Chapter 11
Research Methods for Genetic Studies

Sadeep Shrestha and Donna K. Arnett

Abstract This chapter introduces the basic concepts of fundamental methods for genotype-phenotype association studies and relevant issues in interpretation of genetic epidemiology studies to clinicians. An overview of genetic association studies is provided, which is the current state-of-the-art for clinical and translational genetics. Discussion of future directions in this field is also included.

Keywords Genetic research • Genomics • Hardy Weinberg disequilibrium • Familial aggregation • Linkage disequilibrium • Genome-wide association

This chapter introduces the basic concepts of genes and genetic studies to clinicians. Some of the relevant methods and issues in genetic epidemiology studies are briefly discussed with an emphasis on association studies which are currently the main focus of clinical and translational genetics.

Genetics is the fundamental basis of any organism so understanding of genetics will provide a powerful means to discover hereditary elements in disease etiology. In recent years, genetic studies have shifted from disorders caused by a single gene (e.g. Huntington's disease) to common multi-factorial disorders (e.g. hypertension) that result from the interactions between inherited gene variants and environmental factors, including chemical, physical, biological, social, infectious, behavioral or nutritional factors.

A new field of science, Genetic Epidemiology emerged in the 1960s as a hybrid of genetics, biostatistics, epidemiology and molecular biology, which has been the major tool in establishing whether a phenotype (any morphologic, biochemical, physiologic or behavioral characteristic or trait of an organism) has a genetic component. A second goal of genetic epidemiology is to measure the relative size of that

S. Shrestha, Ph.D., MHS, M.S. (✉) • D.K. Arnett, Ph.D., M.S.
Department of Epidemiology, School of Public Health,
University of Alabama at Birmingham, Birmingham, AL, USA
e-mail: sshresth@uab.edu; Arnett@uab.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_11,

genetic effect in relation to environmental effects. Morton and Chung defined genetic epidemiology as

> a science that deals with the etiology, distribution, and control of disease in groups of relatives, and with inherited causes of disease in populations [1].

In the era of known human genome sequences from multiple individuals, genetic epidemiology methods have been instrumental in identifying the contribution of genes, the environment, and their interactions to better understand disease processes and biological mechanisms.

Genomic scientists have predicted that comprehensive, genomic-based care will become the norm, with individualized preventive medicine, early detection of illnesses and tailoring of specific treatments to an individual's genetic profile. Practicing physicians and health professionals must be knowledgeable in the principles, applications, and limitations of genetics to understand, prevent, and treat any biological disorders in their everyday practice. The primary objective of any genetic research is to translate information from individual laboratory tests to infer the relevance of segments of the human genome in relation to disease risk. This chapter will focus on the fundamental concepts and principles of genetic epidemiology that are important to help clinicians understand genetic studies.

Important Principles of Genetics

In the nineteenth century, long before DNA was known, an Augustinian clergyman, Gregory Mendel, described genes as the fundamental unit that transmits traits from parents to offspring [2]. Based on the observations from his cross-breeding experiments in his garden, Mendel developed some basic concepts on genetic information which still provides the framework upon which all subsequent work in human genetics has been based. Mendel's first law, referred to as the "*The principle of segregation*", basically states that alleles (alternate forms of the gene or sequence at a particular location of the chromosome) at one of the parent's genes segregate independently of the alleles from another parent. Mendel's law, therefore, states that alleles transmitted to an offspring are random (i.e., a matter of chance). It is now known that segregation of alleles occurs during the process of sex cell formation, known as meiosis. His second law is referred to as "*The principle of independent assortment*" which states that two genetic factors are transmitted independently of one another in the formation of gametes. As a result, new combinations of genes can be present in the offspring that are otherwise not possible in either of the parents. These two principles of inheritance and the concepts of dominance and recessive alleles established the foundation of our modern science of genetics. However, Mendel's law is not always true and there are exceptions to these rules, e.g. loci in the same chromosomes tend to transmit together, a key concept in modern genetic epidemiology.

All human cells except the red blood cells (RBC) have a nucleus that carries the individual's genetic information organized in chromosomes. Chromosomes are composed of molecules called deoxyribonucleic acid (DNA) which contain the basic instructions needed to construct proteins and other cellular molecules. Given the diploid nature, each human inherits one copy of the chromosome from the father and the other from the mother. Humans have 22 pairs of autosomal chromosomes and 2 sex-specific chromosomes (X and Y), where males have XY and females have XX chromosomes.

At the molecular level, DNA is a linear strand of alternating sugars (deoxyribose) and phosphate residues with one of four types of bases attached to the sugar. All information necessary to maintain and propagate life is contained within these four simple bases: adenine (A), guanine (G), thymine (T), and cytosine (C). In addition to this structure of a single strand, the two strands of the DNA molecule are connected by a hydrogen bond between two opposing bases of the two strands (T always bonds with A and C always bonds with G) forming a slightly twisted ladder, also referred as double helix. It was not until 1953 that James Watson and Francis Creek described this structure of DNA which became the foundation for our contemporary understanding of genes and disease.

The basic length unit of the DNA is one nucleotide, or one base pair (bp) which refers to the two bases that connect the two strands. In total, the human DNA contains approximately 3.3 billion base pairs and any two DNA fragments differ only with respect to the order of their bases. Three base units, together with the sugar and phosphate component (referred to as **codons**) translate into amino acids. According to the central dogma of molecular biology, DNA is copied into single stranded ribonucleic acid (RNA) in a process called transcription, which is subsequently translated into proteins. With the knowledge of underlying molecular biology, "*gene*" is defined as the part of the DNA segment that encodes a protein which forms the functional unit of the "hereditary" factor. It is now estimated that there are approximately 27,000 genes. The encoded proteins make intermediate phenotypes which regulate the biology of all diseases, so any difference in the DNA sequence could change the disease phenotype. In many species, only a small fraction of the total sequence of the genome encodes protein, and the function and relevance of the remaining noncoding sequences are still unknown. For example, over 98 % of the human genome is noncoding. However, the Encyclopedia of DNA Elements (ENCODE) project recently reported that over 80 % of DNA in the human genome has some biochemical function, most of which is still unknown. We are still in the infant stage of understanding the significance of the rest of these non-coding DNA sequence; however, the sequence could have structural purposes, or be involved in regulating the use of functional genetic information.

