Although there are many possible causes of human disease, family history is often one of the strongest risk factors for common disease complexes such as cancer, cardiovascular disease (CVD), diabetes, autoimmune disorders, and psychiatric illnesses. A person inherits a complete set of genes from each parent, as well as a vast array of cultural and socioeconomic experiences from his/her family. Family history is thought to be a good predictor of an individual’s disease risk because family members most closely represent the unique genomic and environmental interactions that an individual experiences (Kardia et al., 2003). Inherited genetic variation within families clearly contributes both directly and indirectly to the pathogenesis of disease. This chapter focuses on what is known or theorized about the direct link between genes and health and what still must be explored in order to understand the environmental interactions and relative roles among genes that contribute to health and illness.

GENETIC SUSCEPTIBILITY

For more than 100 years, human geneticists have been studying how variations in genes contribute to variations in disease risk. These studies have taken two approaches. The first approach focuses on identifying the individual genes with variations that give rise to simple Mendelian patterns of disease inheritance (e.g., autosomal dominant, autosomal recessive, and X-linked) (see Table 3-1; Mendelian Inheritance in Man). The second approach seeks to understand the genetic susceptibility to disease as the con-
Each of these approaches will be discussed below.

In general, diseases with simple Mendelian patterns of inheritance tend to be relatively uncommon or frequently rare, with early ages of onset, such as phenylketonuria, sickle cell anemia, Tay-Sachs disease, and cystic fibrosis. In addition, some of these genes have been associated with extreme forms of common diseases, such as familial hypercholesterolemia, which is caused by mutations in the low-density lipoprotein (LDL) receptor that predispose individuals to early onset of heart disease (Brown and Goldstein, 1981).

Another example of Mendelian inheritance is familial forms of breast cancer associated with mutations in the BRCA1 and BRCA2 genes that predispose women to early onset breast cancer and often ovarian cancer. The genes identified have mutations that often are highly penetrant—that is, the probability of developing the disease in someone carrying the disease susceptibility genotype is relatively high (greater than 50 percent). These genetic diseases often exhibit a genetic phenomenon known as allelic heterogeneity, in which multiple mutations within the same gene (i.e., alleles) are found to be associated with the same disease. This allelic heterogeneity

### TABLE 3-1 Online Mendelian Inheritance in Man (OMIM) Statistics (as of May 15, 2006), Number of Entries

<table>
<thead>
<tr>
<th></th>
<th>Autosomal</th>
<th>X-Linked</th>
<th>Y-Linked</th>
<th>Mitochondrial</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene with known sequence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10,772</td>
</tr>
<tr>
<td>Gene with known sequence and phenotype</td>
<td>10,215</td>
<td>472</td>
<td>48</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>Phenotype description molecular basis known</td>
<td>349</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>380</td>
</tr>
<tr>
<td>Mendelian phenotype or locus, molecular basis unknown</td>
<td>1,710</td>
<td>153</td>
<td>2</td>
<td>26</td>
<td>1,891</td>
</tr>
<tr>
<td>Other, mainly phenotypes with suspected Mendelian basis</td>
<td>1,384</td>
<td>134</td>
<td>4</td>
<td>0</td>
<td>1,522</td>
</tr>
<tr>
<td>Total</td>
<td>15,723</td>
<td>9,353</td>
<td>56</td>
<td>63</td>
<td>16,777</td>
</tr>
</tbody>
</table>

often is population specific and can represent the unique demographic and mutational history of the population.

In some cases, genetic diseases also are associated with locus heterogeneity, meaning that a deleterious mutation in any one of several genes can give rise to an increased risk of the disease. This is a finding common to many human diseases including Alzheimer’s disease and polycystic kidney disease. Both allelic heterogeneity and locus heterogeneity are sources of variation in these disease phenotypes since they can have varying effects on the disease initiation, progression, and clinical severity.

Environmental factors also vary across individuals and the combined effect of environmental and genetic heterogeneity is etiologic heterogeneity. Etiologic heterogeneity refers to a phenomenon that occurs in the general population when multiple groups of disease cases, such as breast cancer clusters, exhibit similar clinical features, but are in fact the result of differing events or exposures. Insight into the etiology of specific diseases as well as identification of possible causative agents is facilitated by discovery and examination of disease cases demonstrating etiologic heterogeneity. The results of these studies may also highlight possible gene-gene interactions and gene-environment interactions important in the disease process. Identifying etiologic heterogeneity can be an important step toward analysis of diseases using molecular epidemiology techniques and may eventually lead to improved disease prevention strategies (Rebbeck et al., 1997).

As opposed to the Mendelian approach, the second approach to studying how variations in genes contribute to variations in disease risk focuses on understanding the genetic susceptibility to diseases as the consequence of the joint effects of many genes, each with small to moderate effects (i.e., polygenic models of disease) and often interacting among themselves and with the environment to give rise to the distribution of disease risk seen in a population (i.e., multifactorial models of disease). This approach has been used primarily for understanding the genetics of birth defects and common diseases and their risk factors. As described below, several steps are involved in developing such an understanding.

