Impact of the Human Genome Project on Epidemiologic Research

Darrell L. Ellsworth,1 D. Michael Hallman, and Eric Boerwinkle

INTRODUCTION

Traditionally, the fields of human genetics and epidemiology were independent disciplines with minimal interaction between them. Over the years, however, genetic concepts were slowly integrated with epidemiologic methods to capitalize on the advantages of their diverse perspectives and expertise (1). By the late 1970s, genetic epidemiology emerged as a formal discipline and blossomed throughout the ensuing decades as advances in molecular biology advanced our understanding of human genetic disease (2-5). Genetic epidemiology relates genetic characteristics that may be influenced by environmental exposures to the distribution of disease among relatives and within diverse human populations. The primary objective of this field is to better understand the genetic etiology of disease in order to facilitate early prediction and design more effective intervention strategies.

The discipline of genetic epidemiology has greatly expanded the applicability and utility to the public of genetic advancements, including the mapping of genes responsible for Mendelian diseases (such as cystic fibrosis) (6-8) and the development of models to predict disease (such as the multiple-step mechanism of carcinogenesis) (9, 10). The greatest challenges confronting genetic epidemiology, however, are the common chronic diseases with late age-of-onset which exert a tremendous burden on public health as measured by morbidity, mortality, and cost. The greatest impact and benefit to public health from genetic epidemiologic research will likely come from uncovering and better understanding the genetic etiology of the common chronic diseases (such as coronary artery disease and diabetes) and the common forms of cancer (such as breast and colon cancer).

Until recently, genetic epidemiology made inferences primarily from statistical analyses of the distribution of disease or other traits among family members. Direct measures of genetic information were rare and, with the exception of the human leukocyte antigen (HLA) complex, were limited to red cell antigens and polymorphic red cell and plasma enzymes (11). For the field of genetic epidemiology to achieve its full potential and better characterize the genetic etiology of the common chronic diseases, high quality genetic markers were necessary for gene mapping, and improved methods needed to be developed to detect and quantify functional alleles. The Human Genome Project and other developments in molecular biology are providing the necessary tools for epidemiology and genetic epidemiology to uncover the molecular mechanisms for variation in the distribution of disease among families and populations. Although there have been other conceptual and technical advances (particularly in the area of computational methods) that have advanced the field of genetic epidemiology, progress in this area is not the subject of this review. In this presentation we briefly review the objectives of the genome project, the type of information provided by the genome initiative, and its utility to epidemiologic research. Our thesis is that the Human Genome Project is providing valuable tools to further the objectives of genetic epidemiology while simultaneously broadening its scope. However, it is incumbent upon both genetic and other epidemiologists to actively acquire this information and utilize it to its full advantage.
THE GENOME PROJECT

The Human Genome Project is a cooperative multinational initiative with the ultimate goal of determining the complete DNA sequence of the human genome as well as the genomes of several model organisms (12–15). DNA is a macromolecule which represents the molecular basis of heredity and consists of a linear array of deoxyribonucleotides. Each deoxyribonucleotide is composed of sugar (deoxyribose) and phosphate groups as well as a nitrogenous base which can be either a purine (adenine (A) or guanine (G)) or a pyrimidine (cytosine (C) or thymine (T)). The precise order of the nitrogenous bases along the DNA encodes the genetic information in a code that is universal among organisms. The human genome consists of a linear arrangement of approximately three billion deoxyribonucleotides partitioned into 22 autosomes and two sex chromosomes (X and Y), as well as a small amount of DNA present in the mitochondria (16). The protein-coding portions of the estimated 100,000 genes in the human genome represent only a fraction (5–10 percent) of our genetic material (17, 18). Sequences that do not encode protein, such as introns, intergenic regions, pseudogenes, and repetitive elements whose functions (and importance) are not completely understood at present, comprise the remaining 90 to 95 percent of the genome.

