

REVIEW ARTICLE

GENOMIC MEDICINE

W. Gregory Feero, M.D., Ph.D., and Alan E. Guttmacher, M.D., *Editors*

Genomic Medicine — An Updated Primer

W. Gregory Feero, M.D., Ph.D., Alan E. Guttmacher, M.D.,
and Francis S. Collins, M.D., Ph.D.

Cathy, a 40-year-old mother of three, arrives in your office for her annual physical. She has purchased a commercial genomewide scan (see the Glossary), which she believes measures the clinically meaningful risk that common diseases will develop, and has completed her family history online using My Family Health Portrait (www.familyhistory.hhs.gov), a tool developed for this purpose by the U.S. Surgeon General. Her genomewide scan suggests a slightly elevated risk of breast cancer, but you correctly recognize that this information is of unproven value in routine clinical care. On importing Cathy's family-history file, your office's electronic health record system alerts you to the fact that Cathy is of Ashkenazi Jewish heritage and has several relatives with breast cancer, putting her at heightened risk for the hereditary breast and ovarian cancer syndrome. The system prompts you to discuss Cathy's risk of breast and ovarian cancer during the visit. Considering both her family history and ancestry, you refer Cathy to a health care professional with advanced genetics training for consultation.

In the coming months Cathy elects to have her DNA tested for mutations in *BRCA1* and *BRCA2*, the genes associated with hereditary breast and ovarian cancer syndrome, and to undergo a mammographic examination. Although the results of her genetic tests are negative, her mammogram reveals a suspicious abnormality. A biopsy is performed, and breast cancer is detected. Surgery is successful. Pathological examination of tissue from the excised tumor reveals that it is positive for estrogen-receptor protein and negative for human epidermal growth factor receptor type 2 (HER2); the lymph glands are free of cancer cells. Genetic-expression profiling of the tumor indicates a relatively high risk of recurrent cancer, and Cathy elects to receive adjuvant chemotherapy followed by treatment with tamoxifen. Five years later, the cancer has not recurred.

From the National Human Genome Research Institute (W.G.F.), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (A.E.G.), and the Office of the Director (F.S.C.), National Institutes of Health, Bethesda, MD; and the Maine Dartmouth Family Medicine Residency Program, Augusta, ME (W.G.F.). Address reprint requests to Dr. Feero at the National Human Genome Research Institute, National Institutes of Health, Bldg. 31, Rm. 4B09, 31 Center Dr., Bethesda, MD 20892, or at feerow@mail.nih.gov.

N Engl J Med 2010;362:2001-11.

Copyright © 2010 Massachusetts Medical Society.

REMARKABLE ADVANCES HAVE BEEN MADE IN UNDERSTANDING THE HUMAN genome's contribution to health and disease since the first Genomic Medicine series was launched in the *Journal* in 2002.^{1,2} The vignette about Cathy illustrates the strengths and limitations of these advances. Completion of the Human Genome Project in 2003^{3,4} was a major driver for the current period of biomedical discovery, and the pace continues to accelerate. This project spurred the development of innovations with extraordinary benefits. Initially, clinically useful discoveries derived from the Human Genome Project yielded improvements in "genetic medicine" — that is, the use of knowledge about single genes to improve the diagnosis and treatment of single-gene disorders. However, our increased understanding of the interactions between the entire genome and nongenomic factors that result in health and disease is paving the way for an era of "genomic medicine," in which new diagnostic and therapeutic approaches to common multifactorial conditions are emerging.

As a result of genomic discoveries, increasing numbers of clinical guidelines now suggest incorporating genomic tests or therapeutics into routine care. In some cases, the rapidity of translation has sparked debate regarding the level of evidence of clinical benefit needed to introduce new, and potentially costly, medical technologies.^{5,6} Although the effect of genomic discovery on the day-to-day practice of medicine has not been well quantified, it probably remains small in primary care and nonacademic settings as compared with, for example, oncology practice in an academic medical center. Regardless of where medicine is practiced, genomics is inexorably changing our understanding of the biology of nearly all medical conditions. How can any clinician understand the diagnosis and treatment of breast cancer, much less explain it to a patient such as Cathy, without a rudimentary understanding of genomic medicine?

Here, laying the groundwork for the rest of this series, we review key conceptual and technological advances in genomics that have occurred since the first series appeared in 2002. Readers who wish to review core principles of genetics and genomics are encouraged to revisit that first primer.¹ A glossary of key terms appears in this article and will be updated throughout the course of the Genomic Medicine series.

DEFINING THE GENE AND ITS REGULATION

The question of how genes are defined and regulated is deceptively simple. However, the answer has become increasingly complex and remains a work in progress. Gaining a comprehensive understanding of gene structure and regulation is substantially more than an academic exercise — it provides avenues to develop better diagnostic, prognostic, preventive, and therapeutic approaches to both rare and common diseases.

The gene was traditionally defined as a unit of heredity. Once DNA had been identified as the basis of heredity, and the central dogma of molecular biology (DNA→RNA→protein) had been established, the gene was defined as a segment of DNA encoding a protein.⁷ But with the discovery of new classes of RNA, the traditional definition of a gene has required re-examination. The emerging picture of gene regulation depicts interdependent layers and webs of control consisting of interactions of DNA with regulatory proteins and

RNA molecules that are akin to the interactions that occur in computer circuitry. This development has led to the rise of sophisticated “systems biology” approaches to understanding regulation.⁸

The rigorous comparison of the full genome sequences of organisms from bacteria to the chimpanzee has contributed greatly to our understanding of human gene structure and regulation. Table 1 provides an abbreviated list of organisms whose genomes have been fully sequenced.^{9–18} “Comparative genomics” relies on the fact that DNA sequences that are critical to gene function are typically conserved across species. Such cross-species comparisons have shown that although the human genome contains approximately 20,000 protein-coding genes^{9,19} (a relatively small number — both cows²⁰ and mustard plants²¹ have more), we make protean use of them. Close inspection reveals that some human genes are nested within other genes, that genes can occur on forward and reverse strands of the same DNA sequence, and that a single gene can encode multiple proteins or RNA molecules. Moreover, a great diversity of regulatory sequences in DNA can be located anywhere from inside the gene they affect to a great distance from that gene.²² Protein–DNA and RNA–DNA interactions and chemical modifications of DNA that do not affect the primary sequence also affect gene expression. Increasingly, the three-dimensional structure of DNA is recognized as playing an important role in regulating gene expression.²³ Regulation of the transcription of DNA to RNA is only the first layer of control of gene expression in humans — alternative splicing (the processing of newly synthesized RNA molecules into the final functional RNA molecule) and regulation of translation (the process by which ribosomes read messenger RNA [mRNA] molecules to create proteins) are also tightly regulated (Fig. 1).

