Genetic epidemiology: an expanding scientific discipline[1]

Diego F. Wyszynski[2]

ABSTRACT *Genetic epidemiology is a relatively new discipline that studies the interaction between genetic and environmental factors in the etiology of human diseases. Taking advantage of genetic markers provided by molecular biological research, complex computerized algorithms, and large databases, the field of genetic epidemiology has undergone significant development over the past 10 years. Using concrete examples from recent scientific literature, this article describes the objectives and methodology of genetic epidemiology.*

The notion that environmental factors interact with the genome in the production of diseases emerged around the middle of the 19th century, when certain individuals were observed to be more resistant than others to communicable diseases. Almost 100 years passed, however, before epidemiologists interested in genetics and geneticists interested in epidemiology were able to develop the first analytic methods to identify environmental and genetic factors involved in the pathologic process (*1*).

Although expressions such as "epidemiologic genetics" (*2*) and "clinical population-based genetics" (*3*) had already been coined, Morton and Chung (*4*) were forerunners in associating the term genetic epidemiology with the discipline that strives to control and prevent illness by identifying the role of genetic factors, in interaction with environmental factors, in the etiology of human disease (*5*).

Prevention can take place at the primary, secondary, and tertiary levels. Primary prevention refers to reducing the incidence of a disease in a population (*6*). The best known example of primary prevention is immunization to prevent certain infectious diseases. In the scope of genetic epidemiology, avoiding an environmental risk factor (maternal smoking, for example) that interacts with genetic susceptibility (genotype *A2* of the *TGFα TaqI* marker in the fetus), thereby leading to a certain pathologic process (cleft palate), is an example of primary prevention (*7*). Secondary prevention refers to prevention of the clinical manifestations of a disease through early detection and effective intervention in the preclinical stage (*6*). Well-known examples of secondary prevention include early detection and intervention in cases of congenital hypothyroidism and phenylketonuria. Finally, tertiary prevention consists of minimizing the effects of a disease by reducing the complications and damage it causes. An example of tertiary prevention of a genetic disease is the use of prophylaxis with antibiotics and immunization for individuals with sickle cell trait to prevent certain bacterial infections that could endanger the life of the patient.

Genetic mutations are the basis of variation in the population (*8*). Like other clinically expressed or manifested traits (phenotypes), diseases involve genetic factors in three ways, which are not always mutually exclusive:

1. The mutation may be directly harmful to the individual. This category includes the many disorders transmitted in an autosomal dominant manner through a single gene, such as achondroplasia and Marfan syndrome.
2. The mutation may be harmful, but it may remain dormant for genera-

[1] This article has been published in Spanish in this journal (Vol. 3, No. 1, 1998, pp. 26–34) under the title "La epidemiología genética: disciplina científica en expansión."

[2] The Johns Hopkins University, School of Hygiene and Public Health, Baltimore, Maryland, USA. Mailing address: Department of Epidemiology, School of Hygiene and Public Health, The Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205, USA. Tel: (410) 955-7961; Fax: (410) 955-0863. E-mail: dfw@welchlink.welch.jhu.edu

tions. For instance, certain metabolic disorders of the newborn, such as cystic fibrosis, appear only when an individual inherits two copies (alleles) of the mutated gene, that is, one from each parent.

3. The mutation may be harmful only when it interacts with other genetic or environmental factors (*1*). For example, individuals who have both mutated alleles for phenylketonuria or congenital hypothyroidism manifest these diseases only when they are exposed to elevated concentrations of phenylalanine or reduced concentrations of thyroid hormone, respectively.

The goals of genetic epidemiology contrast with those of "traditional" epidemiology and population genetics. "Traditional" epidemiology studies the relationship between the environment and the incidence of a given disease, although it recognizes the significance of the host and his or her genetic makeup. Population genetics, on the other hand, seeks to predict the influences of population structure and selection and mutation on bodily phenotypes and diseases. Finally, genetic epidemiology studies the way environmental risk factors interact with the genetic makeup of a given population.

METHODS OF GENETIC EPIDEMIOLOGY

Genetic epidemiology uses two types of research strategies: descriptive and analytic. The descriptive strategy, at the population as well as at the family level, is based on the study of time, location, and the individual. Some questions that exemplify this strategy are as follows: What is the prevalence at birth of achondroplasia among the population, and what is the mutation rate for this disease? What are the frequencies of blood groups and of histocompatibility antigens in different population groups? Do geographic differences exist in the prevalence of a given genetic factor? In contrast, analytic studies seek to identify the role of genetic factors in the natural history of diseases, both in populations and families. Analytic studies answer the "why?" and "how?" questions of genetic epidemiology.

Family recurrence studies

A fundamental aspect of genetic epidemiology is the study of aggregation (or recurrence) of certain diseases in given families. King et al. (*9*) proposed three questions to help identify the scope of studies of family recurrence:

1. Are there diseases that affect various members of the same family?
2. Is this familial aggregation associated with common environmental exposure, hereditary susceptibility, or cultural inheritance of risk factors?
3. If there is genetic susceptibility, how is it inherited?

