American Gastroenterological Association (AGA) Future Trends Committee Report: The Application of Genomic and Proteomic Technologies to Digestive Disease Diagnosis and Treatment and Their Likely Impact on Gastroenterology Clinical Practice

KONSTANTINOS N. LAZARIDIS and BRIAN D. JURAN
Center for Basic Research in Digestive Diseases, Division of Gastroenterology and Hepatology, Mayo Clinic College of Medicine, Rochester, Minnesota

The American Gastroenterological Association (AGA) Future Trends Committee was created in 2004 to further the AGA Strategic Plan by identifying and characterizing important trends in clinical practice and scientific-technological developments in the world in general and medicine and gastroenterology in particular that potentially will impact the AGA and/or its members in the coming 3–5 years or beyond and to make strategic recommendations to the Governing Board on how AGA should deal with those trends and developments. These trends and developments may be economic, demographic, practice-based, scientific/technological or political in nature.

Specifically, the committee is charged with preparing a report (or reports) for the AGA Governing Board that describes the trends or developments it has identified, postulates their impact on gastroenterology practice and/or research as appropriate, and presents specific recommendations for action by the AGA in terms of policy and programs. The committee is also asked to monitor these trends and technologies as they play out over time.

In July 2004, the AGA Leadership Cabinet suggested several topics that the Future Trends Committee should address. Realizing that the Future Trends Committee could not realistically consider all of them, criteria were developed to prioritize the topics and others that might be added in the future. These criteria were as follows:

- Time variable, that is, “when will gastroenterology be affected?”
- Scale and magnitude
- Does the trend or development represent a threat or opportunity (or both) to gastroenterology?
- Effect on patient care quality and safety
- Effect on AGA members and the AGA per se
- Implications to reimbursement
- Impact on gastroenterologists’ training and education.

In October 2004, a crude Delphi process was used to determine the trends and developments that should be the focus of the committee’s work. Committee members were asked to assign priority scores to the items in the following list, which was based on the suggestions of the Leadership Cabinet and supplemented by AGA staff and others. This process was done via the mail.

- The application of genomic and proteomic technologies to digestive disease diagnosis and treatment
- Major changes in the US health care system and reimbursement
- Increased median age of the population
- Changes in the ethnic and racial makeup of the US population
- Patients’ involvement in their own care
- New colorectal cancer screening and diagnostic technologies
- Biomedical research funding changes
- Changes in academic health centers
- Changes in physician education and training
- Obesity-related disease incidence and prevalence
- Computerization and digitization of gastroenterology practice

**Abbreviations used in this paper:** AGA, American Gastroenterological Association; BE, Barrett’s esophagus; CS, celiac sprue; CUC, chronic ulcerative colitis; FAP, familial adenomatous polyposis; GI, gastrointestinal; HGP, Human Genome Project; IBS, irritable bowel syndrome; kb, kilobase; LD, linkage disequilibrium; MALDI-TOF, matrix-assisted laser description ionization-time-of-flight; MS, magnetic sectors; NHGRI, National Human Genome Research Institute; PCR, polymerase chain reaction; PSC, primary sclerosing cholangitis; SNP, single nucleotide polymorphism.

© 2005 by the American Gastroenterological Association

0016-5085/05/$30.00
doi:10.1053/j.gastro.2005.06.047
Committee members were asked to score each item against each of the priority criteria noted previously using a scale in which 1 represents large effect and 3 represents small effect (on gastroenterology practice and research). The total scores of each topic were then summed and ranked. The 4 highest priority scores that resulted from this ranking were as follows:

New colorectal cancer screening and diagnostic technologies
Obesity-related disease
Aging of the population
Genomic and proteomic technologies.

Because the AGA was already investigating the ramifications of the obesity epidemic, the Future Trends Committee decided to concentrate on the other 3 topics.

The committee determined that preparing the 3 reports on its own was not feasible. Hence, it decided that it would solicit proposals from potential qualified authors to draft the reports and would modify and supplement the drafts as necessary. A request for proposal was prepared and disseminated in December 2004. The authors, who were paid for their work, were chosen by the committee from among the responses to the request for proposal. The manuscripts submitted by the authors were reviewed by the committee in February 2005. Among the changes to the draft reports were recommendations for action by the AGA; these were developed primarily by the committee. At its review meeting, the committee also developed a uniform format for the 3 reports. Revised manuscripts based on the committee’s critiques were completed in March 2005. The committee also had each report evaluated by an outside expert reviewer for completeness and to ensure that the authors had not made any egregious error that may have been overlooked.

This report represents the committee’s recommendation for action by the AGA on this important topic. However, it is not the committee’s final word on the topic. Genomic and proteomic technologies will advance rapidly over the coming years, and the committee will revisit this subject periodically.

**Executive Summary**

Medicine is on the verge of an unprecedented prospect as a result of advances in the discipline of genomics and recent progress in the emerging field of proteomics. Human genomics is the study not just of single genes but also of the functions and interactions among all genes in the genome of humans. The significant evolution that has occurred in genomic science over the past 5 years holds promise to change our ability to better understand, diagnose, treat, and potentially prevent human illness. This unique opportunity stems from the completion of the Human Genome Project (HGP) and the development of novel technologies, several of which involve high-throughput automated assay systems (ie, genotyping platforms) and the application of bioinformation science (ie, bioinformatics). In a similar vein, the discipline of human proteomics is the study of the interactions among the various constituents of the entire proteome of humans. Human proteomics represents an extension of traditional biochemistry coupled with novel technologies (ie, tandem mass spectrometry) that seeks to take a more global approach to the assessment of the library of proteins specific to humans and understanding these proteomic relationships to health and disease.

Because of the inevitable interconnection of the genome and proteome, genomics and proteomics should be viewed as complementary, rather than antagonizing, scientific fields. Nevertheless, the structure and function of the human proteome are far more complicated than those of the human genome. Simply stated, the genome (ie, genomic DNA) is a static, unwavering entity regarding its sequence and own duplication. In contrast, the proteome is a dynamic, ever-changing unit. For instance, the protein expression profiles in different human cells are highly variable, dependent on external or internal stimuli, not to mention the unique protein expression during the distinct stages of a person’s life cycle. In contrast, the genome of each human cell remains, in general, steady and unaltered over generations.

To date, genomic science has mainly dissected the single-gene inherited diseases known as Mendelian disorders. Yet, the greatest promise and impact of genomic and proteomic research lie in their future application to complex or multifactorial diseases. In antithesis to Mendelian disorders, complex diseases arise due to the interplay of multiple genetic variants with environmental factors. Currently, the science and technology of human genomics greatly exceed the discipline of proteomics with respect to research discoveries and applications on science and influence on medical practice. The ultimate question addressed in this report is whether genomics, along with proteomics, will have an impact on future gastroenterology and hepatology clinical practice. Although we are unable to easily articulate a comprehensive answer to this vital and multifaceted question, we believe that genomics and proteomics hold vast potential and almost certainly will shape the way we diagnose and treat patients with digestive and hepatic disorders in years to come.

For eons, humans have understood that heredity, along with the environment, shape our phenotypic diversity and contribute to disease. Yet, it was not until 2003 that the
entire sequence of the human genome was elucidated via the HGP, providing the biologic basis for a better understanding of heredity and likely the interaction of environment on the genetic material. To date, the enormity of information regarding our own blueprint of life awaits meticulous investigation, the overall aim of which is to identify the genetic components of disease susceptibility, thus improving diagnosis, therapy, and disease prevention. To this extent, well-designed translational studies in gastrointestinal (GI) and liver diseases are urgently needed to transform such colossal genomic knowledge into a meaningful outcome that can be incorporated into clinical practice and benefit the sick.

Since the early 1990s, we have made momentous progress in unraveling the genes causing digestive and hepatic Mendelian diseases such as familial adenomatous polyposis (FAP), hereditary hemochromatosis, and hereditary pancreatitis, to name a few. Although the prevalence of such diseases is low, the understanding of the relevant genetic defects has unquestionably shed light on the biology of the GI system and liver. In the genome era, however, the tools are becoming available to begin dissecting common complex diseases such as irritable bowel syndrome (IBS), nonalcoholic fatty liver disease, inflammatory bowel disease (IBD), and many others. Recognizing the unparalleled genetic and environmental intricacy of those complex diseases, we have to realize that greater challenges lie ahead. In this scientific struggle, technological innovations will be a strong ally. Newly available genotyping platforms now compete to provide greater data quality at higher throughput. Importantly, federally funded initiatives are currently in place to develop state-of-the-art whole genome sequencing technologies at relatively affordable cost.

Dr. Harold Varmus, Nobel Laureate and former director of the National Institutes of Health, wrote in a 2002 editorial in the New England Journal of Medicine that, “the publication last year of nearly complete sequences of the human genome in known as the HGP, which aimed at sequencing the complete genome of Homo sapiens. This large-scale biologic endeavor occurred over 13 years (1990–2003). If the “pregenomic era” was concluded by the complete sequencing of the 3.2 billion nucleotides of the human genome in 2003, we have already entered the “genome era.”

In closing, we bear in mind that the proposed transformation toward genomic medicine is not solely dependent on scientific discovery. What has to be clearly understood and acted on is the concern that the assumed impact of genomics and proteomics in clinical practice will not be attained without educating gastroenterology trainees, gastroenterologists and hepatologists, health care professionals, and our patients not only about the clinical applications but also the limitations and threats of using genetic information in medical practice. To this end, confidentiality and equality of genetic information, accuracy of genetic testing, prevention of genetic discrimination, and the psychological impact of knowing the genetic susceptibility to disease will continue to confront healthcare providers, patients and their relatives, and society alike.

Literature Review and Background

About 50 years ago, when James Watson and Francis Crick reported the discovery of the double helical structure of DNA, perhaps hardly any gastroenterologist paid attention to their article in the journal *Nature.* Since then, this seminal finding has transformed the biologic sciences. In the past 2 decades, this scientific breakthrough provided the cornerstone of an international scientific collaboration known as the HGP, which aimed at sequencing the complete genome of *Homo sapiens.* This large-scale biologic endeavor occurred over 13 years (1990–2003). If the “pregenomic era” was concluded by the complete sequencing of the 3.2 billion nucleotides of the human genome in 2003, we have already entered the “genome era.”

As stated provocatively by Francis Collins, Director of the National Human Genome Research Institute (NHGRI), “all disease—aside from most cases of trauma—have a genetic component.” Genetic and environmental contributions have been proposed in several GI and liver diseases for more than a century. To date, we have observed published work on the genetic contributions in colon cancer, hereditary hemochromatosis, IBD, and IBS, to mention a few. This trend will only continue to increase in the genome era, with implications for practicing or experimental gastroenterologists and hepatologists alike, regardless of our intellectual prejudicial approaches to understand the patho-
genesis of, diagnose, treat, and prevent digestive and hepatic disorders.

To successfully treat illness, it is imperative to first understand how a disease state is caused. Unraveling the pathogenesis of a disorder provides a means to intersect with disease processes and hopefully to alter the natural history of an illness toward cure or prevention. As we know it, there are 2 elements, whose interaction leads to many human diseases. The first is the inevitable environmental exposures/risks to which we are constantly exposed, even before birth. The second is the genetic predisposition we have inherited from our parents. These 2 components have to be dissected to shed light on disease pathogenesis if we wish to improve our current diagnostic methods and treatments. To evaluate the environmental component of an illness is challenging. To assess the genetic inclination to disease is a daunting but attainable goal, because the genetic material is generally considered steady and today can be tested in the laboratory. To this end, the genetic information derived from completion of the HGP will facilitate such efforts by thrusting forward the discipline of genomics and aiding the emergence of the field of proteomics. These 2 interrelated scientific subjects and other associated disciplines such as genetic epidemiology and bioinformatics will definitely advance basic studies and translational clinical investigations to elucidate the genetic predilection and environmental factors causing disease.

Recognizing the need to exemplify the mission of the above scientific fields, a few definitions are offered. Human genomics is a scientific field that examines the structure, function, and interaction of all the genes and genetic elements in the human genome.

Human proteomics is an emerging discipline that aspires to study the human proteome in health and disease using a variety of methodologies. Genetic epidemiology is the study of the role of genetic factors and their interaction with environmental elements in the occurrence of disease in human populations. Bioinformatics is the use of computer science and methods for the purpose of speeding up and enhancing biologic research.

Amid the current exhilarating scientific progress, we pose 2 questions. (1) Will genomics and proteomics have an impact on the manner that we will use to diagnose and treat GI and liver disorders in the future? (2) If so, when will this transformation occur? A short answer to the first question is "almost positively," but a reply to the second query is not easy. Most likely this revolution will steadily ensue in the forthcoming decades, although variations in lead-time are anticipated among different digestive and hepatic diseases.

In the following pages of this report, an effort is made to assess the application of genomic and proteomic technologies to the diagnosis and treatment of GI and liver diseases and their proposed impact on the clinical practice of physicians who treat such patients. Our given task is challenging. To this extent, we would like to make 3 statements concerning the structure and scope of this report.

First, the genomic era and the related scientific fields of human genomics and proteomics are currently at their developmental stage. A plethora of genomic data have been generated and analyzed as a result of the HGP. However, additional appealing scientific questions have risen from this effort. To inform the reader concerning the status quo of genomic and emerging proteomic science, the following pages provide a review of the facts regarding the human genome and proteome, including an overview of the contemporary scientific theories that attempt to explain human inheritance in health and disease.

Second, before exploring the possible applications and impact of genomic and proteomic technologies on gastroenterology and hepatology clinical practice, paradigms and updates of genomic discoveries in selected digestive and liver diseases are presented. This background knowledge is to serve as a framework on the discussion about the impact of genomic and proteomic technologies in the practice of our subspecialty.

Third, we believe that educating the human factor is a key element to successfully applying such unprecedented data on genetic information to better diagnose and treat those with digestive and hepatic disorders. To this direction, we discuss the need to educate health professionals concerning the application(s) and limitations of genetic information and relevant testing. An outline of the ethical/legal and social issues of genomic medicine that pertain to both the patient and the health care provider is also presented.

Finally, a basic glossary of genomics, proteomics, and related fields is available as a reference aid regarding the terminology used in this report (see Appendix 1).