One of the more remarkable stories to unfold in biology since 2002 is the diverse and ubiquitous role of small RNA molecules in gene regulation.²⁴ Evidence suggests that this class of molecules contributes to disease pathogenesis, particularly in cancer²⁵ and diseases caused by dysregulation of the immune system.²⁶ Tests based on expression patterns of microRNAs (miRNAs) in tumors can augment traditional pathological techniques for determining the cell type giving rise to tumors²⁷; miRNAs are endogenous noncoding RNA molecules, usually 22 nucleotides in length, that inhibit translation of their target RNAs. Similar in

concept to the miRNA is the small, or short, interfering RNA (siRNA), which binds to complementary mRNA molecules and represses translation through degradation. Rationally designed synthetic siRNA molecules are currently being tested in advanced clinical trials.²⁸

GENOMIC VARIATION

Given the diversity of the human species, there is no “normal” human genome sequence. We are all mutants. Specific locations in the human genome where differences between individual people are found are generally referred to as variations, and the term “normal” or “wild type” is often used to refer to the most common variant at a location in a given population group. In its simplest form, the variation has two different spellings, referred to as “alleles.” If the frequency of the minor allele is greater than 1%, such variants are called polymorphisms. The word “mutation” is generally reserved for changes in DNA that are believed or known to be pathologic (e.g., the mutations in the gene that causes cystic fibrosis — cystic fibrosis transmembrane conductance regulator, or *CFTR*) or for changes that are recent (e.g., a DNA base change in a cancer that is not present in the patient’s germ-line DNA). A person’s complete genomic sequence (genotype), acting in concert with environmental influences, creates individuality (phenotype). A collection of alleles arranged linearly along a person’s DNA molecule is known as a haplotype. Humans are very similar at the DNA sequence level; about 99.6% of base pairs are identical from person to person.²⁹ Given the size of the genome (approximately 6 billion bp in every nucleated non-germ-line cell), there is substantial latitude for individual genetic variation, since the difference between any two people is about 24 million bp.

Deleterious mutations occurring in the DNA of germ-line cells become ubiquitous mutations in the developing body because they are present in every cell. These mutations can give rise to the classical mendelian patterns of inheritance. New mutations in somatic cells are not transmitted from generation to generation and are critical to the development of cancer. Alterations in mitochondrial DNA, which are heritable through the maternal lineage, give rise to a variety of uncommon conditions that typically affect energy metabolism. The familial aggregation of disease seen in complex conditions such as diabetes and coro-

nary artery disease — disorders in which phenotype is determined by both genes and environment — can be the result of contributions from both common and rare genetic variations.

Events contributing to genomic variation fall into three categories: single-base-pair changes (or point mutations) that disturb the “normal” DNA nucleotide sequence (e.g., substitution of adenine for guanine); insertions and deletions of nucleotides from the DNA; and structural rearrangements that reshuffle the DNA sequence, thus changing the order of nucleotides (Fig. 2).

Although these broad categories for describing variation have not changed over the past several years, our understanding of the content of the categories has changed. Consider simple single-base-pair changes in the DNA sequence. The HapMap Project, completed in 2005,³⁰ provided a genomewide map of common single-base-pair variations (also known as single-nucleotide polymorphisms, or SNPs) in persons from a variety of population groups. This project has shown that SNPs are very common throughout the human genome, are often correlated with their neighboring SNPs, and occur, on average, approximately every 800 bp. The HapMap Project has also provided a powerful tool (the so-called genomewide association study) that facilitates the identification of genetic associations with complex conditions.

Over the past 5 years, genomewide association studies have made it possible to measure the associations between mapped SNPs and the presence of common complex conditions in large patient cohorts, thereby revolutionizing the study of many traits and diseases.³¹ In 2003, a mere handful of gene variants were known to be associated with common complex conditions; the number of well-validated associations is now in the hundreds, and the list grows each week.³² Counterintuitively, most of the SNPs that have been found to be associated with common complex conditions to date are outside the protein-coding DNA sequence (exons) of genes. In fact, some SNPs, such as those on chromosome 8q24, which are associated with an increased risk of prostate cancer, occur at great base-pair distances from any known protein-coding sequence.³³ We have much to learn about the function of DNA between known genes.

Most SNPs associated with common diseases explain a small proportion of the observed contribution of heredity to the risk of disease — in many cases less than 5 to 10% — substantially