Units of Genetic Measure

Different genetic markers, which are a segment of DNA with a known physical location on a chromosome with identifiable inheritance, can be used as measures for genetic studies. A marker can be a gene, structural polymorphisms (e.g. insertion/deletion) or it can be some section of DNA such as short tandem repeat (STR)

Table 11.1 Some significant DNA sequence variants

Sequence variations	Description
Short Tandem Repeats (STR)	Tandemly repeated simple sequence motifs of 2–7 base lengths
Single Nucleotide Polymorphism (SNP)	Variations in a single nucleotide occurring in >1 % of the population
Structural variants	Variation in the structure of the chromosome, that includes deletions, inversions, rearrangements, copy number variations

and single nucleotide polymorphism (SNP). Recent advancements in molecular technology have resulted in the discovery of numerous DNA markers and the database of each marker is increasing daily. Polymorphism (poly = many and morphism = form) is a DNA sequence variation at any locus (any segment or region in the genome) in the population that has existed for some time and observed in at least 1 % of the population, whereas a mutation is often recent and the frequency in populations is less than 1 %. The terms mutation and polymorphism are often used interchangeably but mostly defined in the context of frequency. Variants within coding regions may change the protein function (missense) or predict premature protein truncation (non-sense) and as a result can have effects ranging from beneficial to mutual to deleterious. Likewise, although introns (intragenic regions between coding sequences) do not encode for proteins, polymorphisms can affect intron splicing or regulation of gene expression. To understand the role of genetic factors with any phenotype, it is important to understand these sequence variations among those with and without the phenotype within (population) and between (family) generations. We briefly describe the commonly used markers for genetic testing (Table 11.1).

Short Tandem Repeats (STRs)

STRs are tandemly repeated simple DNA sequence motifs of 2–7 bases in length that are arranged head-to-tail and are well distributed throughout the human genome, primarily in the intragenic regions. They are abundant in essentially all ethnically and geographically defined populations and are characterized by simple Mendelian inheritance. STR polymorphisms originate due to mutations caused by slipped-strand mispairing during DNA replication that results from either the gain or loss of repeat units. Mutation rates typically range from 10^{-3} to 10^{-5} events per gamete per generation, compared to single nucleotide rates of mutation of 10^{-7} to 10^{-9}. In humans, STR markers are routinely used in gene mapping, paternity testing and forensic analysis, linkage and association studies, along with evolutionary and other family studies. STRs have served as valuable tool for linkage studies of monogenic diseases in pedigrees, but have limited utility for candidate gene association studies.

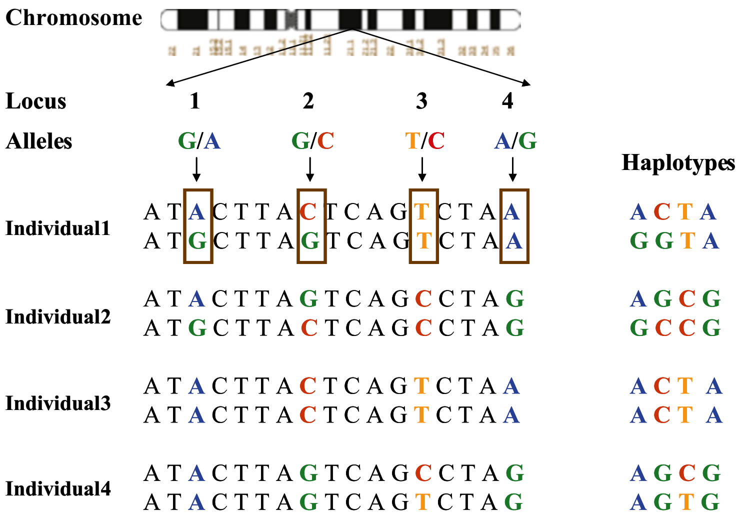

Chromosome
Locus
1
2
3
4
Alleles
G/A
G/C
T/C
A/G
Haplotypes
Individual1
A T A C T T A C T C A G T C T A A
A T G C T T A G T C A G T C T A A
A C T A
G G T A
Individual2
A T A C T T A G T C A G C C T A G
A T G C T T A C T C A G C C T A G
A G C G
G C C G
Individual3
A T A C T T A C T C A G T C T A A
A T A C T T A C T C A G T C T A A
A C T A
A C T A
Individual4
A T A C T T A G T C A G C C T A G
A T A C T T A G T C A G T C T A G
A G C G
A G T G

Fig. 11.1 Alleles and genotypes determined for bi-allelic Single Nucleotide Polymorphisms at four different loci and the corresponding haplotypes. At locus 1, G and A are the alleles; Individuals 1 and 2 have AG heterozygote genotype and Individuals 3 and 4 have AA homozygote genotype. If the phase is known as shown above, the haplotypes for individual 1 would be ACTA and GGTA. However, in most cases, the variant loci are not physically close and the assays may not be able to partition the phase, thus haplotypes are usually estimated with various methods

Single Nucleotide Polymorphisms (SNPs)

SNPs are the variations that occur at a single nucleotide of the sequence. Ninety percent of the polymorphisms in the genome are single nucleotide polymorphisms (SNPs). It has been estimated that there are over 17 million SNPs (1 in every 180 base pairs on average). Most of these variants have been identified through massive efforts of the International HapMap Project (2003) and the 1000 Genomes Project (2008). SNPs are the markers of choice for association studies because of their high frequency, low mutation rates and the availability of high-throughput detection methods. Most SNPs are found in the non-coding region and often have no known biological function, but may be surrogate markers or be involved in regulation of gene (e.g. expression and splicing). With few exceptions, the majority of the SNPs are bi-allelic and the genotypes (genetic makeup at both chromosomes) can be heterozygote (different allele in each chromosome) or homozygote (same allele in both chromosomes) for either allele (Fig. 11.1). All SNPs are catalogued centrally in major databases such as the dbSNP at the National Center for Biotechnology Information (NCBI) and given unique identifiers (rs#) for standard reference.

Structural Variants

The human genome consists of a myriad of structural variants that include deletions, duplications, inversions, translocations and copy number variations (CNVs) that can influence the functions of the encoded proteins. CNVs are the most common structural variants and have been associated several phenotypes and diseases.

It was generally thought that genes occurred in two copies in the genome. Recent studies have suggested that large segments of DNA, ranging from 1 kb to several million bp can vary in copy number, some of which contain several genes. Such CNVs are more common in the human genome than originally thought and can have dramatic phenotypic consequences as a result of altering gene dosage, disrupting coding sequences, or perturbing long-range gene regulation [3]. These regions are estimated to cover 5–20 % of the whole genome.

Although there are different genetic markers (as described above), SNPs are the most frequent variant in the genome and are widely used in genetic studies, so we will refer to SNP polymorphisms to explain the basic concepts in genetic epidemiology, especially in the context of association studies.

