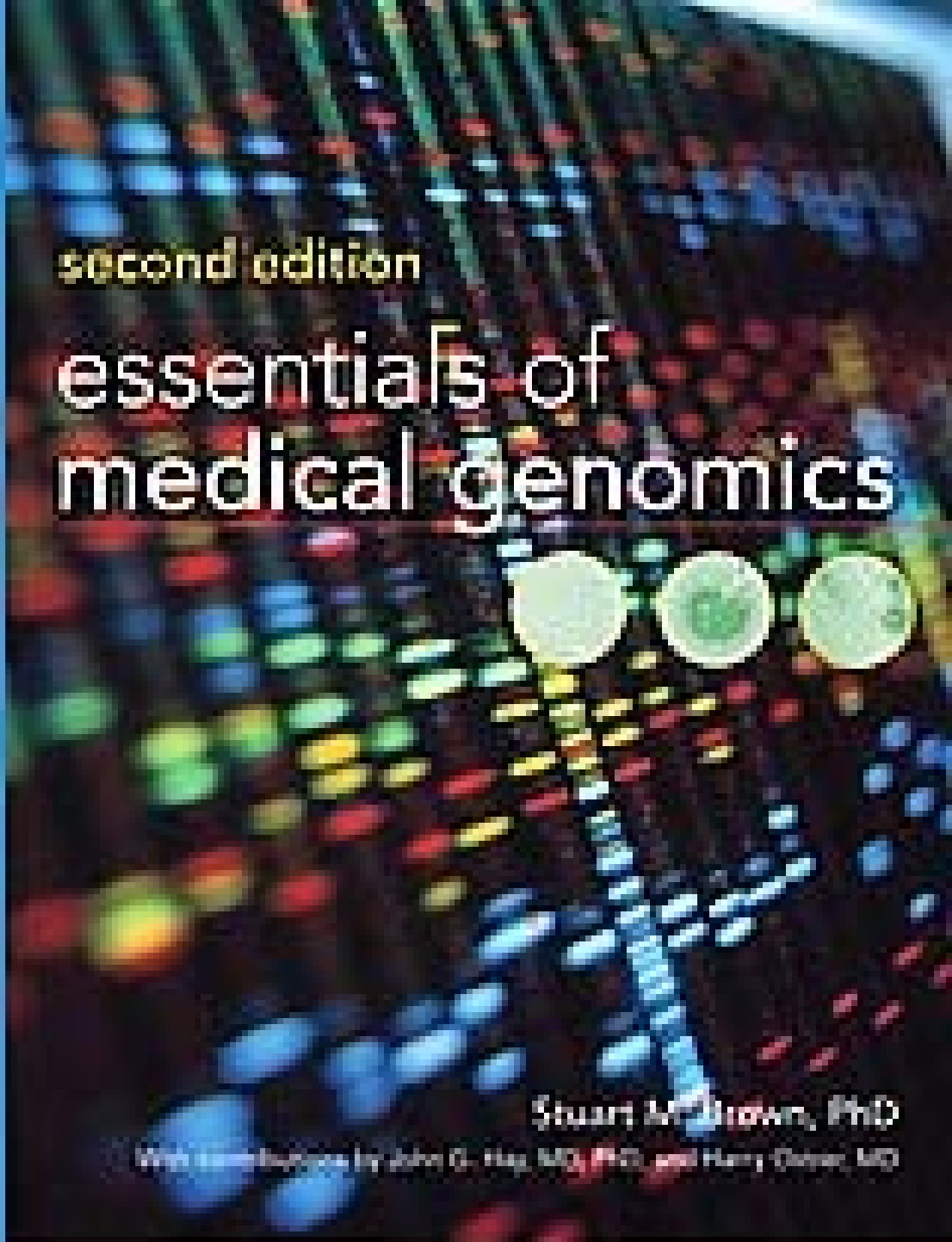
second edition

essentials of medical genomics



Stuart M. Brown, PhD

With contributions by John C. Hay, MD, PhD, and Harry Ostrer, MD

WILEY-BLACKWELL

Essentials of Medical Genomics

Essentials of Medical Genomics

Second Edition

Stuart M. Brown
NYU School of Medicine
New York, NY

with Contributions by

John G. Hay and Harry Ostrer

WILEY-BLACKWELL

A John Wiley & Sons, Inc., Publication

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical, and Medical business with Blackwell Publishing.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Essentials of medical genomics / Stuart M. Brown ; with contributions by John G. Hay and Harry Ostrer.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-14019-2 (cloth)
1. Medical genetics. 2. Genomics.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Preface

Medical genomics might seem like a rather specialized topic, of interest to just a few researchers and genetics experts, but I believe that it is a technology that is already having an impact on the practice of most primary care physicians and biomedical researchers. Genetic tests are now in use as a diagnostic aid for various types of cancer and will soon be commonplace as an aid to prescribing psychiatric drugs. Some drugs are currently under development that will require a genetic test before they can be prescribed. Consumer genomics is a new development that is disrupting the usual flow of health care information. A patient may arrive at his or her physician's office armed with a detailed report on their allelic status for thousands of genetic markers that may or may not be relevant to each health care decision. Therefore, I have tried to make this book as accessible and comprehensive as possible in order to provide a working knowledge of medical genomics both for biomedical professionals and consumers of health care.

However, writing a book about genomics is truly a Sisyphean task, since the goal of reporting current technologies is constantly receding. The book writing process takes about a year from the initial outline to page proofs, and the past year has seen exceptionally rapid progress in genomics technologies. While I was paying attention to genome tiling, copy number, and SNP chips, the revolution in Next-Generation DNA sequencing has snuck

up on me. High-throughput DNA sequencing is the kind of disruptive technology that enables new kinds of scientific research. The cost of sequencing a whole human genome has dropped from several million to about $100,000, and it is likely to be cut by tenfold again by 2009. New and unexpected applications for sequencing technology are being developed almost daily.

I have included as an appendix to this book, a short report written by Amanda K. Sarata of the Congressional Research Service. It is valuable not just because this report provides a nice summary of the science that underlies genetic testing and the related public policy issues, but because it also demonstrates the level of genetics information to which our Congressional Representatives have been exposed.

The Genetic Information Nondiscrimination Act (GINA) was finally passed by the US Congress and signed into law by President Bush in May of 2008. We all await the many social ramifications of this legislation.

STUART M. BROWN
New York

CHAPTER

1

Introduction to Molecular Genetics

The Human Genome Project is a bold undertaking to understand, at a fundamental level, all of the genetic information required to build and maintain a human being. The human **genome** is the complete information content of the human cell. This information is encoded in approximately 3.2 billion base pairs of DNA contained on 46 **chromosomes** (22 pairs of **autosomes** plus the two sex chromosomes—see Figure 1.1). The completion, in 2001, of the first draft of the human genome sequence was only the first phase of this project (Venter et al. 2001; Lander et al. 2001).

To use the metaphor of a book, the draft genome sequence gives biology all of the letters, in the correct order on the pages, but without the ability to recognize words, sentences, and punctuation, or even an understanding of the language in which the book is written. The task of making sense of all of this raw biological information falls, at least initially, to **bioinformatics** specialists who make use of computers to find the words and decode the language. The next step is to integrate all of this information into a new form of experimental biology, known as **genomics**, that

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

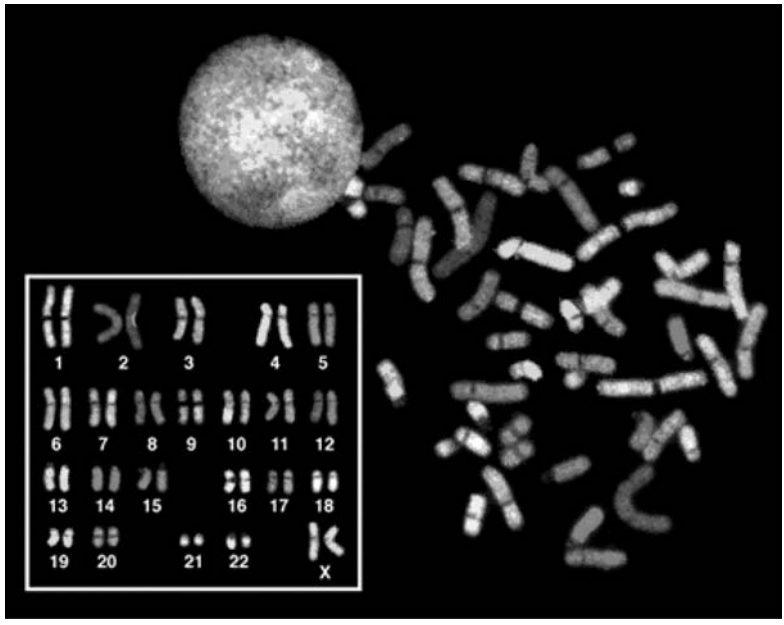


FIGURE 1.1. Human karyotype—SKY image: available at `http://www.accessexcellence.org/AB/GG/sky.gif`; credit to Chroma Technology Inc. (See insert for color representation.)

can ask meaningful questions about what is happening in very complex systems where tens of thousands of different genes and proteins are interacting simultaneously.

The primary justification for the considerable amount of money spent on sequencing the human genome (from governments and private corporations) is that this information will lead to dramatic medical advances. In fact, the first wave of new drugs and medical technologies derived from genome information is currently making its way through clinical trials and into the healthcare system. However, to effectively utilize these new advances, medical professionals need to understand something about genes and genomes. Just as it is important for physicians to understand how to Gram-stain and evaluate a culture of bacteria, even if they never actually perform this test themselves in their

medical practices, it is important to understand how DNA technologies work in order to appreciate their strengths, weaknesses, and peculiarities.

However, before we can discuss whole genomes and genomic technologies, it is necessary to understand the basics of how genes function to control biochemical processes within the cell (molecular biology) and how hereditary information is transmitted from one generation to the next (genetics).

THE PRINCIPLES OF INHERITANCE

The principles of genetics were first described by the monk Gregor Mendel in 1866 in his observations of the inheritance of traits in garden peas ["Versuche über Pflanzen-Hybriden" (Mendel 1866)]. Mendel described "differentiating characters" (*differierende Merkmale*) which may come in several forms. In his monastery garden, he made crosses between strains of garden peas that had different characters, each with two alternate forms that were easily observable, such as purple or white flower color, yellow or green seed color, smooth or wrinkled seed shape, and tall or short plant height. (These alternate forms are now known as alleles.) Then he studied the distribution of these forms in several generations of offspring from his crosses.

Mendel observed the same patterns of inheritance for each of these characters. Each strain, when bred with itself, showed no changes in any of the characters. In a cross between two strains that differ for a single character, such as pink versus white flowers, the first generation of hybrid offspring (the F_1) all resembled one parent—all pink. Mendel called this the **dominant** form of the character. After self-pollinating the F_1 plants, the second-generation plants (the F_2) showed a mixture of the two parental forms (see Figure 1.2). This is known as **segregation**. The **recessive** form that was not seen in the F_1s (white flowers) was found in one-fourth (25%) of the F_2 plants.

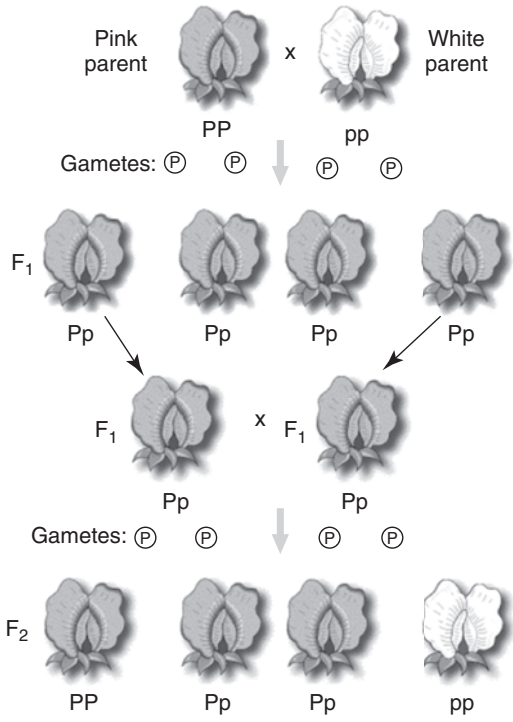


FIGURE 1.2. Mendel observed a single trait segregating over two generations. Pink and white parents have all pink F_1 progeny (heterozygous), but one-fourth of the F_2 generation are white and three-fourths are pink.

Mendel also made crosses between strains of peas that differed for two or more traits. He found that each trait was assorted independently in the progeny—there was no connection between whether an F_2 plant had the dominant or recessive form for one character and which form it carried for another character (see Figure 1.3).

Mendel created a theoretical model ("Mendel's laws of genetics") to explain his results. He proposed that each individual has two copies of the hereditary material for each character, which may determine different forms of that character. These two copies separate and are subjected to independent assortment

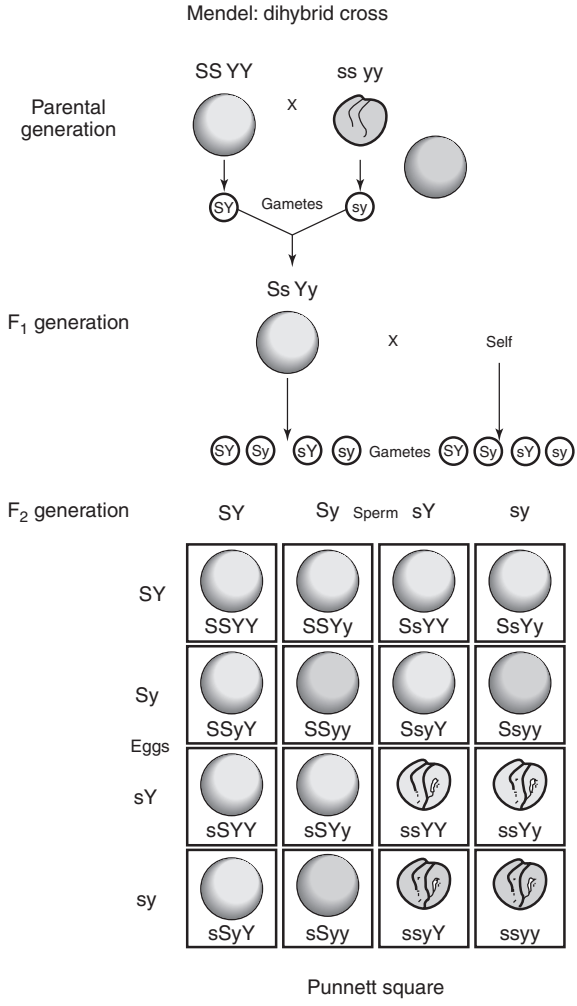


FIGURE 1.3. A cross where two independent traits are segregating (Y = yellow; S = smooth).

during the formation of gametes (sex cells). When a new individual is created by the fusion of two sex cells, the two copies from the two parents combine to produce a visible trait depending on which form is dominant and which is recessive. Mendel did not propose any physical explanation for how these traits were

passed from parent to progeny; his characters were purely abstract units of heredity.

Modern genetics has completely embraced Mendel's model with some additional detail. There may be more than two different alleles for a gene in a given population, but each individual has only two, which may be the same **(homozygous)** or different **(heterozygous)**. In some cases two different alleles combine to produce an intermediate form in heterozygous individuals, so that red and white flower alleles may combine to produce pink or type A and type B blood alleles, which in turn combine to produce the AB blood type.

GENES ARE ON CHROMOSOMES

In 1902, Walter Sutton, a microscopist, proposed that Mendel's heritable characters resided on the chromosomes which he observed inside the cell nucleus (see Figure 1.4). Sutton observed that "the association of paternal and maternal chromosomes in

FIGURE 1.4. Anaphase chromosomes in a dividing lily cell. (See insert for color representation.)

pairs and their subsequent separation during cell division . . . may constitute the physical basis of the Mendelian law of heredity" (Sutton 1903).

In 1909, the Danish botanist Wilhelm Johanssen coined the term "gene" to describe Mendel's heritable characters. In 1910, Thomas Hunt Morgan found that a trait for white eye color was located on the X chromosome of the fruitfly and was inherited together with a factor that determines sex (Morgan 1910). A number of subsequent studies by Morgan (1919) and others showed that each gene for a particular trait was located at a specific spot or **locus** on a chromosome in all individuals of a species. The chromosome was perceived as a linear organization of genes, like beads on a string. Throughout the early part of the twentieth century, a gene was considered to be a single, fundamental, indivisible unit of heredity, in much the same way as an atom was considered to be the fundamental unit of matter.

Each individual has two copies of each type of chromosome, having received one copy from each parent. The two copies of each chromosome in the parent are randomly divided into the sex cells (sperm and egg) in a process called segregation. It is possible to observe the segregation of chromosomes during **meiosis** using only a moderately powerful microscope. It is an aesthetically satisfying triumph of biology that this observed segregation of chromosomes in cells exactly corresponds to the segregation of traits that Mendel observed in his peas.

Recombination and Linkage

In the early twentieth century, Mendel's concepts of inherited characters were broadly adopted by practical plant and animal breeders as well as experimental geneticists. It rapidly became clear that Mendel's experiments represented an oversimplified view of inheritance. He must have intentionally chosen characters in his peas that were inherited independently. In the breeding

experiments where many traits differ between parents, it is commonly observed that progeny inherit pairs or groups of traits together from one parent far more frequently than would be expected by chance alone. This observation fits nicely into the chromosome model of inheritance—if two genes are located on the same chromosome, then they will be inherited together when that chromosome segregates into a gamete, and that gamete becomes part of a new individual.

However, it was also observed that "linked" genes do occasionally separate. A theory of **recombination** was developed to explain these events. During the process of meiosis, it was proposed that the homologous chromosome pairs line up and exchange segments in a process called **crossing over**. This theory was supported by microscopic evidence of X-shaped structures called **chiasmata** forming between paired homologous chromosomes in meiotic cells (see Figure 1.5).

If a parent cell contains two different alleles for two different genes, then after the crossover, the chromosomes will contain new combinations of alleles. For example, if one chromosome contains alleles A and B for two genes, and the other chromosome contains alleles a and b, then without crossovers, all progeny must inherit a chromosome from that parent with either an A–B or an a–b allele combination. If a crossover occurs between the two genes, then the resulting chromosomes will contain the A–b and a–B allele combinations (see Figure 1.6).

Morgan, continuing his work with fruitflies, demonstrated that the chance of a crossover occurring between any two linked

FIGURE 1.5. Chiasmata visible in electron micrograph of meiotic chromosome.

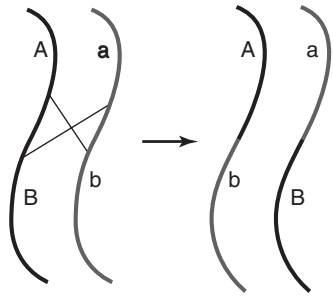


FIGURE 1.6. Schematic diagram of a single crossover between a chromosome with A–B alleles and a chromosome with a–b alleles to form A–b and a–B recombinant chromosomes. (See insert for color representation.)

genes is proportional to the distance between them on the chromosome. Therefore, by counting the frequency of crossovers between alleles of a given pair of genes, it is possible to create genetic maps of chromosomes. Morgan was awarded the 1933 Nobel Prize in Medicine for this work. In fact, it is generally observed that on average there is more than one crossover between every pair of homologous chromosomes in every meiosis, so that two genes located on opposite ends of a chromosome do not appear linked at all. On the other hand, alleles of genes that are located very close together are very rarely separated by recombination (see Figure 1.7).

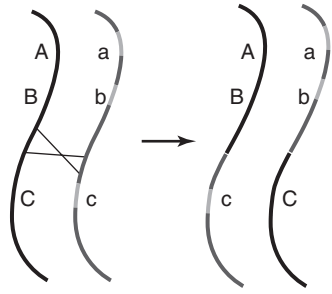


FIGURE 1.7. Genes A and B are tightly linked so that they are not separated by recombination, but gene C is farther away. After recombination occurs in some meiotic cells, gametes are produced with allele combinations ABC, abc, ABc, and abC. (See insert for color representation.)

The relationship between the frequency of recombination between alleles and the distance between genes on a chromosome has been used to construct genetic maps for many different organisms, including humans. It has been a fundamental assumption of genetics for almost a hundred years that recombinations occur randomly along the chromosome at any location, even within genes. However, more recent data from DNA sequencing of genes in human populations suggest that there are recombination hotspots and regions where recombination almost never occurs. This creates groups of alleles from neighboring genes on a chromosome, known as **haplotypes**, that remain linked together across hundreds of generations.

Genes Encode Proteins

Beadle and Tatum (1941) showed that a single mutation, caused by exposing the fungus *Neurospora crassa* to X rays, destroyed the function of a single enzyme, which interrupted a biochemical pathway at a specific step due to the loss of function of a particular enzyme. This mutation segregated among the progeny exactly as Mendel's traits did in peas. The X-ray-induced damage to a specific region of one chromosome destroyed the instructions for the synthesis of a specific enzyme. Thus a gene is a spot on a chromosome that codes for a single enzyme. In subsequent years, a number of other researchers broadened this concept by showing that genes code for all types of proteins, not just enzymes, leading to the **one gene–one protein** model, which is the core of modern molecular biology. Beadle and Tatum shared the 1958 Nobel Prize in Medicine.

Genes Are Made of DNA

The next step in understanding the nature of the gene was to dissect the chemical structure of the chromosome. Crude

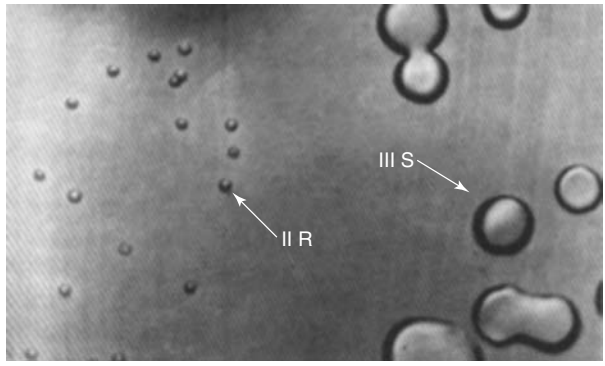


FIGURE 1.8. Transforming experiment: rough (II R) and smooth (III S) *Streptococcus pneumoniae* cells. (From Avery et al., 1944.)

biochemical purification had shown that chromosomes are composed of both protein and DNA. Avery et al. (1944) conducted the classic experiment on the "transforming principle." They found that DNA purified from a lethal S (smooth) form of *Streptococcus pneumoniae* could transform a harmless R (rough) strain into the S form (see Figure 1.8). Treatment of the DNA with protease to destroy all of the protein had no effect, but treatment with DNA-degrading enzymes blocked the transformation. Therefore, the information that transforms the bacteria from R to S must be contained in the DNA (McCarty 1985).

Hershey and Chase (1952) confirmed the role of DNA with their classic "blender experiment" on bacteriophage viruses. The phage were radioactively labeled with either ^{35}S in their proteins or ^{32}P in their DNA. They used a blender to interrupt the process of infection of *Escherichia coli* bacteria by the phage. Then they separated the phage from the infected bacteria by centrifugation and collected the phage and the bacteria separately. They observed that the ^{35}S-labeled protein remained with the phage while the ^{32}P-labeled DNA was found inside the infected bacteria (see Figure 1.9). This proved that it is the DNA portion of the virus that enters the bacteria and contains the genetic instructions

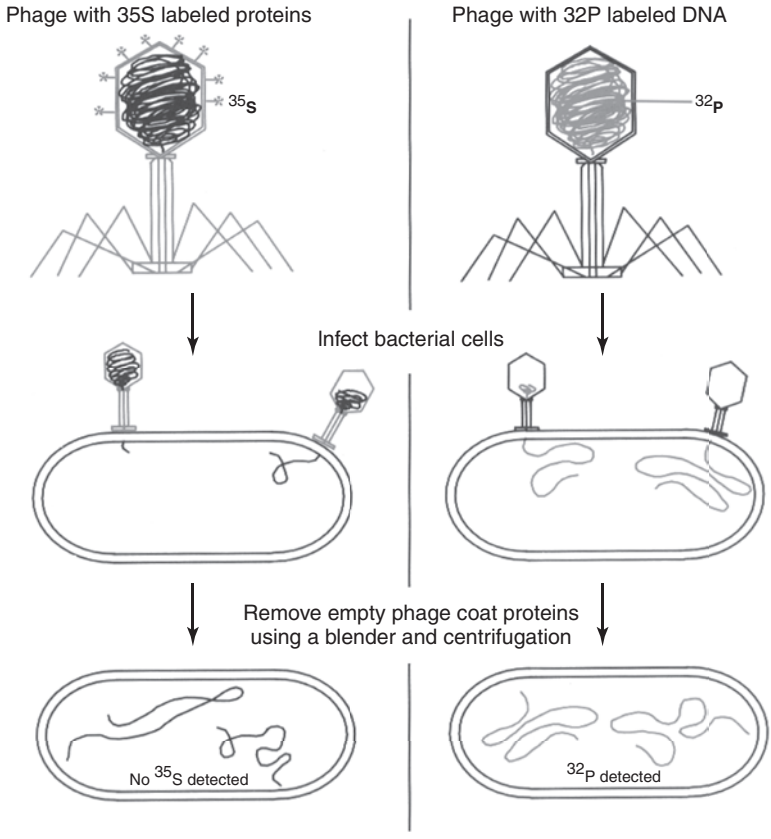


FIGURE 1.9. Hershey–Chase blender experiment. *Escherichia coli* bacteria are infected with phage with ^{35}S-labeled proteins or ^{32}P-labeled DNA. After removing the phage with a blender, the ^{32}P-labeled DNA but not the ^{35}S-labeled protein, is found inside the bacteria. (From Micklos and Freyer, *DNA Science*, Cold Spring Harbor Press, 1990.)

for producing new phage, not the proteins, which remain outside. Hershey was awarded the 1969 Nobel Prize for this work.

DNA Structure

Now it was clear that genes are made of DNA, but how does this chemically simple molecule contain so much information?

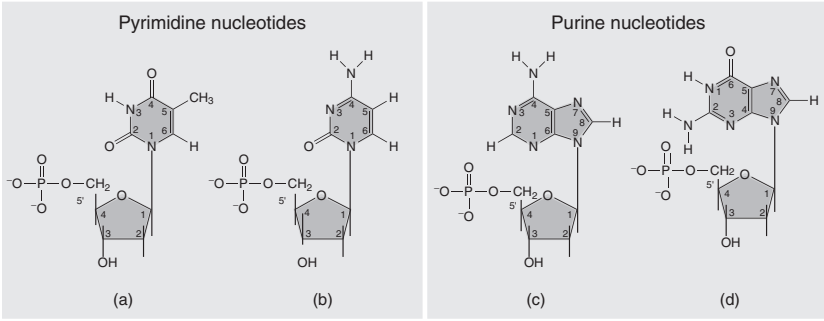


FIGURE 1.10. Chemical structures of the four DNA bases: (a) deoxythymidine monophosphate (dTMP); (b) deoxycytidine monophosphate (dCMP); (c) deoxyadenosine monophosphate (dAMP); (d) deoxyguanosine monophosphate (dGMP).

DNA is a long polymer molecule that contains a mixture of four different chemical subunits: adenine, cytosine, guanosine, and thymine (abbreviated as A, C, G, and T). These subunits, known as **nucleotide bases**, have similar two-part chemical structures that contain a deoxyribose sugar and a nitrogen ring (see Figure 1.10), hence the name deoxyribose nucleic acid. The real challenge is to understand how the nucleotides fit together in a way that can contain a lot of information.

Chargaff (1950) discovered that there was a consistent one-to-one ratio of adenine to thymine and guanine to cytosine in any sample of DNA from any organism. In 1951, Linus Pauling and R. B. Corey described the α-helical structure of a protein (Pauling and Corey 1951). Shortly thereafter, Rosalind Franklin (Sayre 1975) provided X-ray crystallographic images of DNA to James Watson and Francis Crick (see Figure 1.11); this form of DNA was very similar to the α-helix described by Pauling. Watson and Crick's crucial insight (1953) was to realize that DNA formed a double helix with complementary bonds between adenine–thymine and guanine–cytosine pairs.

The Wastson–Crick model of the DNA structure resembles a twisted ladder. The two sides of the ladder are formed by strong

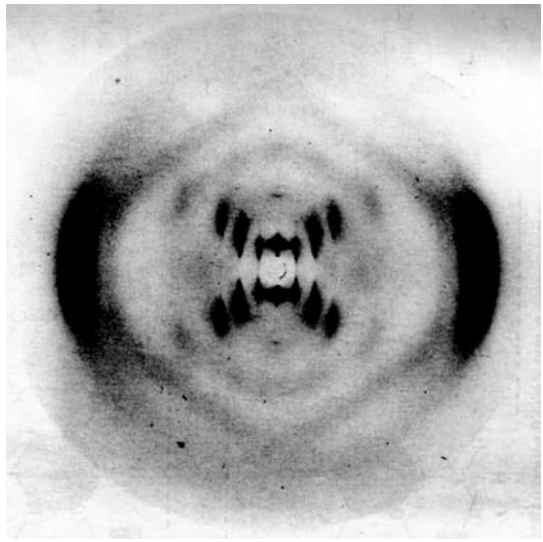


FIGURE 1.11. Rosalind Franklin's X-ray diffraction image of DNA.

covalent bonds between the phosphate on the 5′ carbon of one deoxyribose sugar and the methyl side groups of the 3′ carbon of the next (a phosphodiester bond). Thus, the deoxyribose sugar part of each nucleotide is bonded to the one above and below it, forming a chain that forms the backbone of the DNA molecule (see Figure 1.12). The phosphate-to-methyl linkage of the deoxyribose sugars give the DNA chain a direction or polarity, generally referred to as **5′ to 3′**. Each DNA molecule contains two parallel chains that run in opposite directions forming the sides of the ladder.

The rungs of the ladder are formed by weaker hydrogen bonds between the nitrogen ring parts of pairs of nucleotide bases. There are only two types of base pair bonds: adenine bonds with thymine, and guanine bonds with cytosine. The order of nucleotide bases on both sides of the ladder always reflects this complementary base pairing—so that wherever there is an A on one side, there is always a T on the other side, and vice versa.

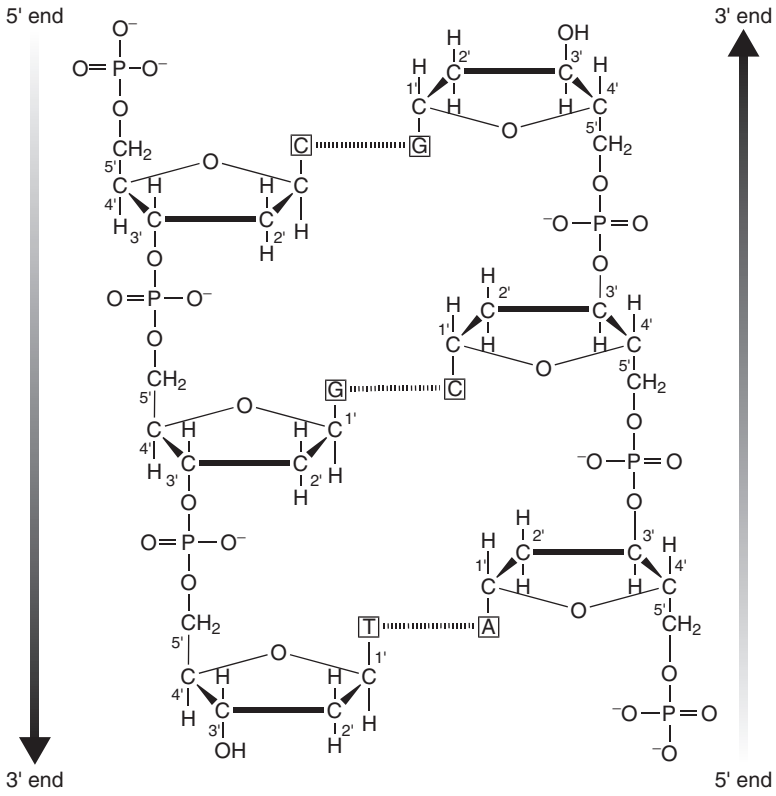


FIGURE 1.12. DNA phosphate bonds.

Since the A–T and G–C units always occur together, they are often referred to as **base pairs**. The G–C base pair has three hydrogen bonds, while the A–T pair only has two (see Figure 1.13), so the bonds between G–C bases are more stable at high temperatures than are A–T bonds. The nucleotide bases are strung together on the polydeoxyribose backbone-like beads on a string. It is the particular order of the four different bases as they occur along the string that contains all of the biological information.

Watson and Crick realized that this model of DNA structure contains many implications (see Figure 1.14). First, the two strands of the double helix are complementary, not identical.

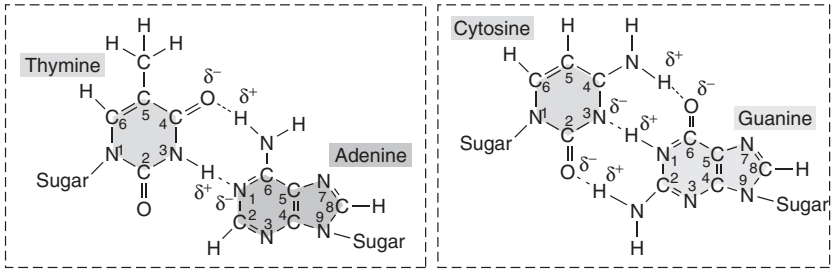


FIGURE 1.13. DNA hydrogen bonds in (a) A–T and (b) G–C base pairs.

Thus one strand can serve as a template for the synthesis of a new copy of the other strand—a T is added to the new strand wherever there is an A, a G for each C, and so on—perfectly retaining the information in the original double strand. In 1953, in a single-page paper in the journal Nature, they said, with a mastery of understatement: "It has not escaped our attention that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material" (Watson and Crick 1953).

So, in one tidy theory, the chemical structure of DNA explains how genetic information is stored on the chromosome and how it is passed on when cells divide. That is why Watson and Crick won the 1962 Nobel Prize (shared with Maurice Wilkins).

If the two complementary strands of a DNA molecule are separated in the laboratory by boiling (known as **denaturing** the DNA), then they can find each other and again pair up, by re-forming the complementary A–T and C–G hydrogen bonds (**annealing**). Bits of single-stranded DNA from different genes do not have perfectly complementary sequences, so they will not pair up in solution. This process of separating and rematching complementary pieces of DNA, known as **DNA hybridization**, is a fundamental principle behind many different molecular biology technologies.

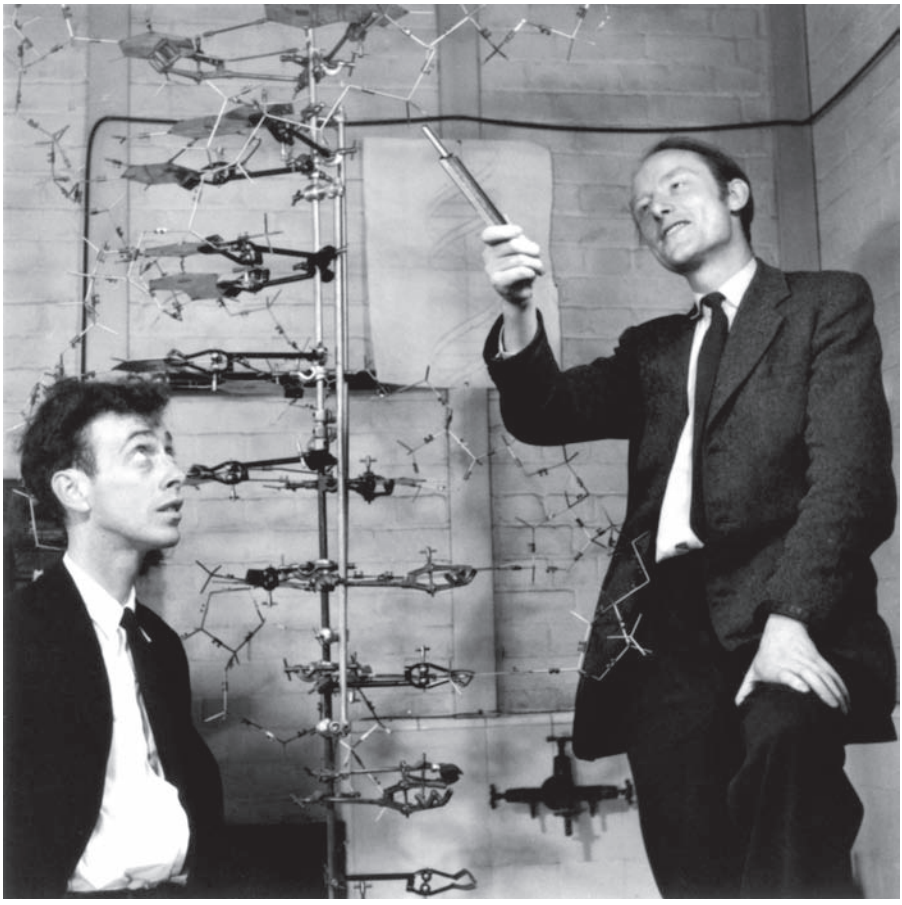


FIGURE 1.14. James Watson (left) and Francis Crick demonstrate their model of the DNA double helix. (From Watson J. 1968. *The Double Helix*, p 125. Atheneum, New York. Courtesy of Cold Spring Harbor Laboratory Archives.)

FIGURE 1.15. The central dogma of molecular biology (as described by Crick in 1957): DNA is transcribed into RNA, which is translated into protein.

THE CENTRAL DOGMA

Crick followed up in 1957 with a theoretical framework for the flow of genetic information in biological systems (Crick 1957). His theory, which has come to be known as the "Central Dogma" of molecular biology, is that DNA codes for genes in a strictly linear fashion—a series of DNA bases corresponding to a series of amino acids in a protein. DNA is copied into RNA, which serves as a template for protein synthesis. This leads to a nice, neat conceptual diagram of the flow of genetic information within a cell: DNA is copied to more DNA in a process known as **replication**, and DNA is **transcribed** into RNA, which is then **translated** into protein (see Figure 1.15).

DNA REPLICATION

Every ordinary cell (**somatic cell**) in an organism has a complete copy of that organism's genome. In mammals and other **diploid** organisms, that genome contains two copies of every chromosome, one from each parent. As an organism grows, cells divide by a process known as **mitosis**. Before a cell can divide, it must make a complete copy of its genome so that each daughter cell will receive a full set of chromosomes. All of the DNA is **replicated** by a process that makes use of the complementary nature of the base pairs in the double helix.

In DNA replication, the complementary base pairs of the two strands of the DNA helix partially separate and new copies of both strands are made simultaneously. A **DNA polymerase** enzyme attaches to the single-stranded DNA and synthesizes new strands by joining free DNA nucleotides into a growing chain that is exactly complementary to the template strand (see Figure 1.16). In addition to a template strand and free nucleotides,

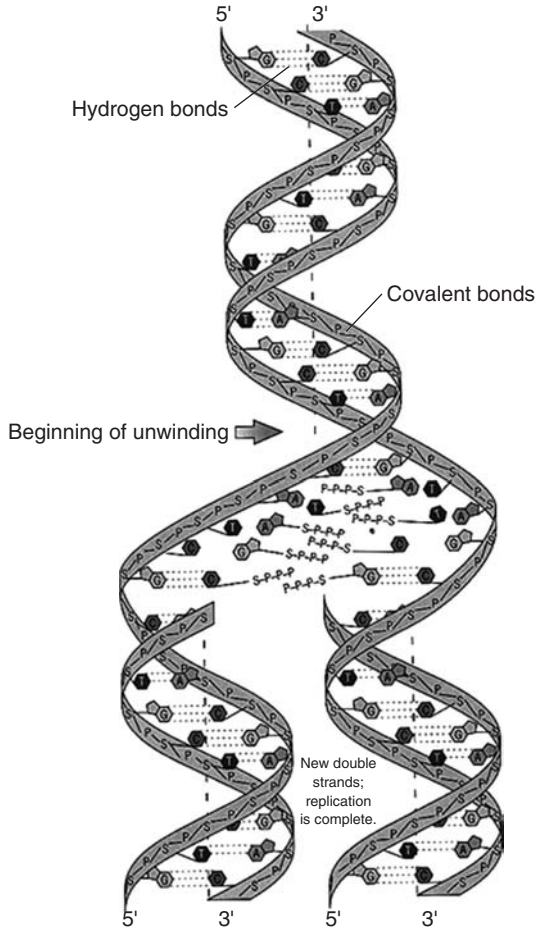


FIGURE 1.16. Diagram of DNA replication showing synthesis of two complementary strands at a replication fork.

the DNA polymerase also requires a primer—a short piece of DNA that is complementary to the template. The primer binds to its complementary spot on the template to form the start of the new strand, which is then extended by the polymerase, adding one complementary base at a time, moving in the 5′→3′ direction. In natural DNA replication, the primer binds to specific spots on the chromosome known as the **origin of replication**.

This semiconservative replication process was demonstrated quite eloquently by the famous 1958 experiment of Meselson and Stahl. They grew bacteria in a solution that contained free DNA nucleotides that contained heavy ^{15}N atoms. After many generations, the bacterial DNA contained heavy atoms throughout. Then the bacteria were transferred to a growth medium that contained normal nucleotides. After one generation, all bacterial cells had DNA with half heavy and half light nitrogen atoms. After two generations, half of the bacteria had DNA with normal nitrogen and the other half had one heavy and one light DNA strand (Meselson and Stahl 1958). After every cell division, the two daughter cells both have chromosomes made up of DNA molecules that have one strand from the parent cell and the other strand that has been newly synthesized. This method of semiconservative DNA replication is common to all forms of life on earth from bacteria to humans.

This mechanism of DNA replication has been exploited in modern DNA sequencing biochemistry, which often uses DNA polymerase from bacteria or other organisms to copy human (or any other) DNA. Key aspects of the replication process to keep in mind are that the DNA is copied linearly one base at a time from a specific starting point (origin), which is matched by a short primer of complementary sequence. The primer is extended by the reaction as new nucleotides are added, so that the primer becomes part of the newly synthesized complementary strand.

TRANSCRIPTION

The DNA in the chromosomes contains genes that are instructions for the manufacture of proteins, which in turn control all of the metabolic activities of the cell. In order for the cell to use these instructions, the genetic information must be moved from the chromosomes inside the nucleus out to the cytoplasm where proteins are manufactured. This information transfer is done using messenger RNA (**mRNA**) as an intermediary molecule. RNA (ribose nucleic acid) is a polymer of nucleotides, chemically very similar to DNA, but with three distinct differences: (1) RNA is a single-stranded molecule, so it does not form a double helix; (2) RNA nucleotides contain ribose rather than deoxyribose sugars; and (3) RNA uses uracil in place of thymine, so the common abbreviations for RNA bases are A, U, G, and C. As a result of these chemical differences, RNA is much less stable in the cell. In fact, the average RNA molecule has a lifespan that can be measured in minutes while DNA can be recovered from biological materials that are many thousands of years old.

The transcription of DNA into mRNA is similar to DNA replication. A single strand of DNA is copied one base at a time into a complementary strand of RNA. The enzyme RNA polymerase catalyzes the incorporation of free RNA nucleotides into the growing chain (see Figure 1.17). However, not all of the DNA is copied into RNA—only those portions that encode genes. In eukaryotic cells, only a small fraction of the total DNA is actually used to encode genes. Furthermore, not all genes are transcribed into mRNA in equal amounts in all cells. The process of transcription is tightly regulated so that only those mRNAs are manufactured that encode the proteins that are currently needed by each cell. This overall process is known as **gene expression**. Understanding the process of gene expression and how it differs in different types of cells or under different conditions is one of the fundamental questions driving the technologies of genomics.

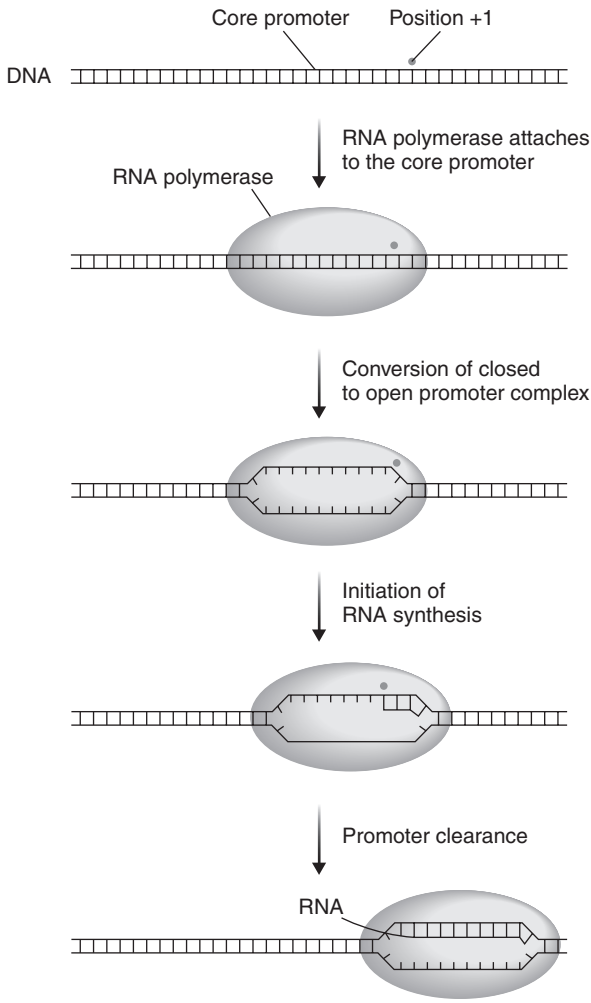


FIGURE 1.17. RNA polymerase II attaches to the promoter and begins transcription.

The primary control of transcription takes place in a region of DNA known as the **promoter**, which occupies a position "upstream" (in the 5′ direction) from the part of a gene that will be transcribed into RNA (the **protein-coding region** of the gene). A huge variety of different proteins recognize specific DNA

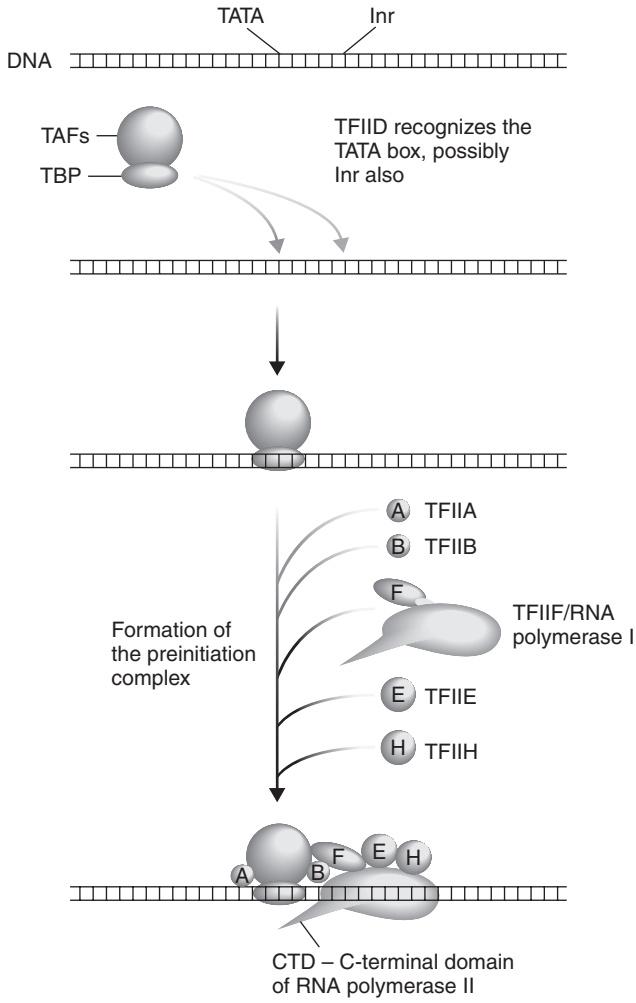


FIGURE 1.18. RNA polymerase II is actually a complex structure composed of many individual proteins.

sequences in this promoter region and bind to the DNA and either assist or block the binding of the RNA polymerase enzyme (see Figure 1.18). These DNA binding proteins work in concert to provide very fine-grained control of the expression of each gene depending on the type of cell, where it is located in the body, its

current metabolic condition, and its responses to external signals from the environment or from other cells.

In fact, the factors governing the assembly of the set of proteins involved in regulating DNA transcription is much more complicated than the sum of a set of DNA sequences neatly located in a promoter region 5′ to the coding sequence of a gene. In addition to the double helix, DNA has tertiary structures that involve twists and supercoils as well as winding around histone proteins. These three-dimensional (3D) structures can bring distant regions of a DNA molecule into close proximity, so that proteins bound to these sites may interact with the proteins bound to the promoter region. These distant sites on the DNA that may effect transcription are known as **enhancers**. The total set of DNA binding proteins that interact with promoters and enhancers are known as **transcription factors**, and the specific DNA sequences to which they bind are called **transcription factor binding sites**.

RNA Processing

Once a gene is transcribed into RNA, the RNA molecule undergoes a number of processing steps before it is translated into protein. First a 5′ cap is added, then a polyadenine tail is added at the 3′ end. In addition, eukaryotic genes are broken up into protein coding **exon** regions separated by non-protein coding **introns**, which are spliced out. This splicing is sequence-specific and highly precise, so that the final product contains the exact mRNA sequence that codes for a specific protein with not a single base added or lost (see Figure 1.19).

Each of these posttranscriptional processes may serve as a point of regulation for gene expression. Capping, polyadenelation, and/or splicing may be blocked, or incorrect splicing may be promoted under specific metabolic or developmental conditions. In addition, splicing may be altered in order to produce different mRNA molecules.

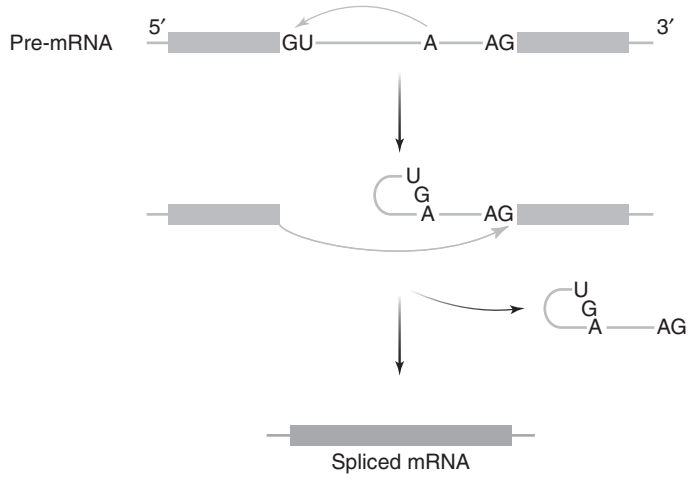


FIGURE 1.19. Model of intron splicing to form a mature mRNA from a pre-mRNA transcript.

ALTERNATIVE SPLICING

Each gene does not encode a single protein, as was originally suggested by the studies of *Neurospora* enzymes by Beadle and Tatum (1941). In many cases, there are several alternate forms of final spliced mRNA that can be produced from a single pre-mRNA transcript—potentially leading to proteins with different biological activities. In fact, current estimates suggest that most genes have multiple alternate splice forms. Alternate splicing my involve the failure to recognize a splice site, causing an intron to be left in, or an exon to be left out. Alternate splice sites may occur anywhere, either inside exons or introns, so that the alternate forms of the final mRNAs may be longer or shorter, contain more or fewer exons, or portions of exons (see Figure 1.20). Thus, each different splice form produced from a gene is a unique type of mRNA, which has the potential to produce a protein with different biochemical properties.

It is not clear how alternative splicing is controlled. The signals that govern RNA splicing may not be perfectly effective, or

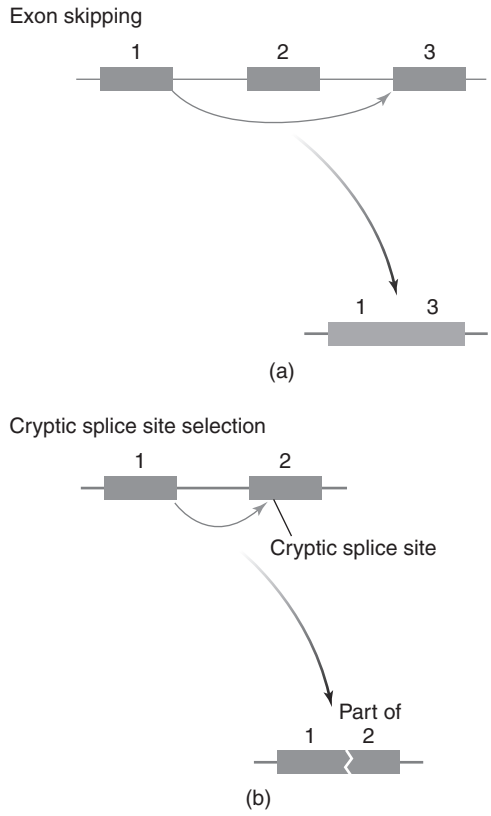


FIGURE 1.20. Two forms of alternative splicing: (a) exon skipping; (b) cryptic splice site selection.

RNA splicing may be actively used as a form of gene regulation. It is entirely possible for the products of other genes to interact with RNA splicing factors—perhaps in conjunction with external signals—to alter RNA splicing patterns for specific genes. The net result will be many different forms of mRNA, some produced only under specific circumstances of development, tissue specificity, or environmental stimuli. Thus, under some conditions a different protein with an added (or removed) functional domain will be produced from a gene, resulting in different protein function.

"Alternative splicing increases protein diversity by allowing multiple, sometimes functionally distinct proteins to be encoded by the same gene" (Sorek and Amitai 2001). The totality of all of these different mRNAs is being called the "transcriptome," which is certainly many times more complex than the genome. The relative levels of alternate splice forms for a single gene may have substantial medical significance. For example, there are 60 kinase enzymes that have alternate splice forms that do not include their catalytic domains, creating proteins that may function as competitive inhibitors of the full-length proteins (Sorek and Amitai 2001).

TRANSLATION

In order for a gene to be expressed, the mRNA must be translated into protein. This theory behind this process was encapsulated quite neatly in 1957 by Crick's diagram of the Central Dogma, but the details of the information flow from DNA to mRNA to protein took another decade to work out. It was immediately clear that the cell must solve several different problems of information storage and transmission. Huge amounts of information must be stored in the simple 4-letter code of DNA, it must be translated into the quite different 20-letter code of amino acids, and a great deal of punctuation and regulatory information must also be accounted for. The problem of encoding 20 different amino acids in the 4-letter DNA/RNA alphabet intrigued information scientists, and physicists as well as biologists and many ingenious incorrect answers were proposed. The actual solution to this problem was worked out with brute-force biochemistry by Har Gobind Khorana (Soll et al. 1965) and Marshall W. Nirenberg (Nirenberg 1965) by creating an *in vitro* (test tube) system where pure pieces of RNA would be translated into protein. They then fed the system with RNA molecules of very simple sequence and analyzed the proteins that were produced. With several years of

Universal Genetic Code

TTT phe F	TCT ser S	TAT tyr Y	TGT cys C
TTC phe F	TCC ser S	TAC tyr Y	TGC cys C
TTA leu L	TCA ser S	TAA OCH Z	TGA OPA Z
TTG leu L	TCG ser S	TAG AMB Z	TGG trp W
CTT leu L	CCT pro P	CAT his H	CGT arg R
CTC leu L	CCC pro P	CAC his H	CGC arg R
CTA leu L	CCA pro P	CAA gln Q	CGA arg R
CTG leu L	CCG pro P	CAG gln Q	CGG arg R
ATT ile I	ACT thr T	AAT asn N	AGT ser S
ATC ile I	ACC thr T	AAC asn N	AGC ser S
ATA ile I	ACA thr T	AAA lys K	AGA arg R
ATG met M	ACG thr T	AAG lys K	AGG arg R
GTT val V	GCT ala A	GAT asp D	GGT gly G
GTC val V	GCC ala A	GAC asp D	GGC gly G
GTA val V	GCA ala A	GAA glu E	GGA gly G
GTG val V	GCG ala A	GAG glu E	GGG gly G

FIGURE 1.21. Translation table for the eukaryotic nuclear genetic code.

effort (1961–1965), they defined a code of 64 three-letter RNA **codons** that corresponded to the 20 amino acids (with redundant codons for most of the amino acids) and 3 "stop" codons that caused the end of protein synthesis (see Figure 1.21). Also in 1965, Robert W. Holley established the exact chemical structure of **tRNA** (**transfer RNA**), the adapter molecules that carried each amino acid to its corresponding 3-base codon on the mRNA (Holley 1965). There is one specific type of **tRNA** that binds each type of amino acid, but each tRNA has an **anti-codon** which can bond to several different mRNA codons. Holley, Khorana, and Nirenberg shared the 1968 Nobel Prize in Physiology or Medicine for this work.

The translation process is catalyzed by a complex molecular machine called a **ribosome**, which is composed of both protein and **rRNA** (ribosomal RNA) elements. Proteins are assembled from free amino acids in the cytoplasm that are carried to the site of protein synthesis on the ribosome by the tRNAs. The tRNAs contain an anticodon region that matches the three-nucleotide codons on the mRNA. As each tRNA attaches to the anticodon,

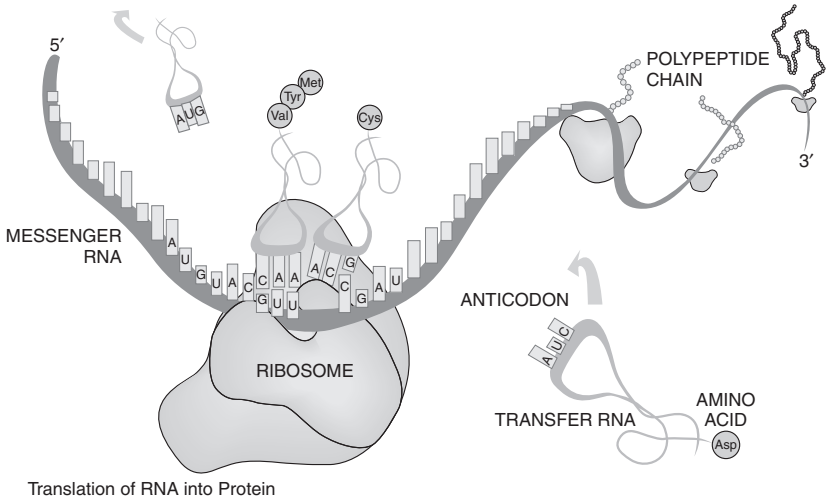


FIGURE 1.22. A diagram of the ribosome interacting with tRNAs as it translates an mRNA into a polypeptide chain.

the amino acid that it carries forms a bond with the growing polypeptide chain; then the tRNA is released and the ribosome moves down the mRNA to the next codon. When the ribosome reaches a stop codon, the chain of amino acids is released as a complete polypeptide (see Figure 1.22).

REFERENCES

Avery OT, MacLeod CM, McCarty M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J Exp Med* **79**:137–158.

Beadle GW, Tatum EL. 1941. Genetic control of biochemical reactions in Neurospora. *Proc Natl Acad Sci USA* **27**:499–506.

Chargaff E. 1950. Chemical specificity of nucleic acids and mechanisms of their enzymatic degradation. *Experientia* **6**:201–209.

Crick FHC. 1957. Nucleic acids. *Sci Am* **197**:188–200.

Hershey AD, Chase M. 1952. Independent functions of viral proteins and nucleic acid in growth of bacteriophage. *J. Gen Physiology* **36**:39–56.

Holley RW. 1965. Structure of an alanine transfer ribonucleic acid. *JAMA* **194**:868–871.

Lander ES et. al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.

McCarty, M. 1985. *The Transforming Principle: Discovering that Genes Are Made of DNA*. Norton, New York.

Mendel, G. 1866. Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines, Abhandlungen, Brünn* **4**:3–47.

Meselson M, Stahl FW. 1958. The replication of DNA in Escherichia coli. *Proc Natl Acad Sci USA* **44**:671–682.

Morgan TH. 1910. Sex-limited inheritance in Drosophila. *Science* **32**:120–122.

Morgan TH. 1919. *The Physical Basis of Heredity*. Lippincott, Philadelphia.

Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, Rottmann F, O'Neal C. 1965. RNA codewords and protein synthesis. VII. On the general nature of the RNA code. *Proc Natl Acad Sci USA* **53**:1161–1168.

Pauling L, Corey R. 1951. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proc Natl Acad Sci USA* **37**:235–240.

Sayre A. 1975. *Rosalind Franklin and DNA*. Norton, New York.

Soll D, Ohtsuka E, Iones DS, Lohrmann R, Hayatsu H, Nishimura S, Khorana HG. 1965. Studies on polynucleottides. XLIX. Stimulation of the binding of aminoacyl-sRNAs to ribosomes by ribotrinucleotides and a survey of codon assignments for 20 amino acids. *Proc Natl Acad Sci USA* **54**:1378–1385.

Sorek R, Amitai M. 2001. Piecing together the significance of splicing. *Nat Biotechnol* **19**:196.

Sutton W. 1903. The chromosomes in heredity. *Biol Bull* **4**:231–251.

Venter JC et al. 2001. The sequence of the human genome. *Science* **291**:1304–1351.

Watson JD, Crick FHC. 1953. A structure for deoxyribose nucleic acid. *Nature* **171**:737.

CHAPTER

2

Molecular Biology Technology

Cut, Copy, and Paste

Genomics technology is all about the application of scaling up, automation, and massively parallel systems to molecular biology. However, in order to understand these new high-throughput technologies, it is necessary to understand the basic molecular biology techniques on which they are based. If we extend the metaphor of the genomic DNA sequence as a book, introduced in the previous chapter, then molecular biology provides the **Cut**, **Paste**, and **Copy** operations needed to edit the text. Traditional molecular biology works on one gene (word) at a time, while genomics technologies allow operations on all of the genes at once (global search and replace).

Restriction Enzymes

It turns out that cutting DNA at specific positions is quite simple. Bacteria do it all the time using proteins called **restriction endonuclease enzymes**, which were discovered by Herb Boyer in

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

1969 (Boyer and Roulland-Dussoix 1969; Roulland-Dussoix and Boyer 1969). It is possible to grow almost any common strain of bacteria in a flask of nutrient broth, collect all the bacterial cells, grind them up, and extract active restriction enzymes from the resulting goop of cellular proteins. A number of companies now specialize in purifying these enzymes and selling them to molecular biologists (at surprisingly inexpensive prices, considering their remarkable powers).

Each strain of bacteria makes its own characteristic restriction enzymes that cut DNA at different specific sequences, known as **recognition sites**, which are typically 4, 6, or 8 bases long. Not surprisingly, bacteria protect themselves from their own restriction enzymes by avoiding the use of the recognition sequence in their own DNA and/or by the action of sequence-specific DNA methylase enzymes, which modify the DNA at the recognition site so that it cannot be cut.

In addition to having a specific DNA sequence that it recognizes as its cleavage site, each type of restriction enzyme cuts the DNA in a specific pattern, leaving a characteristic shape at the free ends. Most restriction enzymes cut the two strands of the DNA double helix unevenly, leaving a few bases of overhang on one strand or the other (see Figure 2.1). The overhanging bases from the two freshly cut ends are complementary in sequence, so under the right conditions, they can pair up to form new hydrogen bonds. These are known as "sticky ends."

Another bacterial enzyme, known as **DNA ligase**, can recreate the phosphate bonds of the DNA backbone across a pair of rejoined sticky ends, effectively pasting together a new DNA molecule. With enzymes that can cut and paste, it is not a tremendous leap to the concept of cutting two different pieces of DNA with the same restriction enzyme, then swapping fragments and splicing them with ligase in a new combination (see Figure 2.2). Paul Berg made the first artificial recombination in 1972 between

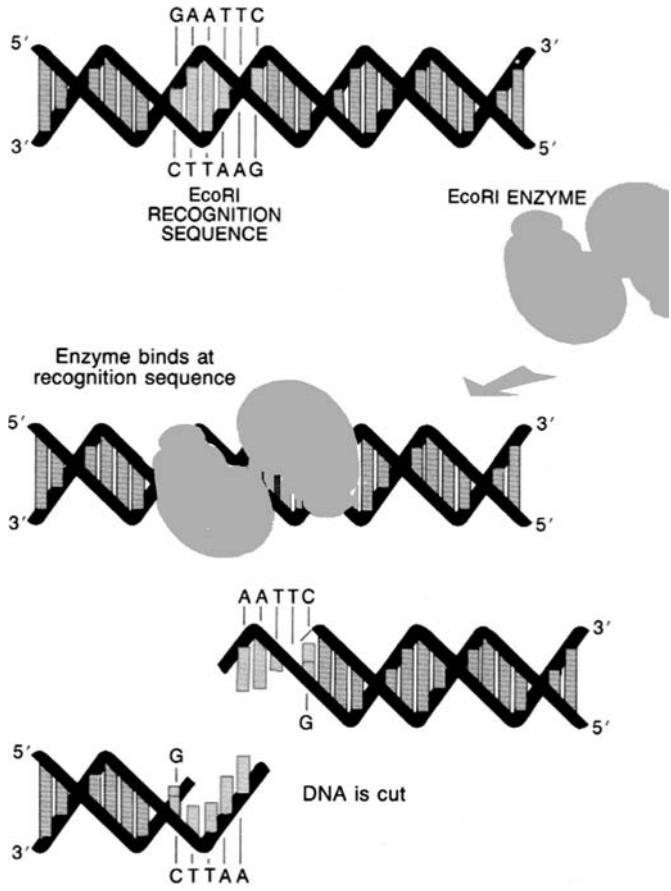


FIGURE 2.1. The EcoRI restriction enzyme produces sticky ends. (Art concept developed by Lisa Shoemaker.)

a piece of SV40 virus and a piece of *Escherichia coli* chromosomal DNA (Jackson et al. 1972).

DNA Cloning Is Copying

Cloning is a process of making identical copies by biological duplication. DNA cloning makes use of bacteria as a host to grow unlimited copies of a single piece of DNA. Bacteria have a single circular chromosome, but Stanley Cohen discovered in 1968

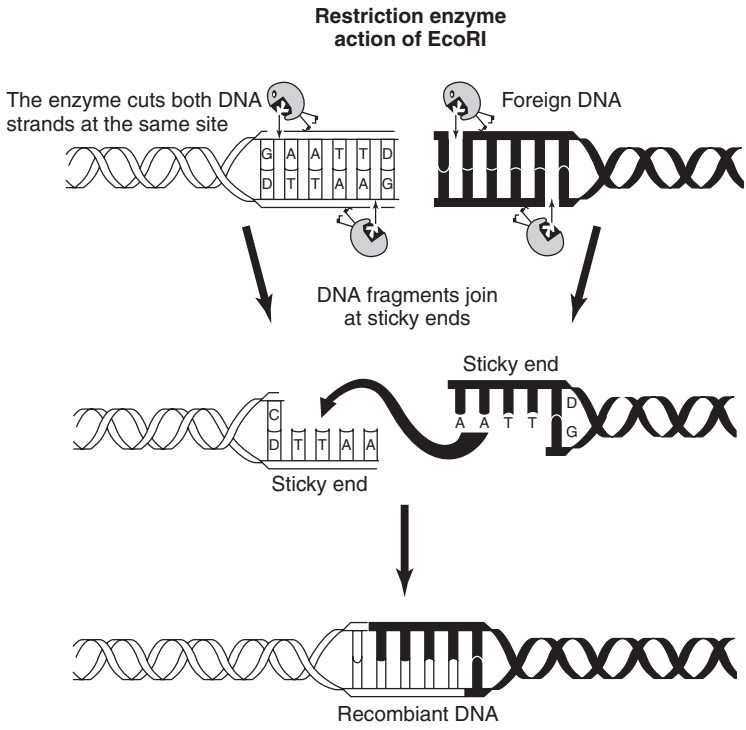


FIGURE 2.2. Two different pieces of DNA cut with EcoRI and ligation of the fragments to create a recombinant molecule.

that they also have some additional small circles of DNA called **plasmids** (see Figure 2.3), which carry genes for functions such as antibiotic resistance that evolve rapidly (Chang and Cohen 1974). Each bacterial cell has just one copy of its chromosome, but it may have hundreds of copies of a plasmid. Bacteria have a natural mechanism to transfer plasmids from one cell to another or to take up plasmids from the environment.

Cohen and Boyer used the recombination concept developed by Berg to make use of bacterial plasmids as **cloning vectors** to carry fragments of DNA from other organisms and make millions of copies of these fragments (Cohen et al. 1973). The cloning process begins by isolating DNA from an organism of interest and

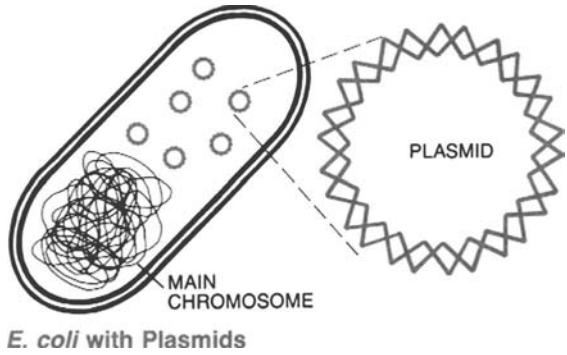


FIGURE 2.3. A plasmid is a circle of DNA maintained independently from the chromosome within a bacterial cell.

cutting it using a restriction enzyme, which cuts the DNA at a specific sequence, such as GAATTC. Plasmid DNA is isolated from bacteria, and it is also cut with the same restriction enzyme. Then the cut fragment of interesting DNA is mixed with the cut plasmid DNA and they are joined together, using DNA ligase, into a new circular molecule that contains both plasmid and the target DNA. This new molecule is called a **chimeric** plasmid because it is made up of DNA from two different types of organisms. The first chimeric clone was created in 1973 by splicing DNA from a frog into an *E. coli* plasmid (Cohen et al. 1973). Cohen, Boyer, and coworkers followed up by demonstrating that *E. coli* will produce foreign proteins from genes cloned into plasmids (Morrow et al. 1974).

