our own summaries for referring to data cited in these two books. Thanks to the courtesy of Drs. Mourant and Tills, we could obtain photocopies of Tills' updated tables several months in advance of their publication. We found that much of the material we had collected was duplicated in their work. Material for *GM* by Steinberg and Cook (1981), and for hemoglobins by Livingstone (1985) was also found to be partially duplicated.

Another important compilation of data by Roychoudhury and Nei (1988) lists 362 loci, including polymorphic and monomorphic ones, for a variable number of populations (a total of about 180 when data are available). They also include 50 world geographic distributions of gene frequencies and show their numerical frequency values. Unfortunately, this work was published too late for use in forming our data bank.

Various circumstances forced us to delay the beginning of our analytical work, and it therefore became necessary to update our own files. This was done by covering the period between the summers of 1982–1986 by systematic analysis of the journals that had proved richer in relevant articles in the previous search. The first search had used articles from a total of 136 journals, of which 12 provided 66% of all articles. In the second search, we used articles from the following 12 journals: *American Journal of Human Genetics*, *American Journal of Physical Anthropology*, *Annals of Human Biology*, *Annals of Human Genetics*, *Genetics* (Russian), *Human Biology*, *Human Genetics*, *Human Heredity*, *Japanese Journal of Human Genetics*, *Journal of Human Evolution*, *Tissue Antigens*, and *Vox Sanguinis*. This second period is therefore covered less thoroughly than the period before 1982. Altogether, more than 2900 articles were indexed by us, of which only 777 survived after eliminating duplications with other tabulations in books or reviews. These form the alphabetized list of references given in the Appendix, cited by numbers in the bibliography accompanying the list of populations.

An important question—but one that is difficult to answer precisely—is, how much information is missing from our data bank (or any other). It must vary considerably from country to country. The proportion missing may be high for several European countries, since many medical journals in languages other than English are not easily available in American or English libraries. For Russian—in particular, Siberian—data, Michael Crawford was especially helpful in personally requesting Russian colleagues, whom he visited before Perestroika, to send us reprints. Data from the Iberian peninsula were very thoroughly searched in Spanish and Portuguese journals by J. Bertranpetit; the number of these missing from our files was not small. However, data from Italy were easily available to us for obvious reasons, and the bias for this country is likely to be in the other direction.

Even if no files are complete, these are probably the most detailed available at the moment. They are limited to genes represented fairly widely. We hope to make the full files available to research workers in the future, in suitable computer form, if there are enough requests. It is also possible that the data and bibliography published in this volume may satisfy most needs. They contain the gene frequencies of the populations, sample sizes, names and locations of the populations as given by the author, and the latitude and longitude of the place of residence (or birth, when indicated) of the individuals forming the samples. The genotype and phenotype frequencies are not given in the files. When the place of origin indicated in the original paper is simply the country, the data were omitted if other data on the same gene and country were available with more precise places of origin. Otherwise, the data were included and assigned the geographic coordinates of the country's capital.

At the time of writing, we hope to update the data bank in the same format, although no specific plans have yet been made. In any case, we would appreciate it if errors and omissions were made known to any one of the authors.

1.10. Methods of Analysis

The data were analyzed according to a variety of methods, which are briefly defined here, and explained in a little more detail in the following sections where key references will also be given. Table 1.10.1 collects a number of formulas employed in the analysis.

Genetic distances (sec. 1.11). Genetic distances are used to measure the global genetic difference between two populations. There are many genetic distances; they are all highly correlated. In this book we have used almost exclusively one measurement of distance, which is explained in detail in the next section.

Trees (sec. 1.12). Trees are the most common method of phylogenetic analysis. A dichotomous tree is constructed by using a matrix of distances between all the possible pairs of n populations. In the intention of the authors who first suggested this approach for evolutionary purposes (Cavalli-Sforza and Edwards 1964, 1967; Edwards and Cavalli-Sforza 1964), the tree should represent the history of fissions (splits, separations) having taken place in the human species or parts of it. Possible sources of error demand internal and external controls for the historical interpretation for a sequence of fissions to be acceptable. There are many methods for reconstructing trees; some give rooted trees (the root or origin is the earliest dichotomy), some give unrooted ones. A root can be added to an unrooted tree on the basis of evidence internal or external to it; the latter is likely to be more robust. Rooted-tree methods force all branches from the origin to the populations being analyzed to be equal in length. This is a necessary consequence of the hypothesis of constant evolutionary rates underlying such methods. Unrooted trees are not bound by these constraints, but

Table 1.10.1. Summary of Major Formulas Most Frequently Employed in the Numerical Analysis

A. Homozygosity, Heterozygosity, and F_{ST}

For a given locus with L alleles, where p_{ij} is the gene frequency of allele i in population j,

$$\sum_{i=1}^{L} p_{ij} = 1 .$$

The homozygosity of population j is

$$H_j = \sum_{i=1}^{L} p_{ij}^2 ;$$

and the heterozygosity of population j is

$$h_j = 1\text{-}H_j .$$

The average gene frequency of allele i in a cluster of s populations is

$$\bar{p}_i = \sum_{j=1}^{s} p_{ij} / s .$$

Weighting by sample size,

$$\bar{p}_i = \sum_{j=1}^{s} n_j p_{ij} / \Sigma n_j ,$$

where n_j is the number of individuals in a sample of population j.

The heterozygosity of a population cluster is expressed as

$$h = 1 - \sum_{i=1}^{L} (\bar{p}_i)^2 , \qquad (1)$$

where subdivisions in s populations are ignored.

The average heterozygosity of s populations is:

$$h_s = \sum_{j=1}^{s} h_j / s . \qquad (2)$$

F_{ST}, a measure of variation of gene frequencies of populations, is calculated from

$$F_{ST} = (h - h_s) / h . \qquad (3)$$

F_{ST} can be rewritten for a single allele i as

$$F_{STi} = \text{var}\,(p_i) / [\bar{p}_i (1 - \bar{p}_i)] , \qquad (4)$$

where $\text{var}\,(p_i) = \sum_{j=1}^{s} (p_{ij} - \bar{p}_i)^2 / (s - 1)$.

Cumulatively, over all alleles at a locus,

$$F_{ST} = \frac{\sum_{i=1}^{L} \bar{p}_i (1 - \bar{p}_i) \cdot F_{STi}}{\sum_{i=1}^{L} \bar{p}_i (1 - \bar{p}_i)} .$$

Under drift alone, in a population of N individuals, F_{ST} increases with time as

$$F_{ST} = 1 - e^{-t/2N} . \qquad (5)$$

Hence, $-\log (1 - F_{ST}) = t/2N$. (6)

If t is small with respect to $2N$,

$$F_{ST} \cong t/2N .$$

B. F_{ST} Unbiased Genetic Distance (Reynolds et al. 1983)

F_{ST} for s populations (for the distance between two populations, $s = 2$) is

$$b = 2 \sum_{j=1}^{s} n_j h_j / s (2\bar{n} - 1) , \qquad (7)$$

(continued)