The genome initiative in the United States is coordinated by the National Human Genome Research Institute (NHGRI) at the National Institutes of Health and by the US Department of Energy (19, 20). Goals to be accomplished in the first 5 years (1990–1995) of the Human Genome Project included: 1) completion of genetic maps with markers spaced 2–5 centimorgans (cM) apart (a centimorgan expresses relative distance on a genetic map and is equal to 1 percent recombination, or approximately one million base pairs) and development of technology for rapid genotyping; 2) development of physical maps with a resolution of 100 kilobase pairs (1 kilobase pair (kb) is equal to 1,000 base pairs); 3) complete sequencing bacteria, yeast, nematode, and fruit fly genomes, as well as developing the capability to sequence 10 megabase pairs (1 megabase pair (Mb) is equal to 1 million base pairs) of human DNA per year at a cost of $0.50 per base pair; 4) improving methods of identifying and mapping genes; 5) creating and refining computer databases to handle the enormous amounts of data generated from the genome initiative; and 6) exploring ethical issues related to molecular diagnostics (21).

The genome project has adopted a hierarchical approach to proceed through increasing levels of resolution and detail. An immediate focus is the completion of several types of genomic maps that reflect the organization and coordinate positioning of specific genome landmarks (sequence-based markers). Low resolution cytogenetic maps are produced by examination of photomicrographs (karyotypes) depicting the number, size, and morphology of an individual’s chromosomes. Karyotypes are prepared during metaphase of mitosis (cell division) when chromosomes are sufficiently condensed so as to become visible. Differentially-stained regions (bands) of the chromosomes provide an efficient system for assigning genes or other features of the DNA to specific chromosomal regions. Although chromosomal rearrangements have limited utility in localizing genes influencing complex diseases, some cytogenetic alterations have been useful in the identification of genes responsible for simply inherited genetic disorders such as fragile X syndrome (22) and chronic granulomatous disease (23). In addition to localizing disease genes, the readily distinguishable banding patterns that comprise cytogenetic maps constitute a framework for the construction of more detailed maps.

Genetic (or linkage) maps orient markers relative to one another such that distances between markers are expressed as recombination frequencies rather than as true physical distances measured in base pairs. Genetic distances are inferred from an analysis of the inheritance patterns of marker genotypes in a large number of families. Restriction fragment length polymorphisms (RFLP) are DNA sequence variations detectable by restriction enzymes (enzymes that cut the DNA molecule at specific sequences) that were used in the construction of early low-resolution genetic maps. However, due to their relatively low information content for genetic linkage analysis (the site of cleavage is present or absent), restriction fragment length polymorphisms have been largely supplanted by microsatellites which consist of two to four base pair sequences (such as CA or GATA) that are tandemly repeated. Microsatellites typically exhibit high levels of variability due to differences in the number of repeats (24). Continuous refinement of comprehensive genetic maps increases the density of high quality markers. For example, the 1994 Généthon human linkage map contained 2,066 short tandem repeat polymorphisms (25), while the latest version (1996) contains 5,264 microsatellites with an average interval of 1.6 cM (26).

Physical maps, which contain ordered DNA markers at known (or closely approximated) distances (in base pairs rather than centimorgans), are constructed by assembling sets of overlapping DNA fragments (contigs) using sequence-tagged sites (STS) (27). A sequence-tagged site is a unique DNA sequence (typically 200–500 base pairs in length) that is readily
detectable by a polymerase chain reaction (PCR) assay and is used to identify long DNA fragments that are assumed to overlap because they share a common sequence-tagged site. Physical maps are extremely useful for determining the positions of genes associated with disease and can provide a primary scaffold for initiating large-scale DNA sequencing. A recent physical map of the human genome with extensive long-range continuity contains more than 15,000 sequence-tagged sites with an average spacing of approximately 200 kilobase pairs (28).

To rapidly identify expressed human genes and provide key resources for gene mapping, recent efforts have focused on isolating and partially characterizing genes that are transcribed as messenger RNA (mRNA). Messenger RNA molecules are transcribed from DNA (gene) sequences in the nucleus and then function out in the cytoplasm by specifying the sequence of amino acids in nascent polypeptides (translation). The partial sequences of messenger RNA molecules, which represent expressed genes, are known as expressed sequence tags (ESTs) and can be localized to specific chromosomal regions and integrated into transcription maps. Transcription mapping is an expedient method for localizing potential disease susceptibility genes because it simultaneously provides location and sequence information. Catalogs of unmapped expressed sequence tags also represent powerful tools for assessing human gene diversity and determining gene function through patterns of expression. Data are currently available on more than 292,878 expressed sequence tags derived from at least 37 distinct organs and tissues (29).

