CHAPTER 5

Genetics and Disease

JAMES KELLEY, PHD
ROBERT P. KIMBERLY, MD

- Most rheumatic disease are caused by a combination of genes and environment, with genetic variation predisposing or protecting and environmental factors initiating and maintaining a disease state.
- Variations in genes can occur as single nucleotide polymorphisms in coding or noncoding regions and leading to different alleles. Point mutations are rare variations occurring at less than 1% minor allele frequency. Deletion, insertion, repeated sequences of different lengths, and copy number polymorphisms are also responsible for differences in genes.
- Haplotypes, blocks of polymorphisms inherited together more often than expected by chance, can be used to identify disease-causing variants and provide information on recombination, population structure, and evolutionary pressures.
- Association of genes with disease can be performed using linkage studies or association studies. Association studies can determine the odds ratio of a particular gene variant being associated with a particular disease, but require large numbers of samples from affected and unaffected individuals.
- Linkage studies are most useful for monogenic traits with high penetrance where extended family information is available. If a particular gene has only a subtle effect, linkage studies are limited in use.

Relationships between genes and diseases have long been hypothesized. The association of a disease with a gene dates back in Western medicine as far as Hippocrates, who hypothesized epilepsy was caused by a singular hereditary unit of biological material. However, with technological advances and the completion of the human genome sequence (1), scientists can now associate specific genetic variations with clinical conditions. Genetic associations provide informative clues for developing new diagnostic and therapeutic techniques to improve patient care. Understanding the principles that underlie genetic studies will become an essential skill for clinicians if we are to appreciate the complexity of genetic contributions to disease and its treatment (Table 5-1).

THE CAUSE OF DISEASE

In most cases, diseases are not caused by either genes or the environment but by a combination of the two. Genetic variation confers a susceptible or protective effect towards an illness for a specific person when compared with a population. In this paradigm, genetic variations predispose an individual towards a particular outcome while environmental factors, such as infectious agents, chemicals, tobacco smoke, and diet, actually initiate and maintain a disease state in the presence of a set of genetic variants (2). The relative contributions of genetic and environmental factors to disease can be thought of in terms of a sliding scale (Figure 5-1). On one side of the scale, some conditions could be attributed almost entirely to the environment, for example, a car accident. On the other end of the spectrum, there are primarily genetic disorders, such as cystic fibrosis or hemoglobinopathy. Most clinical conditions, though, ranging from heart disease to rheumatoid arthritis (RA) to the common cold, involve some causal component from both an individual's genetic background and environment. For example, genes may predispose an individual to develop type II diabetes mellitus; however, diet and exercise habits ultimately lead to disease. Remembering this paradigm is important for appreciating both the advantages and limits of applying genetic research to the understanding of both health and illness. Neither genetics nor environmental factors should be evaluated alone.

When determining the genetic contribution to the cause of disease, the question arises: does one gene or many lead to illness? Mendelian, or monogenic, diseases, such as Huntington's disease or cystic fibrosis, can be associated with a single genetic variant. Most diseases, however, are complex diseases because they derive their genetic components from a combination of genetic variants, each providing subtle, additive, and personalized effects. The varying genetic components

TABLE 5.1. GLOSSARY OF SELECTED GENETIC TERMS.

TERM	DEFINITION
Admixture	Amount of genetic variation present in an individual due to descending from a particular population
Allele	One of the genetic forms possible at a specific locus, when variation at that locus occurs
Ancestry informative marker	Polymorphism occurring at varying allele frequencies between different populations
Complex (disease or trait)	Involving more than one gene
Haplotype	Group of polymorphisms that are inherited together and are observed together more often than expected by chance
Linkage disequilibrium (LD)	Likelihood that two or more polymorphisms will be inherited together as part of a haplotype
Locus (loci)	Defined position within the genome
Minor allele frequency (MAF)	The frequency of the less common allele in a population
Penetrance	Tendency for a trait to be expressed
Polymorphism	"Many bodies," genetic variation
Population structure	Background genetic variation common within and unique to a group due to a similar and isolated evolutionary history
Recombination	Rearrangement of alleles that is due to the nuclear sequence breaking and recombining during the crossing-over phase of meiosis

of complex diseases explain the different possible clinical manifestations present in one condition, such as the 11 defined possible criteria for patients with systemic lupus erythematosus (SLE), of which only four are required for diagnosis (3).