Glossary

- Allele:** One of two or more versions of a genetic sequence at a particular location in the genome.
- Alternative splicing:** Use of different exons in the formation of messenger RNA from initially identical transcripts, which can result in the generation of related proteins from one gene, often in a manner specific to a type of tissue or a developmental stage.
- Base pair (bp):** Two nitrogenous bases paired together in double-stranded DNA by weak bonds; specific pairing of these bases (adenine with thymine and guanine with cytosine) facilitates accurate DNA replication; when quantified (e.g., 8 bp), bp refers to the physical length of a sequence of nucleotides.
- BRCA1 and BRCA2:** Tumor-suppressor genes associated with inherited forms of breast cancer and ovarian cancer. In women with mutations in either gene, there is a much higher risk of breast and certain other cancers than in women without such mutations.
- CFTR:** The gene encoding cystic fibrosis transmembrane conductance regulator, a chloride channel expressed in epithelial cells that causes cystic fibrosis when mutated.
- Complex condition:** A condition caused by the interaction of multiple genes and environmental factors. Examples of complex conditions, which are also called multifactorial diseases, are cancer and heart disease.
- Copy-number variation:** Variation from one person to the next in the number of copies of a particular gene or DNA sequence. The full extent to which copy-number variation contributes to human disease is not yet known.
- Deletion mutation:** A mutation that involves the loss of genetic material. It can be small, involving a single missing DNA base pair, or large, involving a piece of a chromosome.
- DNA:** Deoxyribonucleic acid, the molecules inside cells that carry genetic information and pass it from one generation to the next.
- Epigenetic change:** A change in the regulation of the expression of gene activity without alteration of genetic structure.
- Exon:** The portion of a gene that encodes amino acids.
- Forward and reverse strands:** The two strands of a double-stranded DNA molecule. The complementary nucleotide sequence of both strands may encode important information.
- Gene:** The fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product (i.e., a protein or an RNA molecule).
- Gene chip:** A solid substrate, usually silicon, onto which a microscopic matrix of nucleotides is attached. Gene chips, which can take a wide variety of forms, are frequently used to measure variations in the amount or sequence of nucleic acids in a sample.
- Genome:** The entire set of genetic instructions found in a cell. In humans, the genome consists of 23 pairs of chromosomes, found in the nucleus, as well as a small chromosome found in the cells' mitochondria.
- Genomewide association study:** An approach used in genetics research to look for associations between many (typically hundreds of thousands) specific genetic variations (most commonly single-nucleotide polymorphisms) and particular diseases.
- Genomewide scan:** An assay that measures hundreds of thousands to millions of points of genetic variation across a person's genome simultaneously, either for research or for clinical application.
- Genotype:** A person's complete collection of genes. The term can also refer to the two alleles inherited for a particular gene.
- Haplotype:** A set of DNA variations, or polymorphisms, that tend to be inherited together. A haplotype can refer to a combination of alleles or to a set of single-nucleotide polymorphisms found on the same chromosome.
- HapMap:** The nickname of the International HapMap (short for "haplotype map") Project, an international venture that seeks to map variations in human DNA sequences to facilitate the discovery of genetic variants associated with health. The HapMap describes common patterns of genetic variation among people.
- Histone:** A class of proteins that provide chromosomes with structural support. Histones can undergo reversible covalent modification and are associated with the regulation of gene expression.
- Human Genome Project:** An international project completed in 2003 that mapped and sequenced the entire human genome.
- Insertion mutation:** A type of mutation involving the addition of genetic material. An insertion mutation can be small, involving a single extra DNA base pair, or large, involving a piece of a chromosome.
- Mendelian inheritance:** Patterns of inheritance characteristic of organisms that reproduce sexually, as described by Austrian monk Gregor Mendel in the mid-19th century.
- Methylation:** The attachment of methyl groups to DNA at cytosine bases. Methylation is correlated with reduced transcription of the gene and is thought to be the principal mechanism in X-chromosome inactivation and imprinting.
- Microarray:** A technology used to study many genes at once. Thousands of gene sequences are placed in known locations on a glass slide. A sample containing DNA or RNA is deposited on the slide, now referred to as a gene chip. The binding of complementary base pairs from the sample and the gene sequences on the chip can be measured with the use of fluorescence to detect the presence and determine the amount of specific sequences in the sample.
- microRNA (miRNA):** A short regulatory form of RNA that binds to a target RNA and generally suppresses its translation by ribosomes.

Glossary (Continued.)

Mitochondrial DNA: The small circular chromosome found inside mitochondria. Mitochondria, and thus mitochondrial DNA, are passed from mother to offspring.
Messenger RNA (mRNA): RNA that serves as a template for protein synthesis.
Mutation: A change in a DNA sequence. Germ-line mutations occur in the eggs and sperm and can be passed on to offspring, whereas somatic mutations occur in body cells and are not passed on.
Next-generation sequencing: DNA sequencing that harnesses advances in miniaturization technology to simultaneously sequence multiple areas of the genome rapidly and at low cost.
Nucleotide: The basic building block of nucleic acids. RNA and DNA are polymers made of long chains of nucleotides. A nucleotide consists of a sugar molecule (either ribose in RNA or deoxyribose in DNA) attached to a phosphate group and a nitrogen-containing base. The bases used in DNA are adenine (A), cytosine (C), guanine (G), and thymine (T). In RNA, the base uracil (U) takes the place of thymine.
Oligonucleotide: A short, typically synthetic, polymer of nucleotides.
Pharmacogenomics: A branch of pharmacology concerned with using DNA sequence variation to inform drug development and testing. An important application of pharmacogenomics is the correlation of individual genetic variations with drug responses.
Phenotype: The observable traits of an individual person, such as height, eye color, and blood type. Some traits are largely determined by genotype, whereas others are largely determined by environmental factors.
Point mutation: An alteration in DNA sequence caused by a single-nucleotide base change, insertion, or deletion.
Polymerase chain reaction (PCR): A laboratory technique used to amplify DNA sequences. Short, synthetic complementary DNA sequences called primers are used to select the portion of the genome to be amplified. The temperature of the sample is repeatedly raised and lowered to facilitate the copying of the target DNA sequence by a DNA-replication enzyme. The technique can produce a billion copies of the target sequence in just a few hours.
Rearrangement: A structural alteration in a chromosome, usually involving breakage and reattachment of a segment of chromosomal material, resulting in an abnormal configuration; examples include inversion and translocation.
Ribosome: A cellular particle made of RNA and protein that serves as the site for protein synthesis in the cell. The ribosome reads the sequence of the mRNA and, using the genetic code, translates the sequence of RNA bases into a sequence of amino acids.
RNA: Ribonucleic acid, a chemical similar to DNA. The several classes of RNA molecules play important roles in protein synthesis and other cell activities.
Single-nucleotide polymorphism (SNP): A single-nucleotide variation in a genetic sequence; a common form of variation in the human genome.
Small (or short) interfering RNA (siRNA): A short, double-stranded regulatory RNA molecule that binds to and induces the degradation of target RNA molecules.
Somatic cell: Any cell of the body except sperm and egg cells. Somatic cells are diploid, meaning that they contain two sets of chromosomes, one inherited from each parent. Mutations in somatic cells can affect the person in which they occur but are not passed on to offspring.
Systems biology: Research that takes a holistic rather than reductionist approach to understanding organism functions.
Transcription: The synthesis of an RNA copy from a sequence of DNA (a gene); a first step in gene expression.
Translation: During protein synthesis, the process through which the sequence of bases in a molecule of messenger RNA is read in order to create a sequence of amino acids.

limiting the use of these markers to predict risk. It thus comes as no surprise that as yet there are no evidence-based guidelines that recommend the use of SNP markers in assessing the risk of common diseases in clinical care.³⁴ Considerable resources are being invested in discovering these unknown sources of heritable risk.³⁵ Improved risk-analysis models incorporating genomic and nongenomic factors will emerge when an increased percentage of heritable risk can be measured.