As a first step, study participants are asked to provide a detailed family history to assess the presence of familial aggregation. If individuals with the disease in question have more relatives affected by the disease than individuals without the disease, familial aggregation is identified. While familial aggregation may be accounted for through genetic etiology, it may also represent an exposure (e.g., pesticides, contaminated drinking water, or diet) common to all family members due to the likelihood of shared environment.

When there is evidence of familial aggregation, the second step is to focus research studies on estimating the heritability of the disease and/or its risk factors. Heritability is defined as the proportion of variation in disease
risk in a population that is attributable to unmeasured genetic variations inferred through familial patterns of disease. It is a broad population-based measure of genetic influence that is used to determine whether further genetic studies are warranted, since it allows investigators to test the overarching null hypothesis that no genes are involved in determining disease risk. Twin studies and family studies are frequently used in the study of heritability.

Twin studies comparing the disease and risk factor variability of monozygotic and dizygotic twins have been a common study design used to easily estimate both genetic and cultural inheritance. Studies of monozygotic twins reared together versus those reared apart also have been important in estimating both genetic and environmental contributions to patterns of inheritance. The modeling of the sources of phenotypic variation using family studies has become quite sophisticated, allowing the inclusion of model parameters to represent the additive genetic component (i.e., polygenes), the nonadditive genetic component (i.e., genetic dominance, as well as gene-environment and gene-gene interactions), shared family environment, and individual environments. The contributions of these factors have been shown to vary by age and population.

When significant evidence of genetic involvement is established, the next step is to identify the responsible genes and the mutations that are associated with increased or decreased risk, using either genetic linkage analysis or genetic association studies. For example, in the study of birth defects, this often involves the search for chromosomal deletions, insertions, duplications, or translocations.

**GENETIC LINKAGE ANALYSIS AND GENETIC ASSOCIATION STUDIES**

The human genome is made up of tens of thousands of genes. With approximately 30,000 genes to choose from, assigning a specific gene or group of genes to a corresponding human disease demands a methodical approach consisting of many steps. Traditionally, the process of gene discovery begins with a linkage analysis that assesses disease within families. Linkage analyses are typically followed by genetic association studies that assess disease across families or across unrelated individuals.

**Genetic Linkage Analysis**

The term *linkage* refers to the tendency of genes proximally located on the same chromosome to be inherited together. Linkage analysis is one step in the search for a disease susceptibility gene. The goal of this analysis is to approximate the location of the disease gene in relation to a known genetic
marker, applying an understanding of the patterns of linkage. Traditional linkage analysis that traces patterns of heredity of both the disease phenotype and genetic markers in large, high-risk families have been used to locate disease-causing gene mutations such as the breast cancer gene (BRCA1) on chromosome 17 (Hall et al., 1990).

Because the mode of inheritance is often not clear for common diseases, an alternative approach to classic linkage analysis was developed to capitalize on the basic genetic principle that siblings share half of their alleles on average. By investigating the degree of allelic sharing across their genomes, pairs of affected siblings (i.e., two or more siblings with the same disease) can be used to identify chromosomal regions that may contain genes whose variations are related to the disease being studied. If numerous sibling pairs affected by the disease of interest exhibit a greater than expected sharing of the known alleles of the polymorphic genetic marker being used, then the genetic marker is likely to be linked (that is, within close proximity along the chromosome) to the susceptibility gene responsible for the disease being studied. To find chromosomal regions that show evidence for linkage using this affected sibling pair method typically requires typing numerous affected sibships with hundreds of highly polymorphic markers uniformly positioned along the human genome (Mathew, 2001).

This approach has been widely used to identify regions of the genome thought to contribute to common chronic diseases. However, results of linkage analyses have not been consistently replicated. The inability to successfully replicate linkage findings may be a result of insufficient statistical power (that is, including an inadequate number of sibling pairs with the disease of interest) or results that included false positives in the original study. An alternate explanation could be that different populations are affected by different susceptibility genes than those that were studied originally (Mathew, 2001). Without consistent replication of results it is premature to draw conclusions about the contribution of a gene locus to a specific disease.

Upon the confirmation of a linkage, researchers can begin to search the region for the candidate susceptibility gene. The search for a single susceptibility gene for common diseases often involves examination of very large linkage regions, containing 20 to 30 million base pairs and potentially hundreds of genes (Mathew, 2001). It is also important to note, however, that while linkage mapping is a powerful tool for finding Mendelian disease genes, it often produces weak and sometimes inconsistent signals in studies of complex diseases that may be multifactorial. Linkage studies perform best when there is a single susceptibility allele at any given disease locus and generally performs poorly when there is substantial genetic heterogeneity.
Genetic Association Studies

Technological advances in high-throughput genotyping have allowed the direct examination of specific genetic differences among sizable numbers of people. Genetic association techniques are often the most efficient approach for assessing how specific genetic variation can affect disease risk. Genetic association studies, which have been used for decades, have perpetually progressed in terms of the development of new study designs (such as case-only and family-based association designs), new genotyping systems (such as array-based genotyping and multiplexing assays), and new methods used for addressing biases such as population (Haines and Pericak-Vance, 1998).