The existence of familial aggregation can be determined by observing the prevalence of a given disease in family members of the index case (the index case is the affected individual who introduces the family into the study) and of controls (individuals who are not affected). Such an aggregation exists when relatives of affected individuals run a higher risk of suffering from the disease than relatives of individuals who are not affected. This method is efficient and inexpensive, but one of its limitations is that information about characteristics of family members and controls may give rise to bias. For example, if the researcher is aware that the disease is present in the participant's family, he or she may overdiagnose it. Family members' knowledge of characteristics of the disorder and their ability to recall them may also be greater if they have an affected relative. Table 1 demonstrates a simple method of calculating relative risk (RR) through the use of a 2×2 table, illustrated in the study of Mettlin et al. (*10*), who investigated familial history of breast cancer in 779 patients and 1 558 controls admitted to the Roswell Park Memorial Institute in Buffalo, New York. The RR of suffering from breast cancer associated with a positive family history was 1.62 (95% CI: 1.28 to 2.06) (see Table 1). When the analysis was broken down by age of the cases and the controls (<55 or ≥55 years of age), the RRs were 1.34 (95% CI: 0.94 to 1.92) and 1.88 (95% CI: 1.37 to 2.58), respectively (Table 2). This difference reveals a limitation in family-based case-control studies, especially when the illnesses that are studied appear at a later stage in life, because family members of young patients tend to be younger than those of the controls.

Other methods, such as cohort analysis, regressions, and generalizable estimation equations, allow calculations to be broadened to include more complex situations. It is important to point out that a high family aggregation does not prove the existence of a genetic mechanism producing the disease, just as a low recurrence does not exclude the possibility that such a mechanism exists.

Although the comparison of family members of patients and of controls may be considered to be an "epidemiologic" technique, it is also possible to identify a familial aggregation by means of "statistical genetics." In this case, the degree of aggregation of a

TABLE 1. Relative risk of suffering from breast cancer associated with a positive and a negative family history, based on a group of 779 breast cancer patients and 1 558 controls

		Cases	Controls	Total cases	Total controls
Other relative affected[a]	Yes	a	b	144	191
	No	c	d	635	1 367
Relative risk (95% CI)			ad/bc	1.62 (1.28 to 2.06)	1.00

Source: Reference *10*.

[a] Other relative affected refers to any first-degree relative (mother, daughter, sister) with breast cancer.

TABLE 2. Relative risk of suffering from breast cancer associated with a positive and a negative family history, based on a group of 779 patients and 1 558 controls, by age

		Age			
		<55 years old		≥55 years old	
		Cases	Controls	Cases	Controls
Other relative affected[a]	Yes	58	90	86	101
	No	300	626	335	741
Relative risk (95% CI)		1.34 (0.94 to 1.92)	1.0	1.88 (1.37 to 2.58)	1.0

Source: Reference *10.*
[a] Other relative affected refers to any first-degree relative (mother, daughter, sister) with breast cancer.

disease in a family is expressed as λ_R, which is defined as the quotient between the risk among relatives of the cases of having the disease and the prevalence of that disease in the overall population. This method requires λ to be calculated for each degree of relation. Table 3 shows the results of the study by Slater and Cowie (*11*), who analyzed data from the first published familial studies on schizophrenia. It can be observed that λ approaches 1 as the degree of relationship becomes more distant. It is important to point out that such an association is not sufficient in linking schizophrenia to a purely genetic origin.

In the case of multifactorial hereditary diseases, two components can be distinguished in the covariance or correlation between blood relatives: that attributable to genetic differences and that produced by differences in environmental exposure. For discrete phenotypes (affected as opposed to not affected), the statistical model is based on the premise that there is a continuum of liability, with normal distribution, which determines the risk of suffering from the disease. According to this model, when the threshold is surpassed, the disease appears. Both the susceptibility and the threshold can be inherited, and mathematical properties of the normal distribution allow the parameter λ to be predicted. Analysis of the multifactorial model focuses on estimating the risk correlation among family members (*12*). This model does not distinguish genetic influences from environmental ones, and heritability can be given too much weight, especially when there are environmental factors that greatly influence the risk among family members. The multifactorial linear model also can be applied to phenotypes that are expressed as continuous variables, such as blood lipid or blood glucose concentrations, blood pressure, and hormone levels. Analyses of the variance components, or alternatively path analysis, are also useful methods for studying these phenotypes.

Once there is evidence of familial aggregation and genetic control of a disease, a third question emerges: How can the genetic marker involved be identified? To respond to this question, various methods have been developed over the past 20 years, thanks to the many new molecular biological techniques, as well as computers and complex statistical algorithms. The most common methodologies are described below, with examples from recently published studies.