Terms and Basic Concepts in Genetic Epidemiology (Table 11.2)

Hardy-Weinberg Equilibrium (HWE)

HWE is one of the key concepts of population genetics that can be used to determine whether a genetic variant could be a valid marker in genetic epidemiology studies. In HWE, allele and genotype frequencies are related through the Hardy-Weinberg law which states that if two alleles, "A" and "a", at any locus with frequencies "p" and "q", respectively, are in equilibrium in a population, the proportions of the genotypes, "AA" homozygotes, "Aa" heterozygotes and "aa" homozygotes will be p^2, 2pq, and q^2, respectively. This law holds as a consequence of random mating in the absence of mutation, migration, natural selection, or random drift. One of the implications of HWE is that the allele frequencies and the genotype frequencies remain constant from generation to generation maintaining equilibrium in overall genetic variations. Extensions of this approach can also be used with multi-allelic and X-linked loci. Deviation from these proportions could indicate (a) genotyping error (b) presence of non-random mating, thus bias in the control selection (c) existence of population stratification (as described later) or (d) recent mutation, migration or genetic drift that has not reached equilibrium. Cases are more likely to represent the tail of a distribution of disease, and any putative genetic variant for that disease may not be in HWE; therefore, it is generally recommended to assess HWE in non-diseased (control) groups.

Table 11.2 Some commonly used genetic terms

Term	Brief description
Hardy-Weinberg Equilibrium (HWE)	Used to determine whether a genetic variant could be a valid marker
Linkage	When two genetic loci are transmitted together from parent to offspring more often than expected
Linkage Disequilibrium (LD)	The extent of non-random association between two genetic loci
Haplotype (Fig. 11.1)	A specific combination of alleles along a chromosome, one from the father and one from the mother
Epigenetic changes	Biochemical alterations in DNA that affect gene expression and function without altering DNA sequence
Transmission Disequilibrium Test (TDT)	Alleles of parents are used as "virtual control" genotypes
LOD score	Logarithm$_{10}$ of odds-the likelihood of observing a segregation pattern of recombination frequency compared to chance

Hardy-Weinberg equilibrium: The stable frequency distribution of genotypes, AA, Aa, and aa, in the proportions p^2, 2pq, and q^2 respectively (where p and q are the frequencies of the alleles, A and a, respectively) that results from random mating in a population in the absence of mutation, migration, natural selection, or random drift

Linkage: co-segregation of alleles at two or more loci (family-based)

Linkage disequilibrium: the extent and associations of non-randomness of alleles at two/more loci in a population

Haplotype: A set of closely linked genetic markers present on one chromosome which tend to be inherited together (e.g. Fig. 11.1 – ACTA and GGTA for individual 1)

Epigenetic Changes: genetic control of the expression and activation of genes that involves factors other than changes in DNA sequence

Transmission Disequilibrium Test (**TDT**): a test that measures overtransmission of alleles from parents to offspring with the disease/trait (more frequently than expected by chance)

LOD Score: Logarithm$_{10}$ odds of likelihood of observing the segregation pattern of the marker alleles at a given recombination frequency (linked) to the likelihood of the same segregation pattern in the absence of linkage (by chance)

Linkage and Linkage Disequilibrium (LD)

Linkage and LD are the *sine qua non* of genetic epidemiology. While genes in different chromosomes segregate, Thomas Hunt Morgan and his co-workers observed that genes physically linked to one another on chromosomes of drosophila tended to be transmitted together. This phenomenon, where two genetic loci are transmitted together from parent to offspring more often than expected under independent inheritance, is termed linkage. Linkage was first demonstrated in humans by Julia Bell and J.B.S Haldane who showed that hemophilia and color blindness tended to be inherited together in some families [4]. Two loci are linked if recombination (exchange of genetic information between two homologous chromosomes during meiosis) occurs between them with a probability of less than 50 %. Recombination is inversely related to the physical distance between the two loci. However, after several generations, successive recombinations (especially in regions of recombination hotspots) may lead to complete independence even between loci that may be physically close together.

In population genetics, LD is defined as the extent of non-random association between two genetic loci such that the presence of one allele at a locus provides information about the allele of the other loci [5]. The level of LD in a population is influenced by several factors including genetic linkage, the rate of recombination, mutation, random genetic drift, selection, non-random mating and population admixture. Many different measures of LD have been proposed in the literature, most of which capture the strength of association between pairs of SNPs. Although concepts of LD date to early 1900s, the first commonly used LD measure, D’ was developed by Richard Lewontin in 1964. D’ measures the departure from allelic equilibrium between separate loci on the same chromosome that is due to the genetic linkage between them. The other pairwise measure of LD used in association studies is r^2 also denoted as Δ^2.

For two loci with alleles A/a at the first locus and B/b at the second allele, D is estimated as follows:

$$D = p_{AB} - p_A p_B \tag{1}$$

The disadvantage of D is that the range of possible value depends greatly on the marginal allele frequency. D′ is a standardized D coefficient and is estimated as follows:

$$D' = \frac{D}{D_{\max}} \tag{2}$$

If $D > 0$, $D_{\max} = \min\ [P_A(1-P_B), P_B(1-P_A)]$
If $D < 0$, $D_{\max} = \min[P_A P_B, (1-P_A)(1-P_B)]$s
and r^2 is the correlation between two loci and is estimated as follows:

$$r^2 = \frac{D^2}{p_A p_a p_B p_b} \tag{3}$$

Both D′ and r^2 range from 0 (no disequilibrium) to 1 (complete disequilibrium), but their interpretation is slightly different. In the case of true SNPs, D′ equals 1 if just two or three of the possible haplotypes are present and is <1 if all four possible haplotypes are present. On the other hand, r^2 is equal to 1 if only two haplotypes are present. Association is best estimated using the r^2 because it acts as a direct correlation to the allele at the other SNP. Additionally, there is a simple inverse relationship between r^2 and the sample size to detect association between susceptibility loci and SNPs.

Haplotype

Haplotype is a specific combination of alleles along a chromosome, one inherited from the mother and the other from the father (Fig. 11.1). Recent studies have shown that the human genome can be parsed into discrete blocks of high LD interspersed by shorter regions of low or no LD. Only a small number of characteristic ("tag") SNPs are sufficient to capture most of the haplotype structure of the human genome in each block. Tag SNPs are loci that can serve as proxies for many other SNPs such that only a subset of loci needs to be genotyped to obtain the same information and power obtained from genotyping a larger number of SNPs. The SNPs within the same block show a strong LD pattern while those in different blocks generally show a weak LD pattern. This advantage, along with the relatively smaller number of haplotypes defined by tag SNPs in each block provides another way to resolve the complexity of haplotypes.