Analysis of the effects of genetic variation typically involves first the discovery of single nucleotide polymorphisms (SNPs) and then the analysis of these variations in samples from populations. SNPs occur on average approximately every 500 to 2,000 bases in the human genome. The most common approach to SNP discovery is to sequence the gene of interest in a representative sample of individuals. Currently, sequencing of entire genes on small numbers of individuals (~25 to 50) can detect polymorphisms occurring in 1 to 3 percent of the population with approximately 95 percent confidence. The Human DNA Polymorphism Discovery Program of the National Institute of Environmental Health Sciences’ Environmental Genome Project is one example of the application of automated DNA sequencing technologies to identify SNPs in human genes that may be associated with disease susceptibility and response to environment (Livingston et al., 2004). The National Heart, Lung, and Blood Institute’s Programs in Genomic Applications also has led to important increases in our knowledge about the distribution of SNPs in key genes thought to be already biologically implicated in disease risk (i.e., biological candidate genes).

Impressive and rapid advances in SNP analysis technology are rapidly redefining the scope of SNP discovery, mapping, and genotyping. New array-based genotyping technology enables “whole genome association” analyses of SNPs between individuals or between strains of laboratory animal species (Syvanen, 2005). Arrays used for these analyses can represent hundreds of thousands of SNPs mapped across a genome (Klein et al.,

---

1 An SNP is the DNA sequence variation that occurs when a single nucleotide (A, T, C, or G) in the genome sequence is altered (Smith, 2005).

2 A candidate gene is a gene whose protein product is involved in the metabolic or physiological pathways associated with a particular disease (IOM, 2005).
This approach allows rapid identification of SNPs associated with disease and susceptibility to environmental factors. The strength of this technology is the massive amount of easily measurable genetic variation it puts in the hands of researchers in a cost-effective manner ($500 to $1,000 per chip). The criteria for the selection of SNPs to be included on these arrays are a critical consideration, since they affect the inferences that can be drawn from using these platforms. Of course, the ultimate tool for SNP discovery and genotyping is individual whole genome sequencing. Although not currently feasible, the rapid advancement of technology now being stimulated by the National Human Genome Research Institute’s “$1,000 genome” project likely will make this approach the optimal one for SNP discovery and genotyping in the future.

With the ability to examine large quantities of genetic variations, researchers are moving from investigations of single genes, one at a time, to consideration of entire pathways or physiological systems that include information from genomic, transcriptomic, proteomic, and metabonomic levels that are all subject to different environmental factors (Haines and Pericak-Vance, 1998). However, these genome- and pathway-driven study designs and analytic techniques are still in the early stages of development and will require the joint efforts of multiple disciplines, ranging from molecular biologists to clinicians to social scientists to bioinformaticians, in order to make the most effective use of these vast amounts of data.
group 2) gene were genotyped in 1,091 lung cancer cases and 1,240 controls. When the cases and controls were stratified into heavy smokers versus nonsmokers, Zhou et al. (2003) found that nonsmokers with the mutant XRCCI genotype had a 2.4 times greater risk of lung cancer than nonsmokers with the normal genotype. In contrast, heavy smokers with the mutant XRCCI genotype had a 50 percent reduction in lung cancer risk compared to their counterparts with the more frequent normal genotype. When the three mutations from these two genes were examined together in the extreme genotype combination (individual with five or six mutations present in his/her genotype) there was a 5.2 time greater risk of lung cancer in nonsmokers and a 70 percent reduction of risk in the heavy smokers compared to individuals with no mutations. The protective effect of these genetic variations in heavy smokers may be caused by the differential increase in the activity of these protective genes stimulated by heavy smoking. Similar types of gene-smoking interactions also have been found for other genes in this pathway, such as ERCC1. These studies illustrate the importance of identifying the genetic variations that are associated with the differential risk of disease related to human behaviors. Note that this type of research also raises many different kinds of ethical and social issues, since it identifies susceptible subgroups and protected subgroups of subjects by both genetic and human behavior strata (see Chapter 10).

The study by Zhou et al. (2003) also demonstrates the increased information provided by jointly examining the effects of multiple mutations on toxicity-related disease. Other studies of mutations in genes involved in the Phase II metabolism (GSTM1, GSTT1, GSTP1) also have demonstrated the importance of investigating the joint effects of mutations (Miller et al., 2002) on cancer risk. Although these two studies focused on the additive effects of multiple genes, gene-gene interactions are another important component to develop a better understanding of human susceptibility to disease and to interactions with the environment.

To adequately understand the continuum of genomic susceptibility to environmental agents that influences the public's health, more studies of the joint effects of multiple mutations need to be conducted. Advances in bioinformatics can play a key role in this endeavor. For example, methods to screen SNP databases for mutations in transcriptional regulatory regions can be used for both discovery and functional validation of polymorphic regulatory elements, such as the antioxidant regulatory element found in the promoter regions of many genes encoding antioxidative and Phase II detoxification enzymes (Wang et al., 2005). Comparative sequence analysis methods also are becoming increasingly valuable to human genetic studies, because they provide a means to rank order SNPs in terms of their potential deleterious effects on protein function or gene regulation (Wang et al., 2004). Methods of performing large-scale analysis of nonsynonymous SNPs...
MECHANISMS OF GENE EXPRESSION

Identifying genes whose variations are associated with disease is just the first step in linking genetics and health. Understanding the mechanisms by which the gene is expressed and how it is influenced by other genes, proteins, and the environment is becoming increasingly important to the development of preventive, diagnostic, and therapeutic strategies.