TABLE 3. The first studies conducted on familial risk of suffering from schizophrenia

Years	Studies	Relation	Incidence[c]	λ[a]
1928–1962	14	Parents	336/7 675 = 4.36%	5.45
			(adjusted value[b] = 14.12%)	17.65[b]
1928–1962	12	Siblings	724/8 504 = 8.51%	10.6
1921–1962	5	Offspring	151/1 226 = 12.31%	15.4
1930–1941	4	Aunt/uncle	68/3 376 = 2.01%	2.5
1916–1946	3	Half siblings	10/311 = 3.22%	4.0
1926–1938	5	Niece/nephew	52/2 315 = 2.25%	2.8
1928–1938	4	Grandchildren	20/713 = 2.81%	3.5
1928–1941	4	Cousins	71/2 438 = 2.91%	3.6

Source: Reference *11.*
[a] Values are calculated assuming a population prevalence of 0.8%.
[b] The adjustment is made because the patient rarely has children once schizophrenia has become clinically overt.
[c] Calculated by the following formula:

$$\frac{\text{Individuals with X relationship who develop the illness during the given time period}}{\text{Total individuals with X relationship during the given time period}}$$

Twin studies

Twin studies have typically been used to determine whether genetic factors play a role in the etiology of certain diseases. Such studies consist of comparing the difference in concordance between identical or monozygotic twins (MZ) and fraternal or dizygotic twins (DZ). MZ twins share 100% of their genetic material, whereas DZ twins share, on average, 50% of their genes. If sets of twins are being studied, and the MZ twins are found to be concordant (both have the same disease, for example) with greater frequency than the DZ twins, it is possible to conclude that genetic factors are at least partially involved in the etiology of that disease (*13*). It is important to note, however, that genetic differences may exist between MZ twins. They may differ, for example, in the series of T-cell antibodies and receptors, in the number of mitochondrial deoxyribonucleic acid (DNA) molecules, in somatic mutations in general, and in the inactivation pattern of the X chromosome in female twins (*14*). It is also well known that MZ twins may differ from DZ twins as a result of environmental factors.

One of two calculations is normally made in twin studies, based on the

method used to select the twins: (1) pair concordance rate, which describes the proportion of twin pairs where both siblings are affected; and (2) index case concordance rate, which is the proportion of affected individuals among the co-twins of those selected as index cases. Although the pair concordance rate is the simplest method of determining whether genes affect a specific phenotype, it does not measure the magnitude of such an effect. For that purpose, use of the index case concordance rate is preferable.

Twin studies are limited by several factors, in particular those associated with the way participants are selected for the studies. For example, it has been observed that studies that depend exclusively on volunteers have a greater proportion of MZ twins, female pairs, and participants who are concordant for the phenotype under study. Such differences may influence the concordance rate that is calculated, which is why several countries—Sweden is a prime example—have launched population-based twin registries. Another limitation, especially in behavior studies, is that MZ twins tend to share environmental factors more frequently than DZ twins.

Gene-environment interaction studies

The existence of interactions between genetic and environmental factors has been widely described in the last half century. Phenylketonuria is a classic example. This recessive metabolic disorder manifests itself only in individuals who are homozygous for the mutation and who have been exposed to phenylalanine (an amino acid present in milk and other food products). Xeroderma pigmentosum is another example; affected individuals increase their risk of developing skin cancer when they expose themselves to ultraviolet rays. Ottman (*15*) has reviewed other similar examples.

Because of advances in the Human Genome Project, the case-control method is often used to describe possible genetic-environmental interactions. As seen in Tables 4 and 5, maternal cigarette smoking in the first trimester of pregnancy interacts with the fetal phenotype (*A2* allele of the genetic marker known as the transforming growth factor-alpha [*TGF*α]) in the formation of nonsyndromatic cleft palate (odds ratio obtained in the case-control study [OR_{cc}] = 5.5 [95% CI: 2.1 to 14.6]). This finding was later confirmed by Shaw et al. (*16*) in a study of isolated cases of cleft lip and palate. Khoury and Flanders (*17*) recently described case-only studies as an alternative to case-control studies. In a case-only study, the 2 × 2 table is reconfigured as illustrated in Tables 6 and 7. The odds ratio calculated in the case-only study (OR_{co}) is similar to the OR_{cc}. Although both methods are statistically powerful and relatively simple to perform, it is difficult to interpret the results. The existence of gene-environment interaction is, in itself, a statistical association, which is not necessarily causal. However, it is important to emphasize that both methods are useful instruments in the analysis of gene-environment interac-

TABLE 4. Outline for gene-environment interaction analysis in the context of a case-control study

Environmental exposure[a]	Genetic susceptibility	Cases	Controls	Odds ratio
−	−	a	b	1.0
−	+	c	d	OR_g = bc/ad
+	−	e	f	OR_e = be/af
+	+	g	h	OR_{ge} = bg/ah

Source: Reference *17*.
[a] −: absent; +: present; interaction under an additive model: $OR_{ge} = OR_g + OR_e$; interaction under a multiplicative model: $OR_{ge} = OR_g \times OR_e$.