The chimeric plasmid are then put back into a bacterial cell using a process called **transformation**—basically utilizing the bacteria's natural ability to take up plasmid DNA from a solution. The bacteria carrying the chimeric plasmid are then put into a medium where they can grow, and as the bacteria multiply, so does the plasmid (see Figure 2.4). Then the bacteria are harvested and large quantities of plasmid with cloned DNA can be purified. The interesting DNA fragment can be removed from the plasmid DNA by cutting again with a restriction enzyme. Some bacteria

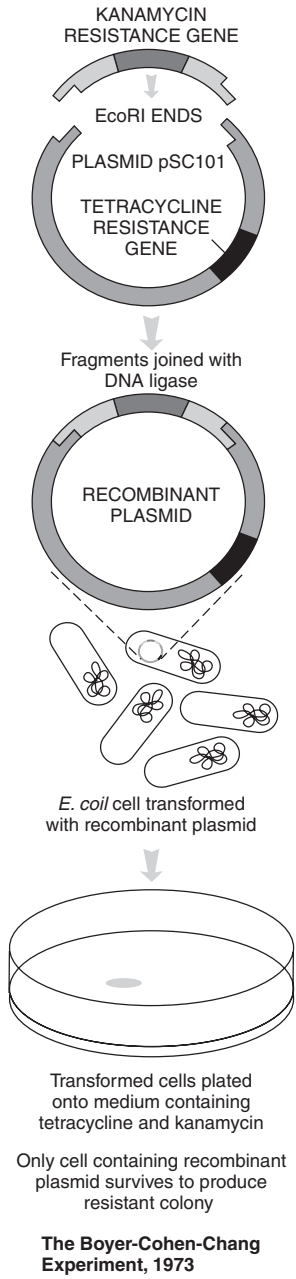


FIGURE 2.4. Ligation of a foreign gene into a plasmid and cloning in *E. coli* cells.

containing the chimeric plasmid can also be frozen so that more cloned DNA from the plasmid can be obtained whenever it is needed.

These methods for cutting, pasting and copying DNA can be used to construct complex DNA molecules that have parts from several different genes or from different organisms. Taken together, this technology is called **recombinant DNA cloning** or **genetic engineering**. Some viruses that grow in bacteria (bacteriophage) can also be used as cloning vectors by replacing part of the natural virus DNA with some other piece of DNA. In order to work with very large fragments of DNA, vectors have been developed that act as artificial chromosomes in bacteria [bacterial artificial chromosomes (BACs)] or in yeast [yeast artificial chromosomes (YACs)].

PCR Is Cloning without the Bacteria

Molecular biology is not a discipline for people who expect to spend a few years learning a set of skills and then sit back and use them for a few decades. Just when you think that you know all the basic chops in the lab, somebody comes along and reinvents the entire field. The polymerase chain reaction (PCR) was such a technical revolution. The basic concept is very simple: use the DNA polymerase enzyme that organisms use to copy their own DNA as a tool to copy specific pieces of DNA. Target the copying process by using a short primer that is complementary to one end of the desired sequence, then make copies of the other strand by using a second primer that is complementary to the other end of the desired sequence. Then make another copy in the forward direction, make another copy in the reverse direction, and repeat for many cycles. After the first cycle, the newly synthesized fragments serve as templates for additional rounds of copying. The net result is that each round of copying doubles the number of copies of the desired DNA fragment, leading to

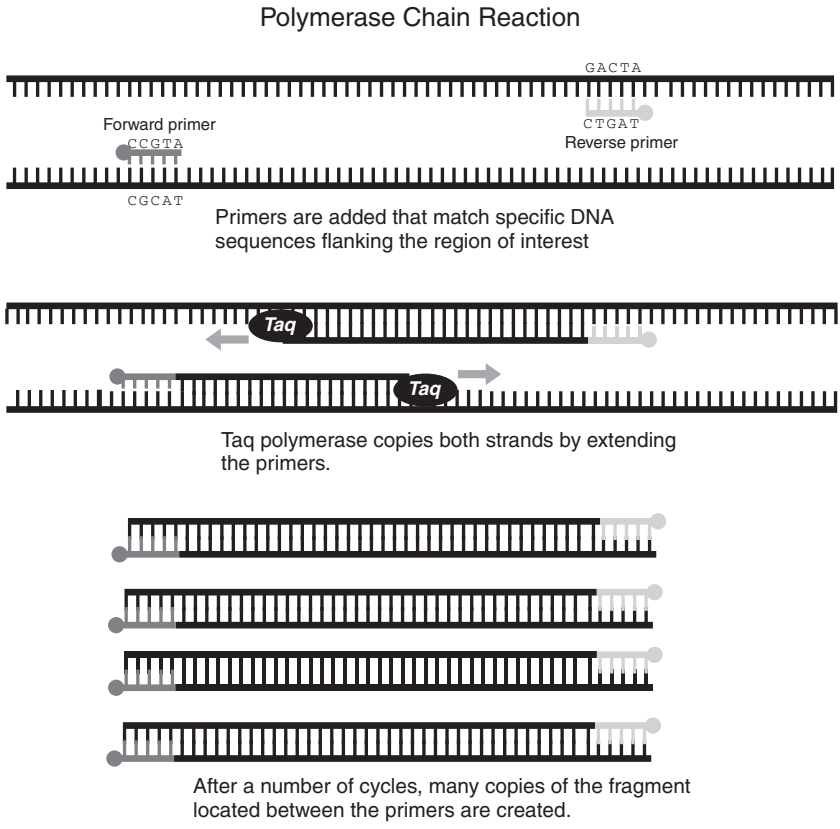


FIGURE 2.5. The PCR process: A pair of primers complementary to both ends of a target DNA sequence bind to the DNA, and complementary copies are synthesized by *Taq* DNA polymerase. Each new strand serves as a template for additional rounds of synthesis leading to creation of large amounts of the target fragment.

an exponential amplification (see Figure 2.5). Even if the amplification is not perfectly efficient, millions of copies are created in about 20 cycles.

The ingredients for the PCR reaction were available in the typical molecular biology laboratory for 10–15 years before Kary Mullis worked out the technique in 1983 (Saiki et al. 1985). Many scientists were slapping themselves on the forehead when

Mullis picked up the Nobel Prize in 1993. The one additional element that made PCR simple and user-friendly was the discovery that bacteria such *as Thermophilus acquaticus* that live in hot springs and deep ocean thermal jets have heat-resistant DNA polymerase enzymes (i.e., ***Taq*** polymerase). These heat-stable enzymes allow the PCR reaction to proceed for many cycles by simply heating and cooling a tube with the target DNA, the two primers, polymerase, and the free G, A, T, and C nucleotide triphosophates.

The beauty of the PCR process is not just that it makes many copies of a DNA fragment in a simple, single-tube reaction. It can also be used to pull out a single specific DNA fragment from a complex mixture—such as an entire genome. PCR can also be used to amplify substantial amounts of specific DNA fragments that can be used for various other molecular techniques from very tiny and impure samples—such as in clinical diagnostics for infectious agents, in forensic investigations, or from fossil remains. The basic requirement is that about 20 bases of sequence must be known at either end of the fragment of DNA that is to be amplified in order to create the forward and reverse primers. However, molecular biologists have worked out dozens of ingenious methods that allow PCR amplification when only a single primer sequence is known, when the sequences flanking the desired DNA segment are only partially known, or when new flanking sequences are attached in some complex cloning scheme.

Polymerase chain reaction is an essential ingredient in many different DNA diagnostic tests. It allows a very small sample of patient DNA (or RNA) to be used as source material to generate sufficient quantities of specific DNA fragments that can be sequenced, identified by mass spectroscopy, or detected by a variety of other labeling, probe, or visualization schemes (see Figure 2.6).

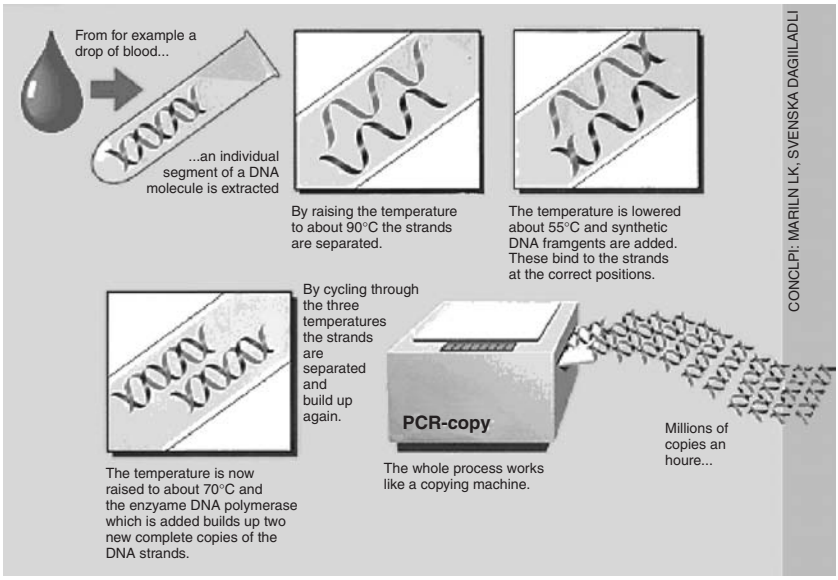


FIGURE 2.6. Simple diagram of the PCR process for a diagnostic test. (Nobel Foundation Website.)

DNA Sequencing

It is a bit of a conceptual leap from the discovery of the structure of DNA to the sequencing of the human genome, but one leads directly to the other. The double-helix model of DNA led to an understanding of how the DNA is duplicated as cells grow and divide. This process of DNA replication was then harnessed as a tool for the Sanger method of determining the sequence of a piece of DNA (Sanger et al. 1977).

Modern DNA sequencing technology is based on the method of controlled interruption of DNA replication developed by Fred Sanger in 1977 (for which he was awarded the Nobel Prize in 1980 together with Walter Gilbert and Paul Berg). Sanger combined the natural DNA replication machinery of bacterial cells with a bit of recombinant DNA technology and some clever biochemistry to create an *in vitro* system where a single cloned fragment of DNA is copied, but some of the copies are halted at each base

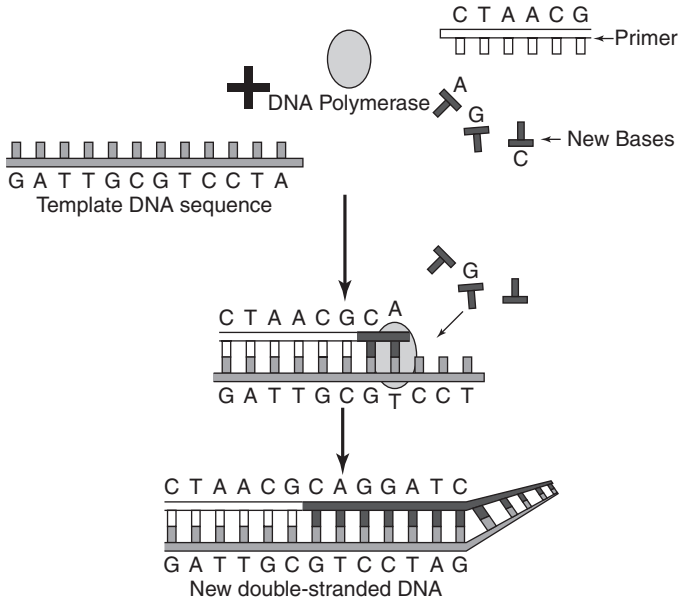


FIGURE 2.7. DNA polymerase uses a primer and free nucleotides to synthesize a complementary strand for a template DNA sequence.

pair along the sequence. Natural DNA replication uses a DNA polymerase enzyme that copies a template DNA sequence (one-half of the DNA double helix) and creates a new DNA polymer, complementary to the template, by joining free deoxynucleotides into a growing DNA chain. The replication reaction also requires a primer—a short piece of DNA that is complementary to the template—to which the polymerase can affix the first added base (see Figure 2.7).

THE SANGER METHOD

The Sanger sequencing method makes use of specially modified **di**deoxynucleotides that stop (terminate) the replication process if they are incorporated in the growing DNA chain instead of the normal deoxynucleotide. For each template, four separate sequencing reactions are set up, each containing one of the

dideoxynucleotides (ddG, ddA, ddT, and ddC) as well as a full set of normal deoxynucleotides, the primer, and the DNA polymerase enzyme. For example, in the reaction mixture containing dideoxyadenine, some of the growing strands are stopped when they reach each adenine in the template sequence. The resulting set of DNA fragments form a nested set, all starting at the same point, but ending at different A residues. Similar reactions are set up that stop replication at G, T, and C residues.

In order to visualize the resulting DNA fragments, it is necessary to incorporate some type of labeled molecule, usually radioactive, in the replication reaction. It is possible to label the primer, the deoxynucleotides, or the dideoxynucleotides. In any case, the fragments are separated by length using polyacrylamide gel electrophoresis (PAGE), with one gel lane for each of the 4 different dideoxynucleotide reactions. Then the gel, containing the radioactively labeled DNA fragments, is placed on top of a sheet of X-ray film so that the radioactive bands of DNA can expose the film (see Figure 2.8). Then the sequence is manually read off of the X-ray film from the positions of the bands and typed into a computer (see Figure 2.9).

The value of determining DNA sequences was immediately obvious to many biologists, but the laboratory techniques of the Sanger method are both laborious and technically demanding. DNA sequencing became a rite of passage for many molecular biology graduate students in the 1980s and early 1990s. Initially some kits were developed to simplify and standardize the biochemistry. These kits evolved to include superior types of polymerase enzymes, and various minor improvements were made in the polyacrylamide gel apparatus, but the essential technique remained unchanged for about 15 years.

Automated DNA Sequencing

The first major innovation to improve DNA sequencing was Leroy Hood's development of fluorescently labeled nucleotides

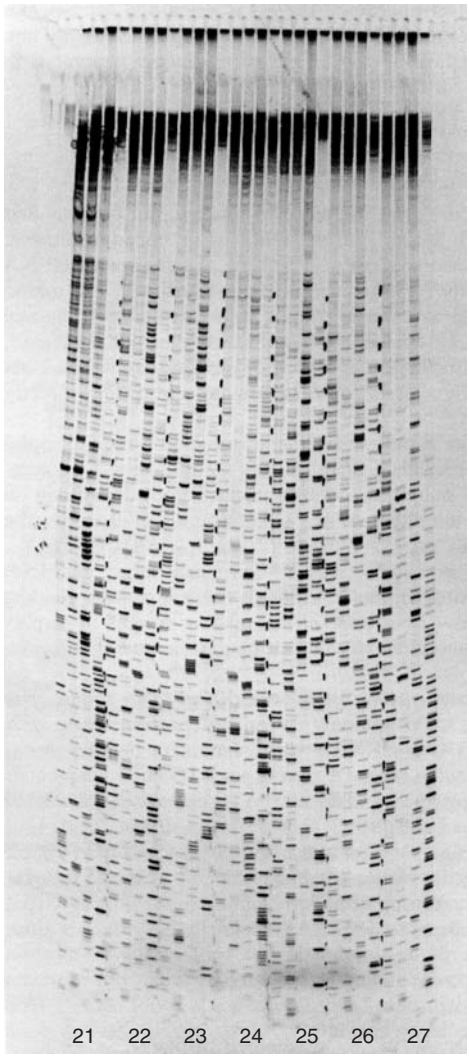


FIGURE 2.8. An autoradiogram (X-ray film) of a DNA sequencing gel. Each sequence requires Four lanes, one for each base.

in 1985 to replace the standard radioactive labels (Smith et al. 1986). The fluorescent labels could be measured directly in the acrylamide gel as DNA fragments passed by a laser/detector, thus eliminating both the radioactivity and the X-ray film. In addition, Hood used four differently colored fluorescent labels, one

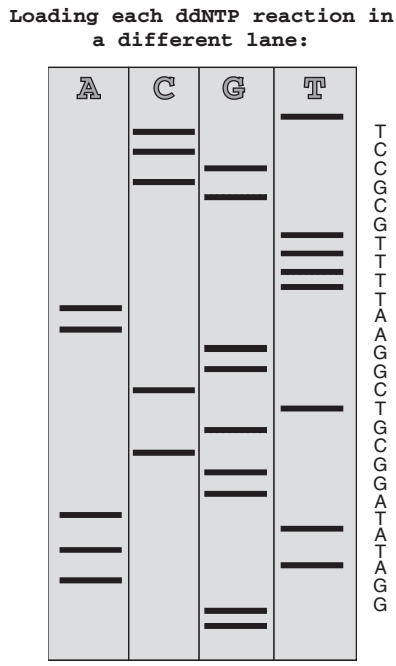


FIGURE 2.9. Diagram of a sequencing gel showing bands in four gel lanes that represent DNA fragments produced by the four different dideoxy sequencing reactions. The final DNA sequence is shown at the right.

for each of the four DNA bases, so that after the sequencing reactions, the four sets of fragments could be run in a single lane of an acrylamide gel and the base determined by the color of each fragment (see Figure 2.10).

Hood also directly connected the fluorescent detector to a computer so that the fluorescent signal was automatically collected and converted to DNA sequence (see Figure 2.11). Hood, together with Lloyd Smith, Michael Hunkapiller, and Tim Hunkapiller, founded Applied Biosystems Inc. (ABI) to manufacture a commercial version of his fluorescent sequencer, which became available in 1986.

Since 1986, ABI (in cooperation with the PerkinElmer Corporation) has consistently improved their machines and dominated

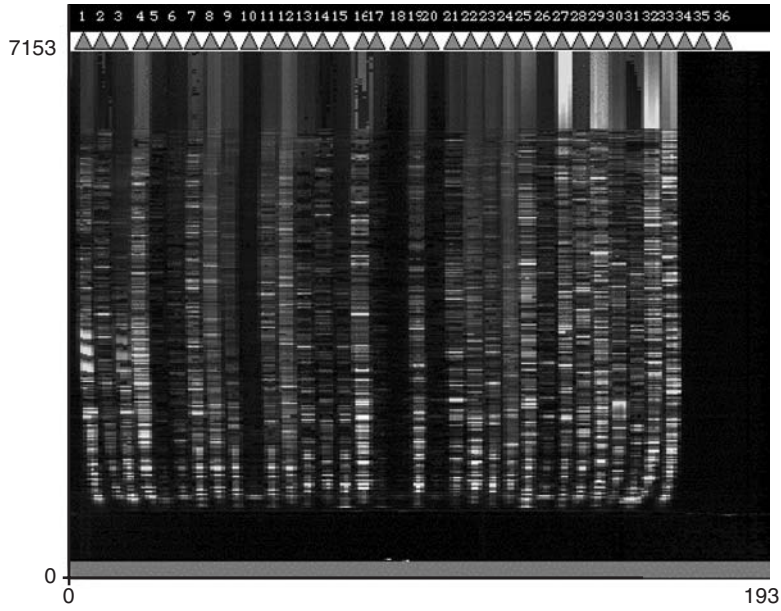


FIGURE 2.10. A fluorescent sequencing gel produced on an ABI automated sequencer. Each lane contains all 4 bases (in different colors). (See insert for color representation.)

the commercial marketplace for automated sequencers. Essentially all of the Human Genome Project and absolutely all of Celera Genomics' sequencing was done on ABI machines. However, ABI machines still have many of the limitations of the original Sanger method. They still rely on DNA polymerase to copy a template DNA sequence and PAGE to separate the fragments.

SUBCLONING

One of the key limitations of the Sanger/ABI method is that DNA sequences can be determined only in chunks of 500–800 bases at a time (known as "reads"). Larger fragments cannot be resolved by PAGE. As a result, determining the sequence of large pieces of DNA requires sequencing many overlapping fragments, and then assembling them. There are a variety of strategies for breaking

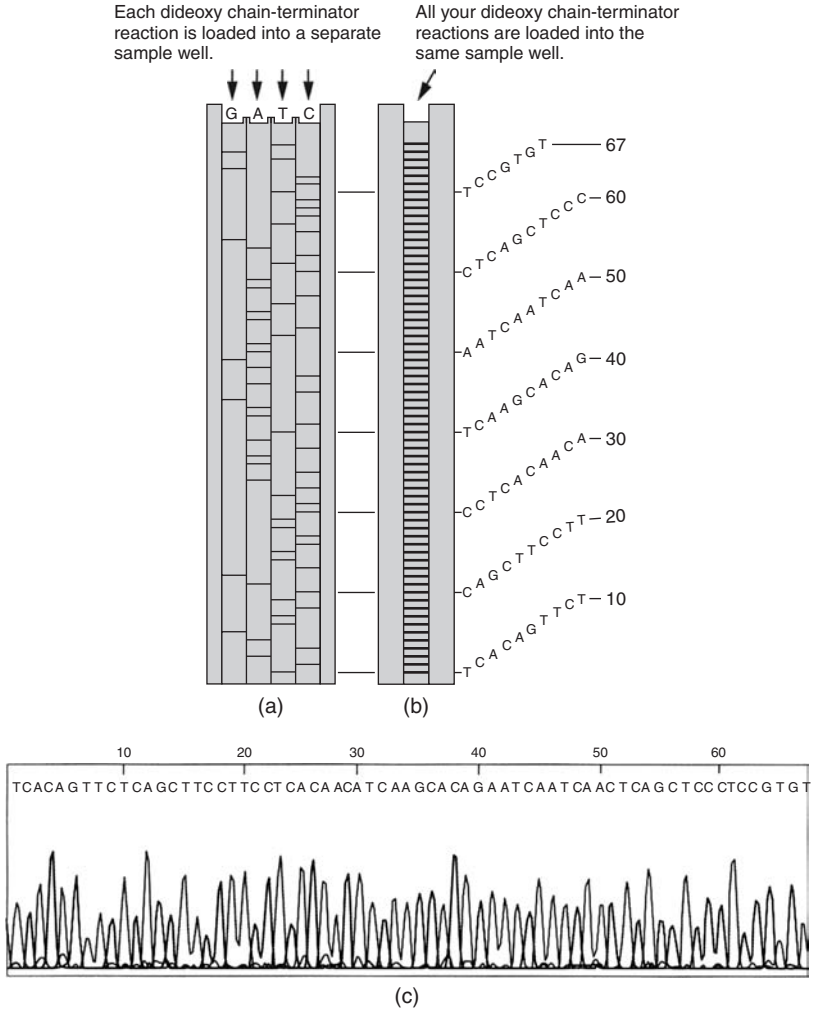


FIGURE 2.11. ABI fluorescent sequencers allow all 4 bases to be sequenced in a single gel lane and feature automated data collection. (See insert for color representation.)

up large DNA molecules—such as human chromosomes—into overlapping fragments for sequencing.

Early DNA sequencing projects proceeded very methodically. A scientist would first map a chunk of DNA for various restriction enzyme cut sites, and clone these restriction fragments into

plasmids (subcloning), carefully tracking how they would reassemble. This process needed to be done at least twice to generate two sets of overlapping fragments since the sequence at the ends of fragments tends to be filled with errors. Alternately, sets of nested fragments could be generated by chewing away at one end of a cloned fragment with a DNase enzyme and stopping the reaction at various timepoints. Then the chewed fragments must be subcloned, their sizes determined, and finally sequenced.

All of this laboratory cloning work was very time-consuming and could not be scaled up for larger projects. As the cost of automated sequencing decreased and the speed and throughput of the machines increased, it became necessary to find faster methods of generating small fragments to be sequenced. In theory, if a chunk of DNA is copied many times (cloned), and all of these copies are broken up into many random fragments ("shotgun" subcloning), and these fragments are sequenced, then eventually a complete set of overlapping fragments can be found whose sequences can be assembled into the sequence of the original chunk. This ends up being something of a statistical game. In order to find a set of overlapping fragments that completely cover a large chunk of DNA by sequencing random shotgun subclones, it is necessary to sequence a total amount of DNA that is equal to much more than the overall length of the original DNA chunk. The shotgun clones form a Poisson distribution—sort of like trying to hit all of the squares of a chessboard by hanging it on the wall and throwing darts: you will hit some squares many times before all of them are hit.

In the mid 1990s, The Institute for Genomic Research (TIGR) became a champion of the shotgun style of sequencing. They streamlined the process of randomly cloning thousands of fragments from large DNA molecules and efficiently feeding clones into a room full of automated sequencers, minimizing the number of scientist-hours spent generating the sequence reads. Then they devoted the bulk of their efforts into developing good sequence

assembly software and "finishing" the assembly of sequences using a combination of computer-savvy scientists to help the software, and occasionally going back into the lab to resequence a troublesome spot. The success of this method was dramatically demonstrated by the publication by TIGR of the complete sequence of the *Haemophilus influenzae* genome at a time when the rest of the scientific community was convinced that such a large project could not be done by shotgun sequencing and was in fact beyond the reach of current technology (Fleischmann et al. 1995).

Sequence Assembly

The assembly of shotgun fragments is obviously a job for computer software, but there are some problems associated with the data from automated sequencers. DNA sequencers are not perfectly accurate, and mistakes are much more common at the beginning and the ends of each sequence read—precisely the regions that are needed in order to find overlaps. DNA contains many types of repeats ranging from long tracts of a single base or a simple repeat of 2 or 3 bases to tandem (or inverse) repeats hundreds to thousands of bases long. Repeats make it very difficult to assemble overlapping sequences unless a single read spans an entire repeat and includes nonrepetitive sequence on both sides.

There are some additional problems with the assembly of huge genome sequences, particularly eukaryotes, that did not affect scientists working on the sequencing of smaller pieces of DNA. Eukaryotic genomes have duplicated genes, and even duplications of entire sections of chromosomes. Eukaryotic chromosomes contain centromeres and telomeres that consist of nothing but thousands of repeats of the same short sequences. The human genome also contains about 100,000 copies of a sequence called **ALU**, which is an inactive transposon 147 bases long. Clearly a sequence fragment that ends in the middle of an ALU sequence could overlap and assemble with any other ALU-containing

fragment, possibly from another chromosome, leading to incorrect assemblies.

Sequencing the Human Genome

The public Human Genome Project Consortium has relied primarily on a map-based strategy to sequence the human genome (Lander et al. 2001). They spent several years cloning 100,000–1,000,000 bp chunks of human DNA into large vectors called **bacterial artificial chromosomes** (BACs), then painstakingly assembling a complete set of overlapping BACs that covered every chromosome. Only after this BAC map was completed did they start large-scale sequencing of each BAC by breaking it into many small fragments that could be directly sequenced by the Sanger method on automated DNA sequencers. Finally, all of these short sequences must be assembled back into complete chromosomes using computer programs.

In contrast, the Celera Genomics human genome sequencing strategy relied entirely on a shotgun sequencing approach (Venter et al. 2001). Rather than carefully building a set of large overlapping BAC clones that covered the entire genome, they randomly cut genomic DNA from several different people into fragments, and sequenced a number of fragments equivalent to 6 times the total size of the human genome (i.e., 18 billion bases of sequence information). Celera then used computer programs to assemble overlapping fragments.

Both the HGP and the Celera approaches led to an uneven sequence of the genome. Some areas are sequenced many times over, and there are gaps. There is also a point of diminishing returns where more random sequencing is not likely to fill some stubborn gaps that might be caused by special properties of the DNA in that location. For example, centromeres and telomeres contain highly repetitive sequences that resist cloning and cannot by assembled by computer algorithms. At that point, which

has already been reached for the human genome sequence, the only way to fill the gaps is to change strategy to a much more painstaking, hands-on approach. Small teams of biologists must tackle the unique problems of each gap. It will be quite a few years before the human genome sequence is perfectly complete without any gaps.

REFERENCES

Boyer HW, Roulland-Dussoix D. 1969. A complementation analysis of the restriction and modification of DNA in *Escherichia coli*. *J Mol Biol* **41**:459–472.

Chang AC, Cohen SN. 1974. Genome construction between bacterial species *in vitro*: Replication and expression of *Staphylococcus* plasmid genes in *Escherichia coli*. *Proc Natl Acad Sci USA* **71**:1030–1034.

Cohen SN, Chang AC, Boyer HW, Helling RB. 1973. Construction of biologically functional bacterial plasmids *in vitro*. *Proc Natl Acad Sci USA* **70**: 3240–3244.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CI, Tomb JF, Dougherty BA, Merrick JM et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**:496–512.

Jackson DA, Symons RH, Berg P. 1972. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: Circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. *Proc Natl Acad Sci USA* **69**:2904–2909.

Lander ES, Linton LM, Birren B et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.

Morrow JF, Cohen SN, Chang AC, Boyer HW, Goodman HM, Helling RB. 1974. Replication and transcription of eukaryotic DNA in *Escherichia coli*. *Proc Natl Acad Sci USA* **71**:1743–1747.

Roulland-Dussoix D, Boyer HW. 1969. The *Escherichia coli* B restriction endonuclease. *Biochim Biophys Acta* **195**(1):219–229.

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**(4732):1350–1354.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**(12):5463–5467.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature* **321**:674–679.

Venter JC, Adams MD, Myers EW et al. 2001. The sequence of the human genome. *Science* **291**:1304–1351.

CHAPTER

3

Genome Databases

Genome Sequencing

Through the Human Genome Project, world governments have made a substantial investment in genome sequencing. From its outset, the genome sequencing effort has involved international collaboration among many different laboratories and research centers. The Human Genome Organization (HUGO; `http://www.hugo-international.org`) was founded in 1988 to coordinate international collaboration on human genetics and genome sequencing. One product of this collaboration has been an emphasis on extremely open data sharing policies. Scientists have always worked in a tradition of open access to the data that support published research work, but genomics technologies have taken this process to an entirely new scale. The raw sequence data have been released on open Websites and file transfer protocol (FTP) servers almost as soon as it comes off the sequencing machines. Assembled and partially annotated data have also been released long before they reached publication quality. The availability of genome data has allowed different sequencing centers to work together on

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

overlapping segments of the genome and basic scientists to grab bits of data that are relevant to their work as soon as they are created.

The Genome Project has also included a substantial effort to create publicly accessible Web databases to enable everyone to access these data, regardless of their scientific credentials. Genome sequence data constitute an essential basic reference that is relevant to basic science, medicine, and high school biology reports. The primary source for genome data is in the three databases of the International Nucleotide Sequence Database Collaboration (INSDC): GenBank in the United States, EMBL in Europe, and DDBJ in Japan. These three databases each maintain complete copies of the entire worldwide DNA sequence collection. The INSDC databases enforce uniform data standards, exchange updated data on a daily basis so that they can provide more rapid access to scientists spread across the world, as well as back up each other for short-term outages or disasters. All data in these databases are provided free with unrestricted access to all data records. No use restrictions or licensing requirements are included in any sequence data records, and no restrictions or licensing fees are placed on the redistribution or use of the data by any party. All database records submitted to the INSDC will remain permanently accessible as part of the scientific record.

GenBank is maintained by the National Center for Biotechnology Information (NCBI), a branch of the National Library of Medicine at the US National Institutes of Health (NIH). GenBank is the complete, definitive public collection of DNA sequence information. As of mid-2007, GenBank contains over 70 billion bases of DNA sequence, which amounts to over 250 GB of data, and about 2 billion bases are being added every 2 months. Since its inception, GenBank has doubled approximately every 14 months (see Figure 3.1).

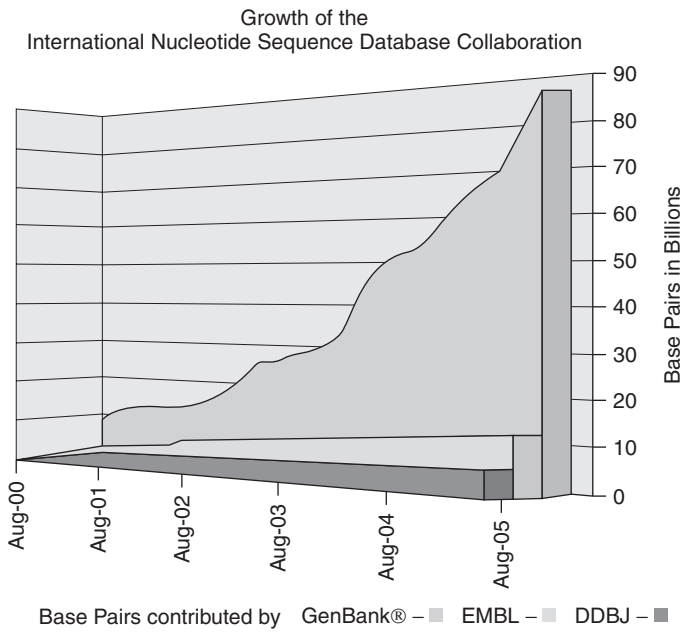


FIGURE 3.1. Growth of DNA sequences in GenBank (`http://www.ncbi.nlm.nih.gov/Genbank/index.html`).

ENTREZ

GenBank is most easily accessed with a Web browser at `http://www.ncbi.nlm.nih.gov`. This Webpage is the entry point to a very sophisticated set of searching tools called **Entrez**, which is the integrated, Web-based search–retrieval system used at NCBI for databases, including PubMed/MEDLINE, nucleotide and protein sequences (GenBank), protein structures, gene expression (GEO), mutations (OMIM and dbSNP), and taxonomy. Queries can be made for gene names, diseases, and individual scientists who submitted sequences; and limited by organism, chromosome, submission date, and other categories. Protein sequences are linked to their corresponding DNA sequences and to locations on the genome (when known). PubMed is the

comprehensive database of all medical and scientific literature maintained by the National Library of Medicine.

It is impossible to overstate the value of free online access to PubMed for medical and biology professionals, as well as for students and the general public. A comprehensive and up-to-date list of published articles can be found for any subject in the fields of medicine, nursing, dentistry, veterinary medicine, and molecular biology. PubMed contains bibliographic citations and author abstracts from about 4600 biomedical, general science, and chemistry journals published in the United States and 70 other countries. The database contains about 12 million citations (as of mid-2007) dating back to the mid-1960s. An increasing number of recently published articles can be accessed directly through Web links to full-text online journal articles (some free, some requiring a subscription to the journal, some for sale at a per-article fee).

The entire concept of evidence-based medicine requires physicians to be aware of the current research on outcomes of various treatments applied to a specific disease. Patients are increasingly showing up at physicians' offices with printouts of journal articles about diagnostics and treatments for their (putative) diseases that they have found via PubMed. Physicians benefit by dealing with a more informed patient, but they are also inspired to keep up (or keep ahead) of this rising tide of information. PubMed allows the medical professional to actively seek out information on subjects of interest across both common and obscure journals, rather that passively waiting for articles (or summaries of articles) to appear in the magazines and newsletters to which they subscribe.

A simple search can be conducted from the PubMed homepage by entering terms in the query box and pressing ENTER from the keyboard or clicking on the GO button on the screen. It is possible to construct complex search strategies using Boolean operators, **`Limits`** that restrict a search term to a specific field

(author, journal title, publication date, article type, journal type, etc.), and a **History** function that allows the user to combine previous searches (by intersection or union). **Clinical Queries** is a set of search filters developed for clinicians to retrieve clinical studies of the etiology, prognosis, diagnosis, prevention, or treatment of disorders (available at `http://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.shtml`). An Entrez feature called `My NCBI` allows the user to save searches and send email updates on the search at regular intervals to help keep current on topics of interest.

Another extremely valuable feature of PubMed is the **Related Links**. Every article is indexed by keywords (manually annotated by librarians), and other articles that share many of these keywords are automatically linked. So, once an interesting journal article is found, the reader can browse a network for many related articles. PubMed citations may also contain links to entries in other Entrez databases including nucleotide and protein sequences in GenBank, OMIM entries about human genetic diseases, GEO gene expression data, and so on. The links work in all directions among the databases, so that a query for a gene or protein or disease will also produce links to relevant journal articles in PubMed. A graphical view of the interlinks among the various Entrez databases is available on the ABOUT ENTREZ Webpage (`http://www.ncbi.nlm.nih.gov/Database/datamodel`; see Figure 3.2).

Since Entrez and PubMed are stable resources, other Webpages and Internet-enabled software applications can include permanent hypertext links to genes, proteins, and journal articles in the database. This allows every term paper, magazine article, blog, and set of lecture notes (in fact, every Word, PowerPoint, and PDF document) to contain live links to the full citations and abstracts for each scientific fact and gene or protein sequence that is mentioned. Genomics software can mine these databases to automatically annotate large lists of genes.

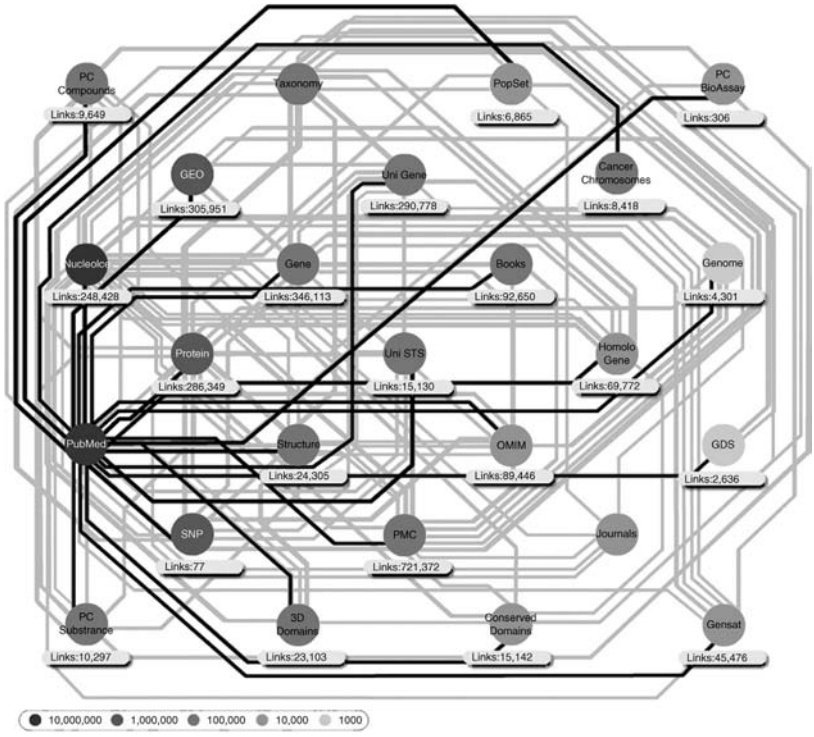


FIGURE 3.2. Links between databases in the Entrez system. (See insert for color representation.)

BLAST

While text-based searching is the most logical way to find journal articles, DNA and protein sequences are not always easy to find by their names. The NCBI also maintains a tool called **BLAST** (basic local alignment search tool) that searches the DNA and protein databases by making comparisons between a sequence supplied by the user and all database sequences (see Chapter 4, on bioinformatics tools). It is also possible to use BLAST to compare DNA to protein sequences using automatic translation. Considering that there are many millions of sequences in GenBank (billions of bases of DNA), BLAST is amazingly fast. The proper interpretation of the results of a

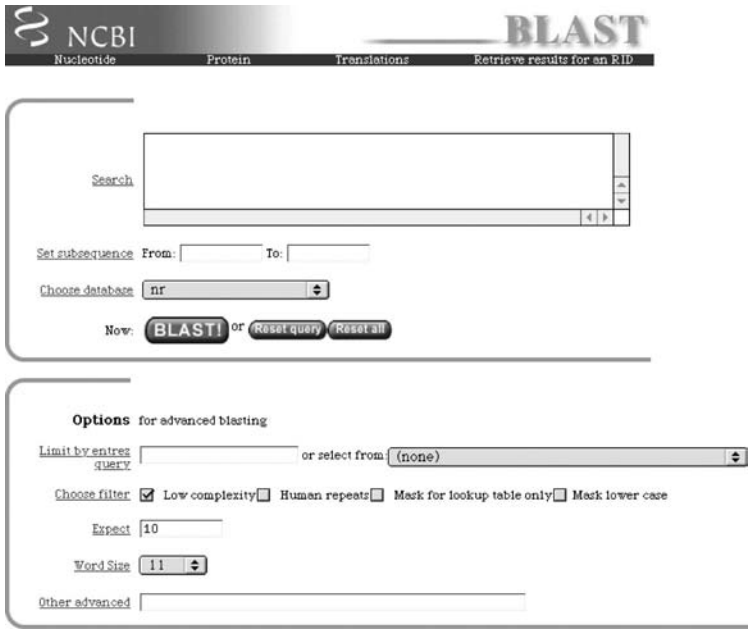


FIGURE 3.3. The BLAST query page on the NCBI Website for DNA–DNA searches.

BLAST search requires a thorough understanding of molecular biology, evolution, and statistics; thus, it is not really intended for use by the general public, but the tools are free for anyone to use at `http://www.ncbi.nlm.nih.gov/blast` (see Figure 3.3).

GENOME ANNOTATION

The initial products of genome sequencing are millions of bases of DNA—just an endless stretch of seemingly random G, A, T, and C letters (see Figure 3.4). The current status of the human genome, and other "complete" genome sequences is quite good, in terms of accuracy. Except for some highly repeated regions (such as centromeres and telomeres), these genome sequences are more than 99.9% accurate. However, this sequence is meaningful only

FIGURE 3.4. Unannotated genome sequence.

when the genes are found and their biological functions are described. Unfortunately, the current status of genome annotation is not nearly as good as that for genome sequencing. Ideally, we would like to have a genome that is deeply annotated with multiple layers of relevant information. We would like to know the locations of all protein-coding regions, intron–exon structure, alternative splicing sites (and the corresponding transcripts), promoters and transcription factor binding sites, protein structure and function (including roles in metabolic and regulatory pathways and protein–protein interactions), and gene expression data (which might include information from a wide variety of healthy and diseased tissue types and responses to various drugs and environmental perturbations). In a sense, the annotated genome should become a central tool for organizing knowledge about genes.

All of this annotation rests fundamentally on the ability to find genes in the genome sequence. That turns out to be a very challenging problem, partly because of the difficulty in defining what a gene is in a manner that satisfies the majority of biologists and can be specified in unambiguous computer code, and partly

because the genome is full of sequences that look like genes but are not (pseudogenes), while other genes are hidden. Another, more fundamental, problem is in the basic concept of pattern recognition. We know the sequences of some genes, so we look for other genes that resemble the ones that we know. This similarity may be based on direct sequence–sequence matching (i.e. BLAST), or it may rely on more subtle statistical properties of the DNA that can be found to differentiate "gene coding" regions from "noncoding" regions. Either way, we are relying on what we know to shape what we are looking for, so it is inevitable that we will miss things that are novel and different. Yet biology is always full of exceptions to the rules, variations on a theme, and new rules that govern new subclasses that can be discovered only using fundamentally different methods.

GenBank does not yet contain a comprehensive list of all human genes and their functions. In fact, there is still not a solid scientific consensus on how many genes are in the human genome. Before the human genome was sequenced (in the year 1999), the consensus among most molecular biologists was that humans had about 100,000 genes (Feng et al. 2000). Immediately after the first draft of the genome was published in 2001, the "official estimate" of genes in the human genome was dropped to about 32,000. In 2003, when the "final draft" of the human genome was published by the International Human Genome Sequencing Consortium, the official gene count was set at 23,299 by the EMBL **Ensembl** genome annotation system. In April 2007, the Ensembl system recognized 21,662 "known genes" plus 3994 genes that encode functional RNAs (tRNAs, rRNAs, and microRNAs). This steady decline in the number of human genes reflects the use of overly optimistic gene finding programs in the early stages of annotation. As gene annotations are studied more carefully, many predicted or "hypothetical" genes that were annotated on the basis of similarity to an mRNA sequence, turn out to lack essential features

(promoters, open reading frames, introns, etc.). At the same time, new genes, and new categories of genes, are also being discovered and added to the database. For example, microRNA genes were not known in 2001.

At NCBI, a team of curators has developed a hand-annotated set of genes known as RefSeq: "The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms." The goal of the RefSeq project is to provide a standard dataset of genetic information for a variety of scientifically important species. Each RefSeq record is a synthesis of available information in public databases about a specific gene. As of April 2007, RefSeq contains 19,271 human genes (full-length mRNAs mapped to the genome), but RefSeq includes only protein-coding genes; it does not include genes that encode functional RNAs. Many "whole genome" gene expression tools (microarrays) are built using RefSeq as the primary data source for human gene sequences from which to design probes.

Genome Browser

While it is possible to directly download the sequence of a single gene (or all the genes) directly from GenBank (or EMBL or DDBJ), is it often more informative to view a gene in the context of the genome—where it lies on the chromosome, what other genes and genetic features lie nearby, and so forth. Each major genome database has made an effort to design user-friendly Web interfaces to the genome data, but the best genome browser is probably the one developed by the UCSC Genome Bioinformatics team (David Haussler, Jim Kent, et al.) at `http://genome.ussc.edu`. There are many innovations in the UCSC Genome Browser, but the fundamental good idea was to lay the genome out horizontally across the top of the screen,

and then add tracks of annotation information that run across the screen underneath the genome. This allows for an intuitive set of navigation controls that include buttons to zoom in, zoom out, and move left or right (see Figure 3.5). The user can then choose among a huge number of different tracks of annotation information to add to the display, including primary sequence information [chromosome band, gaps and contigs (see Glossary at end of this book for definition of **contig**) in the primary sequence assembly, GC content, estimates of local recombination rate, repeats, etc.], gene predictions from a variety of different computer algorithms, gene sequences from authoritative sources

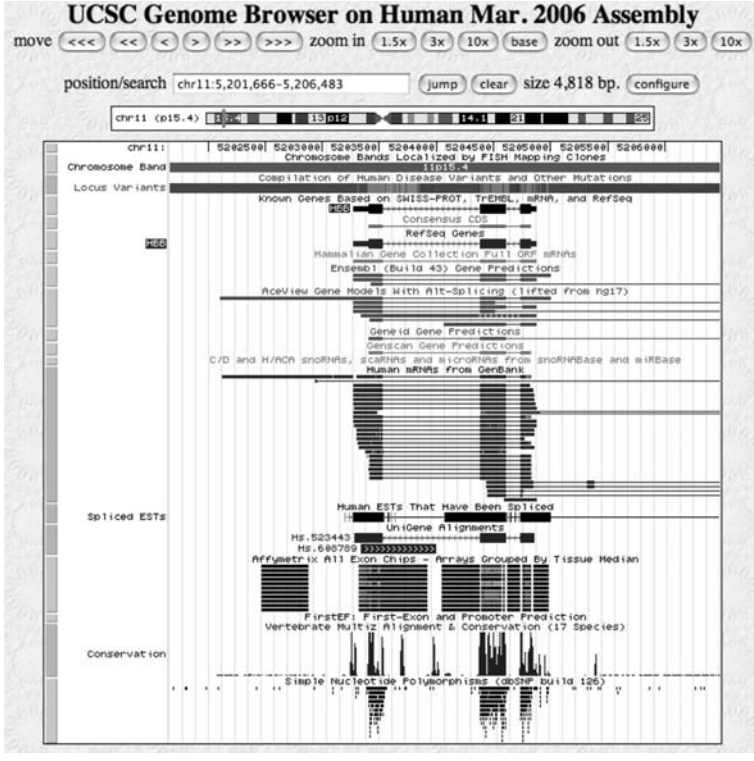


FIGURE 3.5. UCSC Genome Browser view of human β-hemoglobin gene (accessed on April 4, 2007).

(including Ensembl and RefSeq), mRNA and EST sequences from GenBank, gene expression (from microarray studies), and comparisons with the genomes of other organisms.

One excellent property of this UCSC Genome Browser is that the fundamentally confusing nature of the data used for genome annotation is revealed. The raw data (mRNA sequences), the comparisons with other genomes, and the computer annotations are all self-contradictory and disagree with each other. Even for extremely well studied human genes such as β-hemoglobin, there is confusion about exactly where the mRNA begins and ends, and why there should be small regions of highly conserved sequence with other species inside introns that presumably do not code for protein.

There are a number of ways to locate a gene on the UCSC Genome Browser. The simplest way is to type the name of the gene (DNA or protein, using any of a wide variety of database nomenclatures) in the box labeled POSITION OR SEARCH TERM and then hit the SUBMIT button. If the name is ambiguous, a list of possible matches is presented along with their chromosomal locations. It is also possible to enter a chromosome location or region directly (e.g., **`chr3:1000000-1500000`**). Another way to locate genes on the genome browser is by direct sequence similarity. BLAT is a "BLAST-like alignment tool," designed by Jim Kent, that very quickly matches up an input sequence with similar regions of the genome. BLAT allows for large gaps, so an mRNA or protein sequence can be matched to genomic DNA with no penalty for introns.

The UCSC Genome Browser also contains other eukaryotic genomes including chimpanzee, Rhesus monkey, mouse, chicken, rat, dog, cat, cow, yeast, a frog, four fish, three worms, two mosquitoes, and 11 species of fruitfly. The twenty-first-century enthusiasm for genome sequencing shows no sign of letting up, so the collection of fully sequenced species in the UCSC Genome Browser is certain to grow.

OTHER VIEWS OF THE HUMAN GENOME SEQUENCE

In addition to the UCSC Genome Browser, both GenBank/NCBI and EMBL/EBI have created their own Web-based tools to view human (and other species) genome data. The NCBI "map viewer" provides a graphical map of the human genome (`http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606`), which shows the location of all known genes, expressed sequence tags (ESTs), cytogenetic bands, and many other genetic markers including genetic maps build with classic "radiation hybrid" (Deloukas et al. 1998) technologies (see Figure 3.6). However, the vertical layout of the chromosome makes it difficult to scroll across multiple tracks of data. The controls to add and remove tracks are located on another page accessed by the MAPS AND OPTIONS button at the left of the screen. Color is not used to distinguish different types of information, and many links are presented as obscure abbreviations (`HGNC, sv, dl, ev, mm`).

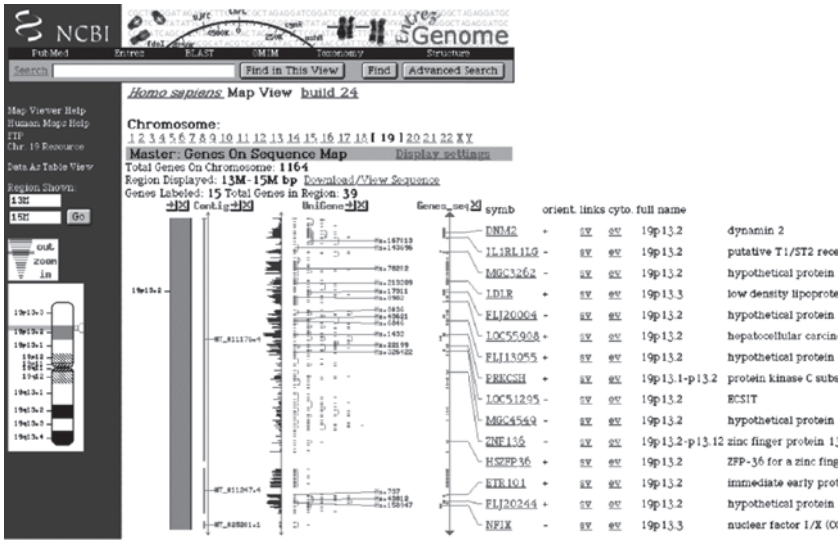


FIGURE 3.6. A section of human chromosome 19 on the NCBI map viewer.

Overall the interface is less informative and more confusing than the UCSC Genome Browser. The NCBI map viewer contains genome data for a much larger set of genomes than UCSC (over 70), including many plants, fungi, and a few protozoa.

The EMBL genome viewer is called the **Ensembl Genome Browser** (`http://www.ensembl.org`). **Ensembl** is an automated genome annotation software system as well as a Web-based genome viewer. Ensembl creates consensus gene predictions for new genomes through an automated pipeline, which makes it possible to produce annotations and gene lists for more genomes, including partially sequenced genomes (the hedgehog), more rapidly than the manual curation process used by RefSeq at the NCBI. The graphical view of the genome is horizontal, and less complex than that of the UCSC Genome Browser, but this does not prevent it from showing contradictory gene models (see Figure 3.7). The Ensembl Website allows the user to locate a specific sequence (DNA or protein) on the genome of any of the included organisms (33 organisms as of mid-2007, which is similar but not identical to the species in the UCSC Genome Browser) by BLAST or SSAHA similarity search tools. It also makes it quite easy to perform cross-species sequence comparisons at various levels of sensitivity. The BioMart tool provides an interface that allows the user to make complex queries across a huge number of different types of genome annotation data and then extract lists of genes, sequences, exons, introns, polymorphisms, or other data features.

Human Genetic Diseases

Rather than start a genome search with a piece of DNA sequence or a complex query in the scientific literature, a physician (or a patient) is more likely to start with a known disease. The NCBI has built a companion to the GenBank/PubMed database

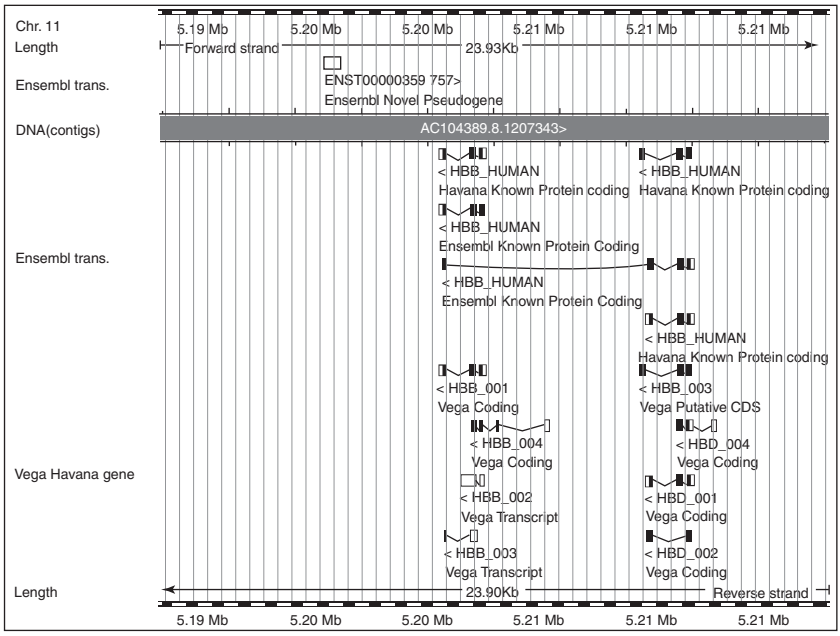


FIGURE 3.7. Human β-hemoglobin shown in the Ensembl Genome Browser.

called **Online Mendelian Inheritance in Man** (OMIM) that is completely focused on human genetic diseases (see Figure 3.8). OMIM is authored and edited by Dr. Victor McKusick and his colleagues at Johns Hopkins University. OMIM contains a short description of each gene and extensive excerpts and summaries of a wide range of scientific literature on the gene and the disease—including clinical reports, all known alleles and mutations, an extensive bibliography, including direct links to PubMed citations for each paper, links to the GenBank entries for both the gene and the protein, links to a cytogenetic map that is in turn linked to the NCBI's human genome map (which contains the complete genomic sequence, neighboring genes, and known mutations in that region). OMIM entries also have links to other disease-specific databases that contain relevant information.

FIGURE 3.8. The entry for the CTFR gene in the Online Mendelian Inheritance in Man (OMIM) Website.

A System for Naming Genes

Unfortunately, searching genome databases is not always as simple as typing in the name of a disease or gene and hitting a SEARCH button. The problem is not inherent in the structure of computer databases or even in the complexity of biology, but simply that people do not always call the same thing by the same name. Scientists working with fruitflies like to name genes after the appearance of flies with a mutation in that gene. However, mice or humans with a mutation in the corresponding gene will not have the same appearance (wingless or bent antennae or whatever), so that gene will have a different name in each species. Later on, another similar gene might be found in a worm or a plant—now,

will it be described as similar to the fly gene or similar to the human gene?

This problem has worsened as more genes have been described in more species and studied in different contexts. Fortunately, a group of geneticists and database curators have begun a project called the **Genome Ontology** (GO) to sort out all of the names into a consistent system. The names of all organisms have been organized into a single taxonomy, so that with a single name a scientist can communicate unambiguous information about the identity of an organism, and at the same time indicate the relationship of that organism to all other living things. A gene ontology would include unambiguous names for all genes as well as a consistent vocabulary to describe the features of genes that is not specific to any one type of organism or any particular scientific discipline (at the expense of others). The Genome Ontology is organized around three general principles that are common to all eukaryotic organisms: molecular function, biological process, and cellular location. In addition to developing an internally consistent vocabulary that can apply equally well to all organisms and biological disciplines, the GO project has taken on the task of reannotating all of the existing gene, protein, and species databases by mapping all terms used in current gene feature descriptions to equivalent GO terms. Once these equivalencies are established, then GO can serve as an intermediary to translate the annotation terms of any database to any other database. Then, if scientists can learn to phrase their database queries using GO terms, the same query will work equally well in any database.

MODEL ORGANISMS (COMPARATIVE GENOMICS)

This discussion of ontology across species brings up an interesting question. Are there equivalent genes for all functions in

all organisms? Intuitively one would say "No; after all, different organisms have very different amounts of DNA and different numbers of genes." On the other hand, all organisms have more or less the same biochemical processes at a cellular level (energy metabolism, growth, reproduction, movement, etc.). At the level of protein sequences, quite a lot of similarity (homology) can be found across distant branches of the tree of life. Yet some groups of organisms have unique structures or unique metabolic processes. Within the group of mammals, there appears to be a common set of genes, some of which may be duplicated or lost in any particular species, but overall, **orthologs** (homologous genes with identical function in different species) can be found for almost every gene.

So, how relevant are gene homologies among various model organisms to the practical aspects of medicine? First, it must be remembered that most drugs are tested on animals before they begin clinical trials on humans, for reasons of safety and also for better control of the experiments. There are many reasons why basic research relies on animal models, but it all boils down to the fact that the human is a poor experimental subject. In humans, it is not possible to make controlled mutants or gene knockouts, nor to make controlled breeding experiments. There is not even a comprehensive collection of mutants. In contrast, the mouse is an almost perfect experimental subject. It is small and can easily be grown in the laboratory. It has a short generation time and is highly prolific. There are thousands of pure-bred strains that contain individual, well-characterized mutations, and any individual gene can be knocked out using standard procedures.

Animal models, particularly the mouse, promise to be extremely important in the next phase of annotating the human genome, primarily, in discovering all of the genes, and then in defining their functions in increasingly fine detail. The mouse genome is at the forefront of the genome sequencing race. Celera Genomics officially completed a version of the mouse genome

sequence in 2001. Celera simultaneously sequenced three different mouse strains (29X1/SvJ, DBA/2J, and A/J) and reported 2.5 million sequence differences (**polymorphisms**) between the strains. The Mouse Genome Sequencing Consortium published a draft assembly (96% complete) for the public mouse genome in December 2002 (Mouse Genome Sequencing Consortium 2002). The mouse genome is slowly being completed up to the "finished" standard (99.9% accurate without any significant gaps). A mixed strain assembly of the Celera data is available on the NCBI map viewer.

HUMAN–MOUSE SYNTENY

Interestingly, among related organisms, such as between mice and humans, not only are there orthologs (an identical gene in a different species), but there are extensive sections of chromosomes that contain these similar genes in the exact same order. In fact, there are a few hundred blocks of **synteny** (conserved gene order) that account for a huge fraction of the mouse and human genomes. In other words, all the genes on the human chromosomes can essentially be reorganized into the order that they are present on the mouse chromosomes with a few hundred cut and paste operations. The tens of thousands of genes have *not* been randomly shuffled by all of the millions of recombination events that have occurred since evolution separated the two species. For example, human chromosome 19 has about 1200 genes, which are found in essentially the same order in 15 segments spread across the mouse chromosomes. (Dehal et al. 2001) (see Figure 3.9).

These human–mouse syntenic segments are the focus of significant activity among genome scientists. Where a similar gene is found in both mouse and human, gene function can be investigated in the mouse by making a gene knockout. In some cases, a mouse or a human gene is known from a cDNA database, but no corresponding gene can be found in the other species by a simple

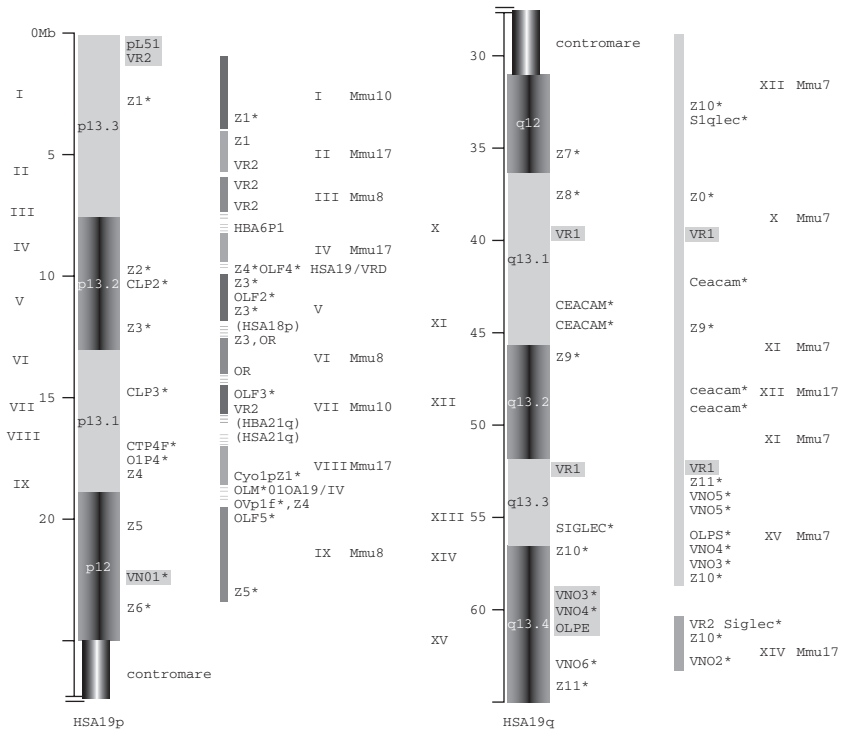


FIGURE 3.9. Human chromosome 19 and the 15 corresponding syntenic segments of mouse chromosomes.

homology search. The corresponding gene can be found in the other species by defining known genes on either side and then looking for ORFs in the corresponding region of genomic DNA. When syntenic segments of mouse and human genomic DNA are aligned, protein-coding sequences show much greater sequence similarity than do noncoding regions. It is possible to identify new genes that are not known in human or mouse simply by finding regions within aligned syntenic segments that show increased sequence similarity (and that have the potential to code for protein).

The comparison of genome sequences from humans and mice has led to a number of fascinating evolutionary insights. Essentially all of the single copy genes are equivalent between human

and mouse, with protein sequences that are generally 90–95% identical, and located in the same order on their respective chromosomes. However, some genes are tandemly duplicated in one or both genomes, leading to families of similar genes. About 30% of humans and mice genes are members of these tandemly duplicated gene clusters. Syntenic gene clusters between humans and mice contain different numbers of genes and appear to be the result of differences in the founder genes that were duplicated, differential gene loss, and independent selection in each conserved cluster since the divergence of primate and rodent lineages. A complex pattern of lineage-specific gene duplication and loss is evident. Some gene copies may become inactivated by mutations (pseudogenes), and others may develop unique tissue-specific or developmentally regulated gene expression patterns. Many of the breaks between the 15 syntenic segments occur in the middle of the tandemly duplicated clusters, so these clusters may also play a key role in large-scale genome evolution.

Some online databases have been established to explore the relationship between the mouse and human genomes—at the levels of individual genes as well as entire chromosomes. The NCBI has a nice human–mouse homology map at `http://www.ncbi.nlm.nih.gov/Homology`. It can be anchored on the human genome map (with cytogenetic positions) to show the corresponding mouse chromosomal segments, or show matching human segments on the mouse chromosome map. Known genes are indicated by name and are linked to descriptions in the **LocusLink** database. The presence of sequence-tagged sites (STSs) is also noted (see Figure 3.10).

The Jackson Laboratory, a large-scale producer of genetically pure strains of mice, also has a mouse–human genetic map on its Website (`http://www.informatics.jax.org`). This map is part of the Mouse Genome Database (Blake et al. 2001), so it is organized around mouse, rather than human, genetic information (see Figure 3.11).

Human STS	Cytogen Pos	Human Symbol	Mouse chr	Mouse Symbol	cM Position	Mouse STS
●	19p13.1-p13.2	PRKCSH	9	Prkcsh	6	●
●●	19p13.2	ELAVL3	9	Elavl3	5	
●●	19p13.2-p13.1	CNN1	9	Cnn1		●
●●	19p13.3-p13.2	ACP5	9	Acp5	6	●
●●	19q13.3	ASNA1		Asna1		
●●	19p13.2	JUNB	8	Junb	38.6	
●●	19p13.2	DNASE2	8	Dnase2	38.6	
●	19p13.13-p13.12	KLF1	8	Klf1	38.6	
●●	19p13.2	FARSL	8	Farsl		
●	19p13.3-p13.2	CALR	8	Calr	37	
●●	19p13.2	RAD23A	8	Rad23a		
●●	19p13.2-p13.1	CACNA1A	8	Cacna1a	38.5	
●●	19	ETRI01	8	Ier2	38.4	●
●●	19p13.3	NFIX	8	Nfix	38.6	
●	19p13.1	RFX1	8	Rfx1	38	
●	19p13.11	WSX-1	8	Wsx1-pending		
●●	19	EEF1D	8	5730529A16Rik		●
●	19p13.2	*RPS28* *	8	Rps28		
	19p13.2	*LYL1* *	8	Lyl1	38.5	
	19p13.2	*KIAA0973* *	8	Sast-pending	38.6	
●●	19p13.2	*GCDH* *	8	Gcdh	38.6	●●●
●●	19p13	CD97	8	Cd97	38	●
●●	19	DDXL	8	2610307C23Rik		
●	19p13.1-p12	PRKCL1	8	Prkcl1	38	
●	19p13.1	PTGER1	8	Ptger1	38	●●
●●	19p13.1	C19ORF3		Rgs19		
●	19p13.2	DNAJB1		Dnajb1		
●	19p13.12-p13.11	NDUFB7		1110002H15Rik		●
●	19	GPSN2		A230102P12Rik		●
	19p13.1	OR7C1		Olfr57		
●●	19ptcr-p13.3	SLCIA6	10	Slc1a6	42.3	
●●	19ptcr-p13.3	NAKAP95	10	Nakap95-pending		
●●	19p13.1-q12	AKAP8	10	Akap8		●
●	19p13.1	BRD4	10	Brd4		

FIGURE 3.10. A section of human chromosome 19 and its syntenic mouse segments shown in the NCBI human–mouse homology viewer.

The Sanger Centre (headquarters of the UK genome sequencing effort) has produced a nice synteny map of human chromosome 22 and the mouse genome showing that just 8 syntenic segments span all of chromosome 22 (see Figure 3.12).