High LD between adjacent SNPs, also result in a much smaller number of haplotypes observed than the theoretical number of all possible haplotypes (2^n haplotypes for n SNPs). There is also biological evidence that several linked variations in a single gene can cause several changes in the final protein product and the joint effect can have an influence on the function, expression and quantity of protein resulting in the phenotype variation. The most robust method to determine haplotypes is either pedigree analysis or DNA sequencing of cloned DNA. Both of these methods are limited by data collection of families or intensive laboratory procedures, but the **phase** (knowledge of the orientation of alleles on a particular transmitted chromosome) of the SNPs in each haplotype can be directly determined. Haplotypes can also be constructed statistically, although constructing haplotypes from unrelated individuals is challenging because the phase is inferred rather than directly measured. Unless all SNPs are homozygous or at most only one heterozygous SNP is observed per individual, haplotypes cannot be discerned. To account for ambiguous haplotypes, several statistical algorithms have been developed [6]. Three common algorithmic approaches used in reconstructing population-based haplotypes are (i) a parsimony algorithm, (ii) a Bayesian population genetic model that uses coalescent theory, and (iii) a maximum likelihood approach that is based on expectation-maximization (EM) algorithm. The details of these methods are beyond the scope of this book, but readers are referred to the book "Computational Methods for SNPs and Haplotype Inference" [6] for further discussion. Recent haplotype estimation methods often use a hybrid approach of EM and Bayesian models.

Biological Specimens

Although the focus of this chapter is not on the laboratory methods of specimen collection, we briefly describe the samples used in clinical studies and their importance. Clinicians deal with different biological organs and tissues in their everyday practice. Most of these however may not be an efficient or convenient source for DNA, the most commonly used resource for genetic studies. Based on factors including cost, convenience for collection and storage, quantity and quality of the source, DNA is commonly extracted from four types of biological specimens: (1) dried blood spots collected in special filter paper (2) whole blood collected in ethylenediaminetetraacetic acid (EDTA) or other anticoagulants such as heparin and acid citrate dextrose (ACD) (3) lymphocytes isolated from whole blood and EBV-transformed for unlimited source of DNA and (4) buccal epithelial cells collected from swabs or mouth-washes (non-invasive and child-friendly). In certain circumstances, samples derived from surgery or other treatment or therapy procedures can also be used for extracting DNA. For instance, formalin-embedded samples of biopsies can be used; however, special laboratory protocols or reagents may be needed (for instance to process the DNA crosslinking).

Ethical, Legal and Social Implications (ELSI)

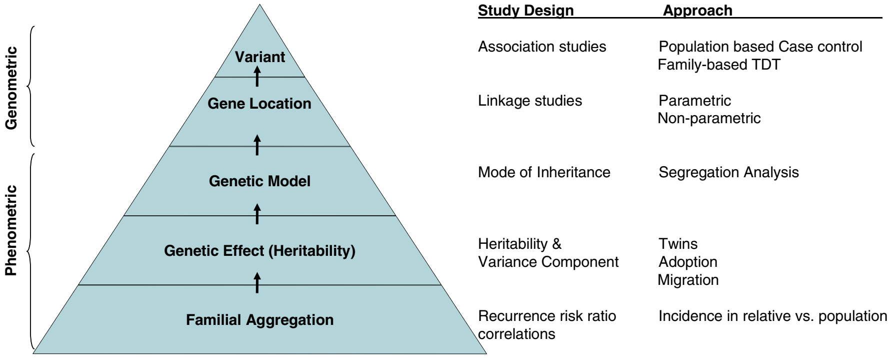
Even for well-intentioned research, one can raise legitimate concerns about the potential misuse of genetic data in regard to social status, employment, economic harm and other factors. A significant amount of work has been done on ethical, legal and social implications (ELSI) research of genetics and policies, but ethics remains an area of major concern. All research protocols can only be conducted upon approval from an institutional review board (IRB) with an appropriate informed consent from the participants. Pediatric genetic research often can be cumbersome as it may require approval from both parents or the legal guardians. It is a routine practice to label the samples with unlinked coded identifiers rather than personal identifiers, so that the individual's identity is masked when linking to phenotypic, demographic, or other personal information. The confidentiality of the DNA results needs to be maximized to protect individual privacy. All reports of genetic studies including manuscripts and grants often require detailed description of ethical concerns and data protection.

Measurable Outcome and Phenotype

Phenotype is an observable and measurable trait which can be defined qualitatively or quantitatively and does not necessarily have to be related to a disease. Some traits or diseases, like the simple Mendelian traits, have a distinctly measurable phenotype definition. However, other illnesses (e.g. psychiatric disorders) are complex to define and require various symptoms and clinical criteria that may have different biological system and pathways combined. The misclassification of cases and controls can be a major problem in any study that can easily introduce biases and inconsistencies between studies. Phenotypes can be defined qualitatively (absent or present) or measured quantitatively. A qualitative trait can be categorized into two or more groups. For example, qualitative traits can be dichotomous (e.g. HIV^+ vs. HIV^-), ordinal (low, average and high blood pressure group) or nominal (green, black, blue eyes) based on certain distinct criteria. On the other hand, measurable physiological quantities such as height, blood pressure, serum cholesterol levels, and body mass index (BMI) can vary among different individuals. Often it may be difficult to examine the genetic effect of quantitative measures; however, they can be transformed into meaningful qualitative values where the genetic effect can be more distinct. To make the quantitative traits more interpretable through statistical analyses, the overall distribution in a given population is viewed graphically. Often these distributions produce a familiar bell-shaped curve (normal distribution), where several statistical methods can be used to assess the effect of genotypes. For example, the individuals at the extreme tails of the curves can have different genetic distributions. Some diseases may also have intermediate phenotypes that can be measured with molecular markers, while others are strictly based on clinical diagnoses. For example, blood cholesterol levels which can be precisely measured may be a better intermediate outcome of cardiovascular disease than a self reported "headache" where the symptoms may be heterogeneous in the population and the measurement is subjective. Other measures, including exposures (e.g. HIV viral load) can define a phenotype better than the clinical symptoms since virally infected individuals can be asymptomatic for undefined period of time. In that specific example, everyone positive for HIV virus test could be defined as the outcome of interest (cases) while in another scenario specific clinical symptoms of HIV infection (e.g. immune cell counts or viral load) could define case status. Even among phenotypes with clinical diagnoses, some have distinct symptoms or signs, with high sensitivity tests, whereas others do not. Some diseases, like Alzheimer's, can have phenotypic heterogeneity, where the same disease shows different features in different families or subgroups of patients. Like in any other clinical study, the key to a genetic study is a clear and consistent definition of the phenotype with underlying biology. Since the main interest in conducting genetic study is to see how variants that change the expression and encoding of protein is related to the biology of the disease, the phenotype has to be clearly defined.

General Methods in Clinical Genetic and Genetic Epidemiology Studies

In the past 20–30 years, epidemiologic methods and approaches have been integrated with those of basic genetics to identify the role of genetic factors in disease occurrence in families and populations [7]. Family studies examine the rates of diseases in the relatives of proband cases versus the relatives of internally matched controls. For a quantitative trait, such as blood pressure, we can measure correlation of trait values among family members to derive estimates of heritability. Mendelian diseases are transmitted in families and recur in the relatives of affected individuals more frequently (10^3–10^6 fold) compared to the general population. In contrast, diseases such as cancer, Alzheimer's Disease, and myocardial infarction are quite common among older adults; however, their occurrence does not follow Mendelian inheritance patterns, but rather are multifactorial with several interactions between environment and genetic factors.