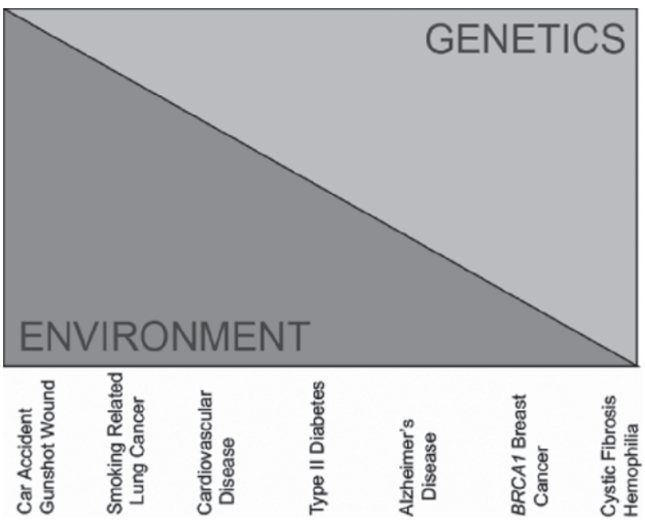

FIGURE 5-1

Relative contributions of genetic and environmental factors to disease. Each disease is caused by varying degrees of genetic and environmental factors.

TYPES OF GENETIC VARIATION

The genetic component of disease can be attributed to genetic variations present (or lacking) in affected individuals of the population. Genetic variation is found by sequencing the genomes of several individuals from a population to detect the most frequent allele and any variants at that specific locus. Single nucleotide polymorphisms (SNPs), the most common type of genetic variant, occur when more than one nucleotide is present at a single position. Nonsynonymous SNPs are those present in coding regions that change the protein's sequence. Synonymous SNPs are those present in exons that do not alter protein sequence. SNPs present in untranslated regions and introns can also affect protein function by altering splicing sites, affecting transcription factor binding, changing promoter sites, and influencing gene expression.

Point mutations are rare variants occurring at a single basepair locus with less than 1% minor allele frequency (MAF). (SNPs occur at greater than 1% MAF.) Point mutations are much more difficult to associate with disease than SNPs without larger sample sizes, which are logistically difficult to collect (2). The frequency at which a variant appears in the population is important because a disease found in a high proportion of the population should be associated with a genetic variant also occurring in a high proportion of the population.

This idea, the common disease–common variant hypothesis (4), directs researchers to which polymorphisms are more likely to influence a disease.

Deletion/insertion polymorphisms (DIPs or "indels") result from the removal or incorporation of nucleotides into the genome sequence. While most DIPs occur outside exons, they are likely to be important in complex traits and diseases, due to their potential for influencing gene expression.

Repeated sequences, or repeat elements, are another form of genetic variation. Interspersed repeat sequences, which account for almost half of the human genome sequence (1), are sections of DNA copied and distributed randomly throughout the genome. Tandemly duplicated elements, such as microsatellites (e.g., CACACACACA), are repeat sequences that, at their time of origin, were copied in a unique pattern and then translocated immediately near their original sequence. However, once present, tandemly duplicated elements are generally inherited through generations in a stable manner. These unique patterns of tandemly duplicated elements provide genetic markers that are both specific and consistent to a population or group of descendants.

Copy number polymorphisms arise when an entire gene or gene segment has been duplicated or when a gene is absent in some individuals. Entire gene duplications allow new genes with new functions to evolve while keeping a functional, backup copy of the original, ancestral gene (5). Examples of copy number polymorphisms are found in natural killer cell receptor gene families and in the major histocompatibility complex (MHC) (6), both regions important in clinical immunology.

RELATIONSHIPS BETWEEN VARIANTS: HAPLOTYPES AND LINKAGE DISEQUILIBRIUM

Haplotypes are "blocks" of polymorphisms that are inherited together more often than expected by chance. These blocks of genetic variants are often separated by regions of high recombination. Haplotypes are useful in identifying disease-causing variants in association studies and can provide information on recombination, population structure, and evolutionary pressures. Because haplotypes define groups of polymorphisms that occur together, experimentally obtaining information about one polymorphism also provides information on the other polymorphisms in the haplotype. Therefore, when attempting to associate genetic variants with disease, testing one or a selection of polymorphisms per haplotype, in a process called *haplotype tagging*, can save time and resources (7).

Assigning polymorphisms to a haplotype, when possible, requires experimental data. About half the genome contains variants that cannot be placed in haplotypes (8). Sequencing multiple samples from a population identifies the combinations and frequencies of polymorphisms possible in that population, which allows researchers to predict which polymorphisms are inherited together and belong in a common haplotype. Haplotypes are defined based on statistical predictions, not absolute certainties. Therefore, polymorphisms assigned to the same haplotype may not be inherited together in all individuals, even though it is likely that they will.