TABLE 5. Interaction between fetal phenotype with TGFα and maternal smoking associated with cleft palate

Maternal smoking	Phenotype TGFα (*A2* allele)	Cases	Controls	Odds ratio (95% CI)
No	No	36	167	1.0
No	Yes	7	34	1.0 (0.3 to 2.4)
Yes	No	13	69	0.9 (0.4 to 1.8)
Yes	Yes	13	11	5.5 (2.1 to 14.6)

Source: Reference *7*.

TABLE 6. Gene-environment interaction analysis in the context of case-only studies

	Susceptible genotype	
Exposure	No	Yes
No	a	b
Yes	c	d

Source: Reference *17*.

TABLE 7. Review of data by Hwang et al. (*7*) in the context of case-only studies

	TGFα phenotype (*A2* allele)	
Maternal smoking	No	Yes
No	36	7
Yes	13	13[a]

Source: Reference *17*.
[a] OR_{co}: ad/bc: (36 × 13) / (7 × 13) = 5.14 (95% CI: 1.68 to 15.71).

tion, because they identify factors that could become significant in the prevention of the disorder being studied.

Complex segregation analysis

Complex segregation analysis is a useful technique for determining whether a specific phenotype (represented by a continuous or discrete variable) has a mendelian transmission pattern in a genealogical group (*1*). The algorithm used provides probability estimates for various genetic factors: for the mendelian models, these include transmission probabilities, gene frequencies, and penetrance parameters; for polygenic models, heritability, sample averages, and variances; and for what is known as the mixed model, both types of parameters (*18*). For example, Newman et al. (*19*) showed that the degree of family aggregation of breast cancer in 1 759 families was consistent with autosomal dominant inheritance as a result of the action of an uncommon allele (0.06%). This allele was implicated in 4% of all cases except 20% of affected mother-daughter pairs, within the larger context of multifactorial causation. Other examples of phenotypes studied by this technique are asthma and atopy (*20*), obesity (*21*), plasmatic apolipoprotein (*22, 23*), dyslexia (*24*), and labiopalatine clefts (*25*).[3] The primary limitation of this method is its sensitivity to the process by which individuals are selected. If the selection is biased, which usually occurs when cases come from a clinical setting, the results tend to be spurious. Furthermore, segregation studies are long and costly.

The methods described above indicate the relative importance of genetic factors in a disease or phenotype, but they do not identify the specific causal factor. To identify the genes that might be involved in the origin of diseases, "positional cloning" techniques are used, including allelic association analysis and linkage analysis.

Allelic association studies

The primary goal in allelic association studies is to compare the frequency of different risk factors in a group of individuals affected by a given disease and in a control group (*27*). The risk factor assessment may include environmental exposure or genetic traits. Genetic traits may be both genetic products, such as proteins or enzymes, or genetic markers based on DNA sequences. Genetic markers, known as restriction fragment length polymorphisms (RFLPs), are obtained by using restriction enzymes, which cut DNA at specific sites. In recent years, another type of genetic marker has been developed—the so-called microsatellites—which, in most cases, can offer more genetic information than traditional RFLPs (*8*).

Statistical analysis in an association study is simple and can be summarized in a 2×2 table. The challenge, as in most case-control studies, lies in selecting the controls. Allelic associations have yielded a better understanding and earlier diagnoses of certain autoimmune diseases. The allele HLA-B27, for example, is present in 90% of patients with ankylosing spondylitis, but it is found in only 9% of the general population (*28*). HLA alleles have also been associated with type I diabetes, rheumatoid arthritis, multiple sclerosis, celiac disease, and systemic lupus erythematosus (*29*). Associations have recently been identified between the angiotensin I-converting enzyme (ACE) and cardiovascular disease (*30*), between angiotensinogen and hypertension (*31*), between apolipoprotein E and Alzheimer's disease (*32*), and between the insulin gene (*INS*) and type I diabetes (*33*).

The interpretation of a positive association should not be taken lightly. Associations can arise for three reasons, one of which is completely artificial (*34*):

1. The allele in question is actually the cause of the phenotype.
2. The allele does not cause the phenotype but is in linkage disequilibrium with the causal allele. Linkage disequilibrium takes place when the causal allele of the phenotype is physically close (or linked) to the allele being studied. This is often observed in young, typically isolated populations (the Finnish population is a good example of a stable group in which allelic association studies often produce positive results).
3. The population is mixed. In a mixed population, any phenotype common to an ethnic group would appear to be positively associated with any allele that is also common in that particular ethnic group. Lander and Schork (*34*) give an amusing example of an association resulting from a mixed population group:

   "... suppose that a would-be geneticist set out to study the 'trait' of ability to eat with chopsticks in the San Francisco population by performing an association study with the HLA complex. The allele HLA-A1 would turn out to be positively associated with ability to use chopsticks—not because immunological determinants play any role in manual dexterity, but simply because the allele HLA-A1 is more common among Asians than Caucasians."

For this reason, the study of relatively homogeneous populations allows such spurious associations to be avoided.