The ultimate goal of the Human Genome Project is to establish the entire 3 billion base pair sequence of the human genome. Only a fraction of the total genomic sequence has been determined to date; however, current methods have enabled complete DNA sequences to be obtained for several genomic regions (30, 31). The largest contiguous segment in humans for which the complete nucleotide sequence has been determined is the B T-cell receptor region (685 kb) containing a complex family of immune recognition genes (32). Refinements in high throughput DNA sequencing technologies, such as automated fluorescence methods (33, 34) and energy transfer primers (35), are anticipated to permit megabases of DNA to be accurately and reliably sequenced within an acceptable length of time and at reasonable cost. Technologic advancements in sequencing methodology (36) and the emergence of novel strategies (37) will significantly increase sequencing speed and efficiency. DNA sequence data will eventually reveal a wealth of information on the organization and biologic complexity of the human genome.

INFORMATION TRANSFER

To take full advantage of the wealth of information generated by the Human Genome Project, epidemiologists and genetic epidemiologists must have the ability to easily access the data, as well as a working knowledge of the retrieval process to properly query, analyze, and interpret the desired information. A primary component of the genome initiative is the development of comprehensive computer databases to assimilate the tremendous amount of mapping and DNA sequence data and to provide links to the scientific and medical literature. Numerous databases have been established to provide organized storage and efficient dissemination of the genome mapping and sequencing data (38, 39). The informatics movement has generated global computer networks with on-line access over the internet that permit remote access and retrieval of raw or computed data. The seemingly unlimited potential of this technology is readily available to epidemiologic researchers who may be unfamiliar with genetics or genome informatics through simplified accession programs and database helplines. These databases are invaluable to epidemiologic research and should be familiar to all genetic epidemiologists because they contain vast amounts of information regarding the genes and molecular defects that contribute to human disease, methods for rapid detection of mutations and polymorphisms (if available), comprehensive descriptions of disease phenotypes, and the status of treatment and intervention strategies.

Annotated DNA sequence information for humans is currently available from more than 677,205 entries in the National Institutes of Health genetic sequence database known as GenBank (www.ncbi.nlm.nih.gov) (release 99.0 in February 1997). The Unique Human Gene Sequence Collection (UniGene) is assimilating DNA sequences to identify and map new human genes (40). Information is currently available on over 55,000 sequence clusters representing the transcription products of distinct genes. The most recent Online Mendelian Inheritance in Man (OMIM) catalog of human genes and genetic disorders (41) contains 8,408 entries that include 5,439 established gene loci and descriptions of 398 inherited disease phenotypes. As the official repository for genomic mapping data resulting from the human genome initiative, the Genome Database (GDB) (www.gdb.org) organizes and stores data, including map locations of DNA markers and genetic disease locus and probe information submitted by genome researchers worldwide, and provides this information electronically to the scientific community.
Genome Database currently contains information on more than 1.5 million clones (physically-isolated DNA fragments) and nearly 18,000 polymorphisms which are accessible with state-of-the-art genomic map viewing software.

LOCALIZING DISEASE GENES

Genetic linkage

Genetic linkage analysis is a common method for localizing genes contributing to human disease to an approximate chromosomal region (42). By definition, linkage exists when two genes cosegregate from parents to offspring more often than would be expected by chance and is due to the close proximity of the two genes along a chromosome. In the search for genes associated with disease, linkage is inferred to exist when cosegregation has been detected more often than expected by chance between a marker (usually a highly polymorphic microsatellite) and a gene affecting disease susceptibility. Traditional linkage methods for single gene disorders utilize a likelihood approach to evaluate the strength of evidence for linkage relative to that for no linkage by calculating LOD scores (decimal log likelihood ratios) which require the mode of inheritance for the disease to be specified a priori. However, modes of inheritance for the common chronic diseases are complex and often heterogeneous among families. Because LOD scores applied to multifactorial diseases may have inflated error rates, robust “nonparametric” methods of linkage analysis are preferred for complex diseases (43–45). Methods using affected relative pairs are available for qualitative traits (e.g., carotid artery atherosclerosis) but are often limited by low power. Analyses of quantitative traits (e.g., cholesterol levels) in entire pedigrees may prove to be more appropriate and informative.

