Review Article

Study Designs in Genetic Epidemiology

**Leyla Sahebi,¹ Saeed Dastgiri,² Khalil Ansarin,¹
Roya Sahebi,³ and Seyyed Abolghasem Mohammadi⁴**

¹ Tuberculosis and Lung Disease Research Center, Tabriz University of Medical Sciences, Iran

² Tabriz Health Services Management Research Centre, Department of Community and Family Medicine, School of Medicine, Tabriz University of Medical Sciences, Tabriz 5166615739, Iran

³ Tabriz University of Medical Sciences, Iran

⁴ Department of Plant Breeding and Biotechnology, School of Agriculture, University of Tabriz, Iran

Correspondence should be addressed to Saeed Dastgiri; saeed.dastgiri@gmail.com

Received 16 March 2013; Accepted 15 April 2013

Academic Editors: M. M. DeAngelis, J. Moreaux, and X.-X. Zhang

Copyright © 2013 Leyla Sahebi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Genetic epidemiology, as a relatively new issue, aims to explore the independent role of genetic-environmental determinants of diseases. Genetic epidemiology studies, depending on the objective, encompass the most preliminary surveys from the attempts to find family history in the occurrence of diseases to the most advanced surveys including specific strategies by clinical trials in the prevention of genetic diseases. Different objectives in genetic epidemiology studies require special methods and study designs. In this review, chief designs including familial aggregation, heritability, segregation, linkage, and association are evaluated; likewise, the purpose of diverse kinds of studies and analyses is briefly discussed. The utilization of study designs and related analyses according to the aims are the main issues and necessary in the accurate implementation of the study. Some methodological issues in relation to studies on tuberculosis are also reported. Attention to these issues might be useful in the implementation of these methods in the studies designed for the prevention and treatment of genetic disorders.

1. Introduction

Epidemiology is the study of distribution and determinants of disease frequency in human populations and the use of this information to control and promote health [1]. The goal of epidemiologic research is to collect valid and precise information on the causes, prevention, and treatment of disease [1]. Genetic epidemiology is the study of the role of genes and their interaction with environmental factors in the occurrence of disease in human populations [2]. The branch of genetic epidemiology is still quite young, although the parents of that (epidemiology and genetics) have rather long history [3]. The objectives of epidemiological studies in genetic science are to determine the risks related to allelic variants of candidate genes, to map more accurately regions of the genome for which there is evidence of linkage to disease susceptibility, and to contribute cases to a genome-wide search for susceptibility genes [4].

2. Study Designs in Classic Epidemiology

The selection of one design over another in studies depends on the particular research question [3] and also on cost, time, and ethical considerations. The most common types of studies are listed with brief explanations about them in Table 1 [1, 5–8].

3. Study Designs in Genetic Epidemiology

Similar to classical epidemiology, observational studies in genetic epidemiology are divided into descriptive and analytical studies. In descriptive studies, the pattern of variation in disease or behavior among immigrants, familial groups, as well as racial/ethnic groups, social classes, and temporal, age, and gender variations is surveyed and can provide clues to whether genetic or environmental factors are involved [8]. In analytical studies, the effect of genetic exposure on

TABLE 1: Main study designs in classic epidemiology.

	Case reports	Case series	Ecological studies	Cross sectional studies	Case control	Cohort	Intervention trials
Unit of study	(Single case) individual	(>1 cases) individual	Population	Individual	Individual	Individual	Individual
Description	Describes unusual characteristics of a case	A group of cases	Comparing populations in different places at the same time or, in a time series	Study of population at, "point-in-time,"	Study of two groups of subjects: (case; disease of interest and control; disease-free)	Study of two groups of subjects (exposure and non-exposure groups)	Study and examine two groups of subjects (intervention and control groups)
Direction	Present	Present	Present	Present	Reverse	Forward	Forward
Type of measurement	Reporting, description	Reporting, description	Correlation	Prevalence, association	Odds ratio	Prevalence, incidence, relative risk, attributable risk	Prevalence, incidence, relative risk, attributable risk
Advantages	Quick, having clinical importance, opportunities for physicians to exchange of thoughts	Quick, having clinical importance, opportunities for physicians to exchange of thoughts	Quick, inexpensive, group-level studies may also be the only way to study the effects of group-level constructs, for example, laws	Easy, inexpensive, useful for investigating fixed exposures such as blood group, most convenient in outbreaks of disease	Relatively inexpensive good for rare diseases Efficient in resources and time	Better for rare exposures, ability to determine causality relations	The strongest evidence for causality, control of unknown confounders, fulfils the basic assumption of statistical hypothesis tests
limitation	Inability to determine statistical relations and analysis	Inability to determine statistical relations and analysis	Ecological fallacy	Susceptible to selection bias and misclassification, difficult to establish a putative, "cause," Not good for rare diseases or rare exposures	Susceptible to selection bias and misclassification bias, may be difficult to establish that, "cause," preceded effect	Costly and time consuming, susceptible to selection bias. Relatively statistically inefficient unless disease is common.	Expensive, time consuming, sometimes ethically generalizability problem

