CHAPTER 1

History of This Investigation and Structure of This Book

1.1 INTRODUCTION

The study of inbreeding, the consequence of the mating of relatives, has an important place in genetics. The similarity of the paternal and maternal contributions caused by the mating of relatives leads to increased genetic homogeneity of inbred individuals. A table of the expected effects of inbreeding in successive generations of selfing (crossing with self), the closest mating possible, which often occurs spontaneously in many plants, appears in Mendel's article, the founding paper of genetics. Thus, Mendel was also the first population geneticist.

Human societies are unique in keeping records of their own ancestry, sometimes, though very rarely, for thousand of years or more. Some breeders of domestic animals, however, have kept records of their animals' ancestry for even longer periods, measured in terms of numbers of generations.

The genetic effect of inbreeding can be estimated by the increase in average homozygosity over that expected by random mating. *Homozygosity* is the average percent frequency of homozygotes—individuals receiving the same form of a gene from both parents. The complement to 100% of homozygosity is *heterozygosity*.

Matings of close relatives (often called consanguineous matings), if repeated for many generations, increase homozygosity to a point of seriously decreasing fertility and individual survival, making prolonged inbreeding incompatible with continuation of life. To obtain stocks of high genetic homogeneity, animal breeders often make re-

course to systematic parent–child or brother–sister matings for a number of generations (twenty or more). In the process they lose many inbred lines because of loss of fertility or survival, and at best end with stocks of weak or delicate constitution (inbreeding depression). There are examples of repeated brother–sister marriages in ancient Egyptian and Persian dynasties; but, for reasons probably not of genetic nature, no dynasty lasted for periods long enough to expect serious inbreeding depression. In any case we have no records from which to observe it. But apart from these examples of encouragement of brother–sister marriages, which was still popular in the Egypt of Cleopatra's time, unions between brother and sister and parent and child are condemned and avoided in practically all human societies. The term "incest" refers to these tabooed unions. They occur, though rarely. Children of incest are very rare, of the order of 1/10,000 in an estimate in Michigan (Adams and Neel 1967).

The dangers of close inbreeding must have become known to early humans fairly soon, since practically every society has rules that tend to avoid close consanguineous marriage, and sometimes extend prohibition to remote consanguinities. This is especially true of very small communities, like those of Eskimos living in extreme northern latitudes (Sutter and Tabah 1956), which are under greater risk of reaching high levels of inbreeding. In one remote, very small, and highly isolated Greenland Eskimo community, no marriage closer than third cousins was found in genealogies from living individuals.

Animals also tend to avoid close inbreeding, by social customs that seem to have been favored by natural selection in response to inbreeding depression. In most Primates, social groups are fairly permanent, but young males reaching puberty tend to leave the group into which they were born and join other groups. This custom is clearly effective in limiting close consanguineous matings. Chimpanzees are the only exception, as here it is the females that leave the group. Among humans one likewise observes a greater tendency of females to leave their birthplace at marriage. Wives tend to move to their husbands' residence in 70% of traditional societies (Murdock 1967). This custom has important genetic consequences: mitochondrial DNA (mtDNA), which is transmitted by the maternal line, should show less geographic clustering than Y chromosomes, which are transmitted by the male line. This expectation was confirmed by

observation (Seielstad et al. 1998). In fact, the Y chromosome shows greater genetic variation between populations than nonsexual chromosomes and mitochondrial DNA.

In addition, at least in chimpanzees and other mammals, some tendency to avoid brother–sister or parent–child mating is observed. Among humans, the social custom of avoiding marriage of close relatives is paralleled by a similar constitutional safeguard against brother–sister mating, in the form of the so-called Westergaard effect: a tendency to avoid sexual contacts between brother and sister, or, in general, children who have been brought up together. Puberty seems to be the dividing line between social contacts that are unfavorable (before puberty) or favorable (after it) to interest in establishing sexual relations between individuals of opposite sex. Research has shown that children brought up in the same kibbutz, where they were usually raised together, marry very rarely, if ever. An old Chinese custom, which survived in Taiwan until recently, is the adoption of a young girl by a family in which a son was born, so that this girl becomes his future wife. These so-called "minor" marriages have been shown to be, on average, less fertile and less long-lasting than ordinary marriages with girls not brought up in the family (Wolf 1980).

There are, however, social exceptions to the rule of avoidance of close consanguineous marriages, less close than brother–sister. In certain social groups such marriages may be much more popular than would be expected by chance, undoubtedly because of a social preference. In West and South Asia two types of consanguineous marriages are especially common: uncle–niece marriages comprise up to 20% of all marriages in several north Indian tribes, and first-cousin marriages reach 50% or more in many Middle Eastern ethnic groups (Arabs, North Africans, and some Jewish groups). First-cousin marriages are or were high in Japan, especially at a time when marriages were mostly arranged by parents. High consanguinity customs spread around with the people who developed them. Perhaps as a remote consequence of Arab domination in Sicily and southern Italy in the eighth to the eleventh centuries, the frequency of uncle–niece and first-cousin matings became high in these regions and is currently especially high in Sicily. These relatively moderate degrees of inbreeding do not seem to have had a truly damaging effect. They may

have contributed to lowering the current frequency of lethal and semilethal genetic diseases, at least in Japanese populations (Cavalli-Sforza and Bodmer 1971, 1999).

By definition, recessive genes are those that are manifested only in homozygotes. In inbred families increased homozygosity is expected, leading to a higher probability of observing recessive inheritance. The study of consanguineous marriages, therefore, has merit for the detection of recessive genes and for the study of their frequency.

In this chapter we summarize the salient points of the history of this investigation, which started in 1951 and is now coming to an end. We then briefly summarize the main properties of consanguinity, inbreeding, and drift, as well as inbreeding effects in humans and some special projects that were part of the investigation. Finally, we summarize the structure of this book.