The Human Genome Project has created an invaluable tool for genetic linkage studies—an integrated map of microsatellite markers that are highly informative for detecting linkage and are abundant throughout the genome. Constant refinement of genetic maps and an increasing density of reliable markers enhance our ability to accurately pinpoint the locations of disease genes and facilitate gene identification by defining regions that can be further characterized by physical mapping techniques.

Genome-wide linkage analysis utilizing the human genetic map recently identified a chromosomal region believed to contain a gene contributing to a late-onset common chronic disease. Non-insulin-dependent (type 2) diabetes mellitus (NIDDM) is characterized by hyperglycemia due to defects in insulin secretion and/or action (46). Affecting 10–20 percent of the population over 45 years of age, NIDDM is a leading cause of morbidity and mortality in developed countries. Although considerable progress has been achieved in identifying genes responsible for Mendelian (early-onset) forms of diabetes, little is known about the genes contributing to the common late-onset form(s) of NIDDM that are believed to be influenced by numerous genes as well as environmental factors. Nearly 500 highly polymorphic markers with an average distance between adjacent markers of less than 10 cM were utilized to search for genes contributing to NIDDM in 330 Mexican-American affected sibling pairs (47). A number of candidate genes throughout the genome showed no evidence of linkage in this sample. Conversely, a single microsatellite marker showed highly significant linkage to NIDDM and may indicate the presence of a gene that is a major contributor to disease susceptibility in this population. Linkage analyses that consider multiple markers simultaneously (multipoint analyses) indicated that 71 percent of the genome could be excluded as containing a locus having an effect large enough to increase the relative risk of disease (λc) by 1.6 in individuals possessing the susceptible genotype. However, only 5 percent of the genome could be excluded as containing a locus having an effect large enough to increase the relative risk of disease by 1.2.

Association and transmission disequilibrium tests

Association studies compare the frequency of alleles (alternate forms of a given gene which differ in DNA sequence) between unrelated affected (case) and unaffected (control) individuals. A given allele is considered to be associated with the disease if that allele occurs at a significantly higher frequency among cases relative to controls. Khoury et al. (4) provide a more complete description of methods for genetic association studies. Association analysis may be more sensitive than linkage methods when the genes being sought contribute to disease susceptibility but are neither necessary nor sufficient to cause disease. When the relative risk of disease given the susceptible genotype is small, detecting genetic linkage becomes increasingly difficult (48).

Transmission disequilibrium tests (TDT) have recently been introduced to avoid some of the limitations and pitfalls inherent in most linkage and simple association studies (49, 50). Most transmission disequilibrium statistics consider parents who are heterozygous for an allele hypothesized to be associated with disease and evaluate the frequency with which that allele (or its alternate) is transmitted to affected offspring. Transmission of the disease-associated allele to af-
fected individuals will be greater than expected under Mendelian (random) segregation if the locus is located near the disease gene. Transmission disequilibrium tests are linkage-based association tests that may be appropriate for complex diseases because they greatly reduce the likelihood that any allele frequency differences between affected and nonaffected individuals might be due to poorly-chosen controls or unsuspected genetic differences among subgroups within the population.

Association studies and transmission disequilibrium tests are likely to play critical roles in mapping complex disease genes. However, such methods will require large population samples and detailed, accurate maps to localize genes that may contribute no more than 5 percent to the total genetic variance of a complex disease. Current physical maps may contain markers at sufficient densities (average spacing of 10 cM) for locating monogenic disease genes and initiating DNA sequencing, but localization of genes contributing to multifactorial diseases will likely require denser maps (on the order of 1 to 3 cM between adjacent markers) because linkage disequilibrium decays quickly with distance. Constant refinement of genetic maps by verifying marker order and increasing marker density will be critical to the successful identification of multifactorial disease genes with small individual effects.