disease is analyzed, and if it is proved that there are causal relationships, the genetic region responsible for the disease is identified. Therefore, relying on familial relationship in genetic epidemiology is one of the main differences between classic and genetic epidemiology. In Table 2, main designs in genetic epidemiology and process of surveys in genetic diseases are summarized [8–16].

4. Analytical Study in Genetic Epidemiology

4.1. Familial Aggregation. The first step in the study of a potentially genetic characteristic is to realize whether it tends to aggregate in families, without having any specific genetic model in mind [11]. In other words, familial aggregation is the tendency for disease to cluster in families [17]. A simple approach to assess aggregation in families is to identify a group of affected case subjects and a group of healthy control subjects and compare the odds of a positive family history in case subjects who report a disease to the odds of a positive family history in the control group. Finally, an odds ratio can be estimated according to Table 3 [1].

While this approach is valuable, there is no way to control for individual or environmental risk factors for each relative, which might be driving the aggregation (e.g., cooccurrence of smoking behaviour and lung cancer aggregation in families); it is also imprecise since the probability of a positive family history of disease increases with age and the number of relatives considered [17]. One way to control confounding variables is to use categorizing methods such as logistic regression modeling [17].

4.2. Heritability. Heritability is the proportion of variation in a trait that is attributable to genetic differences. One of the simplest designs in heritability study is the "twin study." This study design uses the variance component framework and its frequency to estimate heritability; in this way, if the ratio of genetic variance to phenotypic variance (general heritability) is close to 1, this is evidence for a genetic component [18].

4.3. Segregation Analysis. Segregation analysis aims to determine whether the pattern of disease occurrence in families fits a particular type of inheritance. For example, Huntington's

TABLE 2: Main study designs in genetic epidemiology.

Research question/problem	Type of study	Unit of study	Type of measurement	Aim
Is there evidence of phenotypic aggregation within families?	Familial aggregation	Case-parent Case-parent-grandparent	Recurrence risk ratio Correlation Odds ratio Conditional regression logistic	Identifying new disease genes
Is the pattern of aggregation reliable with an effect of genes?	Heritability	Twin study	Variance component heritability	Identifying new disease genes
Is there evidence of a gene with subnational enough effects to justify expensive studies?	Segregation	Familial pedigree	Maximum likelihood estimation Binomial distribution	Identifying new disease genes
Where in the genome is a causative gene most likely to lie?	Linkage	Case; disease of interest and control group; disease-free	Parametric: maximum likelihood Nonparametric: mean, proportion, chi square, likelihood test	Identifying new disease genes
Is there a causative polymorphism?	Association Linkage disequilibrium Cohort Case only study	Related—case control Unrelated—case control Exposure and non-exposure group Case groups	Transmission disequilibrium test Chi square independence Odds ratio Conditional regression logistic Correlation Lewontine's D prime Relative risk, Attributable risk Exposure odds ratio estimate	Characterizing known disease genes
Estimate allele frequency	Cross-sectional		Prevalence Correlation	Characterizing known disease genes
Estimating penetrance	Association Cohort	Related—case control Familial cohort	Odds ratio Conditional regression logistic Risk ratio Attributable risk	Characterizing known disease genes
Evaluating strategies for prevention of genetic diseases	Experimental Retrospective cohort	Clinical trial (intervention and control group) Exposure and non-exposure groups	Risk ratio Attributable risk Risk ratio Attributable risk	Characterizing known disease genes

TABLE 3: OR (odds ratio) = $(a * d)/(b * c)$.