When genes are expressed, the chromosomal DNA must be transcribed into RNA and the RNA is then processed and transported to be translated into protein. Regulating the expression of genes is a vital process in the cell and involves the organization of the chromosomal DNA into an appropriate higher-order chromatin structure. It also involves the action of a host of specific protein factors (to either encourage or suppress gene expression), which can act at different steps in the gene expression pathway.

In all organisms, networks of biochemical reactions and feedback signals organize developmental pathways, cellular metabolism, and progression through the cell cycle. Overall coordination of the cell cycle and cellular metabolism results from feed-forward and feedback controls arising from sets of dependent pathways in which the initiation of events is dependent on earlier events. Within these networks, gene expression is controlled by molecular signals that regulate when, where, and how often a given gene is transcribed. These signals often are stimulated by environmental influences or by signals from other cells that affect the gene expression of many genes through a single regulatory pathway. Since a regulatory gene can act in combination with other signals to control many other genes, complex branching networks of interactions are possible (McAdams and Arkin, 1997).

Gene regulation is critical because by switching genes on or off when needed, cells can be responsive to changes in environment (e.g., changes in diet or activity) and can prevent resources from being wasted. Variation in the DNA sequences associated with the regulation of a gene's expression are therefore likely candidates for understanding gene-environment interactions at the molecular level, since these variations will affect whether an environmental signal transduced to the nucleus will successfully bind to the promoter sequence in the gene and stimulate or repress gene expression. Combining genomic technologies for SNP genotyping with high-density gene expression arrays in human studies has only recently elucidated the
extent to which this type of molecular gene-environment interaction may be occurring.

Cells also regulate gene expression by post-transcriptional modification; by allowing only a subset of the mRNAs to go on to translation; or by restricting translation of specific mRNAs to only when and where the product is needed. The genetic factors that influence post-transcriptional control are much more difficult to study because they often involve multiprotein complexes not easily retrieved or assayed from cells. At other levels, cells regulate gene expression through epigenetic mechanisms, including DNA folding, histone acetylation, and methylation (i.e., chemical modification) of the nucleotide bases. These mechanisms are likely to be influenced by genetic variations in the target genes as well as variations manifested in translated cellular regulatory proteins. Gene regulation occurs throughout life at all levels of organismal development and aging.

A classic example of developmental control of gene expression is the differential expression of embryonic, fetal, and adult hemoglobin genes (see Box 3-1). The regulation of the epsilon, delta, gamma, alpha, and beta genes occurs through DNA methylation that is tightly controlled through developmental signals. During development a large number of genes are turned on and off through epigenetic regulation. One of the fastest growing fields in genetics is the study of the developmental consequences of environmental exposures on gene expression patterns and the impact of genetic variations on these developmental trajectories.

---

3The sickle cell example is abstracted from a commissioned paper prepared by Robert J. Thompson, Jr., Ph.D. (Appendix D).
normal β-globin gene and one copy of the sickle variant (Hb AS) (Ashley-Koch et al., 2000). Four major β-globin gene haplotypes have been identified. Three are named for the regions in Africa where the mutations first appeared: BEN (Benin), SEN (Senegal), and CAR (Central African Republic). The fourth haplotype, Arabic-India, occurs in India and the Arabic peninsula (Quinn and Miller, 2004).

Disease severity is associated with several genetic factors (Ashley-Koch et al., 2000). The highest degree of severity is associated with Hb SS, followed by Hb s/β0-thalassemia, and Hb SC. Hb S/β+ -thalassemia is associated with a more benign course of the disease (Ashley-Koch et al., 2000).

Disease severity also is related to β-globin haplotypes, probably due to

---

**BOX 3-1**

**Gene Expression and Globin**

The production of hemoglobin is regulated by a number of transcriptional controls, such as switching, that dictate the expression of a different set of globin genes in different parts of the body throughout the various stages of the development process. This transcriptional regulation of globin genes is a result of many different DNA sequences and methylation of those sequences. The process begins shortly after conception when the yolk of the egg sac expresses genes that are responsible for the embryonic hemoglobin are deactivated, while the genes responsible for producing fetal hemoglobin in the liver are activated. Upon birth, the adult globin genes are activated and the bone marrow stem cells begin to produce adult hemoglobin and red blood cells (Rimoin et al., 2002).

A group of diseases that are the result of defective switching among the globin genes during the development process are called thalassemias. This class of diseases results in the decreased capacity to carry oxygen due to the complete absence of hemoglobin or the production of abnormal hemoglobin. Two types of thalassemias, alpha and beta, are the product of ineffective gene regulation. The percentage of total globin synthesis is shown in the graph below.

[Graph of globin synthesis percentages]

---

The graph above shows the percentage of total globin synthesis at different stages of development. The x-axis represents postnatal age (weeks), and the y-axis shows the percentage of total globin synthesis. The graph highlights the switch from embryonic to fetal hemoglobin and then to adult hemoglobin.