An important yield of genomewide association

studies is information about the role of specific proteins and biologic pathways in pathogenesis; these proteins and pathways are candidate targets for the development of preventive and therapeutic methods for disease management. Examples include the complement factor system in age-related macular degeneration and the autophagy pathway in inflammatory bowel disease.

Insertions and deletions are common and range in size from one to thousands of base pairs; like SNPs, they can be benign (having no effect on phenotype) or can confer a risk of disease. A class

Table 1. Representative List of Organisms with Fully Sequenced Genomes.

Organism	Classification	Genome Size*	Estimated No. of Protein-Coding Genes†
Human (<i>Homo sapiens</i>)	Placental mammal	3.2 Gb	19,042
Chimpanzee (<i>Pan troglodytes</i>)	Placental mammal	2.7 Gb	19,000
Mouse (<i>Mus musculus</i>)	Placental mammal	2.6 Gb	20,210
Dog (<i>Canis familiaris</i>)	Placental mammal	2.4 Gb	19,300
Platypus (<i>Ornithorhynchus anatinus</i>)	Monotreme	2.2 Gb	18,527
Rice (<i>Oryza sativa</i>)	Plant	389 Mb	37,544
Mosquito (<i>Anopheles gambiae</i>)	Insect	278 Mb	15,189
<i>Plasmodium falciparum</i> (organism causing malaria)	Protozoa	22.8 Mb	5,300
Yeast (<i>Saccharomyces cerevisiae</i>)	Fungus	12.1 Mb	6,607
<i>Escherichia coli</i>	Bacterium	4.6 Mb	3,200
Human immunodeficiency virus	Retrovirus	9.1 Kb	9

* Kb denotes kilobases (10^3), Mb megabases (10^6), and Gb gigabases (10^9).

† Because the process of predicting protein-coding sequences is complex, estimates of gene numbers vary in the literature and change over time.

of these changes known as copy-number variations is associated with a growing list of disorders that includes autism and schizophrenia.^{36,37} Copy-number variations may account for part of the heritability of common diseases that cannot currently be accounted for by SNPs. Smaller structural rearrangements, not yet easily measured in large numbers of people, may also have a substantive role in pathogenesis, although much remains to be learned about how and to what extent these changes contribute to heritable risk of disease, particularly in the case of common disorders.

Our understanding of the role of epigenetic changes in the regulation of gene expression has advanced substantially over the past decade. Epigenetic changes are chemical alterations of the DNA molecule that do not affect the primary base-pair sequence.³⁸ An example is the enzymatic methylation of cytosine nucleotides in regions of DNA (often referred to as DNA methylation) that are not being actively transcribed in differentiated cells. This process has the effect of maintaining repression of the transcription of genes in the vicinity of the methylated DNA. Cancers frequently show markedly abnormal patterns of DNA methylation,³⁹ and drugs targeting aberrant methylation pathways are being studied in clinical trials. Chemical modification of DNA-binding proteins (e.g., histones) also affects transcription. Epigenetic changes vary with time and are in-

fluenced by the environment as well as genetic determinants.⁴⁰

MEASUREMENT OF VARIATION

The landscape of molecular diagnostics is changing rapidly. In the pregenomics era, genetic diagnosis focused mainly on conditions caused by mutations in single genes that required the detection of just one or a handful of mutations. Now the focus is shifting to highly multiplexed tests that detect thousands — even millions — of variants at once.⁴¹ Computer-chip technologies combined with microscopical fluorescence detection allow the mass manufacture of highly efficient and accurate gene chips. In 2010, a single silicon chip can detect well over 1 million different base-pair variations in a person's genome in a few hours for a few hundred dollars. Generating the same amount of information in 2002, when the first article in the Genomic Medicine series appeared, would have required months of work by a laboratory team and hundreds of electrophoresis gels, at a cost of more than \$500,000.

Clinical diagnostic laboratories continue to use assays based on the sequence-specific binding of short complementary DNA probes (oligonucleotides) to DNA samples from patients in order to detect mutations associated with single-gene disorders. Generally, these assays rely on the polymerase chain reaction (PCR) to amplify regions