Family history	Case	Control
Positive	a	b
Negative	c	d

disease is controlled by an autosomal dominant allele. Thus, if one parent has “Aa” genotype and another genotype, “AA,” due to the possible genotypes of “AA” and, “Aa” in the offspring, it is expected that all the children will be affected. Therefore, if the risk of a heritable disease in a large population is 50% among girls and boys, it can be concluded that heritability is of autosomal dominant type. Binomial distribution can be used for analysis. Thus, if n is the sample

size, x is the number of affected offspring, and p is the probability for a case to be affected (supposed to be 50%), by assuming that $H_0: p = 1/2$, we will have

$$(X = x) = \binom{n}{x} p^x (1 - p)^{(n-x)}. \quad (1)$$

4.4. Linkage Analysis. The goal is to find the approximate location of the responsible gene or genes [10, 19, 20]. In linkage analysis, two-point LOD scores evaluate the evidence for linkage between the disease locus and only a single marker; when more than one marker is considered, multipoint LOD scores are reported [11].

Two broad types of linkage analysis exist: parametric and nonparametric. If there is enough information for knowledge

of parameters (mode of inheritance and DNA from multiple members of informative families), it is possible to use model-based (parametric) linkage [17]; however, when the genetic model is unknown, nonparametric analysis should be used [17]. Parametric linkage analysis is a powerful strategy for mapping genes with a simple Mendelian form of inheritance [11]. A result of linkage analysis is usually expressed in terms of an LOD score (logarithms to the base 10). The LOD score is a function of the recombination fraction (θ) [21]. Recombination fraction estimates probability of recombination between two markers. Although a probability, its maximum value is 0.5 indicating a 50:50 chance of recombination, or that two loci sort independently and are unlinked but a recombination fraction less than 0.5 indicates that two loci are not sorting independently and there is linkage between them [22]. The LOD score is computed by comparing the likelihoods for a range of value θ and comparing with the likelihood when θ is equal to 0.5.

Therefore, with assuming, L , the likelihood probability and, θ_i , recombination fraction and maximum recombination fraction, denoted as θ ($\theta = 1/2$), LOD is assessed in the following manner [17, 23–25]:

$$\text{LOD} = \log_{10} \frac{L(\text{family}/\theta_i)}{L(\text{family}/\theta = 1/2)}. \quad (2)$$

An LOD score of 3 (which represents odds of 1000:1) or greater in favour of linkage is used to indicate statistically significant linkage. If this score is minus 2 or less, linkage is unlikely [22].

4.5. Association Study. Association studies are similar to case-control studies except that the disease associated “exposures” that one seeks to identify are variant alleles of genes [11]. These kinds of studies are used to find more common genetic variations that are highly prevalent in the general population [17]. In practice, frequencies of variant alleles among affected individuals are compared to unaffected individuals [26].

There are two types of association studies. The first one is the candidate gene study [17, 27], which focuses on the specific gene, in addition to what may be accomplished by the identification of a gene product such as a specific protein rather than the gene itself [27]. The second one is genome-wide study (GWAS), comprising a wide search of the genome for genes that are related to the disease [27]. The genes discovered thus far using GWAS have been common loci of small effects, but many genetic epidemiologists believe that, when taken together, these genes of small effects may cumulatively have vital effects on the risk for complex diseases [28]. Association studies are subdivided into two types of analysis, direct and indirect [17]. In direct association studies, the candidate gene has been designated and association is tested directly; but in the indirect one, candidate gene has not been identified and is linked to marker genes [16]. The direct method uses an actual causative mutation at a particular gene in the test for association with the disease phenotype. The indirect method uses linkage disequilibrium method. Linkage disequilibrium (LD) describes the strength

TABLE 4: $OR = (a * d)/(b * c)$.

	Case	Control
Exposure	a	b
Not exposure	c	d

of a relationship between alleles at different loci [23, 28]. More specifically, if knowledge of an allele at one locus can predict the allele that will reside at a second locus, then linkage disequilibrium exists between the alleles. However, if knowledge of an allele at the first locus cannot help predict the allele that is at the second locus, then linkage equilibrium (LE) exists [11, 16, 29]. There are many statistical measures for LD, and the more common metrics are Lewontin's D prime index, odds ratio (logistic regression), correlation, and χ^2 independence [30–33].