1.2 HISTORY OF THIS RESEARCH

In 1951 Luca Cavalli-Sforza started teaching a course in genetics at the Faculty of Sciences of the University of Parma, Italy. Among his students was Antonio Moroni, a priest who is now professor of ecology at the University of Parma. At that time Moroni taught natural sciences at the Seminary of the Parma Archbishopric and made Cavalli-Sforza aware of records in the Roman Catholic archives that could be of interest for human genetics: essentially dispensations for consanguineous marriages and parish books of deaths, marriages, and baptisms. In an almost 100% Catholic country, baptisms are a close equivalent of birth records. Newborns are usually baptized very soon after birth, and a very small fraction, probably less than 1%, dies without a chance to be baptized. Moroni was also instrumental in obtaining permission from the higher religious authorities to use these records for genetic investigations. Our investigations began at the bishopric of Parma. They were soon extended to other bishoprics, and eventually to the whole country of Italy. To help with our investigations, a letter from the highest Catholic authorities was sent by the Vatican to all parish priests,* asking them to make parish records

* A letter dated 15 December 1960 by His excellency Monsignor Cesare Zerba,

available for scientific purposes. Genetic research using Roman Catholic records was also started in France by Jean Sutter and his collaborators, at about the same time as ours.

Consanguinity records are to be found in various Catholic archives. Consanguinity itself is very carefully defined and Roman Catholic legislation prescribes with great precision which marriages are completely forbidden, which ones are permitted under dispensation from a higher religious authority, and which do not require dispensation. Priests receive formal teaching about these rules in seminaries in which they also learn to evaluate accurately the degree of consanguinity of candidates for marriage. The need for dispensation has changed through time and is now reduced to a minimum. In earlier times even remote degrees of consanguinity were forbidden and it was essential to ask for dispensations before marriage could be celebrated. A consanguineous marriage celebrated without the requested dispensation would be null and void, generating a very serious social problem for the families.

Chapter 2, on the history of consanguinity regulations in the Roman Catholic Church, examines the historical knowledge available. There are also geographic differences in rules for obtaining dispensation. In peninsular Italy the parish priest must check every pair of prospective spouses for the existence of recent relationship, and if one requiring dispensation is discovered, he must request it from the bishopric. A copy of the request is then sent from the bishopric to the Vatican, and is returned to the bishopric with the Vatican response

---

Secretary of the Congregation for Sacraments, gave permission for the use of consanguinity dispensations held in the Archives of the Congregation of Sacraments and suggested that Italian bishops would allow the use of consanguinity dispensations held by Diocesan Archives, if the need to complete the investigation would arise:

> The Reverend Dr. Antonio Moroni, Professor at the University of Parma, intends to investigate the consequences of marriages among consanguineous persons and among minors that celebrated their wedding after receiving dispensation from their otherwise forbidding condition. The Congregation of Sacraments, believing that the results of the investigation will help better understand the Church norms that discourage the celebration of such weddings, grants permission to Reverend Prof. Moroni to have access to the needed data that are held in the Archive. For similar permission to access the Diocesan Archives he will also address their excellencies the Italian Bishops, who most certainly will not miss the social, pastoral, and juridical usefulness of the scientific investigation started by the University of Parma.

and then to the parish before a marriage requiring dispensation can be celebrated. In a few regions other than peninsular Italy, dispensations for at least some less close consanguinities could be given by a local Catholic authority other than the Vatican.

At least in peninsular Italy, therefore, there are three sources of records: the priest is supposed to indicate on the parish marriage book that a specific marriage required a dispensation, but we found that in some parishes this was not always carefully done. Folders keeping full records of dispensations are kept in each bishopric, and a slightly less complete set of records is available in the Vatican archives. It is rare that a dispensation is not approved by the Vatican, if it is dispensable. Genealogical trees of the candidate spouses reconstructed by the parish priest are extremely common in the bishopric archives that we investigated, and are available for practically every dispensation requested, but much less frequently in the Vatican archives. These genealogical trees are essential for checking the consanguinity degree calculated by the priest and for testing hypotheses on age and migration effects, to be described in a later chapter. The presence of genealogies in the dispensation folders made it possible to check for errors in the calculation of consanguinity degrees. Errors were nonexistent or exceptional, not surprisingly, since the method of computation is regularly taught to priests at seminaries.

Our work was based initially on bishopric records. After a full study of dispensations deposited in the archives of the Archbishopric of Parma—the diocese of the city where the university in which Cavalli-Sforza taught from 1951 to 1962 is located—those of the two adjacent dioceses of Piacenza and Reggio Emilia were also investigated. The territory of these three dioceses and cities has almost identical ecological structure, extending from the Appennine mountains to the lowest part of the plain of the Po River. The three dioceses form the northern moiety of the administrative Italian region called Emilia, practically at the center of the Po valley, in the northern part of Italy. The Po River flows just north of the city of Piacenza, and continues eastward toward the Adriatic Sea. Parma and Reggio Emilia are on a major Roman road, the *Via Aemilia*, an almost straight line in the Po valley leading east-southeast from Milan to Bologna. The region around the Po and the *Via Aemilia* is a very

fertile plain. Proceeding from each of the three cities toward the south, one enters first a hilly region and then a mountainous one. Population density is maximum near the cities, which are all located not far from the center of a very prosperous agricultural region. The mountainous region at the southern end of each diocese has the lowest population density. It is a part of the Appennine chain, the crest of which separates Emilia from Tuscany. The hilly region, intermediate between the plain and the mountains, has intermediate population density. The size of villages is on average highest near the cities and decreases regularly toward the mountains. This variety of environments of each diocese has helped in the investigation of the effects of the relevant ecological and demographic variables.

Bishopric records of the islands of Sardinia and Sicily, as well as of some other islands and inland regions of special interest, were also investigated, showing similar effects of demographic variables, along with other characteristics of each region. Records of individual dioceses of the islands were published earlier; their analysis has been repeated by partially new approaches for the purpose of preparing this book.

In later years it was possible to establish a team of young students who copied the consanguinity records of the Vatican archives under direction of Father M. Bracco. These records included all of peninsular Italy from 1911 to 1964, except for Sicily, which had independent rules. It was necessary to visit directly the bishoprics of Sardinia and Sicily, but our survey of Sicilian dioceses was not complete.

Full names of consanguineous couples were available. They were kept confidential, but we had permission to use them for linking them with other records to study the effect of consanguinity on certain phenotypes. The records were eventually transferred to computer tapes, analyzed statistically, and ordered alphabetically.

Results on some of the bishoprics were in part published earlier. But the major analysis, that of the Vatican records, is published for the first time in this volume. A number of other new calculations were carried out on the available records and are included in this book. Other socioeconomic investigations done in Italy were studied and correlated with the consanguinity data. Socioeconomic information came from the Istituto Centrale di Statistica.