Other analytic techniques developed in recent years do not seem to be affected by the makeup of the target population (*35*). One such technique is the transmission disequilibrium test (TDT) (*36*).[4] A hypothetical example is a genetic marker with two alleles, M_1

[3] Segregation calculations can be done on various computer programs accessible through the Internet (see reference *26*).

[4] Readers interested in other methods can refer to the work by Thomson on the haplotype relative risk method (*37*).

TABLE 8. Combinations of transmitted and nontransmitted marker alleles M_1 and M_2 among parents (2*n*) of affected cases (*n*)

	Nontransmitted allele		
Transmitted allele	M_1	M_2	Total
M_1	a	b	a + b
M_2	c	d	c + d
Total	a + c	b + d	2*n*

Source: Reference *36*.

and M_2, so that the possible combinations are M_1M_1, M_1M_2 (or M_2M_1), and M_2M_2 (Table 8). The case group is selected based on the presence of a given phenotype, and the genotype of these cases and their parents is determined. The frequency with which the M_1 or M_2 allele is transmitted to each affected individual is then assessed.

Families may be triads (the affected individual and parents) or they may be more complex (various affected family members plus parents). The method is statistically sound, even in mixed population groups. The TDT examines the hypothesis that the marker and the phenotype are not genetically linked. The theory used is derived from the Neyman-Pearson method (*38*) and uses only the *b* and *c* observations (see Table 8) from heterozygous parents (M_1M_2). The formula $(b - c)^2/(b + c)$ reveals whether there is an equal number of M_1 and M_2 transmissions from heterozygous parents to their affected offspring. If linkage exists between the marker and the phenotype, in addition to allelic association, *b* and *c* will tend to be different. The test for statistical significance of the TDT is the χ^2 (McNemar asymptotic test) or Fisher's exact test (*36*). A considerable difference confirms that the marker is linked to the phenotype locus. The TDT may be used with genetic markers with more than two alleles and may incorporate covariables (*39, 40*). It is important to point out that when the TDT is conducted on families with a recurrent phenotype (the so-called "multiplex families" in which more than one member is affected), a finding of allelic association is not valid. This is because the χ^2 test assumes that observations are independent, but this is not the case when participants are related. The linkage test, however, is valid even under these conditions (*35*).

Linkage analysis

Linkage analysis is a valuable resource used to identify genes that may have a causal association with the phenotype in question, because it allows one to assess whether the loci in a chromosome are transmitted together more often than expected during meiosis. Statistical tests of linkage estimate the recombination fraction (ϕ) between two loci. If $\phi = 0$, then there is complete linkage, which implies that the alleles at the two loci are always transmitted together. A finding of positive linkage between the locus of a genetic marker (known) and the locus of the phenotype under study (unknown) allows the investigator to determine the chromosomal location of the locus that produces the latter. In this manner, causal genes of more than 60 mendelian disorders have been identified, and the list grows daily (*41*).

If $\phi = 0.5$, then there is no linkage. In other words, each of the loci is transmitted independently of the other (as occurs with loci on different chromosomes). The value ϕ also serves as a measure of the physical distance between two loci; the greater the value of ϕ, the greater the distance from one locus to another. Linkage analysis may also be used to establish the sequence of loci on one chromosome if more than two loci are being investigated.

Linkage between two loci is not an actual occurrence but rather a hypothesis to be tested statistically. For this purpose, the maximum likelihood method is used (*42*). The likelihood of a hypothesis, called $L(H)$, is proportional to the probability of the experimental observation under this hypothesis, $\text{Prob}(O \mid H)$. In this case, the hypothesis is ϕ (linkage or no linkage); thus, maximum likelihood is expressed as $L(O \mid \phi)$. The relative likelihoods of the two hypotheses (linkages, or $\phi < 0.5$; or no linkage, $\phi = 0.5$) are calculated based on the likelihood quotient $LQ = L(O \mid \phi < 0.5) \,/\, L(D \mid \phi = 0.5)$. To produce a significance value, the *LQ* should be transformed logarithmically into an LOD score (logarithm for the likelihood of odds quotient of linkage, represented by *Z*). Algebraically, this is expressed as $Z = \log_{10}(LQ)$.[5]

LOD scores for the different values of ϕ are usually illustrated in a table (*45*). When $\phi = 0.5$, *Z* is always 0—because they divide two identical probabilities—and $\log_{10}(1) = 0$. For recombinant fractions less than 0.5, the referent LOD scores are 3.00 and −2.00. An LOD score ≥ 3.00 ($P = 10^{-4}$) is evidence of linkage, whereas an LOD score < −2.00 rejects the linkage hypothesis. Recently, Lander and Kruglyak (*46*) suggested that linkage should be considered significant once an LOD score of ≥3.3 ($P = 5 \times 10^{-5}$) is reached.