The Human Genome

Simply put, the genome is the program that encodes life. This program, embedded in strings of nucleic acids known as DNA, takes shape in the double helix. This elegant configuration provides the means for accurate duplication of the DNA during each cell division and allows for vertical transmission of the heritable information required to perpetuate the species. While the DNA encoding the genome takes on an incredibly complex 3-dimensional structure when packed into the nucleus of a cell, we are able to understand and analyze this molecule as the sequence of the 4 nucleo bases (A [adenine], C [cytosine], G [guanine], and T [thymine]) that comprise it. Ascertainment of the approximate sequence of the 3.2 billion nucleic acid bases of the human genome was the extraordinary accomplishment
of the HGP, heralding the genome era. The knowledge of this “blueprint” of the human genome sequence provides a solid basis for the study of genes and genetic differences between individuals and will likely have a dramatic impact on the way we approach the study of both health and disease.

HGP: facts learned about the human genome. The sequence of the human genome has been determined by sequencing the genomic DNA from a handful of individuals. This sequence has been mapped to the specific 23 chromosomes comprising the human genome and thus provides man with an “instruction code” of the prototypical human. While this overall genomic structure is what distinguishes us from other forms of life, variation of this sequence differentiates us from each other. As a result of the HGP, it is now estimated that there are approximately 200,000 protein-coding genes in the human genome. These genes are comprised of coding regions (ie, exons) interspersed with noncoding regions (ie, introns) and flanked by enhancer, promoter, initiation, and termination elements. The presence of intervening introns within each gene allows for “alternative splicing” of exons to occur, which is a common mechanism to derive several protein isoforms from an individual gene. In fact, alternative splicing seems to be prevalent in humans, with lower estimates of approximately 35% of genes undergoing this phenomenon.

The protein-coding genes comprise <2% of the human genome; the remaining 98% has traditionally been thought of as “junk DNA” with no function. Interestingly, the sequence of many of these intergenic regions seems to be highly conserved, and as much as half of the genome sequence of many of these intergenic regions seems to be “junk DNA” with no function. Interestingly, the genome; the remaining 98% has traditionally been thought of as “junk DNA” with no function. Interestingly, the genome; the remaining 98% has traditionally been thought of as “junk DNA” with no function.

The sequence of the human genome has been determined by sequencing the genomic DNA from a handful of individuals. This sequence has been mapped to the specific 23 chromosomes comprising the human genome and thus provides man with an “instruction code” of the prototypical human. While this overall genomic structure is what distinguishes us from other forms of life, variation of this sequence differentiates us from each other. As a result of the HGP, it is now estimated that there are approximately 200,000 protein-coding genes in the human genome. These genes are comprised of coding regions (ie, exons) interspersed with noncoding regions (ie, introns) and flanked by enhancer, promoter, initiation, and termination elements. The presence of intervening introns within each gene allows for “alternative splicing” of exons to occur, which is a common mechanism to derive several protein isoforms from an individual gene. In fact, alternative splicing seems to be prevalent in humans, with lower estimates of approximately 35% of genes undergoing this phenomenon.

The protein-coding genes comprise <2% of the human genome; the remaining 98% has traditionally been thought of as “junk DNA” with no function. Interestingly, the sequence of many of these intergenic regions seems to be highly conserved, and as much as half of the genome consists of repeated sequences. These genomic repeats may play an important, yet undefined, structural and functional role. The genome is highly condensed in the cell and sequestered in the nucleus. This nuclear packaging is possible by close association of the DNA with various proteins like histones and allows for tight regulation of processes such as DNA replication and gene expression. Hence, the genome exists in vivo as a vibrant molecule constantly undergoing conformational change in response to internal and external stimuli, allowing the individual cell to react to its environment and perform its specialized functions. At the same time, the genome is a stable entity that is faithfully duplicated in the trillions of cells making up each individual and in transmitting the genetic information to the next generation. Thus, the genome acts as the program encoding life; within the structural “blueprint” of our genome are the instructions to build a human, and within the variation of its sequence lay the differences making each of us a unique individual.

Human genetic variation. Humans are a relatively diverse species, as is easily noticed when looking around most any street corner. Kruglyak and Nickerson eloquently summarized this observation by stating, “genetic variation is the spice of life.” The variation in facial features, body form, and certain mannerisms as well as the predilection toward development of disease are driven in part by the genomic sequence differences between individuals. This genetic variation has arisen through a combination of nucleotide substitutions, recombination events, and random matings that have occurred throughout the history of our species. The extent of genetic variation between any 2 unrelated individuals is estimated to be approximately 0.1%. Although this number may seem small and hardly capable of producing such diversity, it still accounts for millions of genetic sequence variants. Most of these variants take the form of single nucleotide polymorphisms (SNPs), in which the sequence differs at a single base pair. Other important variations in the genome sequence include microsatellites and insertions/deletions. A brief overview of these types of genetic variation is provided.

SNPs: Stated succinctly, a SNP is a specific position in the genome where alternate nucleotides can (and do) exist between 2 individuals or a population (Figure 1). For example, a SNP such as A/C (adenine or cytosine) is a position in the genome that can harbor one of 2 nucleotide alleles (A or C). The allele appearing more frequently in the population is known as the major allele and the less frequent is known as the minor allele. An arbitrary cutoff for minor allele frequency of 1% has been used to define SNPs, because this rate is the traditional definition of polymorphism. We would like to clarify that SNPs should not be viewed as point mutations. SNPs are simply normal variations of sequence and do not necessarily cause disease; in contrast, point mutations are directly capable of causing disease.
SNPs are stable, greatly abundant, and comprise the vast majority (~90%) of polymorphic loci in the human genome. It is estimated that more than 10^7 SNPs exist in the genome of the human population. SNPs are evenly distributed across the genome, generally dependent on the frequency of the minor allele. For example, we expect to observe one SNP with 1% minor allele frequency every 300 bases of genomic sequence and one SNP with 40% minor allele frequency only every 3300 bases. These estimates are based on the observed differences between 2 “haploid genomes” and extrapolated by the neutral theory of population genetics, which assumes a randomly mating population of constant size, with no selectivity of alleles. While this theory is simplistic and does not take into account population expansion or natural selection, more detailed treatments arrive at similar estimates, suggesting that a large percentage of the observed SNPs are rather innocuous in regard to selective pressures and perhaps functional significance. This finding is not surprising considering that 99% of SNPs are located in non–gene-coding regions of chromosomes.

The location of each SNP within the genome may determine its potential functional significance and relative risk on disease impact (Table 1). For instance, SNPs within the coding region of a gene (ie, coding SNPs) may result in an altered protein likely causing a significant impact on gene function and thus influencing the risk of disease. Coding SNPs can be further subclassified as synonymous and nonsynonymous, whether they alter or do not alter the specific amino acid of a protein. Of interest, the structural consequence of nonsynonymous coding SNPs can be predicted (ie, changed amino acid); therefore, these are best suited for association-based candidate gene studies (see discussion of linkage analysis in the following text). How ever, the prevalence of coding SNPs is extremely low, estimated at <0.1% of all SNPs (it is anticipated that there are about 50,000 coding SNPs in the human genome). Despite this fact, some have hypothesized that rare minor alleles may play an important role in disease, such that investigating only SNPs with minor allele prevalence ≥1% may miss important disease associations (see discussion of the common disease/common allele versus the common disease/rare allele hypotheses in the following text).

Somewhat more prevalent are SNPs that may potentially affect the expression or alternative splicing of genes. These include SNPs in the promoter region and exon/intron splice boundaries. Although potentially informative, we are not yet able to predict a priori if such SNPs will actually result in altered expression or splicing. While this category of SNPs is still useful for gene mapping studies, real-time polymerase chain reaction (PCR) and/or “isoform profiling” of messenger RNA (mRNA) may provide a more efficient approach for candidate gene studies involving altered expression. By far, the most prevalent SNPs are those located between genes (ie, intergenic regions). Based on calculations, there are more than 7 × 10^6 intergenic SNPs in the human genome. These SNPs may affect disease risk by interfering with enhancer functions or some other unknown mechanisms, but the likelihood of this is considered exceedingly rare. These SNPs are most useful for positional cloning studies to map the disease-causing gene or genetic variants (see discussion of linkage analysis in the following text).

**Table 1. SNP-Derived Variation and Risk to Phenotype**

<table>
<thead>
<tr>
<th>Type of SNP (from high to low risk to phenotype)</th>
<th>Location</th>
<th>Affect on function (from low to high frequency in genome)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsense</td>
<td>Exon, coding</td>
<td>Premature termination of protein sequence</td>
</tr>
<tr>
<td>Insertion/deletion</td>
<td>Exon, coding</td>
<td>Changes the frame of the protein coding region, can drastically alter the sequence</td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>Exon, coding</td>
<td>Changes an amino acid in the protein, can alter function</td>
</tr>
<tr>
<td>Synonymous</td>
<td>Exon, coding</td>
<td>Does not change protein, can alter splicing or expression level</td>
</tr>
<tr>
<td>Regulatory (expression, alternative splicing)</td>
<td>Promoter, 5’ or 3’, UTR intron near exon</td>
<td>Does not change protein, can alter splicing or expression level, timing, location</td>
</tr>
<tr>
<td>Intronic</td>
<td>Intron</td>
<td>Unknown, could alter splicing or stability</td>
</tr>
<tr>
<td>Intergenic</td>
<td>Between genes</td>
<td>Unknown, could affect expression</td>
</tr>
</tbody>
</table>

**Microsatellites.** Among the many repeat sequence motifs of the genome, microsatellites are the most well studied and have been successfully used as genetic markers for positional cloning of disease-associated genes. Microsatellites are brief strings of tandem dinucleotide, trinucleotide, or tetranucleotide repeats of DNA sequence (eg, CACACACACACACACACACACACA) (Figure 1). The number of these repetitive motifs in each microsatellite is variable, with multiple alleles generally detected in a population. There are estimated to be approximately 10^5 microsatellites in the human genome, spread evenly across chromosomes. Of those, only approximately 10^4 have been studied. While microsatellites generally do not directly affect gene function, they have been successfully used in gene mapping strategies. With the advent of PCR,
microsatellites have been used as genetic markers and typed across the entire human genome, creating whole genome scans that helped successfully identify chromosomal regions linked to specific diseases.

Insertions and deletions. Insertions and deletions are less frequent variants and account for <5% of human genome variation. Insertions take place when one or more nucleotides are introduced into a sequence of DNA (Figure 1). Deletions occur when one or more nucleotides are lost from the sequence (Figure 1). When located in the coding region of a gene, insertions and deletions are apt to have serious consequences on its structure (ie, frameshift, early termination), resulting in gross alteration of protein. A relevant paradigm of a frameshift insertion causing disease is the case of CARD15 as a susceptibility gene of Crohn’s disease (CD; see discussion of IBDs in the following text). Insertions and deletions within the regulatory region (ie, promoter) and intron/exon boundaries of a gene may also have an effect on protein expression and function.

Haplotypes of genetic variation. Humans have relatively limited genetic diversity, even though more than 10 million unique genetic variants may be present in our current population. This limited diversity relates to the young age (approximately 100,000 years) of our species and the fact that our genetic material has only been transmitted through a small number of generations (approximately 5000) from our ancestral origin. Genetic variation in the population is introduced via nucleotide substitution, meiotic recombination, and random mating. Recombination is a natural phenomenon that occurs in gametogenesis, during which regions between pairs of equivalent chromosomes are exchanged through the process of crossing over, generating discrete differences between the parental and offspring chromosomes. Due to recombination, ancestral and contemporary chromosomes will share domains of different length. To this end, alleles of adjacent loci on a chromosome have a tendency to stay together despite recombination; thus, the alleles are said to be in linkage disequilibrium (LD) (Figure 2). In contrast, as we humans go through many more generations and thus recombinations, we expect our ancestral and contemporary alleles to approach a state of equilibrium in which no linkage between ancestral alleles will exist.

The relatively limited genetic diversity of humans can simplify the study of the human genome and our quest to identify disease-causing genetic variants. For example, if we take a neighboring set of 5 SNPs located on the same chromosome, the number of possible combinations, or “haplotypes,” of these is 32 (5 SNPs of 2 alleles each; $2^5 = 32$). However, due to LD, assessment of these same 5 SNPs in a population will often show that only a few (ie, 3–5) of 32 expected haplotypes are present in a majority (~90%) of the population. Moreover, the pattern of LD across the genome is not uniform because recombination appears to occur more readily in certain regions known as recombination hot spots, thus creating regions of low LD (ie, high recombination frequency) interspersed between the regions of high LD (ie, low recombination frequency). These regions of high LD take the form of block-like structures and, as such, have aptly been termed “haplotype blocks” (Figure 3). The size and distribution and the allelic diversity of these blocks are highly variable between races and seemingly compatible with ancestral recombination throughout the genome.
the migratory history of our population. For instance, studies have estimated that the haplotype blocks found in populations of African ancestry are much shorter than in European, Japanese, and Chinese populations. This observation is consistent with the theory that human migration out of Africa resulted in “population bottlenecks” of limited founders, leading to reduced genetic diversity in these populations.

Initial efforts using SNPs to identify and assess the structure of human haplotype blocks have shown that while these regions may contain many SNPs, a few of these can be used to describe most of the genetic variation in the population. This limited allelic diversity found in haplotype blocks has the great potential of minimizing the number of SNPs required for comprehensive association studies of complex diseases, greatly facilitating the undertaking of such studies (see discussion of association studies in the following text). In fact, an international effort is underway to define SNP-based haplotype blocks in a number of populations, known as the Human Haplotype Map (HapMap) Project. The goals of this $100 plus million effort are to describe the common haplotype blocks found in the human race and its subpopulations and identify the informative “haplotype-tag SNPs” for discernment of these haplotypes (Figure 3). The early findings of this project have estimated that the average size of haplotype blocks ranges from 11 kilobases (kb) in African populations to 22 kb in Western European populations, with more than half of the genome existing in larger blocks of approximately 22 and 44 kb, respectively. Additionally, within these blocks, a small number of haplotypes (3–5) will typically describe more than 90% of all chromosomes in each particular population. Using these data, it has been estimated that an upper limit of 300,000 to 1 million (non-African/African) haplotype-tag SNPs will be required to perform high-powered, whole genome association studies for complex diseases. The number of SNPs required for such studies will also significantly decrease if disease is to be studied in populations that have undergone more recent population bottlenecks (such as isolated Norwegian or Icelandic populations). As haplotype blocks become larger due to greater extent of LD in a population, it becomes a greater task to pinpoint the disease-causing alleles via fine mapping studies.