Closely connected with the analysis of consanguinity, studies of demography from parish books, mostly of the Archbishopric of Parma, were started. We realized that the parish books gave us access to demographic evaluations that could be used for comparing the amount of observed and expected "random genetic drift" (genetic variation between villages generated by chance effects, calculable on the basis of information on village size and migration among villages). At the time, the importance of drift in determining genetic variation in humans was considered by many geneticists to be trivial without real proof. Here was thus an opportunity to compare expected and observed genetic variation in real cases, and possibly to evaluate the relative role of drift versus other evolutionary factors. Genetic data were obtained on blood groups of inhabitants of most of the 92 villages of the valley of the Parma River, and correlated with the genealogical studies that form the basis of the present book. Many other investigations have since been done following the same scheme, but none is as large and complete. Only summaries of the Parma valley drift investigation have been published until now; full data are available for the first time in this volume.

Further use of the data collected in the consanguinity studies was made possible by methods we developed for studying surnames as genetic markers. Very recently it has been established that the Y chromosome, which is transmitted from fathers to sons and the presence of which determines the male sex, provides an extremely useful set of genetic markers. The differences in migration at marriage between males and females, mentioned earlier, make genetic diversity of the Y chromosome greater than that of the other chromosomes. Until recently, very few genetic markers of the Y chromosome existed, but this situation is now changed (Underhill et al. 1997, and later papers). Surnames, however, are Y-chromosome markers—they have on average a much younger age than DNA mutants, but are very useful for certain purposes. Parish books and consanguinity data have provided us with a number of surname data that could also be used for the study of genetic variation and give valuable estimates of drift. Thus, we could examine in detail the strong correlations existing between three major phenomena: consanguinity, inbreeding, and drift. We will examine these correlations in some detail.

1.3 CONSANGUINITY

The word consanguinity derives from the mistaken notion that blood (in Latin "sanguis") is the basis of inheritance. Two consanguineous people have at least one ancestor in common. In principle, one can prove that any pair of human individuals has common ancestors, but to find those of two individuals taken at random one would usually have to go back for a great number of generations. It is thus necessary to introduce a limit to how ancient common ancestors can be for considering two individuals as consanguineous. This limit varies with custom and is, of course, arbitrary. There is, therefore, no fixed limit. When Cavalli-Sforza and Bodmer discussed the problem in chapter 7 of the book *The Genetics of Human Populations* (1971, 1999), they mentioned as an example that the bond of two individuals who have a great great-great-grandparent in common (and are, therefore, as we shall see, fourth half-cousins) is very tenuous, indeed. The probability that they share a gene they inherited from their common ancestor is about one in a thousand. The most remote consanguineous pair usually considered is that of third cousins, full or half, for which the probability of sharing a gene transmitted from common ancestors is 1/256 or 1/512, respectively (see figure 1.1).

The case of half-cousins (descending from only one common ancestor) is less frequent than the normal case of full cousins. The common ancestor of half-cousins usually married twice and the two consanguineous mates descend each from one of his or her two spouses. Consanguinity can be simple or multiple, depending on whether there is only one chain of descent from a couple of married common ancestors or there are chains of descent from more than one couple. For instance, two consanguineous spouses can be first cousins from one pair of ancestors and second cousins from another.

The most important simple consanguinities are shown in figure 1.1. The three relationships involving equal length of the two branches leading to the husband and wife (1st, 2nd, and 3rd cousins) are usually the most common degrees. Roman Catholic terminology is shown in the first column of degrees of relationship. Civil law definitions usually follow the Napoleonic code, given in the next col-

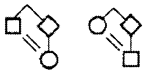
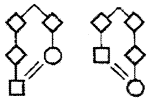
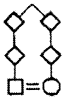
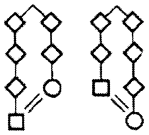
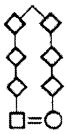
Type	*Symbol*	*Degree of Relationship*		*Inbreeding Coefficients (F)*	
		Roman Catholic Usage	*Napoleonic Code*	*Full*	*Half*
Uncle-niece; aunt-nephew		I in II	III	1/8	1/16
First cousins		II	IV	1/16	1/32
First cousins once removed($1\frac{1}{2}$)		II in III	V	1/32	1/64
Second cousins		III	VI	1/64	1/128
Second cousins once removed ($2\frac{1}{2}$)		III in IV	VII	1/128	1/256
Third cousins		IV	VIII	1/256	1/512

FIGURE 1.1. The most common types of consanguineous matings, their symbols, and inbreeding coefficients: □, male; ○, female; ◇, an individual of either sex. "Full" and "half" refer to the two sibs starting the chains of descent, who are the two top individuals in each pedigree. Full sibs have both parents in common, and half-sibs only one parent. *From Cavalli-Sforza and Bodmer [1971, 1999].*

umn. The measurement of genetic similarity among spouses is expressed by the inbreeding coefficient *F*, shown in the last two columns for half- and full cousins. Its meaning is described in the next section.

1.4 INBREEDING MEASUREMENT

An individual who received from his parents two identical forms of a given gene is said to be homozygous (or a homozygote). In a heterozygote the two gene forms of paternal and maternal origin are different. The indication that an individual is homozygous or heterozygous refers specifically to a single gene (traditionally, a DNA segment with a specific function, but sometimes a DNA segment specified in another way). Different forms of a gene are called *alleles*. Until one could examine DNA directly the number of detectable alleles was very small. Our current ability to examine DNA directly allows us to distinguish all mutations that have occurred and are maintained in the population examined. Today the physical identity of two genes can be ascertained with practical certainty by appropriate tests of DNA. This is called *identity by state* or *identity by nature*.

Inbreeding may cause another type of identity, *identity by descent*, when two descendants receive the same gene from a common ancestor, but his/her two alleles of that gene are not necessarily identical by state or nature. The possibility that two genes identical by descent are not identical by state arises if a gene passed from a common ancestor to two descendants undergoes mutation in at least one of the two chains of descent, leading to the two descendants. When this happens the particular gene being considered will be different by state in the two individuals, although it is identical by descent. This distinction is important for theoretical work.