SEQUENCING OTHER GENOMES

The sequencing of the human genome has received the most attention, but the genomes of many other organisms are also medically important. Many human diseases are caused by pathogens, parasites, and their vectors. Complete genomic sequences have been available for several years for many bacteria. In fact, the first

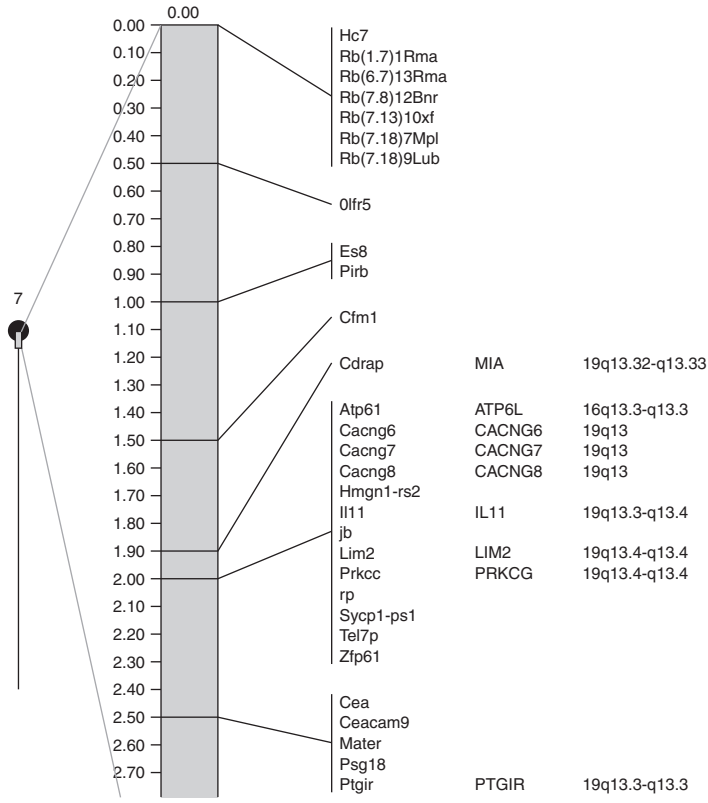


FIGURE 3.11. The Mouse Genome Database at the Jackson Laboratory.

nonviral genome to be fully sequenced was *Haemophilus influenzae* in 1995 (Fleischmann et al. 1995). In theory, protein sequences found in these genomes can be used as targets for the design of new drugs or for the development of vaccines. However, this has proved to be more difficult than expected. While computer algorithms for the detection of genes in prokaryotic genomic sequences have been quite successful, determining the function of the proteins encoded by those genes has been problematic. Surprisingly, about 40% of the genes discovered in each newly sequenced organism have no known role in cell metabolism or physiology. Even the identification of immunologically

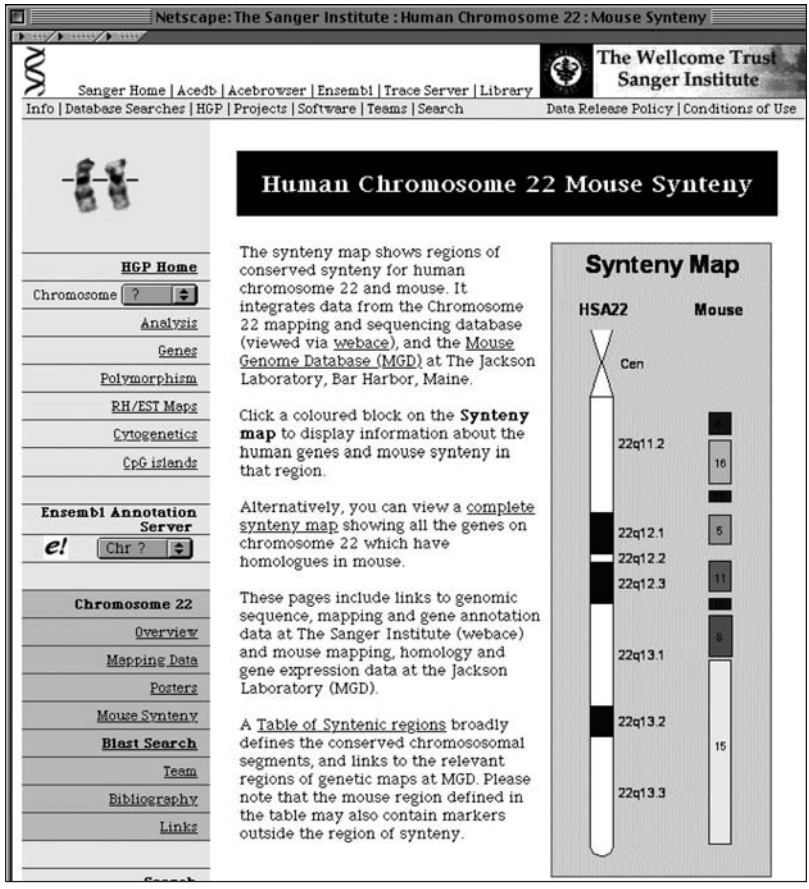


FIGURE 3.12. The Sanger Institute synteny map of human chromosome 22 and the mouse genome.

active proteins such as those that are excreted or present at the cell surface has been only moderately successful. Perhaps expectations for the use of genomic sequences to develop drugs and vaccines should be tempered by the history of the battle against viruses. The genome of the influenza virus has been known since 1982, yet millions of people are still infected each year.

The UCSC, Ensembl, and NCBI Genome Browsers are designed to provide access to complete, annotated eukaryotic genomes. Genome sequencing has also been proceeding

vigorously for many prokaryotes (bacteria) and small eukaryotes (protists and other parasites). These include many human (or plant or animal) pathogens and organisms that play critical roles in ecology. The Institute for Genomic Research (TIGR) has been in the forefront of the effort to sequence microbial genomes. The TIGR Comprehensive Microbial Resource (`http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi`) contains genome information on 370 species of bacteria and 28 archaea: "In addition to the convenience of having all of the organisms on a single website, common data types across all genomes in the CMR make searches more meaningful, and cross genome analysis highlight differences and similarities between the genomes based on a variety of criteria, including sequence homology and gene attributes." Additional TIGR genome databases include about a dozen important eukaryotic parasites (*Plasmodium, Entamoeba, Toxoplasma, Trypanosoma, Schistosoma, Theileria parva*), scientifically and economically important plants (rice, wheat, maize, potato, loblolly pine, *Arabidopsis*), and some fungal species.

The NCBI has its own microbial genome Website (`http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi`), which contains 479 complete genomes and 707 genome in progress (as of April 2007) and a similar eukaryotic genome Website (`http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi`) with 345 speices (26 complete, 319 in progress).

All of these online resources include the ability to search for sequences by similarity and to download specified regions or entire genomes.

References

Blake JA, Eppig JT, Richardson JE, Bult CJ, Kadin JA, and the Mouse Genome Database (MGD) Group. 2001. The Mouse Genome Database (MGD): Integration nexus for the laboratory mouse. *Nucleic Acids Res* **29**: 91–94.

Dehal P, Predki P, Olsen AS, et al. 2001. Human chromosome 19 and related regions in mouse: Conservative and lineage-specific evolution. *Science* **293**:104–111.

Deloukas P, Schuler GD, Bentley DR, et al. 1998. A physical map of 30,000 human genes. *Science* **282**(5389):744–746.

Feng L et al. 2000. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat Genet* **25**:239–240.

Fleischmann RD, Adams MD, White O et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**:496–512.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915):520–562.

CHAPTER

4

Bioinformatics Tools

Patterns and Tools

The success of the Human Genome Project and the foundation of the entire field of genomics is based on the automation of biochemical laboratory methods. This is the same basic concept used for robotic welders in automobile factories. However, the product of automated high-throughput genomics laboratories is information rather than cars—vast quantities of information. This information requires specialized tools for storage and analysis—and that is what bioinformatics is all about.

Bioinformatics is the use of computers for the acquisition, management, and analysis of biological information. It is a hybrid discipline that requires skills from molecular biology, computer science, statistics, and mathematics. In practice, bioinformatics specialists may also need to be fairly knowledgeable about computer hardware, networking, robotics, image processing, and anything else that impacts on the collection, storage, analysis, and distribution of biological information.

The average biologist has been forced to learn a lot about bioinformatics over the past 5–10 years. The use of DNA and

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

protein sequence data has become part of the routine daily work in most biology labs—an essential component of both experimental design and the analysis of results. It seems unlikely that use of computer tools for the manipulation of sequence data is destined to become part of the routine work of the typical physicians. However, just as physicians must know the procedure to test a throat culture for strep (*Streptococcus* infections) even when the actual microbiology labwork is done at a commercial lab, so will they need to understand how DNA sequences are used for diagnostic and therapeutic purposes—both the theoretical and the practical aspects of the technology. With that in mind, in this chapter we will describe a variety of bioinformatics tools in sufficient detail to allow a solid understanding of how they work, but it will not be a detailed tutorial on their use.

Just as the laboratory tools of the molecular biologist can be summarized as **cut**, **copy**, **paste**, and **read** (the DNA letters), the function of bioinformatics tools can be summarized as **similarity**, **alignment**, **pattern finding**, and **clustering**. The current set of commonly used bioinformatics tools was not derived from some coherent set of fundamental theoretical principles. On the contrary, the bioinformatics toolkit is a hodgepodge collection of unrelated algorithms that have been borrowed from many different branches of mathematics, linguistics, computer science, and other disciplines, and then modified through multiple generations of trial-and-error improvement. The current set of bioinformatics tools are what seem to work best to solve a number of different practical problems, but at any time a new tool may emerge from some totally unexpected theoretical background that works better for some specific task. Bioinformaticians are not picky; they will use whatever tool works best for a specific job.

In no particular order, these tools include programs to (1) draw maps of plasmids containing cloned genes, (2) search DNA sequences for sites (short patterns) that can be cut by specific restriction enzymes, (3) design PCR primers that can be used to

target a specific fragment of DNA, (4) compare one DNA or protein sequence to another or search an entire database of sequences for similarity to one query sequence, (5) line up two sequences or a group of sequences (multiple alignment), (6) join overlapping fragments of DNA sequence (sequence assembly), (7) predict the chemical and structural properties of a protein from its amino acid sequence, (8) predict the function of a new protein based on its containing subsequences conserved in known protein families (motifs), and (9) calculate a tree of evolutionary relationships among a set of sequences (phylogenetics).

Listed like this, there seems to be no commonality to these tools whatsoever. That is not true; there are a few consistent ideas common to most of the tools. The most basic is the idea of pattern recognition (see Figure 4.1). The pattern may be as simple as the 4–6 bases that define a restriction enzyme recognition site, or as complex as a conserved structural domain in G-protein-coupled receptors. The problem of computer pattern recognition has been tackled independently by many different disciplines ranging from military remote sensing to voice recognition for collect calls (reverse-charge) from a pay phone. Many of these different approaches have been tried out for various bioinformatics problems. The current crop of tools used in bioinformatics have undergone some degree of evolution and selection, but they are by no means optimal solutions to these many diverse problems. In many cases, bioinformatics experts will repeat an analysis with

FIGURE 4.1. With computers, it's easy to find patterns, even if they are not really there. These letters can be found in butterfly wings. (Kjell B. Sandred, Butterfly Alphabet, Inc., Washington, DC.) (See insert for color representation.)

several different tools because they don't trust any one tool to do the best job in all situations.

One thing that is not consistent is the amount of data that the various bioinformatics programs work on and the amount of computing power that they require to execute their tasks. Some operations, such as finding restriction sites in a plasmid, can be handled in few seconds on any desktop computer, while making a similarity search of a sequence against all of GenBank requires many gigabytes of hard-disk space and gigaflops of computing power.

SEQUENCE COMPARISON

One of the most basic questions in bioinformatics is "How similar are these two sequences?" Calculating sequence similarity is a deceptively difficult problem. For two very short sequences, you might just write them down on two slips of paper, or type them in on two lines of a word-processing program, and then try to slide them by each other to see if there is any group of letters that line up (see Figure 4.2).

The best overall similarity might require some mismatches or inserting some gaps in one or both sequences. Is a short identical region better than an overall match with a scattering of mismatches and gaps? You would think that some mathematicians must have worked on this problem and come up with an optimal way of calculating similarity; and indeed they have. In fact, calculating a similarity score for two sequences and finding the best alignment between them turns out to be one and the same

GATGCCAT**AGAGCGTAGTC**GTTCCCT ←
→ CTAG**AGAGCGTAGTC**AGAGTGTCTTTGAGTTCC

FIGURE 4.2. Alignment of two sequences by hand.

problem, and a good solution to this problem has been available since the early 1970s (Needelman and Wunch 1970).

Similarity scores can be calculated for both DNA and protein sequence pairs with one modification. DNA–DNA similarity is almost always scored as either an identical match, or a mismatch between two bases, with some penalty for inserting a gap in either sequence. However, when aligning two proteins, there can be various shades of similarity between pairs of amino acids that are not identical. Some pairs of amino acids have similar chemical properties, while other pairs have similar codons—so they are separated by just a single mutation event. In practice, the best solution was found to be based on the frequency that one amino acid replaces another at a given position in sets of closely related proteins. This table of natural mutation rates between every possible pair of amino acids can then be used as a scoring matrix for amino acid sequence alignments (Dayhoff et al. 1978).

Needelman and Wunch described the use of the computational technique of **dynamic programming** for sequence alignment in a paper published in 1970, and this method was rapidly implemented in several simple computer programs. The Needelman–Wunch method finds the best overall alignment between two whole sequences, but this does not always find the best alignment if just a small part in the middle of one sequence matches a part in the middle of another sequence. An improved method of **local alignment** using dynamic programming was developed by Smith and Waterman (1981). A local alignment finds the best subsequence match between a pair of sequences (see Figure 4.3).

Local alignments are also used to answer another common bioinformatics question: "What sequences in a database are most similar to this sequence?" Or more generally: "Is this sequence like anything that anyone else has ever seen before?" This requires making a pairwise comparison between a query sequence and every sequence in the database, then choosing the

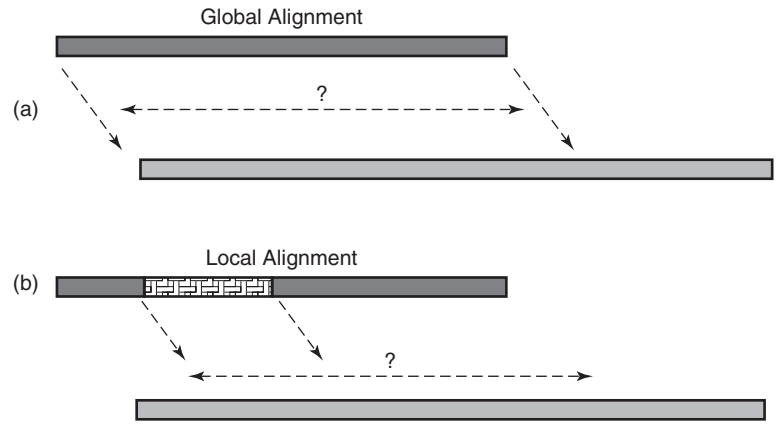


FIGURE 4.3. The difference between global alignment (a) and local alignment (b).

database sequences that give the best overall alignment. The Smith–Waterman method finds the optimal match in a comparison between two sequences, but it is a very slow program when it is used to compare a sequence against all of the other sequences in a large database. In order to make faster comparisons, some shortcuts had to be found. So far, the best methods for comparing a single sequence to a database involve quickly scanning each database sequence for short bits (known as "words") that match the query sequence, then throwing away all of the sequences that do not have any good short matches. Then each short match is used as the start of a local alignment that is extended in both directions until no more matches are found. A score is calculated for each alignment, then the database sequences that have the regions that align best are shown as the results of the search. This type of database search is called a **heuristic** because it is approximate. The tradeoff for greater speed is that it is possible to miss some important matches—particularly those with moderate overall similarity, but no short regions of really high matching—and to get some (or many) false-positive matches.

Within this general area of heuristic searching for local alignments, a number of different computer programs have

been developed. Pearson and Lipman (1985) created FASTA, a fast alignment program, in the mid-1980s and have continued to refine it through many generations of optimization and improved functionality (Pearson 1990). Altschul, Gish, Lipman, and others at the NCBI created a rival program called **BLAST** (basic local alignment search tool) (Altschul et al. 1990), which is even faster than FASTA, if sometimes a bit less sensitive.

Another important feature of both BLAST and FASTA is that they return an ***E* value** (expected value), which is a statistical measure of the quality of each match. *E* values are actually a measurement of how likely it would be to find a match of a similar quality if a search were done with a randomly generated sequence the same length as your query sequence. In more formal terms, an *E* value is a measure of the probability that an observed match between two sequences is due to chance. In common language, an *E* value is a measurement of the likelihood that your match is bogus. However you look at it, a smaller number is a better match. *E* values are generally small fractions expressed in exponential notation (e.g., 3.2×10^{-56}), so the larger the negative exponent, the smaller the *E* value, and the more significant the match. Typically, matches with *E* values smaller than 0.05 are considered significant—just as *P* values are used in classic statistical tests. However, the *E* value of a match depends on a number of factors, including the length of the query sequence (short queries cannot give highly significant matches) and the size of the database being searched (the larger the database, the greater the chance for bogus matches).

BLAST has undergone many generations of change and optimization (Altschul et al. 1997), and it is currently the most popular tool for comparing sequences. The NCBI operates a free BLAST server on its Website that is used by many thousands of scientists each day. No one has suggested that either BLAST or FASTA provides an optimal, or even a good solution for similarity searching of databases, they are simply the best tools that are currently available. Some bioinformatics groups with

lots of money have resurrected the Smith–Waterman search method in custom built supercomputers in order to make more precise and (they hope) more sensitive searches.

Using BLAST to compare pairs of sequences or to compare one sequence to a database is a very common research task, but it is not typically used in diagnostic or forensic medicine. BLAST is generally used when a query sequence is completely unknown. Most medical applications of DNA sequence comparison involve looking for small changes in known sequences. If a direct comparison of sequences is made, it will usually be done by multiple alignment.

MULTIPLE ALIGNMENT

Sometimes it is necessary to make comparisons among a group of related sequences, a **multiple alignment** (see Figure 4.4). This is often important in the study of protein families and motifs. It is also the starting point for evolutionary studies (**phylogenetics**). In clinical medicine, it may also be done in the study of a particular gene in samples collected from many people. Since good algorithms and computer programs are available for aligning two sequences, one might expect that aligning groups of

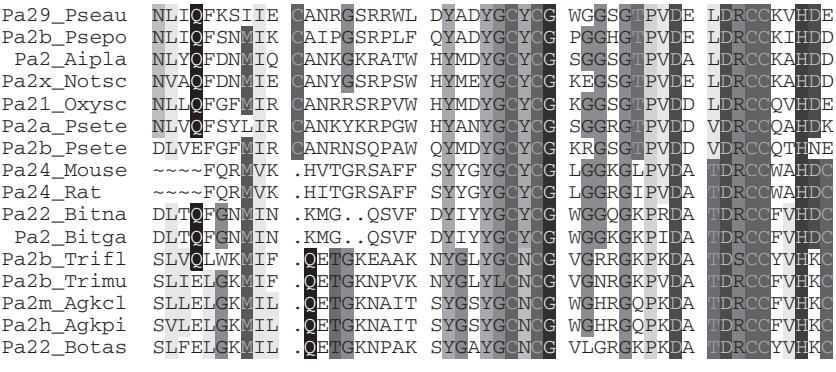


FIGURE 4.4. A multiple alignment of the *Pa* gene from 15 different species.

sequences would just be an extension of these same methods. However, from a computational perspective, it is surprisingly difficult to align a group of sequences. The problem is that each additional sequence added to the alignment requires that adjustments be made to every other sequence (inserting new gaps, shifting mismatches to accommodate a new consensus, etc). In fact, if a dynamic programming approach is used, each new sequence *exponentially* increases the amount of computing required to build the optimal alignment. The amount of computing gets huge if more than about 10 sequences are aligned.

Instead of using a dynamic programming approach to calculate the absolutely optimal alignment of a group of sequences, a shortcut approach called **progressive pairwise alignment** has been developed (Feng and Doolittle 1987). This method relies on a quick set of pairwise comparisons among all of the sequences to be aligned to estimate the relative amounts of similarity. Then the most similar pair of sequences is aligned (using a dynamic programming algorithm). A consensus sequence is generated from this alignment, and the next most similar sequence is aligned with this consensus. Then a new consensus is calculated, and this process is repeated until all of the sequences are incorporated into the alignment. This process works fairly well, but the final alignment produced is approximate rather than optimal. Also, the sequences to be aligned must all be about the same length and have a pretty high level of similarity throughout. It is possible to detect and align small similar regions located within a group of larger sequences, but this requires a combination of pattern detection and multiple alignment algorithms.

Understanding the difference between an optimal alignment and an approximate one is less important than understanding the difference between a computationally optimal solution and a biologically meaningful one. A computational algorithm finds a maximal or minimal score for some set of rules, but these rules, however complex, always represent a simplified model of the

true biological situation. Biology is full of exceptions to rules and special situations. It is often the case that the alignment produced by a computer program will need to be adjusted by hand to preserve biologically important regions (i.e., no gaps inserted in the middle of an enzyme's active site).

PATTERN FINDING

Another broad class of bioinformatics tools are used for pattern finding. Many different kinds of patterns are present in DNA and protein sequences. Some are very simple, such as the DNA sites recognized by restriction endonuclease enzymes (for example, **GAATTC** is recognized by the enzyme EcoRI). Others, such as conserved protein domains that fold into functional three-dimensional structures (**motifs**), are very complex. Therefore, the various bioinformatics tools used to detect these pattern reflect these different levels of complexity.

Simple patterns such as restriction enzyme sites in DNA can be found by an exact pattern-matching tool – just like the `find` command in any word-processing program. A slightly more sophisticated pattern-matching tool can include a mismatch at a specified location, or anywhere in the pattern. However, many biological patterns such as promoter sequences in DNA and protein functional domains require more flexible pattern-searching tools. A more complex pattern can allow for a list of different letters that can be considered to match at each position in the pattern (ambiguities) and regions of variable size where any letter is allowed: `[LIVMFY]-x(2)-[STG]-G-x(2,4)-[ST]-C`. In this example, dashes separate positions in the pattern, letters in brackets can be substituted at that position, `x` means that any letter is allowed at that position, `x` followed by a number such as `(2)` denotes any two letters, and `x` followed by a range `x(2,4)` means that a variable number of letters is allowed (ranging from 2 to 4).

Several databases of biological patterns have been created. Promoter sequences are stored in the Eukaryotic Promoter Database (EPD: `http://www.epd.isb-sib.ch`), and transcription factors (a somewhat broader category that includes enhancers and other regulatory elements that may be located some distance from the coding region of a gene) are collected in the TransFac database (`http://transfac.gbf.de/TRANSFAC`). Patterns for conserved protein domains can be found in Prosite, a dictionary of protein sites and patterns (Hoffman et al. 1999) (`http://www.expasy.ch/prosite`). The essential quality of a pattern is that it is built by hand by an expert biologist who has spent a lot of time scrutinizing a conserved motif in a group of related sequences. A variety of simple computer programs that can run on any PC, UNIX machine, or on free Webpages are available to search any given DNA or protein sequence for matches with the appropriate set of patterns. These pattern-searching programs are very fast and they do not use much computer power, even for searches with thousands of different patterns.

PROFILE SEARCHES

Pattern searches, even with ambiguities and variable-sized gaps, have some serious limitations. It is a form of exact matching, so only those variations of a real biological pattern that have been specifically included in the description of the pattern will be found. So by definition, new sequences that have unexpected variations of this pattern will not be found. This is like not being able to find a file on your computer's hard drive with the `find file` utility because you don't remember how to spell the file's name. The ability of BLAST and FASTA to find *similar* sequences would be very useful in pattern finding, particularly the ability to use a matrix of amino acid similarities when evaluating protein patterns.

The concept of recurring patterns is particularly well developed in the study of protein sequences. If all of the protein sequences in any of the major databases (GenBank, SwissProt, PIR, etc.) are compared to each other using a similarity tool such as BLAST or FASTA, it is immediately obvious that many of them fall into groups. Furthermore, each of these groups contains proteins with similar functions such as kinases, methylases, and cell surface receptors, so the protein groups are actually protein families. Detailed inspection of the sequences within each protein family reveals certain regions that are highly conserved among all of the proteins in that group. Furthermore, many of these conserved regions (known as **motifs**) form three-dimensional (3D) structures that play an essential role in the function of those proteins—an active site for an enzyme or a crucial protein fold that binds with a ligand or in a protein–protein interaction. The functions of new genes can be predicted more reliably using databases of known protein families and pattern-finding tools. A pattern-finding tool may be able to use the similarity of these motifs to identify new members of each protein family that may have too little overall similarity to be identified with BLAST or FASTA. Also, a new protein may have similarity to two or more motifs—which can provide useful information about its potential function. These multiple conserved regions would be difficult to interpret in the results of a BLAST or FASTA search (Lesk 1988):

> In some cases, the structure and function of an unknown protein which is too distantly related to any protein of known structure to detect its affinity by overall sequence alignment may be identified by its possession of a particular cluster of residues classified as a motif. The motifs arise because of particular requirements of binding sites that impose very tight constraint on the evolution of portions of a protein sequence.

An enhanced pattern searching method called **profile analysis** was developed by Gribskov (Gribskov et al. 1990). A profile is a mathematical description of a conserved region, domain,

or motif—usually in a protein, but DNA profiles can also be constructed—which is built from a set of sequences that all share this motif. All of the sequences are aligned (a multiple alignment), and then the frequencies of each letter are calculated for each position (the number of times that each letter appears in a column of the multiple alignment). The result is a **position-specific scoring matrix** that fully describes that motif across all of the sequences used in the multiple alignment. The matrix can be filled out with zeroes for all possible amino acids (or DNA bases) that do not occur in that position in any of the sequences in the set, or the matrix can be filled out with values chosen from a table of natural amino acid mutation rates (a weighted average of the values for each amino acid in the alignment at that position). Then the matrix can be used in place of a single sequence for a modified BLAST/FASTA-type similarity search, against either a single sequence or an entire database of sequences. Alternately, a database of profiles can be created, such as for protein domains, and a new sequence can be searched for similarity to all of these profiles.

A number of profile-based databases of protein motifs have been created. The simple patterns in ProSite have been supplemented with profiles based on multiple alignments of the conserved regions of a selected set of proteins for each family. This set of ProSite profiles has been expanded in the BLOCKS database (`http://www.blocks.fhcrc.org`) by using the ProSite profiles to search all of the proteins in the SwissProt database for more members of each protein family. The ProDom database (`http://prodes.toulouse.inra.fr/prodom`) is built in a completely automatic fashion, first clustering all of the proteins in SwissProt using a similarity program (a form of BLAST), then building a profile for each cluster (see Figure 4.5). This is surprisingly useful, since many clusters are identified that contain proteins that have no known function (yet!), or small conserved domains present in a group of otherwise unrelated proteins.

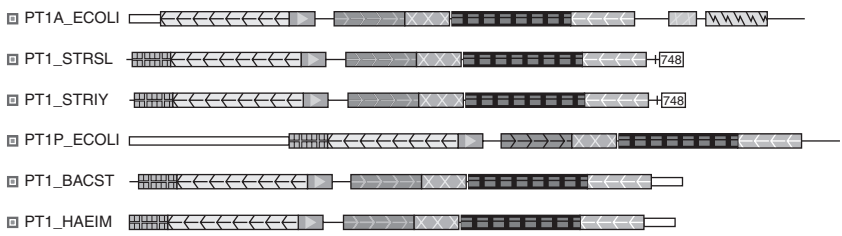


FIGURE 4.5. A set of conserved domains from the ProDom database.

HIDDEN MARKOV MODELS

The concept of using pattern-based searching for conserved domains among related proteins makes intuitive sense. As we learn about a family of proteins, information about a conserved site, such as the catalytic domain for a family of enzymes with related biochemical functions or a DNA-binding domain for a family of regulatory proteins, should be useful in identifying additional proteins in that family. However, this does not always turn out to be the case. A standard, brute-force BLAST search is sometimes more sensitive at discovering new members of a protein family than a profile-based search method. There are a few reasons for this. First, profiles consider information about only a small region of a protein (the conserved motif), but there may be subtle bits of information located elsewhere in the protein—either near the motif or distant from it—that can contribute to the BLAST similarity score. Also, profiles evaluate each position in a motif separately—how well does this letter match the corresponding column of the matrix—and ignore interactions between different positions.

A more sophisticated pattern search tool known as a **Hidden Markov Model** (HMM) has been developed that factors in the influence of neighboring amino acids in constructing and searching for motifs in proteins. The math behind HMMs is rather complicated, so let it suffice to say that this is a pattern

analysis technique that was developed in linguistics, but has been adopted by bioinformatics and that it uses a lot of computer power to make searches. The Pfam database contains about 3000 HMM profiles for protein domains. Most of the profiles in Pfam were created from hand-built multiple alignments of conserved domains from groups of related proteins (protein families). An additional set of domains come from an automated clustering of all of the proteins in SwissProt. These profiles can then be compared to an unknown protein (or a predicted protein from a genome sequence) to identify functional domains.

Many proteins contain more than one functional domain. Sometimes each of these domains match unrelated protein families. Walter Gilbert (1978) has suggested that new proteins evolve by switching functional modules (motifs). In many cases these motifs correspond to exons, so that recombinations that occur between introns can easily create new multidomain proteins. A similarity search with BLAST or FASTA will find regions of alignment with other proteins that have each of these functional domains, and try to extend these alignments across the entire sequence of the protein. The similarity programs rank these alignments by their overall score according to the percentage of identical and similar amino acids and the length of the aligned region between the query protein and the most similar protein in the database. This can be quite confusing when a protein shares just one domain with one protein family and another domain with another protein family. A Pfam search deals sensibly with these "shuffled chunks" of proteins by showing where each profile matches the query sequence and giving a score for how well just that region matches the profile of that conserved domain. HMM searches with Pfam profiles have been very helpful in the initial stages of annotating all of the proteins discovered in whole-genome sequencing projects, including the human genome.

PHYLOGENETICS

Databases of protein domains and motifs are created from clusters of proteins that contain regions of similar sequence. It is important to distinguish between a similarity based on true homology—which is the result of genes that shared a common ancestor, as opposed to a strictly functional similarity based on protein regions that contain a disproportionate amount of one or two amino acids (i.e., a proline-rich region or a membrane-spanning region that contains many nonpolar amino acids). It can also be confusing to look at a family of proteins that share a common domain and see several proteins from a single species and a number of others from different species. In order to make some sense out of the relationships between the members of a protein family, it is necessary to understand something about the process of evolution at a molecular level.

New genes are created by two distinctly different processes: gene duplication and speciation. Genes (or entire chromosomes) can be duplicated in a number of different ways during the processes of replication and recombination. Once a species contains two copies of a gene, random mutation events will cause independent changes in the two sequences. These mutations may lead to one gene copy taking on new a function while the other copy provides the original function. These two gene copies are known as **paralogs**. Alternately, if two populations of a species are separated for a long time, each population will accumulate different mutations in its genes until the two populations form two distinctly different species. Now each gene in species A has a similar but not identical match in species B. These are known as **orthologs**.

Gene families are created by complex combinations of gene duplication and speciation, so it is not always obvious when looking at two gene sequences from two species whether they are orthologs or paralogs. This is extremely important if we are

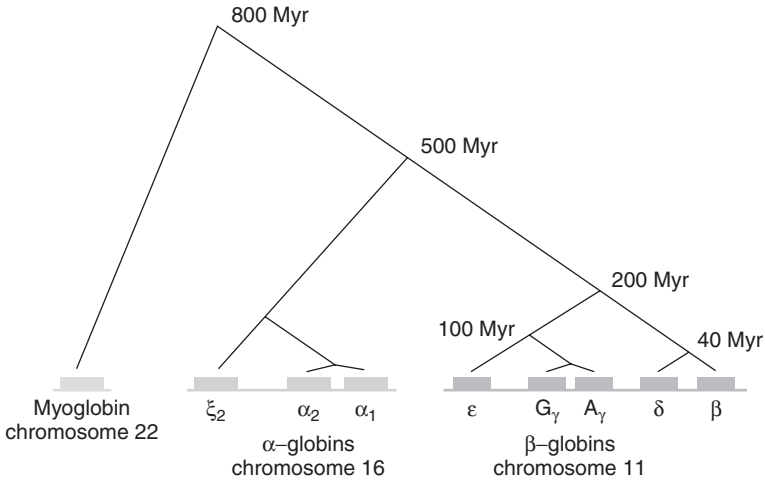


FIGURE 4.6. The globin gene family has been created by gene duplications.

going to rely on comparisons between species to assign functions to genes—such as mouse knockout experiments to explore the function of corresponding human genes. The only reliable method for determining the relationships between the members of a multigene family is to calculate an evolutionary tree. Figure 4.6 illustrates a tree containing the family of paralogous globin genes in humans. There is an entire branch of biology called **taxonomy** which is devoted to defining the evolutionary relationships between species, and the majority of work in this field in the past few decades has focused on methods for the use of DNA and protein sequence data.

Scientists working in numerical taxonomy or **phylogenetics** have created some very complex mathematical approaches to calculating the relationships between gene sequences, which goes far beyond the scope of this book. However, unlike the work on sequence similarity, multiple alignment, or protein domains, there are no broadly accepted phylogenetic methods that consistently achieve good results. Some methods focus on grouping sequences on the basis of their absolute numbers of similar

and different bases (or amino acids for proteins). This creates a good quantitative estimate of current similarities, but ignores evolutionary processes that may have significantly changed some genes in short periods of time. Other methods focus on re-creating the exact process of mutations from a common ancestor to create the current set of observed gene sequences, but without data from all of the intermediate forms (many of which are presumably extinct), there can never be an absolutely accurate calculation. None of these methods can adequately account for the messy realities of evolution, which may include the mixing of different genes by recombination, hybridization between different species, and other phenomena that don't follow clean mathematical models.

Despite these theoretical hazards, phylogenetic software tools can be used to define the relationships between a cluster of genes in closely related species with reasonable reliability. The basis of all phylogenetic calculations is a multiple alignment of the relevant genes. Within a multiple alignment, pairwise similarities are calculated, then a tree is built from the branches back to the base by joining the most similar pairs, then the next most similar, and so on. In general, orthologs cluster together most tightly unless there has been a recent gene duplication event that affected one species but not others.

BIOTECHNOLOGY EXERCISE

DNA cloning technology has become a routine technical skill that can be mastered in a few months of laboratory training. Many advanced high school biology classes use restriction enzymes, gel electrophoresis, PCR, plasmids, and bacterial transformation in laboratory exercises.

The availability of genome information has created shortcuts that were not available to scientists in the late 1990s. As a result, the informatics aspect of labwork has become increasingly important. A few hours on the computer can save weeks (or months) of time in the lab.

For example, the entire human genome sequence is available online at the NCBI Website (`http://www.ncbi.nlm.nih.gov/projects/genome/guide/human/`), the UCSC Genome Browser (at `http://genome.ucsc.edu`), and with the Ensembl Genome Browser (at `http://www.ensembl.org/Homo_sapiens`). Each of these sites contains a comprehensive list of mRNA (cDNA) sequences for known (and predicted) human genes. So, to clone a human gene, all that is needed is to look up the gene, design a pair of PCR primers, have the primers made for you at a commercial DNA synthesis lab, and run a PCR reaction with a tube of human genomic DNA or cDNA (which can be obtained from a number of different commercial suppliers). If you are clever, you will include restriction enzyzme recognition sites in the PCR primers that are not present in the gene sequence. Then, after PCR amplification, your gene can easily be inserted into an appropriate plasmid vector.

The following example follows the methods used by G. Lee et al. (2004) to clone the human casein kinase II β gene. They used known genome sequence to design PCR primers, then amplified a DNA fragment from a human brain cDNA library and cloned it into an expression vector. *Quiz question*: Why use cDNA rather than genomic DNA? (*Answer*: The mRNA is generally much smaller since it lacks introns, it can be expressed in

bacteria that do not splice out introns, and the lack of a promoter sequence makes it easier to express the gene under the control of another promoter.)

Start by looking up `CKII BETA` in the NCBI human genome database. This should bring you to the gene entry for `CSNK2B`, which is located on chromosome 6 at `6p21.3`. The RefSeq messenger RNA for this gene is `NM_001320`. You can also get to the `NM_001320` sequence on the UCSC genome browser by typing "casein kinase II beta" on the human genome query page (`http://genome.ucsc.edu/cgi-bin/hgGateway`) and then clicking on `CSNK2B`, and then on `NM_001320`.

Have a look at the GenBank entry for `NM_001320`. This is a 1128-bp mRNA, but the protein-coding sequence runs from base 341 to 988. The sequence from 1 to 340 is the 5′ untranslated region (5′ UTR) of the gene and 989–1128 is the 3′ UTR. The 5′ and 3′ UTRs contain sequences that control the translation efficiency and the posttranscriptional regulation of gene expression, but they make an ideal location to use for the design of a pair of PCR primers to amplify this gene from a cDNA library. The first and last 150 bases will be adequate to use for primer design.

There are many tools available for the design of PCR primers. The Primer3 Website at MIT (`http://frodo.wl.mit.edu/primer3/input.html`) (Rozen and Skaletsky 2000) is one of the best. Paste the `NM_001320` mRNA sequence in the text input box (see Figure 4.7). We want the primers to be located in the 150 bases at the ends of the sequence. One way to accomplish this is to use the TARGETS field to specify a target that begins at base 150 and extends for 828 bases ("`150, 828`"). It is also useful to specify the product size range from 828 to 1128. All of the other parameters can be left at their default values for now, but it is useful to know that primer size and annealing temperature can be adjusted, and that primers are screened for self-complementarity. Hit the **PICK PRIMERS** button. Copy and

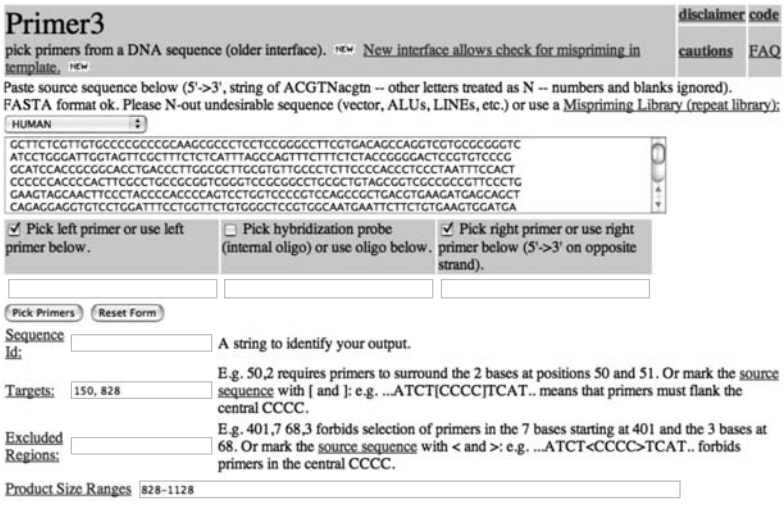


FIGURE 4.7. The Primer3 PCR primer design tool.

save the primers designed by Primer3, and also take note of the position of these primers in the `NM_001320` sequence.

Once you have the sequences for a pair of primers that will amplify the target gene, it is useful to add a few bases that will provide restriction sites that will make it easy to clone the PCR product into the vector of your choice. For this exercise, we will use the common plasmid `pUC19`. The goal is to choose two restriction enzymes from the `pUC19` polylinker region that are not found in the `NM_001320` sequence. The polylinker is a set of enzymes that cut the plasmid only once and interrupt the coding sequence for the B-galactosidase gene, blocking the formation of a blue color when the plasmid is transformed into bacteria and the bacteria are grown on plates containing X-gal. Using this system, it is easy to pick white colonies that contain the plasmid with the cloned sequence and avoid blue colonies that contain only the `pUC19` plasmid with no inserted DNA.

There are many tools available for mapping restriction enzyme sites in DNA sequences. The NEBcutter2 Website (`http://tools.neb.com/NEBcutter2`) is very well designed and easy

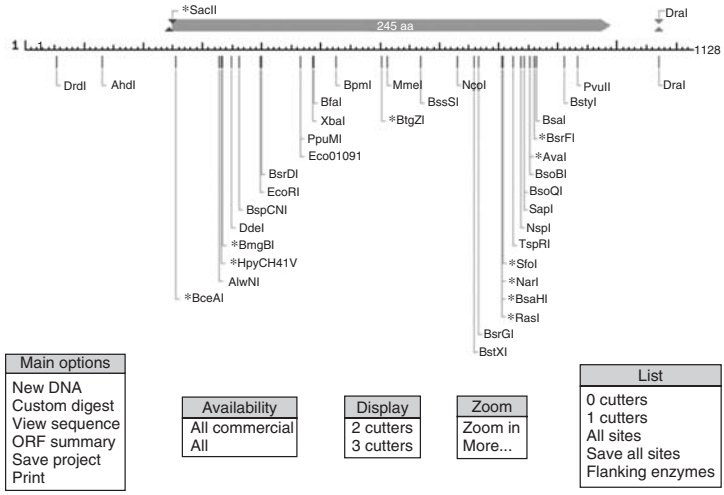


FIGURE 4.8. The NEBcutter2 display of a restriction map for the sequence of NM_001320.

to use (see Figure 4.8). Paste the NM_001320 sequence into the text input box for NEBcutter2 and hit the **SUBMIT** button. Enzymes that do not cut this sequence are listed as "0 cutters."

A restriction map of pUC19 is available from the pulldown menu of "standard sequences" at the right side of the NEBcutter window. The pUC19 polylinker contains unique sites for 13 common enzymes: HindIII, SphI, PstI, SalI, XbaI, HindII, BamHI, Sma, KpnI, EcoRI, AvaI, SacI, and ApoI (see Figure 4.9). Find two enzymes that cut the pUC19 polylinker, but do not cut the NM_001320 sequence. Use the NEBcutter2 Website to look up the recognition sequence for your two enzymes, and add these sequences to the 5′ ends of the forward and reverse PCR primers.

```
pUC19 Polylinker region:
 5' ATG.ACC.ATG.ATT.ACG.CCA.AGC.TTG.CAT.GCC.TGC.AGG.TCG.ACT.CTA.
                          HindIII SphI    PstI    SalI    XbaI
                                                  HindII
                                                  AccI

    GAG.GAT.CCC.CGG.GTA.CCG.AGC.TCG.AAT.TCA.CTG.GCC. 3'
      BamHI  SmaI KpnI    SacI    EcoRI
                                  ApoI
```

FIGURE 4.9. The polylinker sequence of cloning vector pUC19.

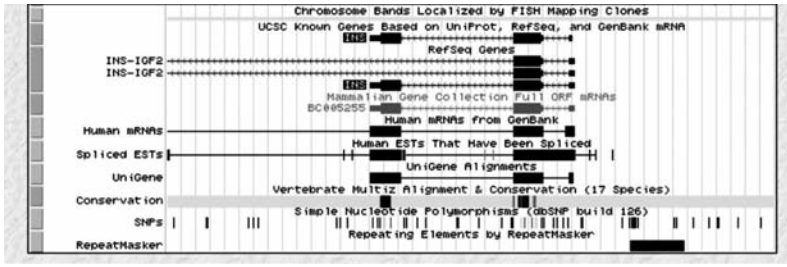


FIGURE 4.10. The human insulin gene (INS) shown in the UCSC Genome Browser.

Once the informatics work is done, the experimental procedure in the laboratory is straightforward. Set up a PCR reaction with the primers and a human cDNA library (available from several commercial sources). Run the products of the PCR reaction in a gel and isolate the amplified fragment as a band that corresponds to the predicted size (the region between the primers plus the primers themselves). Isolate the DNA from the gel, then cut with the two restriction enzymes. Cut `pUC19` with the same two restriction enzymes. Ligate and transform competent *E. coli* bacteria. Spread on ampicillin + X-gal plates, incubate overnight, and pick the white colonies. Then you can grow batches of the bacteria containing the recombinant plasmids, isolate plasmid DNA, and you have unlimited amounts of cloned casein kinase II gene. You can cut the CKII gene back out of the plasmid using the same two enzymes and put it into another vector with a promoter designed to express the protein in whatever type of target cell might be useful.

Repeat this cloning exercise with the human **insulin** gene. It will be easier to find the correct mRNA sequence if you look up the gene symbol "INS" rather than the word "insulin" (see Figure 4.10).

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**:403–410.

Altschul SF, Boguski MS, Gish W, Wootton JC. 1994. Issues in searching molecular sequence databases. *Nat Genet* **6**(2):119–129.

Altschul SF, Madden TL, Schäffer AA et al. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* **25**:3389–3402.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins, matrixes for detecting distant relationships. In Dayhoff, MO (ed), *Atlas of Protein Sequence and Structure*, Vol 5, pp 345–358. National Biomedical Research Foundation, Washington, DC.

Feng DF, Doolittle RF. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* **25**:351–360.

Gilbert W. 1978. Why genes in pieces? *Nature* **271**:501.

Gribskov M, Luethy R, Eisenberg D. 1990. Profile analysis. *Meth Enzymol* **183**:146–159.

Hofmann K, Bucher P, Falquet L, Bairoch A. 1999. The PROSITE database, its status. *Nucleic Acids Res* **27**:215–219.

Lee G, Tanaka M, Park K, Lee SS, Kim YM, Junn E, Lee SH, Mouradian MM. 2004. Casein kinase II-mediated phosphorylation regulates alpha-synuclein/synphilin-1 interaction and inclusion body formation. *J Biol Chem* **279**(8):6834–6839.

Lesk AM. 1988. In Lesk AM (ed), *Computational Molecular Biology*, pp 17–26. Oxford University Press, Oxford, UK.

Needelman SB, Wunch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**:443–453.

Pearson WR, Lipman DJ. 1985. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85**:2444–2448.

Pearson WR. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. In Doolittle RF (ed), *Methods in Enzymology*, Vol 183, pp 63–98. Academic Press, San Diego, CA.

Pearson WR. 1997. Identifying distantly related protein sequences. *CABIOS* **13**:325–332.

Rozen S, Skaletsky HJ. 2000. Primer3 on the WWW for general users and for biologist programmers. In Krawetz S, Misener S (eds), *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, pp 365–386. Humana Press, Totowa, NJ.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol* **147**:195–197.

CHAPTER

5

HUMAN GENETIC VARIATION

Much of what makes us unique individuals is represented by the differences in our DNA sequence from other people. The data from the Human Genome Project indicates that on average any two people have 99.9% identical DNA sequences. Yet these 0.1% of differences, spread over 3.2 billion bases of DNA, amount to a significant number of distinct genetic traits that uniquely distinguish the genome of every person. In fact, the Human Genome Project now estimates that there are just 24,000 functional genes in the human genome. For each of these genes, there exist many different variant forms (known as **alleles**) in the human population, and each person has a unique combination of these forms.

MUTATION

Many heritable diseases are caused by a defect of a single gene. Most of these genetic diseases are recessive—a mutation has created a dysfunctional allele that does not produce an essential protein, and therefore a person who inherits two such mutant alleles cannot make the protein. For some genetic diseases, a particular mutant allele exists at an elevated frequency within a genetically

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

isolated population (e.g., Tay–Sachs disease), while other diseases can be caused by any combination of a number of different mutant alleles of the gene (e.g., phenylketonuria). Other heritable diseases, such as specific forms of cancer and heart disease, are caused by a combination of alleles from number of different genes—either in an additive fashion, or as a specific combination of interacting alleles. Taken by itself, each allele that contributes to a multigene disease may not be an obviously disfunctional mutation, but rather a variant allele that is present in the population at some measurable frequency.

In evaluating the medical significance of these variable genetic traits, it is important to keep in mind the mechanisms by which they arise and are spread through populations. DNA is an extremely stable molecule, and the cellular machinery that copies DNA during the processes of growth and sexual reproduction work with very good fidelity. However, errors do creep in, and these are called **mutations**. External factors such as radiation and chemical mutagens can damage the DNA molecule, which may lead to changes in the sequence of bases (see Figure 5.1).

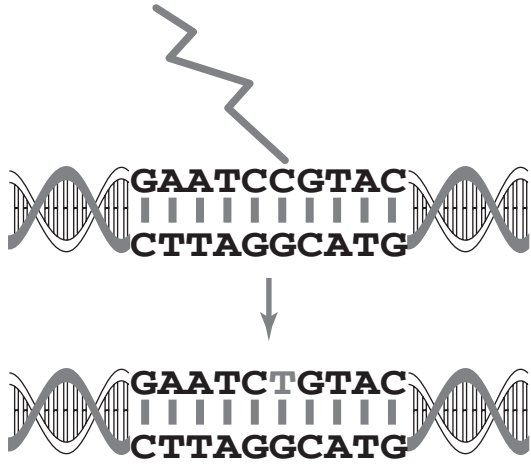


FIGURE 5.1. Mutations can be caused by external factors.

There are internal DNA repair mechanisms that detect and repair mismatched bases, but these can sometimes have the result of changing a base to match its mutated complementary partner. There is also a low level of errors created by the DNA replication machinery itself. If just one in a billion bases is copied incorrectly, then every new cell (and every new person) will have a few brand-new mutations.

Mutations that occur in somatic cells (all the cells of the body except the sex cells) affect only that cell and its direct progeny within the body of one individual. Mutations in the sex cells (sperm and eggs, or the germline stem cells that produce the gametes) can be passed on to the children and become part of the genetic diversity of the human population. In fact, every human is born with a few new mutations, and many more occur in their somatic cells throughout their lifetime. Since the human population is constantly expanding, just by chance, many new mutations are constantly being added to its store of genetic diversity.

The vast majority of these mutations have no effect on proteins. Remember that only about 1% of the DNA in the human genome is transcribed into RNA and that much of that is spliced out as introns. For those differences that do occur in exons, some lead to no change in amino acids (silent mutations), while others do change amino acids, but do not affect the function of the protein (conservative changes). It is those rare mutations in coding regions that change the function of a protein that typically constitute the different observed alleles of genes such as blue or brown eyes or a single-gene genetic disease such as cystic fibrosis or Huntington's disease. In fact, genetic disease could be considered to be evolutionary selection in action against mutations that cause significant damage to essential proteins. Many different mutant alleles may exist for a single gene. These may be regarded as different misspellings of the protein sequence, all of which lead to a nonfunctional protein product, which is essential for some physiologically significant process.

Mutations in non-protein-coding regions of DNA can still have phenotypic effects if these mutations affect intron splicing or gene expression. A mutation in an intron that causes incorrect splicing can produce a nonfunctional protein product just as surely as a coding mutation. A mutation in a promoter or transcription factor binding site that causes changes in gene expression may have significant phenotypic effects and be just as important in evolution as those that change protein sequences. In rapidly evolving systems, such as defense against pathogens, which can rapidly generate new strains, changes in gene expression may provide a more subtle and flexible response than changes in protein sequence. Genetic medicine has concentrated on single-gene inherited diseases that have extremely obvious phenotypes, which are often caused by mutations that result in complete loss of function for a critical protein. However, even among these extreme situations, there are a number of mutations that have been traced to non-protein-coding regions. It is likely that regulatory mutations will play a larger role in multigene complex disorders.

Not all mutations or variants in gene sequences can be cleanly defined as either deleterious or beneficial. In some situations, a gene variant may offer an advantage while in other situations it is a weakness. The sickle cell mutation of the β-hemoglobin gene is a perfect example. People with a single copy of this allele (heterozygotes) have substantial resistance to malaria, while two copies of this allele (homozygotes) experience damaging and sometimes lethal malformations of their red blood cells. In regions of Africa where the incidence of malaria is very common, the protection offered to people with a single copy of the sickle cell allele has balanced the damaging effects on people who receive two copies, so that the trait has been maintained in the population. There are obviously other situations where the tradeoff between beneficial and harmful effects of various gene alleles is less striking. The important concept to keep in mind is that the advantage or

disadvantage of a particular mutation depends on the environment in which the individual is living.

Variant alleles that have no apparent phenotypic effect can have an important survival value under some other set of environmental circumstances. For example, the δ-32 variant of the CCR5 gene provides substantial resistance to HIV. The range of genetic variations present in the whole human population can be considered a reservoir of potential solutions and adaptations accumulated over millions of years, which may become useful in some new environment. Advantages can also occur only with a particular combination of alleles from several different genes. Our current understanding of human genetics is not sophisticated enough to detect these types of interactions, but discovery of these complex traits is one of the primary objectives of plant breeding.

For the large number of mutations that have no effect on protein sequences, splicing, or expression, variants move randomly thorough the human population following the fate of the section of chromosome on which they reside. These are often called **neutral mutations**, since they do not directly affect natural selection. However, mutations that do not affect protein coding regions or gene expression can still be useful diagnostic markers if they are located near genes with important medical effects.

Single-Nucleotide Polymorphisms

A base change at a specific position on the genome is officially considered to be a **polymorphism** in the population when the frequency of the most common base at that position is less than 99%. These single-base changes are known as **single-nucleotide polymorphisms** (SNPs). Sometimes small insertions and deletions are also designated as SNPs. SNPs are very common in the human population. Between any two people there is an average of one SNP every 1250 bases. About 10 million SNPs exist in the

human population, where the rarer SNP allele has a frequency greater than 1%.

These SNPs are potentially very valuable as defined markers to track specific regions of a chromosome, and possibly as genetic tests. One of the key objectives of the Human Genome Project (both the publicly funded project as well as the private effort by Celera Genomics) has been to identify a large number of human SNPs throughout the genome. This requires comparing the DNA sequence of the same region from many different people. The Celera Genomics human genomic sequence comes from five different people, whereas the public Human Genome Project used a much larger number of DNA donors. However, even in the genomic sequence of a single person, SNPs can be identified between the two homologous chromosomes.

An unlikely consortium of pharmaceutical and computer companies have formed a group called the "SNP Consortium" to pool their resources and develop a large database of human SNPs in the public domain:

> The SNP Consortium Ltd. is a non-profit foundation organized for the purpose of providing public genomic data. Its mission is to develop up to **300,000** SNPs distributed evenly throughout the human genome and to make the information related to these SNPs available to the public without intellectual property restrictions. The project started in April 1999 and is anticipated to continue until the end of 2001.
>
> —Lincoln Stein (SNP Consortium Website, `http://snp.cshl.org`)

The January 2001 data release from the SNP Consortium consisted of 856,666 SNPs, all of which have been submitted to GenBank and anchored to the human genome by "*in silico*" mapping to the genomic working draft (UCSC "golden path"). Together with the International Human Genome Sequencing Consortium, the SNP Consortium published a paper in *Nature* (February 2001) describing 1.42 million SNPs that have been archived in a public SNP map base. These SNPs

are publicly available on the Web in the integrated genome maps at the NCBI (`http://www.ncbi.nlm.nih.gov/`), the Ensembl project of the European Molecular Biology Laboratory (`http://www.ensembl.org`), the "Golden Path" genome viewer at the University of California at Santa Cruz (`http://genome.ucsc.edu`), and the SNP Consortium Website at Cold Spring Harbor Laboratory (`http://snp.cshl.org/db/snp/map`). This huge collection of SNPs spans the entire genome fairly evenly, so that there is a 98% chance of a SNP located within 5 kilobases (kb) of every expressed gene.

Other Mutations

There are several other types of common mutation that are not confined to a single nucleotide. There are some highly variable sites in the genome that experience frequent mutations. These sites are known collectively as a **variable number of tandem repeats** (VNTRs). There are several different types of VNTRs, which are categorized by the size of the bit of DNA that repeats. **Microsatellites** have repeating units of 2–9 bp, while **minisatellites** have repeating units of 10–100 bp. These sites are characterized by extremely high heterozygosity in the population and instability of the sequence—mutations occur every few generations. These repeats seem to be subject to an inherent flaw in the DNA replication machinery that "slips" in repeat regions.

These VNTR loci are very useful as identity markers, such as in the forensic DNA testing that is discussed so frequently in the news and on TV crime shows. However, they have limited utility in genomic medicine since they are so extremely variable—they cannot be used as markers to reliably track other genes. These repeat regions are usually located in noncoding DNA; however, a few genes contain VNTRs, which tend to create genetic instability. **Fragile X** syndrome is an example of a 3-base repeat (GCC) in the FMR1 gene (located on the X chromosome), which has a tendency

to expand during meiosis from "normal" alleles with 20–50 copies up to 200 or more copies. Proteins manufactured from the mutant alleles show loss of RNA-binding activity, which leads to mental retardation. Other diseases that have been attributed to similar trinucleotide repeats include Huntington's disease, myotonic dystrophy, and spinal and bulbar muscular atrophy.

Linkage

A mutant allele does not always remain on the same intact chromosome. In addition to the segregation of chromosomes when cells divide, parts of chromosomes cross over between homologous **sister chromatids** in a process known as **recombination**. The result of this process is that the set of alleles on a single chromosome gets mixed up over many generations. However, the frequency of recombination between alleles of two genes is a function of the distance between those genes on the chromosome. Therefore, two genes that are located close together on a chromosome will have alleles that are **linked**, since they will rarely be separated (see Figure 5.2). Thus one allele can be used as a marker for another. This is particularly relevant to genetic medicine since it is becoming quite easy to find SNP markers linked to genetic disease (or risk factor) alleles without the (usually) lengthy process of identifying the precise mutation that is responsible for the disease.

More than one crossover event typically occurs between each pair of chromosomes in each meiosis. However, there seems to be a limit to how close together two recombinations can occur in one meiosis. Therefore, a pair of flanking markers on either side of a target gene can be used in order to track an important allele with even greater certainty than can a single linked marker (if markers on both sides are present, the chance of a double recombination that removes the central allele is extremely low).

A significant amount of basic medical genetics is required in order to find the gene or genes responsible for a heritable disease.

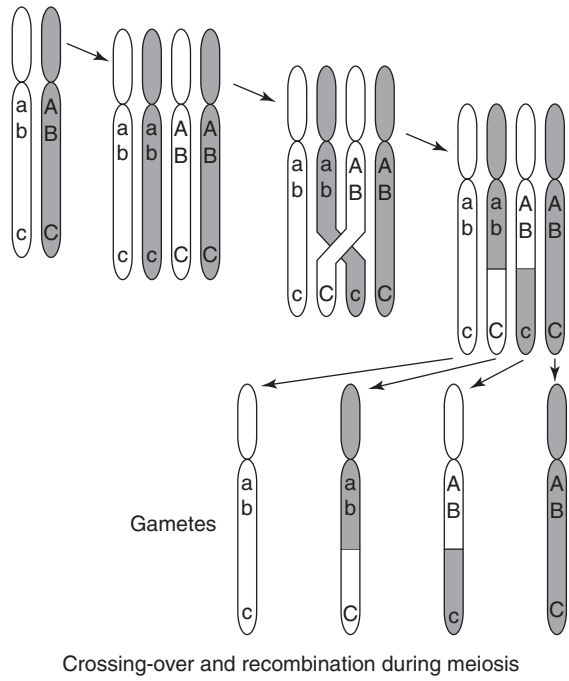


FIGURE 5.2. Cross over and recombination during meiosis. The alleles of genes A and B are linked, so they do not recombine, but the alleles of gene C are swapped, creating new combinations of alleles in two of the progeny gametes.

The starting point for associating a heritable disease with one or more genes is a **linkage analysis**. This is usually done by collecting DNA samples from members of families that have multiple instances of the disease—from both affected and unaffected individuals. Then these samples are screened with markers that span all regions of all chromosomes, looking for linkage. **Linkage** is defined very simply as markers that occur in affected individuals more frequently than in unaffected individuals. These markers can be anything that defines a particular allele of a gene—a phenotype, an enzymatic activity, a protein of a specific size and chemical identity, or a specific DNA sequence (SNP). This work generally requires large families with multiple generations of affected individuals in order to rule out chance associations and to find markers that are linked very tightly to the disease gene

(located nearby on the chromosome). It is often desirable to perform linkage analysis on several different families to gain greater statistical power and to confirm results, but it can be difficult to ensure that a phenotypically similar disease in different families is caused by the exact same genetic defect.

Multigene Diseases

Many of the most common diseases have been shown to have a significant heritable component, yet cannot be traced to alleles of a single gene. These complex diseases, such as asthma, heart disease, and cancer are the result of interactions between many genes (or perhaps just a few). Each of these genes can be considered to be a risk factor for the disease—each contributes, but no single one is either necessary or sufficient to cause disease on its own, or at least not in a significant fraction of patients. However, it is possible to use SNP markers to scan the genomes of families that show inheritance of the disease, or matched groups of people with and without the disease, in order to discover markers that are linked to these risk factors. In fact, these markers can be used to predict disease susceptibility without ever discovering the identity or function of the culprit genes.

Some complex (i.e., multigene) diseases may actually be caused by any of several different genes scattered across different chromosomes. So it is not that the disease really has multiple causes, but rather that there are multiple diseases that show similar symptoms. SNP linkage analysis can distinguish among these various genetic factors and lead to more precise diagnoses than can classical molecular genetic approaches.

Genetic Testing

Once SNPs associated with increased disease risk are documented, it will then be possible to screen ordinary patients for

these markers. The key limiting factor for the adoption of this technology into routine medical practice is the cost of performing a test for multiple SNPs in a single patient's DNA sample. There are a number of promising technologies currently under intensive development, particularly using DNA microarray approaches. It is also not yet clear how these tests will be integrated into the healthcare system. Is it sensible to make a single massive SNP array that will used to profile each newborn and provide a readout of all disease risks, or would it make more sense to create specific panels of SNPs that reveal information about susceptibilities to a particular disease, and then allow patients and physicians to use the tests when they have a concern about that disease? The presence of a SNP marker that is statistically linked to an increase risk for some disease is not enough information to make important medical decisions, but it does provide a justification for pursuing more thorough genetic testing as well as other types of diagnostic procedures.

The frequency of a disease-causing allele (and markers linked to it) may vary in different human subpopulations. These different gene frequencies may be the result of founder effects, genetic drift, or some form of selection (see Figure 5.3). A number of genetic diseases have well-defined frequencies in particular "at risk" populations such as sickle cell anemia among Africans and Tay–Sachs disease among European Jews. In order to use a

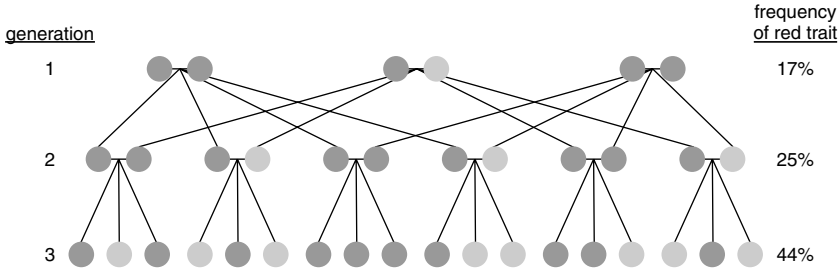


FIGURE 5.3. Gene frequencies change over generations.

particular SNP as a genetic marker, it is necessary to establish the frequencies of the various alleles in different populations. The National Human Genome Research Institute (NHGRI) has established a set of 450 DNA samples that are representative of the genetic diversity found in the US population, which can be used to measure the frequencies of SNP polymorphisms. It will be very useful in establishing standards for genetic tests to validate markers against a common standard.

Every allele of every gene has its own independent evolutionary history (and future!) and exists at different frequencies in each subpopulation. Against this background of constant variation and independently evolving alleles, it is important to keep in mind that there is no single "correct" sequence for any gene. The most common allele for each gene varies in different populations, and it is subject to change over time. Every person has a unique combination of alleles of all human genes plus variation throughout the entire noncoding sequence.

SNP Chips

The technology for producing microarrays has reached the point where millions of probes can be placed on a singe chip. Since there are not millions of human genes to be measured in gene expression assays, the companies that manufacture these arrays have turned their attention to sequencing and genotyping. Microarray technology is well suited to SNP detection. Each SNP allele can be encoded in a single 25-base oligonucleotide, and genotypes can be assigned with extremely high accuracy (99.8%; Rabbee 2006) by detection of hybridization differences in genomic DNA (heterozygotes and homozygotes can also be identified). Large numbers of SNP-specific oligonucleotides can be placed on a single chip and assayed in parallel on a single DNA sample, so the overall cost per genotype becomes very low. This technology allows a set of samples to be profiled for SNPs that cover the

entire genome at high resolution, a "whole-genome association study."

There are approximately 10 million SNPs across the worldwide human population, and over 5 million of these have sequences available in dbSNP at the NCBI Website (`www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi`). Affymetrix currently (in 2008) produces arrays of 10,000, 100,000, 500,000, and 1 million human SNPs. Using a different technology, Illumina offers a HumanHap300-Duo Genotyping BeadChip that contains a total of over 634,000 SNPs, including 317,000 tag SNPs from the HapMap Project.

The benefit (and the downside) of a high-density SNP array is that it produces a lot of data—so much data, that it cannot be managed using ordinary computer tools. Computing the association of one or more medical phenotypes with 1 million SNPs across hundreds or thousands of patients is a very challenging task. Even if computational tools are available that can handle this amount of SNP data, there is a risk that the number of chance associations (false positives) found in a set of a few hundred patients would drown out the true signals. Biomedical researchers are adopting this technology somewhat more slowly than might be expected, given its inherent power, but the initial results are promising. The Affymetrix 100,000 mapping set has been used to find a key gene for age-related macular degeneration and to identify 80 genes associated with multiple sclerosis.

THE HAPMAP PROJECT

Linkage is common and powerful in the human population, particularly in genetically isolated subpopulations, so that a group of alleles for neighboring genes on a segment of a chromosome are very often inherited together. Such a combination of linked alleles is known as a **haplotype**. When a new mutation occurs in a single

individual and is passed down to his or her descendants, it does not move on its own, but is carried on a specific chromosome. Recombination events near the mutant gene are rare, so specific alleles for neighboring genes on that chromosome will remain linked to the mutant gene. Every mutation can be traced back to a single founder chromosome, and it is more likely to be linked to the alleles of nearby genes that were present on that original chromosome than to alternative alleles. The more ancient the origin of the mutation, the less this original linkage can be detected. In theory, over an infinite amount of time, with random breeding of all humans, this linkage would break down. But over periods of hundreds to tens of thousands of years, a state of **linkage disequilibrium** is maintained. So even though every person has a unique combination of alleles of all genes, these alleles are not inherited on a completely random basis—they come in bunches, that is, haplotypes.

Linkage disequilibrium can be observed in any genetically isolated subpopulations, and to a lesser extent throughout the entire human population. Subpopulations that were founded by a small number of individuals, or that have passed through a bottleneck where their numbers were greatly reduced, have more linkage disequilibrium than do larger, more thoroughly mixed populations; thus a medically important allele can more easily be detected by its linkage to other markers within this subpopulation. Alternately, the predictive value of a particular DNA marker that is linked to disease allele in one population could be quite different in a different population.

Eric Lander and coworkers at the MIT Genome Center (Reich et al. 2001) have studied the linkage between SNP markers in human populations and have come up with some surprising results. In a population of pure northern European descent, most SNPs that were 60 kb apart showed significant linkage, and half of all SNPs 80 kb apart were linked. They also found great variation in the size of linked blocks of DNA across the genome. For example

linkage was detected for 155 kb around the Wiscott–Aldrich syndrome-like (WASL) gene, but for less than 6 kb around the Protein C inhibitor (PCI) gene. It seems that the genome is composed of a mosaic of recombination hotspots and coldspots, or that each locus on the genome has an intrinsic level of recombination that it supports.

Building on the success of the SNP Consortium, the International HapMap Project (`www.hapmap.org`) has been launched to map linkage among SNPs across the entire human genome (Thorisson et al. 2005). The goal of the HapMap Project is to develop a haplotype map of the human genome that will describe the common patterns of human DNA sequence variation. The HapMap Project officially started in October 2002 and published its preliminary results in 2005.

For each region of each human chromosome, the SNP markers are not randomly mixed. Most chromosome regions have only a few common haplotypes (each with a frequency greater than 5%), which account for most of the variation in the human population. A haplotype block may contain hundreds or thousands of SNPs that are coinheritied. So the task of genotyping a person can be greatly simplified once these haplotype blocks are defined. The HapMap defines haplotype blocks and specifies a few "tag SNPs" that can be used to identify which of the common haplotype variants is present in each person in specific chromosomal regions. Most of the common haplotypes occur in all human populations; however, their frequencies differ among populations. Therefore, genotype frequency data on a total of 270 people from several populations (Nigeria, Japan, China, Europe) have been evaluated by the HapMap Project in order to choose optimal tag SNPs.

All genotype and linkage data from the HapMap Project is publicly available. Using the Generic Genome Browser (Donlin 2007), it is quite easy to survey recombination frequency along each chromosome, identify SNPs (with known genotype frequencies) within and near genes of interest, and find tag SNPs for

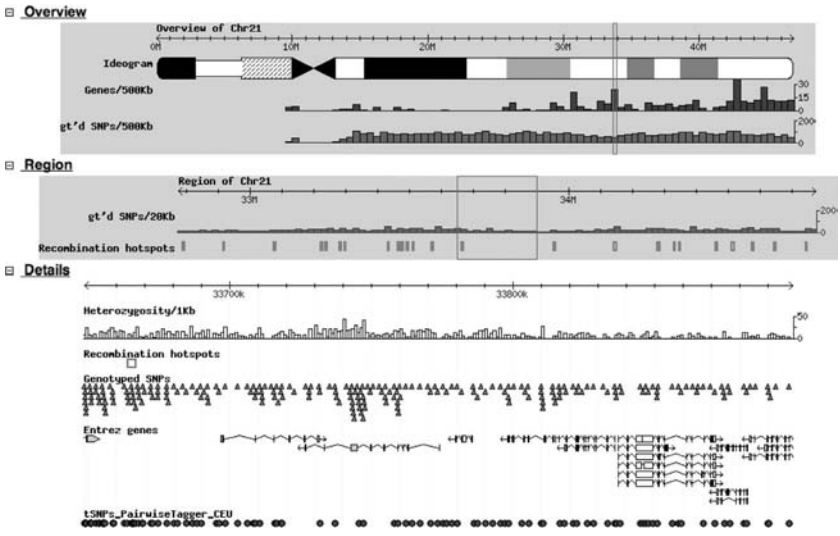


FIGURE 5.4. HapMap data shown in the Generic Genome Browser.

almost any gene (see Figure 5.4). Lincoln Stein's laboratory at Cold Spring Harbor has produced a tutorial on the use of the Generic Genome Browser at the HapMap.org Website (Thorisson et al. 2005).

HapMap data have become the basis for many studies of human genetic diversity and evolution, recombination, and associations of SNPs with various medical phenotypes. At this point (in 2007), the HapMap tag SNPs are not available on a single high-density microarray chip, but many of the SNPs assayed by the current high-density SNP chips have been genotyped across the 270 individuals used in the HapMap project.