The probability that polymorphisms assigned to the same haplotype will occur together is called linkage disequilibrium (LD). In complete or strong LD, these linked alleles are inherited together within one segment of genomic information. Therefore, any evolutionary pressure or association with disease for one linked allele will inadvertently be observed as present in all polymorphisms of the same haplotype. In weak LD, variants are inherited independently, due to recombination, and a genetic event influencing one allele will not affect the other.

Linkage disequilibrium is most commonly measured with the statistics D' and r^2. The values of these measures range from 0 to 1, with 0 showing weak LD and 1 referring to strong or complete LD. Generally, polymorphisms within a defined haplotype will have a correlation coefficient or r^2 value of at least 0.8. To illustrate how LD is determined and haplotypes are defined from experimental data, an equation for calculating D' between two loci is seen below.

$$D = P_{AB} - (P_A X P_B) \qquad D_{max} = P_{AB} \qquad D' = |D/D_{max}|$$
$$D' = 1 \rightarrow \text{Complete LD}$$

The measure D is equal to the probability of both polymorphisms occurring together in an individual (P_{AB}) minus the product of the probabilities that only one of the polymorphisms will occur in an individual (i.e., P_A is the probability that only polymorphism A will occur). Please note that the probabilities mentioned above are equivalent to the appropriate allele frequencies determined from the sequencing data experimentally obtained. For example, if both polymorphisms occur together in 80% of samples and each polymorphism is observed individually in 10% of samples, then $D = 0.8 - (0.1 \times 0.1) = 0.79$. D' is equal to the absolute value of dividing D by D_{max}. D_{max} equals the probability that both polymorphisms occur in the same sample. Again, in this instance, $D' = | 0.79/0.8 | = 0.9875$, showing that strong LD is present between these two loci. Strong LD in the above situation is intuitive because both variants occur together in 80% of samples.

POPULATION STRUCTURE: CONSIDERING ETHNIC DIFFERENCES

Factors causing genetic variation do so in response to the development, migration, and structure of populations. Due to human history, each population has been exposed to different environments, likely creating different evolutionary pressures to preserve or delete genetic variants in their genomes (9). Cross-population studies on the genomic organization of polymorphisms have shown that Yoruban Africans (Nigeria) have shorter haplotypes and more variants than do Europeans and Asians (8). The wider range of genetic diversity in Africans occurs because humans originated there and only a subgroup of the ancestral human population (and therefore subgroup of genetic variants) migrated to other continents, leaving more genetic diversity in Africa to evolve.

Genetic variations common to a population but not to the whole species is known as *population structure*. These population-specific variants can create phenotypic alterations, including some associated with disease. Ethnic differences in allele frequencies at disease-associated loci and in disease prevalence are commonly reported, such as the increased presence of SLE in patients of African ancestry or an increase in RA in Native Americans when compared with Caucasians. Due to the presence of ethnic-specific genetic contributions, matching cases and controls by ethnicity can help prevent any false genetic associations created by population structure (10).

In addition to self-identification, samples segregated by population can be tested empirically for population admixture. Population admixture is the measurement of the number of discrepancies in allele frequencies between two populations that have been historically isolated. Therefore, admixture quantifies the proportion of an individual's genome that is unique and attributable to an ethnic background (i.e., 20% European ancestry). Population structure varies greatly among groups whose immediate ancestors no longer remained in isolation, such as Latin Americans and African Americans. Admixture can be measured by comparing evolutionarily stable microsatellites, which are exclusive to a particular population (11), or by evaluating ancestry informative markers (AIMs). AIMs are polymorphisms (or the genes containing such polymorphisms) that vary distinctly in allele frequencies between populations.

DETERMINING GENETIC COMPONENTS OF DISEASE

Two main types of studies are used to identify disease-causing genes: linkage studies and association studies. Linkage studies use standardized genetic markers, which do not necessarily produce a phenotypic effect, distributed throughout the genome to detect regions that may contain a variant influencing disease. Such studies rely on LD between these markers and a disease-associated variant. Linkage studies have proven most useful in detecting monogenic diseases, such as Huntington's disease and cystic fibrosis.