Gene identification

Phenomenal progress in identifying and isolating genes for monogenic diseases has been achieved through “positional cloning” strategies that locate genes using genetic and physical mapping techniques with only minimal information about the function of the gene or the basic biochemical defects. Linkage and association analyses in families affected with the disease are typically used to define an initial candidate region in which the responsible gene is believed to be located. The candidate region may be narrowed using information from patients carrying large cytogenetic rearrangements and/or deletions. Fine-structure genetic mapping (a high-resolution analysis of exchange events between chromosomes that occur during meiosis) may further delimit the interval which can then be characterized with physical maps. After a thorough inventory of genes and expressed sequences within the region, mutation screening must be conducted to identify the causative gene. Each candidate is surveyed for mutations and the responsible gene is identified by alterations in individuals affected with the disease. To date, positional cloning efforts have led to the successful localization and characterization of more than 40 human disease genes. Although many genes identified thus far (for monogenic diseases) may be of limited interest to epidemiologists, these initial successes using genome technology provide a foundation for the development of techniques and approaches for localizing genes contributing to complex diseases.

Most single gene disorders are characterized by a low frequency of the disease allele in the general population and high penetrance (a large proportion of individuals with the disease allele show symptoms of the disease). Complex disorders are characterized by high levels of genetic complexity, difficulties in early stage diagnosis, late onset of clinical symptoms, and probable gene-by-environment interactions. Alleles associated with increased susceptibility to multifactorial diseases are often common in the general population, and a given gene may contribute only a small proportion to the total genetic variance underlying the affliction. Therefore, traditional approaches for localizing Mendelian disease genes may not be feasible for genetically complex disorders. Locating a single gene within a chromosomal region that is typically 2–5 megabase pairs in length is severely hampered by the absence of chromosomal rearrangements or deletions that define the candidate region and by the often subtle nature of functional sequence variation which may be located outside the coding region.

Once a complex disease gene has been localized to a defined genetic interval, genes previously mapped within the critical region become strong candidates. This positional candidate approach involving linkage analysis of multiple affected family members to localize susceptibility genes to chromosomal regions, followed by an intensive search for logical candidates within the interval, may prove to be an efficient strategy for locating genes contributing to the common chronic diseases. Explosive growth in the construction and refinement of transcription maps through the efforts of an international consortium is expediting the discovery and characterization of genes mutated in human disease. A recent analysis indicates that 71 percent of human disease genes isolated by positional cloning are represented by at least one expressed sequence tag in a publicly-accessible database. The expressed sequence tags division (dbEST) of GenBank, which is part of the International Nucleotide Sequence Database Collaboration, now contains information on hundreds of thousands of expressed sequence tags derived from numerous human tissues or cell types that may be retrieved electronically over the internet.
DNA SEQUENCE VARIATION AND FUNCTIONAL MUTATIONS

Once potential candidate genes have been identified, an exhaustive search must be conducted for DNA variation within the candidate region. A variety of scanning techniques have been utilized for the initial detection of unknown mutations within relatively large genomic regions (61). Most methods detect the differential migration under electrophoresis of single-stranded DNA elements that differ in sequence (62, 63) or alterations in the melting points of double-stranded DNA molecules (64). These techniques detect, with varying efficiencies, the presence of mutations, but they do not identify the precise location or nature of the structural change (65).

Modern genome sequencing technologies now permit the search for DNA variation in linked regions to proceed by direct DNA sequencing (66) without the need to conduct the single-stranded conformation (or similar) analyses mentioned above. Direct sequencing methods both locate and characterize all DNA variation within a region. Characterizing the structure of newly discovered genes and identifying new sequence polymorphisms in previously uncharacterized regions will greatly expedite the search for DNA sequence variation implicated in genetic disease.

Methods of mutation detection identify DNA variants but provide no information regarding their biologic significance. Sequence alterations, such as single nucleotide substitutions, may or may not be functionally relevant. The task of distinguishing DNA variants that contribute to disease from neutral polymorphisms is one of the most intellectually challenging problems confronting human geneticists. Several approaches (67, 68) have been developed to help pinpoint causative variants or reduce the number of potential candidates that require further investigation.