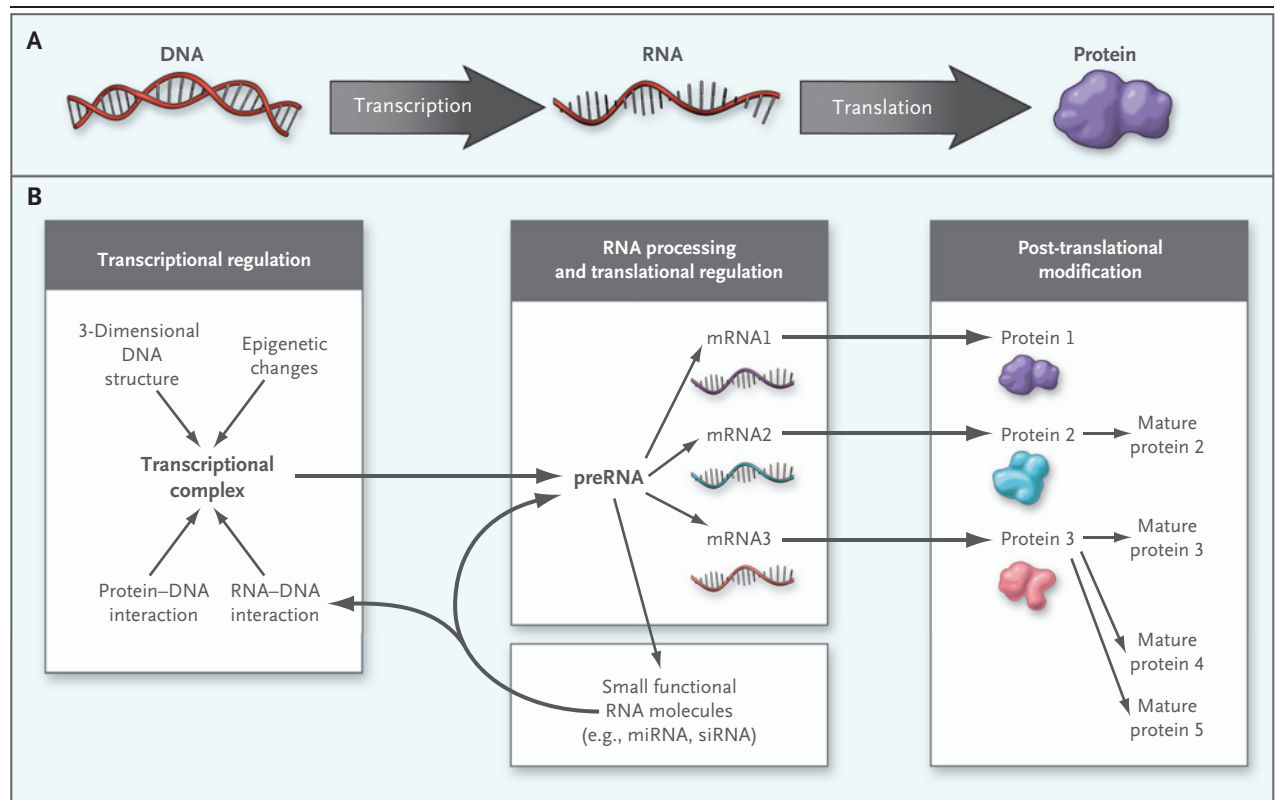


Figure 1. The Increasing Complexity of the Central Dogma of Molecular Biology.

The flow of genomic information from DNA to RNA to protein remains the basis for understanding genomic function (Panel A). A single gene can yield an extensive array of gene products, depending on the environment in which it is expressed, thereby expanding the repertoire of the 20,000 or so genes in the human genome (Panel B). The initial event of gene expression, transcription, is regulated by means of a complex choreography of events involving the three-dimensional DNA structure, covalent chemical, or epigenetic, modifications of the DNA backbone, and interactions between protein and DNA and between RNA and DNA. Translation is similarly complex and tightly regulated by interactions between messenger RNA (mRNA) and proteins. Processing of single-precursor RNA (preRNA) molecules can yield multiple RNA products, including microRNA (miRNA) and small interfering RNA (siRNA) molecules. Post-translational modification of proteins also contributes greatly to the diversity of the output of the human genome through modifications of individual immature proteins (e.g., folding, cleavage, and chemical modifications), which yield an array of related protein products.

of interest in a patient's DNA; the PCR product is then analyzed for the presence or absence of mutations. New approaches, including the use of gene chips and sequencing, are rapidly eclipsing traditional methods of detecting human genetic variation and mutations.

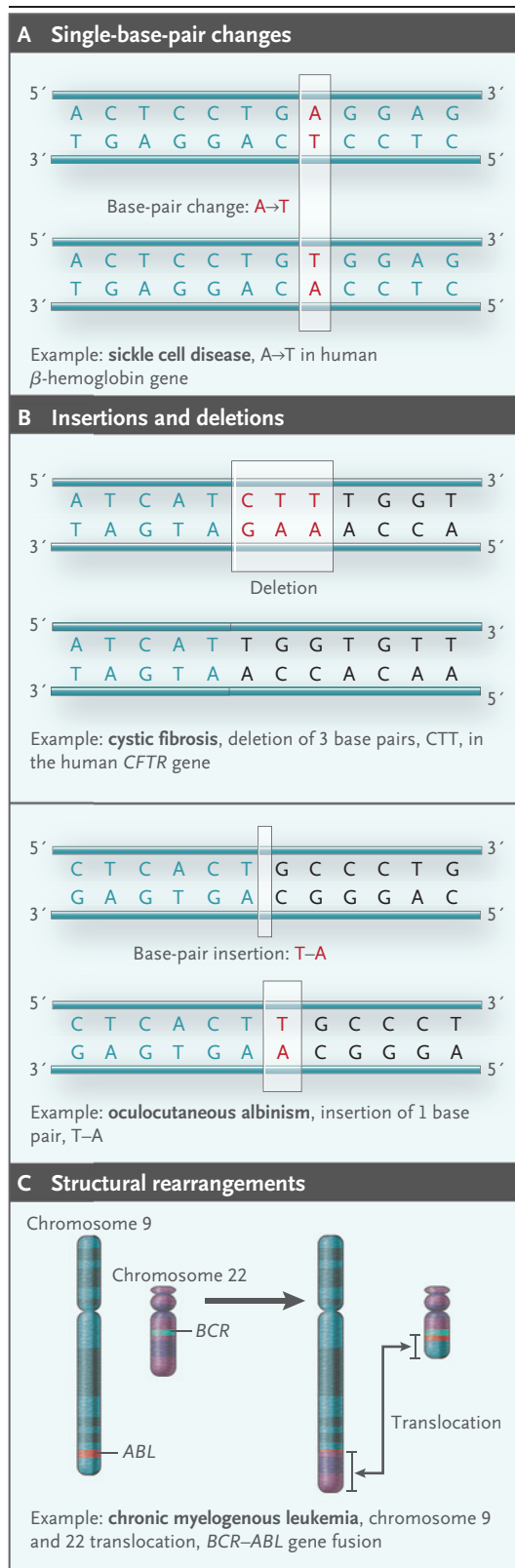
Gene chips consist of a highly ordered microscopic matrix of sequence-specific oligonucleotides tethered to a solid surface, known as a microarray (Fig. 3). To perform a genomewide SNP scan such as the type purchased by Cathy, DNA is isolated from a sample obtained from a patient, cut into small fragments, labeled with a fluorescent dye, and then incubated with the silicon chip. The fragments bind to the tethered oligonucleotides in a sequence-specific manner, and sophisticated scanning hardware and signal-processing soft-

ware analyze the pattern and intensity of the fluorescence signal to determine the sequences present in the sample. Barring human error in the laboratory, current clinical assays typically provide analytical sensitivities and specificities of more than 99.5%. Chip technologies have been adapted for use in a variety of clinical applications, including measurement of changes in structural DNA in patients with unexplained mental retardation⁴² and DNA expression profiling in the analysis of tumor samples.⁴³ High-throughput genotyping can also be achieved with technologies based on the use of microfluidics and microscopic beads coated with oligonucleotides.⁴⁴