Some semantic confusion may be generated by the fact that the word "gene" is used to indicate both loci and alleles. Historically, loci (plural of locus) are positions in a chromosome, as established by studies of crossing over. The word "locus" is used more loosely today. In genes of known sequence, a locus may be the position of a single nucleotide, the smallest element of DNA (sometimes a longer segment), but this is more properly called *nucleotide site*. Alleles are

the alternative forms of DNA observed at a given locus. They often differ one from the other because of a single nucleotide, but sometimes because of a longer segment. There is possible confusion between alleles defined at the level of a whole gene (a DNA segment with a precise function) and at the level of DNA. In pre-DNA times the distinction among alleles was not as sharp as is possible today, when we know the DNA structure of a gene. All differences in state giving rise to different alleles are caused by one or more past mutations.

An individual homozygous by nature need not be homozygous by descent, and usually is not. The test of inbreeding by nature may require knowledge of the genealogy for very many generations to establish common descent. In principle, one can also distinguish homozygosity by descent from homozygosity by state also empirically, by looking at nearby genes. There usually is variation in DNA very close to that being studied; genes identical by descent, even if they are not identical by nature, will have identical neighbors, except for earlier mutations, which are not frequent; genes identical by nature but not by descent may have different neighbors. This diagnosis requires thorough advanced knowledge of neighboring genes at the DNA level, ideally the full sequence.

The inbreeding coefficient F measures the probability that an individual receives at a given locus two genes identical by descent, calculated on the basis of the relationship of his/her parents. It is the same for all genes of autosomal chromosomes and can be calculated from the pedigree of an individual. Figure 1.2, taken from figure 7.1 of Cavalli-Sforza and Bodmer (1971, 1999), shows the calculation for a simple example, a half-uncle–niece marriage, which has $F = 1/16$. The genes of the uncle (A) are labeled a and b, while those of his two spouses are assumed to be different from a and b and between themselves. As there is no need to distinguish them, they are all labeled $+$. The two children B and C can be either $a/+$ or $b/+$, with probability 1/2. If C is $a/+$, then D, the niece of B marrying him, is also $a/+$ with probability 1/2. If D is $a/+$ and B is $a/+$, then their child E can be homozygous (a/a) with probability of 1/4. Altogether, the probability that E is a/a is the product of 1/2 (the probability that B is $a/+$) times 1/2 (the probability that C is $a/+$), times 1/2 (the probability that D is $a/+$ if C is $a/+$), times 1/4 (the proba-

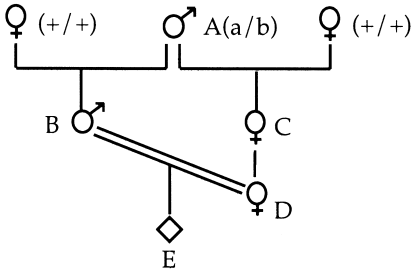


FIGURE 1.2. Half-uncle–niece marriage. Coefficient of inbreeding $F = 1/16$.

bility that a child of B and D is *a*/*a* if both are *a*/+). This product is 1/32. There is an equal, independent probability of having a *b*/*b* child. The total probability that the child is homozygous *a*/*a* or *b*/*b* is the sum of the two probabilities, $1/32 + 1/32 = 1/16$.

Full uncle–niece (UN) marriages have twice that value, and aunt–nephew (AN) marriages give the same value, 1/8 (these values differ for X chromosomes). Adding a generation to one of the two chains of descent leading to the spouses multiplies F by 1/2. For more complex pedigrees a simple method of computation is shown in most population genetics textbooks (see also Cavalli-Sforza and Bodmer [1971, 1999], chapter 7, from which this example and many of the following ones in this chapter are taken).

1.5 INBREEDING EFFECTS

Inbreeding increases homozygosity above the values expected under random mating in a population of finite size. Frequencies of genotypes of ordinary populations are predictable under assumptions of random mating in an infinite population. These give rise to the *Hardy-Weinberg* distribution of genotype frequencies. For alleles A_1, A_2, . . . , A_k with frequencies p_1, p_2, . . . , p_k this distribution is given by

$$(p_1A_1 + p_2A_2 + \cdots + p_k A_k)^2$$

The result is that homozygotes A_1A_1, A_2A_2, . . . , A_kA_k have frequencies ${p_1}^2$, ${p_2}^2$, . . . , ${p_k}^2$, and heterozygotes A_1A_2, . . . , $\mathrm{A_jA_k}$, . . . have frequencies $2p_1p_2$, . . . , $2p_jp_k$, ...

But individuals with inbreeding coefficient F have expected frequencies $Fp_i + (1 - F)p_i^2$ if homozygotes, and $(1 - F)2p_jp_k$ if heterozygotes A_jA_k. Therefore, all heterozygotes have a lower expectation under inbreeding than under random mating by a factor $1 - F$, and homozygotes a correspondingly higher one. If a population is completely inbred, then $F = 1$, and only homozygotes exist.

The average inbreeding coefficient of a population is called α and an approximate estimate can be obtained by averaging the inbreeding coefficients of its individuals. If the frequencies of consanguineous matings in the population m_i and their inbreeding coefficients F_i are known, then

$$\alpha = \Sigma m_i F_i$$

α values vary considerably with space and time (see www.igm.cnr.it/Zei/freqcons.html).

One usually neglects in these calculations more remote consanguinities, because they are not known, and it is inevitable to wonder if this computation leads to serious underestimates. An exact answer is difficult. It might seem that in most instances the inaccuracy is not serious, given that third cousins already contribute almost negligibly to the final estimate. The inaccuracy is likely to be more serious in highly inbred populations that are very small and have remained small for a very long time, so that inbreeding is more likely to have accumulated over time. One of the highest α values (0.044) was found in Samaritans, who number only a few hundred individuals and have essentially no gene flow from outside populations. Small religious isolates in North America, who also have negligible gene flow from outside, except for other colonies of the same denomination, show α values up to 0.025. In one case (Mange 1964) some idea of the underestimation thus incurred in a growing human isolate was possible, and was not entirely trivial. The fact that this isolate was growing rapidly introduced a further factor, however. We have already mentioned that careful avoidance of inbreeding in a very small group of Polar Eskimos has kept α rather low (less than 0.003 [Sutter and Tabah 1956]).