An indirect result of the Human Genome Project has been the rapid expansion of protein sequence databases that aid in the quantification of functional mutations. Resources such as the Molecular Modelling Database and computer software programs such as RASMOL (69) are particularly useful to geneticists and molecular biologists in locating amino acid substitutions that alter protein structure (and possibly function) and determining the spatial position of variants relative to known functional sites within the protein. The accumulation of protein structure and sequence data will continue to provide a wealth of information on the biologic functions of these macromolecules. Using the tools of modern molecular biology and human genetics, a primary objective of genetic epidemiology should be focused on defining the underlying functional mutations and exploring possible disease mechanisms that culminate in clinically apparent disease.

COMPARATIVE GENOMICS AND ANIMAL MODELS

The “Human” Genome Project is actually a diverse initiative that includes the parallel mapping and sequencing of selected model organisms, including bacteria (Escherichia coli), yeast (Saccharomyces cerevisiae), nematode (Caenorhabditis elegans), fruit fly (Drosophila melanogaster), and mouse (Mus musculus), whose genomes increase progressively in size and structural complexity. Critical nucleic acid structures and protein functions are frequently conserved throughout evolution across a diverse array of organisms. Detailed comparisons among a variety of species are therefore useful in deciphering structural information encoded in the DNA and provide insight into the functional significance of genomic sequences. A large number of genes present in humans have counterparts in other species, allowing sequence homology between species to be used to detect genes and regulatory elements in newly-characterized segments of human DNA sequence (70). Important similarities in chromosomal structure and gene function between study organisms and humans will prove invaluable in the difficult process of determining gene functions and mechanisms of genetic disease etiology (e.g., 71).

Gene targeting technology permits specific alterations to be made in selected genes within the genomes of model organisms (particularly the mouse). Targeting has been used to disrupt native genes thereby generating “knockout” animals completely lacking the product of a particular gene. Knockouts are created with targeting plasmids (extrachromosomal genetic elements) containing an altered version of the gene of interest which can be introduced into embryonic stem (ES) cells. Through homologous recombination, a portion of the native gene is replaced by the introduced variant thereby disrupting its structure and normal functioning (72). Transgenic animals containing a functional copy of a foreign gene (such as a human gene) may be produced by assembling a DNA construct containing the gene of interest along with regulatory elements necessary for expression followed by microinjection into fertilized mouse oocytes (one cell stage embryos) which are then implanted into pseudopregnant females (73, 74). Properly designed gene targeting studies have the ability to evaluate the physiologic effects of precise genetic changes while simultaneously eliminating or minimizing environmental effects as well as the effects of other genes.

The ability to manipulate the genomes of model organisms by disrupting native genes and/or introduc-
ing human genes has significantly advanced our understanding of numerous monogenic disorders in humans. However, dissecting genetic factors contributing to disease and distinguishing between causation and correlation are more difficult for complex diseases with multiple genetic and environmental components. Although we do not fully understand the degree to which animal knockout models are biologically relevant to human diseases (particularly those with complex etiologies) or whether information from single gene disruptions can be extrapolated to multifactorial conditions, recent advances in transgenic technology (74, 75) have proven useful for examining the often modest effects of complex disease genes, independent of other genes and environmental factors that may influence susceptibility (76–78). Transgenic animals provide information on the metabolic functions of genes and the relation between genetically determined alterations in gene dosage and predisposition to disease. We anticipate the ability to test human functional variants in animal systems as refinements in gene replacement technology allow intact disease-associated or nonassociated alleles to be introduced in defined copy number and as carefully-structured breeding programs eliminate phenotypic effects caused by polymorphic differences among animals.