The technology used to determine the complete DNA sequence of an individual person is also moving forward rapidly (Fig. 4, and the interactive



An interactive graphic showing DNA-sequencing technologies is available at NEJM.org

**Figure 2. Human Genetic Variation.**

Human genetic variation can be grouped into three major categories: single-base-pair changes (Panel A), small and large insertion and deletion events (Panel B), and structural rearrangements (Panel C). The scale and consequences of such changes can vary dramatically, depending on where and when such a variation occurs. For example, the change of a single base pair can have profound health consequences (e.g., the substitution of a thymidine base for an adenine base in the human β -hemoglobin gene), whereas a large, balanced translocation event (in which the genetic information on an entire arm of a chromosome may switch places with the arm of another chromosome) may have no direct consequences for the affected person.

graphic available with the full text of this article at NEJM.org). Until very recently, even the most sophisticated automated-sequencing devices relied on strategies and chemistry invented in the 1970s. Commercial sequencing machines now make use of diverse methods.⁴⁵ Each has strengths and weaknesses, but all are markedly faster and less expensive than the methods used to generate the first complete human genome sequences. Newer approaches have improved fidelity by deliberately sequencing the same section of DNA multiple times in a highly parallel manner. The sequence data are then assembled with the use of sophisticated computer programs, a feat that requires considerable computational power. New sequencing methods based on nanotechnologies may further enhance accuracy while reducing costs.⁴⁶ The goal of completely sequencing a human genome for \$1,000 is in sight. Affordable, high-throughput genome sequencing will be a key tool in biomedical research for the foreseeable future. One can envision how sequencing might augment areas of clinical practice as disparate as newborn screening, drug selection (pharmacogenomics), and risk-reduction strategies for common complex conditions.

Two clinical axioms familiar to every medical intern should be heeded when contemplating the clinical use of high-throughput SNP genotyping or DNA sequencing. First, “Don’t order a test unless you know what to do with the result.” For instance, years of experience with sequencing the *BRCA1* and *BRCA2* genes for clinical purposes has yielded thousands of sequence variants in these two genes, many of which are rare and of unknown significance. Population-level sequencing

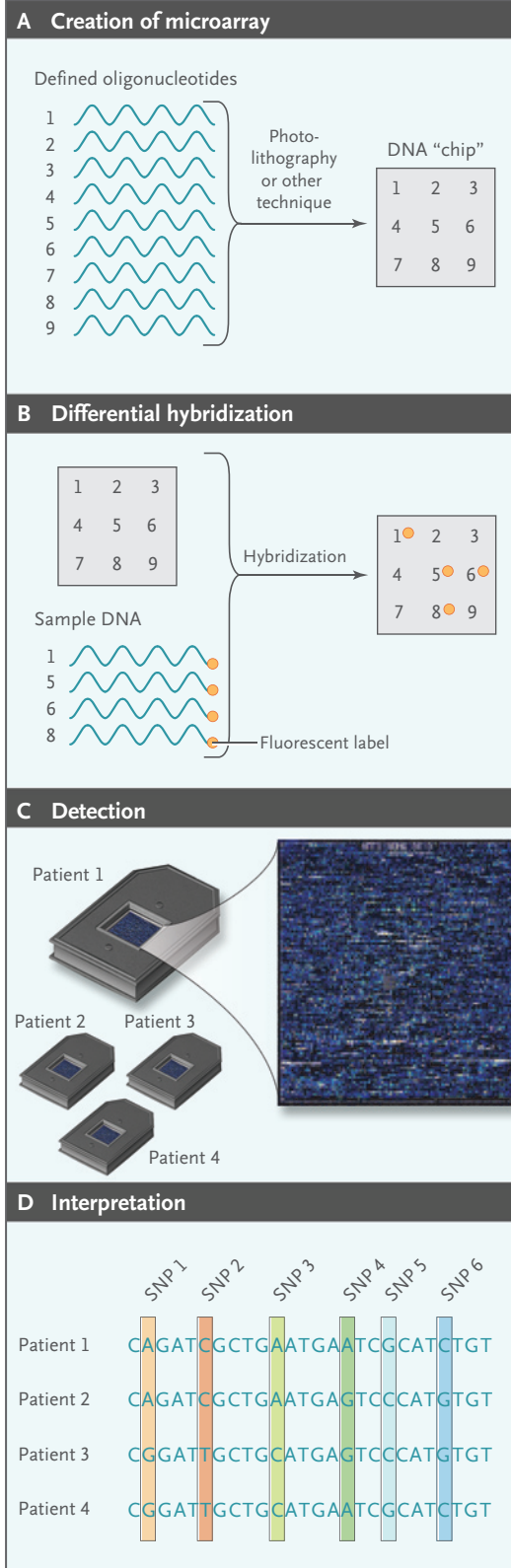
Figure 3. Microarray Technologies.

Microarray technologies, or gene chips, are at the heart of many of the most important scientific and clinical advances of the past 5 years. With current technologies, information on more than a million unique sequence variants can be provided on a single chip. Microarray technologies used to detect single-nucleotide polymorphisms (SNPs) share some common features: first, creation of a structured microscopic arrangement (array) of oligonucleotides of defined sequence on a substrate such as a silicon chip (Panel A); next, hybridization of a fluorescently labeled patient DNA sample to the array (Panel B); subsequently, scanning of the array to detect the location and amount of sequence-specific binding in the sample (Panel C); and, finally, computational processing of the raw image data from the array to yield an interpretable readout of SNP data (Panel D).

will provide the first truly unbiased look at 20,000 genes that have rarely been sequenced in “normal” persons, and it will probably reveal that we understand genotype–phenotype correlations even less well than we currently suppose. Second, “When you order 20 tests, each with 95% specificity, you are likely to get at least one false positive result.” Even if sequencing is 99.9999% accurate, a full diploid genome sequence will contain 6000 errors. Separating the wheat from the chaff will be a considerable challenge. Highly accurate sequencing will be required, and approaches are being developed to achieve this through very-high-fidelity primary-sequence production coupled with redundant sequencing of the same region to check for errors. Such approaches will not obviate the need for large population-based studies to help define genotype–phenotype correlations. Harmonizing research and clinical informatics across the U.S. health care system to create a longitudinal living laboratory for observational research would facilitate such studies. On a practical level, advanced medical informatics systems will be critical for clinicians hoping to use sequence information in patient care.