Current progress and future direction. To date, more than 10 million SNPs have been reported in the dbSNP public database. These can be accessed at http://www.ncbi.nlm.nih.gov/SNP/. Of these 10 million SNPs, 5 million have been validated via various means and only 500,000 have any population frequency estimations (ie, percent of minor or major allele in a population). Although these numbers are impressive, the disparity between the number of validated SNPs and those with estimated population frequencies reflects the dichotomy between our ability to detect SNPs and assess them. Increased knowledge of the location, type, and extent of genomic variation in the human population will no doubt lead to effective approaches to further understand the causes and risk factors of disease. The recent past has seen exponential progress in capacity afforded to genotyping of SNPs, greatly reducing the cost of each genotype from $1.00–$2.00 to less than $0.10 using high-throughput approaches primarily amenable to disease-gene mapping studies. Additionally, customizable medium- to high-density SNP typing assays have entered the market, which are more applicable to candidate genes and “functional” (coding) SNP association studies.

Although our current efforts to identify and describe the variation found in the human genome will certainly be useful in the identification of genetic components involved with disease, the complexity of handling the data regarding individual variants is a challenge for many researchers and the future of genomics. To this end, whole genome resequencing comes as the ideal proposal to understanding how variation of the genome affects human health, because it would provide a global platform (not dependent on the selection of individual SNPs) for such studies. However, current costs prohibit this approach. In this regard, the NHGRI has launched an aggressive 15-year program to develop the necessary technologies to dramatically reduce these costs to $1000 per complete human genome sequenced. The success of such an effort will likely be the major factor behind advances in the study of human genetic variation and its association with disease in years to come.

Human Genomics and Proteomics

Genomics. Human genomics is the discipline that seeks to understand how the genome directs cellular activity through the expression of genes in response to internal (other genes) and external (environmental) stimuli to produce and maintain human life. From the earliest stages of life, the human genome is selectively activated, guiding the development and function of a multitude of cell types organized into higher-order structures such as tissues and organ systems. Understanding such processes requires a vast repository of knowledge not just regarding the genome but of all the potential mRNA gene transcripts (ie, the transcriptome) and the subsequent proteins (ie, the proteome) encoded within. Advancements in computer technologies to identify and map expressed sequences to the genome have provided us with a framework for understanding the transcriptome, and technologies are available to globally “profile” these expressed sequences. However, the function of many of these messages remains either not annotated or
unknown. Elucidation of the proteome is even farther behind due to the historical lack of high-throughput technologies for the study of proteins and our inability to predict form or function from sequence a priori.

Recent accomplishments, such as completion of the HGP and current efforts in this field, are quite impressive but amount to only the “tip of the iceberg”; indeed, we are in the infancy of a discipline with the potential to change medicine as we know it. Interpretation of the biologic function and meaning coded in the genome is a daunting task that may never be fully complete. However, the potential of this knowledge to society is overwhelming and must be explored. In addition to cataloging human genetic variation and its association with disease, a number of genomics “off-shoots” such as comparative genomics, functional genomics, and proteomics are being used.

**Comparative genomics.** Comparative genomics uses the knowledge gained from sequencing the genomes of many species to better understand the function of the human genome and the genomes of a wide variety of organisms that have a medical and an economic impact on human society. This bioinformatics approach generally uses powerful computers and elegant statistical algorithms to identify and compare conserved genomic regions representing evolutionary favorable genes and gene motifs involved with various cellular functions throughout the spectrum of species. To date, the complete genomes have been published for 200 bacteria, 20 archaea, and 17 eukaryotes (including 7 vertebrate species such as mouse, rat, and chimpanzee). Complete lists of genome sequencing projects and simple tools for genome viewing are freely available via the National Center for Biotechnology Information Web site (http://www.ncbi.nlm.nih.gov/ entertain/query.fcgi?db=genomeprj).

An example of one potential medical application of comparative genomics would be to search for conserved homologous regions in the genomes of pathogenic bacteria to identify potential target sites for powerful new antibiotics or conversely to identify pathogen-specific sequences for specific bacterial targeting. Genomic comparison also facilitates the functional classification of newly discovered or predicted human gene products through comparison with homologous regions in well-studied model systems. This bioinformatics-based, comparative subfield of genomics offers to greatly advance our efforts toward a fuller understanding of the human genome and offers numerous medical and socioeconomic rewards. To reach these ends, we must first expand our catalog of genome sequences, especially those of vertebrate mammals, because these are evolutionarily closest to humans. In this regard, the NHGRI currently supports a number of large-scale sequencing centers, which are working to add an additional 20 mammalian genomes to our growing collection. Additionally, technological advances in computing power and comparative algorithms as well as analytical innovation will need to keep pace with our ability to generate the vast quantities of data produced by these projects for the full potential of comparative genomics to be realized.

**Functional genomics.** Functional genomics seeks to identify, locate, and characterize all functional elements in the genome and further attempts to apply associations between such elements. The scope of this approach is to go beyond traditional protein-coding gene function, exploring the intergenic DNA sequences for potential function (such as regulation of structure, temporal control of expression, or other cell behavior). This said, functional genomics also includes the study of the relationships of these functional elements, including the flow of information through traditional, protein-encoded gene networks. Past efforts in this arena have often focused on the observation of cellular dynamics in response to experimental perturbation (ie, drug treatment, gene product modifications, transgenics, and so on) of model systems and have delivered the bulk of the biologic knowledge we know today. Global means to such approaches, including gene expression profiling via DNA arrays, have advanced recently due to the completion of the HGP and subsequent mapping of thousands of expressed sequence tags to the genome.

One example of functional nonprotein coding sequences that have been discovered is the noncoding RNAs. These are genes that are transcribed to RNA from the genome but are not translated into protein.23,24 Instead, they can act as gene regulatory elements through binding the genomic DNA in cis, effectively silencing expression.24 To date, more than 2500 of these noncoding RNAs have been identified.23 One well-studied mechanism involving noncoding RNAs is X-chromosome inactivation, the process of inactivation of one X chromosome in females to provide for dosage compensation of X-linked genes between the sexes.25

The NHGRI has recently launched 2 programs to facilitate the study of the functional genome and its networks: ENCODE (ENCyclopedia of DNA Elements) and “Molecular Libraries.” The ENCODE program seeks to perform an exhaustive determination of all functional elements in the human genome. The first effort of this program will be to characterize and improve the tools necessary for such exploration and to perform a pilot study to identify all functional elements in a 30 million-base pair region of the genome (approximately 1% of the human genome). The Molecular Libraries program aims to provide access to a library of 500,000 small molecules for use as chemical probes toward the study of molecular pathways. This program is expected...
to present a paradigm shift for academic investigators who to date have not had such a powerful research tool at their disposal. Both of these programs offer the potential of exciting new insights into the function and behavior of the human genome and the gene networks encoded within.

Although hundreds of thousands of scientific papers have been published in the past 50 years regarding the genomic phenomenon, our ability to make global connections between them is limited due to the sheer numbers, inconsistent gene nomenclature, and lack of consistent computer-accessible formats. Freely available bioinformatics technology with the ability to leverage this enormous resource would be a great boon to the study of functional genomics. A first step toward addressing these inconsistencies was the creation of the Human Genome Organization, which has been given the task of providing and approving unique gene symbols for all human genes. This process is being performed by the Human Gene Nomenclature Committee of the Human Genome Organization, which can be accessed at http://www.gene.ucl.ac.uk/nomenclature/. To date, the Human Gene Nomenclature Committee has provided unique gene symbols for more than 20,000 human genes. The Human Genome Organization has undertaken the creation of additional tools for the study of the genome, including a comprehensive annotation of human/mouse orthologues and a comprehensive annotation of all human genes; these can be accessed through links on the Human Gene Nomenclature Committee home page. An additional source for gene annotation is available through GeneCards sponsored by the Weizman Institute and accessed at http://bioinfo.weizmann.ac.il/cards/index.shtml. This site is freely available to academic users and provides information for all Human Genome Organization–approved genes automatically extracted from numerous databases.

Proteomics. The discipline of proteomics is simply the study of the proteome (ie, the sum of human proteins and how they interact with themselves and the environment). While considered by some a field of study separate from genomics, we (and others) believe this is not the case because proteins are but one “means” of the genomes “end” (albeit, the ends do not always justify the means). In our perspective, genomics seeks to understand how the genome directs cellular activity through the expression of genes in response to internal (other genes) and external (environmental) stimuli to produce and maintain life. Thus, logically, the agents of the genome (proteins) and their numerous potential interactions fall under the auspices of genomics. Arguments aside, proteomic investigation yields great potential for increased knowledge and is necessary to unlocking the secrets of the genome.

Although the genome sequence is stable, essentially the same in each cell of an organism at all times, the proteome is in constant flux as a reaction to its milieu. This produces an exponentially increased level of complexity to the proteome compared with the genome. Complicating these matters is the transcriptome, which is the sum of all expressed mRNA messages in a cell at a given time, which is also in constant flux mediated by the same environment. These mRNA messages are the templates from which individual proteins making up the proteome are translated and are thus intermediaries between the genome and the proteome. However, transcription of mRNA and subsequent translation into protein are controlled by numerous mechanisms, with the result that increased transcription of protein and, conversely, increased translation of a protein is not necessarily dependent on increased transcription of mRNA. Furthermore, the activity of each individual protein in the proteome is dependent on its subsequent 3-dimensional structure and often on its association with other protein molecules. Thus, increased presence of any individual protein does not necessarily indicate increased functionality of such protein. Because life has persisted for millions of years, eventually evolving to beings capable of pondering and elucidating these very mechanisms, we must assume that this occurs in a highly organized way and for a specific reason. It is in understanding this multitude of interactions that the secrets of the genome, proteome, and life itself will be unlocked.

The technology used to study protein is much different from that used to study DNA and RNA and to this point in time has not proven nearly as amenable to high-throughput approaches. Thus, the study of the proteome is far behind that of the genome. Although we have successfully sequenced the human genome (and many others), providing a solid basis on which to study structure variation and function, no such luxury is afforded the proteome. A first step toward better understanding the proteome will be to catalog the complete transcriptome, including all common alternative splice variations. Even though translation is not always dependent on transcription, this is a good initial effort toward understanding gene expression because the functionality of these messages is generally known (ie, provide a template for protein translation); thus, completion of such a project is foreseeable. This will provide a map of the potential protein products that we should expect to find. This map will allow us to assess the variation of protein expression, structure, and functions encoded in each gene and eventually ascribe associations and function to each form.

Many strides have been made of late to provide for greater throughput in the study of proteins. Innovations in mass spectroscopy have proved the greatest advance in the field, providing a means to distinguish specific proteins
without the need of specific ligands. Additionally, mass spectroscopy has the potential to identify and quantitate thousands of biomarkers at once (although this is a long way off and is dependent on the creation of advanced spectral algorithms and prior knowledge of potential spectra). Current progress with mass spectroscopy has had the greatest implications on its application to 2-dimensional gel electrophoresis in identifying proteins differentially expressed between 2 samples. However, the identification of the specific differential proteins is dependent on previous spectral knowledge of them. Moreover, protein microarrays are another emerging field of proteomic technologies with proposed broad applications for discovery and quantitative analysis of protein-based biomarkers in cancer.

As with the genome, standardization of proteomic data is necessary for useful future endeavors. To this regard, the Human Proteome Organization was established in 2001. The Proteomics Standards Initiative of the Human Proteome Organization has been tasked with the development and implementation of these standards. Included are standards for reporting molecular interactions, mass spectroscopy data, and creation of a general proteomics standard. The Proteomics Standards Initiative Web site can be accessed at http://psidev.sf.net. In addition to the standards initiative, members of the Human Proteome Organization have undertaken large-scale proteomics efforts to elucidate the proteomes of the liver, plasma, and brain. These projects are nearing the end of the pilot phase.

**Human Genomics in Disease**

For millennia, humans have recognized the role of heredity in disease. In 1865, Gregor Mendel introduced the concept that an element is passed down from the parent to offspring in an untouched form, causing an observable phenotype. However, it was not until 1953 that the chemical base of heredity was discovered and the double helix of DNA was elucidated. Yet, until the early 1980s, the discovery of a disease-causing gene was solely based on its functional characteristics. “Functional cloning” refers to identification of a gene causing a disease by knowing or postulating its biochemical function or defect, without reference to its structure and position in the genome (ie, mapping) (Figure 4). Functional cloning was used to clone the gene of sickle cell anemia. Nonetheless, for most human disorders, the biochemical origin or deficit, is unclear, and thus the process of disease pathogenesis remains elusive. To this end, in the mid-1980s, it was proposed that the gene of specific disease could be cloned and subsequently identified exclusively on the basis of its chromosomal position within the genome (ie, mapping). This revolutionary approach, termed positional cloning, assumes no functional information regarding the disease of interest and was used successfully for the first time in 1986 to clone the gene causing chronic granulomatous disease (Figure 4). In the genome era, however, a novel strategy called position candidate approach is expected to surpass the positional cloning method. This new method relies on positional cloning to identify a possible chromosomal area, followed by evaluation and screening of the candidate genes in the genomic area of interest using the publicly available, annotated maps of human genes.

In the following section, the fundamental concepts pertaining to the genomics of human disease are described.

**Mendelian and complex diseases.** Due to completion of the HGP, interest in translational genetic research has recently increased. There are 3 classes of genetic
disorders: Mendelian (ie, single gene), complex (ie, multifactorial), and chromosomal. Interestingly, the prevalence of each disease class changes from birth to adulthood (Figure 5). Although chromosomal disorders affect approximately 1% of live-born deliveries, we only discuss the concepts related to understanding the genetic contribution in Mendelian and complex diseases in this report due to space limitations. In fact, as illustrated in Figure 5, the complex diseases are the most prevalent among adults. As well, they include many of the disorders that gastroenterologists and hepatologists encounter daily in clinical practice.

Mendelian diseases display familial patterns of inheritance, including autosomal recessive, autosomal dominant, or X-linked transmission of the disease-related alleles. In the case of dominant inheritance, only a single trait-causing allele is necessary to express the disease phenotype. In contrast, an autosomal recessive phenotype requires both parental alleles to express the disease or trait. Generally, in a Mendelian disorder, the disease is caused by a few rare mutations of a single gene, although exceptions of more than one gene have been described. In a specific pedigree, the same mutation is accountable for causing the disease phenotype. However, among affected families, several or hundreds of different mutations of the same gene may be detected. Overall, Mendelian diseases are uncommon in the population. The most frequent, hereditary hemochromatosis, affects 1 in every 300 individuals. From a genetic standpoint, Mendelian diseases are considered simple because of the direct correspondence of a genotype to a phenotype (Figure 6). To date, more than 1000 Mendelian diseases have been elucidated; the catalog of human genes linked to these disorders is available online at Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/omim). Yet, additional modifier gene(s) may exist that likely affect the Mendelian disease “penetrance” (ie, the likelihood that a person with a specific genotype will express a certain phenotype) or “expressivity” (ie, the degree of phenotypic expression of a gene).