---

**References**

Ashley-Koch et al., 2000; Quinn and Miller, 2004; Rimoin et al., 2002.
variations in hemoglobin level and fetal hemoglobin concentrations. The Senegal haplotype is the most benign form, followed by the Benin, and the Central African Republic haplotype is the most severe form (Ashley-Koch et al., 2000).

Thus, although sickle cell disease is a monogenetic disorder, its phenotypic expression is multigenic (see Appendix D). There are two cardinal pathophysiologic features of sickle cell disease—chronic hemolytic anemia and vasoocclusion. Two primary consequences of hypoxia secondary to vasoocclusive crisis are pain and damage to organ systems. The organs at greatest risk are those in which blood flow is slow, such as the spleen and bone marrow, or those that have a limited terminal arterial blood supply, including the eye, the head of the femur and the humerus, and the lung as

[source: Figures adapted from Thompson et al., 1991; Bridges, 2002; and Rimoin et al., 2002.]

globin genes activated during fetal stages of development are often not completely deactivated following the birth of individuals affected by thalassemia. Although it is not nearly as effective as the hemoglobin produced by the bone marrow, the remaining globin activation in the liver cells offers an additional source of necessary oxygen suppliers to the cells (Rimoin et al., 2002).
ASPECTS OF HEALTH INFLUENCED BY GENETICS

Because every cell in the body, with rare exception, carries an entire genome full of variation as the template for the development of its protein machinery, it can be argued that genetic variation impacts all cellular, biochemical, physiological, and morphological aspects of a human being. How that genetic variation is associated with particular disease risk is the focus of much current research. For common diseases such as CVD, hypertension, cancer, diabetes, and many mental illnesses, there is a growing appreciation that different genes and different genetic variations can be involved in different aspects of their natural history. For example, there are likely to be genes whose variations are associated with a predisposition toward the initiation of disease and other genes or gene variations that are involved in the progression of a disease to a clinically defined endpoint. Furthermore, an entirely different set of genes may be involved in how an individual responds to pharmaceutical treatments for that disease. There also are likely to be genes whose variability controls how much or how little a person is likely to be responsive to the environmental risk factors that are associated with disease risk. Finally, there are thought to be genes that affect a person’s overall longevity that may counteract or interact with genes that may otherwise predispose that person to a particular disease outcome and thus may have an additional impact on survivorship.

In many ways, we are only at the beginning of developing a true understanding of how genomic variations give rise to disease susceptibility. Indeed many would argue that, without incorporating the equally important role of the environment, we will never fully understand the role of genetics in health. As progress is made through utilizing the new technologies for measuring biological variation in the genome, transcriptome, proteome, and metabonome, we are likely to have to make large shifts in our conceptual frameworks about the roles of genes in disease. Global patterns of genomic susceptibility are likely to emerge only when we consider the influence of the many interacting components working simultaneously that are dependent on
contexts such as age, sex, diet, and physical activity that modify the relationship with risk. For the most part, we are still at the stage of documenting the complexity, finding examples and types of genetic susceptibility genes, understanding disease heterogeneity, and postulating ways to develop models of risk that use the totality of what we know about human biology, from our genomes to our ecologies to model risk.
Pharmacogenetics

It has been well established that individuals often respond differently to the same drug therapy. The drug disposition process is a complex set of physiological reactions that begin immediately upon administration. The drug is absorbed and distributed to the targeted areas of the body where it interacts with cellular components, such as receptors and enzymes, that further metabolize the drug, and ultimately the drug is excreted from the body (Weinshilboum, 2003). At any point during this process, genetic variation may alter the therapeutic response of an individual and cause an adverse drug reaction (ADR) (Evans and McLeod, 2003). It has been estimated that 20 to 95 percent of variations in drug disposition, such as ADRs, can be attributed to genetic variation (Kalow et al., 1998; Evans and McLeod, 2003).

Sensitivity to both dose-dependent and dose-independent ADRs can have roots in genetic variation. Polymorphisms in kinetic and dynamic factors, such as cytochrome P450 and specific drug targets can cause these individuals susceptibilities to ADRs. While the characteristics of the ADR dictate the true significance of these factors, in most cases, multiple genes are involved (Pirmohamed and Park, 2001). Future analyses using genome-wide SNP profiling could provide a technique for assessing several genetic susceptibility factors for ADRs and ascertaining their joint effects. One of the challenges to the study of the relationship between genetic variation and ADRs is an inadequate number of patient samples. To remedy this problem, Pirmohamed and Park (2001) have proposed that prospective randomized controlled clinical trials become a part of standardized practice to ultimately prove the clinical utility of genotyping all patients as a measure to prevent ADRs.

Here we review some of the current work in pharmacogenetics as an example of what might be expected to arise from rigorous study of the interaction between social, behavioral, and genetic factors. Researchers have provided a few well-established examples of differences in individual drug response that have been ascribed to genetic variations in a variety of cellular drug disposition machinery, such as drug transporters or enzymes responsible for drug metabolism (Evans and McLeod, 2003). For example:
• With the knowledge that the HER2 gene is overexpressed in approximately one fourth of breast cancer cases, researchers developed a humanized monoclonal antibody against the HER2 receptor in hopes of inhibiting the tumor growth associated with the receptor. Genotyping advanced breast cancer patients to identify those with tumors that overexpress the HER2 receptor has produced promising results in improving the clinical outcomes for these breast cancer patients (Cobleigh et al., 1999).