Haplotypes could be used to greatly simplify the potential task of genetic testing. A few common haplotypes (chromosome chunks containing a specific set of alleles) may be associated with increased risk for various diseases in specific populations. These haplotypes can be detected with simple SNP markers. Then genetic disease risks can be evaluated in the context of hundreds of common haplotypes rather than considering all of the individual

interactions of tens of thousands of genes, each with their own unique distributions among various subpopulations. Technologies for haplotype testing are currently under development by many biotechnology and genomics companies.

There are many other clinical manifestations of human genetic variation. In fact, *all disease has a genetic component, and all therapies should consider genetic variations* (perhaps that can become the motto for the new era of genomic medicine). The physician should be aware of the genetic components of susceptibility versus resistance to various pathogens, variations in disease severity or symptoms, reactions to drugs (**pharmacogenomics**), and the variable disease course and prognosis that emerges as a synthesis of all of these factors.

Research Uses of SNP Markers

In addition to the direct medical uses of SNPs in genetic testing (see Chapter 6), these markers also have a valuable role in biomedical research. There are approximately 5000 human diseases known to be caused by the malfunction of a single gene, but only a few hundred have so far been fully characterized as to the normal and mutant gene sequences [see Online Mendelian Inheritance in Man (OMIM) database at `http://www.ncbi.nlm.nih.gov/omim`]. The remaining diseases are extremely rare, making it difficult to find sufficient numbers of families with multiple instances of the disease in order to conduct classic linkage analysis. However, with large numbers of SNP markers available, it becomes possible to conduct much more precise genome-wide screens for common mutations among small populations of affected people, even if they are not related. Common SNP markers, together with a complete human genome sequence, allows researchers to take a positional candidate approach: chromosome regions that seem to be common among people with a particular disease can be investigated gene by gene to look for a

common mutation in a gene with a function that could be related to the disease phenotype.

Pharmaceutical companies are very interested in common diseases with multiple "risk factor" genes such as heart disease or cancer. Once again, populations of people who have a disease can be compared to each other and to a group of healthy people to identify SNP markers that are correlated with disease. Since the locations of the genes that contain risk factor mutations for disease or drug reaction are unknown, it is necessary to scan the entire genome (i.e., with high-density SNP chips) for SNPs that might be associated with the medical phenotype. The chromosome regions near these markers can be examined for genes that are likely candidates for involvement in disease (on the basis of the known function of the protein product, or the similarity to proteins of known function). A candidate gene is then sequenced in a small population of diseased and healthy people to see if the diseased people share a common mutation.

Each gene that is identified in this way as contributing to a disease becomes a target for drug development research. Drugs can be designed to enhance or block transcription of that gene, to interact with the mRNA or protein product of that gene to enhance or block its function, or to interact with the other molecules in the cell with which that protein normally binds. This type of "rational" drug design promises a great improvement in the efficiency of the drug development process, and thus improved profits for drug companies and more effective drugs for patients.

Ethnicity and Genome Diversity

The Human Genome Project and other related efforts to sequence human DNA have produced some significant findings about human genetic variation, which have direct bearing on common notions of ethnic distinctions among groups of people. The most obvious conclusion, as noted above, is that the human species as a whole has a much lower level of genetic variation than do many

other species. This can be directly attributed to the extremely recent and very rapid increase in human population.

In the human population, genetic heterogeneity is broad, but shallow. There are many sites for genetic variation, but for each locus, a small number of common polymorphisms explain the bulk of the heterozygosity. This distribution indicates that most of the SNPs in the human population are the result of ancient mutational events that created new alleles that have become broadly distributed throughout the population. Since the human population has grown very quickly, most of the mutations that it carries date back to a time when there were far fewer humans. In other words, two individuals who share a variant allele have a single common ancestor who was the source of that mutation, even if those two individuals are members of different modern subpopulations. In general, the higher the frequency of a SNP allele in the entire human population, the older the mutation that produced it. Most of the SNPs present at high frequencies in the population were present before humans expanded out of Africa; therefore it is unlikely that a particular SNP marker will uniquely distinguish any particular ethnic group. Eric Lander, Director, MIT Center for Genome Research, stated in 2001 that

> Modern Europeans, and possibly populations in other regions of the world, are descended from a group of just a few hundred Africans who left their homeland as recently as 25,000 years ago. In fact, this small bottleneck may represent all the different groups of people leaving Africa.

Pääbo (2001) stated that

> The gene pool in Africa contains more variation than elsewhere, and that the genetic variation found outside of Africa represents only a subset of that found within the African continent. From a genetic perspective, all humans are therefore Africans, either residing in Africa or in recent exile.

The total genetic diversity among the members of any ethnic group is much greater than the diversity between different groups. In fact, 85% of the total human genetic variation is present

within any ethnic group. In other words, "It is often the case that two persons from the same part of the world who look superficially alike are less related to each other [in terms of total DNA similarity] than they are to persons from other parts of the world who may look very different". (Pääbo, 2001).

Despite all of the data showing that human races do not exist in any meaningful genetic context, we can see characteristic differences between various ethnic groups, in obvious traits such as skin hair and eye color or body shape, as well as subtle traits such as the prevalence of hereditary diseases such as Tay–Sachs disease and sickle cell anemia. These group traits can be traced to several factors that can influence the genetic makeup of isolated populations over relatively short periods of time. Many ethnic groups were founded by small populations of migrants, or they may have passed through bottlenecks where the population was greatly reduced by plague, war, or some other adverse events. These small numbers of ancestors possessed a greatly reduced sample of the total genetic variation of the human species, including some relatively rare alleles, and some common ones at higher and lower frequencies than found in other populations. These **founder effects** lead to some unique genetic characteristics in the population of their descendants.

Certain environmental conditions can produce very strong selective pressures which, in small populations, can effect substantial changes in just a few generations. For example, strong sun exposure creates a great advantage for genes that darken the skin to prevent burning, while weak sun exposure favors light skin to favor absorbing sunlight for vitamin D synthesis. It is also possible, in small populations, for sexual selection (the favoring by one sex for mates having a certain trait) to have a significant effect over a relatively short period of time. It is possible to use SNP data to identify chromosomal regions with unusually low levels of diversity in a subpopulation. These are likely to be regions that contain alleles of genes that are currently

undergoing selection for favorable alleles—a favorable allele that is increasing in frequency in the population will "carry along" linked markers leading to a decrease in heterozygosity in that region.

While genome sequence data do not support notions of genetically distinct human "races," the data do contain historical information of tremendous value. The pattern of distribution of specific polymorphisms can indicate the movements of people over thousands of years. Ancient alleles are widely distributed; more recent mutations have smaller circles of distribution (groups of people descended from a common ancestor). The patterns of sequence variation seen today among various groups of people can be used together with historical, linguistic, and archaeological data to reconstruct their social and genetic histories (Ostrer 2001).

Social Implications of Genetic Diversity Data

The human genome is more than a set of instructions for the construction and maintenance of a human being; it is also a historical document. In her or his genome, each person contains a complete genealogical record of their ancestors, going all the way back to the origins of life. All of the evolutionary selections and the random factors that allowed for the reproductive success of their ancestors are recorded in the DNA sequence of each person's genome, but this information can only be fully interpreted in the context of comparisons with the genomes of many other human beings and other species.

This concept was well expressed by the Australian Aboriginal poet Oogeroo Noonuccal (Walker 1992).

> Let no-one say the past is dead.
> The past is all about us and within
> Haunted by tribal memories, I know

> This little now, this accidental present
> Is not the all of me, whose long making
> Is so much of the past.

The Human Genome Project is exploring this racial history by sequencing DNA from many different people and cataloging the frequencies of various polymorphisms in different populations. This data can be read forward to predict many genetic factors related to disease, reactions to drugs, and other variables. It can also be read backward to deduce the history of human migrations, the isolation and/or intermarriage of various groups, and a great deal more ethnological and anthropological data.

It is also very important to keep in mind the limits of genetic factors in determining human individuality. While genes do contribute to the expression of complex characteristics such as intelligence and personality, the complexity of human development must not be oversimplified under the misguided notion of genetic determinism. Just as identical twins have unique personalities, the role of environmental influences and personal experiences is extremely important in all aspects of the development of a person. No one is simply a matrix of interacting genes, and genetic explanations of human behavior are likely to lead to serious errors. In particular, genetic aspects of human behavior and psychiatric illness have frequently been overstated far beyond what the data actually support. It is quite important that medical genetics avoid the loss of credibility that comes with overpromising, to ensure that the genetic tests that truly do offer useful diagnostic and therapeutic value will not be equally tainted.

REFERENCES

Donlin MJ. 2007. Using the Generic Genome Browser (GBrowse). *Curr Protoc Bioinformatics* **9**:9.9.

The International SNP Map Working Group. 2001. A map of the human genome sequence variation containing 1. 42 million single nucleotide polymorphisms. *Nature* **409**:928–933.

Ostrer H. 2001. A genetic profile of contemporary Jewish populations. *Nat Rev Genet* **2**:891–898.

Pääbo S. 2001. Genomics and society. The human genome and our view of ourselves. *Science* **291**:1219–1220.

Rabbee N, Speed TP. 2006. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* **22:**7–12.

Reich DE, Cargill M, Lander ES et al. 2001. Linkage disequilibrium in the human genome. *Nature* **411**:199–204.

Stoneking, M. 2001. Single nucleotide polymorphisms: From the evolutionary past. *Nature* **409**:821–822.

Thorisson GA, Smith AV, Krishnan L, Stein LD. 2005. The International HapMap Project Web site. *Genome Res.* **15**:1592–1593.

Walker K (Oogeroo Noonuccal). 1992. *"The Past" in the Dawn is at Hand: Selected Poems*. Marion Boyars, London.

CHAPTER

6

Genetic Testing for the Practitioner

Harry Ostrer

Genetic testing represents a fairly recent addition to the methods that are available to the physician for diagnosing disease or identifying those at risk for developing a disease or having a child affected with a disease. Pharmacogeneic testing has found applications for predicting response to therapy and potential for developing toxicity to specific therapies. Genetic testing has gained wide acceptance in clinical practice. Many genetic testing laboratories now exist worldwide.

Genetic testing is frequently viewed as different from other kinds of laboratory tests. The major reason is that for some diseases genetic testing can identity those who are currently well but may become ill in the future. Because of the transmissible nature of genetic information, the identification of a disease-associated mutation has implications not only for individuals at risk as well as for their family members. For these reasons, understanding the implications and limitations of genetic testing are important not only for the practitioner but also for the patient. Genetic testing

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

also can be used to identify individuals in forensic applications as perpetrators of assault crimes, in genealogic applications as members of specific lineages, and in population genetic applications as members of specific ethnic groups. Because genetic tests can be multiplexed (i.e., many tests can be performed simultaneously), these applications can be combined. This chapter provides an overview of genetic testing for the practitioner, including the clinical situations in which it is used, the conceptual basis for the various methods of genetic testing, and the significance of genetic test results.

Clinical Applications of Genetic Testing

Frequently the question may arise as to whether the patient has a certain disease for which there is a genetic basis. Often among the 10,000 conditions for which a genetic basis has been identified, the diagnosis can be made from evaluation of personal and family history, physical examination, and conventional laboratory tests. A useful database for identifying these conditions is **Online Mendelian Inheritance in Man** (OMIM) (`www.ncbi.nlm.nih.gov/omim`). This catalog is updated regularly and can be searched using multiple terms. The entries provide information about the clinical signs as well as the genetic basis for the condition, if known, including mutations that have been found to cause the condition. To determine whether genetic testing is available for a given condition and to find a laboratory, a useful link is GeneTests, a free online service (`www.genetests.org`). The entries in this catalog indicate the test menus and contact information for the laboratories, as well as whether the testing is provided on a routine or research basis. A very useful adjunct in the GeneTests Website is GeneReviews, which provides succinct summaries about many genetic conditions

and the ways the genetic testing can be used for diagnosing these conditions, including prediction of natural history.

The clinician is likely to encounter many situations in which a genetic test may be useful. Sometimes genetic testing is required from diagnosis when it cannot be made by clinical criteria alone. The fragile X syndrome is the most common genetic form of mental retardation. Although the diagnosis may be suggested by the presence of the characteristic signs—large ears, protruding chin, and large testes—the only way to diagnose fragile X is by genetic testing. For the various forms of spinocerebellar ataxia, there is considerable overlap. Yet, these can be readily distinguished by their specific mutations. Patients with atypical forms of certain diseases may have a negative gold standard test, but positive genetic test. For most patients with cystic fibrosis, the diagnosis can usually be made by a sweat chloride test. However, a number of individuals have been described with pulmonary disease suggestive of this condition for whom the sweat chloride test is normal. For these patients, the diagnosis has been based on observation of mutations in both copies of their CFTR genes.

For some conditions, the signs of disease may not yet have developed, yet on the basis of ons's family history, one may want to know about the risk of developing disease. This is true for the person whose parent(s) may have died from Huntington's disease, a progressive neurodegenerative disease or for the person whose mother and sister may have died from breast or ovarian cancer, suggesting a heritable risk. For these individuals, a positive genetic test result will indicate an increased, although not necessarily absolute, risk for developing the disease.

Genetic testing is used for assessing reproductive risks—by testing the parents for carrier status and by testing the fetus. Individuals with a positive family history of genetic disease (usually autosomal recessive or X-linked) or who come from ethnic groups with an increased prevalence of autosomal recessive or X-linked diseases are candidates for carrier screening. Currently,

carrier screening for cystic fibrosis, fragile X syndrome, and spinal muscular atrophy is recommended in the United States. For people of Mediterranean, African, or South Asian ancestry, hemoglobinopathy screening is recommended. For individuals of Ashkenazi Jewish ancestry, screening for Tay–Sachs disease, Canavan disease, cystic fibrosis, Gaucher disease, Bloom syndrome, Fanconi anemia, Niemann–Pick disease, familial dysautonomia, maple syrup urine disease, glycogen storage disease, and familial hyperinsulinism is available. An individual who is a carrier for a certain condition may choose not to marry another individual who is a carrier for the same condition. Alternatively, if a carrier couple is identified, they may choose to have prenatal diagnosis to determine whether their fetus is affected with this condition. This can be performed either at 10–11 weeks using the procedure of chorionic villus sampling where a bit of placenta is obtained under ultrasound guidance. As another option, an amniocentesis can be performed at 15–18 weeks of pregnancy to obtain cells from the amniotic fluid. These couples might also choose to have preimplantation genetic diagnosis with selection implantation of only those embryos that are deemed unaffected.

Not all genetic testing involves looking for heritable mutations. Sometimes it is used to look for genetic alterations that are confined to a specific population of cells. These alterations may cause certain cells to become cancerous, or if cancerous, to progress to a more aggressive stage. Genetic testing can be used to identify chromosomal translocations between two nonhomologous chromosomal segments and in the process diagnose a specific form of leukemia. For example, the translocation between chromosomes 1 and 19 in leukemic cells is diagnostic of the acute promyelocytic form of this disease and the translocation between chromosomes 9 and 22 is diagnostic of the chronic myelogenous form. The expression patterns of RNA transcribed from many genes can be assessed to predict the natural history of

the disease. This approach has been used to predict breast cancer outcome and whether more or less aggressive therapies should be used to treat patients.

Individuals might also have genetic tests of identity. These might be voluntary and selected to test specific questions, such as whether they are members of a known patrilineal lineage, such a people with a specific surname. These tests analyze a series of polymorphic genetic markers on the Y chromosome. On the basis of the general pattern of markers, or "haplogroup," they may be told of the geographic region where their Y chromosome originated. According to the number of markers that match with people who are suspected to be of the same lineage, individuals may be advised about the common ancestors or other people in that lineage. Such testing is also possible for matrilineal lineages by testing mitochondrial DNA markers. This testing is provided by commercial firms that market directly to consumers. Identity genetic analysis may also be involuntary and used for paternity testing of children or fetuses or for identification of forensic samples in murder, assault or rape cases, in which the perpetrator of the crime left a tissue sample of blood, semen, hair, or other tissue type from which DNA can be extracted.

Methods of Genetic Testing

Direct Mutation Testing

The method of genetic testing has to be geared to the condition that is being detected. For some diseases, a single, or a limited number of mutations can cause the condition. For some population groups, a founder mutation may have occurred and this mutation may have achieved a carrier frequency of 1% or greater. For these conditions, genetic testing is relatively efficient because it is geared to the detection of these specific mutations. For other conditions, a wide variety of different mutations can produce

the condition. In these cases, a method that can detect all of the possible sequence variants is needed.

Most genetic testing involves the use of the polymerase chain reaction (PCR), whereby large numbers of copies of a gene sequence are made using DNA primer molecules that define the ends of the sequence to be amplified. Fragments are amplified only for the primer sequences that define them, and the presence of an amplified fragment can be diagnostic for the presence of a translocation, deletion, or insertion. The size of the fragment, itself, can be diagnostic for whether a deletion of insertion has occurred. Single-nucleotide polymorphisms (SNPs), involving base pair changes, can be detected by a variety of techniques (Figure 6.1). These may involve allele-specific PCR, primer extension, ligation, or hybridization occurring directly at the site of the base change. The specific reaction occurs only when there is a perfect nucleotide match. These techniques can be used to detect both the common and uncommon (or wild-type and mutant) alleles at this site.

Many of these methods now lend themselves to multiplexing, whereby alleles at several different sites are tested at the same time. The multiplexing format takes advantages of the ability to perform PCR on several different fragments at the same time and the ability to test for mutations at one or more sites in each of these fragments. These methods are useful when the mutation sites to be queried are known.

When they are unknown, other methods are available to test for sequence variation in the genome. The gold standard, and most commonly used, method is DNA sequencing. Sequencing can be performed directly on PCR-amplified DNA fragments. This analysis can be quite efficient as up to 800 bp can be analyzed in a single PCR reaction. Other methods take advantage of changes that occur in the physical properties of DNA when a mismatch occurs at a site (as the result of heteroduplex formation by hybridization). The commonly used methods include

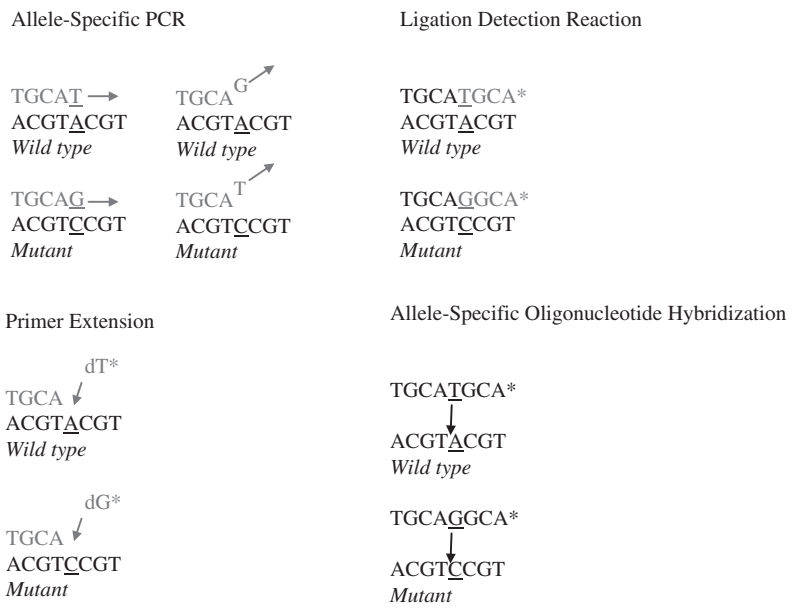


FIGURE 6.1. Methods of detection of single-nucleotide polymorphisms: allele-specific PCR, ligation detection reaction, primer extension, and oligonucleotide-specific hybridization.

single-stranded conformation polymorphisms and denaturing high-pressure liquid chromatography (dHPLC). Generally these are deemed screening methods, and any newly detected polymorphisms are confirmed by DNA sequencing.

One major issue of sequencing-based strategies is the interpretation of novel DNA sequences. If these novel sequences can be shown to disrupt the coding region of a gene by introducing premature terminators, then they are usually deemed to be mutations. If they alter amino acid sequences, then it may be unclear whether they are bona fide mutations that produce the observed phenotype. Their roles may be clarified by case–control studies in which the specified mutation is shown to be present in a high proportion of cases, but a low proportion of controls or by linkage studies in which the specific mutation is shown to

be linked to transmission of the observed phenotype in a family. Other resources that are used to clarify these variants of unknown significance are the Human HapMap, an offshoot of the Human Genome Project that defines the major polymorphisms that are present in different human populations, comparative phylogeny that tests whether the specific residue is invariant in the same gene in other organisms and thus subject to natural selection, and physicochemical prediction programs that attempt to determine whether the properties of the variant protein differ from the wild type.

Linkage Analysis

In the past, linkage analysis was commonly used as an alternative to direct mutation methods. Linkage was used when the map position, but not a disease-causing gene, was known. Generally members of three generations in the family participated in order to ensure the correct assignment of phase for the markers (Figure 6.2). Markers were chosen that flanked the region of

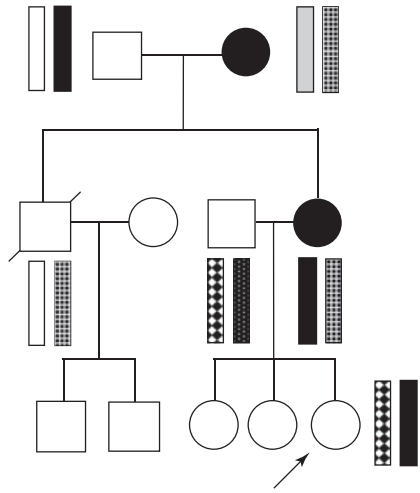


FIGURE 6.2. Linkage analysis for detection of individuals carrying disease-conferring mutation.

interest to account for recombination that could limit the accuracy of prediction. Linkage analysis has fallen by the wayside in clinical applications as the genetic basis for many diseases have been identified. Linkage remains a powerful research tool for identifying new genes for conditions that are transmitted through families. The availability of high-density SNP arrays that can test for linkage to virtually any region of the genome simultaneously has made linkage a readily accessible tool for testing families with many affected individuals.

Array-Based Comparative Genomic Hybridization

Array-based comparative genomic hybridization (arrayCGH) is another application of array technology. These silicon chips or glass slides are fabricated to contain oligonucleotides or bacterial artificial chromosome (BAC) probes, either that span the entire genome or that are targeted to specific regions that have been associated with chromosomal duplication or deletion syndromes. These arrays can be used to detect chromosomal imbalances that are too small to be observed by conventional chromosomal staining techniques. The improved resolution of arrayCGH means that it can be used to detect virtually all cytogenetic aberrations, except for balanced translocations, which, by definition, do not cause any loss of gain of chromosomal material, and low-level chromosomal mosaicism that will not change the intensity of the signal above the threshold for a haploid or diploid chromosome number.

Expression Profiling

Expression profiling examines the expression of a large number of genes simultaneously using microarrays of oligonucleotide or

cDNA probes. The expression can be compared between tissues and between normal and tumor tissue, including tumors known to have favorable and unfavorable outcomes. Multivariate analysis is applied to characterize the genes that are over- and underexpressed. This approach has been used to predict outcomes for women newly diagnosed with breast cancer. For those predicted to have less favorable outcomes, more aggressive courses of chemotherapy can be implemented.

Adequacy of Genetic Testing

Genetic tests should be designed that are both sensitive and specific, that is, a high proportion of cases should be detected by a direct or indirect mutation test in a gene and normals should not be erroneously diagnosed by the same test. Despite the best of intentions, this is not a foregone conclusion. The various methods can preferentially determine sequence alterations of a certain type and may miss others. Some conditions are caused by a variety of different mutations in a single gene or in different genes and thus may be missed. Genetic testing needs to be carried out with appropriate controls so that positive and negative results can be analyzed.

Because of the sophistication required for genetic testing, regulatory measures have been imposed both by state departments of health and the Health Care Finance Administration, specifically under the regulatory authority that was provided by the Clinical Laboratory Improvement Act of 1988 (CLIA '88). This regulation specifies the qualifications for laboratory directors, supervisors, and technicians and the requirements for standard operation procedures for testing, reporting, quality assurance, and quality control. Many jurisdictions now require that testing be performed in CLIA-approved laboratories if the results are to be reported to patients.

The Secretary's Advisory Committee on Genetic Testing (2001), appointed by the Secretary of Health and Human Services, has recommended national standards for genetic testing aimed not only at improving test production but also at test validation. This could lead to the large-scale marketing of test kits, much as occurs for other forms of contemporary laboratory testing. The Food and Drug Administration has a series of recommendations for regulation of genetic testing kits and analytical methods.

Informed Consent

Genetic testing is viewed as exceptional compared to other forms of laboratory testing. This exceptionalism occurs because genetic testing does not vary over time and provides information about an individual's innate heritable predispositions that are not subject to the individual's control. In addition, this information has implications for the individuals tested as well as their family members. For these reasons, many states have passed laws that require individuals to provide informed consent before a genetic test is performed. These laws require that testing be performed only for the conditions for which consent was provided. Some laws allow for residual DNA samples to be available for research purposes, if anonymized.

Genetic Counseling

The significance of genetic testing is frequently unknown to patients. It is critical prior to testing to explain to the patient the reasons for performing a test; the natural history of the condition being tested for, including possibilities for intervention; and the significance of positive and negative results. Following testing, it is important to provide additional counseling about the meaning of a positive or negative result. A positive result does not

necessarily ensure that an individual will develop a condition. Likewise, a negative result does not necessarily eliminate the risk. Genetic counseling should be recommended for those people newly identified as mutation carriers to their family members, for whom similar benefits can then be extended.

Genetic testing for the diagnosis of inherited diseases provides a powerful tool for the practitioner. The technology is advancing rapidly. Diseases are being newly diagnosed, and new methods are being continuously developed. For the practitioner, the best approach is to understand the principles on which the tests are based and apply them on a disease-by-disease basis. Decisions about genetic testing can be facilitated by consultation with an experienced medical geneticist or genetic counselor. Careful application of DNA analysis in the proper setting can improve patient care dramatically.

Clinical Vignettes

Case 1. A 20-year-old man presents with recurrent pneumonias. His chest X ray demonstrates bronchiectasis. Suspecting cystic fibrosis, his physician orders a sweat chloride test that is normal. Subsequently, a genetic test for cystic fibrosis reveals a positive result R117H/delF508, 5T/7T. This test confirms the diagnosis of cystic fibrosis, but of a mild or atypical variety, which explains the normal sweat chloride test. The presence of the 5T allele with the R117H mutation affects the splicing of the CFTR gene. Coupled with the delF508 mutation, the patients's epithelial cells have decreased production of the CFTR chloride transporter. The patient is a candidate for aggressive treatment with antibiotic and mucolytic agents to prevent progression of his lung disease. He is also very likely to have bilaterbal congential absence of the vas deferens, which will cause him to be infertile. Nonetheless, he can reproduce by intracytoplasmic sperm injection (ICSI) of oocytes (which

requires ovarian hyperstimulation, oocyte retrieval, and reimplantation for his partner). His partner should have cystic fibrosis carrier testing prior to undertaking these procedures. If she is found to be a carrier, the couple will be at 50% risk for having affected children and 50% risk for having carrier children.

Case 2. An Italian-American family lost a child with type 1 (severe) spinal muscular atrophy during the first year of life. Genetic testing reveals that both have deletions of the SMN1 gene. They are at 25% risk for having an affected child with type 1 SMA in a subsequent pregnancy. Genetic testing of the fetus can be offered in those pregnancies by either chorionic villus sampling at 10–11 weeks or amniocentesis at 15–18 weeks. Prior to having prenatal diagnosis, both members should be tested for cystic fibrosis, hemoglobinopathy, and β-thalassemia to determine whether they are carriers for these conditions. The woman should also have genetic testing for fragile X syndrome to ascertain whether she is a carrier. If they are found to be at increased risk for having a fetus with these disorders, specific prenatal diagnosis can be offered.

Case 3. A 25-year-old woman was diagnosed with acute promyelocytic leukemia. By PCR analysis, her leukemic cells were found to have a 15, 17 chromosomal translocation, involving the promyelocytic (PML) and retinoic acid α-receptor (RARA) genes. On the basis of this information, she was treated with all-*trans*-retinoic acid and went into remission.

Case 4. A 37-year-old woman was diagnosed with breast cancer. Prior to undergoing lumpectomy and radiation therapy, she had genetic testing. This revealed a novel variant in her BRCA1 gene R1699Q, which was thought likely to be deleterious. Her mother and aunt also had breast cancer. Genetic testing demonstrated that they also had this variant,

strengthening the inference that the mutation was pathogenic. The patient opted to have bilateral mastectomy and oophorectomy to lessen her risk of breast cancer. Other female members of her family chose to have genetic testing to assess their risks.

Case 5. A 7-year-old boy was diagnosed with acute lymphocytic leukemia. Prior to starting treatment with 6-mercaptopurine (6MP), he was tested for thiopurine methyl transferase (TPMT) deficiency and was found to be homozygous for the A154T allele. TMPT deficiency was diagnosed, and the patient was started on a lower dose of 6MP.

References

Andrews LB, Fullarton JE, Holtzman NA, Motulsky AG (eds). 1994. *Assessing Genetic Risks: Implications for Health and Social Policy*. National Academy Press, Washington, DC.

Buetow KH, Edmonson M, MacDonald R, Clifford R, Yip P, Kelley J, Little DP, Strausberg R, Koester H, Cantor CR, Braun A. 2001. High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc Natl Acad Sci USA* **98**:581–584.

Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ, Perou CM. 2006. Concordance among gene-expression-based predictors for breast cancer. *New Engl J Med* **355**:560–569.

Goldgar DE, Easton DF, Deffenbaugh AM, Monteiro AN, Tavtigian SV, Couch FJ; Breast Cancer Information Core (BIC) Steering Committee. 2004. Integrated evaluation of DNA sequence variants of unknown clinical significance: Application to BRCA1 and BRCA2. *Am J Hum Genet* **75**:535–544.

Grody WW, Cutting GR, Klinger KW, Richards CS, Watson MS, Desnick RJ. 2001. Laboratory standards and guidelines for population-based cystic fibrosis carrier screening. *Genet Med* **3**:149–154.

Hacia JG, Brody LC, Chee MS, Fodor SP, Collins FS. 1996. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat Genet* **14**:441–447.

Kronn D, Jansen V, Ostrer H. 1998. Carrier screening for cystic fibrosis, Gaucher disease, and Tay-Sachs disease in the Ashkenazi Jewish population: The first 1000 cases at New York University Medical Center. *Arch Intern Med* **158**:777–781.

Motulsky AG. 1999. If I had a gene test, what would I have and who would I tell? *Lancet* **354**(Suppl 1):SI35–SI37.

Motulsky AG. 1997. Screening for genetic diseases. *New Engl J Med* **336**:1314–1316.

Ostrer H, Hejtmancik JF. 1988. Prenatal diagnosis and carrier detection of genetic diseases by analysis of deoxyribonucleic acid. *J Pediatr* **112**:679–687.

Ostrer H. 2001. A genetic profile of contemporary Jewish populations. *Nat Rev Genet* **2**:891–898.

Pandolfi PP. 2001. Oncogenes and tumor suppressors in the molecular pathogenesis of acutepromyelocytic leukemia. *Hum Mol Genet* **10**:769–775.

Saiki R, Scharf S, Faloona F, Mullis K, Horn G, Erlich H, and Arnheim N. 1985. Enzymatic amplification of B-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**:1350–1354.

Salman M, Jhanwar SC, Ostrer H. 2004. Will the new cytogenetics replace the old cytogenetics? *Clin Genet* **66**:265–275.

Secretary's Advisory Committee on Genetic Testing. 2001. *Enhancing the Oversight of Genetic Tests: Recommendations of the SACGT*. National Institutes of Health, Bethesda, MD, July 2001 (available online at `http://www4.od.nih.gov/oba/sacgt.htm`).

Wiggins S, Whyte P, Huggins M, Adam S, Theilmann J, Bloch M, Sheps SB, Schechter MT, Hayden MR. 1992. The psychological consequences of predictive testing for Huntington's disease. Canadian Collaborative Study of Predictive Testing. *New Engl J Med* **327**:1401–1405.

CHAPTER

7

Gene Therapy

John G. Hay

Historical Perspective

Studies performed in 1928 by Fred Griffith, an English microbiologist, demonstrated that when heat-killed bacteria are mixed with live bacteria, the characteristics of the living bacteria could change (Griffith 1928). Deoxyribonucleic acid (DNA) was subsequently shown to be this "genetic transforming factor" by the studies of Avery, Macleod and MacCarty that were reported in 1944 (Avery et al. 1944). The theory that transfer of DNA can alter the phenotype of an organism was thus established more than 60 years ago.

Subsequent studies established the gene as the unit of inheritance, and over several decades the nature of many gene mutations that lead to specific inherited diseases have been established. The rapid developments in recombinant DNA technology in the 1970s and 1980s set the stage for gene replacement becoming a realistic possibility to cure inherited disease by inserting a functional copy of the mutated gene into a patient's cells. Equipped with the knowledge of mutations within genes that

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

can lead to disease, researchers were able to rapidly translate the discovery of gene mutations into genetic therapy. For instance, the mutant gene that is responsible for the clinical manifestations of the disease cystic fibrosis was identified in 1989, and initial clinical trials using viruses to replace the defective gene were in progress by 1992.

Since 1992, the early enthusiasm for gene replacement has been replaced with a more temperate appraisal and a greater understanding of the problems that remain to be overcome to achieve effective therapy. In recent years (as of 2007), several studies have reaffirmed the hope that gene replacement may become a viable therapy for human disease in the not-too-distant future.

Strategies of Gene Therapy

The most intuitive application for gene therapy is to correct an inherited defect within a single gene. In this circumstance the cause of the disease is clear—a mutation within a single gene—the potential therapy equally apparent—replace the faulty gene with a normal copy. For this strategy to be curative, germ cells or stem cells would require correction with a permanent copy of the normal gene sequence to continually supply progeny cells with the corrected genotype. In view of the ethical and safety concerns of germline transmission of genetic alterations, studies performed so far have focused on somatic cell gene transfer. The consequence of this approach is that gene transfer is not permanent and therapy needs to be repeated.

Diseases that result from single gene mutations that were initially investigated for gene transfer to enable the expression of a normal copy of the mutated gene included cystic fibrosis, familial hypercholesterolemia, and mutations in the adenosine deaminase gene in the lymphocytes of individuals with immune deficiency as a consequence of adenosine deaminase deficiency.

Diseases that result from alterations in expression of many genes that might not at first sight be considered rational targets for gene therapy have also been intensively studied. These conditions include cancer and cardiac and limb ischemia. The strategy in these applications is to deliver a therapeutic gene, rather than correct an inherited abnormal gene.

Another major area of interest is to use genes to render tumors more immunogenic or to enhance the immune response against a tumor. Also genes are being used to induce an immune response against potential pathogens, in essence using genes as vaccines. Finally, also considered within the overall gene therapy umbrella is the use of a genetically modified virus to replicate specifically with tumor cells.

DNA Elements for Gene Expression

Assuming a gene can be delivered to the nuclei of appropriate cells within the target tissues, certain essential DNA elements are required for gene transcription (Figure 7.1). The first essential element is the coding sequence of the gene. Depending on the size of the gene and the capacity of the vector system, the genomic sequence (including the noncoding introns) or the cDNA (excluding introns) is used. In addition to the coding sequence,

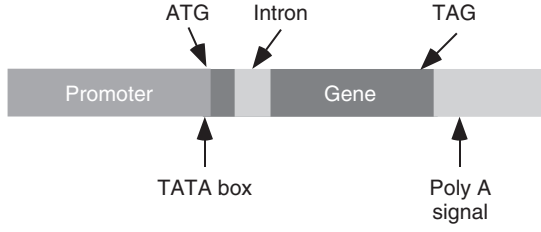


FIGURE 7.1. DNA elements required for gene transcription. The TATA box for RNA polymerase binding, the polyA site responsible for signaling the addition of a polyA tail to the transcript, the ATG codon for initiation of translation, and the TAG translation stop codon are shown.

the flanking sequences that are important for initiation of transcription and translation and RNA processing are essential. The final element to be included is the promoter.

The **promoter** is a stretch of DNA sequence that directs the transcription of the gene. The promoter can limit transcription to a particular tissue, to a particular period of development, to a particular time of the cell cycle, or in response to certain external triggers such as radiation. The promoter is able to perform this function by providing a docking site or TATA box for the RNA polymerase associated protein complex, and then provide sequences that serve as target sites for a variety of proteins that may enhance transcription (transactivators) or repress transcription (repressors) to the transcription unit.

Although it is often highly desirable to have a tightly regulated promoter for the gene to be transferred, limitations of the packaging capacity of vector systems and the effects of surrounding DNA enhancer elements in viral vectors on transcription can provide problems in providing specificity of gene expression. Thus, a powerful constitutive promoter, or "always on" promoter, is often used.

Gene Delivery Systems

As already alluded to, for a therapeutic gene to be expressed, it needs to be located within the nucleus of the target cell. Localization within the nucleus is the consequence of several steps. First, the gene has to be delivered to the cell surface, bind to the cell membrane, traverse the membrane into the cell interior, negotiate the intracellular trafficking pathways to reach the nuclear membrane, translocate the nuclear membrane, and finally the DNA has to be unpacked and released from any carrier within the nucleus ready for transcription (Figure 7.2). Finally, the inserted gene must persist in the nucleus, either by integration into the cell's chromosomes, or as a self-contained genetic element such

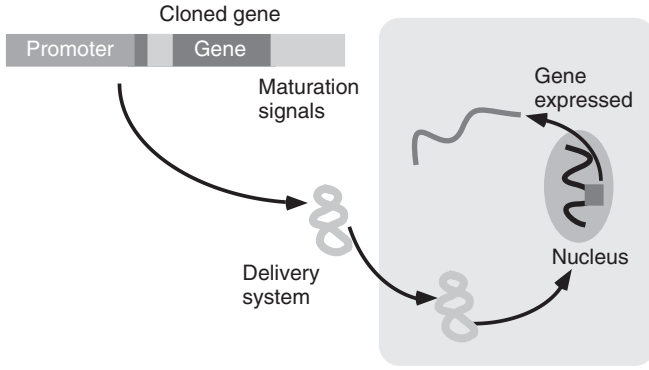


FIGURE 7.2. Localization of the gene to the nucleus is the consequence of several steps: delivery to the cell surface, binding to the cell membrane, crossing of the cell membrane into the cell interior, negotiation of the intracellular trafficking pathways to reach the nuclear membrane, translocation of the nuclear membrane, and finally unpackaging and release of the DNA from the carrier within the nucleus ready for transcription.

as a plasmid. As might be anticipated, many obstacles are present at each step in the path toward the nucleus. The body has many protective mechanisms to prevent foreign material, particularly genetic material, gaining access to the nucleus.

Many delivery systems have been studied, and the predominant problem that has arisen is the inability to achieve sufficient transduction of enough cells in the target tissue to achieve therapeutic levels of gene expression. This low level of gene expression also has to be balanced against any toxicity of the delivery vehicle.

Overall strategies to deliver genes fall in two major groups, strategies that use viruses to deliver the genetic material and nonviral systems.

NONVIRAL DELIVERY SYSTEMS

NAKED DNA The gene cloned into a plasmid backbone that can be amplified in bacteria, purified, and then administered, as "naked DNA" remains the simplest delivery system. This method

of DNA delivery is, however, highly inefficient at transfection, that is, at transporting the gene from the surface of the cell to the nucleus of the target cell. Levels of gene expression are therefore extremely low. However, muscle cells can be transduced, and even at low levels the quantity of protein expressed is sometimes sufficient to induce an immune response within the host. This approach may, therefore, lend itself to the use of DNA as a vaccine. For instance, if a gene that encodes a viral protein from a hepatitis virus is transfected and expressed, it may be possible to induce a protective immune response against that virus.

PARTICLE BOMBARDMENT Ballistic particle-mediated delivery systems or "gene guns" are being evaluated as a method to improve DNA delivery (Figure 7.3). This system is also a reasonably simple approach. The DNA sequence, for which there are no size restraints, is coated onto small heavy particles, either gold or tungsten, and fired at the tissue by a helium pressure device or "gun." The target tissue needs to be exposed, so the skin and wounds remain attractive targets. Transgene expression can be observed in both epidermis and dermal compartments following gene delivery to the intact skin. Efficiency of

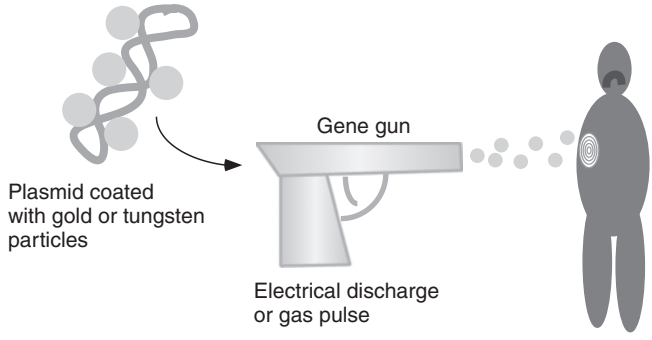


FIGURE 7.3. Particle bombardment. The DNA sequence is coated with heavy particles, either gold or tungsten, and fired at the tissue by a helium pressure device or "gun."

gene delivery is low, but sufficient protein can be expressed to induce an immune response. Applications that are being investigated include transfer of genes to express immunogenic proteins to immunize against influenza, viral hepatitis, HIV, and TB. Delivery of genes for growth factors to wounds and therapeutic genes to melanomas on the skin surface are also promising areas of investigation. This technology has also been used successfully to deliver genes to plants and fish.

Liposome Vectors Lipids with a neutral, negative, or positive charge can complex with DNA. However, positively charged or cationic lipids have shown the most promise in efficiency of gene delivery. Cationic liposomes are able to attach to negatively charged DNA, and still maintain an overall positive charge to facilitate attachment to the negatively charged surface of the cell. A complex of cationic lipid and nucleic acid is referred to as a **lipoplex**. Examples of cationic lipids include *N*-[1-(2,3-dioleoyloxy)propyl]-*N*,*N*,*N*-trimethylammonium chloride (DOTMA) and 3b-[*N*-(*N*,*N*-dimethylaminoethane)carbamoyl] cholesterol (DC-Chol). If a cationic polymer is used in place of a cationic lipid, the DNA conjugate is called a **polyplex**. Examples of cationic polymer include polyethylenimine and polylysine. Combinations of cationic lipid, cationic polymer, and DNA are called **lipopolyplexes**.

The problems that all of these complexes face are similar. First, the complex has to form, and this remains an empiric process dependent on many factors including charge, relative proportions of DNA and lipid/polymer, and ionic strength of the solution. The eventual structure is often unclear, but one favored structure for the lipoplex is for the DNA to be intercalated within two lipid bilayers (Figure 7.4).

Liposomes have shown considerable promise *in vitro*, but many obstacles exist in achieving success in *in vivo* applications. The interaction with serum can affect the colloidal stability of

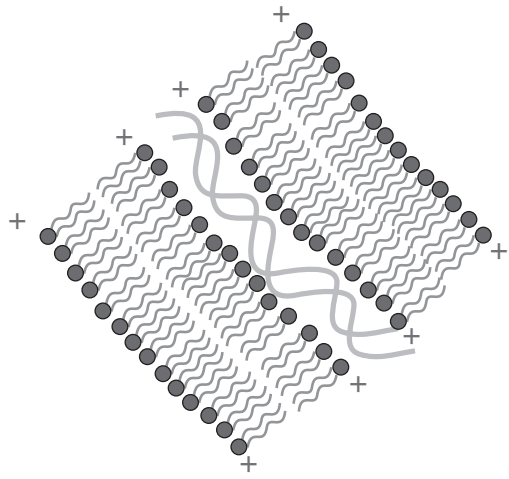


FIGURE 7.4. One favored structure for the lipoplex is for the DNA to be intercalated within two lipid bilayers.

the liposomes, and the liposome can be rapidly eliminated from bloodstream. The particle may also trigger the complement or clotting cascade, leading to toxic effects.

For the liposome to reach the target cells, they must often traverse a tissue matrix that can be a barrier for a charged particle. Once on the cellular surface, the liposome must fuse with the cell surface or undergo endocytosis. Endocytosis through clathrin-coated pits is probably the major pathway of DNA internalization (Figure 7.5). Degradation of the DNA within the endosome or lysosome is a significant problem after internalization, and the DNA has to escape from the endosome. Most liposome formulations, therefore, also include a fusigenic colipid along with the cationic lipid, which functions to facilitate the release of the DNA from the endosome.

The transfected DNA then needs to reach and to cross the nuclear membrane and be disassembled from any remaining carrier before transcription can begin. The pathway taken across the nuclear envelope is not clear, but certain DNA sequences known as **nuclear localization sequences** can facilitate this transport.

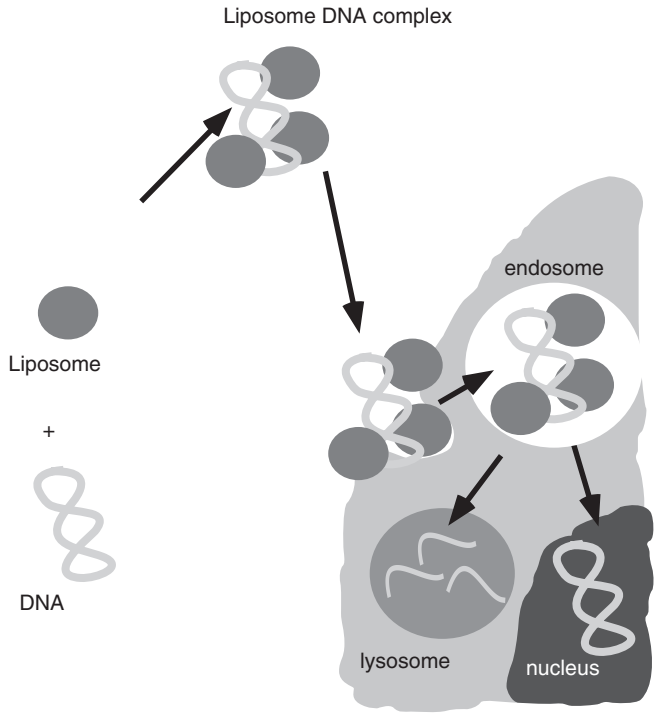


FIGURE 7.5. Liposomes must fuse with the cell surface or undergo endocytosis. Endocytosis through clathrin-coated pits is probably the major pathway of DNA internalization. Once within the cell, the complex has to break out of the endosome and then translocate the nuclear membrane.

The DNA is not stable in the cytoplasm because of the action of nucleases, and the low efficiency of nuclear transport all collude to limit the success of gene delivery. As a consequence of these factors, liposomes have not as yet fulfilled the initial promise seen in *in vitro* settings.

VIRAL DELIVERY SYSTEMS

Viruses are faced with the same problems as naked DNA and liposomes in the delivery of genetic material to the cell nucleus. However, over many millions of years viruses have evolved

mechanisms to overcome many of these obstacles, which make viruses attractive as vectors for gene therapy. Although there are similarities in the methods used by various viruses to transfer their genetic material, it is the distinct properties of specific viruses that predicate suitable gene therapy applications.

Viruses usually infect cells by targeting a cellular receptor that is used by the cell for other functions. This receptor targeting improves internalization of the virus into endosomes within the cell. Many viruses have also developed mechanisms to escape from the endosome in the cytoplasm, avoiding lysosomal degradation. For instance, the adenovirus achieves this by the action of the viral protein VI, which causes endosomal rupture at low pH. Viral DNA often contains nuclear localization signals that facilitate transport to the nucleus.

RETROVIRUSES The retroviruses are RNA-containing viruses, and the most commonly used ones are based on the murine leukemia virus. The main benefit of using the retrovirus as a gene therapy vector is stable integration of the transferred gene into the host chromosome, after reverse transcription of the viral RNA into DNA. The transferred gene becomes inserted into the host cell DNA, and therefore remains in the host cell through subsequent cellular divisions or until cell death. While this is an enormous benefit, it also has associated problems. In particular, there is no control over the site of integration in the host cell genome. This can lead to disruption of a gene at the site of insertion, or the transferred promoter/enhancer sequences can influence the expression of genes surrounding the site of insertion. Disruption of expression of a tumor suppressor gene, or enhanced expression of an oncogene, are clearly serious problems that could lead to the development of a malignancy. The development of replication-competent (pathogenic) retrovirus during manufacture by recombination of the vector with packaging cell components is also a concern.

Another drawback is that the retrovirus, when produced in a murine cell line, is not stable in the circulation and is subject to complement-mediated inactivation by the immune system. This therefore limits delivery options. Further, the target cell needs to be replicating for integration to occur. This requirement of target cell replication impedes application to many tissues with slow cellular turnover, like the respiratory epithelium, but may also be a benefit in specifically targeting cancer cells. Despite persistence of the integrated DNA within the genome of the cell, this persistence of the gene is not always associated with persistence of gene expression. Epigenetic events, in particular DNA methylation, can silence the transferred gene. This may represent a concerted response that has developed as a defense against retroviral DNA.

The steps taken to produce retroviral vectors are shown in Figure 7.6. The cDNA of the therapeutic gene is cloned into a "gutted" retroviral genome that has been deleted of the essential *gag*, *env*, and *pol* genes between the remaining inverted terminal repeats. This modified genome is then transfected into a producer cell line. A producer cell line is engineered to express the essential *gag*, *env*, and *pol* genes that were deleted from the retroviral genome. These essential functions for the production of the new virus are therefore provided "in trans." After transfection of producer cell with the retroviral DNA genome, retroviral RNA is transcribed and packaged into a virus by the producer cells that express the essential retroviral proteins. The resulting virus is infections, and can infect cells in the same manner as wild-type virus. Viral RNA is reverse-transcribed into DNA and the viral DNA integrated into the host genome. However, in the absence of the retroviral structural genes, no new virus can be produced.

The short half-life of retrovirus within the circulation has lead to several different approaches at gene delivery. Host cells, particularly lymphocytes or tumor cells, have been infected *ex vivo* and then readministered to the host. Also the producer cells

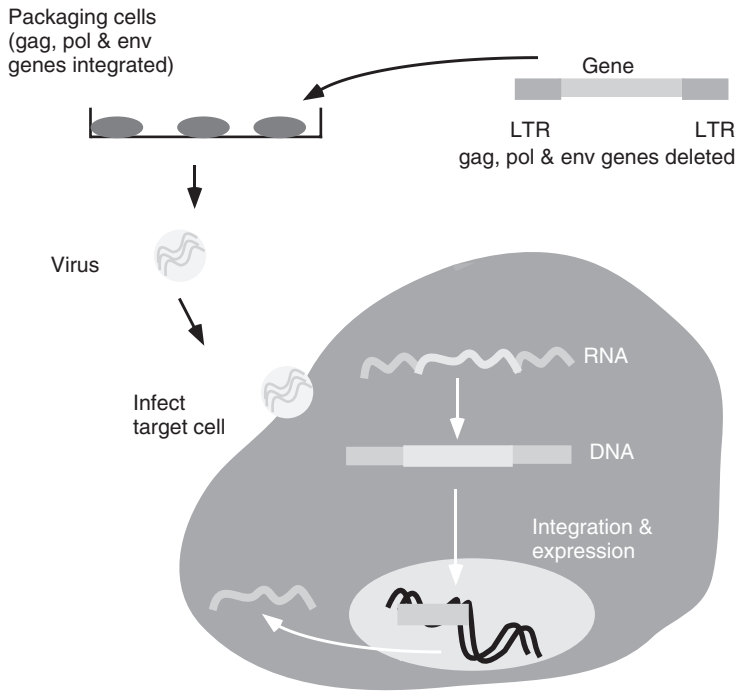


FIGURE 7.6. The steps taken to produce a retroviral vector. The gene to be expressed is cloned into a retroviral cassette devoid of essential retroviral genes, but flanked by the retroviral long terminal repeats. This DNA segment is transfected into a packaging cell line, and infectious but replication-deficient virus is produced.

themselves have been administered directly into certain tumors such as gliomas. The infected producer cells die within the tumor, and the virus is released to preferentially infect the replicating glioma cells.

ADENOVIRUS The biology of the adenovirus provides many properties that are well suited for gene therapy. As a consequence of some of these assets, the adenovirus is at present the most commonly used vector for gene therapy. The adenovirus is a DNA virus with a 36-kb double-stranded DNA genome. Beneficial features of the adenovirus include the ease of production of high

titers of infectious virus, and the capacity for insertion of reasonably large genes. The adenovirus is also very efficient at infecting many cell types, including nondividing cells, and is more stable in the circulation than is the retrovirus.

The virus infects the target cell by two interactions on the cell surface. The adenoviral fiber proteins, which protrude like spikes from the angles of the icosahedral viral particle, interact with the high-affinity adenoviral/coxsackie (CAR) receptor on the cell surface (Figure 7.7). The CAR receptor is widely distributed on the surface of most cell types, although there are tissue-specific variations in levels of expression. Following intravenous administration, preferential deposition from the bloodstream in

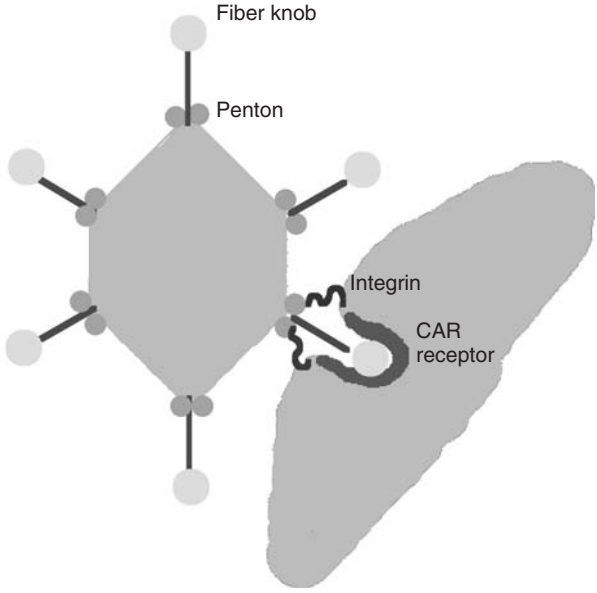


FIGURE 7.7. The adenovirus infects the target cell by two interactions on the cell surface. The adenoviral fiber proteins, which protrude like spikes from the angles of the icosahedral viral particle, interact with the high-affinity adenoviral/coxsackie (CAR) receptor on the cell surface. Viral penton proteins, which are located at the base of the fiber shafts, interact with integrins on the cell surface.

hepatic and pulmonary capillary beds appears to be more important than levels of CAR receptor expression. The second interaction on the cell surface is an interaction between viral penton proteins, which are located at the base of the fiber shafts, and integrins on the cell surface. This interaction between penton and integrin is a lower-affinity interaction, but triggers viral internalization by endocytosis into clathrin-coated pits.

The viral capsid is highly efficient at escaping from the endosome, causing endosomal rupture as the endosomal pH falls. The path to the nucleus is also fast and efficient, utilizing microtubules to deliver the capsid to the nuclear surface. The viral DNA is then injected into the nuclear interior ready for transcription of viral RNA to begin.

The adenovirus DNA does not, however, integrate into the genome but remains episomal. Infection is therefore transient, necessitating repeated administration for persistent gene expression. This is a problem, since the virus induces potent innate, cell-mediated, and humoral immune responses, and neutralizing antibodies may limit the success of repeated administration.

The steps in constructing a virus are shown in Figure 7.8. Replication-deficient viruses were the first to be used. The *E1a* regions of the viral genome, which is essential for viral replication, is deleted and replaced with the cDNA of the therapeutic gene. The viral genome is transfected into a cell line that has the viral *E1a* gene integrated in the host chromosomal DNA, thereby providing the essential functions lacking in the modified viral genome. Infectious but replication deficient virus for gene therapy applications is purified from the cell line.

The induction of an inflammatory response and immune response against viral proteins has been a problem with these first-generation viruses. Several modifications have therefore been introduced, to delete other viral genes such as the *E4* or *E2* genes, which encode proteins that can induce an inflammatory/immune response. Specialized cell lines that express the additional genes

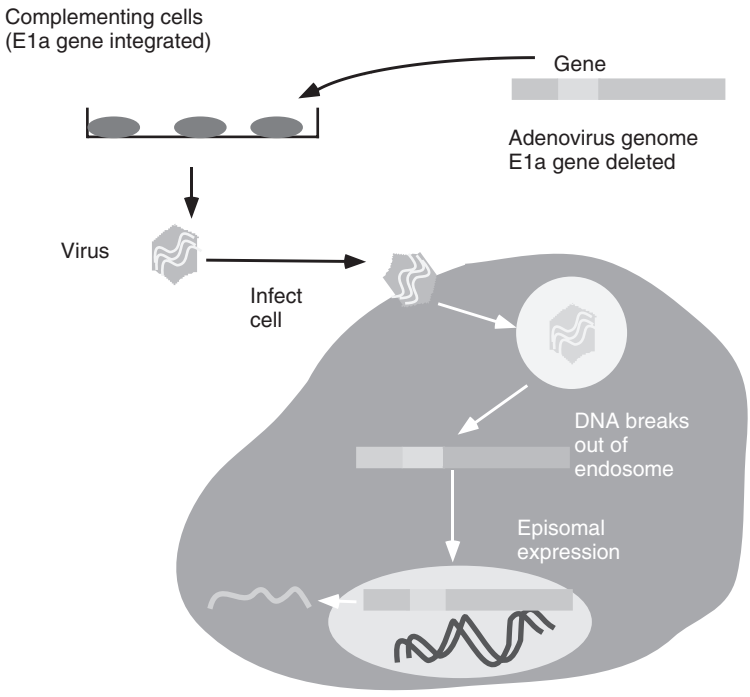


FIGURE 7.8. Construction of an adenoviral vector. The *E1a* region of the viral genome is deleted and replaced with the cDNA of the therapeutic gene. The viral genome is that transfected into a cell line in which the viral *E1a* gene is integrated in the host chromosomal DNA, thereby providing the essential functions lacking in the modified viral genome. Infectious but replication-deficient virus is purified from the cell line.

that have been deleted from the viral backbone (which are often toxic) are required to package these second generation viruses. This line of research has culminated in a viral backbone virtually devoid of all viral genes or a "gutless adenovirus," that is, dependent on a helper virus for production.

More recently, there has been great interest in using conditionally replicating adenoviral vectors in which the genes for replication remain essentially intact, but the virus is designed to replicate only in certain cell types. These strategies will be discussed in the section on targeting.

ADENO-ASSOCIATED VIRUSES Adeno-associated viruses (AAVs) are defective parvoviruses that were first noticed as contaminants in adenoviral stocks. An AAV is a single-stranded DNA virus that is dependent on a helper virus for completion of its lifecycle. Although the role of the helper virus is not completely clear, it appears to have a facilitatory role in rendering a cellular environment conducive to AAV replication. AAVs have several properties that are favorable for gene therapy vectors. They transduce both dividing and nondividing cells with reasonably high efficiency. They do not cause human disease or induce an inflammatory response, and although the human host can generate neutralizing antibodies, the immune response against AAV is limited but not inconsequential.

Adeno-associated viruses bind to heparin sulfate proteoglycans on the cell surface, and both integrin $\alpha v\beta 5$ and human fibroblast growth factor 1 can act as coreceptors. Wild-type AAVs in the absence of helper virus integrate into the host cell genome, and in a site-specific way to chromosome 19. However, integration is dependent on the viral *rep* gene, and when the viral backbone is modified to carry a transgene, integration is rarely seen. The capacity of AAV to accept a transgene is quite limited, at approximately 4.5 kb.

Adeno-associated viruses have been most actively investigation for clinical roles in cystic fibrosis, hemophilia, and ophthalmic disease.

HERPES SIMPLEX VIRUS Herpes simplex virus is a double-stranded DNA virus with a genome of 152 kb. The large size of the viral genome is conducive to the insertion of large transgenes, or alternatively to the insertion of multiple transgenes. The virus is able to undergo a nonintegrated latent infection of neuronal cells that is persistent and occurs without expression of viral lytic proteins. However, a neuronl-specific promoter is active during latency, and thus has the potential for long-term expression of a

transgene. The neural tropism of herpes simplex has predicated neurodegenerative disease as an obvious therapeutic area. The virus can, however, infect a wide variety of cells in both the dividing and resting states. Other disease areas that have been investigated include muscular dystrophy; the capacity of Herpes for large transgenes is an advantage in attempts to transfer the large (14-kb) dystrophin gene. Applications in cancer therapy, both as a toxic gene expressing replication-deficient virus and as a conditionally replicative lytic virus, are also being investigated.

Attempts to target the herpes simplex virus are difficult because of the complexity of the viral structure. A protein tegument and an envelope of glycoproteins surround the viral capsid. The virus attaches to heparin sulfate and glycosaminoglycans on the cell surface by an interaction with envelope glycoproteins. The virus is not endocytosed, but fuses with the cell membrane. Several receptors have been identified that play a potential role in viral internalization.

Vaccinia Virus Vaccinia is a double-stranded DNA virus of the pox family, with a large (200-kb) genome. Vaccinia is unusual in that replication and transcription occur in the cytosol, so host cell integration does not occur. The virus has a broad tropism and the capacity to insert large transgenes. Vaccinia induces a vigorous immune response and, although comparatively safe, can be fatal in immunosupressed individuals. Vaccinia is being investigated as an agent to deliver cytokine genes to various tumors, including bladder tumors and melanoma.

Alphavirus The Sindbis virus is the family member of the alphaviruses that has received the most attention. Alphaviruses are RNA viruses that can infect a broad host range of dividing and nondividing cells. Gene expression is at high levels but transient, as no integration occurs. Alphaviruses are being investigated for gene transfer to tumors, and for vaccine development.

LENTIVIRUS Lentiviral vectors are derived from the HIV virus but, unlike other retroviruses, are able to transduce nondividing cells. Lentivirus vectors can deliver 8 kb of transgenic DNA, which becomes integrated and therefore is expressed long-term in the host cell, without inducing a host immune response. Lentivirus vectors carrying therapeutic genes are being evaluated for β-thalassemia and Parkinson's disease.

HYBRIDS From the preceding discussion it can be inferred that most individual viruses have some highly desirable properties for a gene transfer vector, but also some problems that limit their full utilization. This has led to attempts to produce hybrid viruses to exploit the benefits and minimize the disadvantages of more than one viral system. An example includes an adenovirus–retroviral hybrid, where the retroviral structural genes and the therapeutic gene are packaged within an (usually two-separate) adenoviruses. The adenovirus can be produced at high titer, can efficiently infect cells, and then the retrovirus, produced *in situ*, can infect surrounding cells resulting in genomic integration of the transgene. The adenovirus is also being evaluated in hybrid vector designs as a "carrier" for adeno-associated virus (AAV), lentivirus, and even DNA transposable elements.

EMERGING TECHNOLOGIES Bacteria, in particular *Shigella flexneri*, and other viruses including baculovirus, feline parvovirus and measles virus are all being studied for possible roles in gene delivery.

TARGETING GENE DELIVERY

The two main reasons to target gene delivery are to enhance specificity and efficacy. By specifically limiting transgene expression to the target tissue, safety is improved. By evading delivery of the vector to nontarget tissues, both efficacy and safety are

improved. The obstacles and impediments to deliver DNA to the nucleus are many, and targeting can assist at several levels.

DIRECT ADMINISTRATION

The simplest way to target a tissue is direct administration of the vector to that tissue, and most initial clinical trials have taken this approach. Viruses or liposomes have been directly injected into tumor masses for cancer gene therapy, or applied directly to the surface of respiratory tract for cystic fibrosis. The main advantage is the simplicity, and the avoidance of barriers to vector delivery such as the vasculature. Although these approaches may provide a proof of principle, the hope for the future is clearly to develop molecular targeting approaches so that the vector can be given systemically.

RECEPTOR-MEDIATED GENE DELIVERY

Cell surface receptors provide an avenue for both targeting gene delivery to the cell surface and triggering endocytosis and vector internalization.

LIGANDS

The requirements to target a specific cell type are that the cell in question expresses a unique receptor or overexpresses a more widespread receptor, and that a ligand to that receptor has been identified. Examples of various ligand–receptor combinations that have been used are shown in Table 7.1.

The various ligands listed in Table 7.1 have usually been associated with DNA by using a polylysine bridge. When the complexes are delivered to the cell surface, receptor-binding and often receptor-mediated internalization of the complex can aid gene delivery (Figure 7.9). Ligand–receptor interactions that have been evaluated are shown in the table. Despite the elegance of this

TABLE 7.1. Ligand–Receptor Combinations

Receptor	Ligand	Cell Type	Delivery Vehicle
Asialoglycoprotein receptor	Asialoglycoprotein	Hepatocyte	Polylysine
Transferrin receptor	Transferrin	Several	Polylysine
Epidermal growth factor receptor	Epidermal growth factor	Cancer	
Folate receptor	Folate	Several	Polylysine
Surfactant protein receptor, A and B	Surfactant protein receptors A and B	Airway epithelium	Polylysine
Polymeric immunoglobin A receptor	IgA	Airway epithelium	Polylysine

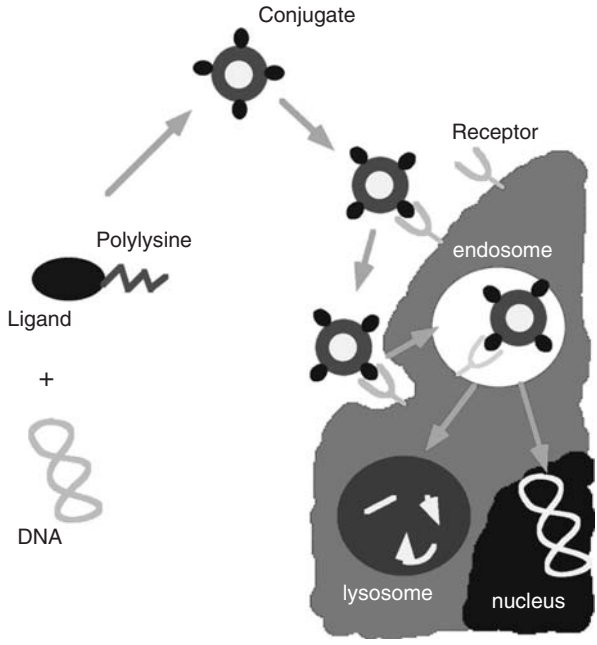


FIGURE 7.9. Ligands are associated with DNA by using a bridge-like polylysine. When the complexes are delivered to the cell surface, receptor binding and often receptor-mediated internalized of the complex can aid gene delivery.

approach, this field of gene therapy has been hampered by poor efficacy of gene expression, and DNA degradation within the cell.

ANTIBODIES

Monoclonal antibodies specific for various cellular targets have been explored in several settings. Antibodies directed against a specific target can be added to lipoplex or polyplex complexes to target the vector to the desired cell. Both retroviruses and adenoviruses have also been retargeted using monoclonal antibodies. To retarget the virus, bispecific antibodies have been constructed that bind to both a virus-coated protein and the desired cellular target.

Viral Protein Modification

Viruses have developed many strategies for efficient delivery of their genetic material into the nucleus of the target cell. The first is an ability to bind to receptors on the cell surface. Different viruses usurp different cellular receptors to enable binding to the cell surface and then internalization into endosomes. The "choice" of receptor targeted by the virus influences the repertoire of tissues that can be infected. For instance, the receptor for the commonly used adenovirus Ad5, which is also shared by the coxsackievirus, is expressed on many epithelial surfaces but at low levels on hematopoetic cells. This therefore emphasizes the epithelium as a target for gene transfer, but limits the competence of the adenovirus as a gene therapy vector for hematopoietic cells. As an aside, this dichotomy of receptor expression has been ingeniously exploited for therapeutic use, by using an adenovirus expressing a toxic gene to specifically infect and kill breast cancer cells within bone marrow cells ex vivo prior to bone marrow transplantation for breast cancer.

The diversity of cells that express receptors for the adenovirus is also a problem for targeting, and this has led to attempts to change the receptor specificity of the virus to both limit infection to certain cells and enhance infection of cells expressing low levels of CAR receptor. The native wild-type virus Ad5 utilizes the coxsackie adenoviral receptor to mediate binding, and an integrin (usually $\alpha v\beta 5$) to mediate internalization into the cell. The virus binds to its receptor via the knob region of the fiber and the intregrin via RGD sequences within the penton proteins at the base of the fiber. If the knob region of the fiber is modified and RGD sequences or polylysine sequence are inserted, the virus is no longer limited to infecting cells that express the CAR receptor. This approach lessens specificity but increase efficacy of viral infections. Conversely, the fiber knob can be modified to express a specific ligand, for example, the gastrin-releasing peptide, and

thereby the virus can be targeted to specific cell types that express the receptor for this ligand. It is also possible to swap the Ad5 fiber for the fiber of other adenovirus serotypes that possess different host cell propensities.

SITE-SPECIFIC REPLICATION

A major limitation in cancer therapy is the inability to achieve sufficiently high levels of the therapeutic agent within the tumor cell to ensure efficacy, but maintain low levels in nontarget tissues to provide safety. The concept of using a virus that could specifically replicate in tumor cells, to increase the local "dose" several thousandfold, is an exciting proposition.