Linkage analysis may be broadened to include more complex systems. For example, multipoint linkage analysis allows multiple genetic markers located in the same chromosome to be assessed simultaneously. As a result of growing identification of the genetic markers present in each chromosome, multipoint linkage analysis has become the technique of choice for the exact location of genes. Given that this technique implies that a large number of markers will be analyzed within the same chromosome, investigators normally apply it only after signs of linkage have been found in a specific chromosomal region.

Linkage analysis may also be conducted when a desired phenotype shows genetic heterogeneity or when it is a result of the interaction of two or

[5] LOD scores can be calculated with programs available for personal computers and networks (*43*) or through the Internet (*44*).

more genes. In the first case, more than one gene acts independently in producing the phenotype. For example, in some families, hereditary breast cancer is attributable to mutations of the *BRCA1* gene; in others, it is due to mutations of the *BRCA2* gene. Finally, in some families the cause lies in mutations of unidentified genes. Phenotypes produced, at least partially, by the synergistic interaction of two or more genes include those observed in multiple sclerosis (*47*) and in total serum immunoglobulin E levels (*48*).

Analysis of shared alleles

The linkage analysis method described here is extremely sensitive to errors in the hereditary transmission models used to explain the phenotype studied and in the variations of the population-based allele frequency values attributed to the families under study. Thus, analytic techniques requiring no models have been developed, based on comparing the alleles that are shared between family members. One such technique, the analysis of affected siblings, evaluates how often a specific copy of a chromosomal region is shared identically by descent (IBD), that is, by being passed down from a common ancestor. For example, two siblings may share none, one, or two IBD copies of any locus (with an expected distribution of 25%, 50%, and 25%, respectively, if random allelic segregation has occurred). The statistical test compares the average number of alleles shared IBD (π) with the expected average (50%). The results are given in *P* values, LOD scores, or *Z* scores (number of standard deviations by which π surpasses the expected 50%). For example, 100 sibpairs who share 61% of the alleles in a genome sector correspond to a *P*-value of 0.001, an LOD score of 2.1, and a *Z* score of 3.1 (*46*). According to those authors, evidence of linkage is obtained with the sibpair method when the LOD score equals 3.6 or more ($P \geq 2.2 \times 10^{-5}$).

This method, using pairs of affected siblings, has been used with positive results in locating chromosomes for several phenotypes, such as those for type I diabetes, essential hypertension, serum immunoglobulin E levels, and bone density in postmenopausal women (*34*). Although much sounder than linkage analysis, the analysis of pairs of affected siblings is limited by the large number of siblings needed to provide sufficient data to perform the statistical calculations (on the order of hundreds or thousands of sibpairs).

FINAL COMMENTS

Academic activity and employment opportunities: what the future holds

Genetic epidemiology is a rapidly expanding discipline. Many academic institutions and government agencies—particularly in England, France, and the United States of America—offer academic and research programs in genetic epidemiology. Employment possibilities for genetic epidemiologists are excellent, especially in the more industrialized nations.

The International Human Genome Project has spurred great interest and controversy. Its primary goal is to obtain a complete map of the human genome by sequential analysis of DNA (*49*).[6] The U.S. National Center for Biotechnology Information announced, in October 1996, that approximately 16 500 genes had been identified, which corresponds to about 20% of all human genes (*50*). The Project is scheduled for completion in the year 2005, when the sequence of the 3 billion nucleotides of human DNA has been mapped. In this context, one task for specialists in genetic epidemiology is to educate the rest of the scientific community, and more importantly, the nonscientific community, regarding the implications and importance of the International Human Genome Project.

Ethical, legal, and social concerns

Since its inception, the planners of the International Human Genome Project recognized that gene identification would have profound implications for individuals, families, and society. Many questions were raised, such as how the genetic information should be interpreted and used; who should have access to it; how individuals could be protected from potential harm; and what is the benefit of genetic research when little, or nothing, can be offered in terms of a cure or prevention.

Genes that cause, or at least partially cause, several diseases have already been identified. Although such diseases can be detected and diagnosed earlier and more accurately, the long-term goal of the International Human Genome Project is to improve their treatment, to prevent them, and to ultimately cure them. In the interim, when early detection is possible but knowledge is limited and treatment is not yet available, there is a period marked by critical ethical, legal, and social controversy.

Since 1989, the National Human Genome Research Institute of the United States has housed the Working Group on Ethical, Legal, and Social Implications of the Human Genome Project. As a multidisciplinary and interinstitutional group, it is interested in the following four domains (*51, 52*):

1. Privacy and fairness in the use and interpretation of genetic information. It seeks to assess the mechanisms for preventing the discrimination and stigmatization that result from the misuse (and misinterpretation) of this information.
2. Clinical integration of genetic technology. In this context, the effect of

[6] The home page for the U.S. National Research Center of the Human Genome can be accessed on the Internet at: http://www.nhgri.nih.gov; the Genome Database, one of the main databases for localized genes, can be found at: http://gdbwww.gdb.org; the comprehensive genetic map can be seen at: http://www.ncbi.nim.nih.gov/SCIENCE96/; and the catalogue of genes and congenital defects is located at: http://www3.ncbi.nim.nig.gov/omin/

the availability of genetic testing in medical practice will be examined, together with the mechanisms for its assessment.