Study Design	Approach
Association studies	Population based Case control Family-based TDT
Linkage studies	Parametric Non-parametric
Mode of Inheritance	Segregation Analysis
Heritability & Variance Component	Twins Adoption Migration
Recurrence risk ratio correlations	Incidence in relative vs. population

Fig. 11.2 Systematic designs and approaches in genetic epidemiology studies to identify the genetic and non-genetic causes of disease

The first step in clinical or epidemiologic genetic studies is to determine whether a phenotype of interest is controlled by a genetic component. There are five key scientific questions that are addressed in sequence in genetic epidemiologic studies (Fig. 11.2): (1) Is there familial clustering? (2) Is there evidence of genetic effect? (3) Is there evidence for a particular genetic model? (4) Where is the disease gene? (5) How does this gene contribute to disease in the general population? The first three questions do not require DNA data and are referred as phenometric studies, but the latter two depend on DNA and referred as genometric studies.

Familial Aggregation

The first step to determine whether a phenotype has a genetic component is to examine the clustering within families. Familial aggregation estimates the likelihood of a phenotype in close relatives of cases compared to the non-cases. If the phenotype is a binary trait, familial aggregation is often measured by the relative recurrence risk. The recurrence risk ratio is the ratio of prevalence of the phenotype in relatives of affected cases to the general population. Greater risk associated with closer degrees of relatedness could also indicate the genetic component. If the prevalence of the phenotype is higher in 1st degree relatives (father, mother, siblings) versus 2nd degree relatives (uncle aunt, cousins) it would suggest a genetic component since the 1st degree relatives share more genetic information than the 2nd degree relatives. For example, cancer and heart disease tend to run in families, as measured by measurements such as relative risk. On the other hand, assessment of familial aggregation of a continuous trait, such as height, can be estimated with a correlation or covariance-based measure such as intrafamily correlation coefficient (ICC). The ICC indicates the proportion of the total variability in a phenotype that can reasonably be attributed to real variability between families. Disease or traits may cluster in families; however, this does not necessarily mean that they share the common genetic factors. Since families often share the same household or geographic region they share common cultural attitudes, socioeconomic status, diet and environmental exposures – all of which can be known or unknown and may not be easily measured. It is difficult to disentangle the genetic effect from the environmental effect due to this shared physical environment. For example, obesity could be due to shared genes within the family or the eating or physical activity habits in the family.

Genetic Effect

Once the familial aggregation is established, the next step is to distinguish between genetic and non-genetic factors and estimate the extent of genetic effect. Different variance component models estimate heritability, which is defined as the proportion of variation directly attributable to genetic differences among relatives to the total variation in the population (both genetic and environmental). Although traditionally used to estimate the genetic effect in familial aggregation, it is a theoretical concept. Heritability is population-specific and must be used with caution when comparing different populations. Other classical designs for distinguishing non-genetic family effects from genetic effects have been studies of twins, adoptees and migrants.

Twin studies

Studies of twins are useful in estimating the contribution to a phenotype through the comparison of monozygotic (MZ) pairs (who share all genes) with dizygotic (DZ) pairs (who share on average half of their genes). If family upbringing acts equally on monozygotic twins as it does on dizygotic twins, then the greater similarity of phenotypes in MZ than DZ twins is attributed to genetic factors. While MZ twins reared together have the same genetic and environment exposures, MZ twins separated at birth and raised apart will have different environment exposures but same genetics. Thus, such studies will provide insights into the contribution of strong environment factors in common diseases such as substance abuse and eating disorders. In contrast, DZ twins may have a similar genetic makeup as other siblings, but they share the same womb, so early environmental exposure related studies can be conducted with these pairs. Concordance rates are used in twin studies which measures and compares the frequency of disease occurrence between MZ and DZ twins. For example, the concordance rate of sickle cell disease among MZ is 100 %, indicating pure genetic effect; whereas Type I Diabetes is 25–35 % among MZ, 5–6 % among DZ twins or siblings and 0.4 % among the general population, suggesting both genetic and environment effects.

Adoption Studies

This study design examines the similarity and differences in the phenotype in the biological parents and foster parents of adoptees, and in their biological and adopted siblings, respectively. The assumptions are that the similarity between an adopted child and biological parent is primarily due to genetic effects, while the similarity between the adopted child and the adoptive parent or adoptive siblings is mainly due to the shared environment since they do not share genetic background as they are not biologically related.

Migration Studies

While with modern globalization, humans are constantly travelling, we are also moving to new areas in search of better opportunities. Patterns in environmental exposures in different areas among different ethnic groups or related family members can be assessed to make some inferences about genetic and environmental influence in phenotypes or diseases. A similar incidence of phenotype or disease in migrants compared to the aboriginal population's incidence suggests a strong environmental factor, whereas similar incidence to the original ethnic group or relatives in the original residence could suggest a genetic effect. Genes do not change as easily as environmental exposures, so the variation in the phenotype after taking into account all the common and new environmental factors could point to a genetic effect.

Genetic Model

After the genetic basis is established, the next step is to find the mode of inheritance which historically was done using segregation analyses, although these methods are not as common in the era of SNP association studies. Segregation analyses does not use DNA-based genetic data, but rather, the methods test whether or not the observed phenotype follows a Mendelian inheritance in the offspring in the pedigree. Mendelian diseases can be autosomal dominant, autosomal recessive, X-linked dominant, or X-linked recessive (usually with high penetrance and low frequency of risk alleles). Traditional segregation analysis primarily studied simple Mendelian disorders where a single gene mutation is sufficient and necessary to cause a disorder. However, most common chronic diseases are regarded as complex where a large number of genetic variants along with environmental factors interact with each other (necessary or un-necessary but not sufficient) to affect the disease outcomes. These diseases usually cluster in families, but do not follow a traditional Mendelian inheritance pattern. While segregation analyses are powerful to test different modes of Mendelian inheritance in the family, it is not useful for complex traits. Linkage and association analysis, both of which utilize DNA, are more powerful to study genetic effects of complex diseases.

Disease Gene Location

Linkage Studies

Linkage studies are performed based on the principle that alleles at two nearby loci on the genome tend to be transmitted together from parent to offspring. Linkage analysis are often the first stage in genetic epidemiology studies to identify broad genomic regions that contain gene or genes that predispose to the phenotype, in the absence of previous biologically driven hypotheses. Genetic linkage analysis tests whether the marker segregates with the disease in pedigrees with multiple affected individuals, according to a Mendelian mode of inheritance. The approach relies entirely on the tendency for genomic regions that affect the phenotype to be passed on to the next generation intact, without recombination events at meiosis. If a marker is passed down through family generation and occurs more commonly among those with the phenotype, then the marker can be used as a surrogate for the location of the gene.