Several innovative strategies, mainly using adenovirus, but also using measles, herpes, or vaccinia viruses have been pursued to achieve this aim of tumor-specific viral replication. One strategy involves changing the promoter of the adenoviral *E1a* gene. The adenoviral *E1a* gene is essential for viral replication, and the native *E1a* promoter is active in all cell types. Replacement of the viral promoter with the promoter of a gene that is preferentially expressed in a tumor, for example, the α-fetal protein gene promoter for hepatoma or the prostate-specific antigen promoter for prostate cancer, can limit viral replication to the targeted cell type. The target cell can then be killed, either as a result of the lytic viral infections or by transfer of a therapeutic viral gene.

Another approach is to exploit the similarities in viral and tumor cell biology (Figure 7.10). In essence, both tumor cells and adenovirus-infected cells share two common needs: to undergo cell division and to overcome signals for apoptosis to stay alive despite uncontrolled cell division. This duplication of function can be exploited to target viral replication to tumor cells. The adenovirus needs to block the function of the host tumor suppressor gene *p53* to be able to push the host cell into cell divisions and replicate efficiently. However, the majority of tumors already

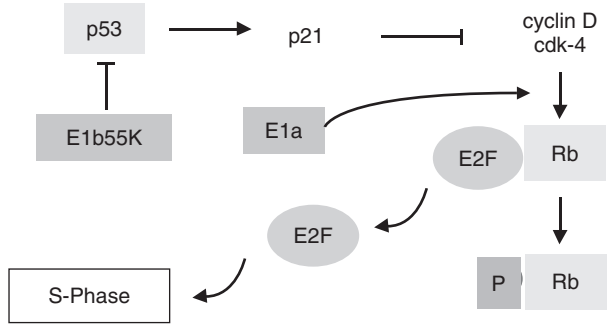


FIGURE 7.10. Two approaches to tumor–specific viral replication: (1) *adenoviral E1b55kD-deleted viruses*—in normal cells, adenoviral *E1b55k* is necessary to inhibit p53 to enable viral replication, whereas in tumor cells with mutated p53, *E1b55k* is not necessary, and an *E1b55k*-deleted virus will replicate only in cancer cells; (2) *adenoviruses with E1a mutations*—in normal cells, *E1a* releases *E2F* from Rb to enable S-phase entry and viral replication, while in tumor cells with mutated Rb, Rb binding of *E1a* is not required; an *E1a*-modified virus should therefore replicate only in cancer cells.

have mutations in the p53 gene that render the p53 protein nonfunctional. A viral mutant with the p53-blocking gene (*E1b55kD*) removed would therefore be blocked from replicating in normal cells with active p53, but not in tumor cells with inactive p53. Similar approaches are also being explored with mutations in the adenoviral *E1a* gene to target Rb-mutated cells.

Natural Targeting

Two viruses that are being evaluated for cancer treatment are naturally targeted to tumors. The reovirus is unable to replicate in normal cells because of normal host cell defense mechanisms that block the translation of viral proteins. However, activation of the *ras* pathway, which is active in many tumor cells, reverses the block on viral translation and enables the reovirus to replicate and lyse the tumor. Vesicular stomatitis virus may also behave in this way. The Sindbis virus targets the laminin receptor on cells for viral entry. Fortuitously, laminin receptors on tumor cells are less

occupied than laminin receptors on the normal cells facilitating the preferential infection of cancer cells.

Formative Years and Initial Clinical Approaches

Since the initial clinical trials in the early 1990s, the field of clinical gene therapy has rapidly expanded. A 2007 review in the *Journal of Gene Medicine* documented 1260 active clinical trials, including 40 phase II/II or phase III trials, that is, trials to determine the effectiveness of therapy. Cancer is the predominant indication (67% of trials) and the adenovirus the most frequently used vector (25% of trials). Genes for cytokines are the most frequently used gene (27%)

Gene Replacement

Adenosine Deaminase (ADA) Deficiency ADA is the cause of severe combined immunodeficiency. In the absence of ADA, a toxic product accumulates in lymphocytes resulting in dysfunction of both T and B cells. A gene therapy approach to restore a normal copy of the ADA gene to lymphocytes was the first gene therapy trial to negotiate the ethical, regulatory, and safety bodies and to be approved in the United States. In the first trial that started in 1990 at the National Institutes of Health, two children received retroviral mediated transfer of the ADA gene to their lymphocytes ex vivo over a 2-year period. Evidence for the persistence of transduced cells over several years have been obtained. However, these patients also received standard therapy for ADA, so confirming evidence of a clinical response from the gene therapy itself has been difficult. In several subsequent trials, peripheral lymphocytes, bone marrow cells, and umbilical cord blood cells of neonates with ADA deficiency have been transduced. Persistence of transgene has been seen frequently.

SCID-X1 Mutation of the γ-cytokine receptor subunit of the interleukin IL-2,4,7,9,15 receptors results in a severe combined immunodefeicney. A gene therapy approach for SCID-X1, utilizing the retrovirus for ex vivo transfer of the corrected gene to CD34+ cells, has resulted in full restoration of function to 12 of 14 infants in two European trials. Unfortunately, T-cell leukemias that were secondary to retroviral vector insertional mutagenesis developed in three patients, in two by vector insertional activation of the *LMO2* transcription factor gene.

CYSTIC FIBROSIS **Cystic fibrosis** (CF) is a homozygous recessive disorder that is the most frequent inherited disease in the Caucasian population. The gene that is mutated in CF encodes a chloride channel termed the **cystic fibrosis transmembrane conductance regulator** (CFTR). The brunt of the disease is manifest on the respiratory epithelium that becomes particularly susceptible to bacterial infection. Frequent infections of the bronchi lead to tissue distortion and destruction, a disease process called **bronchiectasis**. Although the exact mechanism by which abnormal gene function leads to disease is unclear, individuals heterozygous for the gene mutation are free from disease, and experimentation had shown that as few as 1 corrected cell in 10 could reverse the abnormal CF phenotype. In addition, the respiratory epithelium (which includes the nasal epithelium) is reasonably accessible, and efficiency of gene transfer can be measured repeatedly by both molecular and physiological techniques to ascertain functional correction.

For all of these reasons, there was intense interest, and many protocols were initiated to attempt a functional correction of the respiratory epithelium in individuals with CF. Delivery methods that were used include the adenovirus, adeno-associated virus, and liposomes. All of the techniques used showed some evidence of gene transfer and functional correction, but it also became apparatent that all the techniques were faced with considerable

barriers in acieving effective gene transfer. In addition, one individual developed an inflammatory response within the lung following the administration of adenovirus. This was the first clinical suggestion that gene delivery, in this case by the adenovirus, may be associated with adverse effects.

Another problem that this strategy faced was the need for repeated administration, since none of the vector systems resulted in integration of the transgene. This posed particular problems for the first-generation adenovirus, because a brisk cell-mediated and humoral immune response was induced. Newer-generation viruses express much less viral protein and are therefore likely to be an improvement.

Familial Hypercholesterolemia Familial hypercholesterolemia is an autosomal dominant disorder in which the gene encoding the low-density lipoprotein receptor is defective. Reduced levels of this receptor result in high levels of circulating cholesterol, leading to premature atherosclerosis and myocardial infarction. Individuals at particular risk are those who are homozygous for the mutation, in which circumstance life expectancy is markedly reduced. Low levels of receptor expression (>10% of normal) have clinical benefit, supporting attempts at genetic correction. A report published in 1995 described attempts to partially restore receptor expression in the livers of five individuals homozygous for the genetic defect. The vector chosen was a retrovirus, which would lead to stable integration of the transgene. Unfortunately, hepatocytes have a very low basal level of replication. A partial hepatectomy was therefore performed to provide hepatocytes that could be cultured and infected ex vivo and then returned to the patient through a portal venous catheter. This study demonstrated that gene expression could be detected for 4 months, and no adverse effects were seen. The metabolic effects were however variable, and subsequent research has been directed to more efficient means of transgene delivery.

HEMOPHILIA Autologous skin fibroblasts containing a normal copy of the factor VIII gene, which was delivered to the fibroblasts by electroporation, have been given into the peritoneum of six individuals with hemophilia A. Detectable but transient (undetectable at 10 months) serum levels of factor VIII were achieved.

Intramuscular injection into dogs with severe hemophilia B of an AAV vector-expressing factor IX has resulted in a dose-dependent increase and prolonged expression of circulating levels of factor IX, at a level that would be sufficient to provide phenotypic improvement in humans. A clinical trial in seven individuals of AAV-mediated expression of factor IX, administered by portal vein infusion of AAV-expressing factor IX, has resulted in therapeutic but transient serum levels of factor IX. Immune-mediated clearance of the relatively poorly immunogenic adeno-associated virus has proved to be the limiting factor.

LEBER CONGENITAL AMAUROSIS Leber congenital amaurosis causes near total blindness in infancy and can result from mutations in the RPE65 gene. A naturally occurring animal model, the RPE65-/- dog, suffers from early and severe visual impairment similar to that seen in the human disease. Using a recombinant adeno-associated virus carrying wild-type RPE65, visual function has been restored in this canine animal model of childhood blindness. Very recently three patients who have been treated in this way have achieved improvements in retinal function without side effects.

OTHER GENETIC DISEASES Gene therapy approaches for other genetic diseases that have recruited patients into clinical trials include protocols for chronic granulomatous disease, ornithine transcarbamylase deficiency, Canavan disease, mucopolysaccharidosis type I, Gaucher's disease, α-1-antiprotease deficiency, Fanconi anemia, and leukocyte adherence deficiency.

Clinical protocols that have been developed and reviewed by regulatory authorities for Huntington's disease, gyrate atrophy,

muscular dystrophy, Fabry disease, amyotrophic lateral sclerosis, junctional epidemolysis bullosa, and purine nucleoside phosphorylase deficiency.

Expression of Therapeutic Gene

Revascularization Research has progressed in several directions to develop therapies to maintain or develop a vascular supply. One of the main problems with coronary artery vein grafts is the subsequent development of neointimal hyperplasia that narrows the graft lumen and thereby leads to impairment of vascular supply to the heart. Strategies to prevent the accumulation of new intimal cells include the use of oligonucleotides that block the effects of genes that are important in cellular division. A phase III study to assess the efficacy and safety of pretreating vein grafts before coronary artery bypass grafting with **Edifoligide**, an *E2F* oligonucleotide decoy, has been performed. *E2F* is an important transcription factor that regulates genes that increase cellular division. The presence of a decoy that mops up the *E2F* transcription factor might therefore block cellular proliferation. In this randomized, double-blind, placebo-controlled trial of 3014 patients undergoing primary CABG surgery, Edifoligide had no effect on the primary endpoint of vein graft failure.

Attempts have also been made to coat stents that are used for coronary artery stenting with endothelial cells that have been modified to express a therapeutic gene. The therapeutic gene may inhibit clot formation or secrete vascular growth factors to stimulate new vessel formation downstream.

Vascular endothelial growth factor has also been administered using an adenovirus directly to the myocardium to stimulate new vessel formation in ischemic hearts. In a randomized trial for severe angina due to coronary artery disease, the use of an adenovirus expressing the VEGF gene AdVEGF121 given by injection into the myocardium at a minithoracotomy significantly prolonged exercise time at 26 weeks. A plasmid-based approach

using a similar VEGF gene has also shown improvement in myocardial wall motion, although not perfusion.

NEURONAL DISORDERS For Parkinson's disease, the transfer of the gene tyrosine hydroxylase to relieve symptoms or of antioxidant genes to lessen the neuronal deterioration are being explored. In animal models a gene therapy approach in which a recombinant adeno-associated virus carrying the human aromatic L-amino acid decarboxylase (AADC) gene infused into the striatum resulted in an improvement in L-dopa responsiveness.

For Alzheimer's disease, the transfer of nerve growth factor (NGF) to limit neuronal degeneration has received consideration. A clinical trial using genetically modified fibroblasts to express ciliary neurotrophic growth factor has also been performed.

Eight individuals with mild Alzheimer disease had autologous fibroblasts genetically modified to express human NGF implanted into the forebrain. No long-term adverse effects occurred, and there was a suggestion of improvement in the rate of cognitive decline plus improvements in serial positron emission tomography (PET) monitoring of brain activity.

STRATEGIES FOR CANCER THERAPY

For many solid-organ tumors, advances in therapy since the mid-1980s have not led to substantial improvements in cure rates or survival. Gene therapy provides a potentially new therapeutic modality. The problems faced when developing a therapy for cancer include achieving a sufficiently high dose of the agent within the tumor to enable effective tumor cell killing while at the same time targeting the therapy so as to minimize toxicity to normal tissues. Many tumors have low levels of immunogenicity, which limits the natural host defense against the tumor. Tumors are genetic diseases that result from multiple gene mutations; some of the mutations may be inherited, but most are acquired. These mutations can lead to gain of functions; for instance,

mutation of the *ras* gene that leads to the unrestrained activation of the RAS protein are called **oncogene mutations**. Conversely, the loss of function of a gene that represses tumor formation can occur, such as mutations in the p53 gene, and this is called a **tumor suppressor mutation**.

Treatment strategies have fallen into the following main areas: modification of the immune response, expression of a therapeutic drug, correction of any acquired genetic mutations within the tumor, and modification of nontumor stem cells to allow the administration of higher doses of conventional therapy.

IMMUNE RESPONSE MODIFICATION This strategy encompasses two areas (Figure 7.11). One is to try and make the tumor more immunogenic, and the other is to recruit antigen presenting cells to the tumor. To induce an immune response, tumor antigens have to be presented to T cells. Although many tumor cells express antigens, they are not efficiently presented. In addition, tumor cells rarely express the appropriate ligands to provide the

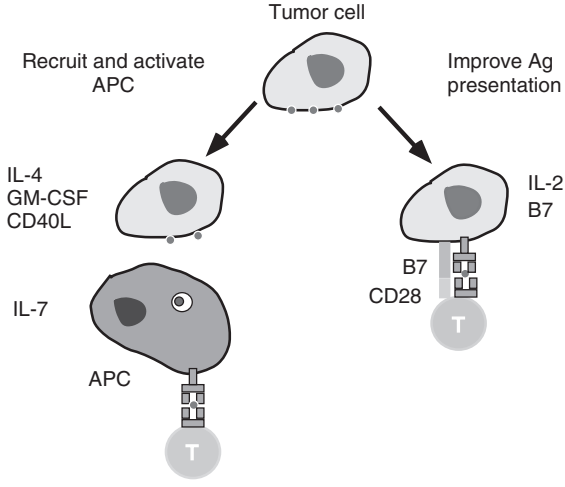


FIGURE 7.11. Strategies to enhance antitumor immunity. Genes can be transferred to tumor cells to either improve antigen presentation or recruit antigen-presenting cells.

second signal required to activate lymphocytes directly, and they are not effective at recruiting antigen-presenting cells to perform this task. Gene therapy vectors have therefore been used to try to improve antigenic presentation by the tumor or to transfer to the tumor a gene expressing a lymphocyte second signal such as the B1 ligand, which binds to the lymphocyte CD28 receptor. This enables the tumor antigen to be presented to T cells in combination with the required second signal to activate the T cell. An alternative is to modify the hosts T-lymphocytes with an engineered T-cell receptor highly specific for the tumor cells. Gene transfer can be effected either by direct injection of the vector into the tumor or by infection of host cells ex vivo, followed by readministration of the modified cells.

Another approach is to transfer to the tumor a gene that will activate and recruit antigen-presenting cells to the tumor to more efficiently present tumor antigen. Genes under investigation, among others, include GM-CSF, CD40L, and IL-1.

As an example, a phase I clinical trial in patients with metastatic non-small-cell lung cancer has been performed. Tumor cells from resected metastases were infected with a replication-defective adenoviral vector encoding GM-CSF irradiated, and returned to the patient by injection under the skin. This "vaccination" elicited immune responses as evaluated by the development of delayed-type hypersensitivity reactions to irradiated nontransfected tumor cells in 18 of 22 patients.

THERAPEUTIC GENES Genes that encode proteins that are cytotoxic to cancer cells have been transferred to tumors, and these include genes for cytokines such as tumor necrosis factor (TNF), or proapoptotic proteins like Bax. Another strategy is to transfer to the tumor a gene that can activate a prodrug to a toxic product. An example of this is the use of the herpes simplex thymidase kinase gene, which, when expressed, confers on the cell the ability to phosphorylate a prodrug gancyclovir to a toxic product.

GENETIC MODIFICATION OF TUMOR As already discussed, tumors develop as a consequence of the accumulation of several genetic mutations. Some of these changes result in an increase in gene activity that induces cell division and tumor-like behavior, yet other mutations can block the effects of genes that normally function to prevent tumor-like changes. Gene therapy strategies to inhibit tumor gene (oncogene) expression include using antisense sequences to block the *ras* gene. Attempts to replace the function of the mutated p53 tumor suppressor gene have been attempted using several vector systems and for several different malignancies. A nonreplicating adenovirus that expresses a normal copy of the p53 gene has been approved for clinical use by regulatory authorities in China.

PROTECTION OF STEM CELLS Bone marrow stem cells are usually very sensitive to the toxic effects of chemotherapy, and this can limit the administration of effective tumor-killing doses. A gene therapy application to try and remedy this is the transfer to marrow stem cells the gene for the multidrug resistance protein or P-glycoprotein. This treatment is intended to increase the resistance of stem cells to the toxic effects of the chemotherapy to enable the administration of more effective doses of chemotherapy.

REPLICATING ADENOVIRAL VECTOR A replicating adenoviral vector targeted to tumor cells with mutated p53 by deletion of the adenoviral *E1b55kD* gene has been shown in combination with chemotherapy to induce some complete response of injected lesions in head and neck cancer. In 2006, a virus of this type was approved for clinical use in China.

THE PROBLEMS

The application of genes as a therapy is faced with the same hurdles as is the application of any new medicine. The therapy must

have the desired therapeutic effect, and this must be achieved with an acceptable level of toxicity. Most of the gene therapy trials have approached diseases that allow a large therapeutic widow, where gene overexpresion is not likely to produce serious adverse effects and a small amount of gene transfer may have beneficial effects. If cystic fibrosis is taken as an example, correction of 1 in 10 cells is probably adequate to provide functional correction of an epithelial layer, and gene overexpression does not appear to be toxic. For a disease like diabetes, where precise regulation of insulin secretion would be required, many more hurdles need to be overcome.

Unlike chemotherapy agents, where the predominate problem is toxicity, the main problem that gene therapy trails have faced is lack of efficacy. The impressive gene transfer and expression that is seen with a variety of vectors in *in vitro* culture systems has not translated well to human trials. Achieving adequate and persistent levels of gene expression at the target site has proved difficult to achieve. This is a consequence of many factors; different vector systems have their own individual Achilles heel. Retroviruses, for instance, are not stable in the circulation, as they are rapidly inactivated in a complement-dependent manner. Early-generation adenovirus has been shown to induce a robust immune response and to be rapidly cleared from animals by immune and nonimmune mechanisms. Genes that have been transferred and become stably integrated also have shown loss of expression over time. This gene silencing probably occurs as a result of promoter methylation.

In contrast to the low efficacy, toxicity has not been a major problem, with the exception of one or two highly publicized adverse events, the most notable of which was the death of a young man with ornithine transcarbamylase deficiency following the administration of an adenoviral vector directly into the portal vein. This incident lead to a rigorous evaluation of the conduct of gene therapy trials and the evolution of the *Office of Human Research Protection* to provide a more detailed guidance on the

performance and monitoring of compliance of clinical trials. This included reevaluation of adverse-event reporting in both gene therapy trials and other investigator-initiated protocols.

The Future

Despite the relatively few therapeutic successes to date, this first 17 years of clinical gene transfer has provided an enormous amount of scientific and clinical information to provide a firm platform from which to move forward to the future. Two adenoviral products for cancer have even been approved for clinical use in China, although presently results of the efficacy of this approach are not fully clear.

The main focus of activity is likely to remain on achieving adequate levels of gene expression in the absence of inflammatory and immune responses to achieve meaningful modulation of human disease. Despite great success in a dog hemophilia model using a poorly immunogenic virus, immune-mediated vector clearance appeared as a problem in clinical trials. The concept of actually exploiting the immune response by using genes as a "vaccination" to induce immune responses against tumors is also being actively investigated. Interest in nonreplicating and replicating adenovirus for cancer treatment remains high with the approval of two products in China. Interest has also increased in using gene therapy together with standard therapy for cancer. There are now commercial interests in gene therapy that are leading to the rapid evaluation of gene therapy products in clinical trials.

References

Acland GM, Aguirre GD, Ray J, Zhang Q, Aleman TS, Cideciyan AV, Pearce-Kelling SE, Anand V, Zeng Y, Maguire AM, Jacobson SG, Hauswirth WW, Bennett J. 2001. Gene therapy restores vision in a canine model of childhood blindness. *Nat Genet* **28**(1):92–95.

Avery OT, Macleod CM, McCarty M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J Exp Med* **79**:137–158.

Blaese RM, Culver KW, Miller AD, Carter CS, Fleisher T, Clerici M, Shearer G, Chang L, Chiang Y, Tolstoshev P et al. 1995. T lymphocyte-directed gene therapy for ADA- SCID: Initial trial results after 4 years. *Science* **270**(5235):475–480.

Cavazzana-Calvo M, Hacein-Bey S, de Saint Basile G, Gross F, Yvon E, Nusbaum P, Selz F, Hue C, Certain S, Casanova JL, Bousso P, Deist FL, Fischer A. 2000. Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science* **288**(5466):669–672.

Crystal RG, McElvaney NG, Rosenfeld MA, Chu CS, Mastrangeli A, Hay JG, Brody SL, Jaffe HA, Eissa NT, Danel C. 1994. Administration of an adenovirus containing the human CFTR cDNA to the respiratory tract of individuals with cystic fibrosis. *Nat Genet* **8**(1):42–51.

Griffith F. 1928. The significance of pneumococcal types. *J Hygiene* **27**:113–159.

Grossman M, Rader DJ, Muller DW, Kolansky DM, Kozarsky K, Clark BJ, 3rd, Stein EA, Lupien PJ, Brewer HB, Jr, Raper SE et al. 1995. A pilot study of ex vivo gene therapy for homozygous familial hypercholesterolaemia. *Nat Med* **1**(11):1148–1154.

High KA. 2001. Gene transfer as an approach to treating hemophilia. *Circ Res* **88**(2):137–144.

Khuri FR, Nemunaitis J, Ganly I, Arseneau J, Tannock IF, Romel L, Gore M, Ironside J, MacDougall RH, Heise C, Randlev B, Gillenwater AM, Bruso P, Kaye SB, Hong WK, Kirn DH. 2000. A controlled trial of intratumoral ONYX-015, a selectively-replicating adenovirus, in combination with cisplatin and 5-fluorouracil in patients with recurrent head and neck cancer. *Nat Med* **6**(8):879–885.

Mann MJ, Whittemore AD, Donaldson MC, Belkin M, Conte MS, Polak JF, Orav EJ, Ehsan A, Dell'Acqua G, Dzau VJ. 1999. Ex-vivo gene therapy of human vascular bypass grafts with E2F decoy: The PREVENT single-centre, randomised, controlled trial. *Lancet* **354**(9189):1493–1498.

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL et al. 1989. Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA. *Science* **245**(4922):1066–1073.

CHAPTER

8

MICROARRAYS

The growth of the term *genomics* both among biologists and in the popular media has been driven largely by the development of microarray technology for the measurement of gene expression. The concept of a gene array is quite simple—a large number of DNA sequences for known genes (targets) are attached in defined locations on a surface (an array of spots). Many different surfaces have been used for DNA arrays, but glass microscope slides coated with a substance that enhances the covalent binding of DNA have become the standard (see Figure 8.1). (These glass slides bearing spots of DNA are often called "DNA chips"—but the term **GeneChip™** is a trademark of the Affymetrix Corporation.) A labeled test sample of mixed RNA sequences extracted from some cells (the probe) is then applied to the surface. Probe sequences bind by RNA–DNA hybridization to complementary target sequences in the array. Then the amount of labeled RNA bound to each target spot is measured. The measurement of RNA probe bound to each target will reflect the amount of that RNA sequence in the test sample, and therefore the level of expression of that gene. A microarray is simply a gene array with the DNA targets applied in very tiny spots.

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

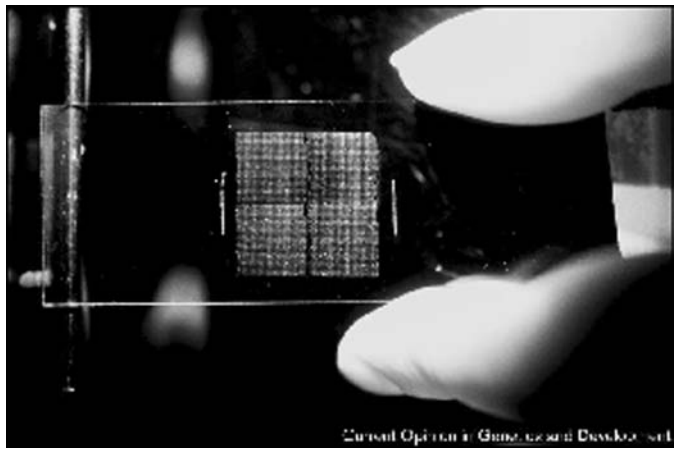


FIGURE 8.1. A DNA microarray on a glass slide (Ferea and Brown 1999).

Microarrays represent an extension of two older molecular biology techniques: the measurement of gene expression by using hybridization to quantify the amounts of specific mRNAs (i.e., "Northern" blots), and the measurement of many genes at once in a "dot blot." Microarrays are a genomic technology because the concept of a dot blot has been scaled up to allow for the measurement of thousands of genes in parallel. Ideally, a microarray would contain a target for every gene in a genome. This has been essentially achieved for the ~6000 genes in yeast (DeRisi et al. 1997), but the genome sequences are not yet sufficiently complete to identify with confidence and make a DNA probe for every gene for humans and other higher eukaryotes.

Microarray experiments are always conducted as a comparison between two (or more) samples. The relative amount of probe bound to each spot in the array is compared between the two samples, and a ratio is calculated. This ratio may be expressed in absolute terms, or more typically, in terms of a fold change—so that a given gene might be observed to increase in expression 2.5-fold when the target tissue is subjected to an experimental condition as compared to the control condition. A different gene

will undergo a different relative change in expression level across the same two samples. Two samples may be compared by making separate measurements of labeled samples on two identical arrays, or two samples may be labeled with different colors of label (different fluorescent makers) and mixed together in a single hybridization reaction on a single array (Shalon et al. 1996). Even if two samples are mixed and hybridized together, the two different probe colors are measured independently by a fluorescence scanner, then the two images are combined to create a false-color image. Typically red is used to represent one probe and green the other. Spots that show up as red or green in the combined image indicate a significantly higher level of mRNA for that gene in one sample as compared to the other. Yellow spots indicate high levels of expression in both samples, and dark spots indicate low expression in both samples (see Figure 8.2).

This measurement of RNA extracted from a tissue sample is actually a snapshot measure of mRNA transcript abundance. This is not exactly the same as a direct measure of gene expression—the actual amounts of each gene product (i.e., protein) being manufactured in the cell at a given moment in time—nor is it a direct measurement of DNA transcription into mRNA. Other factors such as different translation efficiencies, different rates of degradation of mRNA species, and alternative

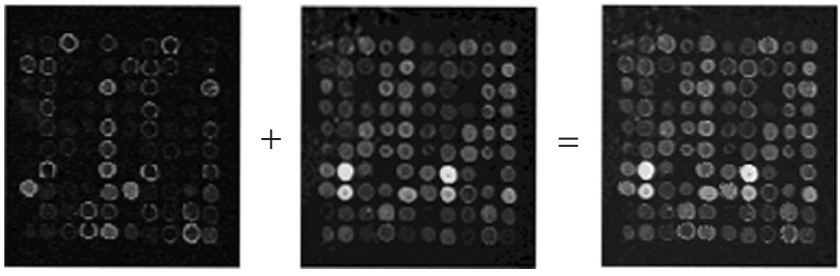


FIGURE 8.2. Two fluorescent images of a microarray with red and green false colors combined to show relative gene expression in two samples. (See insert for color representation.)

splicing play a role in modulating gene expression. However, measurements of mRNA are usually well correlated with both transcription levels and levels of protein abundance in the cell.

Gene expression microarrays composed of DNA targets (gene chips) are intended to report on the transcriptional levels of all genes to facilitate investigation of metabolic differences caused by disease, development, or other variables. However, alternative splicing of mRNA must be accounted for in the design of these chips. At a minimum, alternate splice forms can confuse results since they will show differential binding to probes that bind to exons that are sometimes spliced out. If knowledge of alternate splice forms is incorporated into DNA chip design, then expression of alternate forms can be directly measured and incorporated into the overall gene expression profile.

Spotting versus Synthesis on the Chip

This array hybridization technique is useful for making rough quantitative measurements of the current expression levels of a bunch of genes at once, but as originally conceived, it had several limitations. The first was that only known sequences could be used in the array. That limitation has essentially disappeared now that the human genome has been completely sequenced as well as the genomes of many bacteria, yeast, *Caenorhabditis elegans*, *Drosophila*, mouse, and many other species. The second limitation was the number of genes that can be practically handled—both in the process of building the array and in measuring the binding of labeled sequences. That limitation has been vigorously and rather successfully addressed by several different technical innovations. Genomics is all about the automation of biology and conducting experiments in a massively parallel fashion. Robotic fabrication technologies developed for other industrial applications are well suited to the precise and repetitive work of spotting tiny amounts of DNA onto glass slides

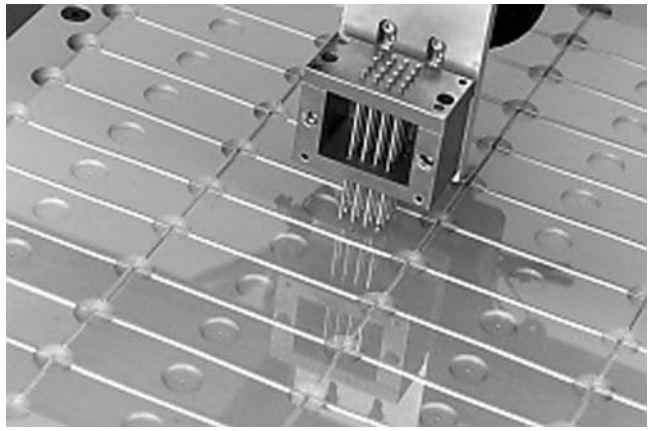


FIGURE 8.3. A microarray-spotting robot making many chips at once.

(Schena et al. 1995) (see Figure 8.3). Automated data collection and analysis software has also been developed to allow high-throughput gene expression measurements with microarrays.

These robotically assembled arrays can contain almost unlimited numbers of very tiny spots of DNA, so they have become known as **microarrays**. Currently there are two fundamentally different microarray technologies, and many variations of these. There are many different robotic systems designed to place spots of cloned cDNAs or PCR amplified fragments onto a substrate—either a glass slide or a nylon membrane. Plasmids carrying clones of cDNAs are generally maintained as frozen stocks, and purified plasmid DNA is stored in microtiter plates. Then each clone is amplified by PCR in order to make DNA fragments for spotting onto chips by the robots. This allows for the expression levels of up to thousands of different genes to be measured simultaneously with high sensitivity. RNA from two different samples—one control and one experimental—are each labeled with different-colored fluorescent dyes. These two labeled RNAs are then combined and hybridized together onto the array. The fluorescence of each label is measured separately,

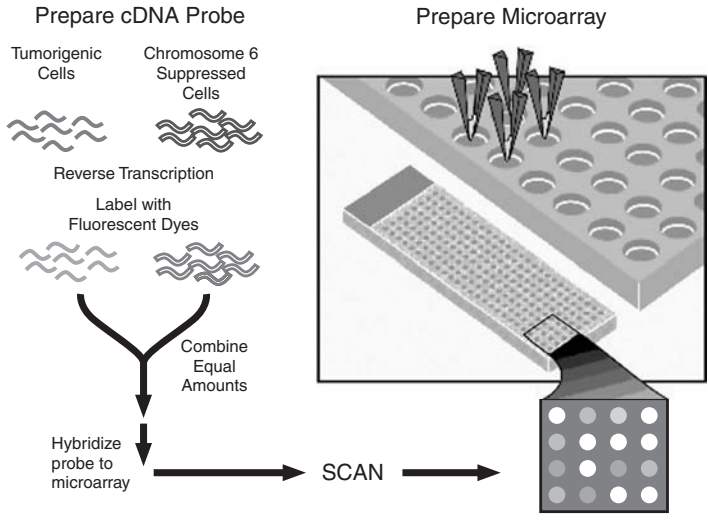


FIGURE 8.4. Schematic of a two-color cDNA microarray experiment.

and the ratio between the two samples is calculated for each spot (see Figure 8.4).

Alternately, it is possible to directly synthesize short DNA oligomers of specific sequence onto known locations on a grid. The Affymetrix Corporation has developed a system called **GeneChip**™ that uses photolithographic technology (similar to that used in the manufacture of computer chips) to simultaneously synthesize hundreds of thousands of different DNA oligomers on a single chip (Lockhart et al. 1996) (see Figure 8.5). The light-activated DNA synthesis is guided by a series of masks with holes that either allow or prevent a given base to be added to the oligonucleotides in each location on the chip.

Each gene is represented by 20 different 25-base-long olionucleotide probes that cover the length of the coding region. In addition, for each probe that matches a region of the gene, a second "mismatch" probe is added that has a single-base mutation in its center (see Figure 8.6). Affymetrix has developed software that calculates the ratio between the binding of labeled RNA to

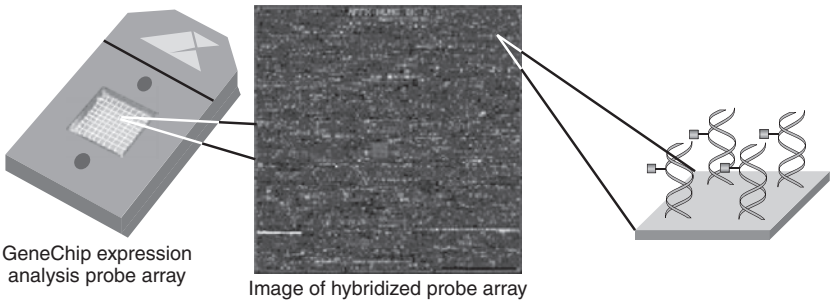


FIGURE 8.5. Affymetrix gene chip with one cell enlarged showing labeled probe bound to oligonucleotide targets.

the perfect-match probe and the mismatch probe and combines this ratio across all 20 probe-pairs to generate a single value to represent the level of mRNA for each gene. In order to make a comparison between two treatments, two RNA samples must be hybridized to separate, identical GeneChips.

An intermediate between the Affymetrix method (using short 25-base targets on the array) and the use of full-length cDNAs as targets, is the use of "long oligos" as targets on the array.

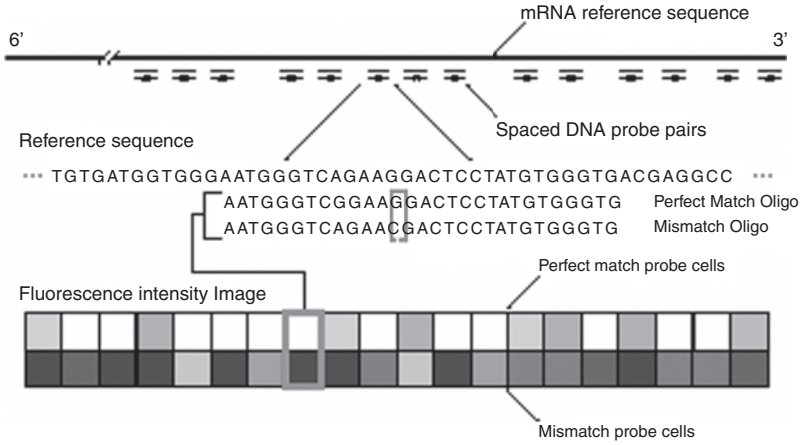


FIGURE 8.6. A schematic diagram of the paired perfect match and mismatch probes used in an Affymetrix GeneChip (Lipshutz et al. 1999).

These are 40–80-base DNA segments from each gene of interest that are synthesized as oligonucleotides and then spotted onto an array. The selection of target sequences is heavily dependent on bioinformatics methods. First, a single representative DNA sequence must be generated for each gene—a synthesis of all available information from cDNA and genomic sequences. Then a single section of the sequence must be chosen that is specific to that gene (i.e., that lacks significant sequence similarity with any other gene) and also must adhere to a variety of primer design criteria such as optimal G+C content and melting temperature and lack of self-complementary sequences.

There is clearly a tradeoff between sensitivity and specificity for targets of varying length. Full-length cDNA targets bind more labeled RNAs of varying lengths and all splice variants so that they can be more sensitive to detect signals in smaller RNA samples and less common RNA species. However, cDNA probes also bind RNAs from homologous genes, such as the members of a multigene family, so they are less specific. The short 25-base oligo probes in Affymetrix GeneChips bind RNA less efficiently, so they are less sensitive, but they can be more specific for one member of a gene family, or even one particular splice variant. Long oligos can combine good sensitivity with the ability to discriminate among similar genes and splice variants.

There are also pragmatic tradeoffs in the choice of cDNA, long oligos, or Affymetrix *in situ* synthesis of short oligos as microarray targets. The Affymetrix method of creating a microarray chip requires a very labor-intensive process of designing custom probes for each gene, and then creating a series of photolithographic masks to allow simultaneous synthesis of all oligos on a chip. It is quite cumbersome and costly to change just a few of the oligos on a chip or to design a custom chip with a new set of genes, so individual scientists rarely create custom Affymetrix GeneChips for experiments that require fewer than thousands of chips. Chips created from spotted cDNAs are much more flexible

in terms of changing a few genes or creating a custom set of genes on a chip for a few experiments. However, when thousands of cDNA probes are being used on a chip, the management of the clones, production of DNA for spotting by PCR reactions, and verification/quality control become challenging for an academic laboratory. Long oligos offer the flexibility to change a few targets in an array or create a custom array, but since there are no clones to grow and no PCR reactions needed to prepare the DNA for spotting onto an array, the potential sources of experimental error in chip construction are less than for cDNA chips. Some vendors of molecular biology reagents are now offering to produce sets of custom oligonucleotides in microtiter format ready for use by chip-spotting robots, or alternately to sell chips spotted with the investigator's choice of long oligos that may be custom-designed from private sequences, chosen from a list of public genes, or chosen in predesigned groups.

Other Types of Arrays

There are many other applications of microarray technology besides the measurement of gene expression. Microarrays can be designed to determine the sequence of a specific fragment of DNA according to its differential binding to targets that contain variant sequences at each position. These sequencing arrays can be designed to provide the entire sequence of a specific fragment of DNA, or simply detect sequence variants at a single known position on a fragment (a SNP). Arrays for the detection of SNPs can also be reversed so that fragments of DNA from specific genes of interest are amplified from a large number of patient samples by PCR and are bound as targets on a chip. Then probes of variant sequences, each carrying a different fluorescent label, are hybridized to the array. The sequence of the target gene in each patient (the SNP genotype) is determined by the color of the probe that binds to each spot on the array.

It is also possible to build arrays of proteins as targets (protein chips)—either by directly attaching purified proteins to a substrate, or by attaching colonies of cells that produce the desired proteins from cloned expression vectors. Then it is possible to use these proteins as targets for labeled probes that might consist of other proteins (to look for protein–protein interactions), or fragments of genomic DNA (to assay DNA-binding properties). Antibodies could also be attached to a chip as targets in an array. Then specific proteins could be quantitated from a labeled mixture according to the amount bound to each type of antibody. The problem with protein arrays is that, unlike mRNA, there is no single universal biochemical approach that can provide global, genomewide profiles for all different types of proteins; they are just too chemically diverse.

Differential Gene Expression

The microarray is a very flexible tool for measuring gene expression in any tissue sample that can be isolated for RNA extraction. The primary research application of this technology has been the study of differential gene expression as the result of development, disease, or the response to chemical agents such as drug treatments. A set of such measurements of gene expression levels across many (or all) genes for a set of treatments is a **gene expression profile** that provides important functional information about many genes. From a broad collection of such profiles created for many different tissue types under many different experimental conditions, an understanding of **gene regulatory networks** will emerge.

It is important to keep in mind that microarray technology measures the *relative* changes in the expression of many individual genes from one treatment to another. It does not measure *absolute* quantities of RNA from each gene, nor does it allow for comparisons of absolute amounts of gene expression from one

gene to another in a single sample. There are many unknown factors that may influence the relative efficiencies of a highly diverse set of molecules in solution (the labeled mRNA from the samples) in binding to a somewhat diverse population of DNA molecules attached to the spots in the array.

Simple microarray experiments aim to identify differences in gene expression (amounts of mRNA from different genes) across a single treatment or tissue type: cancerous versus normal, drug-treated versus untreated, and so on. Data analysis for this type of experiment usually amounts to looking for genes with significant changes in expression levels between the two samples (red vs. green or >twofold change) and making sure that this change is reproducible when the experiment is repeated (see Figure 8.7).

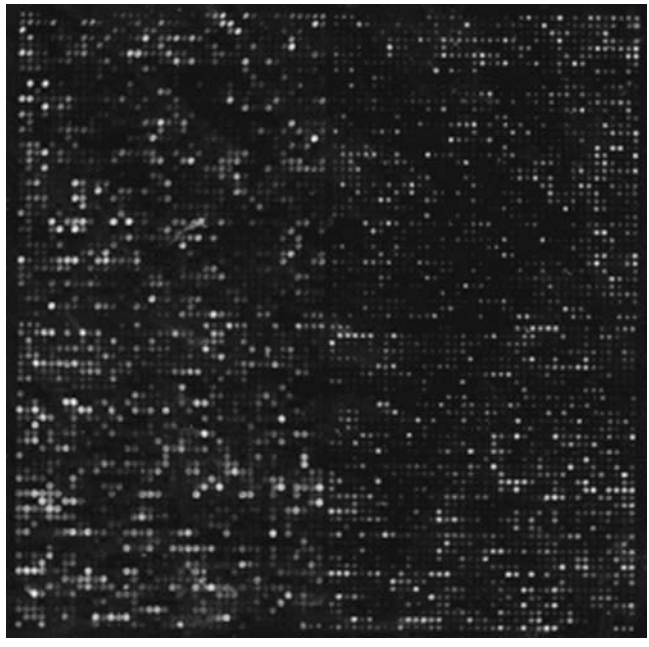


FIGURE 8.7. An example of a spotted cDNA array hybridized with a mixture of two probes with different fluorescent labels and visualized as a red-green false-color image. (See insert for color representation.)

More complex experiments might characterize a time course or a set of related tissues such as subtypes of a particular form of cancer such as leukemia. These more complex datasets contain patterns of expression that are unique for each gene. The challenge is to group the genes into clusters on the basis of similar patterns of changes of the gene expression values across the various samples (see Figure 8.8). Many different clustering techniques have been applied to microarray datasets, but no single analysis method has yet emerged as the best. There is also a challenge in weighing the importance of relative changes in RNA amount (i.e., up threefold from sample A vs. sample C) and absolute changes

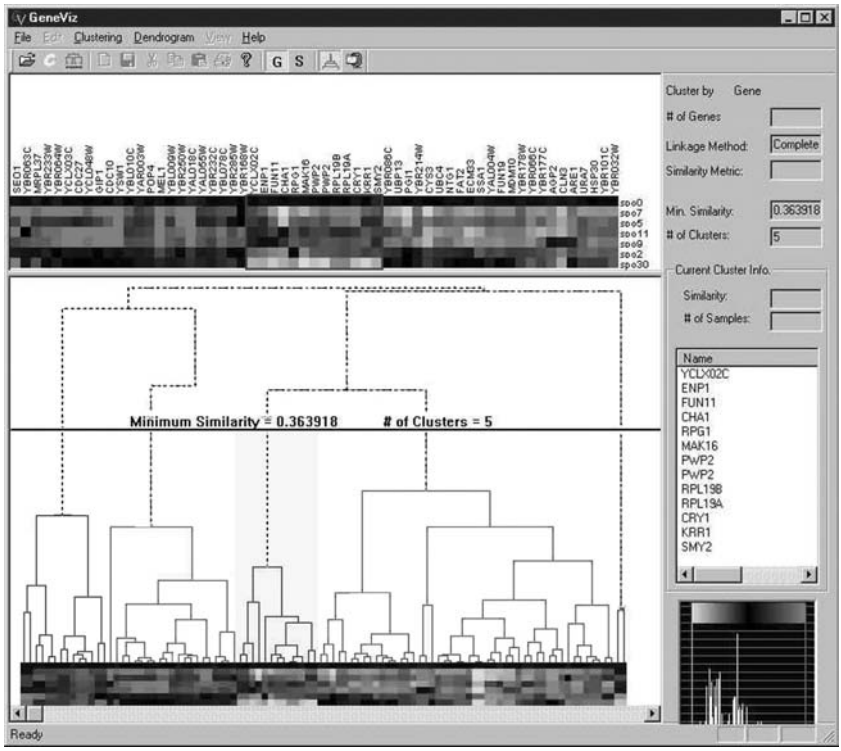


FIGURE 8.8. Clusters of genes that are expressed similarly over different experimental treatments (Seo and Lee 2001).

(i.e., an increase in fluorescence from 1000 to 1500 vs. from 10 to 200) as well as the need to incorporate statistical tests to ensure the validity of these changes.

Microarray measurements are especially useful for identifying correlations between the expression of various genes—gene expression clusters. These patterns of coregulation correspond to metabolic pathways and coordinated cellular responses to developmental and environmental cues. It would be expected that all of the genes in a metabolic pathway, such as synthesis of the amino acid tyrosine, would be turned on and off together. In contrast, the members of a protein family, such as kinases, are involved in many different metabolic pathways that are subject to different regulatory controls. A common characteristic of microarray experiments is that clusters of coregulated genes with previously described functions are found that confirm some hypotheses about dynamic cellular processes, but other clusters are observed that cannot be explained by current biological theories, and thus serve to stimulate fresh thought and new insights into fundamental biological processes.

In each set of coregulated genes identified in a microarray experiment, there are likely to be some genes that have been previously studied and assigned a function as well as some previously unknown genes. In addition, some genes with known functions will be regulated in unexpected ways. By collecting information about differential regulation under a variety of conditions and coregulation with other genes, a detailed picture can be developed for the expression and regulation of every gene.

At the present time, microarray technology is running far ahead of our knowledge of gene functions. It is possible to create arrays on the basis of cDNA sequences or predicted genes from genomic sequence, but after the experiments are completed and clusters of genes with interesting changes in expression levels are found, nothing useful may be discovered about these genes in

public databases. In fact, this is an iterative process. The results of microarray experiments are themselves useful bits of annotation that should be captured in genome databases—tissue-specific, disease-related, or drug-responsive expression is an important aspect of gene function. Then, as new genes are found to be coregulated in common processes, new functional motifs can be defined. Then new drugs may be found that interact with the proteins encoded by these genes (or that modify gene expression), closing the circle from clinical phenotype to molecular genetic analysis and back to clinical therapy.

Classification by Gene Expression

It is also possible to use the coordinate expression of a particular set of genes as a marker for a cellular process or a disease state. For example, various types of cancer are usually diagnosed by a microscopic examination of a tissue biopsy by an expert pathologist (histology), but these diagnoses are not always 100% accurate. This is extremely important for the patient because histologically similar types of cancer with different cellular and genetic origins often respond differently to treatment and have dramatically different prognoses. It is possible to assay samples of these different types of cancer tissues and identify genes that consistently manifest levels that differ between the cancer types. A set of these diagnostic genes on a microarray chip can then be used to categorize an unknown tissue sample with extremely good reliability. This is known as a **class prediction**. The most meaningful set of genes for characterizing a sample into a group may be much smaller than the set of genes that show differential expression. Ideally, genes used for class prediction should consistently differ in expression between the two classes and have low variance within a class. Interestingly, genes that serve to reliably classify samples are often useful as drug targets or provide important insights into the cellular mechanisms of disease.

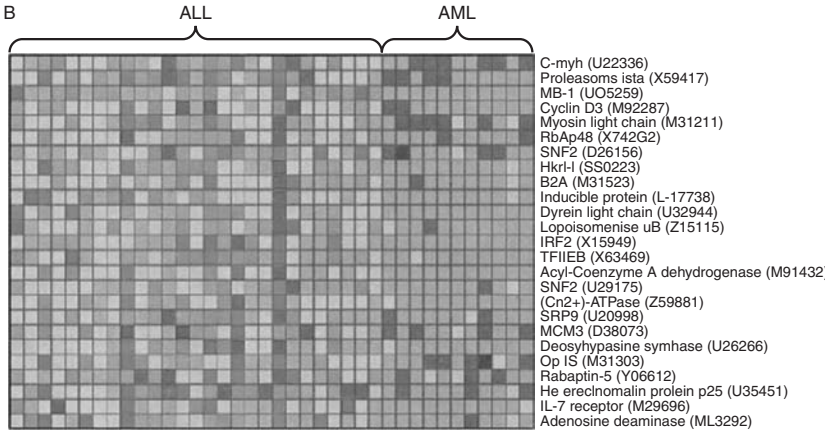


FIGURE 8.9. Genes showing differential expression between ALL and AML (Golub et al. 1999).

Golub et al. (1999) developed a microarray class prediction method to distinguish between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). They used 50 genes (out of a total of 6817 on an Affymetrix GeneChip) to reliably classify a set of unknown samples (see Figure 8.9). One sample diagnosed by classical methods as "atypical AML" did not match the class prediction for either AML or ALL, but an examination of some highly induced genes suggested a muscle origin—which was confirmed as rhabdomyosarcoma by cytogenetic analysis.

For a number of diseases, such as prostate cancer, there is currently no reliable diagnostic standard. A microarray-based assay that could clearly distinguish a prostate cancer from other types of variation in prostate size and production of prostate-specific antigen (PSA) caused by ordinary inflammation or hyperplasia would both save lives and prevent a large number of unnecessary surgeries and aggressive radiation and chemotherapies. A further characterization of tumors as to their aggressiveness and responsiveness to various drugs would revolutionize prostate cancer therapy. A good diagnostic would also

be an important asset in clinical drug trials—limiting trials to patients with aggressive tumors. A change in gene expression profile might also serve as an early indicator of treatment success.

Microarray experiments sometimes reveal additional subcategories of cancers that cannot be distinguished histologically. These subcategories may correlate with different responses to drugs or other treatments, so the more precise diagnosis possible with a microarray is of direct benefit to the patient. Alizadeh et al. (2000) used microarrays to study diffuse large B-cell lymphoma. By clustering gene expression results, they found that tissue samples fell into two classes that represented different stages of B-cell differentiation. These two groups also correlated with patient survival rates.

Once sets of genes have been well defined and verified to predict types of cancer cells or for other diagnostic purposes, the technology could be broadly commercialized. GeneChips are an expensive research tool right now, but they could be mass-produced cheaply. It is important to understand the difference between a gene-expression-based diagnostic test and a classic genetic test that looks at a patient's genomic DNA. Genetic tests measure genomic DNA, which is unchanging and identical in every cell of the body. Gene expression tests measure the current activity in a sample of cells—what proteins are currently being made—and can be used to differentiate cell types or to measure the health of a cell. Gene expression tests are very sensitive to differences in environment, the patient's metabolic status, and the manner in which the sample is collected and processed.

Microarrays are particularly useful in basic research and the early stages of drug development. By studying what genes are induced and repressed in diseased versus normal tissue, researchers can infer key functions and identify potential drug targets. The up- or downregulation of a gene may indicate that it is a cause or a result of disease processes, but in either case, a drug

that returns that gene to normal levels may provide a beneficial therapy.

Error and Reliability

Before microarray technology can be used for routine medical diagnostics, a variety of issues related to sampling accuracy must be resolved. Real patient tissues such as tumor biopsies contain a mix of normal and diseased cell types, including nerve cells, blood cells, immune cells, and vascular and connective tissues. Most patients have multiple tumors with different levels of aggressiveness—and different gene expression profiles. Even cells or tissues that appear histologically normal may be in early stages of developing cancer, or their gene expression may be influenced by nearby cancerous cells. Microarray methods must be robust enough to reliably detect key diagnostic markers in mixed or impure samples.

Microarray experiments produce a large quantity of numerical data that trace the output of fluorescent scanning of the labeled RNA bound to the cells in the array. However, the readout of the fluorescent sensor is an indirect measurement of the amount of each gene's mRNA in the corresponding sample. There are many possible sources of measurement error, including RNA extraction and labeling, hybridization, irregularities in the scanning process, and image processing (finding the boundaries of each spot and integrating its total fluorescent signal, subtracting a background value). Without dwelling excessively on the technicalities of these issues—which are shared by many other data acquisition technologies—it is important to realize that microarray data require some type of standard error calculation. The technique relies heavily on the ratio of gene expression between an experimental and control treatments, but these ratios can be misleading for spots with low fluorescent intensities (i.e., for genes that are expressed at low levels in the sampled cells).

The only way to estimate the reliability of microarray measurements is through the use of replicates. However, Affymetrix GeneChips are very expensive, and in many studies with pathology samples or microdissected tissues, it is extremely difficult to obtain enough material for a statistically valid number of replicates. In addition, it will require greater sophistication in the software used to analyze microarrays in order to accommodate reliability and standard error measurements for hundreds of thousands of data points.

Affymetrix GeneChip experiments do contain some internal validation in the ratio of signal in perfect-match to mismatch probes, but little has been done to utilize this information for improved measurements of reliability for the expression levels of each gene. Similarly, the fluorescent image of each spot on a cDNA array contains additional information about the shape and uniformity of the signal within that spot, but again little has been done to generate a reliability value for each gene from this information.

Another issue in the interpretation of microarray data is the relationship between the probe on the array and the actual mRNAs in the sample. Our current knowledge of human genes is incomplete. Many more cDNAs (and ESTs) exist in the databases than the number of genes predicted in the genome, so if a set of microarray probes are made from these database cDNAs, there is incomplete correspondence with well-defined genes. Whether the probe is a cloned cDNA or oligonucleotides, there are many possible mRNAs that could hybridize. There are sets of closely related genes that share long regions of nearly perfect sequence homology that can produce mRNAs that will cross-hybridize with the probes. These related genes may undergo very different regulation under tissue-specific, developmental, disease, drug response, or other specific conditions. Also, individual genes may undergo many different forms of alternative splicing, which may lead to different mRNAs. These

alternate transcripts may fail to bind to a probe in an array, or bind with different properties—particularly to sets of oligonucleotide probes. As genome information accumulates, microarrays can be designed to compensate for this complexity of mRNA populations—perhaps even to quantitate amounts of various alternately spliced transcripts. However, for now, these must be seen as a source of error and confusion.

Different samples, or replicate samples measured on different days, will yield different values in a microarray experiment. Some of this variation is simply due to systematic changes in experimental conditions—a slightly more efficient buffer or a longer hybridization time, different room temperature, or other variations. These systematic variations can affect all of the values measured on a chip in the same way—everything is brighter or darker. This can be corrected by normalization of the data—scaling the data from each chip to a common value. It might seem logical to use the values of some well-established genes that are known to maintain a steady expression level across all experimental conditions, but after much controversy, no such "universally constant" genes have been proved to exist. Alternately, a DNA sequence that does not correspond to any mRNA in the sample may be included in the array, and a known amount of a matching sequence can be "spiked" into the sample before it is labeled. This creates a positive control that can be used to normalize the intensities of the other spots measured on the array.

Evolutionary Perspectives

Changes in gene expression may be a more subtle and flexible form of evolutionary change than the mutations in protein-coding regions that we have become familiar with in medical genetics. Humans and mice have approximately 92% identical protein sequences; humans and chimpanzees are 99% identical. Data from microarray and other gene expression studies are now

emerging that suggest that there are much larger differences in the expression of some important genes than in their sequences. In an experiment by Svante Pääbo (Max Planck Institute, Germany), differences in relative expression levels of 20,000 genes in humans versus chimpanzees were much more pronounced in the brain than in blood or liver (Pääbo et al. 1995). Intuitively, it seems clear that one can build a larger brain or a smaller brain from the same basic set of molecular components by following a slightly different developmental program, which can be modified by changes in gene expression. Its not what genes you have, but how you use them! Ferea and Brown (1999) state:

> Much of the information encoded in the genome is devoted not to specifying the structure of the protein or RNA that the gene encodes, but rather to controlling precisely in which cells, under what conditions, and in what amounts the gene product is made. Differences in the program of gene expression as opposed to variation in the encoded products may underlie much of the phenotypic variation within and between species.

Differences in gene expression may be a common source of evolutionary adaptation in response to selective pressures such as pathogens. Microarrays are better suited to detecting and characterizing such variation than are tools such as Northern blot or quantitative PCR, which are limited to one or a few genes at a time. In many cases, changes in expression of genes that are not part of an investigators initial hypothesis may turn out to be important in a given biological system.

References

Alizadeh AA, Eisen MB, Davis RE, et al. 2000. *Nature* **403**:503–511.

DeRisi JL, Iyer VR, Brown PO. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**:680–686.

Ferea TL, Brown PO. 1999. Observing the living senome. *Curr Opin Genet Dev* **9**:715–722.

Golub TR, Slonim DK, Tamayo P et al. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**:531–537.

Lipshutz RJ, Fodor SPA, Gingeras TR, Lockhart DJ. 1999. High density synthetic oligonucleotide arrays. *Nat Genet Suppl* **21**:20–24.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**:1675–1680.

Pääbo S, Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**:467–470.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**:467–470.

Seo J, Lee B. 2001. *Dynamic Visualization of Gene Expression Profile*. CMSC 838b Class Project Report, Department of Computer Science, University of Maryland (available online at `http://www.cs.umd.edu/class/spring2001/cmsc838b/Project/Lee_Seo`).

Shalon D, Smith SJ, Brown PO. 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* **6**:639–645.

CHAPTER

9

Analysis of Microarray Data

The technology of gene expression microarrays has become widely available. As an indication of this surge in popularity, the National Library of Medicine (PubMed) reports a total of 18549 scientific and medical journal articles with the word "microarray" in the title, abstract, or keywords as of early 2007, and 5319 articles published in 2006 alone. This is up from just 419 "microarray" articles published in 2000 (Figure 9.1). As microarrays have become more popular, clinical and basic science investigators have come to rely more on commercially manufactured arrays and scanners rather than home-spotted chips. This saves a tremendous amount of effort spent on trivial issues such as coating of glass slides, salt and soap concentrations of buffers, humidity control, pin size, and general manufacturing quality control of home-made arrays. A physician or clinical investigator can develop an idea for an experiment, harvest some tissue from a few patients, extract RNA with a simple kit, and send the samples off to be processed on commercial microarrays as a contract service from core genomics laboratories at many universities and

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

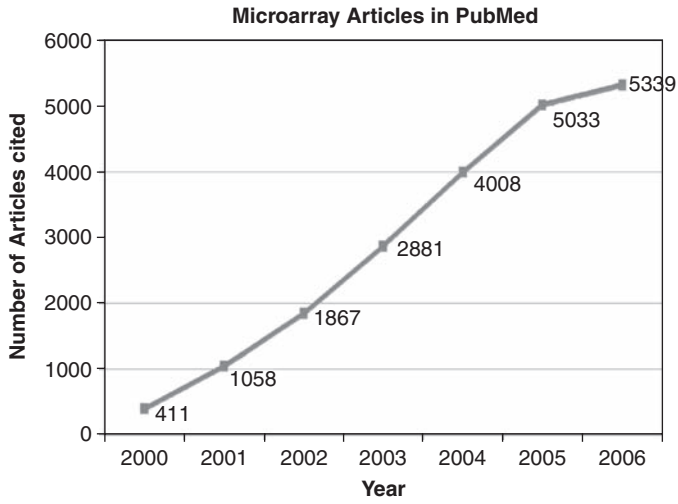


FIGURE 9.1. Number of articles indexed each year in PubMed from years 2000 to 2006 with the term "microarray" in title, abstract, or keywords.

private biotechnology companies. However, the real challenge comes after the data are returned from the lab. A genomewide expression microarray might contain data on tens of thousands of genes (or putative genes). It is not possible to analyze these data using traditional tools like Microsoft Word® and Excel®. A wide variety of freeware and commercial software exist for microarray data analysis, but each tool has a steep learning curve.

Experimental Design

Before embarking on a microarray project, it is useful to review the basic concepts of experimental design and the key steps in data analysis. Microarray projects are generally designed with one (or more) of three basic goals in mind. The most common type of experiments are simply large-scale searches for genes that are regulated in a specific disease, drug reaction, or developmental process. The experiment involves a comparison between

some experimental and some control samples (i.e., two classes). An alternate approach seeks to identify groups of genes that are coregulated by a process, such as a developmental or stress response time course. This implies a clustering of genes across multiple timepoints rather than a simple search for differences between two sets of samples. A third approach is to identify gene expression patterns that can be used as a diagnostic test to differentiate disease from normal or among different subclasses of disease.

Regardless of the research goal, the basic principles of statistics must be respected. Gene expression microarrays measure quantities of mRNAs in cells—an inherently unstable molecule that is subject to rapid changes due to transcriptional regulation by multiple factors, posttranscriptional regulation, and variable rates of degradation by highly active RNAse enzymes. The reality is that there is no typical level of variability that can be predicted for a microarray experiment. Different types of biological material produce very different levels of mRNA fluctuations—tissue cultures are more uniform, while surgical specimens from human patients are more variable. In human subjects, a huge number of variables (that are unrelated to the specific experimental parameters) can affect gene expression at the cellular level, including age, gender, cardiovascular fitness, emotional stress, and what they had for breakfast that morning. In order to obtain a reliable measurement of differences in expression levels for a particular gene between two or more treatments (experimental and control), the measurement must be repeated several times on replicate samples. Even on one set of microarray samples, the variability of some genes may be 10 times greater than that of other genes on the same arrays.

The most frequent question asked by investigators starting to work with microarrays is: "How many replicates do I need for this experiment?" There is no simple answer to that question. A statistician would say, "It depends on the variability [standard

deviation] of your measurements." This is, of course, true, but useless to ask before the experiment is conducted, since the variability of the measurements cannot be predicted in advance. Investigators generally wish to keep the number of replicates down, since both acquiring samples and hybridizing them to microarray chips is expensive. The simplest answer to the "how many replicates" question is that at a minimum, there must be enough replicates to calculate a reliable standard deviation (somewhere between three and five). More replicates will allow smaller differences to be detected between treatments. However, large numbers of replicates (dozens to hundreds) will increase the number of false positives returned by most statistical techniques.

Another important point to keep in mind, before staring at very large spreadsheets of numbers, is that the measurements in gene expression microarray experiments are always comparisons. Microarrays cannot accurately assess the relative amounts of mRNAs for two different genes in a single sample. What they can do is estimate whether each gene increases or decreases its expression between two different samples. The physical chemistry of RNA–DNA hybridization is extremely complex. It is highly dependent on the number of G+C versus A+T base pairs in the probe sequences; internal structure of the probes such as self-complementary sequences that form hairpins; and cross-hybridization of the probes to multiple RNAs due to gene duplication, multigene families, repetitive sequences, and random sequence similarity. Therefore, it is not possible to directly compare the signals from two different probes on the same RNA sample. It is necessary to calculate the ratio of signals for each probe between two different samples.

Early microarray experiments on small spotted arrays were done using two different RNA samples (experimental and control) labeled in two colors, then hybridized together to the array. The gene expression was calculated as the ratio of the

signal of the two colors on each spot. High-density microarrays such as Affymetrix GeneChips™ hybridize each array with only a single labeled RNA sample. Therefore it is necessary to compare two different arrays hybridized with different types of samples (experimental vs. control) in order to calculate this basic expression ratio for each gene.

DATA ANALYSIS WORKFLOW

Once the experiment is designed, the samples collected, RNA extracted, and the microarrays have by hybridized and scanned, the data are returned from the genomics lab to the investigator. At this point, confusion generally ensues. Rather than starting by looking for software that can open these large data files, it is helpful to plan out the data analysis workflow. Very briefly, these steps are: image analysis, normalization, filtering, statistical test or clustering, and functional analysis of a set of selected genes (Figure 9.2).

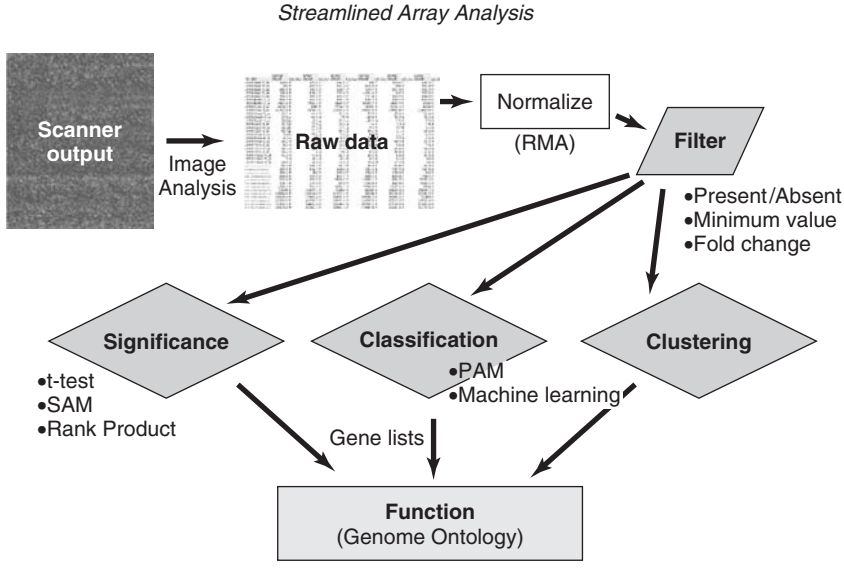


FIGURE 9.2. Microarray data analysis workflow.

Image Analysis

The most basic step in data analysis is the image analysis of the fluorescent signal from the scanned array to find the spots that correspond to probes, then quantify signals from each probe, perform background subtraction, and analyze the combination of signals from multiple probes to give signals for each gene. This is generally handled by software provided by the manufacturer of the array or of the fluorescent scanner. As the density of spots on commercially manufactured arrays has increased, the image analysis process has become more automated. This is a good thing—not necessarily because an automated image analysis program is more accurate than a careful investigator, but because an automated image analysis program is more likely to give similar results on the same scanned image when it is run in two different labs than with two skilled scientists using manual image analysis methods.

Normalization

Once the primary image analysis is completed, the raw data consist of a large spreadsheet of genes (or probes) with a signal intensity and some associated quality scores for each sample. For an Affymetrix chip, this raw datafile might consist of a 10-megabyte (MB) file with over 50,000 rows of data. The next step is to normalize the data across all of the samples. Normalization can be mathematically complex, but in the simplest terms, it sets the average brightness of each chip to be equal and corrects for systematic errors (i.e., the upper left corners of all scans are brighter). Affymetrix chips use multiple probes for the same gene, so the signals from these different probes need to be combined into a single signal for the gene in a way that minimizes the noise caused by variation in binding efficiency among the different probes. For Affymetrix chips, the robust multichip average (RMA) (Irizarry

et al. 2003a) has proved to be the best normalization method. The RMA method is built into a number of different software tools including the free BioConductor package and RMAExpress, a free Windows program (Bolstad 2007), as well as commercial software such as GeneSpring (Agilent) and ArrayAssist (Stratagene Inc.). The Affymetrix software includes a modified form of RMA in its probe logarithmic intensity error (PLIER) probe summarization method.

For two-color spotted arrays, normalization is also needed to correct for dye bias (one dye may give overall stronger fluorescence than the other, or individual probes may give stronger signals with one color of dye than the other) and print tip effects. The best normalization method is generally considered to be locally-weighted regression (LOESS) (Cleveland 1979; Yang et al. 2001) or locally weighted scatterplot smoothing. The LOESS method is available in the free BioConductor tools built in the R statistical language. Carlo Colantuoni, George W. Henry, and Jonathan Pevsner (2002) have developed an easy-to-use Web-based LOESS microarray normalization tool called SNOMAD, which is freely available on the PevsnerLab Website (Figure 9.3) at the Kennedy Krieger Institute (`http://pevsnerlab.kennedykrieger.org/snomadinput.html`).