CLINICAL AND DIAGNOSTIC APPLICATIONS
Molecular diagnostics

A major focus of research on the epidemiology and prevention of disease is the ability to identify individuals at increased risk and to predict disease before onset of clinical manifestations. An increasingly effective strategy for defining individuals at increased risk involves the identification of specific DNA polymorphisms that are associated with disease. Molecular (DNA) diagnostics is a rapidly expanding (in both scope and importance) discipline of medical genetics that encompasses a diverse array of clinical applications from the diagnosis of genetic disorders (79) and neoplastic conditions (80) to the identification of infectious disease agents (81). Technical improvements and sophisticated variations of recombinant DNA technologies are increasingly being applied to detect disease-associated mutations in human genes. For example, diagnostic tests are currently available for numerous single gene disorders such as Duchenne and Becker muscular dystrophies (82) and cystic fibrosis (83) and for various forms of cancer (84, 85). Genetic lesions known to be responsible for human inherited diseases are already being collated in a comprehensive online reference source, the Human Gene Mutation Database (available through OMIM), which provides information of practical diagnostic importance to geneticists, physicians, and genetic counselors.

The majority of human diseases involve multiple genes that may interact with each other and whose effects are often mediated by the environment. Due to the high prevalence of complex diseases, such as cardiovascular disease, diabetes, and certain cancers in the general population, the capacity to identify those at increased risk of disease could lead to preventive measures (lifestyle changes) and targeted intervention strategies designed to modify risk and/or prevent premature onset of disease. Many of the genes contributing to multifactorial diseases have not yet been characterized, and the ability to detect DNA sequence variation predisposing to such diseases is often beyond our diagnostic capabilities. Fortunately, several genes implicated in the occurrence of complex diseases or common forms of cancer have been characterized where specific mutations or variants that are common in the population contribute to disease. One example of such a gene is apolipoprotein (apo) E. Apolipoprotein E is a structural constituent of several lipoprotein species and plays a major role in lipid metabolism through cellular uptake of lipoprotein particles (86). The human apolipoprotein E gene is polymorphic with three common alleles (e2, e3, and e4) (87). Various studies have shown that the effects of this gene are relatively consistent across ethnically and geographically diverse populations—the average effect of the e2 allele is to lower total serum cholesterol levels while the average effect of the e4 allele is to raise total cholesterol levels (88, 89). The e2 allele is hypothesized to have a protective effect on the development of atherosclerosis (90) because it is associated with lower cholesterol levels and is more frequent in patients with no or minimal atherosclerotic involvement (91). Conversely, a number of epidemiologic studies have reported an association of the e4 allele with cardiovascular disease (92–94). The multifactorial etiology of, and environmental influences on, cardiovascular disease make it difficult to accurately predict disease risk for specific individuals. However, efforts to reduce the prevalence of known risk factors in the general population and in particular at-risk subgroups are effective intervention strategies.

The recent identification of genes influencing hereditary breast and ovarian cancers heightened enthusiasm that such discoveries would improve the ability to identify individuals most at risk of developing breast and ovarian cancer and would be key to better understanding all forms of cancer. Breast cancer represents the most common form of cancer among women in westernized countries; cumulative lifetime risk for non-Jewish women in the general population is
approximately 12 percent, but risk may approach 85 percent in families of Ashkenazi Jews that carry disease-associated mutations (95). Breast cancer susceptibility genes that may account for 3–7 percent of all familial breast cancer cases were recently isolated and characterized (BRCA1 in 1994 (96) and BRCA2 in 1995 (97)). Despite these highly publicized discoveries, reliable population screening programs and effective treatment and prevention options are not yet available. Difficulties in relating DNA variation to risk for breast cancer are attributable to our current lack of knowledge regarding the function and regulation of the BRCA genes and a plethora of nonrecurrent (seen in only one or a few families) mutations (98). Researchers have identified more than 235 different sequence variations within BRCA1 and approximately 100 mutations in BRCA2 (99). Prevalence estimates for the various mutations in the general population are not yet available, and the risk of disease imparted by specific mutations remains unknown. The heterogeneity of variation within the breast cancer genes may reduce the effectiveness of potential diagnostic tests, leading to high frequencies of both false positive and false negative results. Commercial tests are now available to detect specific mutations, and numerous other diagnostic tests are undoubtedly in development. However, the ability to detect mutations influencing disease risk may quickly outdistance our ability to develop effective measures for prevention and treatment.

