As researchers have explored the environmental and inherited causes of common diseases, they have often amassed clinical and laboratory data collected from people with common complex disorders. Many have also collected biologic material, including DNA. These resources represent an essential component for ferreting out genes relevant to disease with the use of the genome-wide association study. This method entails the matching of a given human genome sequence with an annotated, high-resolution map of common genetic variation; it benefits from a large collection of DNA samples obtained from a population whose clinical characteristics are well defined, as well as cost-effective genotyping and sophisticated statistical analysis. With all these components now in place, genome-wide association studies are contributing to our understanding of diseases to which there is a genetic predisposition. Early successes include the identification of genes associated with age-related macular degeneration, myocardial infarction, and abnormal cardiac repolarization intervals. A more recent success is the identification of four loci associated with type 2 diabetes; one of these loci contains a gene (TCF7L2) that had been previously implicated in this disease, and a second locus includes the SLC30A8 gene that is expressed only in insulin-producing beta cells.\(^1\)

The human genome consists of approximately 3 billion nucleotides of DNA sequence, most of which have now been identified in their linear arrangement on chromosomes. The genome contains several million individual DNA-sequence variants (or alleles), defined as differences in sequence at identical sites on homologous chromosomes. Many of these alleles are common, and many encode the functional differences underlying protein variants, such as the \(\varepsilon2\), \(\varepsilon3\), and \(\varepsilon4\) variants of apolipoprotein E, which confer different levels of risk for Alzheimer’s disease. Less common variations that contribute to important human single-gene disorders are also determined by nucleotide variation, such as the adenine-thymidine change underlying the alternative amino acid sequences (glutamate to valine) that differentiate the normal and sickle forms of \(\beta\)-hemoglobin. The plotting of these sequence differences on the human genetic map and

---

**Figure 1. Mapping the Relationships among SNPs.**

Each small circle above the magnified chromosome (labeled 5′ to 3′) represents one SNP with its two allelic possibilities. At the intersection between any two of these SNPs, the associations between their variants are shown in various shades from white to red, with the deepest red indicating the strongest association. Patterns of triangular blocks of strong association are separated by short nodes with very little association. One SNP (called a tagging SNP) represented above a deepest-red block — block 1 (3′) or block 2 (8′) — can serve as a surrogate for any variant within its block. Testing for one SNP might provide almost complete genetic information for that block.
the identification of the most common form of variation, the single-nucleotide polymorphism (SNP), have created a dense web of SNPs across the genome at intervals of about one SNP per 300 base pairs of DNA.

The mechanisms of human inheritance are such that if SNPs are in close physical proximity (typically, 50 kb apart or closer) they are more likely than those that lie farther apart to have alleles that travel together in a block when passed from parent to child (see Figure 1). This phenomenon, termed linkage disequilibrium, allows one SNP variant in such a block to serve as a surrogate for the presence of other SNP variants, obviating the need for direct (and expensive) individual testing for the presence of every SNP. Most SNPs reside in fairly well defined blocks, which are separated by nodes containing very few SNPs. The recently completed International HapMap Project delineated the architecture of this arrangement and made it possible to choose a few select SNPs (called “tagging” SNPs) in each block whose presence implies the presence of other SNPs (and any unknown embedded disease-related mutation). This arrangement of genetic variants in a block on a single chromosome is called a haplotype (see Figure 2).

Any new mutation that arises within a block region travels along with other members of the block for hundreds of generations, surrounded by many coinherited SNP alleles. The statistical reliability of the association between a given disease-susceptibility mutation and particular nearby SNPs depends on how many generations have elapsed since that mutation originated. If one can capture a mutation at the point at which it has become relatively common in a population but remains genetically connected to its block, one can use the identification of any member of that block as a surrogate for the presence of the causative mutation in that block.

Researchers engaged in many large studies, such as the Framingham Heart Study, have collected thousands of DNA samples from subjects, along with an array of data on their clinical and laboratory status. Technological developments now permit high-throughput testing of the several hundred thousand individual sequence variants necessary to provide adequate coverage of all the DNA blocks in humans to ensure that if a variant associated with disease is present, it will be found. Coupling the genotypic data with epidemiologic data that include many covariates, one is theoretically able to identify genes or gene–environment interactions that predispose to both normal trait variation and disease processes.