4.5.1. Lewontin's D Prime Index (D') [30, 34]. In mathematical terms, if there is no association or dependence between two alleles (e.g., C and G), then

$$P(\text{haplotype c-g}) = P(\text{allele c}) * P(\text{allele g}). \quad (3)$$

If alleles C and G are associated (in LD), in other words, they are dependent on each other, then

$$P(\text{haplotype C-G}) = P(\text{allele c}) * P(\text{allele g}) + \delta, \quad (4)$$

where δ is the *raw disequilibrium coefficient*. Therefore, if two alleles are in linkage equilibrium, then $\delta = 0$.

The raw disequilibrium coefficient, δ , can be difficult to interpret because it is dependent on allele frequencies at the two loci. D' is a scaled version of d that measures LD as a proportion of the maximum amount of LD possible for the specific allele frequencies at the two loci. It can take values from -1 to $+1$; if D' is equals to 1, it means that there is complete LD as follows:

$$D' = \frac{\delta}{\delta_{\max}}. \quad (5)$$

$\delta_{\max}$ can be estimated by the following equation (p , allele frequency in a locus, and q , allele frequency in another locus:

$$\delta_{\max} = \min [p(1 - q), (1 - p)q]. \quad (6)$$

4.5.2. Odds Ratio [11, 29]. One of the other indicators for the analysis of the relationship is comparing the odds of exposure (genotype or allele) in case group with the odds of exposure in controls group with assuming H_0 : odds ratio is = 1 we have (Table 4).

4.5.3. Correlation [31]. r^2 measures the correlation between alleles with a range of -1 to $+1$; in a situation of linkage disequilibrium, r^2 will be 1 and can be calculated as follows:

$$r^2 = \frac{\delta^2}{p(1 - p)q(1 - q)}. \quad (7)$$

TABLE 5: OR (in matched paired studies) = a/d .

	Patient group		
		E−*	E+**
Parents group	E+	a	b
	E−	c	d

E−*: non exposure.

E+**: exposure.

4.5.4. Odds Ratio in Matched Studies. An allelic variant of a candidate gene or of a genetic marker was associated with increased risk of disease; one would expect that variant to be transmitted from a heterozygous parent to an affected offspring more often than the 50% frequency expected by chance. Assuming a biallelic locus, let b be the number of times; then, the A1 allele is transmitted from a heterozygous A1A2 parent to an affected offspring and c is the number of times the A2 allele is transmitted from heterozygous A1A2 parent. It will be recognized as McNamara's test for a pair-matched case-control study. The "case" is the transmitted allele and the "control" is the nontransmitted allele [35].

McNamara's Test [32]. McNamara's test is a nonparametric method used on nominal data. It is applied to 2×2 contingency tables with matched pairs of subjects [32]. In 2×2 McNamara's table, subjects with offspring alleles E− and parent allele E+ will be in a cell, while subjects with offspring alleles E+ and parent allele E− will be in d cell; see (Table 5).

By assuming H_0 , it is expected that $a + d = 1/2$; then, we have

$$\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(a - (a + d)/2)^2}{(a + d)/2} + \frac{(d - (a + d)/2)^2}{(a + d)/2}. \quad (8)$$

And briefly,

$$\chi^2 = \frac{(a - d)^2}{a + d} \quad df = 1. \quad (9)$$

One potential limitation of association study is the probability of a false positive relationship between markers and genes [33]. Therefore, it should be noted that significant causation test in marker and alleles confirm causation and do not confirm linkage disequilibrium [34, 36]. The two loci may tend to be inherited together without the causality of the disease. This condition is more common in small populations (ethnic or tribal assembly) that have a lot of shared traits [37]. LD can be influenced by several factors, including chance, selection, migration, and isolated populations [38, 39]. Another restriction of association study is variation of test power when the disease allele is recessive compared with dominant allele [11]. Association studies (based on control group) are classified in to two groups: related case-control and unrelated case-control studies.