Deviations from Hardy-Weinberg (HW) in human populations are rare. The extent of deviations caused by inbreeding is rather small, making it unlikely that an inbreeding effect will manifest itself by

deviation from HW. Natural selection is more likely to cause the opposite deviation, making heterozygotes more frequent than expected. But admixture of populations with different gene frequencies, as is unavoidable in very large samples collected over wide geographic areas, can cause a relative decrease of heterozygotes, simulating nonexistent inbreeding. Some improperly call inbreeding the loss of heterozygosity due to recent admixture, as can be observed in heterogeneous populations. In general, deviation from Hardy-Weinberg is not a useful cue for detecting inbreeding effects, as it is more likely that it is the result of admixture of populations because of poor sampling schemes.

Especially in plants, but also in animals, extreme inbreeding may cause loss of fertility, resistance to diseases, and phenotypic deterioration (e.g., loss of height). This happens, however, on α values higher than those observed in humans. Thus, inbreeding depression, as the phenomenon is called, is only moderate in humans. The very low fertility observed in various indigenous groups of the Andaman Islands, who are now reduced to less than 100 individuals, may be one of the worst cases of inbreeding depression, but the F is not precisely known in the absence of long genealogies, and there are other possible causes of loss of fertility.

Ways of analyzing inbreeding depression in humans include the study of *genetic load*. Genetic load is defined in terms of Darwinian fitness, which is the number of children contributed to the next generation by a genotype. Darwinian fitness is usually but not always expressed as relative to that of the genotype with highest fitness (Crow 1958). The genetic load, as first defined by H. J. Muller (1950) is the loss of Darwinian fitness due to deleterious genes maintained in a population by the balance of adverse mutation, producing deleterious genes, and natural selection eliminating them.

New mutants are sometimes deleterious. The relative loss of Darwinian fitness of the homozygotes for a specific mutation, compared with that of the normal homozygotes, is called *s,* and that of a heterozygote for the normal and mutated form of the gene is *hs*. To avoid possible confusion, this means that we call *1* the fitness of the homozygous normal, *1-hs* that of the heterozygote, and *1-s* that of the homozygous mutant. Therefore, h is a quantity specifying the dominance of the mutation, being 0 if the mutant is fully recessive, 1 if it

is fully dominant, intermediate between 0 and 1 if dominance is intermediate, and negative if there is heterozygous advantage (also called "overdominance"). Mutation at that locus, occurring at rate μ per generation, keeps accumulating new mutants, which are eliminated at a rate increasing with s. At equilibrium between mutation and selection the number of new mutants per generation is equal to those eliminated by natural selection. The proportion of mutant genes at that locus existing in the population at equilibrium depends on mutation rate μ and the selection parameters s and h. If $h = 0$, the mutant is fully recessive and the gene frequency of the mutant in the population at equilibrium is $\sqrt{\mu/s}$. The mutant phenotype before selection has frequency μ/s, that is, it is equal to the mutation rate if the mutation is completely lethal ($s = 1$); otherwise, it is higher. If the mutation is completely dominant ($h = 1$), then its equilibrium gene frequency is μ/s and the frequency of the phenotype is twice that value. When the heterozygote is intermediate between 0 and 1, its fitness tends to dominate the picture even if h is small, because heterozygotes for rare mutants have a frequency much higher than that of homozygotes, which are usually very rare.

Morton et al. (1956) applied the Muller approach to estimate the genetic load in human populations. The load L is the loss of Darwinian fitness and is a linear function of the inbreeding coefficient, being, for a specific locus i,

$$L_i = a_i + b_i F$$

where L can be calculated as minus the logarithm of mortality (or other measurements of load).

Summing over all i loci gives

$$L = A + BF$$

provided different loci act independently, so that the joint probability of different damages is their product.

The expected linear relationship was tested in a large number of populations. The best data set seems to be one of the oldest, produced by Schull and Neel (1965) from estimates obtained in Nagasaki and Hiroshima. The mortality examined was that of children from first-, 1½-, and second-cousin marriages, the most common and most easily detectable consanguineous marriages, which are partic-

ularly frequent in older Japan. Here it was found that $B = 0.4$ and $A = 0.04$. Note that the mutation estimates on which these data are based are not related to radiation effects, but express mutants produced in earlier generations, long before atom bombs were used against these two cities during World War II.

The quantity B, or a slightly larger one, included between B and $A + B$ is approximately equal to the sum of sq over all independent loci contributing to mortality, where q is the gene frequency of a deleterious allele at a given locus, and s is the loss of fitness of the mutant homozygote. It has been called "the number of lethal equivalents," that is, the number of lethals that would be found in gametes, assuming that $s = 1$ for all genes. Since s, the mortality due to a deleterious but not fully lethal gene, is smaller than 1, the number of deleterious genes in the gametes of an average individual from the population must be greater than the number of lethal equivalents.

The comparison of different populations shows considerable variation of A's and B's. An old summary (Cavalli-Sforza and Bodmer 1971, 1999) indicated that Japanese values are lower than average; data from Caucasoids tend to be more variable and higher, varying from less than 0.5 to more than 2.5. Limiting the analysis to first-cousin data, which are the most reliable, Bittles and Neel (1994) estimated the mortality due to recessive deleterious genes to be around 4.4%, from which one can calculate that the average person carries 1.4 lethal equivalents.

A recent, thorough meta-analysis by Grant and Bittles (1997) proved that there is considerable heterogeneity among the 42 populations considered. Linearity with F was not tested directly by the authors, but the data they published indicate a serious deviation from linearity at the upper end: the highest degrees tested (uncle–niece or equivalent, i.e., double first cousins, which have $F = 1/8$) had a mortality lower than twice that of first cousins (twice is its approximate expectation). Even if mortality was not perfectly linear with F, there was a clear effect of consanguinity, which was qualitatively in the expected direction. There were no obvious differences in the sensitivity to F among the different types of mortality: prenatal and postnatal in various age ranges, including postreproductive ones.