FILTERING

After image analysis and normalization, it is standard procedure to trim down the huge number of genes/probes under consideration with some filtering method. From a biological perspective, it is unlikely that any single tissue is expressing as much as half of all the genes in the genome. Also, microarray technology is not capable of making accurate and reproducible measurements of very small amounts of mRNA, such as might be expected from some transcription factors and other regulatory genes. Statistical methods can be fooled by the relatively large "fold change"

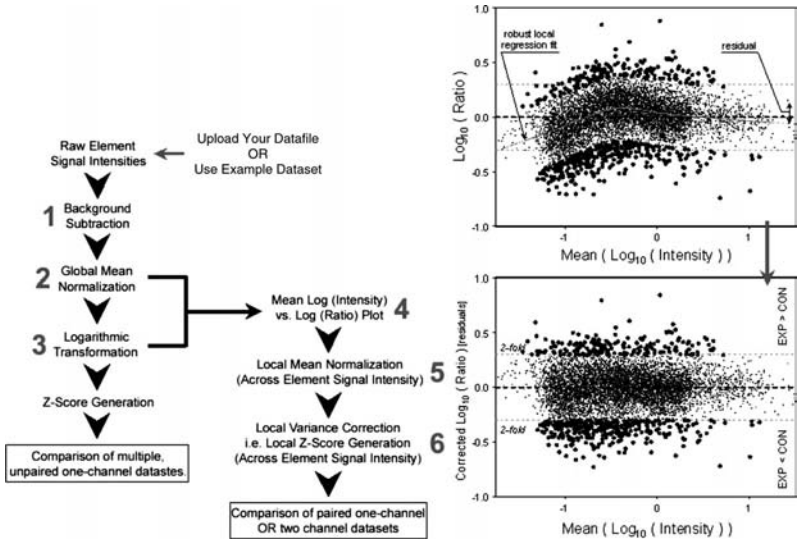


FIGURE 9.3. Normalization of two-color microarray data with the LOESS method using the SNOMAD tool from the Pevesner lab Website.

fluctuations observed from genes measured at very low signal levels. The various Affymetrix probe-level summarization methods have always included a present/absent call for each gene according to the number of probes for a single gene with signal above a threshold, variance of the signal among the probes, and the difference between probes that match the gene (perfect-match probes) and those that contain an intentional mismatched base (mismatch probes). Image analysis of spotted two-color arrays with software such as GenePix (Molecular Devices, Inc.) generates a quality score and a "flag" for poor quality or below background spots. These simple $\pm$ markers can be used to remove from one-tenth to one-fourth of genes with very weak or poor-quality signals from the dataset. Further filtering must rely on essentially arbitrary criteria such as a signal threshold (i.e., signal ≥ 100) or a measure of variance (maximum signal across all chips must be at least 2 times greater than minimum). Failure to impose fairly stringent filtering criteria will inevitably lead to

large number of "false positives" showing up in the final list of regulated genes, but stringent filters will also always remove a few truly interesting genes from the results.

FOLD CHANGE

After filtering, some form of statistical, clustering, or classification analysis is applied to the data. The simplest method to measure differential expression, which was used in many of the earliest microarray experiments, is the **fold change**: the ratio of signal for each gene between treated and control samples. If there are replicates, then for each gene, take the average signal for all replicates of each treatment, and take the ratio of the averages. The weakness of a simple ratio calculated as [average (experimental)/average (control)] is that genes that are upregulated in the experimental samples are given a large positive number for fold change and genes that are downregulated get a fold change value that is a fraction that approaches zero as the differential expression is more extreme. This does not produce a nice graph – since all the downregulated genes are bunched up near the origin. To make the fold change values easier to work with, a $\log_2$ transformation is applied, which has the effect of making the fold change of downregulated genes into negative numbers on the same scale as upregulated genes are positive, and produces a nicely symmetric graph. It also has the advantage of scaling down extremely high fold change values.

Another weakness of using the fold change to identify differentially regulated genes is that it often gives artificially high values for genes with low signals. If the fluorescent signal on the microarray is measured on a scale from 0 to 10,000 with a measurement accuracy of $\pm$ 50, then genes with an average signal of less than 100 are quite likely to show up with a random twofold change from experimental versus control. A real change of a few hundred signal units will show up as a huge fold change

for a low-intensity gene and a tiny change for a highly expressed gene. One interesting approach to this problem is to add a constant value to each signal measurement. For example, if 100 is added to all signals, it will have little effect on the fold change of a gene with an average signal of 3000 in the experimental samples and 1000 in the control (3100/1100 is still a fold change of 3), but it will dramatically reduce the fold change calculated for a gene with an average signal of 150 in the experimental and 50 in the control ($150/50 = 3$, but $250/150 = 1.7$).

For some reason, biologists tend to focus on differentially expressed genes with fold change values of ≥ 2 as being interesting, while statisticians favor the *t*-test, which attempts to determine whether the difference between the means of the experimental and control samples is greater than the standard deviation of the measurements. The standard *t*-test finds significance with a *p* value of 0.05, which implies that a difference of this size (between experimental and control) would only occur by chance 1 time out of 20. If the microarray has tens of thousands of genes, then the *t*-test will find significant differences in many genes due to random fluctuations in the signal measurements. In order to control these false positives, a number of different multiple testing corrections have been used. Unfortunately, none of these methods produces a completely satisfactory result, since microarray data do not match the theoretical assumptions. For example, it is clearly not true that differential expression of each gene is independent of the expression levels of other genes.

A variant of the *t*-test based on permutations was developed for microarry analysis by Tibshirani and coworkers (Tusher et al. 2001) and made freely available (to academic users) as part of the BioCondudtor package and the **significance analysis of microarrays** (SAM) add-in for Microsoft Excel (`http://www.stat.stanford.edu/~tibs/SAM`). SAM is particularly useful to biologists because it provides an intuitive "delta slider" tool as a graphical interface to adjust the

FIGURE 9.4. Delta slider interface for the SAM (significance analysis of microarrays) tool from Tibshirani and coworkers (Tusher et al. 2001).

significance level (the "false discovery rate") so that the user can observe the tradeoff between marking more genes as "significant" and the increase in the number of false positives that are included in that set (Figure 9.4). Most users find a comfort level with the number of false positives set at somewhere between 5% and 20% of the number of "significant" genes. The slider can also be used to demonstrate conclusively that zero significant genes are found unless the number of false positives is allowed to grow too large, or alternately, to set the significance level arbitrarily high to reduce the number of significant genes to a manageable number for functional analysis. The user can also limit the final output to genes that have a fold change greater than a specific amount (such as twofold).

CLASSIFICATION

Classification is based on the idea of "supervised clustering," where the samples are known to represent two or more classes (experimental and control, cancer subtypes, etc.). Rather than ask the simple question "What genes are differentially expressed between these classes?" classification analysis asks "Which genes can most reliably identify samples as members of these classes?"

There is often an overlap between the lists of genes produced by classification and differential expression methods, but the objectives are different. Classification can produce a gene expression profile or a scoring function that can later be applied as a diagnostic test to microarray data from unknown samples and reliably assign them to a class. Classification methods often focus on the smallest possible set of genes that can reliably separate the samples, while differential expression methods often try to capture the largest number of genes with significant changes in expression levels.

A simple classification tool called **prediction analysis for microarrays** (PAM) has been developed as an R module in BioConductor and as a Microsoft Excel® add-in by Tibshirani et al. (2002). The PAM method uses a statistical approach called the **nearest shrunken centroid**, which is the average gene expression for each gene in each class divided by the within-class (intraclass) standard deviation for that gene. This method has the advantage that it can make the classifier more accurate by reducing the effect of noisy genes, and it does automatic gene selection. In particular, if a gene is shrunk to zero for all classes (where differences in expression are less than the standard deviation), then it is eliminated from the prediction rule. Alternatively, a gene may be set to zero for all classes except one, which indicates that high or low expression for that gene characterizes that class. Overall, the PAM method defines the subset of genes that best characterize each class—which is the goal in developing practical gene-expression based diagnostic tests.

CLUSTERING

Clustering is a data analysis tool that can identify patterns of gene expression that can be used to identify groups of genes that are coregulated across several experimental conditions, or to identify groups of samples that show similar gene expression

profiles. Clustering is generally used when there are more than two experimental conditions, since there is little to be learned from a cluster of two treatments that cannot be found with a simple fold change or *t*-test approach. If a number of different treatments are going to be combined in one data analysis, then there must be some common way to calculate expression ratios for all treatments. Generally the experiment must be designed with a single common control or baseline treatment—such as a time-zero timepoint or an untreated sample type. Then all of the different treatments are divided by this baseline in order to produce ratio measurements.

In collaboration with Pat Brown, Mike Eisen developed the microarray data analysis tools Cluster and TreeView, which create a two-dimensional hierarchical clustering of array data and impose a simple red-green color scale that makes it easier for (non-color-blind) investigators to visually identify patterns (Eisen et al. 1998). These tools are freely available for noncommercial users from Eisen's Website (`http://rana.lbl.gov/EisenSoftware.htm`). The basic approach starts with a matrix of data where columns represent samples and rows represent genes (the values are usually expressed as normalized $\log_2$ ratios). The clustering method then calculates the average difference between each pair of columns across all of the genes (rows), and moves the columns so that the columns (samples) with the smallest average distances are next to each other. This process is repeated with the rows—moving genes (rows) that have the smallest average differences across all columns next to each other. The individual cells of the matrix are colored on a red-green scale so that upregulated genes are bright red, downregulated genes are bright green, and genes with no expression change are black. The resulting two-way cluster diagram has come to be known as a **microarray heatmap** or **Eisengram**, and it has been used in many hundreds of published articles (Figure 9.5).

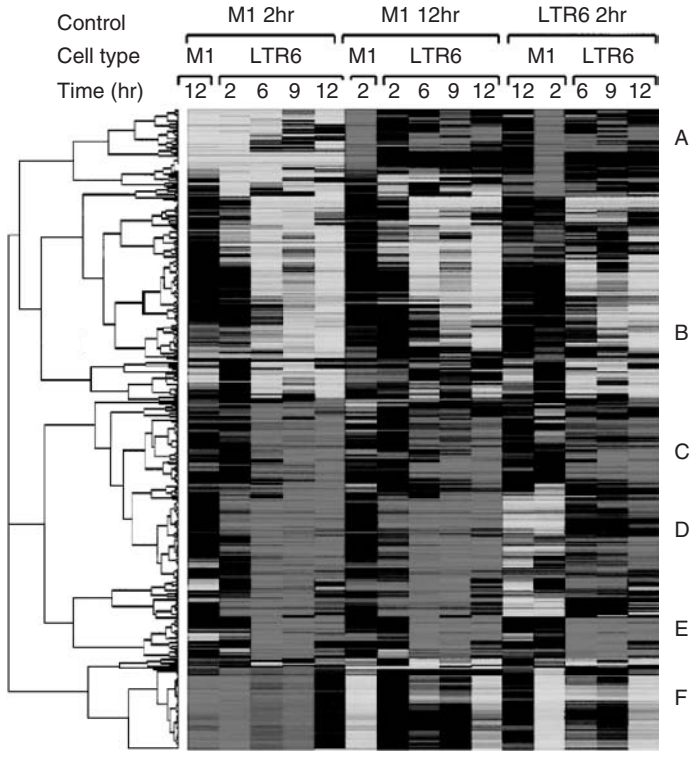


FIGURE 9.5. A two-way clustered heatmap of microarray data as produced by Cluster and TreeView software by Mike Eisen. (See insert for color representation.)

A great deal can be learned from the study of a properly clustered set of gene expression data. On first glance, the replicate samples (columns) should be grouped together. If replicates don't group, then random variation (or systemic error) is a larger overall factor in the gene expression measurements than are the experimental factors separating the different treatments. Next, there are generally some blocks of color—groups of genes that show similar expression patterns across multiple samples. These block patterns are encouraging, but they do not constitute statistically significant proof of coregulation among the genes. The two-way clustering method can produce similar blocks from

random data. Another feature of the cluster is the hierarchical branch diagram (tree) that connects the rows (genes). If a vertical line is drawn through this tree, the genes will be separated into groups or clusters. Some tree diagrams have obvious breakpoints, while others show a range of branching relationships that makes it difficult to determine exactly where the clusters should be broken or how many clusters are present. A variety of other clustering methods, such as K-means and self-organizing maps, are based on the idea of building these clusters of coregulated genes in a statistically robust manner.

Functional Analysis

Once a group of genes has been identified as "interesting" in a gene expression microarray experiment (differentially regulated, part of a classifier, or members of some form of cluster), the next step in the analysis is to look at the biological functions of the genes. Generally the goal at this point is to find common pathways or themes among the genes. If a particular biological process is being induced or repressed as a result of some experimental treatment, then many of the genes that are detected with substantial differences in their expression patterns should have functions related to that process. In addition, some genes with previously unknown function may be associated with the pathways and functional groupings delineated by the better-known members of the coregulated gene sets.

Prior to the era of complete genome sequencing, the description of gene functions in the scientific literature was a mess. Each community of researchers who worked on a particular organism, disease, or metabolic/developmental function had their own vocabulary. The same gene might be known by many different names in different organisms or different research contexts. Even when biologists work in the same field, different journal articles and databases may use different terminology for the same

concepts, such as "translation" versus "protein synthesis." This creates difficulties for manual literature and database searches and makes automated data integration virtually impossible. The Gene Ontology (GO) project (`http://www.geneontology.org`) is a collaborative effort by biologists to address the need for consistent descriptions of gene products. The GO project began in 1998 as a collaboration between three model organism databases, FlyBase (*Drosophila*), the *Saccharomyces* (yeast) Genome Database (SGD), and the Mouse Genome Database (MGD), to develop a single set of consistent terminology to describe the function of genes and gene products. Building on this base of common terminology associated with well-described model organisms with fully sequenced genomes, the genes and gene products of other organisms have been mapped to the same set of functional terms.

The GO project has developed three structured controlled vocabularies (**ontologies**) that describe gene products in terms of biological processes, cellular components, and molecular functions across all species. These terms are then applied (annotated) to specific genes in each species. Within each of the three of ontologies, the terms are organized in a hierarchical structure so that queries can be directed at various levels of specificity. Then each gene from a list of "interesting" genes produced by a microarray experiment can be annotated with functional terms from the Gene Ontology on the basis of its similarity to genes from the model organisms. In a well-designed experiment, the list of differentially regulated genes produced by an analysis of microarray data should include many genes that share similar functions, especially in the "biological process" ontology, which has categories such as signal transduction, DNA repair, and cell cycle regulation. It is possible to use a simple statistical test to determine the extent to which this clustering of genes into functional categories is actually due to true shared biological functions, compared to the clustering of functions that would be found for a random set of genes taken from the same array.

A number of free Web-based tools are available to provide the functional assignment of GO terms to genes from microarrays and to do the significance testing for shared functions. The National Institute of Allergy and Infectious Disease (NIAID) provides the DAVID Website with functional annotation of gene lists (Sherman et al. 2007):

> The Database for Annotation, Visualization and Integrated Discovery (DAVID) provides a comprehensive set of functional annotation tools to understand biological meaning behind large list of genes (`http://david.abcc.ncifcrf.gov`). DAVID tools identify statistically significant enriched biological themes, particularly GO terms and discover enriched functional-related gene groups. The DAVID Functional Annotation Tool, can display genes from a user's list on pathway maps from KEGG and BioCarta to facilitate biological interpretation in a network context.

The following is an excerpt from Al-Shahrour et al. (2004):

> FatiGO (`http://fatigo.bioinfo.cipf.es`) takes a list of genes from a microarray (or other genomics experiment) and converts them into a list of GO terms using the corresponding gene-GO association table. Then a Fisher's exact test (with multiple testing correction) for 2×2 contingency tables is used to check for significant overrepresentation of GO terms with respect to the other genes in the experiment (the entire array).

Berriz et al. (2003) state:

> FuncAssociate (`http://llama.med.harvard.edu/cgi/func/funcassociate`) is a web-based tool that accepts as input a list of genes, and returns a list of GO attributes that are over-represented among the genes in the input list. Statistical significance is calculated using a Monte Carlo simulation approach corrected for multiple hypotheses testing.

Other free GO annotation tools are designed as standalone applications or as modules for various microarray/genomics analysis platforms (Maere et al. 2005):

> BiNGO (`http://www.psb.ugent.be/cbd/papers/BiNGO`) is an open-source Java tool to determine which GO categories are statistically

> over-represented in a set of genes. BiNGO is implemented as a plugin for Cytoscape, which is a software platform for data integration and visualization of molecular interaction networks. BiNGO maps the predominant functional themes of a given gene set on the GO hierarchy, and outputs this mapping as a Cytoscape graph.

Finally, Hosack et al. (2003) state:

> EASE (`http://david.abcc. ncifcrf.gov/ease/ease.jsp`) is a customizable, standalone, Windows© desktop software application that facilitates the biological interpretation of gene lists derived from the results of microarray, proteomic, and SAGE experiments. EASE provides statistical methods for discovering enriched biological themes within gene lists, generates gene annotation tables, and enables automated linking to online analysis tools.

VALIDATION

Gene expression microarrays are used for quantitative measurements of mRNA, a notoriously unstable molecule, subject to regulation by many different cellular systems. High-density arrays contain probes for many thousands of genes, which creates a situation where most statistical tests will produce some false-positive results due to the "multiple testing" effect. Therefore, the results of any microarray experiment need to be validated using some other experimental method.

The most direct method for validation of differential gene expression is to use a Northern blot to estimate the amount of hybridization of a gene-specific labeled probe to RNA from two different samples immobilized on a membrane. The Northern blot uses gel electrophoresis to separate the RNA by size, which confirms the identity of the particular mRNA being bound by the probe. The downside of a Northern blot is that it requires a large amount of very high-quality RNA (typically 10 μg), a separate labeled probe and hybridization reaction for each gene that is being measured, and a couple of days of labwork (Chelly and Kahn

1994). If a microarray experiment produces a list of hundreds of differentially regulated genes, it is not feasible to validate all of them by Northern blot. Quantitative measurements by Northern blot are approximate, since signal is generally measured by densitometry of spots on X-ray film, and comparison between different samples is dependent on loading the same amount of RNA to each well of a gel and the measurement of another "housekeeping" gene as an internal control. The expression of the internal control should be constant across all samples being analyzed.

The most popular method for validation of differential gene expression measurements is quantitative real-time PCR (qRT-PCR). RT-PCR is much more sensitive than a Northern blot, with 100% detection found using just 100 copies of a target gene in a reaction (Lai et al. 2003). Various manufacturers claim accurate RT-PCR results using as little as 100 pg–1 μg of total RNA. RT-PCR uses the enzyme reverse transcriptase to convert mRNA into cDNA, then amplifies the target gene with a pair of primers, just as in a standard PCR reaction. To compare gene expression between different samples, a housekeeping gene is used as an internal control. The use of fluorescently labeled probes allows both the target gene and the internal controls to be coamplified in the same tube (multiplex PCR). At the start of the PCR reaction, primers and reagents are in excess, and amplification proceeds at a constant, exponential rate. However, in the later cycles of the PCR reaction, product renaturation competes with primer binding and the amount of product produced in each cycle is reduced, so that the overall product concentration plateaus. The reaction must be analyzed in the linear range of amplification before the products reach the PCR plateau for either the gene of interest or internal control. Real-time PCR measures the amount of product after each cycle of amplification, giving a wide range of product concentrations where two samples can be compared. Automated RT-PCR equipment calculates a standard curve for the control and copy number for the target gene in each sample.

In addition to greater sensitivity and less hands-on time, RT-PCR can be run in parallel on many genes. The synthesis of RT-PCR primers for multiple genes is relatively inexpensive. Many PCR reactions can be constructed with a single "master mix" that contains all reagents except the target gene primers, then sample RNA and gene-specific primers can be added to individual sets of tubes. It is feasible for a technician to conduct dozens to hundreds of these gene expression assays in a single day. The bottleneck in this operation is that each set of PCR primers for a specific gene must be validated on control RNA to show adequate and specific amplification of the target gene. Several manufacturers of RT-PCR equipment offer validated sets of primers for commonly studied genes, but microarrays contain tens of thousands of genes, any of which may show up as differentially regulated in a specific experiment.

References

Al-Shahrour F, Díaz-Uriarte R, Dopazo J. 2004. FatiGO: A web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics* **20**:578–580.

Berriz GF, King OD, Bryant B, Sander C, Roth FP. 2003. Characterizing gene sets with FuncAssociate. *Bioinformatics* **19**(18):2502–2504.

Bolstad B. 2007. `http://rmaexpress.bmbolstad.com`.

Chelly J, Kahn A. 1994. RT-PCR and mRNA quantitation. In Mullis DB, Ferre F, Gibbs RA (eds), *The Polymerase Chain Reaction*, pp 97–109. Birkhauser, Boston.

Cleveland WS. 1979. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* **74**:829–836.

Colantuoni C, Henry G, Zeger S, Pevsner J. 2002. SNOMAD (Standardization and NOrmalization of MicroArray Data): Web-accessible gene expression data analysis. *Bioinformatics* **18**(11):1540–1541.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **95**(25):14863–14868.

Hosack DA, Dennis G, Sherman BT, Lane C, Lampicki R. 2003. Identifying biological themes within lists of genes with EASE. *Genome Biology* **4**:4.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. 2003a. Summaries of Affymetrix GeneChip® probe level data. *Nucleic Acids Res* **31**(4):e15.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. 2003b. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**(2):249–264.

Lai JP, Yang JH, Douglas SD, Wang X, Riedel E, Ho WZ. 2003. Quantification of CCR5 mRNA in human lymphocytes and macrophages by real-time reverse transcriptase PCR assay. *Clin Diagn Lab Immunol* **10**(6):1123–1128.

Maere S, Heymans K, Kuiper M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. *Bioinformatics* **21**:3448–3449.

Sherman BT, Huang da W, Tan Q, Guo Y, Bour S, Lin O, Stephens R, Baseler MW, Lane HC, Lempicki RA. 2007. DAVID knowledge base: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics* **8**:426.

Tibshirani R, Hastie T, Narasimhan B, Chu G. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA* **99**(10):6567–6572.

Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* **98**:5116–5121.

Yang YH, Dudoit S, Luu P, Speed TP. 2001. Normalization for cDNA microarray data. In Bittner M, Chen Y, Dorsel A, Dougherty ER (eds), *Microarrays: Optical Technologies and Informatics*, Vol 4266. San Jose, CA, pp 141–152.

CHAPTER

10

Pharmacogenomics and Toxicogenomics

Pharmacogenomics

One of the first spinoffs from the Human Genome Project to reach the practicing physician will be genetic tests designed to aid in prescribing drugs. This technology, known as **pharmacogenomics**, promises to be both simple and relatively uncontroversial. **Pharmacogenetics** is the study of how genes affect the way people respond to medicines. All patients want to receive the most effective drugs that will have the fewest side effects, but up to now there was essentially no information to help the physician decide which drug would be best for a specific person. It has been known for some time that certain genetic factors influence the efficacy and side effects of a particular drug in an individual patient, but the physician generally had no way to measure these factors in advance and account for them when writing a prescription. Genomics technology promises to make this information easily accessible.

Pharmacogenomics is generally defined as the use of DNA sequence information to measure and predict the reaction of

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

individuals to drugs. The theoretical basis for this technology is quite straightforward. Many proteins are known to enhance or block the action of specific drugs—through either direct chemical action on the drug molecule (degradation or activation), interaction with a common target molecule (i.e., to block drug binding to a receptor), or by regulation of a metabolic pathway that impacts on drug function. Some genes have also been shown to modulate drug side effects. It is now possible to use SNP markers to identify alleles of these drug interaction genes in test populations, and then screen patients for the markers before prescribing the drug.

Examples of Genetic Traits for Drug Response

There are many examples of drug–gene interactions that have been discovered through the unfortunate experiences of people receiving certain drugs. It has been estimated that over 2 million people are hospitalized each year in the United States because of adverse reactions to drugs that were properly prescribed – but without knowledge of each patient's unique genetic makeup.

In World War II, the US Army discovered that 10% of African-Americans have polymorphic alleles of glucose-6-phosphate dehydrogenase that leads to hemolyitic anemia when they are given the antimalarial drug primaquine. Approximately 0.04% of all people are homozygous for alleles of pseudocholineseterase that are unable to inactivate the anaesthetic succinylcholine, leading to respiratory paralysis.

About 10% of the Caucasian population is homozygous for alleles of the cytochrome P450 gene CYP2D6 that do not metabolize the hypertension drug debrisoquine, which can lead to dangerous vascular hypotension (Kuehl et al. 2001). There are many polymorphic alleles of the *N*-acetlytransferase (NAT2) gene with reduced (or accelerated) ability to inactivate the drug isoniazid. Some individuals developed peripheral neuropathy in reaction

to this drug. Some alleles of the NAT2 gene are also associated with susceptibility to various forms of cancer. This is an important point to which we will return in the discussion of ethical implications of genomic testing—a test for one trait may reveal information about other genetic factors, through either pleiotropic effects of a single gene or alleles of linked genes.

In other cases, drugs are less effective for people with a specific genetic trait. Patients homozygous for an allele with a deletion in intron 16 of the gene for angiotensin-converting enzyme (ACE) showed no benefit from the hypertension drug enalapril while other patients did benefit.

The Use of SNP Markers

In all of these examples, the drug response phenotype is associated with a specific allele of a single-gene. Once that gene is identified and the sequence variation is identified, it is possible to construct a single-gene test, or in some cases a biochemical assay for the variant protein. These are approaches that were available to twentieth-century genetics. In the era of medical genomics, the identification of pharmacogenomic traits can proceed much more rapidly, and multigene effects can be identified almost as easily as single-gene traits. It is possible to use a panel of thousands of SNP markers that cover the entire genome to screen groups of patients receiving a specific drug, and then correlate good and poor drug efficacy and the occurrence of specific side effects with individual SNP markers (or groups of markers). These linked markers can be used directly to predict drug response traits, or they can be used as landmarks on the genome to initiate a second round of more precise screening with additional sets of SNP markers that are focused on those specific chromosomal regions. Then, without ever identifying the genes involved in the process, the linked SNP markers can be used to predict the efficacy and likely side effects of the drug on new patients.

In some cases, differences in drug response between patients identified by pharmacogenomic screening may indicate fundamentally different disease mechanisms. In other words, people with similar symptoms may experience different disease subtypes, and therefore require different treatment. This is especially likely in complex diseases such as asthma or heart disease, where a pharmacogenomic test may reveal similar data to a genetic test on the basis of inherited risk factors. In fact, by using pharmacogenomics to divide patients into subtypes, it may be possible to develop new drugs that specifically benefit only one subclass of patients.

Drug Development Research

Drug companies are also using pharmacogenomic technology to speed up the clinical trials process for new drugs (the most expensive and time-consuming phase of dug development). It is possible to construct genetic profiles of patients in the early-stage clinical trials of a drug and correlate these profiles with drug response and side effects. Then, for later-stage trials, patients can be prescreened to eliminate those likely to respond poorly or to experience side effects. The result of such "stratified trials" will probably be drugs that are approved for use only in conjunction with the genetic test that determines whether it will be effective and safe for each patient. While this represents some loss of profit in comparison with a "one drug fits all" marketing strategy, the drug companies will more than make up for it by the ability to license many drugs that were previously disqualified because of low levels of efficacy on some people or unacceptably high frequencies of side effects. When used together with genetic testing, drugs will be safe and useful for some people. In addition, new, targeted drugs will be sold to niche populations who previously did not benefit from drug treatment (Rothstein et al. 2001).

At a more prosaic level, there are often a number of different drugs available to treat a given condition (high blood pressure, anxiety/depression, migraine headaches, etc.). In the current healthcare system, a patient might receive a prescription for one of these drugs chosen according to whatever their physician has read lately about the incidence of side effects, known negative drug interactions, and so on. After taking the drug for some period of time, the physician assess the effectiveness of the drug and the severity of the side effects and decides whether to continue the prescription, or to change to another drug that perhaps will be more suitable for the patient. In this way, the patient may suffer for many weeks or months (or longer) with one or more ineffective drugs and/or unpleasant side effects when a better drug was available, if only the physician had more information about that patient's genetic drug response characteristics. Pharmacogenomics can provide this information and can also aid the physician in determining appropriate drug dosages. Current methods of basing dosages on weight and age will be replaced with dosages based on a person's genetics—how effective the medicine is in that person's body and the time it takes to metabolize it. This will maximize the therapy's value and decrease the likelihood of overdose.

Genetic Profiles versus Gene Expression

Pharmacogenomics is based on matching drugs to patients on the basis of genetic profiles—identifying specific alleles of known genes, or SNP markers linked to these alleles in each patient. These are permanent characteristics of the genome of each person. There are other situations where it is not the genotype of the patient that determines the effectiveness of a drug, but rather the metabolic state of a particular affected tissue. Different types of cancer tumors respond differently to chemotherapy agents, but it is often difficult to accurately diagnose a tumor with classic

histological pathology methods. However, the gene expression patterns of different tumor types can be distinguished using microarrays that measure levels of mRNAs for various genes (see Chapter 6). This has been demonstrated convincingly for AML versus ALL leukemias (Golub et al. 1999). In other cases a gene expression profile of a tumor can accurately indicate its aggressiveness, which can be useful in determining appropriate drugs and a course of treatment, such as for prostate cancer.

Personalized Medicine

Pharmacogenomic is often described as "personalized medicine" or "designer drugs," but these terms encourage a misunderstanding of the basic technology. Pharmacogenomics will not involve creating a custom drug specifically designed for each patient. Instead, pharmacogenomics offers a form of "mass customization" of drugs, so that the physician can choose from among a panel of available drugs the one best suited to each patient—sort of like a choice of size and colors in a sweater, not like a custom-made suit of clothes.

The National Institute of General Medical Sciences (NIGMS, a branch of the US National Institutes of Health), is currently funding a major research initiative, the Pharmacogenetics Research Network. Research conducted by scientists in the network includes identification of important genetic polymorphisms, functional studies of variant proteins, and studies that relate clinical drug responses to genetic variation. NIGMS is creating a free online database of pharmacogenomic information collected by the scientists and physicians participating in this program, namely, the Pharmacogenetics Knowledge Base (`http://www.pharmgkb.org/`). This online database is hosted and managed by the Stanford Medical Informatics group (SMI), in the Department of Medicine in the Stanford University School of Medicine. Since January 2002, the online database has contained

information about 430 genes that have been shown to affect drug response in clinical studies.

This database is intended for use as a research tool, to help scientists understand how genetic variation among individuals contributes to differences in reactions to drugs and to contribute to the development of new genetically targeted drugs. However, the database is freely accessible on the Web by any physician, patient, or interested member of the public. However, the database is organized for research uses and does not support simple queries by drug name—and even if a gene–drug interaction is found in the database, the existence of a genetic test commercially available for that gene is unlikely.

The Pharmacogenetics Knowledge Base includes health information such as history of disease, physical and physiological characteristics such as height, weight, heart rate, and blood pressure, as well as pharmacogenomic information about any drugs being taken, data on physiologic responses to drugs, and DNA sequences suspected to play a role in these drug responses. However, all personal identifying information has been stripped from these data.

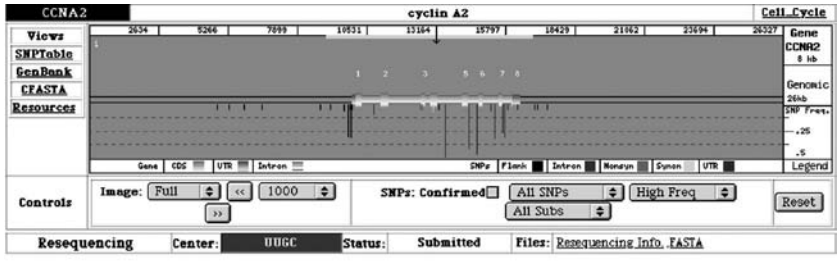
In conjunction with the Pharmacogenetics Knowledge Base, the NIGMS has also created a public education Website called "Medicines for You" (`http://www.nigms.nih.gov/funding/medforyou.html`). This is an excellent resource for patients who are curious about the potential for "personalized medicine."

Environmental Chemicals

Just as people have genetic differences in their responses to drugs, there are genetic differences in their responses to toxic chemicals that occur as environmental pollutants (or as food contaminants, food additives, etc.). Once again, it is possible to collect genetic data on people who demonstrate specific sensitivities to

chemicals. The National Institute of Environmental Health Sciences (NIEHS) has initiated an "Environmental Genome Project" to systematically identify common sequence polymorphisms in genes with suspected roles in determining chemical sensitivity.

Similar to the Pharmacogenetics Knowledge Base, the NIEHS has set up an online database of genetic data linked to susceptibility to environmental chemical exposure. The NIEHS database, GeneSNPs, developed and hosted by The University of Utah Genome Center (`http://www.genome.utah.edu/genesnps/`), contains human genes and sequence polymorphisms related to DNA repair, cell cycle control, cell signaling, cell division, homeostasis, and the metabolism of environmental chemicals. Again, this database is freely available to anyone on the Web, but it is oriented toward the researcher with a specific gene in mind, rather than the physician or layperson with an interest in a particular chemical and its possible genetic effects. However, the display of data for each gene is very impressive—perhaps the best integrated resource for any collection of genetic data (see Figure 10.1). Each gene is shown with its gene name, functional category, coding regions, introns, 5′ and 3′ untranslated



View Frame

All SNPs [CSV]

Implication	Position	Allele	Freq/(Het)	#Chrom.	PopID	Submitter	SS#
proximal promoter	772	T/G	0.06	180	NIH90	UUGC	869790
proximal promoter	925	G/A	0.29	180	NIH90	UUGC	869791
proximal promoter	984	C/T	0.02	180	NIH90	UUGC	869792
proximal promoter	1021	C/T	0.30	180	NIH90	UUGC	869793
proximal promoter	1055	T/C	0.01	180	NIH90	UUGC	869794
flanking	7901	G/C				LEE	1527762

FIGURE 10.1. Genetic data for the cyclin A2 gene in the GeneSNPs database.

regions (UTRs; the part of the mRNA that is not translated into protein), as well as an additional 10 kilobytes (kB) of genomic DNA that flank the coding region on both sides. All known SNPs are annotated as to their location on the gene (5′ UTR, 3′ UTR, intron or exon). For those SNPs that fall occur in an exon, it is further noted whether the mutation is synonymous (e.g., CAG 229 CAA :: Gln 229 Gln) or nonsynonomous (e.g., GTA 163 ATA :: Val 163 Ile) and the amino acid position in the translated CDS.

These data on genetic sensitivity to toxins could be used to help people make lifestyle choices to avoid certain chemicals for which they have a genetic sensitivity. However, they could also be used to discriminate against people, for example, in hiring decisions to limit employer liability for on-the-job exposure to potentially toxic chemicals. This is an ethical gray area—we don't want people to be needlessly exposed to chemicals for which they are particularly sensitive, but we also do not wish to live in a genetic dictatorship where the results of a gene test are used to limit or determine one's employment options.

Another ethical concern is the overlap of these various genetic tests. It is entirely possible, even likely, that if a person were to have a simple genetic test taken to predict the best medicine for a common disorder—perhaps for a prescription for high blood pressure medicine—that test would also reveal information about chemical sensitivity, cancer risk, and other factors. Control over ownership, privacy, and the right to permanently and completely destroy this type of genetic information will be a crucial factor in determining the widespread adoption of this technology in routine medical practice.

Toxicogenomics for Drug Development

DNA microarray technology can be used to measure differences in gene expression (mRNA levels) due to any type of developmental change, response to a pathogen, or environmental

stimulus. This is particularly useful in drug development research where it might be used to detect toxic responses to molecules that are being evaluated as potential drugs. The majority of drug toxicity reactions (adverse side effects) occur in a few well-defined biochemical pathways—common modes of action. These toxic reaction pathways involve induction and/or down-regulation of specific sets of genes. Microarray technologies can be used to detect these changes in gene expression, provided that RNA can be extracted from samples of affected tissues. The gene expression changes observed on a microarray due to the reaction to a specific toxic drug form a **toxicogenomic** profile for that drug and for that pathway of toxic response.

The early stages of testing drug candidate molecules often involves various types of cell culture systems, followed by animal studies (usually mice). Many pharmaceutical researchers are currently working to establish cell culture systems that accurately reflect human toxic reactions to drugs. In such a cell culture system, gene expression profiles can be created for various compounds known to have toxic side effects. Then each new drug candidate can be applied to the cell culture, and the resulting gene expression profile can be compared with the library of known toxicogenomic profiles. Similarly, tissue samples can be collected from affected tissues in mice suffering from known toxic reactions and microarray gene expression profiles compiled. Samples of the same tissues can be collected from mice being treated with experimental drugs and tested on the microarrays. If the gene expression profile of a new drug candidate matches a known toxic profile, then that candidate molecule is eliminated from further testing, saving time, money, and possibly the suffering of unlucky drug testing animals (or humans).

It is also possible to use microarray screening to monitor patients who are currently receiving drug treatment—either as an early alert for toxic effects, or to assay drug effectiveness. The crucial distinction between toxicogenomic technologies and

standard genetic tests it that it is the present state of cellular metabolism that is being measured rather than a permanent genetic characteristic of the patient. A drawback of this technology is that samples must generally be collected directly from the tissues that are the subject of the drug treatment (i.e., a liver biopsy).

Drug Specificity

A drug with effects narrowly targeted on a single tissue is much less likely to produce unwanted side effects than is a drug that produces effects on many different tissue types. Tissue-specific drug effects could be monitored by microarray tests of mRNA extracted from various tissues in an experimental animal treated with the drug. However, this expression profile is only as specific and precise as the tissue used to prepare the RNA, since tissue samples may be composed of many different cell types that can be difficult to separate.

Toxic effects of drugs are sometimes linked to their mode of action—unanticipated metabolic side effects of blocking or stimulating a particular protein or receptor too well—but more often they are the result of interactions between the drug and nontarget molecules. A new genomic technology has been proposed to use microarrays to specifically screen for nontarget effects of drugs. Marton et al. (1998) at Rosetta Inpharmatics (Kirkland, WA) have proposed a method for investigating nontarget effects of drugs. For several drugs, they have created gene expression profiles in a wild-type strain of yeast, and then compared these to gene expression profiles of yeast strains with a mutation in the gene affected by the drug (or another gene in the same metabolic pathway). The concept is that the mutant strain will show little effect when exposed to the drug since its target is absent—thus validating that gene as the true drug target. However, any nontarget effects of the drug will be much easier to detect in the mutant strain when the usual metabolic effects of the drug are absent. This technology is

quite powerful in yeast, which has only 6000 well-characterized genes and for which the production of specific mutants is a simple laboratory procedure. It is much more challenging in mice, and essentially impossible (with current technology) in humans.

Environmental Toxicology

There are also environmental and toxicological applications for microarray technology. Just as gene expression profiles can be established for drugs with toxic effects, similar profiles can be established for various toxic chemicals that might occur in the environment, or be used in industry. These profiles could then be very useful in establishing the mechanism of action for new chemicals with suspected toxic effects. This is important considering that there are approximately 80,000 chemicals in commercial use in the United States with an additional 1000 new chemicals developed each year.

There is no way that industry can afford to conduct complete batteries of animal tests for all of these chemicals, nor could the US Environmental Protection Agency scrutinize complete animal testing results. A preliminary toxicogenomic study of a chemical could be accomplished quickly and inexpensively—perhaps on a small set of well-defined cell cultures. A streamlined preliminary testing procedure based on microarrays would fit in nicely with the current law, known as the Toxic Substances Control Act (TSCA), which requires that a "premanufacture notification" (PMN) be submitted for each new chemical. The PMN is not currently required to contain any toxicity data, which has come under criticism from environmental advocacy groups. Adding toxicogenomic data to the PMN would be both inexpensive and informative.

Toxicogenomic profiles might also be used diagnostically to help determine what type of chemical might be causing an adverse health effect in a person when exposure to toxic chemicals

is suspected. Gene expression microarrays could also provide a more sensitive and earlier assay for exposure to toxic chemicals than do current blood chemistry or physiological tests.

REFERENCES

Golub TR, Slonim DK, Tamayo P et al. 1999. *Science* **286**:531–537.

Holstege FCP, Jennnings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**:717–728.

Klein TE, Chang JT, Altman RB et al. 2001. Integrating genotype and phenotype information: An overview of the PharmGKB project. *Pharmacogenom J*. **1**:167–70.

Kuehl P, Zhang J, Schuetz E et al. 2001. Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. *Nat Genet* **27**(4):383–391.

Marton MJ, DeRisi JL, Friend SH et al. 1998. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med* **4**:1293–1301.

Rothstein MA, Epps PG. 2001. Ethical and legal implications of pharmacogenomics. *Nat Rev Genet* **2**:228–223.

CHAPTER

11

Clinical Research Informatics

Clinical Databases

Genomics technology can rapidly generate very large amounts of gene expression, alternative splice forms, genome sequence, copy number, protein concentrations, and other types of quantitative data. However, in order to apply this technology to medicine, it is necessary to combine these quantitative data with clinical data about patient health, demographics, medical history, drug reactions, and other factors. The development of new therapies and diagnostics from genomics and drug discovery technologies requires a process of translational research and clinical trials. Initially, this means that clinical researchers need to design small experimental trials to test new ideas and new drugs on patients. Each of these trials may involve some unique procedures and data collection methods applied to a small number of patients. This is a difficult situation for database design (research informatics support). Each project is small and has a unique set of data to be collected, and budgets are always tight.

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

Genomics data are generally collected automatically in standard formats by computers, while clinical data are generally collected by hand on survey forms, patient medical charts, log books, and reports from clinical labs. Genomics experiments typically involve just one or two variables (time and treatment, healthy vs. tumor, etc.), while clinical investigations involve many different measures of patient health, history, and reaction to treatment. Medical records at different hospitals (or different clinical units within a single hospital system) may encode and store similar data in different formats. Clinical data are also protected by privacy and consent rules, which creates technical and administrative challenges in order to access hospital information systems to gather information for research. Clinical data are generally recorded as a single number (blood pressure, blood glucose, etc.), or a choice from a pick-list of options (left, right, bilateral, not determined, etc.), while genomics data generally involve thousands of numbers (e.g., expression measurements for thousands of genes). Standard hospital information systems and clinical trial management systems are not designed to handle genomics data (many thousands of measurements for a single sample, unique datafile types). Similarly, genomics data analysis software is not designed to handle the complexity of clinical medicine—where dozens of parameters that are measured with a mixture of numerical and multiple-choice-type values are common. It is extremely difficult to design a database that can hold both of these types of data and allow for all combinations of queries that might be useful to both physicians and researchers. From a computational standpoint, it requires a huge amount of power (or a very clever algorithm) to discover meaningful associations among dozens of multivalued clinical parameters and tens of thousands of genetic markers.

The interaction of genomic and clinical data is especially important for genetic association studies. Large-scale SNP genotyping is becoming common in clinical trials for new drugs and other

therapies (i.e., pharmacogenomics). Thorough analysis of gene expression and genotype data may reveal many combinations of alleles (for genes that were previously not known to interact) that are associated with specific diseases, drug reactions, environmental factors, or treatment outcomes. In order to discover these associations of genotype with phenotype, it is necessary to have data analysis tools that make it possible to correlate any combination of hundreds of clinical parameters (many with text-based categories) with any combination of hundreds of thousands of genotype or tens of thousands of gene expression measurements.

In 2006, the NCBI made its first tentative step into the design of a database for genetic and clinical associations with dbGaP, the database of Genotype and Phenotype. The dbGaP is intended to support genome-wide association studies that explore the relationship between specific genes (genotype information) and observable clinical traits, such as blood pressure and weight, or the presence or absence of a disease or condition (phenotype information). The database will also provide precomputed analyses of the level of statistical association between genes (SNPs) and selected phenotypic variables. The initial design of dbGaP does not include support for gene expression, gene copy number, proteomics, or other types of genomic information.

The initial release of dbGaP contained data on just two studies: the "age-related eye diseases study," a 600-subject study of age-related macular degeneration and age-related cataracts; and the NINDS Parkinsonism study with DNA, cell line samples, and detailed phenotypic data on 2573 subjects. NCBI plans to add studies across a broad range of disease areas, including women's health, neurological disorders, neuropsychiatric disorders, diabetes, and environmental factors in disease. The Framingham Heart Study will provide genotypes for approximately 7000 of the study subjects linked to the numerous types of clinical data collected in the study. At this point, it is unclear how many of the genotype and phenotype variables will be shared across different

studies (common data elements). As of mid-2008, dbGaP contains data from 21 different studies.

While the efforts of the NCBI are valuable, it will remain challenging for quite some time to design efficient data collection and data analysis methods for individual genetic association studies. Comparisons across different studies that use different genomic technologies (or different vendors of similar technologies), assess different sets of genes, and collect heterogeneous clinical data in different formats will remain extremely difficult.

Clinical Trials Management

As clinical projects increase in size (phase II and phase III trials), new data management challenges are encountered. In addition to scaling the size of the database, data must be collected from multiple sites, more complex statistical designs are required, and for studies that span several years, changes in genomics technologies and the availability of reagents may become a problem.

There are many software products available to manage clinical data ranging from those designed to manage patient information in a single physician's office up to those designed for multinational pharmaceutical companies. However, no off-the-shelf software can possibly manage the complexity of all clinical research. Fundamentally, the point of research is to investigate new methods of treating patients, which often involves the collection of new types of data. Genomics technology, with its very large amounts of data, cannot be managed by software designed to manage routine (twentieth-century) medical data. On the other hand, it is also a mistake to design a completely customized database for each clinical project. The effort to design new databases is wasteful, since many of the fields and data types are the same across many studies, and data cannot easily be combined from databases created with ad hoc designs. The obvious solution is a single database with common data fields for

patient demographics and other routine clinical parameters, but with sufficient flexibility to allow for unique user-defined fields and customized data collection and report screens for each study. Then key data fields can be extracted and used for classification of samples for genomics analysis.

IBM has found a solid market for their "life sciences clinical genomics solution," which they claim "has an open, scalable architecture that enables the integration of disparate data sources and provides a range of data mining and analysis tools to mine the integrated information." However this is fundamentally a custom-configured software solution that starts with months of consulting and needs assessment. Unfortunately, this solution is not affordable for every clinical research institute working with genomics data.

There are many vendors of commercial software designed for the collection and management of clinical trial data. Large software companies such as Oracle, SAP, SAS, and Siebel have created customized "enterprise" products for the pharmaceutical/biomedical market. A surprisingly large number of small software vendors are entirely focused on producing products to support clinical trials (Phase Forward, Nextrials, Medidata, Velos, ClinPhone, DataTrak, EResearch Technology). Other companies focus on clinical informatics consulting—adapting various software products to meet the specific needs of client institutions. In some sense, it serves the interests of these companies to maintain a bewildering array of different data formats and standards for clinical information. Once locked into a particular software and data standard, a hospital or research institution will have significant costs to change to a new standard.

Clinical trials management systems (CTMS) must be compliant with the FDA's 21 CFR Part 11 regulation in order to ensure that protocols, investigator brochures, case report forms (CRFs), CTM information, and clinical data remain attributable, traceable, and controlled. However, these FDA requirements fall far

short of enforcing common data standards. There is an industry-sponsored organization, the Clinical Data Interchange Standards Consortium, which develops standards for data in commercial clinical software systems; however, these standards focus mostly on streamlining the process of submission of clinical trials data to the FDA. This organization is handicapped by its need to meet the business goals of its sponsoring companies—so its standards tend to support the current status quo of various existing commercial software architectures, allowing for multiple data models to describe the same types of data (SDTM, LAB, ODM, SEND, ADaM).

The Yale Center for Medical Informatics (ycmi.med.yale.edu) has produced a customizable clinical research database product, known as TrialDB, and is offering it for free to the biomedical community (i.e., open-source, GNU Public License). TrialDB is a generic CDMS/CTMS (clinical trial data system/clinical trial management system) that can manage an arbitrary number of different studies, with no limits on the number of patients per study or the number of parameters that are tracked in each study. It is not necessary to modify the database structure in order to establish parameters for a new type of clinical data. TrialDB includes Web-based data entry forms that can be used to collect data from multiple sites. Via the Web interface, it is also capable of automatically generating CRFs, based on definitions of the data elements in each CRF. As freeware, TrialDB is not designed to function as a simple plug-and-play tool, nor does the Yale CMI provide a tech-support hotline. They do recommend a full-time staff of two (DBA database developer and medical informatics study designer) to maintain the database, add new studies, and provide training to users.

Data Standards and Ontologies

There is a need for standards in data formats so that information from different trials can be combined. In particular, if

the goal of medical genomics is to relate genes (expression and/or mutations) to medical conditions, there is a need for a standard vocabulary to be used to describe medical conditions and gene functions. An **ontology** is a set of terms with definitions and specific logical (semantic) relationships among them. Several groups are working to create standard ontologies to describe clinical parameters, genomics data, and experimental designs. Biologists have traditionally had difficulty developing standard names for things. The anatomical parts and genes of each organism have been named by scientists who specialize in that organism, often without regard to the existence of names for similar structures in other organisms. A single process can be described by several different, synonymous names. For example, the terms "translation" and "protein synthesis" are functionally identical, and used interchangeably in databases and in the scientific literature. This creates difficulties for human searches and makes automated data integration virtually impossible.

The Gene Ontology (GO) project provides a single controlled vocabulary to describe gene and gene product attributes in any organism (`http://www.geneontology.org`). The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of biological processes, cellular components, and molecular functions across all species. These terms are then applied (annotated) to specific genes in each species. Within each of the three ontologies, the terms are organized in a hierarchical structure so that queries can be directed at various levels of specificity. For example, the term "signal transduction" is part of cell communication, which is a cellular process, which is a biological process. The activity of a specific signal transduction protein such as **Rab protein** is four levels down in the hierarchy (a form of "Ras protein signal transduction," which is a form of "small GTPas mediated signal transduction," which is a form of "intracellular

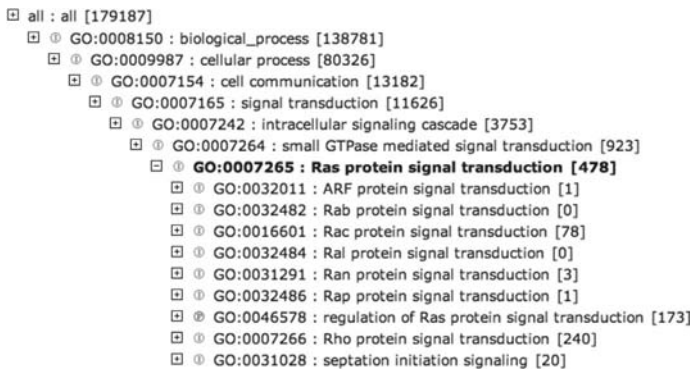


FIGURE 11.1. A hierarchy of "biological process" terms in the Genome Ontology system.

signaling cascade," which is a form of signal transduction; see Figure 11.1).

The MGED (microarray and gene expression data) consortium is approaching this problem from the perspective of standard terminology for gene expression microarray experiments—leading to standard database designs and data exchange formats. The initial products of this effort include the MIAME (minimal information about a microarray experiement) standards, which describe the biology of an experiment and the design of a DNA array, and MAGE-ML, which is a standard XML format for data exchange. Use of these standards has been adopted as a requirement for publication by a majority of peer-reviewed journals.

The National Cancer Institute has put a great deal of work into developing an ontology for cancer research with the specific goal of creating a "biomedical vocabulary that provides consistent, unambiguous codes and definitions for concepts used in cancer research." The result of this project is known as the NCI Theasaurus, which in turn is the basis for the Enterprise Vocabulary Server (EVS) of the caBIG project. The NCI Thesaurus provides broad coverage of the cancer domain, including

cancer-related diseases, findings, and abnormalities; anatomy; agents, drugs and chemicals; and genes and gene products. It combines terminology from numerous cancer-research-related domains, and provides a way to integrate or link these terms together through semantic relationships. The NCI Thesaurus contains over 34,000 concepts, structured into 20 taxonomic trees. The NCI Theasaurus is freely available on the Web via the NCI Terminology Browser and via APIs that allow other programs to access it interactively (http://nciterms.nci.nih.gov/NCIBrowser).

The NCI has funded a National Center for Biomedical Ontology at Stanford Univeristy (`http://bioontology.org`) to "develop innovative technology and methods that allow scientists to create, disseminate, and manage biomedical information and knowledge in machine-processable form." The Center is creating a large collection of biology terms known as the **open biomedical ontologies**. This meta-vocabulary is accessible through a public website known as the BioPortal (`http://bioontology.org/tools/portal/bioportal.html`), where investigators can look up a term for a specific medical or experimental condition, create a link from another application to allow users to choose a term from a specific ontology (lookup list), or download an entire ontology. The BioPortal currently (2007) contains 54 specific ontologies that cover topics that range from human disease and protein modification, to mosquito gross anatomy. The BioPortal also includes many large existing ontologies such as the NCI Theasaurus; NLM/MedLine MESH terms; and the three Gene Ontology (GO) vocabularies of biological processes, cellular components, and molecular functions. The ultimate goal of this project is to create simple, Web-based tools that allow all investigators (and software developers) to easily access biology terms from a set of controlled vocabularies, so that similar objects are consistently described using the same set of terms, and data from different projects and different systems can easily be combined and compared.

Large ontologies such as the NCI Theasaurus are built by combining preexisting vocabularies such as the Unified Medical Language System (UMLS) and Systematised Nomenclature of Medicine (SNOMED), and adding on to them. However, this process may create multiple definitions for the same term and incorporates inconsistencies that existed in these other vocabulary systems.

Tissue Banks

All genomics is based on the analysis of tissue samples – for both the discovery process to find genes that have expression or alleles associated with disease or response to a drug, and the validation process where potential targets or markers are screened on patients. In some cases, groups of patients are recruited for a specific research study and samples are collected and processed immediately for a genomics analysis. Existing repositories of human tissue samples, collected for other purposes, are also a valuable resource for the genomics researcher. However, research progress has been limited by insufficient quantities of well-characterized tissue samples. Scientists need access to high-quality biological samples of disease targets. These samples must be well characterized with sufficient and accurate clinical data to make a connection between genotype and phenotype.

Samples of human tissue are collected every day as part of modern medical practice. Blood samples are taken as a routine part of nearly every medical exam. Biopsy of tissue suspected of disease is the primary method by which diagnostic decisions are made for cancer and many other diseases (hepatitis; vasculitis; amyloidosis; bacterial, viral, or fungal infection; etc.). Pathologists are responsible for processing and reporting on all specimens generated during surgery. Tissue samples are taken from the submitted specimens, stained, and processed for microscopic

evaluation, chemical, antibody (immunohistochemistry), and genetic testing to evaluate disease of any type, and this information is returned to the surgeon via a pathology report. Samples that have been evaluated by pathologists are generally saved in a tissue bank—either frozen in liquid nitrogen or preserved in blocks of paraffin. These stored samples may be used for further evaluation by another pathologist (a second opinion), for additional testing, or for research.

The human tissue (and blood) samples stored at medical centers (tissue banks) have tremendous value for research, both within the context of direct patient care and for additional testing beyond the direct needs of the patient. Specialized tissue banks may be created for specific research projects. The Health Insurance Portability and Accountability Act (HIPAA) ensures that patients control access to their own clinical pathology data and their stored tissue samples. Patients must sign a consent form in order to grant researchers access to their data and tissue samples. This consent form may be very detailed since the patient must be fully informed of the planned research and all the uses in which her or his data and tissue might be involved. In addition, researchers who wish to obtain access to clinical data or stored tissues must have their project approved by the Institutional Review Board (IRB) that governs research at each medical center (see Figure 11.2).

Human biospecimens form the critical bridge between basic science and clinical medicine. Biological paradigms tested in cell lines and animal models can be validated in human specimens with minimal risk to human subjects.

In order to be valuable for research, a tissue bank must maintain an inventory database that matches each sample to a unique patient ID, deidentified patient demographics, and clinical data at the date of sample collection, as well as complete clinical follow-up and outcome data. A well-designed tissue database can serve as a unifying point among diverse medical departments and

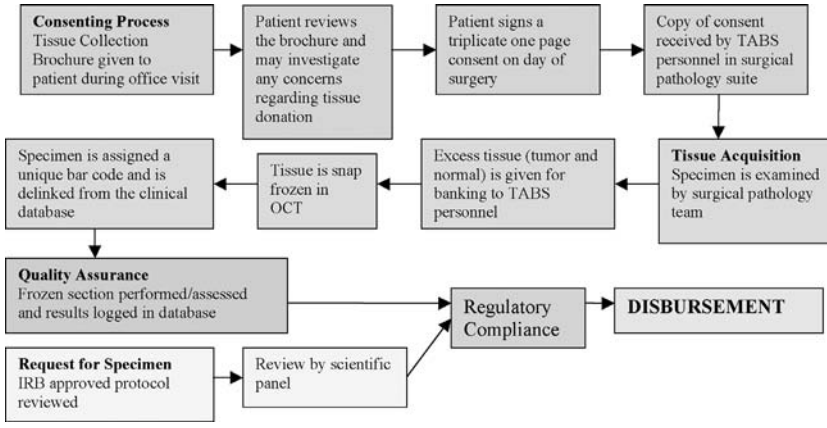


FIGURE 11.2. Tissue acquisition for a tissue bank at a research hospital (Singh 2007).

research groups, creating a set of common data elements (fields and terms to describe patient demographics, sample types, and disease states), and facilitating collaborations between primary care physicians, pathologists, and genomics scientists.

Application to Medical Practice

The problems associated with clinical genomics data storage and analysis during the discovery and early clinical trials process do not directly impact the majority of primary care physicians. However, the concept of genetic association studies is growing increasingly broad. In the very near future, genotype data will be collected from many patients during the normal process of diagnosis and prescribing drugs for many common conditions such as clinical depression and hypertension (pharmacogenomics). These genotype data will be valuable to researchers and pharmaceutical companies in assessing the effectiveness of drugs and diagnostic tests, and may help in discovering additional subclasses (i.e., alleles of specific genes) that have a functional effect on disease progress or drug effectiveness. Patient consent will

be obtained at the time a sample is drawn for genotyping (or gene expression, biomarker, etc.) as part of a diagnostic or drug choice process. Deidentified patient demographic data, results of the genomics tests, and the observations of the physician will be collected by database software operated by biomedical research investigators or by pharmaceutical companies. In essence, all genomics-based drugs and diagnostics will be in extended phase IV clinical trials, and all physicians who use these products will be research collaborators and all patients will be research participants.

REFERENCES

Ceusters W, Smith B, Goldberg L. 2005. A terminological and ontological analysis of the NCI Thesaurus. *Meth Inform Med* **44** (4): 498–507.

Life Science Insights: U.S. Clinical Trial Management Systems. 2004. In Hanover J, Julian EH (eds.), *Vendor Analysis: Leadership Grid and Market Shares*. IDC, Framingham, MA, 2004.

National Cancer Institute, Office of Communications, Center for Bioinformatics. NCI terminology browser, `ftp://ftp1.nci.nih.gov/pub/cacore/EVS/`.

National Center for Biomedical Ontology, `http://bioontology.org`.

NCBI dbGaP, `http://www.ncbi.nlm.nih.gov/entrez/query/Gap/gap_tmpl/dbGaP_HowTo.pdf`.

Open Biological Ontologies, `http://obo.sourceforge.net/`.

Singh B, Director of Tissue Acquisition and Banking Service, NYU Cancer Institute. 2007. Personal communication.

CHAPTER

12

RNA Interference and MicroRNAs

Cell biology is fractal. The more closely each structure and process is examined, the more complexity and detail is discovered. The classic "Central Dogma" model of genome function is that genomic DNA encodes genes, which are transcribed into messenger RNA (mRNA) molecules, which are then translated into protein. Information flows strictly from DNA to RNA to protein. DNA-binding proteins (transcription factors) provide regulation of genes by controlling the rate at which mRNAs are produced. In addition to mRNA, RNA molecules also function in translation as part of the ribosome (rRNA), and transfer RNA (tRNA) molecules act as carriers that bring amino acids to the site of polypeptide synthesis on the ribosome. More recent research has revealed many new functions of RNA, including defense against viruses, gene silencing, and an entirely new class of microRNAs that do not code for protein but regulate the expression of other genes.

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

Antisense RNA

In the standard model of gene expression, the mRNA acts as a single-stranded molecule that carries protein-coding information from the DNA to the ribosome. However, it is possible for one RNA molecule to form a double-stranded molecule with another strand of RNA with a complementary sequence—the **antisense** strand. In theory, the formation of double-stranded RNA (dsRNA) would effectively block the message, prevent translation, and lead to the degradation of the dsRNA by RNAse enzymes. Antisense RNAs have been documented as gene regulation mechanisms in many bacterial and viral systems (Wagner et al. 2002)

The gene silencing effect of antisense RNA has been demonstrated in many *in vitro* and animal experiments, but so far has not been successfully implemented in clinical medicine, with the exception of the antiviral drug fomivirsen. Fomivirsen (trade name Vitravene) is a 21-base phosphorothioate oligonucleotide that is complementary to the mRNA transcribed from the major immediate–early transcriptional unit of cytomegalovirus retinitis. However, it is possible that some of the effective demonstrations of antisense RNA in suppressing gene expression were actually produced by other mechanisms, such as RNA interference (discussed below).

There are a number of challenges for the medical use of antisense RNA technology to block the expression of specific genes:

1. RNA is an inherently unstable molecule, which is actively degraded in all cells as a normal function of the gene expression process. It is not possible to regulate gene expression at the level of transcription unless existing mRNAs have a limited lifespan.
2. All cells and body fluids contain highly active RNase enzymes as a defense against foreign RNAs, such as from viruses.

3. It is difficult to direct a therapeutic RNA to the desired site within the body. RNA is degraded in the stomach and the bloodstream.
4. Since the mechanism of antisense RNA involves direct one-to-one binding of antisense RNA molecules to target gene mRNA, the effective concentration of antisense RNA within the cell must be very high to achieve a substantial reduction of gene expression. Thus the antisense molecule must be administered repeatedly at a high dose.
5. Antisense RNA molecules have nontarget effects. This may be as result of nonspecific RNA–RNA binding, or generalized immune reactions to the presence of these molecules.

RNA Interference

The process of RNA-mediated gene silencing is generally known as **RNA interference** (RNAi). The first hints about this process were discovered in petunia plants. Researchers were trying to use genetic engineering methods to insert additional copies of a pigment gene to produce a darker-colored flower, but instead, some of the transformed plants produced white or variegated flowers (see Figure 12.1) (van der Krol et al. 1990; Napoli et al. 1990). An extra copy of the pigment gene (chalcone synthase) blocked expression of the native gene—a phenomenon that was called **transgene silencing** or **cosuppression**. The mechanism of gene silencing was attributed to the production of double-stranded RNA from the inserted copy of the gene, which led to the shutdown of expression of the endogenous gene.

The details of the process of RNA interference were further studied in the nematode worm *Caenorhabditis elegans* by Andrew Fire and Craig Mello (Fire et al. 1998), for which they were awarded the Nobel Prize in 2006. They found that short (21–22 bp) double-stranded RNA molecules, when injected into the worm, could silence the expression of genes with a

FIGURE 12.1. Petunia flower with variegated pattern caused by cosuppression of chalcone synthase (pigment) gene by RNA interference (photo by R. Jorgensen, reprinted with permission). (See insert for color representation.)

complementary sequence. A similar process was found to occur in plants infected with an RNA virus that contained a portion of a plant gene (Wassenegger et al. 1994).

However, in mammals, the introduction of long dsRNA molecules induces the interferon pathway, which leads to a global shutdown of protein synthesis and cell death through apoptosis. This pathway is thought to function as a defense against viruses, which produce dsRNA. It was found that short dsRNA molecules of 21–23 base pairs, called **small interfering RNA** (siRNA), could be used effectively to block expression of specific genes in mammals without triggering the interferon response (Elbashir et al. 2001; Caplen et al. 2001).

Any double-stranded RNA molecules in the cell may become a substrate for this RNA interference pathway, including viral RNA, RNA produced by transposons, antisense transcripts from protein-coding genes, and the products of RNA-dependent

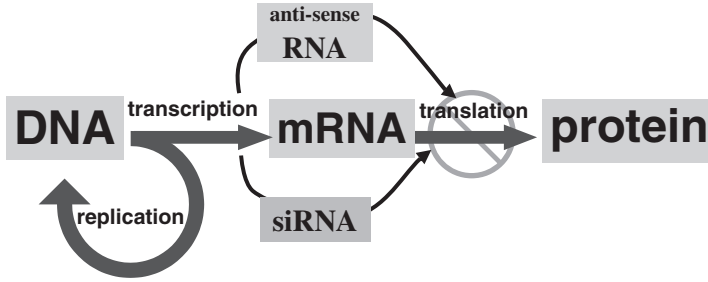


FIGURE 12.2. siRNA and antisense RNAs regulate gene expression by blocking translation of specific mRNAs.

RNA polymerase. These long double-stranded RNA molecules are cut up into 21 or 22 bp fragments by a ribonuclease III enzyme known as **dicer**, which is located in the cytoplasm. These short double-stranded RNA molecules are known as **short interfering RNA** (siRNA). The sense strand of the siRNA is removed, and the antisense strand associates with a protein complex known as **RNA-induced silencing complex** (RISC). The single-stranded siRNA guides the complex to complementary mRNAs, which are cleaved by an RNAse in the RISC. This is a catalytic process, so one RISC with an associated antisense siRNA strand can destroy many mRNA molecules. siRNA may also play a role in transcriptional gene silencing by inducing changes in chromatin structure in the DNA of the target gene.

MicroRNA Genes

RNA interference was initially thought to be a mechanism that reacted to external RNAs, such as a virus, or to the unusual presence of double-stranded RNA, which might result from the transcription of the antisense strand of a gene. However, it was discovered that eukaryotic cells have a native RNAi mechanism that is an important aspect of posttranscriptional regulation of gene expression. MicroRNAs (miRNAs) are encoded by genes that are not

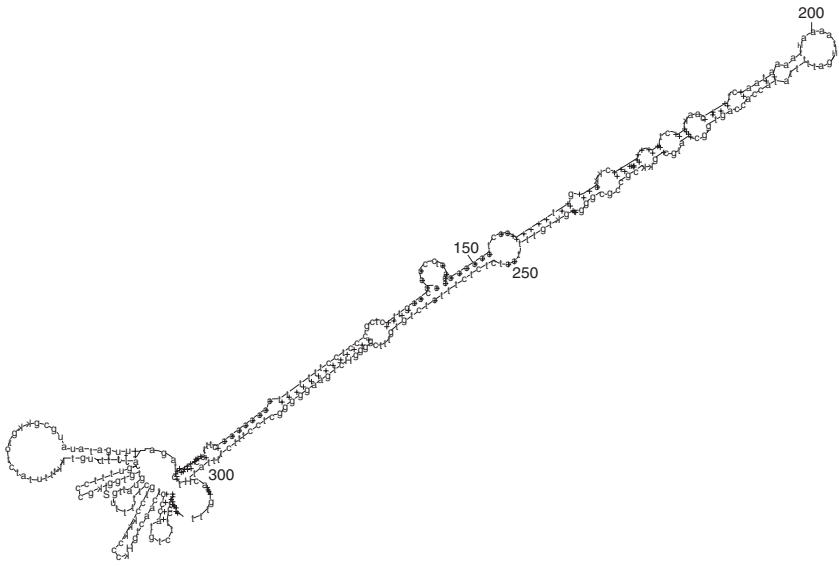


FIGURE 12.3. The predicted stem–loop structure of a precursor microRNA molecule, miR319c.

translated into proteins (noncoding), but regulate the expression of other genes by direct RNA–RNA interactions. The primary transcript of a miRNA gene contains inverse repeats that cause it to fold up to form a double-stranded RNA molecule, the precursor miRNA (see Figure 12.3). This large dsRNA molecule is cut by an enzyme called **drosha** to form a smaller hairpin folded dsRNA that is approximately 50 bases in length (see Figure 12.4). This short hairpin structure is then further processed by the dicer

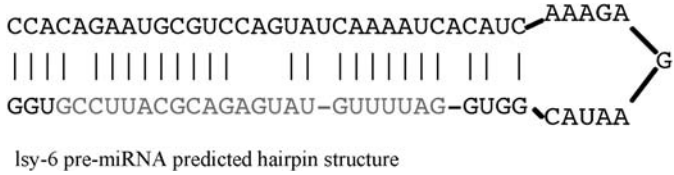


FIGURE 12.4. The 50-base hairpin structure of the lsy-6 pre-miRNA from *Caenorhabditis elegans*.

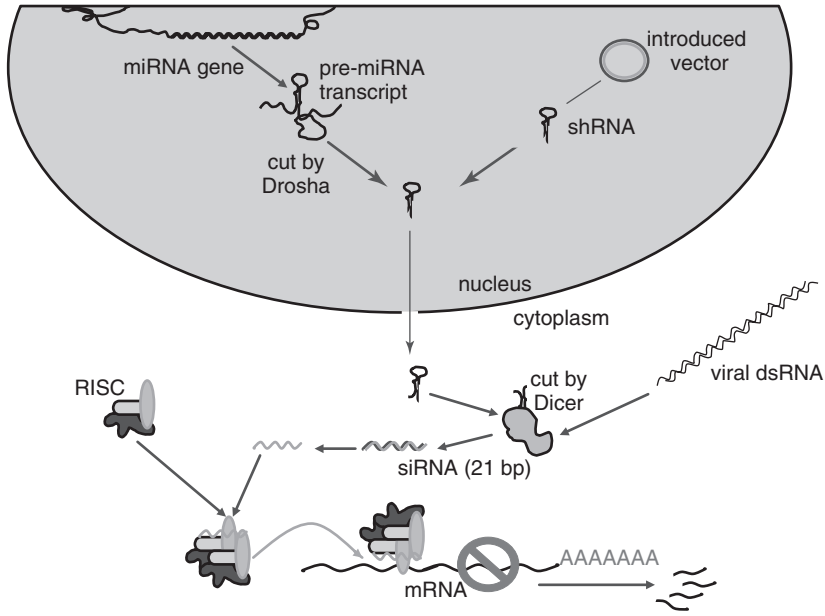


FIGURE 12.5. Diagram of RNA interference mechanism showing dsRNA produced from microRNA genes, introduced vector DNA, or viral RNA, all processed by dicer and bound to RISC, which blocks translation or degrades mRNA.

enzyme to cut off the loop between the paired strands to form a double-stranded miRNA 21–22 bp in length. Then, like siRNAs, the dsRNA fragments associate with RISC, the sense strand is removed, and the antisense strand binds to a complementary sequence on an mRNA. MicroRNAs act as negative regulators by specifically binding to short complementary sequences in the 3′ UTR of the mRNA of target genes and then either targeting the mRNA for degradation, or blocking its translation (see Figure 12.5).