In the related case-control studies, relatives of case patients are used as control subjects. These designs can have various control groups, such as sib-control, cousin control

and pseudosibbling. Although the use of sibling control offers several advantages over population controls (unrelated case-control), such as more willingness to participate and reducing cost and time, it has some disadvantages like the probability of overmatching and limitation in age and sex matching in small families [39]. The advantage of a cousin-control group is that one may be able to obtain closer matching on age and year of birth, with less loss in efficiency because the case and cousin are not as closely matched on genotypes, in addition to less chance of overmatching because the case and cousin have only one side of their families shared [39]. In pseudosibbling designs, no actual controls are selected; instead, genotypic data are obtained on the parents of the case, and the genotype transmitted to the case is then compared with the three genotypes (pseudosibbling) that were not transmitted to the case. The question this design seeks to address is whether a specific allele or genotype occurs more commonly in cases than in their pseudo-sibs [39].

4.6. Estimation of Interaction in Exposures. One of the most functional studies in genetic epidemiology is the case-only which is used for cross-sectional, cohort, or case control patients [35]. The case-only design which is to assess gene-environment interaction was presented by Aalen et al. and then by Hamajima et al. [35]. The case-only study design used to study gene-environment interaction has been criticized for its susceptibility to bias caused by nonindependence between genetic and environmental factors [39–41]. Each person with disease (D) that is coded as positive or negative for a genetic factor (G) and environmental factor (E) can be located in one of the four situations of (D++), (D−), (D±), and (D±). It is noteworthy that when G and E are independent in the source population, the case-only OR is equivalent to the interaction estimate based on RRs, regardless of disease risk. Conceptually, the interaction between G and E refers to the extent to which the joint effect of the two factors on disease differs from the independent effects (effect of each of the factors on disease in the absence of the other factor) [42–44].

Multiplicative interaction is assessed by comparing the joint effect (effect on D due to the presence of both factors) with the product of the independent effects (product of effect on D in the absence of other). For example, if independent effect of G equals 3 and independent effect of E equals 2, then we would expect the joint effect of G and E to be 6 if there is no multiplicative interaction [44].

Given that independence between genetic and environmental factors is critical to the validity of case-only estimates of interaction [40, 41], when G and E are independent, the OR relating G and E are equivalent to the interaction effect of G-E; see Table 6.

In practice, if the sources of nonindependence are measured, classification or adjustment for using multiple models can be used to remove the bias in case-only analysis of interaction [41].

4.7. New Designs in Genetic Studies. Although case-control studies are commonly used for genetic-epidemiologic studies, an increasing number of cohort studies have been

TABLE 6: OR (G-E interaction effect) = $(a * d)/(b * c)$.

	E+*	E-**
E+	a	b
E-	c	d

E+*: exposure.

E-**: non exposure.

established over the past decade [45, 46]. Two progressive studies, the nested case-cohort and nested case-control, have recently been suggested. The major advantage of nested designs is their ability to match controls with cases on follow-up duration [4].

5. Study Designs in Mycobacterium Tuberculosis

Twin studies are one of the primary and inexpensive heritable studies on tuberculosis (TB), which provided valuable and important information about the etiology of TB. Because twins theoretically share the same environment, higher rates of concordance for monozygous (identical) twins than for dizygous (fraternal) twins suggest that genetic factors are important in susceptibility to tuberculosis and provide an estimate of the magnitude of this effect [47, 48]. During the past 15 years, various surveys have been carried out on the genetics of susceptibility to mycobacterial diseases [49, 50]. Etiology effects on tuberculosis have been used in case-control studies, too, like the case-control study carried out in Gambia, which showed that polymorphisms in the NRAMP1 gene were significantly associated with susceptibility to tuberculosis [51, 52]. Another case-control study in London showed VDR gene effect on susceptibility to TB [52]. Using association designs, important pathogeneses of tuberculosis have been discovered, too, such as NRAMP1, vitamin D3 receptor, interferon- γ , interleukin-1 β , interleukin-12, tumor necrosis factor- α , interleukin-4 and interleukin-10 [53–55]. Linkage studies have also shown that there is disease susceptibility gene or genes in the neighbourhood of the marker, and detailed investigation of genes in the region is indicated. Such a genome-wide scan of affected sibling pairs from Gambia and South Africa identified potential susceptibility loci on chromosomes 15q and Xq [49, 56] Deng [56] have reviewed the use of genetic linkage and association studies in the investigation of genetic susceptibility to infectious diseases. Implementation of such studies in developing countries presents some particular challenges. However, it is obvious that since tuberculosis occurs mainly in adults, parents of a case are frequently unavailable for genotyping. But using unaffected siblings as controls is possible [57]. In the study of complex diseases, as TB, because the effects of genes may be modified by environmental (i.e. non-genetic) factors, gene-environment interactions may be explored in study designs, such as case-only, cross-sectional, cohort, and case-control studies and family-based designs, such as case-parent, affected sibling pair and twin studies [57].