FUTURE DIRECTIONS

The wealth of scientific discovery generated over the past 10 years is unparalleled in the history of biomedicine. Moreover, the rate of discovery is accelerating. At the outset of the past decade, the identification of gene mutations causing single-gene disorders was largely a cottage industry, con-



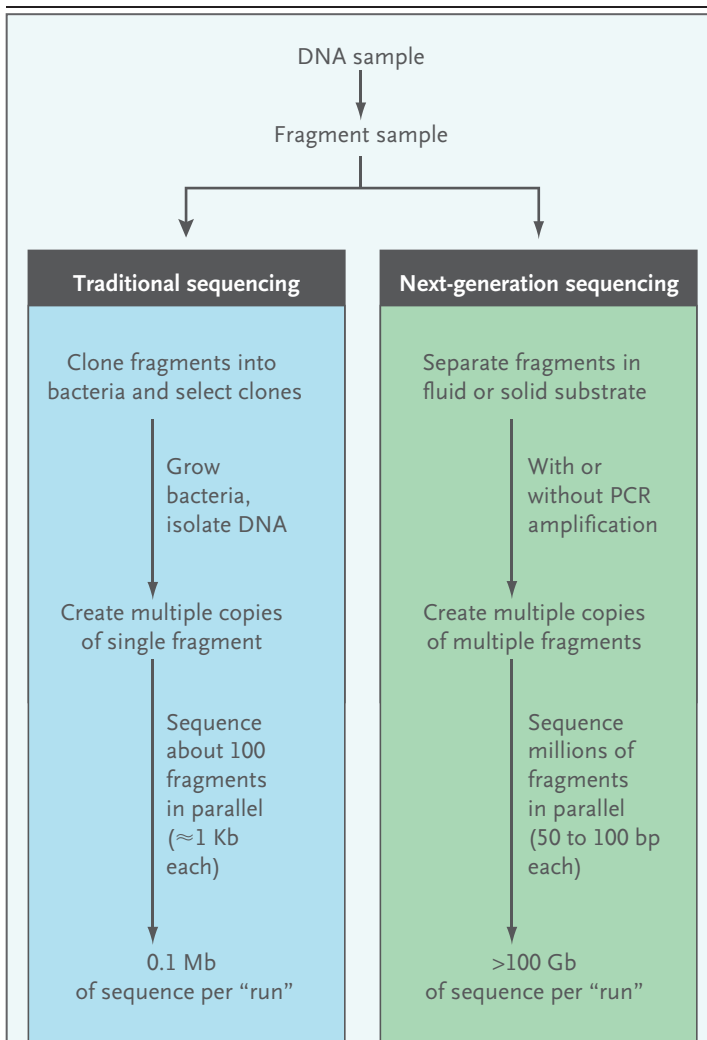


Figure 4. Sequencing Technologies.

Rapid advances in sequencing technologies have brought the cost of sequencing DNA down at a rate that exceeds that of Moore's law. Traditional sequencing technologies rely on a labor- and resource-intensive process of cloning fragments of sample DNA into bacteria, selecting and growing the clones, and then sequencing purified copies of a single fragment. Each traditional sequencing reaction yields approximately 1 Kb (1000 bp) of DNA sequence; typically about 100 reactions are run in parallel. In contrast, next-generation sequencing technologies take advantage of miniaturization and automation to sequence hundreds of thousands to millions of DNA fragments in parallel using extremely small amounts of chemical reagents per reaction. A single sequencing run with next-generation technologies can yield more than 100 Gb (100 billion bp) of DNA sequence in a matter of hours or days, depending on the particular technology used. On the horizon are "\$1,000 genome" technologies that further refine sequencing by allowing direct sequencing of individual DNA molecules. Such refinements have the potential to eliminate the need for sample amplification, further reduce the use of chemical reagents, and produce highly accurate sequence data more rapidly.

ducted by individual laboratories studying extended families. By the end of the decade, genomewide association studies targeting gene associations for complex conditions involved collaborations of research groups spanning the globe. International collaborations of large-scale sequencing centers are generating terabytes of sequence data at speeds and costs that seemed inconceivable 5 years ago. Much discovery now takes place with the use of a computer connected by means of the Internet to a wealth of databases containing information on genotype and phenotype for humans and other model organisms.

The ability to measure human genetic variation reliably and inexpensively in research settings has fueled and shaped the movement toward personalized medicine in health care. Although personalized medicine has many definitions, most share the core idea that any one patient's health is best managed by tailoring preventive measures and treatment to personal preferences as well as to the patient's particular environmental and biologic — including genomic — attributes. There is an inherent, unresolved tension between genomics-enabled personalized medicine and the tenets of population-based, evidence-based medicine.⁴⁷ However, there is no reason that the two approaches to caring for patients should be in opposition. A first step to bridge genomics-enabled medicine and evidence-based medicine would be to collect and store DNA from enrollees in most new clinical trials of drugs and devices. With appropriate consent, such samples could provide a resource for learning about the role of genomic variation in treatment response. There is also a need for well-designed prospective clinical trials that measure patient-oriented outcomes of selected genomic applications (e.g., in the arena of pharmacogenomics) for which the advantages of introducing new technology over using standard care are not immediately evident. Creative solutions are being rapidly developed in the public and private sectors to allow for translational research that will help bridge the gap between the worlds of personalized medicine and evidence-based medicine.^{48,49}

The articles in this series will provide a sampling of the state of the art of genomics in biomedicine, although readers should recognize that the reach of genomics extends to scientific disciplines such as agriculture, paleontology, and renewable energies. The articles will appear regularly and will cover topics from the nuts and

bolts of genomewide association studies to cancer genomics to the ways in which genomics is reshaping our perceptions of race, ethnicity, and ancestry. Each article will emphasize what can be done for patients today and provide a basis for

understanding what will be possible for patients in the years to come.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank Jeff Schloss, Ph.D., of the National Human Genome Research Institute, for his help in preparing Figure 4.