The first microRNA gene, *lin-4*, was discovered in *C. elegans* by Lee and coworkers in 1993, before the concept of RNA interference (RNAi) was understood. The *lin-4* gene produces a a 61-nucleotide RNA transcript that contains a sequence

complementary to sequences in the 3′ untranslated region (UTR) of *lin-14* mRNA, suggesting that *lin-4* regulates *lin-14* translation via an antisense RNA–RNA interaction.

The sequence of the antisense strand of miRNAs generally are not perfectly complementary with their target mRNAs. This enables each miRNA gene to target many different mRNAs with nearly complementary recognition sequences in the 3′ UTR region (Lim et al. 2005). In addition, each mRNA may have multiple recognition sites for miRNA binding—and these sites may be recognized by multiple copies of the same miRNA, or be bound by several different miRNAs, thus allowing for complex combinatorial regulation of protein-coding genes by miRNAs. MicroRNA genes occur throughout the genome in several different configurations. Some microRNAs are located within introns of other genes, which may or may not be regulated by that microRNA. Some have their own promoters, which in turn may be regulated by proteins acting as transcription factors, and some occur in polycistronic clusters where several microRNAs are produced on a single mRNA with a single promoter, then processed into separate short RNAs by drosha.

It is difficult to precisely predict the genes that are the regulatory targets of individual miRNAs using bioinformatics methods because the RNA–RNA binding is limited to a short region of about 8 bp, and it often involves mismatched bases. Experimental verification of miRNA targets is also challenging since it generally requires multiple transgenic cell lines (miRNA knockouts and gene replacements). Microarrays that measure gene expression via mRNA transcript levels of target genes are also imprecise in measuring the effects of miRNAs on gene expression, since translation of an mRNA into protein can be blocked without affecting the overall level of the transcript in the cell. Experimental verification of the effects of miRNAs are generally conducted at the protein level, either by Western blot (using a specific antibody) or by proteomics.

The expression of the ~450 known human microRNA genes can be measured using microarray or quantitative real-time PCR technology (Griffiths-Jones et al. 2006). The expression of microRNAs is highly tissue-specific and consistent for each tissue type. MicroRNA expression profiles provide very accurate profiles of tissue and cell types. This is very useful for cancer pathology since tumor cell type is a strong indicator of disease progression and often determines which drugs will be most effective for treatment. Many cancers are detected as metastasis of unknown origin, and an accurate determination of the cell type is possible only with molecular diagnostics.

The overall levels of microRNA expression have been observed to change during the process of developmental differentiation of cells. Cellular differentiation is largely a process of shutting down the expression of genes. Stem cells and undifferentiated cell types show lower overall microRNA levels than do highly differentiated cell types, and specific differentiated cell types have characteristic profiles of miRNA expression. Cancer cells, like stem cells, have low overall levels of miRNA expression. This suggests that microRNAs play an important role in shutting down unneeded gene transcription and maintaining the differentiated state.

MicroRNA expression profiling can distinguish normal from tumor cells in many different types of tissue, including lung, breast, stomach, prostate, colon, and pancreatic tumors (Volinia et al. 2006). A significant increase in levels of *miR-221*, *miR-222*, and *miR-181b* were detected in human thyroid papillary carcinomas in comparison with normal thyroid tissue (Pallante et al. 2006). In fact, the same miRNA genes are often up- or down-regulated in many different tumor types. MicroRNA expression profiling of human tumors has identified patterns of differential expression of specific miRNAs associated with diagnosis, staging, progression, prognosis, and response to treatment (Calin and

Croce 2006). In several cases, specific deletions or overexpression of miRNA genes have been demonstrated to cause cancer in a mouse model (Croce 2007). A microarray of miRNA genes may provide a more precise and reliable cancer diagnostic tool than can a microarray based on protein coding genes.

Since miRNAs act to inhibit gene expression, a miRNA that inhibits a tumor suppressor gene, such as p53, is itself an oncogene. A miRNA that blocks expression of an oncogene, such as *ras*, is itself a tumor suppressor. The *let-7* microRNA controls timing of cell cycle exit and terminal differentiation in *C. elegans* and is poorly expressed or deleted in human lung cancers. Inhibition of *let-7* in lung cancer cell line A549 leads to increased cell division in lung cancer cells. Conversely, overexpression of *let-7* in cancer cell lines alters cell cycle progression and reduces cell division. An expression profile analysis of cells that overexpress *let-7* reveals that multiple genes involved in cell cycle and cell division functions are repressed. This suggests that the *let-7* miRNA plays an important role in the regulation of cell proliferation and acts as tumor suppressor.

MicroRNAs may serve as a focal point in cancer disease progression. Many different genetic changes (mutations, gene deletions, duplications, translocations, either in the miRNA genes themselves, or in genes that regulate expression of miRNAs) lead to changes in the expression of a few key miRNA genes, which in turn control protein expression in key developmental pathways that lead to cancer phenotypes. As a focal point, miRNA genes are an attractive target for anticancer drugs. A single drug that inactivated a tumor-promoting miRNA, or stimulated a tumor suppressor miRNA, could be useful in many different types of cancer. The miRNA could be targeted via drug molecules that affected its transcription (interacting with specific transcription regulatory proteins), or by antisense oligonucleotides that blocked its ability to interact with mRNAs. A trial in mice of an antisense RNA molecule that binds and inhibits *miR-122* lowered

cholesterol by up to 35% and also seems to protect the liver against hepatitis C infection (Geddes 2007)

Changes in expression levels of miRNAs lead to large changes in cellular phenotype and development because each miRNA has multiple mRNA targets. However, changes in the miRNA-binding sequence of a single gene, such as SNPs within the 3′ UTR region, could affect the regulation of just that one gene, perhaps in a tissue-specific manner. The implications of SNPs in microRNA-binding sites has been explored in genes that affect human behavior. MicroRNA response elements have been identified in the mRNAs from genes encoding the cannabinoid receptor type 1 (CNR 1), serotonin receptor type 1B (HTR1B), catechol-*O*-methyl transferase (COMT), GABA receptor α-2 subunit (GABRA2), serotonin receptor type 2C (HTR2C), and serotonin transporter (SLC6A4). The common SNP rs13212041 disrupts the repression of the serotonin 1B receptor mRNA by *miR-96*, a brain-specific microRNA. Individuals who harbor this variant may lack the ability to repress serotonin receptor 1B levels in response to brain stimuli. In human population studies, this variant correlates with the frequency of conduct disorder behaviors and adolescent risk for substance dependence.

A mutation that creates a new microRNA-binding site can also have dramatic phenotypic effects. A mutation creating a possible microRNA target site in the myostatin gene has been linked to increased muscularity in sheep (Clop et al. 2006). The myostatin GDF8 allele of Texel sheep contains a G→A transition in the 3′ UTR that creates a target site for the *miR-1* and *miR-206* miRNAs that are highly expressed in skeletal muscle. This causes translational inhibition of the myostatin gene, which may contribute to the muscular hypertrophy of Texel sheep. A microsatellite mapping analysis demonstrated tight linkage between this allele and the hypermuscular phenotype. Analysis of SNP databases for humans and mice demonstrates that mutations creating or

destroying putative miRNA target sites are abundant and might be important effectors of phenotypic variation.

The RNA interference (RNAi) pathway acts as another entire layer of posttranscriptional regulation of gene expression, and microRNAs are a completely new set of functional RNA molecules. RNAi may also be involved in transcriptional silencing by inducing changes in chromatin structure (DNA methylation and modification of histones). Evidence has been found that RNA interference is involved in the regulation of a wide variety of cellular processes and biological pathways, including many that are important in the development and progression of cancer. It is now thought that a significant percentage of all mammalian genes (perhaps 25–30%) have mRNAs that are recognized and that their expression is modified by one or more microRNAs. So microRNAs may be as important for the regulation of gene expression as transcription factor DNA-binding proteins.

In an effort to mimic the function of microRNA genes, several investigators have developed vectors that produce a similar hairpin RNA transcript. These are known as **short hairpin RNA** or **shRNA**. The advantage of this system is that, if the vector can be inserted in the target cells, the double-stranded RNA will be produced within the cell, so repeated treatments with synthetic RNA molecules would not be needed. In addition, expression of the shRNA from the vector can potentially be controlled by the use of an inducible promoter.

RNAi has become extremely valuable as a research tool in plants and invertebrates, providing the ability to analyze the functional effects of a "knockout" of any individual gene. Prior to the development of RNAi, it was generally not feasible to conduct knockout studies on humans, and it was technically challenging and expensive to create gene knockouts in mice. Genome-wide RNAi screens have been carried out in *C. elegans* and *Drosophila melanogaster* (fruitfly). RNAi has been used to create gene knockout mice. It is also possible to combine RNAi

with high-throughput genomic methods to construct a library of siRNAs for every human gene and conduct high-throughput screens of the entire library against cell cultures (Mukherji et al. 2006). This is equivalent to the genome-wide gene knockout experiments that have been conducted in the past in bacteria and yeast. In this way, the function of every gene can be tested by removing its protein product and assaying for various phenotypic changes. In the human RNAi screen developed by Mukherji and colleagues, cell cultures were transfected with ~35,000 different siRNAs, allowed to grow for 3 days, then stained with DAPI and assayed by automated single-cell fluorescence microscopy for cell cycle abnormalities (nuclear DNA content, perimeter-to-area ratio, percentage of cells in G1, S, and G2/M phases). Now that the RNAi gene knockout methodology is established for human cells, the effects of individual gene knockouts on many other phenotypes can be screened.

Medical Applications

RNAi is sometimes called a "gene silencing" technology since it blocks the expression of individual genes by specifically degrading and/or blocking translation of the mRNA for a target gene. The direct use of RNAi for human therapies is being pursued to limit angiogeneisis (blood vessel growth) and for viral infections, cancers, and inherited genetic disorders. Since RNAi is highly sequence-specific, it may be possible to block the expression of one allele of a gene in a heterozygous individual. A number of autosomal dominant genetic diseases are caused by expression of a mutant allele of a gene, leading to a disease-causing protein (Huntington's disease, ALS, neurofibromatosis, Marfan syndrome, achondroplasia, etc.), which could be directly targeted with an siRNA to block expression and eliminate (or greatly reduce) the mutant disease-causing protein. There are many other diseases such as cancer, autoimmune

disease, and macular degeneration, where the disease process might be blocked by knocking out the expression of a key gene in a metabolic–regulatory–developmental pathway. There are also some situations where disrupting normal metabolic processes in specific cells is the goal—such as a hair removal cream that disrupts hair growth. There have also been successful experiments that show siRNA can be used to block the replication of viruses (Morrissey et al 2005).

Measurement of microRNA levels (microRNA expression profiles) may provide important diagnostic information such as tumor type, stage, and aggressiveness. Calin et al. (2005) reported that a unique microRNA signature correlates with the prognosis and clinical course of chronic lymphocytic leukemia (CLL). The researchers used a microarray of 190 human miRNA genes to analyze the microRNA profiles of CLL cell samples from 94 patients. They identified a signature of 13 microRNAs that could differentiate between aggressive and nonaggressive CLL. The researchers also found somatic mutations in microRNA genes in 11 of 75 patients with CLL. These mutations were not found in normal controls, which suggests that some microRNAs could function as tumor suppressor genes. The authors conclude that microRNA expression can be used as a diagnostic marker for aggressiveness of CLL. In addition, mutations in microRNA genes are common and might contribute to the mechanism behind this prevalent form of leukemia.

RNAi has been demonstrated to inhibit viral replication *in vitro* for many different pathogenic viruses, including HIV, influenza, hepatitis C, hepatitis delta, rotavirus, respiratory syncytial virus, poliovirus, West Nile virus, foot-and-mouth disease, dengue virus, human papillomavirus, hepatitis B, and herpes simplex. Some specific examples include the use of a 50-bp mhRNA vector, transfected into mammalian cells, which suppresses the replication of multiple hepatitis C viruses (Akashi et al. 2005). Lentiviral and hairpin-type siRNA expression vectors

encoding a fragment of the HIV-1 *env* gene exhibited sequence-specific suppression of target gene expression and strongly inhibited HIV-1 infection in cultured human cells (Hayafune et al. 2006). Transduction with a lentivirus-based short hairpin RNA (shRNA) expression vector containing a portion of the HIV-1 *nef* gene inhibited viral replication in a monocytic cell line and in primary monocyte-derived macrophages (Yamamoto et al. 2006).

An alternate antiviral strategy is to use RNAi to repress host genes that enable viral infection. Lentiviral vectors of short hairpin design incorporating siRNAs for both the CXCR4 and CCR5 cellular coreceptor genes were transduced into cultured human Magi and Ghost cell lines. The transduced cells demonstrated marked viral resistance when challenged with X4 and R5 tropic HIV-1. HIV-1 resistance was also observed in primary PBMCs transduced with the same bispecific lentiviral vector (Anderson and Akkina 2005).

A few *in vivo* studies in mouse suggest that siRNA can be effective in blocking gene expression of native or viral genes. Giladi and colleagues (2003) used hydrodynamic delivery (a rapid injection of a large volume of aqueous solution into the mouse tail vein, creating high pressure in the vascular circulation that leads to extensive delivery into hepatocytes) of siRNA to inhibit levels of hepatitis B viral transcripts, viral antigens, and viral DNA in liver and sera.

With this multitude of targets, pharmaceutical and biotechnology/genomics companies are investing a great deal of research effort into developing siRNA for clinical applications. The primary challenge for safe and effective use of siRNA for therapeutics is the difficulty of delivering the molecules into the correct cells or tissues in sufficient quantities, and the stability of the siRNA inside the cell. RNA molecules are large and negatively charged, so they are not easily taken up by cells across the cell membrane. Unprotected siRNA has a half-life in circulating blood of only a few minutes, due to the high levels of RNase

enzymes in serum. This is essentially the same dilemma faced by gene therapy technologies (see Chapter 7), and some of the same approaches are being tried. Chemical modification of the RNA nucleotides can render siRNA resistant to RNases without significantly decreasing its gene silencing activity. Liposomes aid in moving siRNA molecules across the plasma membrane. Alternately, siRNA can be produced inside the cell from introduced plasmid or viral expression vectors, or genes that produce siRNA could be integrated in the genome. However, random integration of new DNA into the genome may result in insertional mutatgenesis, knocking out or disrupting the regulation of important genes, which could lead to malignant transformation. Viral vectors such as adenovirus may be used that can survive and/or replicate within host cells, but this may lead to host immune response either to viral proteins or directly to the viral DNA/RNA (i.e., the interferon response).

If the delivery challenges can be overcome, RNAi has the potential to be used to develop therapeutic agents that control multiple targets. Many types of human diseases, such as cancer and autoimmune disorders, result from the overexpression of multiple disease-causing genes. Combinations of RNAi agents targeted at several genes can be delivered with little additional difficulty over the delivery of an agent with a single target. Combination therapies (drug cocktails) have been shown to be very effective for treatment of HIV and cancer chemotherapy. Multiple hits on a biochemical or regulatory pathway often have synergistic effects (blocking alternate pathways that might compensate for a single blocked gene).

Other technical challenges include the choice of optimal sequences for siRNA—which include %G+C content, optimal bases at each end of the sequence, and regions of the target mRNA that may be best suited for siRNA binding. Unwanted side effects from RNAi therapies due to suppression of nontarget genes (off-target effects) due to partial sequence complementarity

(cross-hybridization) is also an issue. Natural microRNAs can be effective with as little as 7 complementary bases between the antisense RNA molecule and the target mRNA, but in experimental models, synthetic RNAi molecules often exhibit off-target effects with anything fewer than 15–18 bases of perfect complementarity. Natural microRNAs and their target mRNAs have coevolved, so it is likely that all of the factors involved in target recognition are not yet understood.

RNAi Clinical Trials

The first clinical trial of siRNA in humans involved therapy for age-related macular degeneration of the eye (AMD), conducted in 2004–2005 by Sirna Theraputics, Inc. This trial used a siRNA targeted at the protein vascular endothelial growth factor receptor (VEGFR1), which mediates the development of abnormal blood vessels in the eyes. Of 26 patients in the phase I trial, 25 showed visual acuity stabilization (halted the progression of the disease) and 23% of those patients experienced clinically significant improvement in visual acuity 8 weeks after a single siRNA injection. The patients also showed a decrease in central foveal thickness measured by ocular coherence tomography (OCT), which was the first demonstration of biological activity of a siRNA in humans.

Another siRNA therapeutic that targets the mRNA for the VEGF gene in macular degeneration and in diabetic macular edema has reached phase II clinical trials. Many other siRNA drugs are in animal testing, including a therapeutic agent for type 2 diabetes targeting protein tyrosine phosphatase 1B (PT1B) in the liver, blocking the replication of hepatitis B and C virus; silencing of the mutant form of superoxide dismutase (SOD1), which is the cause of amyotrophic lateral sclerosis (ALS); blocking the expression of the mutant allele of the huntingtin protein in Huntington's disease, and blocking the expression of interleukin-4 receptor in the lungs in asthma.

Inhaled siRNA has been shown to be effective against influenza and respiratory syncytial virus (RSV) in mice (Ge et al. 2004; Zhang et al. 2005). Alnylam Pharmaceuticals, Inc. (2006) started a phase I human clinical trial in 2006 of an inhaled aerosolized siRNA drug candidate against respiratory syncytial virus to assess safety and pharmacokinetics on 80 healthy adult volunteers. This is the largest trial of siRNA in humans to date. Phase II trials to assess protection from RSV are planned for 2007.

Ribozymes

Ribozymes are another form of "active RNA" that can be used to inactivate mRNA molecules within a cell. Ribozymes are RNA molecules that act as sequence-specific RNase enzymes, finding a target mRNA and catalyzing the cleavage of a phosphodiester bond in the sugar–phosphate RNA backbone. Ribozymes are not a native form of gene regulation in eukaryotic cells, but are based on the ability of some mRNA molecules to remove intron sequences by self-splicing in the absence of any protein enzyme. In a laboratory setting, ribozymes have been shown to inactivate expression of specific genes. However, ribozymes have never been used in human clinical trials. The primary obstacles are delivery of sufficient quantity of the ribozyme to target cells, stability of the ribozyme molecule within the cell, and the avoidance of inducing immune response and/or apoptosis (interferon pathway).

References

Akashi H, Miyagishi M, Taira K et al. 2005. Escape from the interferon response associated with RNA interference using vectors that encode long modified hairpin-RNA. *Mol Biosyst* **1**(5–6):382–390.

Alnylam Pharmaceuticals, Inc. 2006. Alnylam Initiates Phase I Clinical Study of Inhaled Formulation of ALN-RSV01, an RNAi Therapeutic for the Treatment of Respiratory Syncytial Virus (RSV) Infection, Cambridge, MA (*Business Wire*), Oct 11, 2006.

Anderson J, Akkina R. 2005. HIV-1 resistance conferred by siRNA cosuppression of CXCR4 and CCR5 coreceptors by a bispecific lentiviral vector. *AIDS Res Ther* **2**(1):1.

Calin GA, Ferracin M, Croce CM et al. 2005. A micro-RNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *New Engl J Med* **353**:1793–1801.

Calin GA, Croce CM. 2006. Micro-RNA signatures in human cancers. *Nat Rev Cancer* **6**(11):857–866.

Caplen NJ, Parrish S, Imani F, Fire A, Morgan RA. 2001. Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems. *Proc Natl Acad Sci USA* **98**(17):9742–9747.

Clop A, Marcq F, Georges M et al. 2006. A mutation creating a potential illegitimate micro-RNA target site in the myostatin gene affects muscularity in sheep. *Nat Genet* **38**(7):813–818.

Croce CM. 2007. Personal communication.

Elbashir SM, Lendeckel W, Tuschl T. 2001. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev* **15**:188–200.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**:806–811.

Geddes L. 2007. New gene therapy targets cholesterol. *New Sci* **2604**:8.

Giladi H, Ketzinel-Gilad M, Rivkin L, Felig Y, Nussbaum O, Galun E. 2003. Small interfering RNA inhibits hepatitis B virus replication in mice. *Mol Ther* **8**(5):769–776.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: Micro-RNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**(database issue):D140–D144.

Hayafune M, Miyano-Kurosaki N, Takaku H, Park WS. 2006. Silencing of HIV-1 gene expression by siRNAs in transduced cells. *Nucleosides Nucleotides Nucleic Acids* **25**(7):795–799.

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294**:853–858.

Lee, RC, Feinbaum RL, Ambros V. 1993. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**:843–854.

Lim LP, Lau NC, Linsley PS et al. 2005. Microarray analysis shows that some micro-RNAs downregulate large numbers of target mRNAs. *Nature* **433**:769–773.

Morrissey D, Lockridge J, Polisky B et al. 2005. Potent and persistent *in vivo* anti-HBV activity of chemically modified siRNAs. *Nat Biotechnol* **23**(8):1002–1007.

Mukherji M, Bell R, Schultz PG et al. 2006. Genome-wide functional analysis of human cell-cycle regulators. *Proc Natl Acad Sci USA* **103**(40):14819–14824.

Napoli C, Lemieux C, Jorgensen R. 1990. Introduction of a chimeric chalcone synthase gene into petunia tesults in reversible co-suppression of homologous genes in trans. *Plant Cell* **2**:279–289.

Pallante P, Visone R, Fusco A, et al. 2006. MicroRNA deregulation in human thyroid papillary carcinomas. *Endocr Relat Cancer* **13**:497–508.

van Der Krol AR, Mur LA, Beld M, Mol JN, Stuitje AR. 1990. Flavonoid genes in petunia: Addition of a limited number of gene copies may lead to a suppression of gene expression. *Plant Cell* **2**(4):291–299.

Volinia S, Calin GA, Croce CM et al. 2006. A micro-RNA expression signature of human solid tumors defines cancer gene targets. *Proc Natl Acad Sci USA* **103**(7):2257–2261.

Wagner EG, Altuvia S, Romby P. 2002. Antisense RNAs in bacteria and their genetic elements. *Adv Genet* **46**:361–398.

Wassenegger M, Heimes S, Riedel L, Sanger HL. 1994. RNA-directed de novo methylation of genomic sequences in plants. *Cell* **76**(3):567–576.

Yamamoto T, Miyoshi H, Yamamoto N, Yamamoto N, Inoue JI, Tsunetsugu-Yokota Y. 2006. Lentivirus vectors expressing short hairpin RNAs against the U3-overlapping region of HIV nef inhibit HIV replication and infectivity in primary macrophages. *Blood* 1**08**(10):3305–3312.

Zhang W, Yang H, Mohapatra SS, et al. 2005. Inhibition of respiratory syncytial virus infection with intranasal siRNA nanoparticles targeting the viral NS1 gene. *Nat Med* **11**:56–62.

CHAPTER

13

Alternative Splicing

The Human Genome Project is widely regarded as a great success. The complete sequence of the human genome has been determined with greater than 99% accuracy and is available online from GenBank, Ensembl, and the UCSC Genome Browser. This sequence information is useful for biomedical research; however, most of the practical value is not derived directly from the raw sequence, but from the annotation and computational analysis of the sequence. Having 3.2 billion bases of genomic sequence is nice, but knowing which gene is responsible for a disease and which bases differ between the normal and disease-causing alleles has direct practical medical value. Therefore, a considerable part of the effort in genomics research has shifted from high-speed sequencing to the analysis of genome data. One major goal of this analysis and annotation effort is to produce an accurate and comprehensive list of human genes and proteins.

Prior to large-scale genome sequencing, most genes were studied by reverse transcription of mRNA into complementary DNA (cDNA), which is then cloned and sequenced. Following the principles of the "Central Dogma," it was assumed that each gene produces one kind of mRNA transcript, which can be captured

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

from the cDNA made from the mRNA of cells that express high levels of the corresponding protein. This approach has worked well, and it has been validated many times by inserting the cloned cDNA sequences into other cells (often bacteria) and demonstrating the production of the target protein. The sequencing of cDNAs was expanded in the 1990s by many expressed sequence tag (EST) projects, where entire libraries of cDNAs from many different tissues were sequenced in bulk without prior knowledge of their protein products, and the sequences were deposited in GenBank as "hypothetical genes" (Adams et al. 1991). GenBank now contains more than 7.9 million human ESTs (NCBI 2007).

The basic biology of eukaryotic gene transcription is understood in some detail. For protein-coding genes, RNA polymerase II and a complex of transcription factor proteins binds to a promoter element on the DNA. The most common promoter sequence is the TATAAA box (Goldberg–Hogness box), which is located approximately 25 bases 5′ of the transcription start site (Lifton et al. 1978). RNA polymerase II synthesizes a primary transcript (pre-mRNA), which is an RNA molecule that is complementary to the template strand of the DNA. This primary transcript can be many tens of thousands of bases long. Both 5′ capping and 3′ polyadenlyation are RNA processing steps that are tightly coupled with the transcription process. Soon after transcription begins, the 5′ end of the transcript is capped by the addition of a single 7-methylguanine nucleotide in an unusual 5′–5′ triphosphate linkage (see Figure 13.1). In vertebrates, the first two nucleotides of the mRNA are also methylated at the 2′-hydroxyl of the ribose.

As the RNA polymerase II reaches the 3′ end of the pre-mRNA, a polyadenylation signal with the sequence AAUAAA indicates the end of the gene. A set of cleavage factor proteins recognizes the AAUAAA signal and cuts the pre-mRNA at a location approximately 35 bases further downstream, releasing the RNA polymerase II. Then the enzyme polyadenylate polymerase

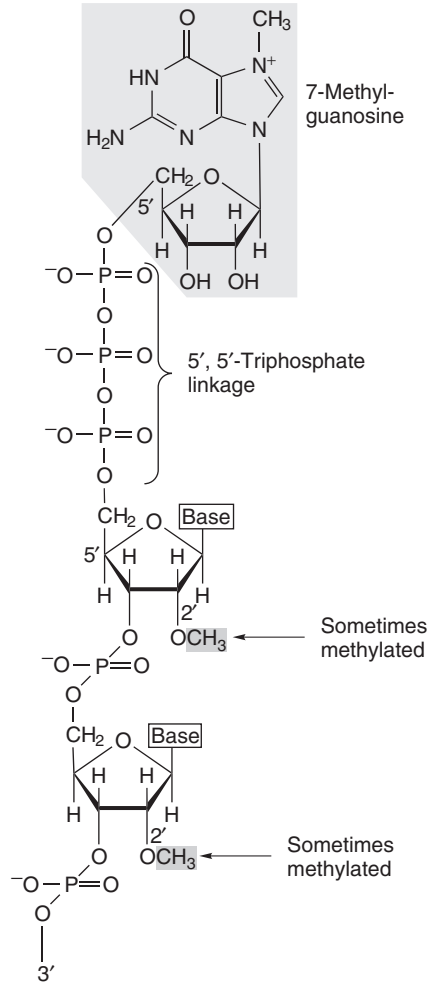


FIGURE 13.1. The 7-methylguanosine cap is linked by a 5′, 5′ triphosphate bond to the 5′ end of the mRNA. The first and second bases may also be methylated at the 2′ position.

(PAP) adds adenosine residues to form a polyA tail 50–250 bases long (see Figure 13.2).

After addition of the 5′ cap and the 3′ polyA tail, the pre-mRNA is further processed by removal of intron sequences via a splicing process controlled by a complex of proteins and small

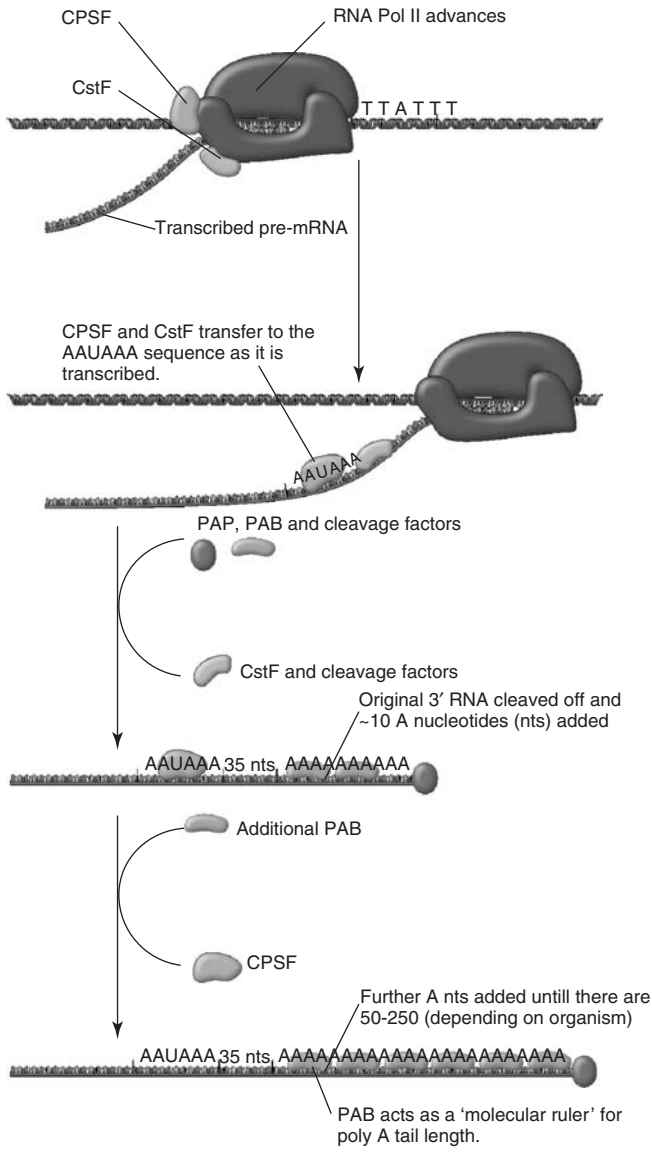


FIGURE 13.2. Addition of the polyA tail is initiated by recognition of AAUAAA signal sequence by cleavage factors. Then polyadenylate polymerase (PAP) adds adenosine residues to form a polyA tail 50–250 bases long. (See insert for color representation.)

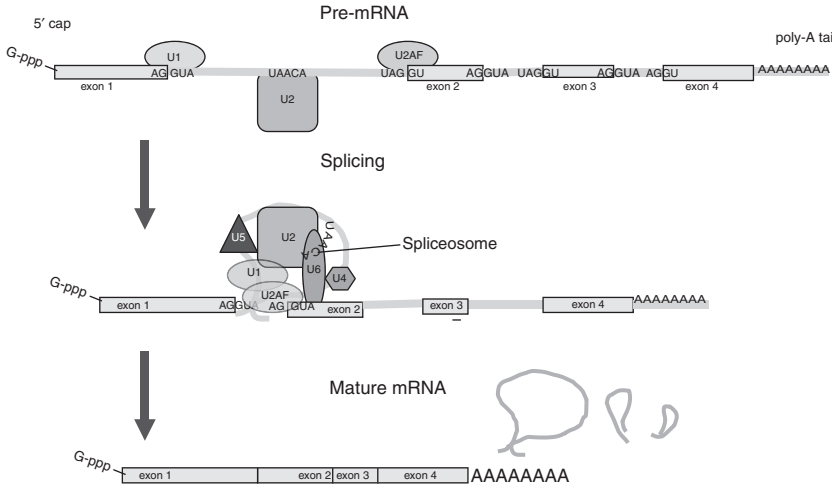


FIGURE 13.3. The pre-mRNA transcript is bound by multiple splicing factors (U1, U2, U2AF, etc.) that recognize specific splicing sequences within the intron and at the intron–exon boundaries. The splicing factors form a spliceosome that catalyzes the excision of intron sequences and joins the exons in a mature mRNA molecule.

RNA molecules known as the **spliceosome**. In mammalian genes, the boundaries of the introns are defined by a consensus sequence of AG-|-*GT* at the 5′ end and *CAG*-|-GT at the 3′ end. The splicing factors recognize and bind to these sequences and initiate a process that leads to cleavage of the RNA molecule precisely at the intron boundary (see Figure 13.3).

On average, human genes have 4 introns, but many genes have dozens of introns and some genes have more than 100. Most introns are 75–100 bases long, but very rarely shorter than 40 bases, and some extremely long introns have been observed that contain more than 100,000 bases. Intron sequences may constitute 90% or more of the total length of the pre-mRNA (Stamm et al. 2005) (see Figure 13.4). Despite the extremely small size (and therefore low information content) of the splice recognition sites, cleavage of the mRNA must be very precise, since the addition or deletion of a single extra base would lead to a frameshift in

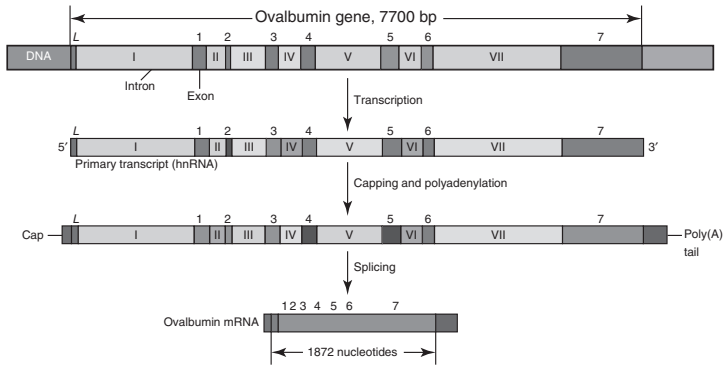


FIGURE 13.4. The ovalbumin gene has 7 introns, which make up more than 75% of the pre-mRNA transcript. (From T.A. Brown, *Genomes*. Copyright 1999 John Wiley and Sons, Inc. All rights reserved.)

the protein-coding sequence, and produce a totally nonfunctional protein. Incorrectly spliced transcripts may also be targeted for rapid degradation through the nonsense-mediated mRNA decay pathway.

The process of making cDNA from mRNA in the laboratory is error-prone. In particular, the reverse transcriptase enzyme uses a polyT primer, which binds to the polyA tail of the mRNA, and the sequence is copied from the 3′ end back toward the beginning (see Figure 13.5). The mRNA molecules are fragile, so they may be broken during the isolation and cloning process. Therefore it

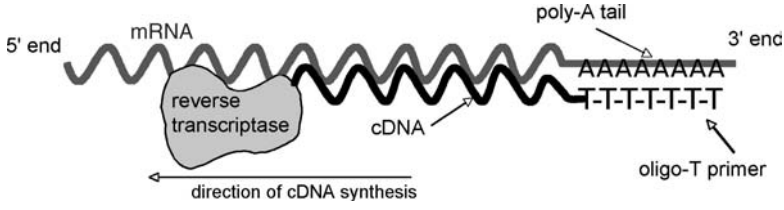


FIGURE 13.5. Complementary DNA is synthesized by the reverse transcriptase enzyme from the 3′ end of a mRNA molecule back toward the 5′ end using a primer complementary to the polyA tail.

is expected that the collection of ESTs in GenBank will contain some sequences that are truncated at the 5′ end or are incomplete.

Many biologists regard a cDNA as full-length when it contains the complete protein coding region, specifically, up to the ATG start codon. However, transcription of genomic DNA into RNA begins some distance upstream of the ATG codon, so a real full-length cDNA must reach up to the transcriptional initiation point and contain a 5′-untranslated region (5′-UTR). A Japanese group led by Suzuki and Sugano (2003) have developed a method for reliably sequencing the 5′ end of cDNAs. When they looked at the well-annotated human RefSeq genes in GenBank, they found that 34% of the sequences did not extend to the true 5′ end of the transcript. It is important to identify the true 5′ end of each gene, because the sequence immediately upstream of the transcription start site is the promoter region, which contains binding sites for RNA polymerase and transcription factors, which regulate the expression of the gene. The 5′ UTR also contains sequences that are recognized by posttranscriptional regulators.

The UniGene database at NCBI (`www.ncbi.nlm.nih.gov/UniGene`) represents one attempt to organize the huge collection of EST sequences into sets that correspond to genes. Comparisons among the sequences of ESTs for a single gene (overlapping fragments of different lengths) show a number of inconsistencies. Not only do the cDNA sequences have different 5′ ends; they also have different 3′ ends— some are short internal fragments of others, and some sequences seem to be missing chunks or have additional chunks in the middle (see Figure 13.6). The portions of cDNAs that were commonly missing often correspond to entire exons, the extra chunks are introns, or portions of introns. The EST sequences indicate that some mRNAs have different patterns of intron splicing (alternative splicing)—missing an entire internal exon, failing to splice out an introns, or splicing at different places, creating an mRNA sequence that includes a bit of extra

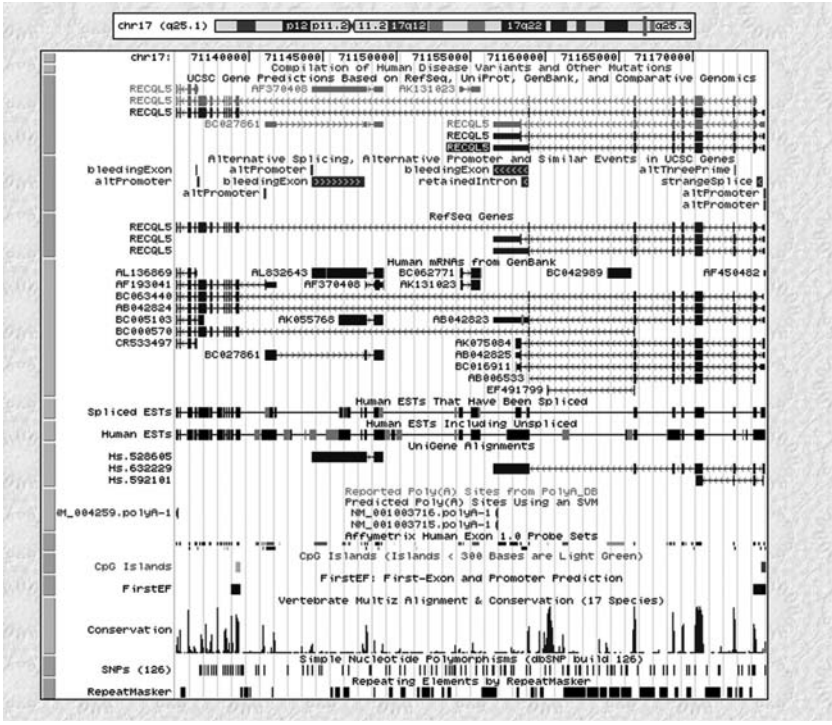


FIGURE 13.6. The UCSC Genome Browser shows multiple mRNAs and multiple predicted genes at the RECQL5 locus due to alternative splicing.

intron or a bit of missing exon (see Figure 13.7). Similarly, the transcription initiation and polyadenylation sites are also variable. Some genes (or many genes) have alternative sites to start and end transcription.

Interestingly, some of these alternative transcripts lead to the production of viable proteins. In fact, some alternative transcripts encode protein products with functions that are different from, or even opposite to, functions of the "standard" protein. The products of many genes involved in apoptosis are alternatively spliced, and this splicing can result in the synthesis of isoforms that antagonize each other by having promoting versus inhibitory effects on apoptosis (Wu et al. 2003). So, if multiple protein-coding

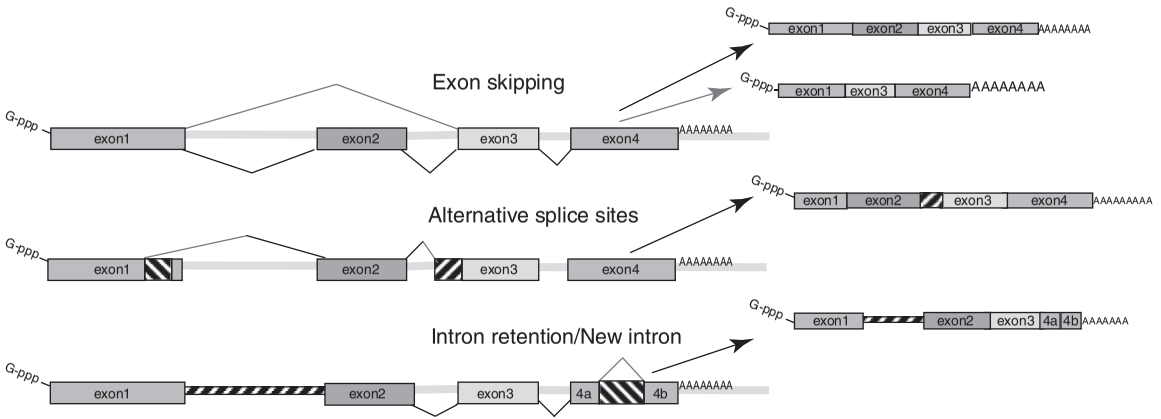


FIGURE 13.7. Alternative splicing can produce many different transcripts from a single pre-mRNA by exon skipping, use of alternative splice sites to extend or shorten introns, or the creation of novel splice sites within exons.

transcripts can be produced from the same region of DNA, then how can we define a gene so that a clear and comprehensive list of genes, transcripts, and proteins can be enumerated for the human genome? Recently published estimates of the number of human genes that are alternatively spliced vary from 38% (Brett et al. 2000) to 70% (Kalnina et al. 2005). This is understandable, since the techniques for the detection of alternately spliced transcripts are unreliable, and alternative splicing of each transcript may respond to different signals that are related to tissue specificity, developmental stage, disease process, or various types of cellular stress.

Currently, the best-annotated collection of human genes is the RefSeq database, created by the curators of GenBank at NCBI. The RefSeq database contains 22,458 transcripts that can be mapped to the human genome sequence (NCBI 2007). Some of these transcripts map to the same locus—representing alternative splicing, or genes contained entirely inside introns of other genes. When all of the transcripts that map to the same locus are merged, a total of 15,783 nonredundant genes is produced (Nakaya et al. 2007).

Exon Arrays

One weakness of the gene expression microarrays designed by Affymetrix and other vendors is that they do not detect alternative splicing. These arrays are composed of sets of oligonucleotide probes for each gene, but these probes are clustered in a few exons at the 3′ end of the gene. This makes sense, since the labeling reaction used to prepare cellular RNA for hybridization with the probe array uses reverse transcriptase with a polyTTTT primer. The reverse transcription reaction starts at the 3′ end of mRNA molecules (the polyAAAA tail) but does not always run to completion, so many of the labeled molecules will be truncated. These 3′-biased probe designs cannot detect

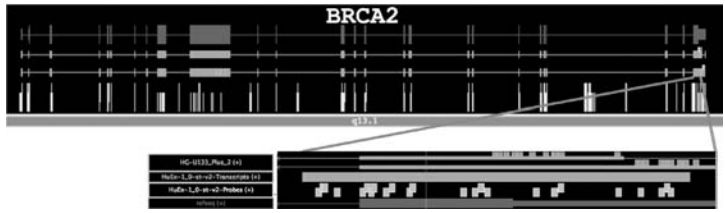


FIGURE 13.8. Affymetrix expression probes (HG-U133_Plus_2) and Exon probes (HuEx-1.0) for the human BRCA2 gene. The diagram also shows three different predictions for the transcription start site and the RefSeq gene model. (See insert for color representation.)

changes in transcription start sites or other types of alternative splicing. A more comprehensive picture of gene expression can be constructed by creating microarray probes for each gene by placing probes in every exon supported by EST data or predicted by annotation software. Affymetix has designed a human genome exon array (GeneChip® Human Exon 1.0 ST Array) that can be used to validate gene models and investigate alternative splicing, which may play a role in some diseases (see Figure 13.8). The Human Exon 1.0 ST array contains 5.3 million 25-base oligonucleotide probes, organized into 1.4 million probe sets (putative transcripts).

Interpretation of data from an exon array are very challenging. The possible variations in alternative splicing, transcription initiation, and termination are almost limitless. Gene models such as RefSeq or Ensembl rely on data from sequenced mRNAs and ESTs in GenBank plus some computational prediction of consensus sites for promoters, exon splice sites, and transcription initiation and termination sites. A transcript may actually start or end upstream or downstream of the predicted sites, in sequence annotated as intron, exon, or untranscribed. Alternative splicing may involve simple exon skipping, failure to splice out introns, the movement of splice sites within introns and exons, or additional exons inside predicted introns or up and downstream from the predicted transcript. Fundamentally, any set of exon-specific

probes reflects a gene model for the location of introns and exons, and it cannot identify an infinite number of splice-form variants.

Medical Applications of Alternative Splicing

The medical implications of alternative splicing are extremely far-reaching. Many genetic diseases are caused by mutations that affect splicing rather than causing changes in the protein-coding sequence. These aberrantly spliced transcripts produce proteins with changed functions, or the transcripts are degraded and no protein is produced.

The most frequent cause of β-thalassemia, which is one of the most common human genetic diseases, is mutations in introns 1 and 2 of the β-globin gene, which result in aberrant splicing of the mRNA. Other examples of human disease genes known to have mutations in splicing control elements include BRCA1 (breast cancer 1, early onset), CFTR (cystic fibrosis transmembrane conductance regulator), HPRT1 (hypoxanthine phosphoribosyltransferase 1), MAPT (microtubule-associated protein tau), and SMN1 (survival of motor neuron 1) (Cartegni et al. 2002). Systematic analyses of transcripts from specific disease genes such as NF1 (neurofibromatosis type 1) and ATM (ataxia telangiectasia mutated) lead to the striking conclusion that as many as 50% of disease mutations in exons may effect splicing (Blencowe 2006).

Genetic diseases that affect proteins involved in the spliceosome or in the regulation of splicing will have effects on the expression of many genes, which may have pleiotropic effects that may be expressed in a tissue-specific or developmentally regulated manner. Nova1, a splicing regulatory protein that is active in neurons, is associated with abnormal motor inhibition in paraneoplastic opsoclonus myoclonus ataxia (POMA). Nova1 is essential for postnatal motor neuron survival, where it binds mRNA in a sequence-specific manner to regulate neuron-specific

splicing of inhibitory receptor pre-mRNAs. Ule et al. (2003) found that Nova1 binds to a specific recognition signal on pre-mRNAs. Of 34 transcripts that were identified as targets of Nova1, three-quarters of these encode proteins that function at the neuronal synapse, and one-third are involved in neuronal inhibition.

Spinal muscular atrophy (SMA) is a disorder characterized by progressive loss of spinal cord motor neurons resulting in paralysis due to the deletion of the survivor of motor neuron gene (SMN1). Deletion of SMN1 prevents assembly of the U1 ribonucleoprotein complexes in the cytoplasm and results in global defects in pre-mRNA splicing (Rossoll et al. 2002). Mutations in the MeCp2 gene are associated with the symptoms of Rett syndrome (slowed brain and head growth, gait abnormalities, seizures, and mental retardation) as well as other disorders such as autism and a variety of X-linked mental retardation syndromes. Loss of MeCp2 function can cause dramatic changes in alternative splicing in neurons.

Changes in alternative splicing patterns have been observed in many types of cancer, and many cancer-associated genes are regulated by alternative splicing. Loss of fidelity, variation of the splicing process, and even controlled switching to specific splicing alternatives may occur during tumor progression and could play a major role in carcinogenesis. Splice variants that are found predominantly in tumors have clear diagnostic value and may provide potential drug targets. More recent studies suggest that splice-form specific assays may provide more informative diagnostic biomarkers than do gene expression technologies that ignore alternative splice forms.

Progression of prostate cancer from an androgen sensitive to androgen insensitive tumor is accompanied by a change in alternative splicing of fibroblast growth factor receptor 2 (FGF-R2). This change results in loss of the FGF-R2(IIIb) isoform and predominant expression of the FGF-R2(IIIc) isoform.

Zhang et al. (2006) observed differential expression in prostate cancer versus normal prostate of splice forms for a number of known prostate cancer-associated genes as well as other marker candidate genes, which span a wide spectrum of biological functional roles, including signal transduction (SIM2 and CDC42BPA), extracellular matrix, and cytoskeleton (CD44, MAPT and ILK). Others appear to be involved in, for example, epidermal differentiation and proliferation (KRT15, IGF1, PGR and HPN), cell growth and development (FGFR2), apoptosis (DBCCR1 and CLU), and lipid metabolism (AMACR). Several genes encoding splicing factors, such as U2AF1, U2AF2, and DHX34, which are themselves alternatively spliced, also show significant differential expression of splice forms in prostate cancer. This is consistent with the observation that the expression of many splicing factors is deregulated in tumors.

Now that high-throughput technology exists for investigating alternative splicing across large numbers of genes and splice variants (splice-variant or exon microarrays), many more experiments are underway to detect the role of splice variants in various diseases. A screen of brain tissue from patients with mesial temporal lobe epilepsy and Alzheimer's disease, with an array designed to detect 1665 possible alternative splicing events, found a total of 221 splicing changes that were identified as statistically significant. These changes were found to exhibit unique and consistent patterns within the disease groups.

Some studies of alternative splicing have found that the majority of genes that have differentially expressed splice forms under a specific pair of conditions (different tissues, cancer versus normal, stress response, etc.) do not show differential expression of overall mRNA levels under those same conditions. This, combined with the concept of common recognition sites in groups of pre-mRNAs that are specifically recognized by splice-modifying proteins, suggests that alternative splicing could exist as a post-transcriptional regulatory system for gene expression. Regulation of gene expression by alternative splicing may work in concert, or

independently from the more thoroughly studied world of promoters and transcription factors. Regulators of splicing become potential drug targets that may be capable of increasing or decreasing the expression of a specific gene, or a coordinated set of genes that share a splicing factor recognition site.

REFERENCES

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde RF, Moreno RF. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**(5013):1651–1656.

Blencowe BJ. 2006. Alternative splicing: new insights from global analysis. *Cell* **126**(1):37–47.

Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, Bork P. 2000. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett* **474**:83–86.

Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* **3**(4):285–298.

Kalnina Z, Zayakin P, Silina K, Line A. 2005. Alterations of pre-mRNA splicing in cancer. *Genes Chromosomes Cancer* **42**(4):342–357.

Lifton RP, Goldberg ML, Karp RW, Hogness DS. 1978. The organization of the histone genes in Drosophila melanogaster: Functional and evolutionary implications. *Cold Spring Harb Symp Quant Biol* **42**:1047–1051.

Nakaya HI, Amaral PP, Louro R, Lopes A, Fachel AA, Moreira YB, El-Jundi TA, da Silva AM, Reisand EM, Verjovski-Almeida S. 2007. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biol* **8**(3):R43.

National Center for Biotechnology Information (NCBI). 2007. Expressed Sequence Tags Database (`http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html`, accessed on 5/1/2007).

Rossoll W, Kroning AK, Sendther M, et al. 2002. Specific interaction of Smn, the spinal muscular atrophy determining gene product, with hnRNP-R and gry-rbp/hnRNP-Q: a role for Smn in RNA processing in motor axons? *Hum Mol Genet* **11**(1):93–105.

Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H. 2005. Function of alternative splicing. *Gene* **344**:1–20.

Suzuki Y, Sugano S. 2003. Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method. *Meth Mol Biol* **221**:73–91.

Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. 2003. CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**:1212–1215.

Wu JY, Tang H, Havlioglu N. 2003. Alternative pre-mRNA splicing and regulation of programmed cell death. *Prog Mol Subcell Biol* **31**:153–185.

Zhang C, Li HR, Fan JB, Wang-Rodriguez J, Downs T, Fu XD, Zhang MQ. 2006. Profiling alternatively spliced mRNA isoforms for prostate cancer classification. *Bioinformatics* **7**:202.

CHAPTER

14

Genome Tiling Chips

Genome Chips

Genomics is a technology-driven field. The innovative technology of the microarray is the idea of an ordered array of DNA probes that correspond to specific genes, arranged in a grid on a solid support (i.e., a glass slide), and hybridized with a solution containing a labeled mix of RNA from an experimental organism. While the first microarrays were created in order to answer specific biological questions about gene expression in *Arabidopsis* (Shena et al. 1995) and yeast (DeRisi et al. 1997), the technology was quickly adopted for many other applications. Once the technology was available, many investigators were able to use it for open-ended data-gathering experiments with loosely defined hypotheses such as "What happens to the expression of thousands of genes in my favorite organism/tissue when I apply treatment X?" or "What changes in gene expression can be observed between diseased and healthy tissue?"

The popularity of microarray experiments among researchers has led to competition among several manufacturers to develop microarray chips that contain larger numbers of individual DNA

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

probes. As the manufacture of microarrays has become more sophisticated, new applications have been developed. In addition to measuring gene expression by hybridization of RNA to the chip, these high-density chips can be used to investigate genomic DNA. With chips that contained tens of thousands of probes, microarrays were designed for resequencing sections of genomic DNA, such as genes with many highly polymorphic mutations (i.e., cytochrome P450), or scanning many single-base mutations (SNPs) scattered across the entire genome. When chips were developed with hundreds of thousands of probes (by manufacturers including Affymetrix Inc., Nimblegen Inc., Perlegen Inc., Agilent Inc.), it became possible to create "tiling arrays" that contain probes that span entire genomes.

RESEQUENCING CHIPS

There are a number of ways to design a genome tiling array. In order to determine the exact nucleotide sequence of a section of DNA, the tiling array must have a set of fully overlapping probes, with every possible variant at every position—four probes for every nucleotide of DNA (see Figure 14.1). The accuracy of such

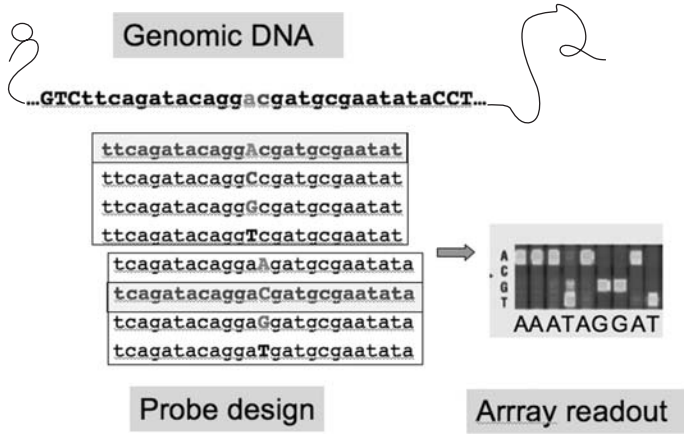


FIGURE 14.1. Probes for sequencing genomic DNA on a microarray.

sequencing arrays is extremely high, similar to the 99.9% standard for automated fluorescent sequencers using modified Sanger sequencing methods. Even the most advanced current technologies do not have the ability to put the entire human genome on a chip at this density (~12 billion probes). However it is possible to fully sequence a bacterial or viral genome (a few million probes). The current (2007) generation of Affymetrix CustomSeq™ arrays can sequence up to 300,000 bases of genomic DNA on a single chip. Sequencing arrays have been used to genotype a variety of pathogens such as the SARS virus, smallpox, and anthrax (*Bacillus anthracis*). This sequence information allows clinicians to relate pathogen subtypes to drug response and patient outcomes, and then to apply this information to diagnostic identification of pathogen strains. Rapid, inexpensive pathogen sequencing to identify species and strain is also very useful for epidemiology.

Whole-Genome Transcription Profiling

Another way to build a genome tiling array is to cover the entire genome with a set of oligonucleotide probes set end to end. It is also common to include probes for the complementary strand (i.e., antisense strand) of DNA, so for a stretch of 1 million nucleotides of genomic DNA, there would be 80,000 probes on the array. Arrays with gaps between the oligonucleotide probes provide nearly as much information as the continuous tiling design, but require fewer oligos on the chip (see Figure 14.2). These nonoverlapping types of array do not accurately determine the sequence of the DNA, but they can be used for several new applications. By hybridizing total cellular RNA to the array, it is possible to discover what parts of the genome are being transcribed (Mockler et al. 2005). The chief advantage of the genome tiling array over the standard microarray design is that it is unbiased. Rather than relying on the sequences of known genes (determined by labwork or predicted by computer

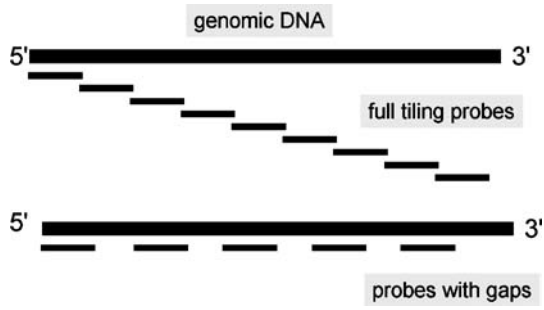


FIGURE 14.2. Probe design for genome tiling arrays, both with and without gaps.

models from genome sequence), it shows all regions of the genome that are transcribed. This would include the many types of small nonmessenger RNAs such as microRNAs, siRNA, and antisense RNA, as well as allowing for the discovery of many new genes. RNA hybridizes not only to probes for many "hypothetical" genes (previously predicted only by computational methods), but also to many regions that are conserved in comparative genomic analyses and also to new genomic regions not previously suspected of containing genes.

Data analysis for a genome tiling array is different from what is typically done on the standard oligonucleotide arrays designed from known expressed gene sequences (ORF arrays). The focus of tiling arrays can be on the gene expression of a single RNA sample, or on a comparative analysis of treated versus untreated samples. Tiling arrays contain more probes for each gene and different numbers of probes for genes of different sizes, and they also contain probes for introns, so it is not possible to create a single expression measure for a gene by a simple combination of values from all probes. Data analysis generally emphasizes common expression patterns in groups of adjacent probes (see Figure 14.3). It is also possible to use genome tiling arrays to identify alternative splicing of genes by comparison of RNA samples from different tissues or different treatments to identify intragenic segments

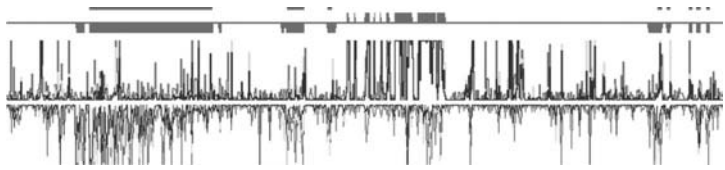


FIGURE 14.3. A portion of the data display for a whole-genome tiling array. (See insert for color representation.)

that show differential expression in the RNA from one sample to another. Since genome tiling arrays are generally built with regularly spaced probes, it is not possible to adjust the probes to have constant G+C/A+T content, so hybridization will be weaker to probes with low GC content.

A complete genome tiling array for the *Arabidopsis* plant (a lab variety of a wild mustard weed) genome shows that up to 10 times more of the genome is transcribed into RNA than can be accounted for by the known genes (Yamada et al. 2003). A human genome tiling array has identified over 10,000 new transcribed regions in the genome. Transcription appears to be active in introns (possible alternate exons), in regions between genes, and on the opposite strand of DNA from known genes (creating RNA molecules known as antisense transcripts). Up to 90% of the genome locations that bind RNA do not correspond to the sequences of known genes (Kapranov et al. 2002). Some of these novel transcribed regions have been validated by RT-PCR (Rinn et al. 2003).

Stolc et al. (2004) created a whole-genome microarray for *Drosophila* that included 180,000 36-mer oligonucleotide probes. This array included probes for all predicted exons, and probes tiled throughout the predicted intronic and intergenic regions of the genome. Their expression results (using mRNA collected from six stages of the fly lifecycle) showed transcription from 77% of all exon probes and 41% of all nonexon probes (introns and intergenic regions). A set of splice junction probes also indicated that 53% of *Drosophila* genes with more than one exon

exhibit exon skipping (alternate splicing) and that these alternate splice forms are differentially expressed at various lifecycle stages. They concluded:

> It is clear that our past understanding of genome-wide RNA transcription has been very limited, because a large proportion of exons show dynamic patterns of differential splicing, and noncoding activity is ubiquitous. Our results indicate that there are thousands of uncharacterized and unannotated transcripts expressed in a developmentally coordinated manner . . . there is considerably more complexity in gene and transcript regulation than was previously known.

In 2006, Li and coworkers built a custom tiling microarray for the rice plant that contained 13 million individual 36-mer oligonucleotide probes tiled throughout the nonrepetitive sequence of the genome on 34 chips (Li et al. 2006). The rice genome produced a similar transcriptional profile as *Arabidopsis.* The array confirmed transcription from 82% of previously annotated genes and identified an additional 5464 probable new genes as transcriptionally active regions that show homology to other plant proteins. Also, 24% of genes showed significant antisense transcription. This study used probes prepared from cDNA, so it did not detect nonmessenger RNAs.

Using a similar method, Bertone et al. (2004) created a genome tiling array for the human genome including a total of 52 million 36-mer oligonucleotide probes. The probes were spaced at an average of every 46 nucleotides on both strands of 1.5 Gb (gigabases) of nonrepetitive human DNA (half of the genome, leaving out the centromeres, telomeres, and highly repetitive regions). These probes were arrayed on 134 chips at a density of 390,000 probes per chip. Using RNA from liver tissue, they detected approximately 10,500 new transcribed sequences that did not correspond to any previously annotated exon. Almost 6000 of these new transcribed sequences were located more than 10 kb (kilobases) from any previously annotated gene.

While this whole-genome expression technology is not yet directly relevant to patient care, it does point out how important

genomics technology has become as a driving force for fundamental discoveries in biology. When transcription is investigated on the entire genome at high resolution, it is discovered that the entire process of transcription needs to be redefined. Genome tiling arrays are a very powerful tool for molecular biology—somewhat akin to a new telescope for an astronomer that is 100 times more powerful than anything previously available. New and unexpected discoveries are produced in almost every experiment.

ChIP-chip

Genome tiling chips have also been used to identify the genomic targets (binding sites) for transcription factors and other DNA-binding proteins. This method, known as **chromatin immunoprecipitation on DNA chips** (ChIP-chip), has also been used to study the interaction of replication and recombination proteins with DNA, as well as chromatin structure. Cells are grown under a specific set of conditions, then treated with formaldehyde, which creates crosslinks between the DNA-binding protein and the genomic DNA. Then the DNA is sheared into small fragments (1 kb or smaller) and DNA–protein complexes containing the protein of interest are purified by immunoprecipitation with an antibody for a specific DNA-associated protein. The DNA fragments are released from the protein crosslinks, then labeled and applied to a whole-genome tiling array (see Figure 14.4). The array identifies protein binding targets by showing hybrdization to probes at specific genomic locations that were bound by the protein (Buck and Lieb 2004).

Before high-density whole-genome tiling arrays were available, ChIP-chip experiments were conducted using arrays of probes designed to match "upstream elements," which are genomic sequences in the putative promoter region spanning ~1000 bp 5′ to the predicted transcriptional start site for known genes. This type of targeted array suffers from the poor quality of

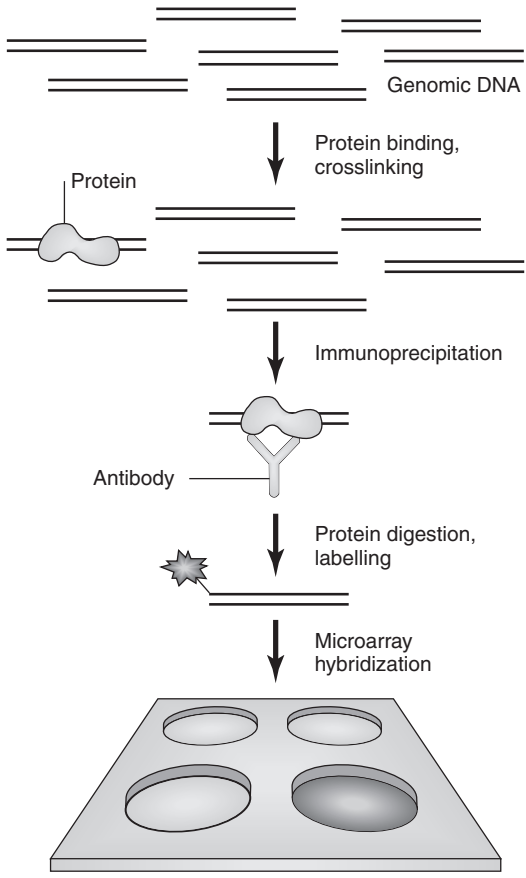


FIGURE 14.4. Diagram of ChIP procedure. Protein is bound to fragments of genomic DNA, then antibody captures the protein–DNA complex. DNA is released from protein, labeled, and identified by hybridization to the genome tiling microarray. (Reprinted with permission from Hoheisel 2006, *Nature Reviews Genetics* **7**:200–210.)

annotation in current genome databases. While transcription start sites have been accurately mapped for most yeast genes, this information is not well established for humans and other animals or plants. Furthermore, the targeted array does not detect DNA–protein interactions outside the 1-kb upstream region, nor does it detect transcription factors that affect previously unknown genes. The use of whole-genome tiling arrays allows the

ChIP-chip experiment to produce an unbiased readout of all binding locations for a specific protein on the genome.

Data analysis for ChIP-chip is problematic for several reasons. First, any immunoprecipitation method has a significant amount of false positives due to nonspecific binding of the antibody. DNA-binding proteins also show some nonspecific binding (i.e., association with nontarget sites on the DNA).

ARRAYCGH

Another application of genome tiling arrays is to identify changes in copy number for genomic segments, known as **comparative genomic hybridization** (CGH). Gene amplifications, deletions, and translocations are very common in many types of cancer and can be diagnostic for particular cancer subtypes. It is important to keep in mind that a tumor is a clone of a mutant cell, and it has a genome that is different than the normal cells of the same person. CGH utilizes the hybridization of differentially labeled tumor and reference DNA to generate a map of DNA copy number changes in tumor genomes (Kallioniemi et al. 1992). Many cancers have mutations that affect DNA repair enzymes, so they tend to accumulate many insertions, deletions, and point mutations. CGH, using quantitative two-color fluorescence *in situ* hybridization of genomic DNA preparations from tumor and normal tissue samples to a normal metaphase karyotype preparation, has been a standard method for cytogenetic analysis since the early 1990s (du Manoir et al. 1993).

The resolution of traditional CGH based on hybridization to whole chromosomes is about 10–20 Mb (megabases; million base pairs). This is sufficient to detect deletions and reorganizations of sizeable segments of chromosomes, but will not detect microamplifications and deletions that may affect single genes involved in disease (i.e., oncogenes and tumor suppressor genes). In 2004, Ishkanian and coworkers created a spotted tiling array

for the complete human genome using a set of 32,000 overlapping BAC clones of approximately 1 Mb in size (Ishkanian et al. 2004). Using this BAC array, many previously undetected gene copy number changes were discovered, including an amplification in lung cancer cell line H526 of a 1.3 Mb fragment at 2p24.3, which contains the MYCN oncogene. Analysis of colorectal cancer cell line COLO320 identified four new microamplifications on chromosome arms 13q, 15q, 16p, and 22q. A 240-kb deletion was detected in the breast cancer cell line BT474, containing genes PRKAR2B, a regulatory kinase, and HBP1, a G1 inhibitory kinase.

A similar method, known as **loss of heterozygosity** (LOH), has also been widely used in oncology. LOH is based on quantitative PCR amplification of specific segments of genomic DNA or the use of polymorphic genetic markers such as microsatellites or SNPs. A critical limitation of LOH studies is that it can detect deletions only at predetermined loci, rather than providing a full-genome scan like CGH. LOH studies generally focus on known tumor suppressor genes, such as the Rb1 gene, the deletion of which is known to lead to retinoblastoma.

Loss of heterozygosity is often used as an example to support the "two-hit model" for cancer (Knudson 1971). In order for specific types of cancer to develop, a certain key "cancer suppressor" gene must be inactivated. For hereditary cancers, one parent may carry a mutant allele of the cancer suppressor gene, then a child who inherits this mutant allele will be heterozygous in all somatic cells. A single mutation or **deletion** in the wild-type copy of the gene in any somatic cell will be a LOH event, which creates a null mutant for the tumor suppressor gene. This mutant cell may now be able to proliferate without the usual genetic controls on cell division, creating a clone of tumor cells. Other mutations may occur in these growing cells that lead to the aggressive behavior of malignant cancer. Hereditary cancers of this type often occur in multiple locations and at an early age, since a single mutation event in a single cell has a high chance of occurrence. This type of

mutation in a tumor suppressor gene is considered to be a dominantly inherited predisposition to cancer, since a single copy of the mutant allele creates the phenotype.

A whole-genome oligonucleotide tiling array allows CGH/LOH analysis at a resolution of 25–100 bp, which is much finer than CGH based on hybridization to whole chromosome preparations or BAC clones. At this resolution, deletions and amplifications can be detected that affect only a single exon within a gene (Pinkel et al. 1998). ArrayGCH provides copy number information about the entire genome rather than just selected markers (see Figure 14.5). Each oligonucleotide probe gives an independent value for the copy number of that locus on the chromosome. Array-based CGH/LOH studies could be done more rapidly and produce data that can be more easily interpreted by the physician than classic cytogenetic methods. The principal drawback to widespread use of genome tiling arrays for cancer diagnosis is the very high cost of high-resolution genome tiling chips.