References

- [1] R. Bonita, R. Beaglehole, and T. Kjellström, *Basic Epidemiology*, WHO Library Cataloguing-in-Publication Data, 2nd edition, 2006.
- [2] J. Last, *A Dictionary of Epidemiology*, Oxford University Press, Oxford, UK, 3rd edition, 1993.
- [3] D. C. Thomas, *Statistical Methods in Genetic Epidemiology*, pp. 3–22, Oxford University Press, New York, NY, USA, 2004.
- [4] C. Lienhardt, S. Bennett, G. Del Prete et al., “Investigation of environmental and host-related risk factors for tuberculosis in Africa. I. Methodological aspects of a combined design,” *American Journal of Epidemiology*, vol. 155, no. 11, pp. 1066–1073, 2002.
- [5] S. Schwartz, “The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences,” *American Journal of Public Health*, vol. 84, no. 5, pp. 819–824, 1994.
- [6] F. D. K. Liddell, “The development of cohort studies in epidemiology: a review,” *Journal of Clinical Epidemiology*, vol. 41, no. 12, pp. 1217–1237, 1988.
- [7] L. Rodrigues and B. R. Kirkwood, “Case-control designs in the study of common diseases: updates on the demise of the rare disease assumption and the choice of sampling scheme for controls,” *International Journal of Epidemiology*, vol. 19, no. 1, pp. 205–213, 1990.
- [8] R. Peto, M. C. Pike, and P. Armitage, “Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design,” *British Journal of Cancer*, vol. 34, no. 6, pp. 585–612, 1976.
- [9] D. C. Thomas, *Statistical Methods in Genetic Epidemiology*, pp. 253–281, Oxford University Press, New York, NY, USA, 2004.
- [10] D. C. Thomas, “Chapter 4. Basic epidemiologic and statistical principles,” in *Statistical Methods in Genetic Epidemiology*, Oxford University Press, New York, NY, USA, 2004.
- [11] W. J. Gauderman, J. S. Witte, and D. C. Thomas, “Family-based association studies,” *National Cancer Institute Monograph*, vol. 26, pp. 31–37, 1999.
- [12] Q. Yang and M. J. Khoury, “Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research,” *Epidemiologic Reviews*, vol. 19, no. 1, pp. 33–43, 1997.
- [13] J. S. Witte, W. J. Gauderman, and D. C. Thomas, “Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs,” *American Journal of Epidemiology*, vol. 149, no. 8, pp. 693–705, 1999.
- [14] J. L. Hopper, G. Chenevix-Trench, D. J. Jolley et al., “Design and analysis issues in a population-based, case-control-family study of the genetic epidemiology of breast cancer and the Cooperative Family Registry for Breast Cancer Studies (CFRBCS),” *National Cancer Institute Monograph*, vol. 26, pp. 95–100, 1999.
- [15] P. R. Burton, M. D. Tobin, and J. L. Hopper, “Key concepts in genetic epidemiology,” *The Lancet*, vol. 366, no. 9489, pp. 941–951, 2005.
- [16] M. Cote, “Study designs in genetic epidemiology,” in *Tumour Biomarker Discovery*, vol. 520 of *Methods in Molecular Biology*, Humana Press, New Jersey, NJ, USA, 2009.
- [17] M. Korkeila, J. Kaprio, A. Rissanen, and M. Koskenvuo, “Effects of gender and age on the heritability of body mass index,” *International Journal of Obesity*, vol. 15, no. 10, pp. 647–654, 1991.
- [18] J. Akey, L. Jin, and M. Xiong, “Haplotypes versus single marker linkage disequilibrium tests: what do we gain?” *European Journal of Human Genetics*, vol. 9, no. 4, pp. 291–300, 2001.