• A therapeutic class of drugs called thiopurines is used as part of the treatment regimen for childhood acute lymphoblastic leukemia. One in 300 Caucasians has a genetic variation that results in low or nonexistent levels of thiopurine methyltransferase (TPMT), an enzyme that is responsible for the metabolism of the thiopurine drugs. If patients with this genetic variation are given thiopurines, the drug accumulates to toxic levels in their body causing life-threatening myelosuppression. Assessing the TPMT phenotype and genotype of the patient can be used to determine the individualized dosage of the drug (Armstrong et al., 2004).

• The family of liver enzymes called cytochrome P450s plays a major role in the metabolism of as many as 40 different types of drugs. Genetic variants in these enzymes may diminish their ability to effectively break down certain drugs, thus creating the potential for overdose in patients with less active or inactive forms of the cytochrome P450 enzyme. Varying levels of reduced cytochrome P450 activity is also a concern for patients taking multiple drugs that may interact if they are not properly metabolized by well-functioning enzymes. Strategies to evaluate the activity level of cytochrome P450 enzymes have been devised and are valuable in planning and monitoring successful drug therapy. Some pharmaceutical drug trials are now incorporating early tests that evaluate the ability of differing forms of cytochrome P450 to metabolize the new drug compound (Obach et al., 2006).

Some pharmacogenetics research has focused on the treatment of psychiatric disorders. With the introduction of a class of drugs known as selective serotonin re-uptake inhibitors (SSRIs), pharmacological treatment of many psychiatric disorders changed drastically. SSRIs offer significant improvements over the previous generation of treatments, including improved efficacy and tolerance for many patients. However, not all patients respond positively to SSRI treatment and many experience ADRs. New pharmacogenetic studies have indicated that these ADRs may be the result of genetic variations in serotonin transporter genes and cytochrome P450 genes. Further study and replication of these findings are necessary. If the characterization of the genetic variations is completed and is fully understood it would be possible to screen and monitor patients using genotyping...
techniques to create individualized drug therapies similar to those discussed above (Mancama and Kerwin, 2003).

A significant challenge to the development of individualized drug therapies is the often polygenic or multifactorial inherited component of drug responses. Isolating the polygenic determinants of the drug responses is a sizable task. A good understanding of the drug’s mechanism of action and metabolic and disposition pathways should be the basis of all investigations. This knowledge can aid in directing genome-wide searches for gene variations associated with drug effects and subsequent candidate-gene approaches of investigation. Additionally, proteomic and gene-expression profiling studies are also important ways to substantiate and understand the pathways by which the gene of interest operates to affect the individual’s response to the drug (Evans and McLeod, 2003). It is not enough to show an association; characterization of the underlying biological mechanisms is an essential component of moving genetic findings into the area of risk reduction. Another key component of utilizing genetics to improve prevention and reduce disease is an understanding of the distribution of the genetic variations in the populations being served.

GENETICS OF POPULATIONS AS RELATED TO HEALTH AND DISEASE

Human populations differ in their distribution of genetic variations. This is a consequence of their historical patterns of mutation, migration, reproduction, mating, selection, and genetic drift. Inherited mutations typically occur during gametogenesis within a single individual and then can be passed on to offspring for many generations. Whether that mutation goes on to become a prevalent polymorphism (i.e., a mutation with a population frequency of greater than 1 percent) is determined by both evolutionary forces and chance events. For example, it depends on whether the original child who inherited the mutation survives to adulthood and reproduces and whether that child’s children survive to reproduce, and so on. The number of children in a family also influences the prevalence of the mutation, and this is often tied to environmental factors that impact fertility and mating patterns that influence the speed with which a private mutation becomes a public polymorphism. There are well-known examples of what are called founder mutations in which this trajectory can be documented. For example, one particular district in what is Quebec (Canada) today was originally founded by only a few families from a particular French province. One of the founding fathers carried a 10kb deletion in his LDL receptor (LDL-R) gene that was passed down through the generations quickly and today is carried by 1 in 154 French Canadians in northeastern Quebec. This mutation is associated with familial hypercholesterolemia, and French Ca-
nadians have one of the highest prevalences of this disease in the world because of the small founding populations followed by population expansion (Moorjani et al., 1989).

There are also a number of examples where mutations that arise in an individual become more prevalent because of the selective advantage they impart on their carriers. The best known example is the mutation associated with sickle cell anemia. The geographical pattern of this mutation strongly mirrors the geographical pattern of malarial infection. It has been molecularly demonstrated that individuals carrying the sickle cell mutation have a resistance to malarial infection. Because many of the selection pressures that may have given rise to the current distribution of mutations in particular populations are in our evolutionary past, it is difficult to assess how much variation within or among populations is due to these types of selection forces.