Complex diseases such as colon cancer, IBD, nonalcoholic fatty liver disease, and IBS are believed to have a multifactorial pathogenesis. It is generally accepted that complex diseases develop as a result of the interplay between several genes or genetic variants with environmental exposures, hence, the idiom “complex” (Figure 7). For the most part, complex diseases are caused by relatively common genetic variants in a number of genes, each of which has a small contribution on the disease trait or phenotype. As a result, the direct correspondence of a genotype to a phenotype that characterizes the Mendelian diseases does not exist in complex disorders. This fundamentally conceptual distinction may explain the observed heterogeneity of complex disease concerning clinical manifestations, progression, and response to treatment. Thus, although complex diseases have a genetic component, which in essence is different than in Mendelian diseases, they demonstrate familial aggregation. In fact, the risk of developing a complex disease among relatives of a proband is greater than the estimated risk of the disease in the general population. To this end, the term relative risk ratio of a sibling ($\lambda_s$) was coined to define the risk of a sibling presenting with a disease if a biological brother or sister is already affected. The $\lambda_s$ is calculated by dividing the prevalence of a complex disease among siblings by the prevalence of the disease in the general population. The higher the value of $\lambda_s$, the greater the evidence for a genetic component to a complex disease.

Mendelian and complex disorders operate in different aspects. Mendelian diseases are usually the result of a single gene with high penetrance of the genotype and low prevalence.
lence in the population. In contrast, complex diseases are caused by modest effects of several genetic variants with low penetrance of the genotype, of which (ie, diseases) may have high prevalence in the population. We believe the apparent epidemiologic and conceptual pathogenetic differences between Mendelian and complex diseases have important ramifications in the discovery, application, and overall impact of genomic and proteomic technologies on gastroenterology clinical practice.

The interplay of genes and environment. Environmental risks are important for any disease state but probably have a higher contribution in complex diseases. On the other hand, the genetic factors in complex diseases should be viewed as “susceptibility gene(s)” rather than “causative gene(s),” because such presence in an individual does not imply development of the disease. Dissection between genetic susceptibility and environmental risk(s) in complex diseases is the future challenge of genomics and proteomics. Recognizing the complexity of some Mendelian diseases such as cystic fibrosis, where hundreds of mutations could cause or affect the severity of the disease phenotype, someone could realize that identifying the susceptibility genes of complex disease will be a slow and challenging process. A putative model that describes the continuous interaction of genetic variants with the environment is shown in Figure 8. In this model, it is postulated that the unique susceptibility genotypes of 2 unrelated individuals (A and B) along with the environmental history (ie, diet, lifestyle, physical activity, smoking, and so on) could predict their future outcome between health and illness. To be in a position to predict whether individual A or B will develop disease, we have to assess both their genetic and environmental risk. To date, we can easily and accurately detect genotypes across the entire human genome. However, it remains extremely difficult to assess the past, present, and future environmental risks. Even if the latter were possible, we still lack robust methods to examine the interplay of genes and environment and how their interaction leads to disease. Nonetheless, this hypothetical model provides the stage to understand the scope and challenges we have to overcome as investigators and physicians before we can make significant progress to better diagnose and treat complex diseases in the genome era.

The common disease/common allele versus the common disease/rare allele hypotheses. At present, many logistic difficulties of genomic science stem from the theoretical challenges of associating the genetic variation of humans with the phenotype of complex diseases. To address this issue, 2 hypotheses have been postulated. The first, “the common disease/common allele hypothesis,” is proposed given the fact that the present human population of approximately 6 billion represents a global expansion from the single sub-Saharan African founding inhabitants of relatively small size (~10,000 people) that took place approximately 100,000 years ago. In this regard, it is expected that contemporary humans share a number of alleles with this small group of founders (Figure 2). As a result, the common disease/common allele hypothesis proposes that common alleles did exist before the global expansion and divergence of humans and contribute significantly to predisposition (ie, susceptibility) to common complex disease. Such alleles may confer moderate risk to complex disease and should occur at relatively high frequencies (ie, >1%) in the present human population. At this frequency of alleles, it is implied that association studies (see following text) using large patient cohorts will result in identification of the susceptibility alleles of common complex diseases. The fact that a limited number of common haplotypes account for the majority of haplotype blocks of the genome supports the optimism that association studies using tag-SNPs will identify common haplotypes predisposing to common complex disease. The common disease/common allele hypothesis was the basis to develop a genome-wide human haplotype map (ie, HapMap) that will describe the major haplotypes and tag-SNPs of the human genome.

On the other hand, the common disease/rare allele hypothesis proposes that most common complex diseases are caused not by common but rather rare alleles. This hypothesis proposes that more than 99% of genetic variants predisposing to common complex diseases arose following the global expansion and divergence of the human population. Additionally, it predicts that common complex diseases will demonstrate allelic heterogeneity (ie, different
disease-causing alleles at the same locus) and locus heterogeneity (ie, disease-causing alleles at separate loci), further complicating the discovery of genetic elements that cause disease. If the common disease/rare allele hypothesis is correct, then genome-wide association studies utilizing common alleles to interrogate a heterogeneous population would prove insufficient to identify the genetic variant leading to increased susceptibility to common complex diseases. To this extent, the ongoing construction of the human HapMap would be inadequate to define the variants of common complex diseases simply because it was developed based on common alleles.

While these 2 hypotheses are certainly at odds regarding the commonality of susceptibility alleles in common complex diseases, the heterogeneity of presentation and outcome of most such diseases has led investigators to believe that both rare and common genetic variants will prove to play an important role in the development of these diseases.

Identifying disease-causing genetic variants: study designs and strategies. Linkage analysis. This established method has proven successful to confine and identify genes causing Mendelian disorders.44 Linkage analysis was developed on the basis that alleles of disease-causing genes and genetic markers (which are evaluated for) if present on the same chromosomes should segregate simultaneously, meaning that they are physically linked. During meiotic recombination, however, chromosomes do not always remain intact. Crossing over between a pair of homologous chromosomes may lead to the separation of disease genes from genetic markers. Thus, the interlocus chromosomal distance is proportionally related to the probability of independent transmission of alleles (in other terms, the closer the alleles are located, the higher the chance they will segregate together because meiotic recombination between them becomes more rare).

Practically, linkage analysis searches for the cosegregation of polymorphic genetic markers, such as microsatellites among affected family members. In a family, the shared chromosomal region(s) of the affected members should contain the gene(s) causing the disease. When linkage between a disease and a genetic marker(s) is documented, then additional markers (ie, SNPs) within this genomic region of interest can be evaluated to more precisely map the location of the disease-causing gene.44 Linkage analysis has proven victorious for identifying genes of Mendelian diseases with high penetrance; nevertheless, its applicability to detect genetic variants of complex diseases has been unfulfilling.

Association studies. Association analysis is a case-control study design that seeks a statistical correlation between particular genetic variant(s) and a disease or trait.55 Large association studies (ie, >1000 participants) have greater statistical power than linkage methods to detect causative genetic variants of small effect on the disease phenotype of complex diseases.36 The genetic variants (more likely SNPs) could be located on genes (ie, candidate genes) or distributed throughout the genome and thus can be used in 2 separate approaches as described below.

The approach of direct association (Figure 9) seeks genetic variant(s) of plausible candidate genes and obeys the following procedure14,37: (1) selection of candidate genes that credibly may be involved in the pathogenesis of the disease of interest; (2) identification of the functional genetic variants with or in close proximity to coding regions of the candidate genes; (3) ascertainment of subjects, including careful phenotyping in cases and well matching of controls; (4) genotyping of chosen genetic variants in cases and controls; and (5) statistical analysis to determine whether significant association exists between the examined variants and the disease. Despite the statistical power of association studies, reproduction of published disease associations to genetic variants is uncommon. The reasons for this disparity include small study sample size, stratification biases, and disease locus heterogeneity (ie, presence of causal variants at different loci among patients).

The strategy of indirect association (Figure 9) is currently not possible. However, it will become achievable once the human HapMap38 is complete. This approach entails a genome-wide association study in which hundreds of thousands of tag-SNPs that cover the entire genome are genotyped in both patients and controls. Subsequently, LD analysis is applied to map the genomic region(s) that are indirectly associated with disease susceptibility genes or variants. This methodology is unprejudiced regarding specific genes or genomic regions.
Nevertheless, biases can still take place because of population stratification.39

Transmission disequilibrium test. To tackle the stratification bias of case-control association studies, statistical geneticists developed the transmission disequilibrium test.40 The transmission disequilibrium test is actually an association study based on a family design. The statistical method uses data on typed genetic markers derived from the probands and both parents. The transmission disequilibrium test evaluates the frequencies of parental alleles that are transmitted to their affected offspring compared with the frequencies of the alleles that are not transmitted. If a disease is associated with a high-risk allele (ie, causative variant), then it is expected that the frequency of this allele will be higher among the alleles transmitted compared with the non-transmitted alleles. Because the transmission disequilibrium test eliminates the concern of population stratification bias, it can be used to verify positive findings of case-control association studies.40

Genetics/Genomics in Digestive and Liver Diseases

To provide a stage for our discussion about the future impact of genomics and proteomics on gastroenterology and hepatology clinical practice, selected paradigms of Mendelian and complex digestive and liver diseases are presented.

Mendelian diseases. FAP. The prevalence of FAP is 2–3 cases per 100,000 individuals, and the disease phenotype is transmitted in an autosomal dominant fashion. The FAP syndrome includes related entities such as attenuated FAP, Gardner’s syndrome, and sometimes Turcot’s syndrome. At age 35 years, approximately 95% of patients with FAP have thousands of colon polyps. Without early diagnosis of FAP and subsequent therapeutic colectomy, colon cancer is expected by the mean age of 39 years.

The cause of FAP, the APC gene located on the long arm of chromosome 5 (5q21-q22), was discovered in 1991 using positional cloning approaches.41 Although <1% of colon cancers are due to FAP, discovery of the APC gene has shed light on understanding the pathogenesis of sporadic colon cancer. Indeed, we now know that more than 80% of all adenomas and cancers of the colonic mucosa exhibit early inactivation of the APC protein.42 APC is a tumor-suppressor gene, and loss of both alleles is necessary to lead to carcinogenesis. To date, more than 800 different germline APC mutations have been described and the vast majority (>90%) cause truncation of APC, resulting in loss of function. Interestingly, the location of the APC gene mutation can predict the number and extracolonic manifestations of the syndrome.43 Moreover, identical mutations among affected individuals could cause distinct FAP phenotypes, suggesting the presence of modifier gene(s).

In addition to APC gene sequencing and linkage analysis in affected families, protein truncation testing is an in vitro assay to assure that a proposed mutation causes FAP. Sequencing of the APC gene is more accurate than protein truncation testing. Of interest, up to 30% of newly diagnosed cases represent de novo APC mutations. Thus, failure to identify a known APC mutation does not exclude FAP. Although much has to be elucidated about the genotype-phenotype associations of APC and FAP, genetic testing in affected families has an impact on early disease diagnosis among relatives and can lead to prevention and treatment of colon cancer and the associated extracolonic manifestations of the syndrome.

Hereditary hemochromatosis. Hereditary hemochromatosis is the most common genetic disease in populations of European ancestry, with a prevalence of 1 in every 300 individuals. Hereditary hemochromatosis is transmitted as an autosomal recessive trait. This is an elusive disease because of the nonspecific nature of its symptoms, making diagnosis difficult at times. The discovery of the hemochromatosis gene (HFE) in 1996 using positional cloning techniques has been critical to understanding the pathogenesis and natural history of the disease and in devising diagnostic strategies.44 The HFE gene is located on the short arm of chromosome 6 and exhibits primarily 2 mutations (ie, C282Y and H63D) that can cause hereditary hemochromatosis. Indeed, C282Y homozygotes and H63D homozygotes account for ~80%–85% and ~1% of the affected individuals, respectively. Approximately 4% of hereditary hemochromatosis cases are due to C282Y/H63D compound heterozygotes. Nevertheless, in the past 5 years, additional rare non-HFE genes have been identified that can predispose to familial iron overload, including ferroportin 1,45 transferrin receptor 2,46 and hemojuvelin.47

Hereditary hemochromatosis was considered the prototype genetic disease for widespread population screening. First, the disease is common and preventable using simple phlebotomies once the diagnosis is made. Second, there is available genetic testing consisting of only 2 mutations. Despite these favorable facts, the initial enthusiasm of population screening was dissipated given the fact that significant numbers of C282Y homozygotes do not express the disease and the lack of evidence of liver disease progression in many C282Y homozygotes.48-49 Interestingly, collected data from 14 studies showed that about 50% of C282Y homozygotes do not exhibit iron
overload.\textsuperscript{50} Nevertheless, long-term observation of such patients indicated accumulation of iron, suggesting the need for follow-up of these individuals, including repeat testing for transferrin saturation. These cases of C282Y homozygosity with lack of penetrance imply the presence of modifier gene(s) that interact with HFE to produce the iron overload phenotype. Once hereditary hemochromatosis is diagnosed, however, genetic testing of the affected pedigrees is crucial. Siblings have the greatest chance of carrying the HFE gene compared with other first-degree relatives.

\textit{Wilson's disease}. Wilson's disease is an autosomal recessive disorder of copper metabolism with a prevalence of 1 in 100,000 individuals. The disease is caused by defective biliary excretion of copper, leading to its accumulation in the liver, brain, and cornea. The corollary of copper buildup in these organs is the eventual development of severe hepatic and neurologic disease. The gene for Wilson's disease, located on the long arm of chromosome 13 (13q14.3-q21.1), was identified by positional cloning approaches in 1993\textsuperscript{51,52} and codes for a copper transporting P-type adenosine triphosphatase (ATP7B). The ATP7B gene product delivers copper to the apoceruloplasmin and mediates the excretion of copper into bile.\textsuperscript{55}

More than 200 distinct genetic defects have been reported in patients with Wilson's disease. The inherited errors of the ATP7B gene include nonsense and missense mutations, insertions, and deletions. Some mutations are associated with more severe, early-onset disease. The His1069Gln missense mutation exists in 30\%–60\% of patients with Wilson's disease and of eastern, central, and northern European ancestry but is rarely seen in patients of non-European origin. The diagnosis of Wilson's disease is made by clinical examination, laboratory tests, and liver biopsy. Mutation analysis of the ATP7B gene is limited given the large number of mutations. To date, gene screening is important for suspected and unsuspected relatives of a proband when the causative ATP7B mutation of the index case is known. Otherwise, haplotype analysis around the disease locus can be used in an affected family to identify carriers of the defective gene. Therefore, the clinical impact of genetic testing is limited to affected pedigrees.