REFERENCES

1. Guttmacher AE, Collins FS. Genomic medicine — a primer. *N Engl J Med* 2002; 347:1512-20.
2. Varmus H. Getting ready for gene-based medicine. *N Engl J Med* 2002;347: 1526-7.
3. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860-921.
4. Collins FS, Morgan A, Patrinos A. The Human Genome Project: lessons from large-scale biology. *Science* 2003;300:286-90.
5. Hunter DJ, Khoury MJ, Drazen JM. Letting the genome out of the bottle — will we get our wish? *N Engl J Med* 2008; 358:105-7.
6. McGuire AL, Burke W. An unwelcome side effect of direct-to-consumer personal genome testing: raiding the medical commons. *JAMA* 2008;300:2669-71.
7. National Library of Medicine. Genetics home reference: gene. (Accessed April 30, 2010, at <http://ghr.nlm.nih.gov/glossary= gene>.)
8. Davidson EH, Levine MS. Properties of developmental gene regulatory networks. *Proc Natl Acad Sci U S A* 2008;105: 20063-6.
9. Church DM, Goodstadt L, Hillier LW, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* 2009;7(5):e1000112.
10. Blattner FR, Plunkett G III, Bloch CA, et al. The complete genome sequence of *Escherichia coli* K-12. *Science* 1997;277: 1453-62.
11. Ratner L, Haseltine W, Patarca R, et al. Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature* 1985;313: 277-84.
12. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 2005;437:69-87.
13. Gardner MJ, Shallom SJ, Carlton JM, et al. Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14. *Nature* 2002;419:531-4.
14. Holt RA, Subramanian GM, Halpern A, et al. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 2002;298:129-49.
15. Warren WC, Hillier LW, Marshall Graves JA, et al. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* 2008;453:175-83.
16. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. *Science* 1996;274: 546, 563-7.
17. Lindblad-Toh K, Wade CM, Mikkelsen TS, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 2005;438:803-19.
18. The map-based sequence of the rice genome. *Nature* 2005;436:793-800.
19. Clamp M, Fry B, Kamal M, et al. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* 2007;104:19428-33.
20. Elsik CG, Tellam RL, Worley KC, et al. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 2009;324:522-8.
21. Yamada K, Lim J, Dale JM, et al. Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* 2003; 302:842-6.
22. Birney E, Stamatoyannopoulos JA, Dutta A, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007;447:799-816.
23. Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* 2009;324:389-92.
24. Carthew RW, Sontheimer EJ. Origins and mechanisms of miRNAs and siRNAs. *Cell* 2009;136:642-55.
25. Bartels CL, Tsongalis GJ. MicroRNAs: novel biomarkers for human cancer. *Clin Chem* 2009;55:623-31.
26. Xiao C, Rajewsky K. MicroRNA control in the immune system: basic principles. *Cell* 2009;136:26-36.
27. Lee YS, Dutta A. MicroRNAs in cancer. *Annu Rev Pathol* 2009;4:199-227.
28. Castanotto D, Rossi JJ. The promises and pitfalls of RNA-interference-based therapeutics. *Nature* 2009;457:426-33.
29. Kidd JM, Cooper GM, Donahue WF, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* 2008;453:56-64.
30. A haplotype map of the human genome. *Nature* 2005;437:1299-320.
31. Manolio TA, Brooks LD, Collins FSA. HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008;118:1590-605.
32. National Human Genome Research Institute. A catalog of published genomewide association studies. (Accessed April 30, 2010, at <http://www.genome.gov/gwastudies/>.)
33. Hindorf LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009;106:9362-7.
34. Kraft P, Hunter DJ. Genetic risk prediction — are we there yet? *N Engl J Med* 2009;360:1701-3.
35. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009;461:747-53.
36. Cook EH Jr, Scherer SW. Copy-number variations associated with neuropsychiatric conditions. *Nature* 2008;455:919-23.
37. Stefansson H, Rujescu N, Cichon S, et al. Large recurrent microdeletions associated with schizophrenia. *Nature* 2008; 455:232-6.
38. Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. *Nature* 2007;447:433-40.
39. Esteller M. Epigenetics in cancer. *N Engl J Med* 2008;358:1148-59.
40. Bjornsson HT, Sigurdsson MI, Fallin MD, et al. Intra-individual change over time in DNA methylation with familial clustering. *JAMA* 2008;299:2877-83.
41. McCarroll SA, Kuruvilla FG, Korn JM, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* 2008;40:1166-74.
42. Sagoo GS, Butterworth AS, Sanderson S, Shaw-Smith C, Higgins JP, Burton H. Array CGH in patients with learning disability (mental retardation) and congenital anomalies: updated systematic review and meta-analysis of 19 studies and 13,926 subjects. *Genet Med* 2009;11:139-46.
43. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. *N Engl J Med* 2009;360:790-800.
44. Gunderson KL. Whole-genome genotyping on bead arrays. *Methods Mol Biol* 2009;529:197-213.
45. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;26:1135-45.
46. Branton D, Deamer DW, Marziali A, et al. The potential and challenges of nanopore sequencing. *Nat Biotechnol* 2008;26: 1146-53.
47. Khoury MJ, Gwinn M, Burke W, Bowen S, Zimmern R. Will genomics widen or help heal the schism between medicine and public health? *Am J Prev Med* 2007; 33:310-7.
48. Teutsch SM, Bradley LA, Palomaki GE, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. *Genet Med* 2009;11:3-14.
49. Khoury MJ, Feero WG, Reyes M, et al. The genomic applications in practice and prevention network. *Genet Med* 2009;11: 488-94.

Copyright © 2010 Massachusetts Medical Society.