Another major force in determining the distribution of genetic variations within and among human populations is their migration and reproductive isolation. According to our best knowledge, one of the most important periods in human evolution occurred approximately 100,000 years ago, when some humans migrated to other continents from the African basin and established new communities with relative reproductive isolation. Genetic differences among people in different geographical areas have been associated with the concept of race for hundreds of years. Although race is still used as a label, the original concept of race as genetically distinct subspecies of humans has been rejected through modern genetic information. For numerous reasons, discussed in the section below, it is more appropriate to reconceptualize the old genetics of race into a more accurate genetics of ancestry.

In addition to distant evolutionary patterns of migration, more modern migration patterns also have had a profound effect on the genetics of populations. For example, the current population of the United States and much of North America is very diverse genetically as a consequence of the mixing of many people from many different countries and continents.

A central reason for studying the origins and nature of human genetic variation is that the similarities and differences in the type and frequencies of genetic variations within and among populations can have a profound impact on studies that attempt to understand the influence of genes on disease risk. For example, some genetic variations, such as the apolipoprotein E protein polymorphisms, are found in every population and have very similar genotype frequencies around the world (Wu et al., 2002; Deniz Naranjo et al., 2004). The variation’s association with increased heart disease and Alzheimer’s disease could be and has been tested in many of the world’s populations. Other mutations such as the 10kb
deletion in the LDL-R gene described above are more population-specific variations.

Furthermore, from a statistical point of view, the effect of a genetic variation on the continuum of risk found in any population is correlated with its frequency. For example, common genetic polymorphisms with frequencies near 50 percent cannot be associated with large phenotypic effects within a population because the genotype classes each represent a large fraction of the population and, since most risk is normally distributed, the average risk for a highly prevalent genotype class cannot deviate from the overall risk of the population to any large degree. This correlation between genotype frequency and effect does not mean that common variations cannot be significant in their effects. The statistical significance of an association between a genetic variant and a disease is a joint function of sample size and the size of the effect. In addition, genetic research among populations that differ in their genotype frequencies can differ in their inferences about which polymorphisms have significant effects even if the absolute phenotypic effect is the same. See Cheverud and Routman (1995) for a more formal statistical explanation of this phenomenon and its impact on assessing gene-gene interactions.

Another key consideration in understanding the relationship between genetic variations and measures of disease risk is the population differences in the correlations between genotype frequencies at different SNP locations. There are two common reasons why the frequency of an allele or genotype at a particular SNP could be correlated with the frequency of an allele or genotype for a different SNP. First, a phenomenon known as linkage disequilibrium creates correlations among SNPs as a consequence of the mutation's history. When mutations arise, they occur on a particular genetic background, which creates a correlation with the other SNPs on the chromosome. Second, the mixing of populations known as admixture that occurs typically through migration means that SNPs with population-specific frequencies will be correlated in a larger mixed sample. In this case, population stratification is the cause of the correlation, and there has been much genetic epidemiological research on this phenomenon and how to control for it. Population stratification is thought to be a possible source of spurious genetic associations with disease (see Box 3-2).

CONCLUSION

In large part, the twentieth century was dominated by studies of human health and disease that focused on identifying single genetic and environmental agents that could explain variation in disease susceptibility. This new century has been characterized by huge advances in our understanding of Mendelian disorders with severe clinical outcomes. However, the Men-
delian paradigm has failed to elucidate the genetic contribution to susceptibility to most common chronic diseases, which researchers know have a substantial genetic component because of their familial aggregation and studies that demonstrate significant heritabilities for these diseases. Likewise, environmental and social epidemiological studies have been wildly successful in illuminating the role of many environmental factors such as diet, exercise, and stress on disease risk. However, these environmental factors still do not, by themselves, fully explain the variance in the prevalence of several diseases in different populations. Researchers are only now beginning to study in earnest the potential interactions between the genetic

**BOX 3-2**

**Population Stratification (Confounding)**

When the risk of disease varies between two ethnic groups, any genetic or environmental factor that also varies between the groups will appear to be related to disease. This phenomenon is called "population stratification" in epidemiologic studies investigating the effect of a genetic factor on disease, and it is a form of confounding. Population stratification refers to the presence of subgroups—for example ethnic groups—in the sample, which could potentially cause a spurious association between genetic variations and trait. Concerns about population stratification have raised doubts about the credibility of some reported findings in candidate gene studies and have led to calls for the routine use of related controls in case-control studies of genetic factors to eliminate the possibility of population stratification (Lander and Schork, 1994; Altshuler et al., 1998). In fact, although population stratification is frequently used as an explanation for nonreplicable associations in the literature, there are few actual examples to support this assumption (Risch, 2000) and many agree that the problem probably has been overstated (Cardon and Bell, 2001). For example, Wacholder et al. (2000) argued that population stratification to an extent large enough to distort results is unlikely to occur in many realistic situations. Population stratification is a manifestation of confounding—that is, the distortion of the relationship between the exposure of interest and disease due to the effect of a true risk factor that is related to the exposure (Wacholder et al., 2000). Thus, in population stratification ethnicity acts as a surrogate for the true risk factor, which may be environmental or genetic. This means that controlling for ethnicity can reduce the confounding bias.