Two types of linkage analysis can be performed: parametric and nonparametric analysis. Parametric linkage analysis involves testing whether the inheritance patterns fits a specific model and is traditionally measured with a statistical test, LOD score (logarithm (base 10) of odds) – $L(\theta)/L(\theta=0.5)$ i.e., the likelihood of observing the segregation pattern of the marker alleles at a given recombination frequency θ (linked) compared with the likelihood of the same segregation pattern in the absence of linkage (by chance). While the approach is very powerful, the study design can be challenging logistically since as it requires recruitment of families (with history of the phenotype) to estimate a number of recombination occurrences in order to calculate the LOD score. STRs with multiple alleles are more powerful for linkage studies than SNPs, which are mostly biallelic. The objective of parametric linkage analysis is to estimate the recombination frequency (θ) and to test whether θ is less than 0.5, which is the case when two loci are genetically linked. The nonparametric approach evaluates the statistical significance of excess allele sharing for specific markers among affected sibs and does not require information about the mode of disease inheritance. With this approach, often the inheritance pattern is measured in terms of identical by descent (IBD), where the same allele is inherited from a common ancestor, and identical by state (IBS), where the allele is the same but not necessarily inherited from the same ancestor. Thus, these methods are based the fact that affected relatives have a higher probability of sharing genes IBD at or near a locus of susceptibility allele/gene to a disease than sharing an unlinked locus. The genes contributing to the phenotypic variation have been successfully localized by linkage analysis for Mendelian diseases that have a strong genetic effect and are relatively rare (e.g. cystic fibrosis, Huntington disease). However, for more complex and common diseases (e.g. cancer, cardiovascular diseases), linkage analysis has had less success. The method of choice for complex genetic diseases has evolved to association studies which are followed by fine-mapping studies to narrow down the putative disease locus.

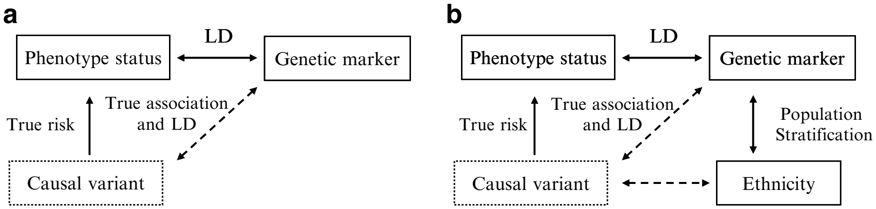


Fig. 11.3 True association, LD and the effect of population stratification. (**a**) Genetic marker that is in LD with causal variant serves as a surrogate of the true association with the phenotype. (**b**) Population stratification is a confounder that leads to spurious association

Association Studies

Genetic association studies aim to correlate differences in allelic frequencies at any locus with differences in disease frequencies or quantitative traits [8]. Genetic association occurs if the specific genetic variant is more frequent in the affected group than the non-affected group. Most association studies represent classical case–control approaches where the risk factor under investigation is the allele at the genetic marker (mostly with SNPs). SNP-based association studies can be performed in two ways: (i) direct testing of an exposure SNP with a known varying function such as altered protein level or structures and (ii) indirect testing of a SNP which is a surrogate marker for locating adjacent functional variant that contributes to the phenotype or disease state (Fig. 11.3a). The first method requires the identification of all "functional" variants in coding and regulatory regions of genes. The latter method avoids the need for cataloguing potential susceptibility variants by relying instead on association between disease and neutral polymorphisms tagging a SNP near a risk-conferring variant. It exploits the phenomenon of linkage disequilibrium (LD) between alleles of closely linked loci within the genomic regions.

Given the diallelic nature of majority of the SNPs, a disease locus may be difficult to identify unless the surrogate marker is closely linked to the disease locus. Apart from a single SNP association strategy, a dense panel of SNPs from the coding and non-coding regions of the gene that form haplotypes can also be tested. Some studies have also demonstrated that the analysis of haplotypes rather than individual SNPs can detect association with complex diseases. It has been suggested that single SNP-based candidate gene studies may be statistically weak as true associations may be missed because of the incomplete information from individual SNPs. For example, haplotypes contain more heterozygosity than any of the individual markers that comprise them and also mark more of the variation in the gene than single SNPs. Several haplotype association studies have shown the power of haplotypes over individual SNPs as it can either combine multiple causal variants or tag a less common causal variant than a more frequent single SNP.

Candidate Gene vs. Genome Wide Association Studies (GWAS)

Candidate gene approaches examine polymorphisms in genes with potential biological mechanisms or pathways related to the phenotype of interest. Some of the candidate genes are also based on physical location or sequence homology to a gene encoding protein that is in the etiologic pathway. As attractive as this hypothesis-driven candidate gene approach is, it focuses exclusively on the relatively few known genes, ignoring many that have not yet been characterized to play a role, suffering from potential publication bias in the process of selection of the genes. One major drawback of candidate gene approach is that *a priori* knowledge of the pathogenesis of the disease is required – when the molecular mechanism is poorly understood or complex, it could lead to selection of the wrong genes. Even with the right genes within the pathway, the challenge is to find variants that influence the regulation of gene function. Candidate gene studies have proven to be more successful when used as a follow-up of linkage studies. For example, APOE4, the most common genetic factor associated with Alzheimer's disease, was primarily discovered by candidate gene approach following the linkage study which mapped to chromosome 19.

Alternatively, with assurance of adequate power, hypothesis-generating genome wide association studies (GWASs) have been widely used. While the study design and methodological approaches are the same as for the candidate gene approach [8], GWAS studies rely on the microarray chips that consists of thousands to millions of genomic variants that has resulted from large projects such as the HapMap, 1000 Genome Project, and continuing sequencing efforts by various groups and investigators. Technological advances have dramatically resulted in cost-effective high-throughput genotyping arrays making GWAS more promising and attractive. GWAS has the advantage in the sense that no *a priori* knowledge of the structure or function of the genes involved is required. Additionally, with complex statistical models, untyped SNPs can also be imputed using GWAS data, which has been proven to be very reliable for common variants. Hence, this approach provides the possibility of identifying variants and genes that influence the phenotype or the disease that had previously not been biologically suspected. A two step design has often been used by researchers where common variation is first screened for association signals using cost-effective typing of tagging SNPs with GWAS followed by denser sets of SNPs in regions of potentially positive signals. If the sample size is large enough, a third stage of validation of association can also be conducted with proper power calculations. Although promising results have been found for different phenotypes with GWAS, analytical considerations are still underway to develop a robust strategy to interpret the findings especially for complex diseases with multiple gene-gene and gene-environmental interactions. Such large datasets still require new methods and approaches to understand the true biology of the phenotype. A lot of emphasis has been made towards using stringent statistical criteria for handling false positive issues; however, new biologically-driven methods are required to dissect such large datasets to understand and identify the complex nature of common diseases.