\textit{Complex diseases. IBDs}. CD and chronic ulcerative colitis (CUC) are chronic inflammatory diseases that involve the GI tract. The prevalence of CD and CUC is 10–100 cases and 35–100 cases per 100,000 individuals, respectively. A genetic epidemiologic approach to evaluate the heritability of a complex disease is a twin study. In such a design, the concordance rate of a disease is compared between monozygotic and dizygotic twins.

Given that monozygotic and dizygotic twins share 100\% and 50\% of their genetic material, respectively, if the disease of interest has a strong genetic element, a greater concordance rate of the disease is anticipated in monozygotic compared with dizygotic twins. In a twin study from Sweden, the index case concordance rate among monozygotic twins was 6.5\% for CUC and 58.3\% for CD\textsuperscript{54} indicating that heredity is stronger in CD than in CUC. In another study, IBD was found to aggregate in families.\textsuperscript{55}

Since the mid-1990s, several whole genome linkage studies have been performed to dissect the genomic regions causing susceptibility to IBD. The first genome-wide linkage study for IBD was reported in 1996, indicating 16q as the specific locus for CD (IBD1).\textsuperscript{56} Since the first report, 6 other whole genome studies have been performed and identified additional loci that fulfill criteria for significant linkage to IBD, including chromosomal regions 1p, 5q, 6p, 12q, 14q, and 19p.\textsuperscript{57}

CD represents one of the successful scientific endeavors to identify a susceptibility gene of a complex disease by pursuing fine mapping approaches of the IBD1 locus. These strategies led to the simultaneous discovery of the \textit{CARD15} gene (N-terminal caspase recruitment domain) as the first susceptibility gene of CD by 2 separate groups of investigators.\textsuperscript{19,58} Three common variants of \textit{CARD15}, namely, Arg702Trp, Gly908Arg, and a Leu1007fsinC, which truncate the CARD15 protein, are associated with CD. The \textit{CARD15} gene, also known as \textit{NOD2}, possesses apoptosis and nuclear factor \textit{kJ}B activation regions (CARD1 and CARD2), a nuclear binding domain, and a bacterial recognition region. The relative risk of developing CD having one susceptibility variant of \textit{CARD15} is 1.5–3.0. However, the relative risk sharply increases to 20–40 when an individual possesses 2 susceptibility variants of \textit{CARD15}. Nevertheless, the absolute risk of developing CD is about 3\% (in other terms, <1 in 25) even for homozygous individuals, supporting the concept that putative modifier gene(s) and environmental exposures do interact with the susceptibility variants of \textit{CARD15} to cause CD. To this end, healthy homozygous carriers of the \textit{CARD15} variants have been reported.\textsuperscript{59}

Interestingly, there are significant differences in the prevalence of \textit{CARD15} variants among patients of diverse ethnic origin with CD. The association of \textit{CARD15} variants with CD is strong in European, North American, and Australian patients but less prominent in Finnish, Irish, and Scottish cohorts.\textsuperscript{60} In contrast, patients with CD from Japan, Korea, and China lack any of the 3 reported \textit{CARD15} variants.\textsuperscript{60} These observations reflect the known genetic heterogeneity (ie, different variants
causing a similar phenotype) and locus heterogeneity (ie, variants at separate loci causing a comparable phenotype) that describe the pathogenesis of complex diseases.

The second whole genome scan in IBD revealed significant linkage within a domain of chromosome 12 (ie, IBD2 locus), which is stronger among CUC pedigrees compared with CD families. Moreover, a European study has dissected another IBD locus on chromosome 6p (IBD3), which confers susceptibility to both CD and CUC. In 1999, a US study identified the IBD5 locus on chromosome 5 (5q31-33) that is specific for CD and not CUC. Additional fine mapping of the IBD5 locus demonstrated a single, highly conserved haplotype of 250 kb in length that was associated with CD.

The above family-based whole genome studies in IBD have proven, as it was expected, that likely many loci contribute to CD or CUC. As we move forward, we will need to compile all these data and apply them in large, well-designed whole genome association studies before being able to better define and verify the several susceptibility variants of IBD. Only then can such observations have an impact on gastroenterology clinical practice.

_Helicobacter pylori_/peptic ulcer disease. _Helicobacter pylori_ is likely one of the most prevalent infectious agents that populate humans. Even in developed countries, 25%–50% of the population is infected. Nevertheless, only 10%–20% of _H pylori_-infected individuals develop GI diseases (ie, gastritis, gastroduodenal ulcer) and have an elevated risk of gastric cancer. Evidence suggests the likelihood of genetic susceptibility to _H pylori_ infection. For example, twin studies have shown that the concordance rate for _H pylori_ infection was 81% for monozygotic twins compared with 63% for dizygotic twins (P = .001). To start dissecting the genetic component(s) of _H pylori_ infection in humans, a genome-wide linkage analysis was performed in 143 infected Senegalese siblings. This study showed linkage on the long arm of chromosome 6, an area where the gene that encodes for chain 1 of the interferon gamma receptor (IFNGR1) is located. Subsequently, sequencing of IFNGR1 gene revealed 3 polymorphisms, [−56 C → T (promoter), H318P (exon 7), and L450P (exon 7)], that were associated with _H pylori_ infection. This study suggested that polymorphisms of interferon gamma play a role in predisposing to _H pylori_ infection in humans.

**Barrett’s esophagus and esophageal adenocarcinoma.** Barrett’s esophagus (BE) is characterized by replacement or metaplasia of the distal squamous esophageal mucosa with a columnar epithelium. When the latter is defined as specialized intestinal metaplasia, there is potential progression to esophageal adenocarcinoma through a metaplasia-dysplasia-carcinoma sequence.

In addition to environmental determinants, genetic variants are likely risk factors for developing BE. First, there are multiple case reports of familial aggregation of BE involving several generations of relatives. Second, reports of BE inheritance in sibling pairs suggest an autosomal recessive inheritance. Third, only a portion of patients with reflux symptoms develop BE. Fourth, the degree of reflux exposure and a prior diagnosis of reflux esophagitis do not predict the development of BE.

Interestingly, first-degree relatives of patients with BE have a higher prevalence (41%) of heartburn symptoms compared with the spouses (12%) of the latter who were used as controls. Moreover, reflux symptoms display greater concordance in monozygotic compared with dizygotic twins, with an estimated heritability of approximately 30%. Finally, in a large pedigree with severe pediatric reflux disease, which was inherited in an autosomal dominant fashion, investigators found a susceptibility locus on chromosome 13q14 using linkage analysis. However, it is not known whether these patients will develop BE in the future.

_Celiac disease._ Celiac sprue (CS) or gluten-sensitive enteropathy is characterized by inflammatory injury of the small intestine mucosa due to ingestion of cereal prolamins. Once believed to be rare in the United States, the prevalence of CS is now estimated to be 1 in 250 individuals. Genetic susceptibility to CS is strongly supported by family studies. For instance, the concordance rate for the disease is 75% for monozygotic compared with 11% for dizygotic twins. Moreover, the prevalence of CS among first-degree relatives of affected individuals is approximately 10%.

CS is strongly associated with HLA-DQ2 and HLA-DQ8. Despite this evidence, additional non-HLA loci may exert susceptibility to CS. Studies have estimated that the HLA-associated risk for developing CS is 20%–40%. To date, several genome-wide scans have been performed in CS-affected families. A study from Ireland showed strong linkage for CS with the HLA region and suggestive linkage in chromosomal regions 6p, 7q31, 11p11, 15q26, 19q, and 22cen. Two independent genome-wide analyses from Italy reported evidence for linkage in chromosome 5q. A study from the United Kingdom showed indicative evidence for linkage in chromosome domains 10q and 16q and less evidence for linkage in 6q, 11p, and 19q. Moreover, in Finnish families with CS, a whole genome scan found strong evidence for linkage in the HLA region at 6p21.3 and suggestive evidence for linkage in 6 other chromosomal locations: 1p36, 4p15, 5q31, 7q21, 9p21-23, and 16q12. CS is the prime paradigm of a GI disease in...
which the genetic and environmental (ie, gluten) components exist and interact in causing the disease phenotype.

IBS. IBS is perhaps the most common disorder that gastroenterologists encounter in clinical practice. IBS is characterized by recurrent abdominal discomfort or pain and is associated with diarrhea and/or constipation. Although pathogenetic abnormalities from visceral sensitivity to abnormal gut motility and putative risk factors such as psychological causes have been postulated, the molecular mechanisms underlying IBS remain elusive. Lately, genetic epidemiology approaches suggest that IBS may have a hereditary component.

Several epidemiologic studies indicate that IBS aggregates in families. Investigators have reported in a group of 100 consecutive outpatient IBS visits to the clinic that 33% of their patients had a family history of IBS compared with 2% of age-, sex-, and social class–matched controls. Moreover, in a population-based study it was reported that a first-degree relative with abdominal pain or bowel problems was associated with a proband stating IBS symptoms (odds ratio, 2.5; 95% confidence interval, 1.5–4.2). Additionally, in a twin study from Australia, the concordance of functional GI symptoms was 33% for monozygotic compared with 13% for dizygotic twins. In the largest twin study from the United States, the concordance rate for IBS was 17.2% for monozygotic twins compared with 8.4% for dizygotic twins.

Several candidate genes have also been examined in IBS using association studies. One investigated genetic variant is an insertion/deletion within the promoter of the serotonin transporter gene (5-HTT or SLC6A4). This polymorphism causes a short (S) and a long (L) transcript of 5-HTT, in which the former results in a positive modulator on the G-coupled protein. In a study of 276 patients with IBS and 120 controls, the α2c del322-325 and the α2a -1291C/G variants were associated with constipation-dominant IBS, with odds ratios of 2.48 (95% confidence interval, 0.98–6.28) and 1.66 (95% confidence interval, 0.94–2.92), respectively.

Beyond identifying susceptibility genetic variants, investigators studying IBS have started to perform pharmacogenomic studies to better understand individual response to treatment. One group genotyped 30 patients with symptoms of diarrhea-dominant IBS for the insertion/deletion within the promoter of the 5-HTT gene. Those patients were taking alosetron (ie, a 5-HT3 receptor antagonist) and were found to have a greater decline of colonic transit studies in LL homozygotes compared with LS heterozygote patients.

Technologies

Many technical breakthroughs were necessary to propel forward the completion of the HGP, the most important of which were the invention of capillary-based sequencing and the vast improvements to PCR. Before completion of the HGP, it was realized that new low-cost, high-throughput technological platforms for the assessment of nucleic acids would be necessary to catalog the genomic variation between individuals. This realization has spurred great innovation in the applicability of classic nucleic acid techniques to high-throughput approaches and inspired completely new techniques toward these ends. A plethora of platforms for both genotyping and expression profiling have entered the market in recent years, too many for complete review in this report. Most of these use the unique property of nucleic acid hybridization to achieve their goals, be it SNP genotyping or gene expression profiling, although other approaches such as mass spectroscopy have been taken. The ability to easily and rapidly reproduce nucleic acids with PCR and detect them using hybridization techniques has led to considerable throughput in these genomics-based technologies. The study of proteins, however, is not as efficient. Proteins cannot be reproduced as easily as DNA, and detection techniques have previously required specific ligands to be produced (quite laboriously) for each protein. These facts and the heavy focus on nucleic acids in the recent past (the HGP) have resulted in a considerable lag in the throughput capabilities of the proteomic technologies. However, recent strides in proteomic technology have been made and are a vast improvement over traditional biochemical approaches. Here we offer an overview of some of the technologies used for

November 2005

American Gastroenterological Association

1737
SNP genotyping, expression profiling, and proteomics. This is not meant to be an exhaustive list but provides some paradigms of the current technological platforms regarding methods, flexibility, throughput, and usefulness. Costs related to these technologies are prohibitive to the traditional single investigator wet bench academic laboratory, mainly due to the significant investment required to acquire the equipment. Thus, large institutions tend to offer access to these platforms through centrally administered core facilities.

**Genotyping platforms.** Here we provide an overview of select platforms for SNP genotyping. Included are companies providing laboratory equipment as well as those offering a genotyping service. An outline of these platforms is provided in Table 2.

**Pyrosequencing.** Pyrosequencing is a quantitative, sequencing-based genotyping platform developed and marketed by Biotage (http://www.pyrosequencing.com). This technology effectively sequences short lengths of DNA via detection of light emissions from a luciferase-catalyzed reaction to quantitatively detect pyrophosphate release during incorporation of individually applied deoxynucleoside triphosphates. Although this assay can be performed on 96 templates at a time, parallel analysis of numerous SNPs is impractical. Thus, this platform is not suitable for approaches requiring large numbers of SNPs to be typed, such as whole genome association. In addition, the low-throughput results in a relatively high per-SNP cost. However, the pyrosequencing technology provides for quantitation of allele frequency and presents SNPs in the context of the surrounding gene sequence, features that are not common to high-throughput platforms, which generally offer an “allele call” for each SNP. The quantitative feature of pyrosequencing lends itself well to the quantification of gene copies and extent of methylation studies, while the sequence component provides assay quality control and allows for the typing of consecutive or closely based SNPs. Thus, pyrosequencing is an important tool when high quality or unique applications are needed. It is an excellent choice for the typing of a small number of SNPs.

**Sequenom.** Sequenom (http://www.sequenom.com) uses a matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectroscopy platform known as MassARRAY. SNPs are genotyped using the MassEXTEND assay, a label-free primer extension assay coupled with specific extension terminators, to produce DNA fragments of differing size depending on the allele of the assayed SNP and discriminated based on mass. This assay is designed for a 384-well format, allowing for a number of assays to be prepared simultaneously. The MassARRAY instrument can also be used for SNP discovery and quantitative gene expression. Mass spectroscopy offers an alternative to the fluorescent-based technologies used to genotype SNPs.