Ardlie et al. (2002) evaluated four moderately sized case-control studies for the presence of population structure and concluded that carefully matched case-control samples in U.S. and European populations are unlikely to contain levels of population stratification that would result in significantly inflated numbers of false positive associations. However, methods have been developed by which unlinked genetic markers can be used to detect stratification and even correct for it when it is present (Pritchard and Rosenberg, 1999; Satten et al., 2001).
and environmental factors that are likely to be contributing to a large fraction of disease in most populations. There is much that can be done to incorporate measures of social environment into genetic studies and to also incorporate genetic measures into social epidemiological studies.

Over the last two decades, progress in identifying specific genes and mutations that explain genetic susceptibility to common conditions has been relatively slow, for a variety of reasons. First, the diseases being studied tend to be complex in their etiology, meaning that different people in a population will develop disease for different genetic and/or environmental reasons. Any single genetic or environmental factor is expected to explain only a very small fraction of disease risk in a population. Moreover, these factors are expected to interact, and other biological processes (e.g., epigenetic modifications) are likely to be contributors to the complex puzzle of susceptibility. An accurate phenotypic definition of disease and its subtypes is crucial to identifying and understanding the complexities of disease-specific genetic and environmental causes.

Second, geneticists only recently have developed the knowledge base or methods needed to measure genetic variations and their metabolic consequences with sufficient ease and cost-effectiveness so that the large number of genes thought to be involved can be studied. With the completion of the Human Genome Project in 2003, many different scientific entities (e.g., the Environmental Genome Project and the International HapMap Consortium) have been working to identify the mutational spectra in human populations, and genetic epidemiologists are just now beginning to understand the extensive nature of common variations (>1 percent population frequency) within the human genome that could be affecting people's risk of disease. The SNP data generated by these initiatives are now centrally located in a number of public databases, including the National Center for Biotechnology Information’s dbSNPs database, the National Cancer Institute’s CGAP Genetic Annotation Initiative SNP Database, and the Karolinska Institute Human Genic Bi-Allelic Sequences Database. At present, the largest dataset on human variation is being generated by the International HapMap Project,4 which is genotyping millions of SNPs on 270 individuals from 4 geographically separated sites from around the world. The International HapMap Project has greatly increased the number of validated SNPs available to the research community to be used to study human variation and is producing a map of genomic haplotypes in four populations with ancestry from parts of Africa, Asia, and Europe. In addition, high-throughput methods of genotyping large numbers of SNPs (thousands) in large epidemiological cohorts are only now becoming available

4See www.hapmap.org.
Unfortunately, high-throughput methods of measuring the environment have not kept a similar pace. For many studies of common disease, a rate-limiting step to increasing our understanding will continue to be the difficult and costly measurement of environmental factors.

Finally, progress also has been hampered because of a lack of adequate investment in developing new methods of analysis that can incorporate the high-dimensional biological reality that we can now measure. The complex genetic and environmental architecture of multifactorial diseases is not easily detected or deciphered using the traditional statistical modeling methods that are focused on the estimation of a single overall model of disease for a population. For example, using traditional logistic regression methods it would be simply impossible to enter all the hundreds of genetic variations that are thought to be involved in CVD risk or in any of the other common disease complexes currently being studied. Beyond the obvious issues of power and overdetermination in such a large-scale model, we also do not know how to model or interpret interactions among many factors simultaneously or how to incorporate the rare, large effects of some genes relative to the common, small effects of others. New modeling strategies that take advantage of advances in pattern recognition, machine learning, and systems analysis (e.g., scale-free networks, Bayesian belief networks, random forest methods) are going to be needed in order to build more comprehensive, predictive models of these etiologically heterogeneous diseases.

The field of human genetics, like many other disciplines, is in transition, and there is much to be gained by joining forces with a wide range of other disciplines that are focused on improving prevention and reducing the disease burden in our populations.

REFERENCES


Genetic, Environmental, and Personality Determinants of Health Risk Behaviors

INTRODUCTION AND OVERVIEW

Tobacco use, obesity, and physical inactivity are the greatest preventable causes of morbidity and mortality in the United States (Mokdad et al., 2004). These behaviors involve motivational and reward systems within the individual that develop through gene interactions with the social environment. Therefore, a better understanding of the genetic, social environmental, and individual determinants of risk behaviors, such as tobacco use, unhealthy eating behaviors, and physical inactivity could contribute to improved strategies for primary, secondary, and tertiary disease prevention.

Models of gene, environment, and behavior interactions in disease have been proposed, one of which has been adapted here to illustrate the central role of health risk behaviors (Rebbeck, 2002). Risk behaviors such as tobacco use, unhealthy eating behaviors, and physical inactivity play an important role in models of genetic and environmental interactions in health outcomes. As illustrated in Figure 4-1, gene-environment interactions contribute to the initiation and maintenance of these risk behaviors, which in turn increase risk for poor health outcomes (pathway a). In addition, gene-environment interactions can modify the effects of these risk behaviors on disease states and health outcomes (pathway b) and also can have direct effects on health outcomes (pathway c) (see also models of gene-environment interactions in Chapter 8 and Appendix E).

The goal of this chapter is three-fold: (1) to provide a brief overview of the epidemiology of tobacco use, unhealthy diet/obesity, and physical inactivity in relation to health outcomes; (2) to describe the genetic and environ-