Risk Quantification

Gene-Gene and Gene-Environment Interaction

A central theme of genetic epidemiology is that human disease is caused by interactions within and between genetic and non-genetic environmental factors. Thus, in the design and analysis of epidemiologic studies, such interaction needs to be explicitly considered. A simple approach would be to create a classic 2×2 table with genotypes at the two loci classified as present or absent and compute odds ratios for all groups with one reference group. The extent of the joint effect of two loci can be compared with the effects for each locus independently. The same approach can be considered for gene-environmental interaction for qualitative measurements. However, as more genes are involved and the environmental exposure is quantitatively measured, the analysis and interpretation of the interaction can be complicated, but various methods are being continuously developed. Large sample sizes are needed to observe true interactions, especially if they are small effects.

Gene Contribution

Once the association of the genetic allele is discovered, it is important to assess the contribution of this variant to the phenotype. The public health relevance of a given polymorphism is addressed by estimating the proportion of diseased individuals in the population that could be prevented if the high-risk alleles were absent (known as attributable fraction, etiologic fraction, or population attributable risk percent). Accurate estimation of the population frequency of the high-risk variant (allele and/or genotype) is important because the **attributable fraction** is a function of the frequency of the high-risk variant in the population and the penetrance (i.e., the likelihood that the trait will be expressed if the patient carries the high-risk variant). Attributable fractions can also be used to estimate the proportion of disease that is a result of the interaction of a genetic variant and an environmental exposure. Genetic variants are not usually modifiable within the longevity of an individual (although very possible evolutionarily over time in populations); therefore the prevention of disease will depend on interventions that target environmental factors that interact with genetic susceptibility to influence the risk of disease.

Additional Applications of Genetic Studies

Most of the genetic studies (candidate or genome-wide) are focused on case–control designs with the underlying goal of understanding the biological cause of the disease. Other time dependent studies can be performed to understand the genetic effect in the natural history or progression of the disease. The outcomes of these studies are helpful for providing counseling to individuals about their offspring (genetic screening) or the interaction between environmental factors. However, there are a growing number of genetic studies examining the differential response to drugs or vaccines, with potential application of translational science. For instance, "**pharmacogenetic**" studies focus on genetic determinants of individual variation in response to drugs, including variation in the primary domain of drug action and variation in risk for rare or unexpected side effects of drugs. Likewise, "**vaccinogenetic**" studies examine the genetic determinants of differential vaccine response (e.g. antibody titer) and side effects between individuals.

Beyond Association Studies

While other factors such as epigenetic and regulatory factors are beyond the scope of this chapter, it is important to understand that association studies itself may not fully delineate the genetic effect on a disease. Epigenetic changes are biochemical alterations in DNA that affect gene expression and function without altering the underlying DNA sequence. DNA methylation is one epigenetic process implicated in human disease that involves methylation of cytosine, usually at CpG dinucleotides. Micro-array methods are available to capture the methylation patterns across genes that could help in addition to the variant findings. Recent insights of the ENCODE project has helped shift focus to complex molecular mechanisms by which genetic factors such as microRNAs (miRNAs) may regulate genes. MiRNAs are evolutionarily conserved small non-coding RNAs (~22 bp) that inhibit translation of proteins by binding to the target transcript in the 3′ untranslated region. It has been estimated that miRNAs contribute to expression of over 60 % of protein coding genes in humans. In this regard, as the testing costs are being lowered, it may be beneficial to perform whole genome sequencing (versus GWAS or even exome-sequencing that targets variants in the exons of all known genes) that will provide information on all the known and the unknown variants of the human genome. Although new approaches and analytical methods are warranted to fully understand the genome, sequencing data will provide both rare and common variants in both genic and non-genic regions which can have regulatory or unknown functions, as suggested by ENCODE.

Major Issues and Limitations in Genetic Studies

In most cases with complex diseases, the effect of any genetic variant is small and can only be observed in studies with a large sample size or the frequency of the allele is rare and has a large relative risk. There are very few common variants (>10 % allele frequency) with a relative risk exceeding 2 (e.g. APOE and Alzheimer's disease). A major concern with respect to genetic association studies has been lack of replication, especially contradictory findings across studies. Replication of findings is very important before any causal inference can be drawn. For example, since 2005, over 1,600 publications have identified more than 2,000 genetic associations with approximately 300 common diseases and traits, but many of these studies need to be replicated. Several study design and statistical issues need to be seriously considered when conducting genetic studies which are briefly described below:

Genetic Heterogeneity

There are several cases where multiple alleles at a locus are associated with the same disease. This phenomenon is known as **allelic heterogeneity** and can be observed with a multi-allelic locus. This may explain why in some studies one allele is associated with the disease and in other studies it is another allele. Likewise, locus heterogeneity may also exist where multiple genes influence the disease independently and thus a gene found to be associated in one study may not be replicated in the other but rather another gene may be associated.

Confounding

One crucial consideration in genetic studies is the choice of an appropriate comparison group. In general, as in any well-designed epidemiological case–control study, controls need to be sampled from the same source population as the cases. The use of convenient comparison groups without proper ascertainment criteria may lead to spurious findings as a result of confounding caused by unmeasured genetic and environmental factors. Population stratification can occur if cases and controls are not matched by ethnicity or if individuals have differential admixture (the proportions of the genome that have ancestry from each subpopulation). Stratification can results when phenotypes of interest differ between ethnic groups (Fig. 11.3b). Although most genetic variation is inter-individual, there is also significant inter-ethnic variation irrespective of disease status. One classic example is reported by Knowler et al. [9] who showed spurious inverse association between variants in the immunoglobulin haplotype Gm3;5,13,14 and non-insulin dependent diabetes mellitus among the Pima-Papago Indians [9]. Individuals with the haplotype Gm3;5,13,14 had a higher prevalence of diabetes than those without it (29 % vs.8 %). This haplotype, however, measured the subjects' degree of Caucasian genetic heritage and when the analysis was stratified by degree of admixture, the association did not exist.

One way to overcome such issue of confounding by population stratification is to conduct family based designs with special statistical analyses such as transmission-disequilibrium test (TDT). Basically, in TDT, alleles of parents not transmitted to the patients are used as "virtual control" genotypes so any population-level allele frequency differences become irrelevant. Several other family-based and population-based methods have also been derived from TDT. While these methods are attractive because they correct false positives from population stratification, family-based samples are difficult to collect and might not be feasible for late-onset diseases where the parents might be deceased. Another approach is to use a "homogeneous" population. In recent years, there is growing interest to study genetically isolated populations such as Finland and Iceland. These populations have been isolated for several years and expanded from a small group of individuals called "**founder population**". Founder population limits the degree of genetic diversity making more or less a homogenous population. One major limitation of finding from such isolated population is the generalizability to other populations which may have different genetic make-ups.