**Affymetrix DNA chips/ParAllele platform.** Affymetrix (http://www.affymetrix.com), the company primarily known for their gene expression profiling products, also offers genotyping arrays based on the GeneChip platform (which is described in detail in the following text). Currently, Affymetrix offers 2 gene-mapping assays, allowing for the simultaneous analysis of 10,000 and 100,000 SNPs per sample, respectively. These assays overcome a primary bottleneck of large-scale SNP genotyping, the need for locus-specific SNP primers and/or probes, by using a generic single primer to generate the fragments necessary for allelic discrimination. This is accomplished through restriction digestion of the target genomic DNA with consequent ligation of adapters to the DNA ends, PCR is performed using an adapter-specific primer, and products are size fragmented before hybridization with the chip. Although this approach greatly simplifies the preparation of targets for allelic discrimination, it is restricted by an inability to custom design assays. To address these shortcomings, Affymetrix has partnered with ParAllele (http://www.parallelebio.com) to provide standard and customizable assays for use with the Affymetrix system. ParAllele uses molecular inversion probe technology, which is basically an oligo ligation assay with a unimolecular mode of action allowing for amplification of targeted SNPs by PCR with common primers and analyzed by universal sequence tag.

### Table 2. Examples of Genotyping Platforms

<table>
<thead>
<tr>
<th>Platform</th>
<th>SNP determination</th>
<th>Suitable applications</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyrosequencing</td>
<td>Sequencing based</td>
<td>Small SNP sets, candidate SNPs, special situations</td>
<td>Equipment only</td>
</tr>
<tr>
<td>Sequenom</td>
<td>Mass spectroscopy</td>
<td>Small SNP sets, candidate SNPs</td>
<td>Equipment only</td>
</tr>
<tr>
<td>Affymetrix/ParAllele</td>
<td>Hybridization to GeneChip</td>
<td>Larger SNP sets, candidate genes, fine mapping, whole genome mapping</td>
<td>Equipment only</td>
</tr>
<tr>
<td>Illumina</td>
<td>Hybridization to beads</td>
<td>Larger SNP sets, candidate genes, fine mapping, whole genome mapping</td>
<td>Equipment and service</td>
</tr>
<tr>
<td>Perlegen</td>
<td>Hybridization to GeneChip</td>
<td>Whole genome mapping</td>
<td>Service only</td>
</tr>
</tbody>
</table>

1738 AMERICAN GASTROENTEROLOGICAL ASSOCIATION GASTROENTEROLOGY Vol. 129, No. 5
DNA arrays. ParAllele currently offers a single-tube 10,000 coding SNP assay and the MegAllele Genotyping 10K eSNP kit in addition to their user-customizable 3K, 5K, and 10K SNP assays. With such products, ParAllele is an excellent choice for functional SNP candidate gene studies.

Illumina. Illumina (http://www.illumina.com) is a “full-service” genotyping company offering both a genotyping service and a commercially available equipment platform based on their proprietary BeadArray technology. This technology uses oligo-coated beads that self-assemble into microwells chemically etched into an array substrate, resulting in a randomly ordered DNA array. This array undergoes a decoding process providing a map used in the downstream analysis of the data. The end result is a highly parallel, customizable assay capable of simultaneously interrogating tens of thousands of SNPs on each sample. The strength of this technology is its flexibility, because it provides a practical platform to affordably assess hundreds to hundreds of thousands of SNPs for each sample. As such, it is a suitable platform for fine mapping and candidate gene studies as well as whole genome association studies. Although incredibly low per-SNP costs can be achieved with this technology, the high cost of the equipment limits its practicality to core genotyping facilities and large genotyping programs. Additionally, the setup of large genotyping panels is only cost effective if many hundreds to thousands of samples will be typed. Some predesigned assays have been announced, including a whole genome association “bead chip” capable of simultaneously assessing 100,000 highly informative SNPs on each sample.

Perlegen. Perlegen (http://www.perlegen.com) is a service-only genotyping company that primarily services pharmaceutical companies and large well-funded genotyping efforts. They use proprietary assays developed on the Affymetrix gene chip platform that are not publicly available, with the current capability to assess 1.5 million SNPs per sample. This company’s services are directly applied toward whole genome association studies and are not readily suitable to candidate gene approaches.

Expression profiling platforms. Affymetrix GeneChips. Affymetrix (http://www.affymetrix.com), maker of the GeneChip, is the leader in gene expression profiling technologies; to date, more than 3000 scientific papers have been published on the use or review of this platform. The process they use to manufacture their chips was adapted from the semiconductor industry and uses photolithography coupled with solid-phase chemistry to “build” oligonucleotides directly onto the chips, providing an extremely accurate, high-density array of oligonucleotide probes. Gene expression profiling with the Affymetrix GeneChip is conceptually simple. Labeled complementary RNA targets are prepared from the mRNA of the experimental samples and hybridized to the oligonucleotides on the GeneChip. The level of fluorescence associated with each location of the GeneChip is assessed using a laser-scanning fluorescent confocal microscope scanner (sold by Affymetrix) and directly reflects the amount of mRNA in the sample. Affymetrix manufactures GeneChips to assess gene expression for more than 20 species, both plant and animal. The Human Genome U133 Plus 2.0 Array is the current flagship of the Affymetrix product line. This GeneChip allows for the simultaneous interrogation of more than 47,000 transcripts selected from the GenBank, dbEST, and RefSeq databases, representing the complete known transcriptome and including many expressed sequences with unknown function. Affymetrix also offers a Human Genome Focus Array based on the U133 chip, representing more than 8500 verified genes as a tool for investigators with limited resources. Additionally, Affymetrix offers an array customization program for those investigators or companies wishing to produce arrays with a focused content, such as pathway-specific or cell-specific genes.

Spotted arrays. Spotted arrays are produced by depositing PCR-produced complementary DNA (cDNA) or long (generally 40–100 base pairs) oligonucleotides onto glass slides for use in gene expression analysis. While potentially more economical than GeneChips when produced en masse, the production of these arrays does require specialized equipment and expertise with the handling of large cDNA or oligonucleotide collections. Quality control of these collections is of utmost importance, because misidentification or contamination can have drastic effects on experimental results. The use of cDNA on spotted arrays is substantially labor intensive due to the necessity of PCR amplification, purification, and verification for thousands of products. Oligonucleotides can be readily produced or procured commercially, somewhat reducing labor. Shorter oligos have less potential for cross-hybridization than longer oligos or cDNAs but tend to give weaker signals. Probe set selection can be a daunting task for the individual, especially if a broad assessment of the transcriptome is desired, because not only do thousands of sequences need to be identified but potential cross-hybridization also needs to be minimized. Array-to-array reproducibility has been traditionally hard to control, but improvements in the equipment to spot the arrays has improved this; however, spotted arrays are not often used to compare a large number of samples. Instead, spotted arrays are generally used to analyze a mixture of 2 differentially
labeled samples, providing a ratio of expression between the 2 samples. There are a few companies offering pre-designed and custom-made cDNA spotted arrays, such as Agilent (http://www.agilent.com) and others offering spotters and scanners, including Affymetrix. The potential pitfalls of spotted arrays coupled with the wide availability and affordable customization of GeneChips preclude spotted arrays as a practical approach to global expression profiling. However, the potential for spotted arrays to be much more sensitive (detect smaller amounts of RNA in a sample) than GeneChips may prove to have future applicability to the study of subtle variation in expression and genes regularly expressed at low levels.

**Proteomics platforms.** Two-dimensional gel electrophoresis. Two-dimensional gel electrophoresis is a powerful method for the separation of proteins by both charge (first dimension) and mass (second dimension) from complex mixtures such as tissues, culture media, or biologic fluids. Once stained, the product of 2-dimensional gel electrophoresis is a 2-dimensional matrix of dots representing different proteins that can be isolated and characterized by mass spectroscopy. Its primary applications are in the global assessment of the proteome of a sample and in the identification of differentially expressed proteins between samples. Proteome assessment involves the isolation and characterization of all protein dots for a single sample. This is a potentially powerful approach but is reliant on staining techniques that can be quite insensitive and will miss many proteins that are at low concentration and thus does not provide a complete proteome assessment. Identification of differentially expressed proteins is accomplished by (1) comparing spot intensities between 2 samples using dyes that stain proteins in a quantitative manner or (2) comparing the matrix for variation in layout (ie, dots present in one sample, missing in another). A number of publicly available databases of 2-dimensional gel results can be found on the Internet, the most complete of which is SWISS-2D PAGE, found at http://us.expasy.org/ch2d/.

**Mass spectroscopy.** Mass spectroscopy is a tool allowing for accurate and sensitive determination of molecular weight and structural information from biomolecules such as proteins and nucleic acids. A mass spectrometer consists of 3 parts: the ionization source, the mass analyzer, and the detection system. Most common mass spectrometers today use electrospray or matrix-assisted laser desorption ion sources because these are the most amenable to the analysis of large biomolecules such as proteins. A number of different mass analyzers are available to separate the molecules based on their mass-to-charge ratio, the most common of which are quadrupoles, TOF, magnetic sectors (MS), and ion traps. These analyzers vary on their detectable mass-to-charge range and compatibility with different ionization methods. Tandem mass spectrometers using multiple analyzers such as quadrupoles-TOF, MS-MS, and TOF-TOF are widely available and generally used where additional ion fragmentation is necessary, such as in the generation of protein structural information. A photomultiplier, electron multiplier, or microchannel plate detector is used, based on the type of detector, to monitor the ion current and transmit the signal (mass spectra) to the computer. The most widely applied proteomic application of mass spectroscopy is the identification of proteins recovered from 2-dimensional gel electrophoresis via peptide mass fingerprinting. Briefly, the recovered proteins are subjected to cleavage with specific proteases to produce products of varying mass and subjected to mass spectroscopy. The resulting masses represent a fingerprint of the protein, based on the sizes of the fragments generated by the protease cleavage. These fingerprints are compared with measured and/or calculated protein fingerprints from a number of publicly available databases to identify the isolated protein. The Swiss Institute of Bioinformatics offers many tools and links for protein analysis via their Expert Proteome Analysis System proteomics server at http://us.expasy.org/. Another commonly used protein sequence database mining tool, Protein Prospector, is offered by the University of California San Francisco and can be accessed at http://prospector.ucsf.edu/.

**Implications for Gastroenterology Clinical Practice**

Genomics and proteomics promise unparalleled advancements in our knowledge of human health and disease. With time, this accumulated information re-

![Figure 10](http://us.expasy.org/ch2d/)

**Figure 10.** The proposed impact of genomics in medicine. Knowing the genetic contribution to disease will facilitate prevention, diagnosis, and novel therapies. Modified and reprinted with permission from Collins and McKusick (the AMA does not hold copyright to this article. It was written by an employee of the United States Federal Government and thus is in the public domain.)
Regarding the structure and function of the human genome and proteome will alter our approach to the diagnosis and treatment of disease and may dramatically change the practice of medicine (Figure 10). Advances in genetic testing for disease risk and early disease detection as well as pharmaceutical responsiveness will ultimately increase life expectancy and quality of life for many individuals while decreasing the overall cost of health care. Down the road, the lessons learned from the HGP will result in better therapies for all diseases, whether through new pharmaceuticals modified specifically to the individual(s) or direct manipulation of the genome to achieve the desired result.

In the previous sections of this report, we have described concepts, facts, and methods regarding human genomics and the emerging field of proteomics. We believe this introduction will facilitate understanding of the potential applications, current limitations, and anticipated impact of these scientific fields and related technologies in the clinical practice of GI and liver diseases. At the present, despite ongoing enthusiasm and great expectations, the clinical application of genomics and proteomics in digestive and hepatic diseases is limited to Mendelian diseases such as hereditary hemochromatosis, adenomatous polyposis coli, and other inherited polyposis syndromes. There are many challenges to overcome before these technologies will be integrated into the practice of medicine. From our perspective, there are 3 key components that are essential to move genomic and proteomic research forward, including implementation in clinical practice: (1) development and execution of translational clinical studies on GI diseases of interest, (2) continued technological innovation, and (3) sufficient funding. These important issues are discussed in the following paragraphs of this section.

Mainstreaming the sciences of genomics and proteomics into clinical practice will require the development and execution of translational clinical trials in GI diseases of interest. First, we must identify and prioritize the GI diseases in which application of genomics and proteomics will be most beneficial, taking into account a number of issues such as (1) burden of the GI disease in the population, (2) GI diseases that currently lack efficient therapies, (3) likely impact on GI practice, and (4) model GI disease with strong genetic component (ie, CD, CUC, primary sclerosing cholangitis [PSC]). Second, we need to develop translational studies in partnership with the National Institute of Diabetes and Digestive and Kidney Diseases for research in genomic and proteomic applications focused on these prioritized GI diseases and relevant projects.

Common complex GI diseases such as IBS or sporadic colon cancer have high prevalence in the population; thus, genomic/proteomic-based discoveries for these disorders will have a strong impact on clinical practice. However, recognizing that biologic phenomena such as penetrance, variable expressivity, genetic and locus heterogeneity, imprinting, and epigenetics cause a high noise-to-signal ratio, we have to accept that it will be difficult to dissect the causative genetic and environmental contribution of common complex diseases. To address these issues, the current generation of physician-investigators will have to develop large (ie, >1000 study participants) research resources of well-phenotyped patients linked to biospecimens (ie, genomic DNA, serum, plasma, tissue of interest) to apply the tools of genomics and dissect the genetic variants predisposing to GI and liver diseases. The overarching goal of such an effort is ambitious and will require many resources, including the possibility of collaboration among medical centers for less common complex diseases. To this extent, we may also have to focus on rare complex GI diseases with a strong genetic component as models because dissection of the susceptibility genetic variants in these is expected to be less challenging than that of common complex GI diseases. Subsequently, the application of findings derived from rare complex GI diseases can be significant in both. For example, in a family of an affected proband, the relative risk of a sibling to develop sporadic colon cancer, CUC, or PSC is 2- to 3-fold, 10- to 20-fold, and ~100-fold, respectively. Because of their higher relative risk, understanding the genetic contribution to PSC or CUC (although not an easy task) will be relatively less demanding compared with dissecting the genetic susceptibility to sporadic colon cancer. Nonetheless, unraveling the genetic predisposition of PSC or CUC could shed light on sporadic colon cancer pathogenesis given the fact that both of the former (ie, CUC and PSC) are independent risk factors for development of colon carcinoma.