The progressive characterization of disease genes through the Human Genome Project, coupled with an improved ability to identify the molecular defects contributing to disease, is expected to revolutionize the molecular diagnosis of genetic diseases. Ultimate improvements in the ability to diagnose genetic disease at the DNA level may advance our knowledge of genetic disease etiology, but such progress requires a concomitant acceleration in therapeutic, intervention, and prevention options. Dissemination of technologic developments associated with genome research (such as polymerase chain reaction and DNA sequencing) to the fields of medical genetics and genetic epidemiology must not only enhance our ability to diagnose and predict genetic disease but should also provide future directions for prevention and treatment.

Ethical and social issues in diagnostic molecular genetics

Prior to the genome initiative, ethical issues in molecular genetics focused primarily on monitoring and regulating experimentation in recombinant DNA and genetic engineering. With the inception of the Human Genome Project, a joint National Institutes of Health/Department of Energy working group, the Ethical, Legal, and Social Issues (ELSI) Program, has been established to examine various issues associated with the generation and dissemination of a vast array of genomic information. High-priority issues initially targeted by the Ethical, Legal, and Social Issues Program for development of policies and guidelines included: 1) the integration and impact of new genetic tests and the debate over population screening; 2) privacy and confidentiality of genetic information; 3) genetic counseling and reproductive decisions influenced by diagnostic results; and 4) public education (100). In recent years, the Ethical, Legal, and Social Issues Program has also emphasized technical problems such as the potential for genetic discrimination, educating physicians in the advantages and limitations of genetic data, quality control in DNA testing laboratories, and defining guidelines for obtaining informed consent for genetic research (101).

As new genetic assays are introduced into clinical practice, rigorous adherence to established protocols and quality control assurance are of paramount importance. Attention must be directed toward the debate over implementing population-wide screening programs as routine practice in clinical medicine to detect those at increased genetic risk. The increasing ability to diagnose individuals at risk for genetic diseases for which there are no therapeutic options will require enactment of measures to 1) prevent insurance and/or employment discrimination against asymptomatic carriers and 2) accommodate the psychologic needs of those who are likely to develop a late-onset condition. Improvement of noninvasive prenatal and preimplantation diagnostic procedures is creating an immediate need to explore the ethical dilemmas and difficult reproductive choices faced by prospective parents known to carry disease-associated genes (102, 103). Increasing public awareness of the availability, benefits, and limitations of molecular diagnostic tests is anticipated to simultaneously improve health care delivery while minimizing the potential for psychologic and social stigmatization. The impending explosion in the number of well-characterized human disease genes and new abilities to diagnose genetic disorders will likely necessitate development of novel avenues for education and genetic testing. Careful integration of genetic information with a practical system for characterizing and resolving ethical and social issues will provide future directions for the fields of molecular genetics and clinical medicine.

CONCLUSIONS

Rapid discoveries of novel genes for a variety of human diseases are anticipated as genomic maps be-
come more detailed and methods for mapping and characterizing disease genes become more refined.
Recent and continuing developments in genome technology and analytic methods provide the tools and raw materials for unraveling the complexities of the common chronic diseases and common forms of cancer. Perhaps the greatest utility of the vast genetic information being generated by the human genome initiative is in primary prevention programs. Methods for identifying asymptomatic individuals at risk for genetic disease and the development of more efficacious intervention strategies are becoming paramount as health care costs escalate and medical genetics shifts to early detection and prevention of disease. With these abilities will come the need to fully integrate genetic information into large prospective studies, and intervention trials to accurately predict disease risk and synthesize new approaches to risk reduction. Understanding the role of genes in human disease will improve our understanding of genetic disease etiology as well as our ability to predict disease. Insight into the genetic basis of chronic disease etiology will have immediate impact by suggesting novel therapeutic approaches and aiding new drug discovery.

ACKNOWLEDGMENTS
This work was supported by grants HL51021 and HL54481 from the National Heart, Lung, and Blood Institute.
The authors thank their collaborators and friends in the Rochester Family Heart Study, the Atherosclerosis Risk In Communities project, and the Starr County Health Studies for years of support and encouragement.

REFERENCES
12 Ellsworth et al.