Studies have shown that there is admixture even within such isolated populations. An alternate method to control for population stratification is to use unrelated markers from the non-functional region of the genome as indicators of the amount of background diversity in individuals. The first approach, referred as "**genomic control**", measures the extent of inflation due to population stratification and this value can be adjusted in the standard analyses. The second approach would be inferring genetic ancestry, by either the structured-association approach where individuals are assigned to subpopulation clusters using model-based clustering program such as STRUCTURE; or infer population structure with principal component analysis (PCA). Either association analyses are performed by stratifying clusters or covariates derived from ancestry information are adjusted in the analyses.

Genotype Error and Misclassification

For family-based studies (trio data for TDT), genotyping errors have been shown to increase type I and type II errors and for population-based (case–control) studies it can increase type II errors and thus decrease the power. Additionally, misclassification of genotypes can also bias LD measurements.

In general, genotyping errors could be a result of poor amplification, assay failure, DNA quality and quantity, genomic duplication or sample contamination. It is important that a quality-check be performed for each marker and the low-performance once be removed from the analysis before the results are interpreted. Several laboratory based methods such as (a) genotyping duplicate individuals (b) genotyping the same individuals for the same marker using different assay platforms or (c) genotyping in family pedigrees to check for Mendelian inconsistency, (i.e. the offspring should share the genetic makeup of the parents and any deviation could indicate genotype error) can be used to assure the quality of the genotypic data. Testing for HWE is also commonly used, however it is important to note that deviation from HWE does not necessarily indicate genotype error and could be due to any of the underlying causes as described earlier.

Multiple Testing

Regardless of whether each SNP is analyzed one at a time or as part of a haplotype, the number of individual tests can become very large and can lead to an inflated (false positive) type I error rate both in candidate gene approach and whole genome approach. If the selected SNPs are all independent, then adjustments to the conventional p-value of 0.05 with Bonferroni correction could account for the multiple testing. However, given the known LD pattern between SNPs, such adjustments would overcorrect for the inflated false-positive rate, resulting in a reduction in power. An alternate method would be to use the False Discovery Rate (FDR) approach which rather than correcting the p-value, corrects for fraction of false-positives with the significant p-value. When a well defined statistical test is performed (testing a null against an alternative hypothesis) multiple times, the FDR estimates the expected proportion of false positives from among the tests declared significant. For example, if 100 SNPs are said to be significantly associated with a trait at a false discovery rate of 5 %, then on average 5 are expected to be false positives. However, the gold standard approach that is being appreciated more is the permutation testing where the groups status of the individuals are randomly permuted and the analysis repeated several times to get a distribution for the test statistics under the null hypothesis but this method can also be computationally intensive and time-consuming.

Concluding Remarks

The completion of the Human Genome Project in 2003 heightened expectations of the health benefits from genetic studies [10]. Other projects such as the HapMap and 1000 Genome projects have complemented knowledge from the Human Genome Project. The markedly low cost to sequence the genome has provided additional information from various projects, which was not possible a few years ago. The ENCODE project has furthered our knowledge that previously thought “junk” DNA sequences are important as they have regulatory and other unknown functions. While the known genetic factors and methods drive our paths ahead, all the unknown factors make us all strive to answer the multitude of important translational questions in the field of clinical research and medicine.

Methods in genetic epidemiology are very powerful in examining and identifying the underlying genetic basis of any phenotype if conducted properly. There are several study designs that can be used with a common goal of finding both the individual effects and interactions within and between genes and environmental exposures that causes the disease. While the technology has provided us better and efficient platforms to conduct the studies, the underlying purpose of genetic epidemiology studies have always remained the same – what genetic variants cause the phenotype or the disease and how can we complement this deficit or control the

Table 11.3 Possible explanations to consider before interpreting the association study results

Outcomes of association studies	Possible explanations to consider
Positive association	True causal association LD with causal variant confounding by population stratification Hardy Weinberg disequilibrium Multiple comparison (false positive)
Negative association	No causal association Small sample size Phenotype misclassification
Multiple genes associated to the same phenotype	Genetic heterogeneity Interactions within and between genes and environmental factor False positive
Multiple alleles at the same gene associated to the same phenotype	Allelic heterogeneity False positive
Same allele in the same gene associated with the same phenotype but in opposite direction	Confounding by population stratification Phenotype heterogeneity False positive

overload of the protein encoded by the variant in the gene to stop the disease? Regardless of the approach, several design and methodological issues need to be considered when conducting studies and interpreting the results (Table 11.3). Although these studies may find association of the phenotype with a genetic variant, the challenge is to meaningfully translate the findings. In most instances the alleles are in the non-coding region and the frequencies are rare but this the stepping stone in the process of understanding the complexity of common diseases. Very rarely can we find a conclusive evidence of genetic effect from a single study, so replication studies with larger samples size should be encouraged to provide insurance against the unknown confounders and biases. To understand the biologic significance of the variants, animal studies and gene expression studies can be conducted as follow-up studies. Of note, most of the loci from the association studies, singly or in aggregate, only explain a small proportion of trait heritability. This “missing heritability” is reflected by small odds ratios and often has limited predictive utility. Overall, clinicians need to be aware of the potential role of genetics in disease etiology and be cautiously familiar with issues and limitations in conducting genetic epidemiology studies before interpreting them for clinical or public health use.

References

1. Morton NE, Chung CS. Genetic epidemiology. New York: Academic; 1978.
2. Mendel G. Versuche über Pflanzen-Hybriden. Verh. Naturforsch. Ver. Brünn 4: 3–47 (in English in 1901, J. R. Hortic. Soc. 26: 1–32)1866.

3. Kehrer-sawatzki H, Cooper DN. Copy number variation and disease. Basel/London: S Krager; 2009.
4. Bell J, Haldane JBS. The linkage between the genes for colour-blindness and haemophilia in man. Proc R Soc (Lond) B. 1937;123:119–50.
5. Hartl DL, Clark AG. Principles of population genetics. Sunderland: Sinauer Associates; 2007.
6. Istrail S, Waterman M, Clark A. Computational methods for SNPs and haplotype inference. New York: Springer; 2004.
7. Khoury MJ, Beaty TH, Cohen BH. Fundamental of genetic epidemiology. New York: Oxford University Press; 1993.
8. Ziegler A, Konig IR. Statistical approach to genetic epidemiology: concepts and applications. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA; 2006.
9. Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet. 1988;43(4):520–6.
10. Khoury MJ, Burke W, Thomson EJ. Genetics and public health in the 20th century. New York: Oxford University Press; 2000.

Additional References and Recommended Readings

1000 Genomes: http://www.1000genomes.org/
dbSNP National Center for Biotechnology Information (NCBI): http://www.ncbi.nlm.nih.gov/SNP/
ENCODE: https://genome.ucsc.edu/encode/
Human Genome Project: http://www.genome.gov/10001772
International HapMap Project: http://hapmap.ncbi.nlm.nih.gov/