Presently, technological innovation is moving rapidly and clinically applicable genomic-based approaches are on the horizon. Technology today provides the first step toward the goal of “personalized medicine” with Food and Drug Administration approval of the AmpliChip CYP450 array (Roche Molecular Diagnostics) in December 2004. Based on the Affymetrix GeneChip platform, the CYP450 array performs a comprehensive analysis of 2 genes of the CYP450 gene family, namely CYP2D6 (29 alleles) and CYP2C19 (2 alleles). These 2 genes are likely responsible for the metabolism of approximately 25% of medications currently on the market, including a number of antidepressants, β-blockers, proton pump inhibitors, and antipsychotics. Genetic
variation of these genes determines how well the relevant drugs are metabolized and hence may provide physicians with a useful tool in choosing the optimum medication and/or dose. Variable response to pharmaceutical treatment remains a significant hurdle in medical practice. Adverse drug reactions are the cause of 2 million hospitalizations and 100,000 deaths annually, at a cost of $100 billion dollars each year.106 The ability to predict such adverse reactions and correctly identify optimal therapies thus has the potential to save human lives and the economy billions of dollars per year in medical costs.

The AmpliChip CYP450 array offers a “predictive phenotype” for each of the 2 genes based on the combination of alleles present in the individual. For CYP2C19, there are 2 such classifications: poor metabolizer (ie, no enzyme activity) and normal metabolizer (ie, normal enzyme activity). There are 4 classifications for CYP2D6: poor (ie, no enzyme activity), intermediate (ie, reduced enzyme activity), extensive (ie, normal enzyme activity), and ultrarapid (ie, higher than normal enzyme activity) metabolizer. Those with poor metabolic capacity for certain drugs are most at risk for developing complications due to the toxic buildup of relevant medications, whereas drugs may not reach beneficial levels when prescribed at recommended doses in the ultrarapid metabolizer. Currently, the high cost of the test, the lack of knowledge to interpret it by physicians, and the necessity for sophisticated laboratory setup to perform the testing ensure that it will be some time before the AmpliChip CYP450 array is widely used in medical practice. However, the potential exists. For instance, genotyping patients with IBS for CYP2D6 may be important to recommend dosage compensation for nortriptyline or amitriptyline so that the benefit of treatment is maximal.

Generous funding is critical for the research enterprise. The federal government and private sector have supported and are committed to continue promoting genomic and proteomic technologies because of anticipated clinical applications. In fact, the main goal of the NHGRI is to sponsor and advance genome-oriented research. For instance, the NHGRI has recently committed more than $38 million in grants to spur development of innovative technologies designed to dramatically reduce the cost of whole genome sequencing, aimed at broadening the applications of genomic information in medical research and health care. The strategic goal of the NHGRI is to initially reduce the cost of sequencing an entire mammalian-sized genome to $100,000. It is expected that such technologies will reach commercial availability 5 years from now, enabling researchers to sequence the genomes of people as part of studies to identify genes that contribute to common complex diseases such as colon cancer and diabetes. Ultimately, the vision of the NHGRI is to decrease the price of whole genome sequencing to $1000 or less, which would allow the sequencing of individual genomes as part of medical care.

The ability to sequence each human’s genome can cost-effectively give rise to more individualized strategies for diagnosing, treating, and preventing disease. To this extent, the AGA, in collaboration with the National Institute of Diabetes and Digestive and Kidney Diseases, should delineate the current and future needs of research direction and initiatives for digestive diseases and subsequently sponsor translational patient-oriented studies poised to leverage these promising technologies as they mature in the near future. Beyond federal funding, however, the private sector has also invested in genomic and proteomic technologies and initiatives. In fact, many of the technological innovations have come from private research and development companies. More importantly, genotyping companies such as Illumina and Perlegen have been actively involved in and are currently performing genomic projects in collaboration with major academic medical centers in a variety of diseases.

From a theoretical perspective, new technologies (ie, microarrays, sequencing) will shed light on the abnormal molecular pathways and cellular events that occur in digestive and liver diseases. Presently, it is difficult to accurately predict the future impact of genomics/pro-

---

**Figure 11.** The putative clinical impact of genomics and proteomics over the next 25 years. Genomics and proteomics are currently in a discovery and research phase, during which time there will be little direct impact on clinical practice. Nevertheless, expected improvements in genomic technology (ie, P450 array, whole genome sequence cost at $100,000 or $1000 per person) and data analysis, coupled with the knowledge gained in the initial period, will result in an exponential increase of clinical impact; use of genetic testing and designer drugs are proposed to be introduced into medical practice in 2015 and 2020, respectively.
teomics in diagnosing and treating GI and liver diseases. Nevertheless, it has been proposed that by 2015 genetic testing for predisposition to selected complex diseases will become available in clinical practice 107 and by 2020 predicting drug responsiveness will be standard of care and gene-based designer drugs will be introduced to medical practice (Figure 11).107 We expect genomic and proteomic technologies to positively affect Mendelian and complex digestive and hepatic diseases alike. However, as stated earlier, the clinical and economic impact will be greater for the latter because of the higher overall prevalence in the general population.

Among complex GI and liver diseases, CD is one that can be used as a paradigm to envision how genomics could be implemented into clinical practice. After a decade of research effort, the first susceptibility gene of CD is a reality.19,58 We now know that mutations of the CARD15 gene are responsible for a small part of the risk (<5%) of developing CD in the Western hemisphere. It appears that the penetrance of CARD15 susceptibility variants, even in homozygotes, is at best 10%. Indeed, there are reports that carriers of the CARD15 susceptibility alleles have no CD.59 Collectively, these findings not only illustrate the complexity of CD pathogenesis but also indicate the need for additional studies to unravel the yet unknown CD susceptibility gene(s), the putative modifier gene(s), and the environmental elements that are critical to the pathogenesis of this disorder. Hence, this case exemplifies the various challenges that we face before implementing the knowledge of genomics into clinical practice.

The importance of identifying the susceptibility genes of complex diseases relates not only to diagnosis but also to better understanding of the pathogenesis, clinical presentation, and natural history of disease. For instance, we now begin to appreciate that patients who carry the CARD15 susceptibility alleles are diagnosed with CD at a younger age, usually present with ileitis, and have a tendency to develop strictures more frequently during the course of disease. For a chronic disease like CD, identifying patients who will develop complications such as intestinal strictures or colon cancer has important clinical implications. To this end, we need to develop translational studies to better define the genotype/phenotype associations of CD.

Beyond complex diseases, however, better understanding of genomics of Mendelian digestive and hepatic diseases is also important to shed light on yet unknown molecular and cellular pathways of relevant diseases. For instance, we now know that more than one gene causes hereditary hemochromatosis, and this fact has significantly improved our knowledge of iron absorption and cellular processing. In another example, the discovery of LDL receptor, as the result of studying an extremely rare disease (ie, homozygote hypercholesteremia),108 led to the development of statins, which have revolutionized the medical treatment of hyperlipidemias. This is a prime paradigm of an exceptionally rare Mendelian disease becoming the basis for prevention of coronary artery disease, a common complex disorder. Indeed, the estimated target population for statins in the United States alone is 25%–50% of adults older than 45 years of age.109

Clearly, there are many challenges ahead before we can apply the plethora of genomic and proteomic technologies and science to patients with digestive and hepatic diseases.

Implications for Education and Training of Gastroenterologists

In the coming years, discovery of genetic variants that influence predisposition to disease will likely have an impact on the diagnosis and therapy of several Mendelian and complex digestive and hepatic diseases. In the appropriate setting, use of tests to screen for genetic variants will become an important component of clinical practice in the future (Figure 11). At that time, medical practitioners will need to correctly interpret the test findings and thoroughly understand the implications of such testing to effectively diagnose and treat the patient and to prevent illness. To date, most physicians have no formal training in medical genetics; therefore, they have difficulty with the interpretation of genetic test findings and lack counseling skills. In fact, a study to evaluate the clinical use of APC gene testing reported that genetic testing results were misinterpreted by physicians in approximately 30% of patients with FAP.110 This finding becomes important in patients at risk for highly penetrant diseases, where incorrect interpretation of results may lead to false-negative diagnosis for an otherwise vastly preventable disease.

Previous sections of this report focused on scientific advances in genomics and proteomics. Nevertheless, we need to remember that established approaches to medical practice, such as a gathering and comprehensive analysis of the family history, will remain important toward screening and treatment of patients with digestive and hepatic diseases and their families. Instructions on obtaining the family history have traditionally been focused on Mendelian diseases. However, in the near future,
family history will be applied more often in the context of complex disorders. As we gain better understanding of the genetic and environmental components of these diseases, family information will become increasingly essential. Health care professionals need to retain the skill of collecting and assessing the family history in addition to understanding the basics of genetics before new genetic testing becomes a widely used tool for the diagnosis, treatment, and prevention of complex disease.

Beyond addressing the genetic susceptibility to complex diseases, new clinical tests will likely focus on assessing individual variation to drug response. Recently, the US Food and Drug Administration approved the first DNA microarray for genotyping 2 genes (ie, CYP2D6 and CYP2C19) of the cytochrome P450 system (see previous text). Variants of these 2 genes can predict poor to ultrarapid metabolizers of medications before even prescribing them. The use of this test could affect decision making on a variety of prescribed medications by defining the optimal treatment dose and/or preventing dangerous drug side effects and interactions.

As a result of all these exciting discoveries, medical students to practicing gastroenterologists alike will be asked to be able to interpret genetic test results and to communicate these findings to patients and relatives, colleagues, and referring physicians. Realizing these educational needs, medical genetics is now incorporated in the curriculum of medical schools. However, to the best of our knowledge, there is no current medical genetics curriculum during internal medicine residency or GI fellowship training programs. In fact, an Internal Medicine Residency Training Program Genetics Curriculum Committee was recently formed and a curriculum outline is now published. Nevertheless, to date, there are no plans on educating the practicing gastroenterologist on genomics and proteomics should be directly related to clinical practice. An appreciation of the basics of genetics will be required, along with ability to interpret genetic tests and to effectively counsel patients and their relatives. These learning objectives can be achieved through self-directed reading, continuing medical education courses, attendance of professional meetings (ie, American Association for the Study of Liver Diseases, AGA), and special courses.

At the beginning of the 21st century, medicine enters a new era. Despite progress in human genomics and the emergence of proteomics, the practicing physician has and will have a pivotal role in assessing and dissecting the phenotype of disease and in applying the technological advances to provide the best possible care to the patient. Postulating that these laboratory technologies will begin to enter clinical practice in the next decade(s), now is the time to begin addressing the needs to educate and train current and future gastroenterologists in the fundamentals of medical genetics and the clinical relevance of genomic medicine. To this end, the AGA is in a unique position to initiate, guide, and supervise such an action that will greatly benefit each of its members and patients suffering from digestive and liver diseases.

**Suggested Key Research Questions**

Suggested key research questions are as follows.

- What is the current and expected significance of genomic and proteomic technologies in gastroenterology clinical practice?
- Are the available genomic and proteomic technologies a research tool and/or currently in use for clinical applications?
- Which element(s) will affect the future clinical implementation and overall impact of genomic and proteomic technologies?
- Which are the current challenges of genomic and proteomic technologies; how can we overcome them?
- What will influence the clinical usefulness of the AmpliChip CYP450 array and other similar technologies in years to come?
Ethical, Legal, and Social Issues

The study of genomics and proteomics offers unprecedented potential toward the understanding and assessment of disease, especially in the identification of those at increased risk of contracting illness, in turn allowing us to tailor our surveillance and treatment methods to the individual and resulting in a health care system that is markedly more efficient and effective than that of today. However, with these advancements come great social and economic implications with the potential of slowing or perhaps limiting the successful implementation of these medically beneficial insights. Although a complete treatise on these ethical, legal, and social issues is outside of the scope of this report, we think it is necessary to briefly address a few key topics.

The main concerns of the public regarding genetic testing are currently focused on 2 issues: insurability and employment. A first step in addressing the societal concerns regarding genetic information and insurability was put forth in the Health Insurance Portability and Accountability Act of 1996. This legislation provided the first federally mandated protections against genetic discrimination in health insurance, most importantly by stating that genetic information in the absence of a clinical diagnosis of illness could not be considered a preexisting condition. However, the Health Insurance Portability and Accountability Act does not prohibit insurance providers from charging higher rates to individuals based on genetic information and does not prohibit insurers from requiring applicants to undergo genetic testing. Furthermore, the Health Insurance Portability and Accountability Act places no restrictions on the disclosure of genetic information to insurance companies, a fact that lends credence to the public’s concerns. For example, individuals potentially benefiting from genetic screening for disease risk may be reluctant to take advantage of this screening for fear that their already-high insurance rates will be increased. Additionally, those identified as at risk through genetic screening may be reluctant to receive adequate follow-up disease surveillance during periods when not insured for fears of clinical diagnosis and thus a “preexisting” condition. The concerns regarding insurability will become exponentially more important as genomic and proteomic knowledge advances.

The second primary concern of the public regarding genetic information is whether employers will be permitted to consider genetic information in making personnel decisions and therefore unjustifiably rule out individuals from employment for reasons that are not related to their ability to perform the tasks of their job but for motive of future economic liability to the organization. Such concern may be warranted, because some companies have adopted discriminatory policies toward certain at-risk segments of the population such as smokers or the obese, who are not protected under the Civil Rights Act of 1964 or the Americans with Disabilities Act of 1990. With the increasing costs of health care, one can imagine that a business focused on the “bottom line” may well see a genetic prescreen as a way to reduce these costs. In turn, these potentially considerable cost savings could be passed on to their employees such that they might not mind such an intrusive policy and willingly become participants in de facto genetic discrimination. While the employers’ role in genetic discrimination is not now a major factor, it may become so as our knowledge of genomics and proteomics grows.

An important legal aspect of genomic and proteomic study relates to the patenting of newly discovered genes and proteins. These natural products have been easily patentable following isolation, purification, and characterization, although recent acceptance guidelines necessitate that some level of feasible utility for use of the gene or protein be shown. To date, there have been more than 3 million genome-related patent applications filed. The patenting of genes or proteins is a strategy to reward scientists for their novel discoveries and to protect scientific inventions without secrecy. While this has been beneficial, providing the means of existence for countless biotechnology companies and contributing significantly to the economy, the current accumulation of patented genes may begin to discourage new investigators, who may consider that all of the “good” genes are “taken.” In realistic terms, genes and proteins have and will be patented. While there are arguments both for and against the practice, biologic patents are now intertwined with the economy; there is no turning back.