3. Methodology of genetic research. It is primarily concerned with determining how to inform potential volunteers of the risks and benefits of participating in a research study and how to obtain the corresponding consent.
4. Education for community members and medical professionals on the scope and importance of the Human Genome Project.

Genetic epidemiology in today's world

The International Genetic Epidemiology Society,[7] which has more than 400 members and is growing rapidly, edits the monthly journal *Genetic Epidemiology* and organizes international conferences annually. In Latin American countries a common medium for the development of genetic epidemiology has not yet been created. Initiatives in this respect have been independent and sporadic. Although many Latin American countries are experiencing widespread social and political changes, academic and scientific institutions need to make room for and give support to new disciplines, such as genetic epidemiology, without neglecting those that already exist. Only in this way will Latin American countries be able to join the circle of scientifically advanced nations.

[7] Its Website can be found at the following Internet address: http://darwin.mhmc.cwru.edu/IGES/index.html

REFERENCES

1. Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of genetic epidemiology*. New York: Oxford University Press; 1993.
2. Neel JV, Schull WJ. *Human heredity*. Chicago: University of Chicago Press; 1954: 283–306.
3. Vogel F, Motulsky AG. *Human genetics, problems and approaches*. 2nd ed. Berlin: Springer-Verlag; 1986.
4. Introduction. In: Morton NE, Chung CS. *Genetic epidemiology*. New York: Academic Press; 1978:3–11.
5. Cohen BH. Chronic obstructive pulmonary disease: a challenge in genetic epidemiology. *Am J Epidemiol* 1980;112:274–288.
6. Introduction. In: Gordis L. *Epidemiology*. Philadelphia: WB Saunders; 1996:5–6.
7. Hwang SJ, Beaty TH, Panny SR, Street NA, Joseph JM, Gordon S, et al. Association study of transforming growth factor alpha (TGFα) TaqI polymorphism and oral clefts: indication of gene-environment interaction in a population-based sample of infants with birth defects. *Am J Epidemiol* 1995;141: 629–636.
8. Mutation and instability of human DNA. In: Strachan T, Read AP. *Human molecular genetics*. New York: Wiley-Liss; 1996: 259–261.
9. King MC, Lee GM, Spinner NB, Thomson G, Wrensch MR. Genetic epidemiology. *Annu Rev Public Health* 1984;5:1–52.
10. Mettlin C, Corghan I, Natarajan N, Lane W. The association of age and familial risk in a case-control study of breast cancer. *Am J Epidemiol* 1990;131:973–986.
11. Slater E, Cowie V. *Genetics of mental disorders*. Oxford: Oxford University Press; 1970.
12. Elston RC. Segregation analysis. *Adv Hum Genet* 1981;11:63–120.
13. Susser M. Separating heredity and environment. *Am J Prev Med* 1985;1:5–23.
14. Ollier WER, MacGregor A. Genetic epidemiology of rheumatoid disease. *Br Med Bull* 1995;51:267–285.
15. Ottman R. An epidemiologic approach to gene-environment interaction. *Genet Epidemiol* 1990;7:177–186.
16. Shaw GM, Wasserman CR, Lammer EJ, O'Malley CD, Murray JC, Basart AM, et al. Orofacial clefts, parental cigarette smoking, and transforming growth factor-alpha gene variants. *Am J Hum Genet* 1996;58: 551–561.
17. Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Am J Epidemiol* 1996;144:207–213.
18. Morton NE, MacLean CJ. Analysis of family resemblance. III. Complex segregation analysis of quantitative traits. *Am J Hum Genet* 1974;26:489–503.
19. Newman B, Austin MA, Lee M, King MC. Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families. *Proc Natl Acad Sci U S A* 1988;85:3044–3048.
20. Panhuysen CIM, Meyers DA, Postma DS, Levitt RC, Bleecker ER. The genetics of asthma and atopy. *Allergy* 1995;50:863–869.
21. Bouchard C. The genetics of obesity: from genetic epidemiology to molecular markers. *Mol Med Today* 1995:45–50.
22. Moll PP, Michels VV, Weidman WH, Kottke BA. Genetic determination of plasma apolipoprotein AI in a population-based sample. *Am J Hum Genet* 1989;44:124–139.
23. Prenger VL, Beaty TH, Kwiterovich PO. Genetic determination of high-density lipoprotein cholesterol and apolipoprotein A-I plasma levels in a family study of cardiac catheterization patients. *Am J Hum Genet* 1992;51:1047–1057.
24. Pennington BF, Gilger JW, Pauls D, Smith SA, Smith SD, DeFries JC. Evidence for major gene transmission of developmental dyslexia. *J Am Med Assoc* 1991;266: 1527–1534.
25. Wyszynski DF, Beaty TH, Maestri NE. Genetics of non-syndromic oral clefts revisited. *Cleft Palate Craniofac J* 1996;33: 406–417.
26. Statistical analysis for genetic epidemiology. Available: http://darwin.mhmc.cwru.edu/pub/sage.html. Accesed 25 September 1997.
27. Hwang S-J, Beaty TH, Liang K-Y, Coresh J, Khoury MJ. Minimum sample size estimation to detect gene-environment interaction in case-control designs. *Am J Epidemiol* 1994;140:1029–1037.
28. Ryder LP, Andersen E, Svejgaard A. *HLA and disease registry: third report*. Copenhagen: Munskgaard; 1979.
29. Braun WE. *HLA and disease*. Boca Raton, Florida: Chemical Rubber Company Press; 1979.
30. Tiret L, Rigat B, Visvikis S, Breda C, Corvol P, Cambien F, et al. Evidence, from combined segregation and linkage analysis, that a variant of the angiotensin I-converting enzyme (ACE) gene controls plasma ACE levels. *Am J Hum Genet* 1992;51: 197–205.
31. Jeunemaitre X, Soubrier F, Kotelevtsev YV, Lifton RR, Williams CS, Charry A, et al. Molecular basis of human hypertension: role of angiotensinogen. *Cell* 1992;71: 169–180.
32. Pericak-Vance MA, Haines JL. Genetic susceptibility to Alzheimer disease. *Trends Genet* 1995;11:504–508.
33. Bain SC, Prins JB, Hearne CM, Rodriguez NR, Rowe BR, Pritchard LE, et al. Insulin gene region-encoded susceptibility to type 1 diabetes is not restricted to HLA-DR4-positive individuals. *Nature Genet* 1992;2: 212–215.
34. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science* 1994;265:2037–2048.
35. Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* 1996;59:983–989.
36. Spielman RS, McGinnis RE, Ewens WJ. Transmission disequilibrium test for link-

age disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993;52: 506–516.
37. Thomson G. Mapping disease genes: family-based association studies. *Am J Hum Genet* 1995;57:487–498.
38. Kendall MG, Stuart A. Volume 2: *Inference and relationship*. In: *The advanced theory of statistics*. 4th ed. London: Griffin; 1979.
39. Duffy DL. Screening a 2 cM genetic map for allelic association: a simulated oligogenic trait. *Genet Epidemiol* 1995;12:595–600.
40. Bickeböller H, Clerget-Darpoux F. Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. *Genet Epidemiol* 1995;12: 865–870.
41. McKusick VA. History of medical genetics. In: Rimoin DL, Connor JM, Pyeritz RE, eds. *Emery and Rimoin's principles and practice of medical genetics*, 3rd ed. New York: Churchill Livingstone; 1996.
42. Edwards AWF. *Likelihood*. Cambridge: Cambridge University Press; 1972.
43. Terwilliger J, Ott J. *Handbook for human genetic linkage*. Baltimore: Johns Hopkins University Press; 1994.
44. An alphabetic list of genetic analysis software. Available: http://linkage.rockefeller.edu/soft/list.html. Accessed 25 September 1997.
45. Morton NE. Sequential tests for the detection of linkage. *Am J Hum Genet* 1955;7: 277–318.
46. Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet* 1995;11:241–247.
47. Tienari PJ, Terwilliger JD, Ott J, Palo J, Peltonen L. Two-locus linkage analysis in multiple sclerosis (MS). *Genomics* 1994;19: 320–325.
48. Xu J, Levitt RC, Panhuysen CIM, Postma DS, Taylor EW, Amelung PJ, et al. Evidence for two unlinked loci regulating total serum IgE levels. *Am J Hum Genet* 1995;57:425–430.
49. Engel LW. The human genome project: history, goals, and progress. *Arch Pathol Lab Med* 1993;117:459–465.
50. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, et al. A gene map of the human genome. *Science* 1996;274: 540–546.
51. Pembrey ME, Anionwu EN. Ethical aspects of genetic screening and diagnosis. In: Rimoin DL, Connor JM, Pyeritz RE, eds. *Emery and Rimoin's principles and practice of medical genetics*. 3rd ed. New York: Churchill Livingston; 1996.
52. Human Genome Project information. Ethical, legal, and social issues (ELSI). Available: http://www.ornl.gov/Tech Resources/Human_Genome/resource/elsi.html. Accessed 25 September 1997.

Manuscript received on 22 April 1996. Revised version accepted for publication on 5 February 1997.

RESUMEN

La epidemiología genética: disciplina científica en expansión

La epidemiología genética es una disciplina relativamente reciente que estudia la interacción entre los factores genéticos y ambientales en el origen de las enfermedades humanas. Valiéndose de marcadores genéticos desarrollados a través de la biología molecular, de complejos algoritmos almacenados en computadoras y de amplias bases de datos, la epidemiología genética se ha desarrollado notablemente durante los últimos 10 años. El presente artículo describe los objetivos de la epidemiología genética y su metodología, empleando ejemplos concretos de la literatura científica reciente.