- [19] M. D. Teare, *Genetic Epidemiology*, pp. 49–78, Springer, New York, NY, USA, 2006.
- [20] A. Kong and N. J. Cox, “Allele-sharing models: LOD scores and accurate linkage tests,” *American Journal of Human Genetics*, vol. 61, no. 5, pp. 1179–1188, 1997.
- [21] S. E. Hodge and D. A. Greenberg, “Sensitivity of lod scores to changes in diagnostic status,” *American Journal of Human Genetics*, vol. 50, no. 5, pp. 1053–1066, 1999.
- [22] K. Yoonhee, “Genetic epidemiology 101: a review of the statistical basis,” *BioWave Journal*, vol. 10, pp. 16–45, 2008.
- [23] D. A. Greenberg, “Inferring mode of inheritance by comparison of lod scores,” *American Journal of Medical Genetics*, vol. 34, no. 4, pp. 480–486, 1989.
- [24] N. M. Gatto, U. B. Campbell, A. G. Rundle, and H. Ahsan, “Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias,” *International Journal of Epidemiology*, vol. 33, no. 5, pp. 1014–1024, 2004.
- [25] L. R. Cardona and J. I. Bell, “Association study designs for complex diseases,” *Nature Reviews Genetics*, vol. 2, pp. 91–99, 2001.
- [26] D. C. Thomas, *Statistical Methods in Genetic Epidemiology*, pp. 61–93, Oxford University Press, New York, NY, USA, 2004.
- [27] K. G. Ardlie, L. Kruglyak, and M. Seielstad, “Patterns of linkage disequilibrium in the human genome,” *Nature Reviews Genetics*, vol. 3, no. 4, pp. 299–309, 2002.
- [28] S. A. Bacanu, B. Devlin, and K. Roeder, “Association studies for quantitative traits in structured populations,” *Genetic Epidemiology*, vol. 22, no. 1, pp. 78–93, 2002.
- [29] B. Devlin and N. Risch, “A comparison of linkage disequilibrium measures for fine-scale mapping,” *Genomics*, vol. 29, no. 2, pp. 311–322, 1995.
- [30] J. K. Pritchard and M. Przeworski, “Linkage disequilibrium in humans: models and data,” *American Journal of Human Genetics*, vol. 69, no. 1, pp. 1–14, 2001.
- [31] R. Bellamy, N. Beyers, and K. P. McAdam, “A genome-wide search for tuberculosis susceptibility genes in Africans,” *Proceedings of the National Academy of Sciences*, vol. 97, pp. 8005–8009, 2000.
- [32] D. C. Thomas and J. S. Witte, “Point: population stratification: a problem for case-control studies of candidate-gene associations?” *Cancer Epidemiology Biomarkers and Prevention*, vol. 11, no. 6, pp. 505–512, 2002.
- [33] S. Wacholder, S. Chanock, M. Garcia-Closas, L. El Ghormli, and N. Rothman, “Assessing the probability that a positive report is false: an approach for molecular epidemiology studies,” *Journal of the National Cancer Institute*, vol. 96, no. 6, pp. 434–442, 2004.
- [34] R. C. Lewontin, “On measures of gametic disequilibrium,” *Genetics*, vol. 120, no. 3, pp. 849–852, 1988.
- [35] N. Hamajima, H. Yuasa, K. Matsuo, and Y. Kurobe, “Detection of gene-environment interaction by case-only studies,” *Japanese Journal of Clinical Oncology*, vol. 29, no. 10, pp. 490–493, 1999.
- [36] J. H. Zar, *Bio Statistical Analysis*, Prentice Hall, New York, NY, USA, 5th edition, 2009.
- [37] H. Campbell and I. Rudan, “Interpretation of genetic association studies in complex disease,” *The Pharmacogenomics Journal*, vol. 2, pp. 349–360, 2002.
- [38] H. H. H. Göring, J. D. Terwilliger, and J. Blangero, “Large upward bias in estimation of locus-specific effects from genome wide scans,” *American Journal of Human Genetics*, vol. 69, no. 6, pp. 1357–1369, 2001.