Translational clinical studies are necessary to prove the medical applicability of genomic and proteomic technologies. To this extent, investigators have stored human biologic samples for decades. Such collections can be obtained for diagnostic and research purposes and could aid our knowledge regarding diseases in humans. Current technologies in biology provide effective approaches to use such biospecimen resources for research, diagnostic, and therapeutic purposes. Nevertheless, ethical issues have been raised about these sample repositories and related technologies. Important issues regarding the storage of human specimens include consent of individuals who provided such materials, methods and duration of stored biologic materials, identification of the material and its source, material linkage to medical data, transfer and exchange of biomedicai material, and transformation of genetic material (ie, immortal cell lines) for future use. It is important that researchers should pursue their scientific goals without com-
promising the rights of the human subjects involved in such research. To this direction, institutional review boards, along with investigators and research sponsors, must exercise great care and social sensitivity in applying the professional guidelines and governmental regulations to protect the subjects of research whose biologic samples are used for research purposes.

The ethical, legal, and social issues pertaining to biomedical research evolve as science, medicine, and society move forward and past practices change or redefine. In the genome era, topics related to privacy of the genetic information, fairness in the use of genetic information, the psychological impact of knowing the genetic predisposition to disease, accuracy of genetic testing, and education of physicians and the public about the capacities and limitations of genetic information will continue to challenge health professionals, patients, and society alike.

### Appendix 1. Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele</td>
<td>One or more alternate forms of a gene at a specific locus</td>
</tr>
<tr>
<td>Alternative splicing</td>
<td>The mechanism by which variation in the use of a gene's exons leads to the production of multiple mRNA species, or isoforms, potentially coding for functionally different proteins</td>
</tr>
<tr>
<td>Association study</td>
<td>A case-control study design that searches for a statistical correlation between particular genetic variant(s) and a disease or disease trait</td>
</tr>
<tr>
<td>Autosomal dominant</td>
<td>A pattern of inheritance in which one dominant allele is capable of exerting its phenotypic effect, regardless of the other allele</td>
</tr>
<tr>
<td>Autosomal recessive</td>
<td>A pattern of inheritance in which 2 copies of the same allele are necessary to exert a phenotypic effect</td>
</tr>
<tr>
<td>Autosomes</td>
<td>Any chromosome except for the sex chromosomes (X and Y)</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td>The use of computer science and methods for the purpose of speeding up and enhancing biologic research</td>
</tr>
<tr>
<td>Candidate gene</td>
<td>A gene potentially harboring genetic variants unique to or more prevalent in those with a particular disease</td>
</tr>
<tr>
<td>Chromatin</td>
<td>The observable (via staining) material in the interphase nuclei, composed of genomic DNA, chromosomal proteins, and some RNA</td>
</tr>
<tr>
<td>Codon</td>
<td>A sequence of 3 DNA or RNA nucleotides coding for a certain amino acid or providing a signal to terminate translation of a protein (ie, stop codon)</td>
</tr>
<tr>
<td>Comparative genomics</td>
<td>Uses the knowledge gained through genomic study of other species to infer the likely function of unstudied genes in humans</td>
</tr>
<tr>
<td>Deletion</td>
<td>Loss of one or more nucleotides from a DNA or RNA sequence</td>
</tr>
<tr>
<td>Dizygotic twins</td>
<td>Twins resulting from the fertilization of separate eggs by separate sperm and thus sharing the same amount of genetic material as nontwin siblings (ie, 50%)</td>
</tr>
<tr>
<td>Epigenetic</td>
<td>A nonmutational phenomenon that alters the phenotype without alteration of genotype (ie, promoter methylation or histone modification)</td>
</tr>
<tr>
<td>ESTs</td>
<td>Expressed sequence tags—short stretches of DNA sequence expressed at the RNA level and detected by means that are not gene specific; ESTs are primarily used to identify new genes and with the completion of the HGP are being mapped to chromosomes</td>
</tr>
<tr>
<td>Euchromatin</td>
<td>The genetically active, gene-rich part of the chromatin</td>
</tr>
<tr>
<td>Expressivity</td>
<td>The degree of phenotypic expression of a gene</td>
</tr>
<tr>
<td>Exon</td>
<td>A segment of DNA that is present in fully processed mRNA</td>
</tr>
<tr>
<td>Gene expression profiling</td>
<td>The large-scale quantitative assessment of gene expression at the mRNA level</td>
</tr>
<tr>
<td>Genetic epidemiology</td>
<td>The study of genetic factors and their interaction with environmental factors in the distribution and determination of disease in the population</td>
</tr>
<tr>
<td>Genotype</td>
<td>The genetic makeup of an individual</td>
</tr>
<tr>
<td>Haploid genome</td>
<td>Genome with only a single chromosome complement, such as are found in gametes</td>
</tr>
<tr>
<td>Haplotypic</td>
<td>A combination of alleles at 2 or more closely linked genetic loci</td>
</tr>
<tr>
<td>Haplotype block</td>
<td>Block-like structures of the genome that are in LD with each other and can be described by a few of the many potential haplotypes</td>
</tr>
<tr>
<td>Heredity</td>
<td>The passage of genetic material from parents to offspring, resulting in the inheritance of similar traits</td>
</tr>
<tr>
<td>Heterochromatin</td>
<td>The genetically inactive, gene-poor regions of the chromatin composed of repetitive DNA sequences</td>
</tr>
<tr>
<td>Heterozygous</td>
<td>Having 2 different alleles at a specific gene locus</td>
</tr>
<tr>
<td>Homozygous</td>
<td>Having 2 identical alleles at a specific gene locus</td>
</tr>
<tr>
<td>Imprinting</td>
<td>Differential expression of a gene or chromosome segment dependent on parental origin, often through epigenetic means</td>
</tr>
<tr>
<td>Insertion</td>
<td>Addition of one or more nucleotides to a DNA or RNA sequence</td>
</tr>
<tr>
<td>Intergenic region(s)</td>
<td>The regions of the genome located between genes</td>
</tr>
<tr>
<td>Intron</td>
<td>A nontranscribed region of a gene that does not code for a protein</td>
</tr>
</tbody>
</table>
## Appendix 1. (continued)

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linkage</td>
<td>The tendency of genetic loci on the same chromosome to be inherited together as a consequence of their physical proximity</td>
</tr>
<tr>
<td>Linkage analysis</td>
<td>A method to trace and measure the cosegregation of a disease or disease traits in a family using marker loci</td>
</tr>
<tr>
<td>Linkage disequilibrium</td>
<td>Particular alleles at 2 or more adjacent loci that occur together more frequently than expected based on their individual frequencies; often termed “allelic association”</td>
</tr>
<tr>
<td>Loci</td>
<td>Plural of locus—the physical location of a gene.</td>
</tr>
<tr>
<td>Major allele</td>
<td>The most prevalent allele for a given locus in a population</td>
</tr>
<tr>
<td>Microsatellites</td>
<td>Small string of tandemly repeated DNA sequence, commonly of 1–4 base pairs; represent approximately $10^5$ uniformly distributed loci in the human genome</td>
</tr>
<tr>
<td>Minor allele</td>
<td>The least prevalent allele for a given locus in a population</td>
</tr>
<tr>
<td>Modifier gene</td>
<td>A genetic factor that modifies the expression or phenotypic consequence of expressing another gene</td>
</tr>
<tr>
<td>Monozygotic twins</td>
<td>Twins resulting from the fertilization of one egg with one sperm, thus sharing 100% of their genetic material</td>
</tr>
<tr>
<td>Mutation</td>
<td>A permanent, potentially heritable modification in the sequence of genomic DNA</td>
</tr>
<tr>
<td>Mutation, frameshift</td>
<td>Mutation caused by deletion or insertion of a nucleotide(s) resulting in an altered open-reading frame of a gene and usually to a truncated protein</td>
</tr>
<tr>
<td>Mutation, missense</td>
<td>Mutation resulting in the incorporation of an incorrect amino acid into a protein</td>
</tr>
<tr>
<td>Mutation, nonsense</td>
<td>Mutation resulting in a stop codon that causes truncation of a protein</td>
</tr>
<tr>
<td>Mutation, point</td>
<td>Mutation of a single nucleic acid, capable of directly causing disease; often confused with SNPs</td>
</tr>
<tr>
<td>Mutation, silent</td>
<td>Mutation in coding sequence that causes no change in the amino acid sequence of a protein</td>
</tr>
<tr>
<td>Nucleotide substitution</td>
<td>Change of a single nucleotide due to slight lack of fidelity in the DNA replication process; is one evolutionary process producing additional diversity in the species</td>
</tr>
<tr>
<td>Oligonucleotide</td>
<td>A short (20–100 base pairs) nucleic acid molecule generally used as a primer or probe through its ability to hybridize with complementary nucleic acid sequences</td>
</tr>
<tr>
<td>Penetrance</td>
<td>The likelihood that a person with a specific genotype will express a certain phenotype</td>
</tr>
<tr>
<td>Phenotype</td>
<td>The observable physical, biochemical, or physiologic features resulting from the expression of one or more genes</td>
</tr>
<tr>
<td>Pharmacogenomics</td>
<td>The study of genetic and environmental factors contributing to the success or failure of pharmaceutical intervention</td>
</tr>
<tr>
<td>Pleiotropy</td>
<td>Multiple, seemingly unrelated and variably expressed phenotypes stemming from the expression of a single gene</td>
</tr>
<tr>
<td>Polymorphism</td>
<td>The existence of 2 or more common alleles at a specific locus</td>
</tr>
<tr>
<td>Population bottleneck</td>
<td>Decreased genetic variation among members of a population expanded from relatively few founders</td>
</tr>
<tr>
<td>Positional cloning</td>
<td>Identification of specific disease genes exclusively based on their chromosomal position within the genome</td>
</tr>
<tr>
<td>Proband</td>
<td>The individual through which a pedigree is discovered and explored</td>
</tr>
<tr>
<td>Recombination</td>
<td>The process during meiosis in which regions between pairs of equivalent chromosomes are exchanged through the process of crossing over; this process results in offspring that are biologically different from their parents and thus provides genetic variation in a population</td>
</tr>
<tr>
<td>Recombination hot spot</td>
<td>An area of the chromosome where recombination takes place more often than expected, resulting in low LD and driving the creation of haplotype blocks</td>
</tr>
<tr>
<td>Relative risk ratio</td>
<td>The risk of an “exposed” population to develop a disease compared with an “unexposed” population; exposure not only includes environmental exposures but also familial, genotypic, or allelic exposure—this ratio is calculated by dividing the prevalence of disease in the exposed population by the prevalence in the unexposed population; the greater the number, the greater the risk of developing the disease in the exposed population</td>
</tr>
<tr>
<td>Relative risk ratio of a sibling ($\lambda_s$)</td>
<td>The risk of a person to develop a disease if his or her biological sibling is affected. The $\lambda_s$ is calculated by dividing the prevalence of disease among siblings with the prevalence of the disease in the general population</td>
</tr>
<tr>
<td>Sibling (sib)-pair analysis</td>
<td>Linkage analysis in which genetic markers are tested for linkage to a disease or phenotypic trait by measuring the extent to which affected sibling pairs share the marker haplotypes</td>
</tr>
<tr>
<td>SNPs</td>
<td>Single nucleotide polymorphisms—a specific position in the genome where alternate nucleotides can (and do) exist between 2 individuals or a population; a minimum minor allele frequency of 1% is often invoked to further define an SNP</td>
</tr>
<tr>
<td>SNP, coding</td>
<td>An SNP located in a gene coding exon</td>
</tr>
<tr>
<td>SNP, nonsynonymous</td>
<td>A coding SNP that changes the amino acid sequence</td>
</tr>
<tr>
<td>SNP, synonymous</td>
<td>A coding SNP that does not change the amino acid sequence</td>
</tr>
</tbody>
</table>
Appendix 2. Selected Online Resources

Genomic projects/riages

National Center for Biotechnology Information—home page
Human Genome Organization—home page
http://www.gene.ucl.ac.uk/hugo/
Human Gene Nomenclature Committee—home page
http://www.gene.ucl.ac.uk/nomenclature/
SNP Consortium—home page
http://snp.cshl.org/
International HapMap Project—home page
http://www.hapmap.org/
NHGRI—home page
http://www.genome.gov/

Genomic databases/tools

Online Mendelian Inheritance in Man—catalog of human genetic disorders
Entrez Genome Project—comprehensive database of genome mapping projects
Basic Local Alignment Search Tool—nucleic acid and protein alignment tool
dbSNP—comprehensive database of SNPs
dbEST—comprehensive database of expressed sequence tags
dbSTS—database of sequence and mapping data for sequence tagged sites

Human Genome Browser—tools for searching and viewing the human genome
http://www.genome.ucsc.edu/
CHIP Bioinformation Tools—tools for accessing SNPs and gene ontology
http://snpper.chip.org/
GeneCards—database of human genes
http://bioinfo.weizmann.ac.il/cards/index.shtml/
Gene Ontology Web site—database and search tool for gene function and annotation
http://www.geneontology.org/

Proteomics sites

SWISS-2D PAGE
http://us.expasy.org/ch2d/
Expert Proteome Analysis System
http://us.expasy.org/
UCSF Protein Prospector
http://protein.ucsf.edu/
Proteomics Standard Initiative of the Human Proteome Organization
http://psidev.sf.net/
Human Proteome Organization
http://www.hupo.org/

Technology

Agilent Technologies
http://www.agilent.com
Illumina
http://www.illumina.com
Perlegen
http://www.perlegen.com
Affymetrix
http://www.affymetrix.com
Sequenom
http://www.sequenom.com
ParAllele
http://www.parallelebio.com
Biotage
http://www.pyrosequencing.com

References


60. Ahmad T, Tammoli CP, Jewell D, Colombel JF. Clinical relevance of advances in genetics and pharmacogenomics of IBD. *Gastroenterology* 2004;126:1533–1549.


Address requests for reprints to: Chair, Future Trends Committee, AGA National Office, c/o Membership Department, 4930 Del Ray Avenue, Bethesda, Maryland 20814. Fax: (301) 654-5920.

This report was prepared by Dr. Lazaridis under the direction of the AGA Future Trends Committee. It was approved by the committee on May 15, 2005.

Members of the AGA Future Trends Committee include Nicholas F. LaRusso (chair), Juan R. Malagelada, Walter J. McDonald, Pankaj J. Pasricha, Suzanne Rose, and Michael Lee Weinstein.