The interpretation of load analysis is not as straightforward as was suggested in the first studies. Theoretically, it can be shown that it is

difficult to distinguish *mutational load*, as the H. J. Muller type of load is called, from *segregational load* (due to polymorphisms responsible for heterozygous advantage). Also, possible nonadditive interactions between genes make the estimate of lethal equivalents uncertain. Nevertheless, if one can confirm that the load estimated in Japanese populations is truly lower than in others, which have had less consanguineous marriages, and if one can assume that the Japanese custom of frequent consanguineous marriage has been going on for a substantial time, it is reasonable to conclude that this practice has freed this population from part of its mutational load. A lethal recessive mutation that has taken place in the past will sooner or later determine one genetic death in a stationary population (and more in a growing population). The regime of consanguineous marriage will affect the delay with which this happens.

1.6 RANDOM GENETIC DRIFT

Drift is the name given by Sewall Wright to the effect of chance on gene frequencies in successive generations. Higher organisms form the next generation by sexual reproduction: gametes unite to form zygotes by fusion of a spermatozoon and an egg cell. If N individuals are formed from those of the earlier generation, $2N$ gametes of that generation must have been used to form the new generation, and they may be viewed as a random sample (in the absence of gametic selection) of the genes present in the earlier generation. Under these conditions, the genes of every new generation are a random sample of those of the last generation, and one can compare the formation of the next generation to the taking of a random sample of gametes from the former generation. The nearest statistical analog of this sampling is called "binomial sampling," which is like drawing a sample of black and white balls from a bag where they are contained in known proportions. The two quantities of importance are the proportion p of black or white balls in the bag and the number N of balls taken out to form every new generation. Actually, there are two bags: one of male gametes and one of female gametes, but in practice they almost always have the same composition. To simulate the very large number of gametes from which are taken the few sorted to form the

next generation, one always puts the black or white ball back into the bag after its extraction, before taking another ball. This is the same as "binomial sampling"; it is not realistic but the deviation thus generated is negligible.

The important conclusion to take home is that chance affects long-term results of evolution, the more so, the less numerous the individuals of a population. A population bottleneck can have drastic effects, especially on pathology. If a population is generated by very few founders, say 10, and one of them happens to have a rare inherited disease, the frequency of that disease will be 10% after the bottleneck and will tend to remain 10% if the diseased individuals are not subject to strong adverse selection. This is especially true of diseases that do not cause prereproductive mortality or loss of fertility. The frequency of the diseases will oscillate, of course, statistically, in successive generations. There are many examples of populations that have abnormally high or low frequencies of certain genetic diseases, specifically because of drift effects due to a history of one or more demographic bottlenecks in the recent past, or to a persistently low population size. The variation may be in both directions: the disappearance of a disease or the presence of it at an abnormally high frequency. Every population will therefore show a different picture of incidence of genetic diseases, which will deviate more from average, the smaller the bottleneck at founding of the population or at any later time. The effect will vary enormously in the same population for different diseases.

Some people have confused drift and inbreeding. These are really two different phenomena, but there is a strong correlation between the two. When the population size N is small, the chance is that one will mate more easily with a consanguineous individual. Small N is likely to generate inbreeding, and of necessity it also generates drift. Thus, a population of a given small size N whose individuals marry at random may be affected by a certain degree of inbreeding even if mating is random (independent of degree of consanguinity of mates). We may call this a *random inbreeding*. But mating systems may keep the inbreeding level high or low with respect to that expected under random mating: if consanguineous matings are avoided, then inbreeding is reduced; and if they are favored, then inbreeding will increase. One of the aims of the present research is to find whether systems of

mating of human populations cause the degree of inbreeding to be higher or lower than expected on the basis of population size.

Use of the word drift in genetics is equivalent to the long-term effect of statistical fluctuations of gene frequencies due to the finite size of populations. The concept involves, therefore, only the effect of random events. But the word drift is used with an opposite meaning in other sciences, like physics and linguistics, where it usually means a definite trend in specific directions. The *Oxford English Dictionary* suggests that drift means a definite tendency: the act of driving in a specific direction, or the condition of being driven, as under the action of a current. Kimura has therefore suggested the use of a longer expression, "random genetic drift", when it is useful to avoid potential confusion.

1.7 RESEARCH ON DRIFT IN THE PARMA VALLEY

The Parma River starts in the Apennines and descends northward toward the Po, the major river in Italy, into which it flows at the center of the northern plain. The valley it forms is separated to some degree from the two parallel valleys, east and west. Its geographic structure seemed ideal for a study of the effect of random genetic drift. It has a multitude of villages that increase regularly in size, on average, as one descends from the highest altitude toward the lowest. Village sizes, and genetic exchange by migration between them, mostly due to marriage, have been relatively stable over the centuries. Almost every village had its parish church, and a great majority of the parishes were already in existence in the eleventh century. Only a few have been founded since that time. Since the end of the sixteenth century, parish books of baptisms, deaths, and marriages were kept in each parish church and it became clear that they could be made available for study of the demography of the region.

In 1955, Cavalli-Sforza decided to begin an analysis of drift in the Parma valley. The plan was to study the demography of the valley on parish books and evaluate theoretically on the basis of these data the amount of drift expected in the various sections of the valley. At the same time, blood samples would be obtained from at least 30–50 adults, as unrelated as possible, from the parishes, and the only ge-

netic markers available at the time, major blood groups, would be tested on them.

Data collection lasted until the end of the 1950s. Statistical analysis went through various phases. It was soon clear that, by and large, the genetic variation of the Parma valley could be explained on the basis of drift, but there were problems tied with the existing methods of analysis. In the 1960s many methods were tried and new methods proposed, including what was probably the first use of population simulations. These difficulties were not completely resolved at the time the book *The Genetics of Human Populations* by Cavalli-Sforza and Bodmer was published (1971, reprinted without change by Dover in 1999). Eventually it was concluded that the most satisfactory method of analysis was population simulation, and final proof that drift could fully explain the results is published in this book.

1.8 GENETIC USES OF SURNAMES

It has long been clear that surnames, being transmitted from fathers to children, contain information potentially of interest for genetics. In particular, in consanguineous marriages both spouses have a specific probability of having the same surname (isonymy). The study of isonymy of spouses has been considered as a method of estimating inbreeding and kinship in a population (Crow and Mange 1965). The relationship is complicated by the fact that some surnames are very frequent, probably because the same surname arose many times independently (polyphyletism); therefore, the identity of surnames of two individuals is no proof that they are related. In general, the method needs corrections on the basis of the frequency distribution of surnames in the specific population, and has shown variable appreciation, depending on authors and on specific populations investigated. Lasker's book (1985) is dedicated especially to the use of surnames, with special but not exclusive interest in isonymy, and its usefulness for investigating genetic population structure. Two other books (the first arranged by K. Gottlieb [1983], the second by Brunet et al. [2001]) are collections of papers using the method and evaluating it, as well as other ways to use surnames for the same purpose.