As with any new method, it is important not to overstate what genome-wide association studies can do. First, the populations under study must be characterized to allow the selection of patients likely to share a genetic cause of disease. Second, thousands of cases and controls may be needed if a study is to have sufficient statistical power to identify the al-

---

**Figure 2. The Nature of Genomic Variation.**

Invariant nucleotide bases (gray circles) are interspersed with SNPs (orange circles). SNPs lying in close proximity in genome regions that tend to be unaffected by genomic shuffling during meiosis are usually inherited together. The inheritance pattern of SNPs 3 and 4 suggests that they are tightly linked to each other (box) — G travels with C and T travels with T — as well as to SNPs 1 and 2 in Figure 1. One tagging SNP may therefore be used as a surrogate for other SNPs in genome-wide analyses.
leles of interest, and some relevant statistical strategies are still being developed. Third, the need to test hundreds of thousands of DNA variants in thousands of subjects creates bioinformatics challenges and raises questions about how to identify true positive signals in a sea of false positives. In addition, a few technological challenges remain, as do a few gaps in the genetic map.

Moreover, not all common and complex diseases have just one or a small number of common variants as contributing factors. In some disorders, such as sickle cell anemia, there is one common allele that could have been identified by genome-wide association studies had it not already been known, but other disorders with public health importance, such as phenylketonuria, may be caused by any of a number of rare mutations. In these cases, the genome-wide association approach would be ineffective, since there is no common mutation tagged by a particular SNP. Direct DNA sequencing and family-based studies will be required to define the genetics of disorders in which there is substantial allelic heterogeneity.

Finally, even when research involving genome-wide association identifies a physical region of interest, finding the specific mutation may not always be straightforward. Mutations that disrupt the normal amino acid sequence of a protein are ready candidates for causation, but such mutations are not always present. Some causal mutations may lie within regulatory elements that can be difficult to identify.

Furthermore, the genome sequence continues to teach us about new forms of genetic variation; in the past few years, for example, copy-number variation has begun to be recognized as a contributor to human disease. A copy-number variant is the deletion or duplication of up to hundreds of thousands of DNA nucleotides and may span one or more genes; the techniques for detecting such variants in large collections of samples are not sufficiently developed to be used in a cost-effective manner. In this issue of the Journal, Lupski (pages 1169–1171) points out that sporadic disease might be found to result from a copy-number variation or two or more variations in combination. Copy-number variants can change the gene dose and thus cause comparatively subtle changes at the level of gene expression; they may influence susceptibility to complex traits.

There are now many large-scale efforts — in the United States, the United Kingdom, Japan, and elsewhere — to uncover genetic effects and gene–environment interactions relevant to disease. For example, the National Institutes of Health established the Gene Environment Initiative (GEI) to perform such studies on samples that are already in hand, if the sample size, available clinical data, and consent from the institutional review board afford the promise of success. GEI plans to make the genotypic and phenotypic data accessible to interested scientists and physicians, permitting data mining and hastening the discovery process. Ethical issues confronting such data releases are still under discussion; the challenge is to maintain confidentiality while observing the wishes of sample donors to have their samples used for medical advances. Public posting of the prevalence of each SNP in a genome-wide association study in the case group, as compared with controls, would permit researchers to confirm their findings rapidly, without the risk of breaching the confidentiality of the study participants.

The genetic risk factors identified by genome-wide association studies are likely to be associated with moderate risks (as is the case with most known environmental risk factors), rather than the extremely high risks associated with single-gene disorders. However, if such a risk factor is common in a population, it could account for a substantial proportion of that population’s cases. Therefore, the downstream functions of such common genetic
variants will be potential targets for lifestyle or medical interventions.

For the clinician, the outcome of these studies will be similar to those of any new investigative technology. The first results will need to be verified in similar populations by independent groups. Then, the usefulness of the variants for clinical practice will depend on their improving diagnostic prediction or fostering changes in prevention or treatment strategies. There is a compelling need for more, and more efficient, epidemiologic studies so that these new approaches can be exploited. To meet this need, scientists and clinicians must collect information, informed consent, and tissue samples in the expectation of future studies that will address as yet unformed questions.

Dr. Murray reports sitting on the National Advisory Council for Human Genome Research.

Dr. Christensen is a professor of epidemiology at the University of Southern Denmark, Odense, Denmark. Dr. Murray is a professor of pediatrics, biology, and public health at the University of Iowa, Iowa City.


Copyright © 2007 Massachusetts Medical Society.