While linkage studies provided early insight into the genetic component of complex diseases, association studies are used with increasing frequency to identify variants involved in complex disorders, such as SLE, RA, and other autoimmune conditions. They compare the frequency of a variant in an appropriate number of patients with the disease to the frequency in unrelated, yet matched, controls. Matched controls have similar ages, ethnicities, and backgrounds to affected patients in an effort to reduce errors from population structure. While such studies associate the likelihood of a variant occurring simultaneously with a phenotype, they do not necessarily provide information on any functional difference leading to disease (12). During an association study, genome-wide polymorphism scans can detect specific polymorphisms that are more common in disease groups than in healthy cohorts. Using high capacity technology, this approach evaluates a panel of thousands of common polymorphisms for differences in allele frequencies between affected individuals and controls. A smaller scale technique for association studies is to select and test candidate genes for association. Candidate gene selection suggests a physiological reason for a gene's possible relationship with a particular disease, thereby focusing which genes and variants to study (13).

Statistical measures facilitate the interpretation and design of association studies. Relative risk (RR) measures the likelihood that an individual who possesses the genetic variant will (or will not) develop the associated disease (i.e., the *risk* an individual possessing the variant has of getting the disease *relative* to those without the variant). A RR of 1.5 means an individual with the associated variant is 1.5 times more likely to express the phenotype. RR can be estimated in association studies with a statistic, the odds ratio (OR). The OR can be calculated with the equation: $(A \times B)/(C \times D)$, where A equals the number of samples with the variant and the disease, B equals the number of samples without either the variant or the disease, C equals the number of samples with the variant but not the disease, and D equals the number of samples with the disease but not the variant (see Figure 5-2). Basically, the OR measures the ratio of the presence or absence of both disease and variant against the appearance of either of the two exclusively.

Statistical power, in association studies, refers to the probability of finding a genetic association when it is in fact true. Power is a function of the number of samples tested, the MAF of the variant, the presence of genetic

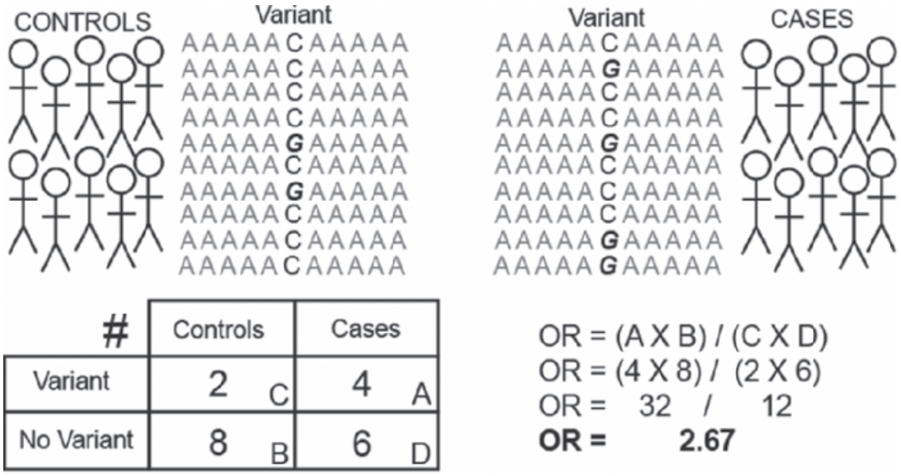

FIGURE 5-2

Associating a genetic variant with disease. Samples taken from a group of affected cases and a group of matched controls are used to determine which allele each individual in the study possesses. Note that not all individuals with disease have the studied variant and not all individuals with the variant have the disease. In this case, the variant is the G allele at position 6. By counting the number of times the variant occurs in each group, it is possible to determine the likelihood that having the variant will correlate with having the disease. This correlation is based on calculating the odds ratio (OR).

features (e.g., dominant/recessive allele), and the OR required to convince scientists the association is meaningful. This OR is arbitrarily set depending on the disease and the level of effect researchers hope to observe. A study that tests too few samples can likely lead to false-positive results, especially when testing a low frequency variant.

There are advantages and weaknesses of both linkage and association studies. Linkage studies are more useful in situations where samples are available from extended families to detect a monogenetic trait with high penetrance. When a variant only contributes a subtle phenotypic effect, linkage studies are limited in use. In this case, association studies should better detect an association; however, the large sample sizes necessary for statistical inference of association with disease can be challenging to obtain. Phenotypes tested in association studies often vary, especially in rheumatic diseases, complicating interpretation of results. For example, in SLE, patients can present a variety of symptoms; therefore, the genetic variants contributing to one individual's disease may differ from the next patient. Therefore, association studies should consider well-defined clinical subgroups during analysis to prevent missing a positive association. Replication, which is finding a positive association in another collection of samples or another population, also can confuse interpretation of association studies' results. During replication studies, positive associations can be lost when larger sample sizes are tested or when testing other populations due to the different evolutionary histories and other genetic variants present in each group. Keeping these issues in mind is important when interpreting genetic studies because many false positives may be in the literature due to a positive publication bias; that is, studies demonstrating an association are more likely to be published than studies that fail to find an association (14).