We chose to avoid the use of isonymy because our interest in

surnames was aroused by other approaches. The rationale is that surnames can be considered as a single gene with many alleles. Such genes are particularly informative for evolutionary purposes, and may be very useful because many surname data on human populations are already available and computerized. They differ from the usual genes (with the exception of those of the nonrecombining portion of the Y chromosome, which behaves in a similar way) in being almost universally transmitted by only one parent, of the male sex only. They therefore are transmitted as in unisexual haploids, a fact requiring some simple corrections in the formulas to be used for estimating evolutionary factors like migration and drift. One weakness is that surnames are affected by illegitimacy to an unknown extent; this will generate a deviation from the usual gene behavior, which can be formally expressed by considering them as Y-chromosome "genes" affected by a high mutation rate. Another limitation of surnames is that they arose mostly recently, and therefore genealogies based on them have validity only for a short lapse of the past. This reason is to some extent confounded with problems due to illegitimacy. However, even in countries where they arose as recently as in Japan during the Meiji restoration (1867–1912) (Yasuda and Furusho 1971) they have been fairly useful for some genetic purposes.

We have used surnames especially to evaluate two types of genetic analysis of populations: random genetic drift and migration. The quantities involved are population size N, and the migration rate m per generation (the proportion of individuals of the population that entered from other populations in the last generation). Both quantities are necessary for evaluating the effect of "population structure" on evolution. The effect of drift is stronger the smaller N and the smaller m. In fact, their product, Nm, can be used as an inverse measure of drift, which increases as Nm grows smaller. Naturally, these quantities are bound to change over time, and what would be necessary is their knowledge over an adequate time period, but this is rarely available. In this respect, records from the Catholic Church are valuable, but their examination is time-consuming.

One of the constant and difficult problems in the study of population genetics is the separation and evaluation of the relative roles of genetic drift and natural selection as they affect the evolutionary change of gene frequencies over time and space. When genes are

studied it is not always easy to distinguish the effects of these two major evolutionary forces. But one can expect surnames to be relatively unaffected, per se, by natural selection, and this is of some advantage with respect to other traits.

The estimation of population size N is easier for human populations than for most other living organisms, and for this reason human populations are among the best organisms for the study of drift. To some extent, the study of migration is also easier in human populations, although it is less easy to estimate than population size. For certain purposes one needs a special formula for the calculation of N, known as effective population size N_e, but in many cases it is not necessary. When this estimation is necessary, in human populations the effective population size is approximately one-third of the global population, as explained in Cavalli-Sforza and Bodmer (1971, 1999). In the case of surnames, N should ordinarily be limited to the number of males. When this is strictly necessary we will make sure that the refinements of using N_e and N of males only are followed, but when examining correlations or studying proportionality to other demographic quantities neither restriction is important. One can obtain a quantity very similar to Nm from a formula involving the number of different surnames and the number of individuals in a population unit. This formula was developed by Fisher (1943) for other purposes, but we have started applying it in population genetics.

The second use is for calculating genetic variation among populations, with formulas taken from the F_{ST} variance of gene frequencies. This is another approach to the study of drift and, more generally, genetic variation among populations. Variances are calculated independently for each surname and averaged over surnames.

The first collection of surnames we used was from consanguineous marriages. We had computerized, for other reasons, around 540,000 consanguineous marriages for which there existed, in the archives of the Vatican, a consanguinity dispensation for the years 1911–1964. The archives excluded Sardinia and Sicily, which were done independently. Later, we were able to obtain surnames on magnetic tapes from telephone books by the SEAT Company for all Italy and, for Sardinia, also from electricity bills of households by the ENEL Company. Telephone listings for the year 1978 were examined for 91 Italian provinces and first used for a study of the migration per

province. We compared the migration that could be obtained from the quantity *Nm* derived from each province from surnames and the migration measured directly by the Istituto Centrale di Statistica (ISTAT). There was good agreement (Piazza et al. 1987). In several other papers (Zei et al. 1983a,b, 1986, 1993, Wijsman et al. 1984, Lisa et al. 1996) we studied specific problems like estimation of genetic parameters, relations with drift and natural selection, and, in general, the use of surnames for genetic problems.

We later extended the analysis of surnames for genetic drift to the approximately 8,000 Italian communes, again using telephone books by SEAT Company from 1993. Conclusions and maps are given in chapter 10. We also carried out several other correlations to cross validate data obtained on surnames and on regular genes, whenever a direct comparison was possible. This analysis confirmed the usefulness of surnames for most genetic analyses. Here we consider especially their use for the analysis of drift and for studying correlations of drift with inbreeding and consanguinity.

1.9 A SUMMARY OF PUBLISHED STUDIES ON CONSANGUINITY AND INBREEDING, WITH SPECIAL REFERENCE TO ITALY

The first published data on consanguinity and kinship in Italy are from Mantegazza in 1868. His pioneer work entitled "Studj sui matrimonj consanguinej" was based on the analysis of the offspring of 512 families (90 in Italy) with different degrees of consanguinity. It was only in the 1930s that there appeared other studies on consanguinity in Italy. As reported by Serra and Soini (1959), these studies regarded isolated areas, such as small communities, but also entire alpine valleys (Cantoni 1931, 1935, 1936, 1938, Gianferrari 1932, 1936). Since World War II, this research has spread.

Sometimes the study of the effects of consanguinity is the reason for collecting great samples of data on marriages often over long periods. Fenoglio (1956, 1969) searched for a relationship between consanguinity and sterility and between consanguinity and blood groups, analyzing all the marriages that occurred in a province of North Italy (Cuneo) during 1901–1960. Bigozzi et al. (1970) studied

morbidity and mortality in the offspring of 300 consanguineous marriages celebrated in Firenze during 1939–1958, compared with an equal number of nonconsanguineous marriages. Recently, Danubio et al. (1999) searched for a relationship between inbreeding and malaria in the southern region of Calabria. Marriage behavior, in particular, the rate of endogamy and inbreeding, is being studied by Guerresi et al. (2001), who are analyzing 4518 marriages that occurred in 7 parishes of the alpine Non valley.