- [39] P. S. Albert, D. Ratnasinghe, J. Tangrea, and S. Wacholder, “Limitations of the case-only design for identifying gene-environment interactions,” *American Journal of Epidemiology*, vol. 154, no. 8, pp. 687–693, 2001.
- [40] M. J. Khoury and W. D. Flanders, “Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case control studies with no controls!,” *American Journal of Epidemiology*, vol. 144, pp. 207–213, 1996.
- [41] A. M. Goldstein and N. Andrieu, “Detection of interaction involving identified genes: available study designs,” *Journal of the National Cancer Institute*, vol. 26, pp. 49–54, 1999.
- [42] C. L. Saunders, C. Gooptu, and D. T. Bishop, “The use of case-only studies for the detection of interactions, and the non-independence of genetic and environmental risk factors for disease (Abstract),” *Genetic Epidemiology*, vol. 21, p. 174, 2001.
- [43] C. L. Saunders and J. H. Barrett, “Flexible matching in case-control studies of gene-environment interactions,” *American Journal of Epidemiology*, vol. 159, no. 1, pp. 17–22, 2004.
- [44] T. A. Manolio, “Cohort studies and the genetics of complex disease,” *Nature Genetics*, vol. 41, no. 1, pp. 5–6, 2009.
- [45] D. W. Haas, “Genetic studies in clinical trials and observational cohorts,” *HIV PGX*, vol. 1, pp. 1–4, 2006.
- [46] B. S. Hulka and B. H. Margolin, “Methodological issues in epidemiologic studies using biologic markers,” *American Journal of Epidemiology*, vol. 135, no. 2, pp. 200–209, 1992.
- [47] R. G. Loudon and S. K. Spohn, “Cough frequency and infectivity in patients with pulmonary tuberculosis,” *American Review of Respiratory Disease*, vol. 99, no. 1, pp. 109–111, 1969.
- [48] M. Möller, E. de Wit, and E. G. Hoal, “Past, present and future directions in human genetic susceptibility to tuberculosis,” *FEMS Immunology and Medical Microbiology*, vol. 58, no. 1, pp. 3–26, 2010.
- [49] B. Simmonds, *Tuberculosis in Twins*, Pitman Medical, London, UK, 1963.
- [50] G. Madico, R. H. Gilman, W. Checkley et al., “Community infection ratio as an indicator for tuberculosis control,” *The Lancet*, vol. 345, no. 8947, pp. 416–419, 1995.
- [51] C. M. Stein, “Genetic epidemiology of tuberculosis susceptibility: impact of study design,” *PLoS Pathogens*, vol. 7, no. 1, Article ID e1001189, pp. 1–8, 2011.
- [52] X. Ma, R. A. Reich, J. A. Wright et al., “Association between interleukin-8 gene alleles and human susceptibility to tuberculosis disease,” *Journal of Infectious Diseases*, vol. 188, no. 3, pp. 349–355, 2003.
- [53] D. López-Maderuelo, F. Arnalich, R. Serantes et al., “Interferon- γ and interleukin-10 gene polymorphisms in pulmonary tuberculosis,” *American Journal of Respiratory and Critical Care Medicine*, vol. 167, no. 7, pp. 970–975, 2003.
- [54] S. Ryu, Y. K. Park, G. H. Bai, S. J. Kim, S. N. Park, and S. Kang, “3’UTR polymorphisms in the NRAMP1 gene are associated with susceptibility to tuberculosis in Koreans,” *International Journal of Tuberculosis and Lung Disease*, vol. 4, no. 6, pp. 577–580, 2000.
- [55] N. Risch, “Linkage strategies for genetically complex traits. II. The power of affected relative pairs,” *American Journal of Human Genetics*, vol. 46, no. 2, pp. 229–241, 1990.

- [56] L. Abel and A. J. Dessein, "Genetic epidemiology of infectious diseases in humans: design of population-based studies," *Emerging Infectious Diseases*, vol. 4, no. 4, pp. 593–603, 1998.
- [57] H. W. Deng, "Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits," *Genetics*, vol. 159, no. 3, pp. 1319–1323, 2001.