MAJOR HISTOCOMPATIBILITY COMPLEX

While scientists have found that genetic variants from all over the genome contribute to complex disease, one region of the genome has been associated with more diseases, including rheumatic diseases, than any other: the MHC (6). A selection of genetic associations of rheumatic diseases with MHC-encoded genes is listed in Table 5-2.

The MHC is a dense cluster of over 260 genes on chromosome 6p21.3 containing a high percentage of immune-related genes, in particular the highly polymorphic human leukocyte antigen (HLA) genes involved in antigen presentation. The MHC is genomically organized into multiple regions. From the telomere to centromere, they are: extended class I region, class I region (*HLA-A*, *HLA-B*, *HLA-C,* etc.), class III region (*C4*, *TNF*, *LTA,* etc.), class II region (*HLA-DR*, *HLA-DQ*, *HLA-DP,* etc.), and the extended class II region. The class IV, or inflammatory region, is located within the class III region and contains a concentration of genes encoding inflammatory-mediating molecules (see

TABLE 5.2. SOME RHEUMATIC DISEASE ASSOCIATIONS WITH THE MAJOR HISTOCOMPATIBILITY COMPLEX.

GENE	DISEASE
HLA	Systemic sclerosis
HLA-B	Ankylosing spondylitis
HLA-B	Behcet's disease
HLA-B	Sarcoidosis
MICA	Behcet's disease
MICA	Rheumatoid arthritis
TNF	Ankylosing spondylitis
TNF	Rheumatoid arthritis
NFKBIL1	Rheumatoid arthritis
BTNL2	Sarcoidosis
HLA class II	Juvenile ankylosing spondylitis
HLA class II	Systemic lupus erythematosus[a]
HLA class II	Systemic sclerosis
HLA-DRB1	Rheumatoid arthritis
HLA-DRB1	Sarcoidosis
HLA-DRB1	Sjögren's syndrome
HLA-DRB1	Takayasu arteritis
TAP2	Rheumatoid arthritis
TAP2	Sjögren's syndrome
TAP2	Systemic lupus erythematosus

[a] Specific associations with HLA class II genes (*HLA-DQ* and *HLA-DR*) may vary with ethnicity.
Examples were taken from the Online Mendelian Inheritance in Man database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim) and the Genetic Association Database (http://geneticassociationdb.nih.gov). Note that MHC encoded genes other than *HLA* have been associated with diseases.

Chapter 6B). The MHC is likely involved with so many diseases because it contains the highest density of genetic variants, areas of strong LD, and highest density of genes in the human genome (6).

Genetics will have an increased role in medicine over the coming years as genes' relationships with disease become more understood and as new genetic-based technologies are translated into the clinic. Understanding the genetic contribution to a clinical condition, both in the MHC and throughout the genome, will allow physicians and researchers the ability to find new markers for detecting and preventing an illness, to develop new diagnostic measures for evaluating potential success of drug therapies, and to predict biological malfunctions underlying a disease.

References

1. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931–945.
2. Hunter DJ. Gene-environment interactions in human diseases. Nat Rev Genet 2005;6:287–298.
3. Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. Arthritis Rheum 1997;40:1725.
4. Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet 2001;17:502–510.
5. Ohno S. Evolution by gene duplication. Berlin: Springer; 1970.
6. Kelley J, Trowsdale J. Features of MHC and NK gene clusters. Transpl Immunol 2005;14:129–134.
7. Johnson GC, Esposito L, Barratt BJ, et al. Haplotype tagging for the identification of common disease genes. Nat Genet 2001;29:233–237.
8. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. Science 2002;296:2225–2229.
9. Bamshad M, Wooding SP. Signatures of natural selection in the human genome. Nat Rev Genet 2003;4:99–111.
10. Clayton DG, Walker NM, Smyth DJ, et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 2005;37:1243–1246.
11. Patterson N, Hattangadi N, Lane B, et al. Methods for high-density admixture mapping of disease genes. Am J Hum Genet 2004;74:979–1000.
12. Daly AK, Day CP. Candidate gene case-control association studies: advantages and potential pitfalls. Br J Clin Pharmacol 2001;52:489–499.
13. Risch NJ. Searching for genetic determinants in the new millennium. Nature 2000;405:847–856.
14. Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG. Genetic associations in large versus small studies: an empirical assessment. Lancet 2003;361:567–571.