More often, the studies were essentially devoted to analyzing frequency and trend over time of consanguineous marriages in some particular area or in the whole of Italy and searching for demographic and socioecomic factors affecting them. Serra and Soini (1959, 1961) examined all the marriages celebrated in three provinces of North Italy (Milano, Como, and Varese) during the period 1903–1953. Barrai and Moroni (1965) gave a preliminary picture of the consanguinity trend in the Reggio Emilia diocese during three centuries (1631–1963) through the analysis of a sample of parishes corresponding to 23% of the total. A long period (1565–1980) was examined for consanguinity in the Upper Bologna Apennine by Pettener (1985), who found the frequencies and trend of consanguineous marriages to be very close to those of the other Emilian dioceses of Parma, Piacenza, and Reggio Emilia, described in this book.

All these studies were based on data collected in ecclesiastic archives in the parishes or dioceses. These data are believed to be reasonably accurate for reasons described in this book (chapter 3). Not so reliable are the other sources of consanguinity data, as, for example, official statistics. Fraccaro (1957) made one the first studies of consanguinity in all regions of Italy based on the data provided by the Istituto Centrale di Statistica for the year 1953. The values of the regional inbreeding coefficient are underestimated by a factor of about three, compared with those obtained from the ecclesiastic sources. The author himself looks with caution at the results, remembering that "the data furnished by the official statistics on consanguinity depend, on one hand, on the *bona fides* of the partners and, on the other, on the *bona voluntas* of the officials" (Serra 1959).

Somewhat more reliable seems the estimate of consanguinity frequencies obtained by Cavalli-Sforza (1960) through the "Special investigation on the consanguinity of the marriages" that was coupled

to a survey of the Istituto Centrale di Statistica in 1959. This pilot study performed on a sample of 5‰ of the Italian families gave values of consanguinity underestimated in comparison with those of ecclesiastic sources, but only by a factor of about 1.5.

A bibliography of recent investigations of consanguinity frequency, and of its effects, outside Italy is available from Alan Bittles at www.consang.net.

1.10 STRUCTURE OF THIS BOOK

Chapter 1 has provided an introduction to the scientific definitions used and the problems that are addressed in this book.

Chapter 2 summarizes the history of rules about consanguineous marriages according to the Roman Catholic Church, how they varied over the centuries, and historical reasons for the changes.

Chapters 3 and 4 are dedicated to an analysis started long ago, but never completed or summarized: the calculation of expected frequencies of consanguinities if marriage is random or, more precisely, is unaffected by the consanguinity degree, and the comparison between observed and expected frequencies. Chapter 3 contains a statistical analysis showing the importance of age factors affecting the frequencies of consanguineous marriages, especially differences of frequencies between degrees that differ for the number of generations leading to spouses in the two branches, husband and wife. There is a basic tendency for husbands to be older than wives in all marriages. We analyze whether the general distribution of ages of husbands and wives can explain every age effect. Statistical analysis of the data shows that the effect of age is strong and can be explained to a large extent by the general correlation for age at all marriages, but two other causes of variation must be taken in consideration. The more important one is differences in movements at marriage between husband and wife, which we call sex-differential migration. A third, minor effect we observed is a probable tendency to maintain family ties by mothers of consanguineous mates. This work was done especially for the cited three dioceses of northern Emilia (Parma, Piacenza, and Reggio Emilia), but there are some extensions to other parts of Italy.

In chapter 4 we describe simple theories to calculate the expected frequency (i.e., the probability) of consanguineous marriages, taking account of the major factors of age and sex differential migration. On their basis, we estimate how observed consanguinity deviates from random expectation, or, in other words, if a real avoidance of consanguineous marriages is supported by the need to require dispensation.

Chapters 5 and 6 are dedicated to the study of the relationship between drift and inbreeding, on the basis of genetic research done in the Parma valley. Chapter 5 describes in detail a genetic study of blood groups of the villages of the Parma valley. Its conclusions were published briefly in a *Scientific American* article (Cavalli-Sforza 1969), but the full details of the data and their analysis were never published before. Chapter 5 evaluates drift from the blood group data and also from surname data obtained from consanguineous marriages, as well as consanguinity and inbreeding data and coefficients, and compares them for the whole valley.

Chapter 6 gives, for the last three centuries, migration and other demographic data relevant to drift for 22 villages of the upper section of the Parma valley, which show the greatest drift. The analysis and comparison of observed and expected drift and consanguinity or inbreeding is made here with a computer simulation of the upper Parma valley population, made up of about 5000 individuals from the 22 villages.

Chapter 7 is dedicated to regions presenting special problems and opportunities. These include the two major islands, Sardinia and Sicily, and some other special isolates. Islands deserved special treatment, since their culture was somewhat different from that of the mainland, and it was necessary to examine the bishopric archives in detail, because not all these data are found at the Vatican archives. In Sardinia it was possible to collect data for all bishoprics over long periods. In Sicily, only first-cousin and closer consanguinities were available in the Vatican. Here a peculiarity of major interest is the very high frequency of uncle–niece matings. There are a great number of bishoprics in Sicily and not all could be examined, but trends are similar in all those examined. Chapter 7 also studies in detail a few isolates: the Aeolian Islands north of Sicily, the most important of which is Lipari.

Chapter 8 summarizes studies based on linking records of individual consanguinity with those of disease and physical traits.

Chapter 9 analyzes correlations between the frequencies of major consanguinity degrees, especially first and second cousins, and the α average inbreeding coefficient in all of Italy, obtained from the Vatican archives and other sources, and available data of demography, economy, and social characteristics studied by the Italian Istituto Centrale di Statistica.

Chapter 10 examines the concept of "deme" and estimates drift in all Italian communes by an analysis of surnames.

Chapter 11 summarizes the major conclusions.

The frequencies of various degrees of consanguineous marriages (percentages of total marriages) in space (92 provinces) and time (1911–1964, in 5-year periods) for the whole country, based on the Vatican records, can be found on the Internet at www.igm.cnr.it/Zei/freqcons.html.