Genomic amplifications and deletions were previously thought to occur primarily in cancer cell lines, but when normal

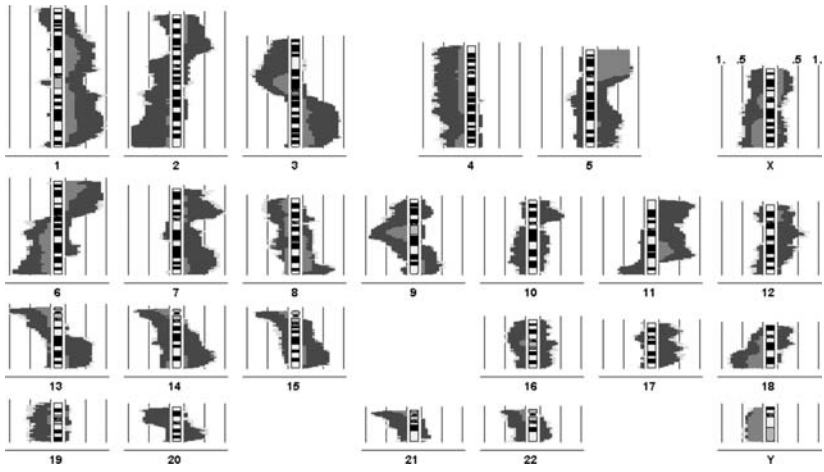


FIGURE 14.5. A display of ArrayCGH data spanning the entire human genome. (See insert for color representation.)

human cells were screened on genome tiling arrays, significant amounts of copy number variation between individuals were found throughout the genome in both coding and noncoding regions. ArrayCGH methods have also been applied to detect small deletions and amplifications in congenital conditions that do not exhibit known mutations (Bar-Shira et al. 2006), such as Ulnar mammary syndrome (Klopocki et al. 2006) and holoprosencephaly (Koolen et al. 2006). Differences in gene copy number may be as important as allelic differences in determining important medical phenotypes such as disease susceptibility and drug reactions.

References

Bar-Shira A, Rosner G, Rosner S, Goldstein M, Orr-Urtreger A. 2006. Array-based comparative genome hybridization in clinical genetics. *Pediatr Res* **60**(3):353–358.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**(5705):2242–2246.

Buck MJ, Lieb JD. 2004. ChIP-chip: Considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83**(3):349–360.

Cawley S, Bekiranov S, Gingeras TR et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**(4):499–509.

DeRisi JL, Iyer VR, Brown PO. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**:680–686.

du Manoir S, Speicher MR, Joos S, Schrock E, Popp S, Dohner H, Kovacs G, Robert-Nicoud M, Lichter P, Cremer T. 1993. Detection of complete and partial chromosome gains and losses by comparative genomic *in situ* hybridization. *Hum Genet* **90**(6):590–610.

Hoheisel JD. 2006. Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet* **7**(3):200–210.

Ishkanian AS, Malloff CA, Lam WL et al. 2004. A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet* **36**:299–303.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**:818–821.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**(5569):916–919.

Klopocki E, Neumann LM, Tonnies H, Ropers HH, Mundlos S, Ullmann R. 2006. Ulnar-mammary syndrome with dysmorphic facies and mental retardation caused by a novel 1.28 Mb deletion encompassing the TBX3 gene. *Eur J Hum Genet* **14**(12):1274–1279.

Knudson, AG Jr. 1971. Mutation and cancer: Statistical study of retinoblastoma. *Proc Natl Acad Sci USA* **68**:820–823.

Koolen DA, Herbergs J, Veltman JA, Pfundt R, van Bokhoven H, Stroink H, Sistermans EA, Brunner HG, Geurts van Kessel A, de Vries BB. 2006. Holoprosencephaly and preaxial polydactyly associated with a 1.24 Mb duplication encompassing FBXW11 at 5q35.1. *J Hum Genet* **51**(8):721–726.

Li L, Wang X, Stolc V, Li X, Zhang D, Su N, Tongprasit W, Li S, Cheng Z, Wang J, Deng XW. 2006. Genome-wide transcription analyses in rice using tiling microarrays. *Nat Genet* **38**(1):124–129.

Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, Ecker JR. 2005. Applications of DNA tiling arrays for whole-genome analysis. *Genomic* **85**(1):1–14.

Pinkel D, Segraves R, Albertson DG et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20**(2):207–211.

Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M. 2003. The transcriptional activity of human chromosome 22. *Genes Dev* **17**: 529–540.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**:467–470.

Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE, Bussemaker HJ, White KP. 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306**:655–660.

Yamada K, Lim J, Ecker JR et al. 2003. Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302** :842–846.

CHAPTER

15

Cancer Genomics

Understanding Cancer Genomics

Cancer is a unique disease because it involves genetic changes in somatic cells. Cancers are groups of cells (tumors) that grow in an unregulated and aggressive fashion. This disorganized growth pattern would not be possible for all the cells of the body—so the genetic patterns found in cancer cells would be lethal in a gamete or an embryo.

Cancer is a disease involving a multistep process of changes in the genome of somatic cells. Mutations in the DNA of somatic cells occur frequently—many cell types, such as epithelial cells, grow constantly and millions of new cells are produced by mitosis every day. Most new mutations in somatic cells have no effect on protein sequences, gene expression, or cell phenotype. A few mutations do change critical proteins, and are lethal to the cell. These single-cell mutations seldom have an overall effect on the body since the mutated cells do not reproduce. However, somatic cell mutations that affect cell growth regulation and DNA repair enzymes can be dangerous. A mutation in a growth regulator may allow the cell to reproduce beyond its normal limits—failing to

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

sense the inhibitor signals from the surrounding tissue. A mutation in a DNA repair enzyme makes the cell (and its progeny) more likely to undergo new mutations. A combination of these two features of growth and high mutation rates can create a population of cells that have the emergent property of **evolution by selection**. Any new trait acquired by mutation in one cell will be passed on to its descendents, creating a clone, and traits that promote additional growth will lead to more cells carrying that trait. Over time, mutations can accumulate that lead to the malignant properties of cancer cells. Any growth regulators or toxic agents applied to the growing clone will create a selective pressure, allowing mutant cells that are resistant to the agent to grow.

If a mutation that destroys the function of a gene leads to cancer, then that gene is considered to be a **tumor suppressor**. Normal expression of this gene blocks cancer. Tumor suppressors include DNA repair enzymes and cell cycle regulators. If a mutation that increases the expression of a gene leads to cancer, then that gene is considered to be an **oncogene** (cancer-promoting gene). Oncogenes usually encode proteins that function as growth factors, growth factor receptors, signal transducers, transcription factors, or regulators of apoptosis.

Another level of complexity in the genetics of cancer in humans (and all other mammals) is the diploid genome. Every gene is present in two copies on the two homologous chromosomes (one inherited from each parent). So a single somatic mutation to a tumor suppressor gene, even in a critical growth regulator or DNA repair enzyme, is unlikely to produce much phenotypic effect. A normal copy of the gene will still be present on the other chromosome to provide the normal protein function. There are dominant mutations, which take effect when a single gene copy is changed, such as a critical regulator, enzyme, or pore protein locked in the ON position, or an oncogene that is overexpressed because of loss of a binding site for a negative regulator. For example, mutations in the *K-ras* gene, causing constitutive

activation of the signal-transducing function of the *ras* protein, have been found in about 30% of lung adenocarcinomas, 50% of colon carcinomas, and 90% of carcinomas of the pancreas (Minamoto et al. 2000).

The diploid nature of the genome means that a dangerous cancer-promoting mutation in a tumor suppressor gene can be inherited as a recessive allele, whose function is compensated by the normal copy of the gene on the other chromosome. A person with such a recessive mutation is vulnerable to any mutation that affects the normal copy of the gene in any somatic cell, creating a knockout for that gene's function. This knockout may be caused by a deletion of the normal copy of the gene, breakage of the chromosome disrupting the gene, or any single-nucleotide change that leads to defects in protein function or expression. Any of these mutation events that affect the single good copy of a gene are called **loss of heterozygosity** (LOH), because the cell is transformed from heterozygous for gene function to homozygous negative. LOH due to gene deletion can be detected with quantitative PCR amplification of specific segments of genomic DNA. Specific point mutations may be detected with SNP markers. A critical limitation of marker-based LOH methods is that they can detect deletions only at predetermined loci, rather than providing a full-genome scan. LOH studies generally focus on known tumor suppressor genes, such as the *Rb1* gene, the deletion of which is known to lead to retinoblastoma. Many heritable cancers are caused by this type of mutation in one copy of a tumor suppressor gene. The most common example is hereditary nonpolyposis colorectal cancer (HNPCC) syndrome; 80% of HNPCC cases have germline mutations in the *MSH2* or *MLH1* genes, both of which encode enzymes involved in DNA mismatch repair (Lynch and de la Chapelle 1999; Yan et al. 2000).

This genetic predisposition to cancer often leads to cancers that occur in multiple locations in the body, due to different instances of somatic cells developing mutations in the single

functional copy of the gene. The inheritance of cancer susceptibility genes reverses the usual concept of dominance and recessive gene function. A single defective copy of a gene that requires mutations that inactivate both homologous copies to produce its cancer promoting phenotype would be considered recessive from a molecular perspective, but the medical condition of hereditary predisposition to cancer is dominant (a person with one defective allele has the trait).

Copy Number Mutations

Cancer cells are often found to have large-scale changes to the structure of their chromosomes (cytogenetic aberrations). This is caused by mutations in the DNA repair genes and/or mutations in the quality control system that usually causes cells with abnormal chromosomes to self-destruct by apoptosis. Virtually all solid tumors, lymphomas, and leukemias have an abnormal karyotype. Entire chromosomes may be lost, or large pieces of a chromosome deleted. Alternately, regions of chromosomes may become amplified, increasing the copy number of some genes many fold. Translocations of large pieces of DNA from one chromosome to another, or inversion of a portion of a chromosome, are also common. The clonal nature of tumor growth allows a chromosomal rearrangement event in a single somatic cell to propagate into a population of cells. It is also common to find different karyotypes among the cells in a single tumor. Once again, these cytogenetic abnormalities would be lethal in a gamete or developing embryo, but it is possible for somatic cells with abnormal chromosomes to survive and reproduce.

Cytogenetic aberrations can lead to cancer in a number of ways. Amplification of an oncogene increases its expression, leading to more of its protein product in the cell. Amplification of members of the *myc*, *erb* (epidermal growth factor receptor), and *ras* gene families are found in a significant number of human

tumors. Approximately 30% of breast and ovarian cancers have amplification of the erbB2 (HER-2/neu) gene (Slamon et al. 1989). Deletion of a chromosome segment that contains a tumor suppressor gene can create a loss of heterozygosity event, or simply losing one copy of that gene may create a lower protein level that changes the regulation of important cellular pathways. Translocations and inversions can split the coding region of a gene, inactivating it, or combine a gene with a new regulatory region that either increases or decreases its expression in a particular cellular context. Some translocation or inversion events can create new combinations of coding regions between different genes, which produce a fusion protein with novel properties. The classic example of a translocation creating an oncogene by gene fusion is the Philadelphia chromosome, which is found in all cases of chronic myelogenous leukemia (CML). A translocation between chromosomes 9 and 22 in CML fuses the *c-abl* gene with the *bcr* gene. The *bcr/abl* fusion encodes a chimeric protein with a novel tyrosine kinase activity (Hermans et al. 1987).

The tyrosine kinase enzyme, known as BCR-ABL, which is created by *bcr/abl* fusion in CML acts as an oncogene, stimulating cell division in white blood cells and inhibiting the apoptosis pathway. Gleevec is a targeted cancer drug, designed to treat CML. Gleevec binds and inactivates the BCR-ABL protein. Interestingly, Gleevec (imatinib) has also been found to be effective as a therapy for gastrointestinal stromal tumors, where it inhibits a different tyrosine kinase that also has cancer-promoting properties. Gleevec is currently being tested as a therapy for other cancers, including small cell lung cancer, chronic mylomonocytic leukemia, and glioblastoma. While Gleevec is the poster child for rational drug design, its development and use is not truly based on genomic technologies. For CML, the only qualifying condition is the presence of the Philadelphia chromosome or the BCR-ABL protein. It may be possible to pre-qualify patients for Gleevec treatment for other types of cancer based on a protein or gene

expression screening for overproduction of other tyrosine kinase targets.

The cytogenetic analysis of cancer cells has benefited from a succession of molecular biology techniques. Initially, cancer cells were karyotyped by microscopic analysis of metaphase chromosomes in dividing cells. Quinacrine–mustard and trypsin–Giemsa stains create characteristic banding patterns in the chromosomes, which allows the visual identification of individual chromosomes and makes it possible to detect large-scale aberrations. Cytogenetics was improved by the use of **fluorescence *in situ* hybridization** (FISH), which uses labeled DNA probes to identify regions of each chromosome in a karyotype. Hybridization with these probes allows the detection of deletions, translocations, and chromosome duplications. However, FISH techniques require prior knowledge of sites of genomic instability in order to design probes to detect changes. A more general technique called **comparative genomic hybridization** (GCH) uses the complete genome as a labeled probe and a normal metaphase karyotype preparation mounted on a microscope slide as the target. Genomic DNA from normal cells is labeled in one fluorescent color, and genomic DNA from tumor cells is labeled in another fluorescent color; then both are mixed and hybridized to the target chromosomes. Regions of the target chromosomes that show a difference in fluorescence between the two colors represent a difference in copy number between the tumor genome and the normal genome, showing both deletions and amplifications (see Figure 15.1). CGH has been a standard method for cytogenetic analysis since the early 1990s.

The resolution of CGH based on hybridization to whole chromosomes is about 10–20 Mb (million base pairs). This is sufficient to detect deletions and reorganizations of sizable segments of chromosomes, but will not detect microamplifications and deletions that may affect single genes involved in disease (i.e., oncogenes and tumor suppressor genes). More recent advances

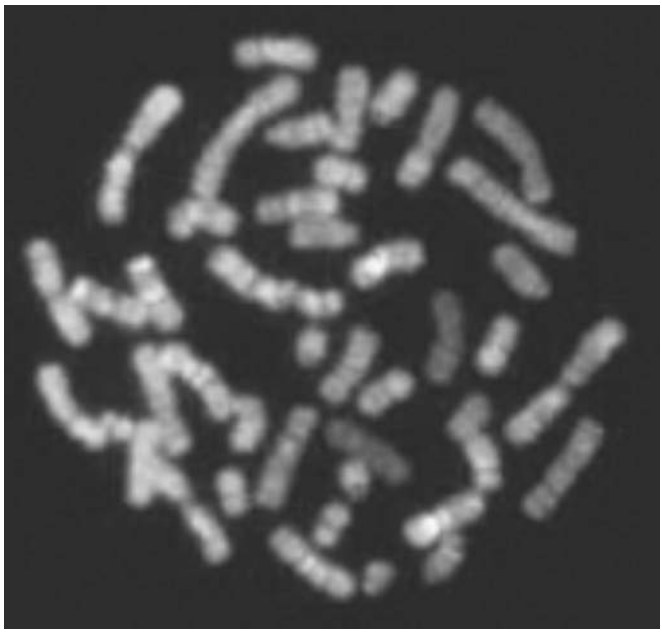


FIGURE 15.1. Hybridization of fluorescent tagged genomic DNA from normal (green) and tumor (red) cells to a chromosome squash from a normal cell. Green regions represent deletions and red regions, amplifications. (See insert for color representation.)

in microarray technology have made it possible to survey the entire genome at very high resolution on a genome tiling chip (arrayGCH), so deletions and amplifications of regions as small as 10,000 bases (10 kb) can now be detected (see Figure 15.2). As arrayCGH chips become less expensive, copy number analysis is a rapidly growing area of genomics that may find applications beyond cancer.

Changes in relative gene copy number detected using CGH may be associated with oncogene amplification or loss of tumor-suppressor gene function. However, CGH cannot detect translocations (the breakage and rejoining of segments from different chromosomes). A genomic hybridization technique known as **spectral karyotyping** (SKY) uses a variety of different fluorescent dyes to create a unique colored label for each chromosome,

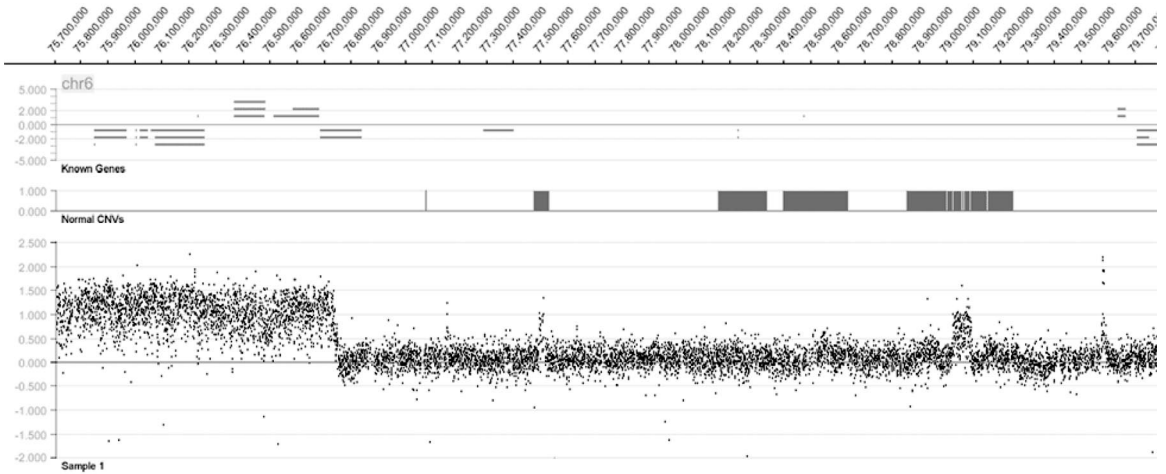


FIGURE 15.2. Array CGH integrated genomic hybridization data for probes that span the entire genome, allowing for the precise location of regions of amplification and deletion.

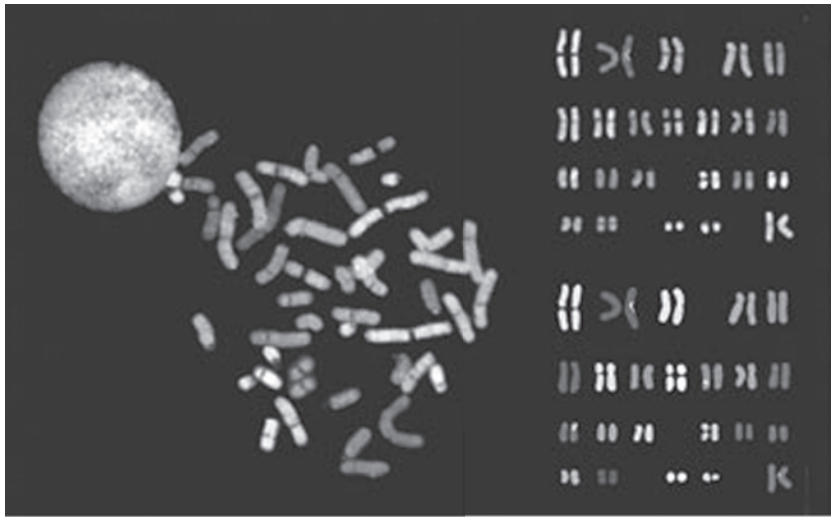


FIGURE 15.3. SKY spectral karyotyping allows for the identification of translocations between different chromosomes. (See insert for color representation.)

and then hybridizes this mixed label to cells of the tumor (see Figure 15.3). The result makes it easy to identify translocations (a hybrid chromosome that contains parts from two different normal chromosomes). ArrayCGH is not well suited for detection of translocations, since the overall copy number of genomic segments may not be changed and only one or a few probes may be affected at the breakpoint of the DNA.

Gene Expression Signatures

One reason why cancer is difficult to diagnose and treat is that it is not one disease. Although medicine typically classifies cancers by the organ or tissue where they occur (breast cancer, lung cancer, leukemia, etc.), not all cancers of a single tissue have the same clinical outcome or response to treatment. This variability reflects both a heterogeneity of cell types where the cancer originated and differences in the specific mutation events that activated the

malignant phenotype. Tumors in a single tissue that appear similar may have mutations in different genes, and therefore respond differently to drugs and produce different clinical outcomes.

Early work by Golub, Lander, and colleagues (Golub et al. 1999) demonstrated that it was possible to use microarrays to reliably distinguish known cancer subtypes, such as acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL). Microarrays have also been used with clustering- and machine-learning-based data analysis techniques to identify *previously unknown* subtypes in diffuse large B-cell lymphoma that correlate to patient survival rates (Alizadeh et al. 2000; Shipp et al. 2002). Similar work has used microarray gene expression profiles to identify classes or subtypes of many cancers that are associated with survival, including prostate cancer (Lapointe et al. 2004), lung adenocarcinoma (Beer et al. 2002), and cervical cancer (Lyng et al. 2006). A diagnostic test that can characterize a cancer as either aggressive (associated with poor survival) or slow-growing can have great clinical significance for the patient. Aggressive cancers can be treated with aggressive therapy methods, including surgery, radiation, and powerful chemotherapy methods; while patients with slow-growing cancers may be spared the severe side effects of these methods while the effectiveness of less disruptive treatments are evaluated.

One of the common features of cancer cells is their aggressive growth. Cells in different tissues are stimulated to grow by different hormones and growth factors. For example, breast and ovary cells grow in response to the hormones estrogen and progesterone. Approximately 75% of breast cancers are found to contain elevated levels of the estrogen or progesterone receptor proteins (ER+). These receptors bind hormones present in the blood and activate other proteins within the cell that act as transcription factors, turning on (or off) the expression of multiple genes that lead to cell growth and reproduction. The hormone receptor

acts as a switch or master regulator that activates a complex cell growth pathway. Higher expression of hormone receptors allows cells to grow in response to hormone concentrations that would not stimulate growth in normal cells. Overexpression of hormone receptors may be due to a copy number amplification of the corresponding genes, a mutation in the regulatory regions of these genes that changes their response to transcription factors or posttranscriptional regulators, or changes in the amounts or functions of transcriptional and posttranscriptional regulators.

Breast tumors that contain elevated levels of hormone receptors can be treated with hormone-blocking drugs such as tamoxifen. The drug binds to the hormone receptors and prevents the hormone from stimulating growth, which often leads to the death of tumor cells. Other antihormone therapies include aromatase inhibitors, which reduce the synthesis of estrogen in muscle and fat cells (in postmenopausal women), drugs that block synthesis of estrogen by the ovaries (for premenopausal women), and surgical removal of the ovaries.

Some breast tumor cells are able to bypass regulation by estrogen/progesterone receptors and turn on growth when no estrogen or progesterone is present, or if estrogen/progesterone receptors are not active. One alternative growth stimulation pathway involves the **epidermal growth factor receptor 2** (also known as **HER2** or **HER2/neu**). Approximately 25% of women with metastatic breast cancer have elevated levels of the HER2 receptor. These cancers are characterized by aggressive growth and a poorer prognosis (De Laurentiis et al. 2005). Elevated levels of the HER2 receptor in tumor cells is usually caused by copy number amplification of the HER2 gene. The drug Herceptin (trastuzumab) is a monoclonal antibody that binds specifically to the HER2 receptor on tumor cells and blocks its growth-stimulating function, while also stimulating immune response against these cells. Most HER2+ cancers can be treated effectively with Herceptin.

In addition to the distinction between ER and HER2-positive and negative phenotypes, the expression of the p53 tumor suppressor gene is an important independent factor in predicting survival in breast cancer and many other types of cancer. Mutations in p53 (and the expression levels of genes regulated by p53) are correlated with poorer survival.

Since the genetics and biochemistry of tumor growth regulation is so complex, it may be difficult to accurately diagnose each cancer from a screening for only a few proteins or genetic markers. The overall pattern of gene expression in tumor cells may be a better indicator of which growth pathways are switched on, regardless of the exact cause of the switch. Gene expression profiling using microarrays has led to the discovery and/or characterization of distinct subtypes for a number of different cancers. Using a multigene clustering approach on breast cancers, ER-negative tumors have been divided into the HER2+ and the basal-like subtypes, and ER-positive tumors are divided into the luminal A and luminal B subtypes. The basal-like subtype is characterized by the low expression levels of the ER-related and the HER2-related group of genes, and therefore is often "triple-negative" on clinical assays for these proteins.

In each case, a profile based on the expression levels of dozens to hundreds of genes can be used to reliably classify tumors into classes that predict disease outcome (such as 5- or 10-year relapse-free survival). Functional characterization of the differentially regulated genes that are identified by these predictive profiles has led to biological insights about the tissue of origin for the cancer and to specific genetic changes (such as amplification of oncogenes). Gene expression profiles have been used to classify breast cancers into as many as five subtypes. The two main subtypes are **luminal A**, which is estrogen-receptor-positive, and **basal**, which is associated with overexpression of genes involved in cell proliferation and differentiation.

Classification of cancers on the basis of these gene expression profiles have proved to be more accurate than those based on histology or single-biomarker tests (estrongen receptor, HER2, etc.). Several breast cancer diagnostic tests based on gene expression profiles of key genes have become commercially available (Oncotype DX™ by Genomic Health Inc., and MammaPrint® by Agendia BV) to help physicians determine what types of therapy are most appropriate for each patient.

In hepatocellular carcinoma (HCC), two distinct subclasses were identified by clustering of gene expression profiles for 80 genes from tumor tissue from different patients (Lee et al. 2006). One cluster shared similar gene expression patterns with stem cells such as fetal hepatoblasts. Patients with tumor gene expression profiles in this cluster had significantly poorer survival than did patients with profiles in the other cluster, which resembled adult hepatocytes.

An important point to keep in mind about gene expression profiles and cancer is that differentiation into tissue and cell types involves permanent changes in the expression of many thousands of genes, while the development of cancer involves changes in the expression of hundreds of genes. There are changes in the regulation of several processes that distinguish cancer: growth, cell cycle, DNA replication, DNA error repair, angiogenesis, and cell motility. But the regulation of these processes is established with a different balance of regulatory molecules in each normal cell and tissue type. Overall, gene expression profiling will always distinguish cell type and tissue type much more easily than it will distinguish cancer from normal.

Cancer Genome Atlas

The NCI has committed to a major investment ($1.5 billion) in a genomics project called the **Cancer Genome Atlas** (`http://cancergenome.nih.gov`) that is designed "to systematically

explore the entire spectrum of genomic changes involved in human cancer." The project will focus on detailed genetic characterization of 12,500 tumor samples from the 50 most common types of cancer. This includes whole-genome gene expression studies and alternative splicing (exon arrays), whole-genome copy number changes and genome rearrangement (translocations), epigenomic changes such as DNA methylation, and whole-genome sequencing of cells from tumor tissue. In each case, tumor cells will be compared to matched normal cells from the same patient. It is hoped that these genomic data will provide insight into the genetic changes in somatic cells that lead to malignant development in cancer and differentiate cancer subtypes.

A number of cancer and genomics scientists have criticized this project as being too costly and poorly designed. The chief problem with the Cancer Genome Project is that the cells to be characterized will come from primary tumors. However, the most damaging cells are the metastatic cells, which move throughout the body and establish secondary tumors. Further complicating the situation, large tumors are not genetically homogeneous. Tumors are composed of cells that are constantly mutating as they grow and reproduce. Most of these mutations are simply byproducts of the defective DNA repair and editing systems in the cancer cell lines. Multiple sections of a single tumor will each contain many different mutations. Therefore the Cancer Genome Project is expected to identify many mutations that do not play a critical role in the development and survival of cancers.

In the United Kingdom, the Sanger Institute is working on its own Cancer Genome Project (`http://www.sanger.ac.uk/genetics/CGP`) with similar goals: to identify somatic mutations critical for the development of human cancers. The Sanger Project is focusing first on the sequencing of coding regions and exon splice junctions in the DNA of primary tumors and cancer cell lines for comparison with the sequence of normal genomic DNA from the same individuals. Early data from this

project published in March 2007 (Greenman et al. 2007) identified more than 1000 mutations in all 518 known genes for protein kinase enzymes, sequenced in 210 human cancers. Another study found mutations in 200 different genes in tumor tissue taken from 11 colorectal cancer patients and 11 breast cancer patients (Lin and Sjoblom 2007). However, it is not possible to definitively distinguish the true cancer-causing mutations from incidental "passenger" mutations in these studies. Very little overlap has been found between mutations in tumors from different types of cancer, or even between samples from different patients with the same type of cancer. If mutations can be identified that play a functional role in cancer development and malignancy, then the corresponding genes become potential drug targets.

Despite criticisms of the cost and experimental design, the Cancer Genome Project may produce useful byproducts. Much of the early work for the project will be targeted at technology development—to allow for the scaleup and cost reduction of the various genomics technologies. One important component of the NCI's Cancer Genome Atlas will be a biospecimen core resource (tissue bank). The biospecimen collections will employ the highest level of ethical, technical, biologic, pathologic, and bioinformatics standards to ensure the quality and quantity of the samples and the associated clinical and consent information. There is a heavy emphasis on informatics including funding for a data coordinating center (DCC). It will be necessary to integrate huge amounts of genomics data from many different technologies plus clinical information on a single set of patients. The DCC will establish public cancer genomics data resources integrated with the NCI's Cancer Biomedical Informatics Grid (caBIG™) and the National Library of Medicine's National Center for Biotechnology Information (NCBI) that scientists can use in their research to generate new insights into the causes and potential targets for interventions in cancer. Once large-scale cancer genome data are

accumulated in a coherent data infrastructure, they may well lead to unanticipated discoveries.

References

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**:503–511.

Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM et al. 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* **8**:816–824.

De Laurentiis M, Arpino G, Massarelli E et al. 2005. A meta-analysis on the interaction between HER-2 expression and response to endocrine treatment in advanced breast cancer. *Clin Cancer Res* **11**:4741–4748.

Golub TR, Slonim DK, Lander ES et al. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**(5439):531–537.

Greenman C, Stephens P, Stratton MR et al. 2007. Patterns of somatic mutation in human cancer genomes. *Nature* **446**(7132):153–158.

Hermans A, Heisterkamp N, von Lindern M, van Baal S, Meijer D, van der Plas D, Wiedemann LM, Groffen J, Bootsma D, Grosveld G. 1987. Unique fusion of bcr and c-abl genes in Philadelphia chromosome positive acute lymphoblastic leukemia. *Cell* **51**:33–40.

Lapointe J, Li C, van de Rijn M, Huggins JP, Bair E et al. 2004. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci USA* **101**:811–816.

Lee JS, Heo J, Libbrecht L, Chu IS, Kaposi-Novak P, Calvisi DF, Mikaelyan A, Roberts LR, Demetris AJ, Sun Z, Nevens F, Roskams T, Thorgeirsson SS. 2006. A novel prognostic subtype of human hepatocellular carcinoma derived from hepatic progenitor cells. *Nat Med* **12**(4):410–416.

Lin J, Sjoblom T. 2007. Genome-wide mutational analyses of breast and colorectal cancers. *Discov Med* **7**(37):13–19.

Lynch HT, de la Chapelle A. 1999. Genetic susceptibility to non-polyposis colorectal cancer. *J Med Genet* **36**:801–818.

Lyng H, Brovig RS, Stokke T et al. 2006. Gene expressions and copy numbers associated with metastatic phenotypes of uterine cervical cancer. *Genomics* **7**:268.

Minamoto T, Mai M, Ronai Z. 2000. K-ras mutation: Early detection in molecular diagnosis and risk assessment of colorectal, pancreas, and lung cancers—a review. *Cancer Detect Prev* **24**:1–12.

Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL et al. 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* **8**:68–74.

Slamon DJ, Godolphin W, Press MF et al. 1989. Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science* **244**:707–712.

Yan H, Papadopoulos N, Vogelstein B et al. 2000. Conversion of diploidy to haploidy. *Nature* **403**:723–724.

CHAPTER

16

PROTEOMICS

Proteomics, a hot new buzzword surfacing in scientific conferences and journal articles as well as among biotechnology investors, can be loosely defined as the measurement and study of all of the proteins in an organism (or in a specific tissue or cell type), namely, the **proteome**. This covers the full gamut of information about proteins from amino acid sequences to tissue-specific expression, three-dimensional (3D) structure, protein–protein and protein–DNA interactions, as well as biochemical/metabolic function. This is closely associated with **functional genomics**, which seeks to understand the function of all the genes in the genome.

Gene expression, as measured by microarray technology, provides some information about where and when genes are expressed, but most genes exercise their biological function through the production of proteins. The proteins are the enzymes, regulatory molecules, and building blocks of cellular structures. Unlike DNA microarrays, which apply a uniform technology based on RNA–DNA hybridization to measure the mRNAs produced by all genes, proteomics involves an assortment of different technologies. Proteomics technologies include quantitative

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

measurements of amounts of different molecular species of proteins (mass spectroscopy), identification of protein–protein interactions, protein structural analysis, 2D gel electrophoresis, and various forms of computational function prediction. The essential distinction between classic protein chemistry and molecular biology approaches versus proteomics methods is that proteomics attempts to address all of the proteins in an organism at once.

Protein Modifications

Proteomics is significantly more complex than DNA- or mRNA-based genomics technologies. While alternate splicing allows one gene to produce several different mRNAs, posttranslational modification of proteins can multiply this complexity manyfold. Proteins can be cut by specific proteases (proteolytic cleavage), crosslinked by disulfide bonds either internally or to other protein molecules (of the same type or of different types), phosphorylated, glycosylated, hydroxylated, carboxylated, amidated at the *C* terminal, acylated, and methylated, with sulfate added to tyrosine residues and farnesyl or geranyl groups added to carboxy terminal cysteine residues. Proteins may also be localized within or excreted from the cell, transported through the body, and bound at cell surface receptors or imported by specific cell types that are located far from the cells that produced them. Proteins can also bind to other proteins or nonprotein cofactors to form complex molecular machines such as ribosomes, membrane pores, and spliceosomes. All of this complexity leads to the inevitable conclusion that identifying and measuring all of the proteins from cells or tissues will produce a very large dataset of chemically diverse molecules.

These posttranslational modifications have a profound influence on the function of a protein in a specific cellular system. Proteins with the same amino acid sequence but that are in different crosslinking, phosphorylation, or glycosylation states may have

very different metabolic activities. Ideally, investigators wish to know the exact amounts of every form of every protein as a precise measurement of the biological state of a tissue or cell type.

Quantitative Approaches

Quantitative measurement of proteins potentially provides the most precise information about a patient's current health and metabolic status. Genomics experiments that use DNA microarrays to measure RNA levels can make only approximate measurements of the levels of proteins being produced from each gene. RNAs are the template for protein synthesis, but there are many forms of posttranscriptional regulation that can affect the amount of protein made from each mRNA molecule. The mRNAs for some genes may be translated into protein more efficiently than others. This may be due to direct sequence-specific differences in the processing of mRNA sequences by the ribosomal translation machinery, or to the actions of capping and other mRNA-processing enzymes, which may themselves have sequence-specific affinities. Alternately, some mRNA molecules may be degraded by RNAse enzymes more rapidly than others, and therefore serve as a template for the synthesis of fewer protein molecules. There may also be a significant time lag between changes in mRNA levels and changes in the overall levels of the corresponding proteins—for instance, large pools of some proteins may exist in the cell, buffering the effect of a rise or fall in the rate of new synthesis. In fact, some proteins may have long lifetimes in a cell, so that rates of gene transcription and protein synthesis may not be closely related to the amounts of that protein in the cell.

Measuring the amounts of many different proteins in a cellular extract is technically much more difficult than measuring mRNA in DNA microarrays because proteins are so chemically diverse. Proteins range from strongly acidic to strongly basic,

hydrophilic to hydrophobic, membrane-bound or soluble, glycosylated, attached to metallic or organic cofactors, bound into dimers or complex multiunit molecular machines, and so on. Any measurement technology that isolates proteins from a cellular lysate must favor some chemical forms over others. Every buffer and reagent will have differential effects on this complex mixture of molecules. There is no single technology that can capture and quantitate all (or even most) of the proteins produced by a cell. Therefore proteomics technologies will inevitably be a composite of many different chemical methods.

Current methods for measuring proteins involve some combination of gel electrophoresis, chromatography, affinity binding, and mass spectrometry. Two-dimensional polyacrylamide gel electrophoresis (2D PAGE) has been the traditional workhorse of protein chemistry for several decades. A sample of proteins (cellular extract) is first separated in one direction by pH in an acrylamide gel by isoelectric focusing. Then proteins are separated according to size in a second dimension (at right angles to the first) by electrophoresis. Finally, proteins are visualized in the gel by staining or autoradiography. If all experimental conditions are kept constant, then the same protein should end up in the same location on the gel in repeated experiments, so that samples from different tissues or experimental manipulations can be compared to see if the amount of protein present in a specific spot increases or decreases. This involves quite complex image analysis software since *many* proteins are present (see Figure 16.1) and no two gels are ever exactly alike. Even with the most perfectly controlled gels and excellent image analysis tools, the intensity of protein spots on a 2D PAGE can provide only a rough estimate of protein amounts.

Once proteins are separated by 2D PAGE, it is possible to identify them using a variety of techniques. Individual proteins can be identified by **immunoblotting**—transferring the proteins from the acrylamide gel onto a nylon membrane, then using a specific

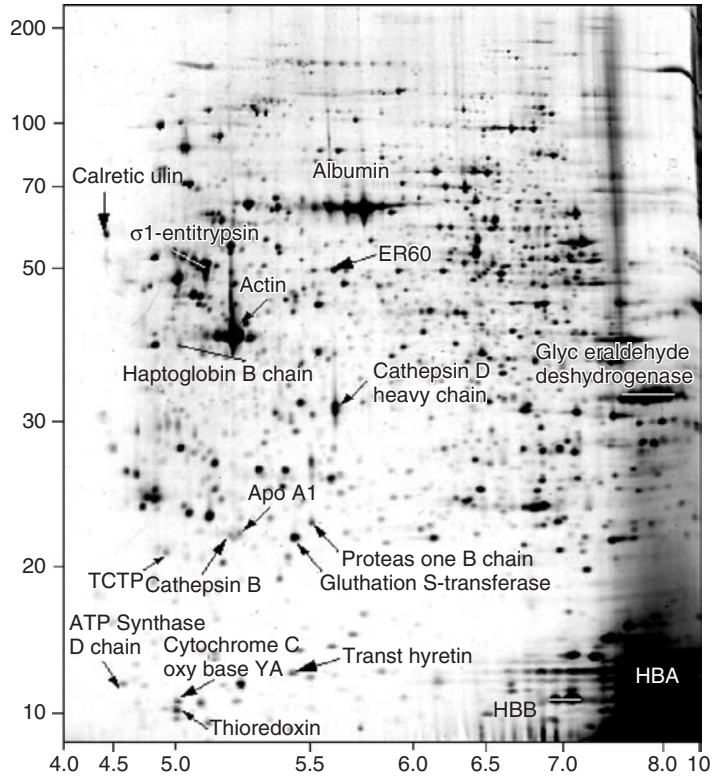


FIGURE 16.1. A two-dimensional PAGE image of proteins from human lymphoma tissue.

antibody to bind the protein of interest and some immunostaining technique to identify where on the gel the antibody binds. Individual spots can also be cut out from the gel and the proteins identified by mass spectrometry or by amino acid sequencing. A number of databases have been created to facilitate the analysis of 2D PAGE that identify the pH and mass locations of various proteins and offer many images for comparison, but the identity of every protein spot must still be confirmed.

Proteomics requires a high-throughput method to identify and quantitate huge numbers of different proteins in parallel, rather than one at a time. New mass spectrometry technologies

such as matrix-assisted laser desorption/ionization–time-of-flight mass spectrometry (MALDI/TOF-MS) offer the most promise. MALDI-TOF works by ionizing protein molecules with a laser, accelerating the ions in an electric field, and then measuring the amount of time required to reach a detector. The mass of a protein molecule is proportional to its time of flight to the detector, and its amount is proportional to the intensity of signal recorded at the detector at a particular moment in time. Proteins can be uniquely identified by the mass signatures of their ionization products. MALDI/TOF can identify a number of mixed proteins in a very tiny and impure sample (femtomole quantities), and the machine has very rapid throughput, but it still cannot sort out tens of thousands of proteins in a heterogeneous mixture. Some separation technology must be applied first.

Various affinity-based separation schemes are currently in use to separate proteins from a cell extract prior to MALDI-TOF analysis. A number of generic chemical or biochemical affinity ligands can be used to bind entire classes of proteins by their intrinsic biochemical properties, or proteins can be labeled in solution with a linker molecule, with that linker later used as an affinity tag to attach the proteins to a substrate. It is also possible to use specific antibodies to fish out individual species of protein molecules from a mixture, or to create tagged proteins by gene fusions that can later be retrieved by affinity binding to the tag.

In order to perform a high-throughput, genome-scale experiment, it would be necessary to set up a single experiment where all proteins are targets. This has been partially accomplished with the **protein chip** by placing a large number of different proteins in an array on a glass slide. This requires that the gene for each protein be cloned into an expression vector; then the resulting protein product is purified and placed in a spot on the array. Then cell extracts can be washed over the array and interactions can be screened in an all-against-all format. So far, these protein arrays have mostly been used to measure the interactions of each

protein on the chip with a single substrate such as calmodulin, streptavidin, or phosphotidylinositide. The protein chips can also be used to characterize the interactions of an array of proteins with a drug molecule. The combination of protein chips with MALDI-TOF should soon lead to a technology that can identify and quantitate every protein (or at least a significant fraction of the proteins) in a cellular extract.

Uses for Quantitative Protein Data

Quantitative protein data can be used for many of the same applications as mRNA-based gene expression data: identification of coregulated proteins, determination of tissue-specific (or subcellular) localization of proteins, identification of quantitative difference in proteins associated with specific genotypes or phenotypes, or association of changes in the abundance of specific proteins with disease or the reaction to drugs or toxic substances (i.e., proteomic signatures that can be used as diagnostic tools). All of these measurements could potentially be more precise than mRNA-based technologies because they measure the actual amounts of proteins active in the cells and because it is possible to discriminate between various forms of a protein (phosphorylation state, posttranscriptional modifications, etc.).

Biomarkers

The application of genomics to clinical medicine is often described in terms of **biomarkers**. Genomics is a broad discovery process that includes a variety of high-throughput assays of DNA (multilocus SNP genotyping, array comparative genomic hybridization), mRNA (gene expression microarrays), proteins (proteomics), hormones, and other metabolites (metabolomics). Most of these research tools are poorly suited for direct clinical use since they produce large complex datasets. The goal of genomics research is generally to identify specific genes, proteins,

or other markers that are associated with disease, which leads to a better fundamental understanding of the disease process, targets for the development of new drugs, and markers that can be used to detect or more accurately diagnose disease. Molecules that are found to be significantly associated with disease become candidate biomarkers, and focused assays are developed for these markers in order to validate the markers in larger populations of patients. In order to enter common clinical practice, a biomarker must be highly predictive of disease, and provide some clinical information that cannot be readily obtained through traditional methods. Biomarkers are particularly valuable if they allow early detection of disease or allow for more precise diagnosis that enables informed treatment decisions.

Since 1998, the NCI has sponsored a large project known as the **Early Detection Research Network** (EDRN) dedicated to the discovery of biomarkers for the early detection of cancer. In 2004, the NCI funded 17 labs to work on the development of new cancer biomarkers. These laboratories are using DNA arrays, protein arrays, and bioinformatics to investigate hundreds of thousands of leads to discover unique signatures for early cancer. The EDRN also includes support for clinical validation trials for potential biomarkers developed in the discovery labs. A phase II EDRN-sponsored clinical trial was begun in 2005 for des-γ carboxyprothrombin (DCP) as a detection marker for early stage liver cancer. Preliminary studies have shown that DCP has close to 90% accuracy for the diagnosis of liver cancer.

Proteins are often the preferred biomarker for disease detection and diagnosis since many rapid, sensitive, and reliable tests for proteins are in common use in clinical medicine. The enzyme-linked immunosorbent assay (ELISA) is one of the most common and reliable methods in the field of diagnostic medicine. The enzyme-linked antibodies used in the assay are stable and inexpensive to produce. The protein targets of the assay are generally considered to be stable in blood and other biological fluids, and the antibody test is both highly sensitive and highly specific for

the protein target. Diagnostic protein biomarkers are often detectable by ELISA in saliva, or urine, which makes it possible to develop a noninvasive test that can be administered by people without advanced medical training (or self-administered by the patient). The ELISA test produces a simple color change, which can be read as a +/− score. This is a simple, unambiguous result, which can easily be interpreted by physician or patient. Essentially any protein can be used to develop an antibody that can be incorporated into an ELISA test—natural proteins from the human body, proteins from pathogens, and antibodies produced by the immune system as a reaction to exposure to foreign proteins.

The drugstore early-pregnancy test is a classic example of a biomarker that has been developed into a widely used diagnostic tool. The early pregnancy test uses an antibody for the protein hormone human chorionic gonadotropin (hCG), which is produced by the embryo after fertilization. The immunoassay, which is rapid and inexpensive, can detect hCG levels in the range of 20–100 mIU/mL, which are typically produced by the second week of pregnancy. The pregnancy test is highly sensitive and specific (accuracy greater than 99%) because hCG levels are negligible in women except during pregnancy (and hCG is rarely produced in males). While the hCG pregnancy test is not a result of genomics research, it is the product of a thorough understanding of the underlying biology. More recently, genomics methods have been used to search for biomarkers that could be used to detect abnormal pregnancy. Researchers at Ciphergen Biosystems, Inc. and Yale University School of Medicine have used proteomics to identify biomarkers that can distinguish women with preeclampsia from healthy pregnant women. Protein biomarkers for preeclampsia include inhibin A and activin A. Women with urine inhibin A levels greater than 90 pg/mg urine creatinine exhibited a 17-fold elevated risk of preeclampsia (Hamar et al. 2006). Interestingly, hCG is elevated in some cancers of reproductive tissues such as choriocarcinoma, embryonal carcinoma, polyembryoma, and mixed germ cell tumors, and it

may be used as a biomarker to aid in detection and diagnosis of these cancers (Gadducci et al. 2004).

Sensitivity and specificity are problematic for some of the existing clinical tests that are based on biomarkers, especially tests that detect changes in levels of proteins that are naturally present in the body. **Sensitivity** refers to the ability of a test to detect a disease when the disease is present. A failure of sensitivity would result in a false-negative result, where the patient was incorrectly diagnosed as healthy but in fact had the disease. Most clinically acceptable tests have a sensitivity (detection rate) of at least 95%, with 99% a common standard. **Specificity** is a measure of the false-positive rate of the test, the fraction of positive test results where the patient does actually have cancer, and this is where many tests in common use perform poorly. Many common clinical tests have specificity below 50% (more than half of positive test results are false). For example, the prostate-specific antigen (PSA) test is commonly used to detect prostate cancer in men over the age of 40. However, the specificity of the PSA test is only about 30%, so that more than two-thirds of patients with a positive test result do not actually have prostate cancer (Hoffman et al. 2002).

The development of biomarkers has proved difficult for diseases that are characterized by abnormal development rather than infection, such as cancer, heart disease, and Alzheimer's disease. In the past, successful biomarker diagnostics were based on research that has moved from a thorough understanding of the basic biology of the disease or condition to the identification of target proteins. The most successful protein biomarkers are not present at all in the healthy or unexposed person, greatly reducing the problem of false positives. Developmental disorders are complex and not well understood on a molecular level. Rather than identifying a unique process associated with disease, researchers have attempted to use proteomics methods for a scatter-shot discovery of biomarkers. They ask questions such as "What proteins can be detected that are different between a tumor and

normal tissue of the same organ, or between aggressive and benign tumors?" This is a difficult discovery process since the observed differences are generally quantitative rather than representative of a complete change from absent to present, and many other biological processes can affect the levels of endogenous proteins.

The ideal tumor marker has been described as follows (Kufe et al. 2003):

1. Specific production by premalignant or malignant tissue early in the progression of disease
2. Produced at detectable levels in all patients with a specific malignancy
3. Expression in an organ site-specific manner
4. Evidence of presence in bodily fluids obtained noninvasively or in easily accessible tissue
5. Levels related quantitatively to tumor volume, biological behavior, or disease progression
6. Relatively short half-life, reflecting temporal changes in tumor burden and response to therapy
7. Existence of a standardized, reproducible, and validated objective and quantitative assay

Unfortunately, no currently available biomarkers meet these excellent criteria.

Proteins are biologically active molecules that are chemically modified in many different ways both during synthesis and as they function within the cell. Alternative splicing of mRNAs may lead to the production of many different polypeptide chains from a single gene. In fact, proteomic methods based on mass or immunological detection of proteins are more likely to detect the results of alternate splicing than hybridization-based assays of RNA. A single mRNA may encode one polypeptide sequence,

but many different protein molecules can be produced that differ in mass, charge, and other chemical properties as a result of phosphorylation, glycosylation, proteolytic cleavage, and other posttranslational modifications. These chemical isoforms may be rapidly interconverted by the action of endogenous enzymes. Therefore, there is a great deal of natural variability in the protein content of biological samples.

On top of this underlying biological variability in proteins, there is a great diversity of analytical methods used for proteomics. Many different manufacturers produce different types of mass spectrographic equipment that are described by a bewildering array of acronyms, including MS/MS, LC-TOF, Q-TOF, MALDI, TOF/TOF, and various combinations of these. These are further complicated by a variety of different preparation and labeling steps prior to mass spectrograph analysis such as iTRAQ and ICAT. The data produced by these various types of equipment and analytical methods are stored in many different formats, which are proprietary to the different manufacturers.

As a result of the underlying biological variability and the difficulty of comparing analytical results between laboratories, progress has been slow in using proteomic methods to identify reproducible markers that can be used to diagnose or classify disease. Many markers that prove to be highly correlated with disease in one study are not validated by other laboratories.

Protein Databases

It is a primary goal of the human genome project to create a single, definitive list of all of the genes and all of the expressed proteins in the human genome and to assign functions to each of these proteins. However, in 2002, this goal seemed quite distant. There were dozens of protein databases that relied on various interpretations of genomic data, each containing some entries that were not shared by the others. The NCBI has created a

"hand-curated" list of human proteins known as **RefSeq NP**, which has 19,959 entries (as of May 2008). **Ensembl** (a genome annotation effort maintained by the European Molecular Biology Laboratory) contains 21,541 known human protein-coding genes, but many of these are different from the proteins listed in RefSeq. **SwissProt** is a high-quality manually annotated database of protein sequences maintained by the Swiss Institute of Bioinformatics (SIB), which contains 19,552 human proteins (out of a total of 385,721 proteins from all species). The SIB also maintains a more comprehensive nonredundant list of proteins called **TrEMBL** (translations of the EMBL DNA database), which has 71,941 human entries, again substantially different from the lists from either RefSeq or Ensembl. The European Bioinformatics Institute has developed a cross-referenced protein database for all of these others known as the International Protein Index (IPI). As of May 2008, it contains 71,884 entries for human proteins.

Protein–Protein Interactions

Another aspect of proteomics is the physical interaction between proteins. Many proteins interact with other proteins—either to form complex multisubunit molecular machines, to regulate the function of other proteins, or to be regulated. The extent and the nature of these interactions are important for the understanding of metabolic and regulatory pathways, as well as for the functional characterization of the many new proteins being discovered by genome sequencing (**functional proteomics**). Many proteins form complex multisubunit structures that may include two or more copies of the same protein (homodimers or homopolymers) or complexes with other proteins (heterodimers or multimeric structures). These multisubunit structures can assume the complexity of full-fledged molecular machines such as ribosomes, histones, DNA and RNA polymerases, or the RNA-splicing complex.

Several different technologies exist to examine protein–protein interactions, but none have the capacity for high-throughput analysis of the complete protein complement of a cell or an organism. The traditional biochemical method of investigating protein–protein interactions was to attach a purified protein (the target protein) to a matrix, such as a resin in a chromatographic column, then pour a cellular extract over the matrix, allowing some proteins to adhere by binding to the target protein. Then the bound proteins would be chemically characterized—generally by mass spectroscopy (MS).

The yeast two-hybrid system improved this method by allowing the protein–protein interaction to take place inside a yeast cell and selecting clones of genes for each protein that binds to the target protein. Even so, the mapping of all interactions between all proteins would require a separate experiment using each different protein as the target.

A more recent innovation in protein–protein interactions has been the use of gene fusions to add affinity tags to the ends of cloned genes for a large number of individual proteins (Ho et al. 2002; Gavin et al. 2002). These tagged proteins can then be expressed in yeast cell lines, where they are used as "bait" for other interacting proteins. Under a given set of metabolic conditions, the tagged protein plus any other proteins bound to it are collected on an affinity column, then the captured proteins are separated from the tagged "bait" protein and identified by mass spectroscopy. So far these technologies have been applied only in yeast, where the construction of thousands of transgenic strains with affinity-tagged protein sequences is quite simple, but in principle they could be applied to any cells that can be grown in culture.

Protein–protein interactions and metabolic pathways can also be predicted computationally. One approach to this problem is to utilize the evolutionary tendency for multiple proteins that function in sequential steps in a metabolic pathway to become fused into a single gene in some organisms. For example, the

α and β subunits of the fungal tryptophan synthetase gene correspond to two separate bacterial genes. Similarly, a set of proteins that function in a single metabolic pathway tend to be conserved across evolution—it is unlikely that a species would keep some members of the pathway and discard others. Thus, as more complete genome sequences accumulate for more species, the functional annotation of human proteins can be filled in. Computational predictions of protein interactions can be combined with experimental data and knowledge of biochemical and signal transduction pathways to form protein interaction maps (see Figure 16.2). These maps can then be used as a tool to

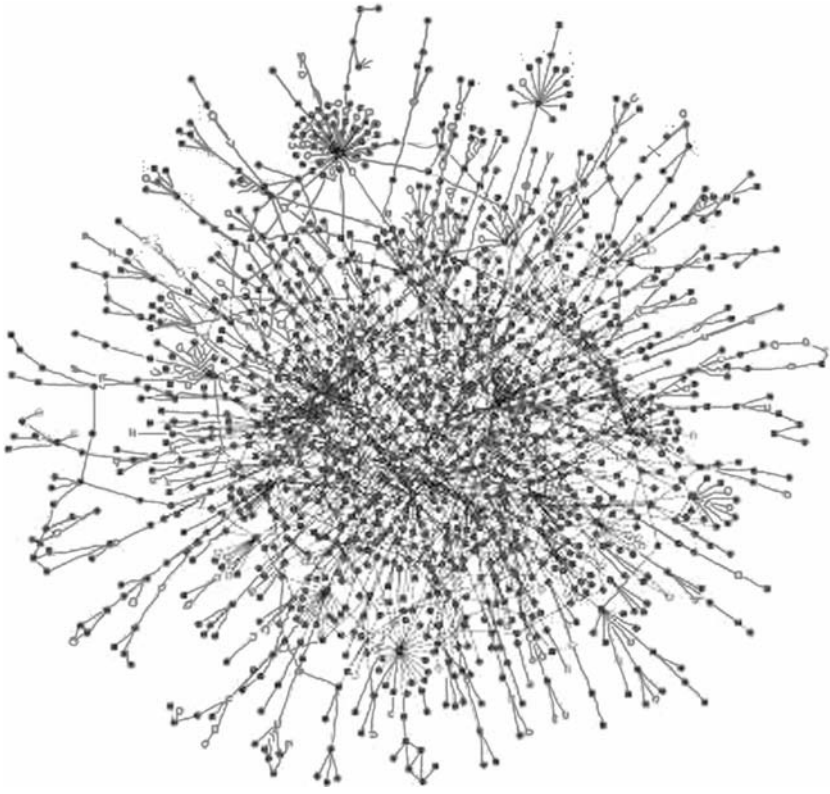


FIGURE 16.2. A map of protein–protein interactions for 1870 yeast proteins (Jeong et al. 2001). (See insert for color representation.)

validate or interpret genomic and proteomic results that indicate coregulation.

DNA-Binding Proteins

Many proteins regulate the transcription of other proteins by binding to the genomic DNA near the coding sequence of the regulated protein and interacting with the RNA polymerase machinery. These transcription factors bind to very specific DNA sequence motifs which may be located in the promoter region—directly upstream from the transcriptional start site, or elsewhere on the chromosome as much as 20 or 40 kb from the coding sequence (enhancers). A given DNA-binding protein may stimulate or repress transcription, or it may have both functions depending on its protein–protein interactions with other transcription factors, which may themselves bind to other DNA motifs. The control of gene expression seems to be the result of a complex combinatorial interaction of sets of DNA-binding regulatory proteins with the promoter sequences of genes.

A variety of molecular techniques such as gel shift electrophoresis and affinity chromatography have been used to catch all of the proteins that bind to a given piece of DNA. Then the individual proteins can be identified using mass spectroscopy. It is also possible to identify the particular sequence to which each transcription factor binds. In fact, these transcription factor binding sites are generally only about 6 bases long, which is not nearly enough information to allow the transcription factors to uniquely target just the promoters of a precise set of genes. (A 6-base sequence will occur by chance once in every 4096 bases of DNA.) However, promoter sequences contain multiple transcription factor-binding sites, sometimes all for the same factor, but often for two or more different factors. Therefore, the binding of each transcription factor protein to DNA may require cooperative interactions with other transcription factors. The overall

interaction of all of the transcription factors with all of the binding sites in the promoter region of a gene allows it to serve as a very complex switching mechanism which is sensitive to small shifts in levels of any of the regulatory proteins. Other possible modes of regulation might include alternate splice forms or chemically modified (phosphorylated, glycosylated, etc.) forms of the transcription factor proteins, which might bind DNA more tightly or serve as competitive inhibitors.

A set of mutually interacting transcription factors may function to regulate many different genes that show similar gene expression patterns, such as the members of a single biochemical pathway. However, these same DNA-binding proteins, with the addition of a few alternates, may also regulate proteins in totally different pathways with different tissue-specific, developmental, and temporal gene expression characteristics. Thus, a relatively small set of transcription factors serve in various combinations to regulate very precisely the expression of all genes.

Structural Proteomics

It is expected that a great deal more can be learned about the biological function of a protein from its 3D structure than from its amino acid sequence or its molecular weight. However, it is not currently possible to predict 3D structures directly from amino acid sequence or mass spectrometry data. Three-dimensional structures can be determined experimentally only from the painstaking work of purifying and crystallizing a protein, then subjecting the crystal to X-ray crystallography or nuclear magnetic resonance (NMR) analysis. However, it is possible to use information about known protein structures to predict the structures of similar proteins. This process, known as **threading**, starts by looking for protein sequence similarity between a new protein and all of the proteins with known structure in a database called the **PDB** (Protein Data Bank), which is maintained by a

group called the Research Collaboratory for Structural Bioinformatics (Westbrook et al. 2002). If a similarity is found in the PDB, then threading software can try to fold up the new sequence into a shape similar to the one in the database, making allowances for some known folding properties of specific amino acids that differ between the two proteins.

Threading is limited to pairs of proteins in which the amino acid sequences are roughly 30% identical. Also, when proteins are compared for sequence similarity, only a portion of any two proteins may be similar (Pieper et al. 2002). These conserved regions often correspond to functional domains or motifs, which are actually distinct substructures of the complete protein. In the language of the PDB, they are protein folds. Interestingly, these functional folded substructures generally correspond to exons in the genomic DNA sequence, validating Walter Gilbert's hypothesis that introns allow functional portions of genes to recombine (Gilbert 1978).

The PDB currently contains about 52,000 structures, but these break down into only about 1000 unique folds. For comparison, the InterPro database recognizes about 4700 protein domains, but in fact the discrepancy is greater. Each protein family in InterPro contains proteins that have less than 30% sequence identity with each other, so more than one experimentally determined structure is needed for each recognized functional domain. While efforts are being made to increase the rate at which protein structures are analyzed, there are significant obstacles to automating and scaling up this process. Crystallizing proteins is still much more of an art form than an industrial procedure; each protein requires its own optimal pH, salt concentrations, and the presence of various biological and inorganic cofactors. Furthermore, the proteome is complex—each gene may produce multiple protein isoforms as a result of alternate splicing and/or postranslational modifications, and each isoform is likely to have a unique structure.

Drug Targets

The application of genomics technologies to developing new drugs is all about targets. A drug target generally refers to a protein in the human body that can be acted upon by a drug in order to treat disease. Until relatively recently, pharmaceutical companies have created drugs against a total of about 500 drug targets. With the help of genomics, this number could eventually balloon to 4000 drug targets. Microarrays are particularly well suited for identification of many new genes that are induced (or downregulated) during the various stages of a disease. The next step is to validate these potential targets: "Is the protein involved in a biochemical or regulatory pathway that is directly involved with the etiology of the disease of interest (or in the manifestation of symptoms)?" Then comes the search for a drug molecule that can interact with this protein target to effect a useful change—such as to block disease development or to ameliorate symptoms. If the 3D structure of a target protein is known (or can be computationally predicted), then databases of small molecules can be tested initially by computer-simulated docking to screen for potential drugs. Again, this can speed the process of searching for a drug.

However, the most time-consuming steps in the drug development process are the late-stage testing on animals and humans. First it must be proved that the new drug is safe and has a useful therapeutic effect; then it must be shown that it is more effective than other existing treatments. Unfortunately, these steps will not be substantially shortened by proteomic technologies.

References

Gadducci A, Cosio S, Carpi A, Nicolini A, Genazzani AR. 2004. Serum tumor markers in the management of ovarian, endometrial and cervical cancer. *Biomed Pharmacother* **58**(1):24–38.

Gavin AC, Bosche M, Krause R et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**:141–147.

Gilbert W. 1978. Why genes in pieces. *Nature* **271**:501.

Hamar BD, Buhimschi IA, Buhimschi CS et al. 2006. Serum and urine inhibin A but not free activin A are endocrine biomarkers of severe pre-eclampsia. *Am J Obstet Gynecol* **195**(6):1636–1645.

Ho Y, Gruhler A, Heilbut A, Bader GD et al. 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**:180–183.

Hoffman RM, Gilliland FD, Adams-Cameron M, Hunt WC, Key CR. 2002. Prostate-specific antigen testing accuracy in community practice. *Fam Pract* **3**:19.

Jeong H, Mason S, Barabási AL, Oltvai ZN. 2001. Centrality and lethality of protein networks. *Nature* **411**:41–42.

Kufe DW, Pollock RE, Weichselbaum RR, Bast RC, Gansler TS, Holland JF, Frei E (eds). 2003. *Cancer Medicine*, 6th ed. BC Decker, Hamilton, Canada.

Pieper U, Eswar N, Stuart AC, Ilyin VA, Sali A. 2002. MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res* **30**: 255–259.

Westbrook J, Feng Z, Jain S et al. 2002. The Protein Data Bank: Unifying the archive. *Nucleic Acids Res* **30**:245–248.

CHAPTER

17

Consumer Genomics and Genealogy

Like Tang™ from the NASA program, technology research often produces unexpected spinoffs. The recent advances in DNA sequencing and genotyping technologies have made it much less expensive to obtain genetic information. As a result, direct-to-consumer applications of DNA sequencing and genotyping have emerged, particularly in the field of genealogy (where little harm can be done), and also in the unregulated areas of healthcare advice and nutritional supplements. Many of these genetic/genomic tests are marketed directly to consumers and sold via the Internet. People purchasing these tests generally have no communication with licensed genetic counselors.

Genealogy

The application of DNA technologies to genealogy is based on three different types of genetic information: genomewide SNPs, mitochondrial DNA sequences, and Y chromosome sequence polymorphisms. In all cases, the fundamental concept is based

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

on ancestry by descent. Although humans have a lot of DNA (3.2 billion bases), and mutations occur all the time (each person has a few novel mutations), a specific mutation at one site in the genome is unlikely to occur independently in different people. A group of people who share a common mutation (a SNP allele) are very likely to all share a common ancestor who was the originator of this mutation. Geographically isolated populations tend to accumulate distinct mutations over time (markers); then the movement of groups of people to new locations can be tracked by these markers.

Using an approach similar to the HapMap project, randomly chosen, highly polymorphic SNP markers have been characterized in ethnically distinct populations of people from different geographic regions. A cluster analysis of these markers allows for the creation of SNP profiles for the various populations. These markers can then be used to screen DNA samples from people of unknown ancestry, producing a percentage of match to each ethnic/regional profile. The company DNAPrint Genomics offers a test called **AncestryByDNA™** that follows this genome-wide SNP approach. The test uses a panel of 175 SNPs described as "ancestry informative markers" (AIMs) that have been characterized in a large number of well-defined biogeographic population samples. These markers are selected on the basis of showing substantial differences in frequency between population groups. For example, the Duffy Null allele (FY*0) is very common (with an allele frequency of close to 100%) in all sub-Saharan African populations, but is found in low frequency in other groups. Thus, a person with this allele is very likely to have some level of African ancestry.

AncestryByDNA can be used to resolve the following four biogeographic groups:

- *Native American*. Populations that migrated from Asia to inhabit North, South, and Central America.

- *European*. European, Middle Eastern, and South Asian populations from the Indian subcontinent, including India, Pakistan, and Sri Lanka.
- *East Asian*. Japanese, Chinese, Mongolian, Koreans, Southeast Asians, and Pacific Islanders, including populations native to the Philippines.
- *African*. Populations from sub-Saharan Africa such as Nigeria and the Congo region.

An additional test, using 320 SNP markers, is available to resolve members of the European group into four subgroups: northern European, southeastern European (Mediterranean), Middle Eastern, and South Asian (Figure 17.1).

After analysis of the AIMs, in a sample of a person's DNA, the likelihood (or probability) that a person is derived from any of the parental populations and any of the possible mixes of parental populations is calculated. The population (or combination

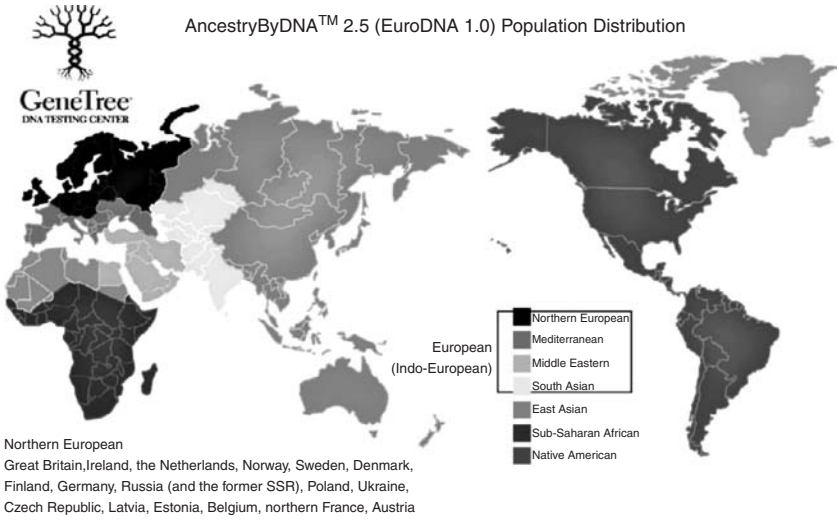


FIGURE 17.1. AncestryByDNA map of human haplotype population distribution. (See insert for color representation.)

of populations) where the likelihood is the highest is then taken to be the best estimate of the ancestral proportions of the person.

An alternative approach to DNA-based genealogy relies on an analysis of sequences of mitochondrial DNA (mtDNA) pioneered by Dr. Brian Sykes of Oxford University. Every cell in the human body has a number of mitochondria, the organelle responsible for converting sugar into ATP for energy. Each mitochondria has its own DNA, organized on a single small, circular chromosome (similar to bacterial DNA). Mitochondrial DNA is inherited directly from the mother (in the cytoplasm of the egg cell), so there is no mixing or recombination across the generations. If a germline mutation occurs in the mtDNA of a woman, then all her children will have the new sequence, and all of her daughters will pass on the new sequence to their children, radiating outward into future generations by maternal lineage. Mutations occur rarely in mtDNA, so the sequence present in any person can be traced unchanged back through generations of maternal ancestors. Mutations present in many people, widely distributed across the world, are very ancient. Mutations present in few people, locally clustered, are likely to be recent.

Sykes found that European people could be divided into seven clusters of mtDNA sequences. He described these clusters as "tribes," each descended from a single female ancestor (Sykes 2002). A total of 36 clusters have now been found in the worldwide human population. By phylogenetic analysis, a single original mtDNA sequence must have existed between 70,000 and 150,000 years ago in a single female, "Eve," who is one of the ancestors of all living humans.

The company Oxford Ancestors® provides mitochondrial DNA analysis to consumers on the basis of the work of Dr. Sykes. Customers submit a swab of cheek cells and 400 bases of

their mitochondrial DNA are sequenced and compared with the company's database of mtDNA from ~17,000 people from around the world. Based on the presence of characteristic mutations for each cluster, people are assigned to their ethnogeographic "tribe".

Mitochondrial DNA is also useful for forensics. All children of one mother, and their mother and grandmother, will have the identical mtDNA sequence. This technique was used to identify the bodies of the royal Romanov family from Russia (Ivanov et al. 1996).

The genetics of the Y-chromosome is similar to that of mtDNA, but it is inherited by males directly from their fathers. Y-chromosome testing can be applied only to males, since females do not have a Y-chromosome. Again, there is no mixing or recombination across generations, however, mutations occur more frequently than in the mtDNA. The Y-chromosome contains several regions of short tandem repeats (Y-STRs), which are subject to frequent mutations (addition or deletion of some copies of the repeat). These Y-STR profiles are often used for forensic applications in order to match individuals to crime-scene DNA samples or to prove paternity. The DNA tests supporting the claim that Thomas Jefferson fathered a child with his slave Sally Hemmings were based on Y-chromosome DNA samples.

Y-chromosome profiling can trace paternal lineages for genealogical and kinship testing. A worldwide database of Y-STR profiles, known as the **Y-chromosome haplotype reference** (YHRD), has been constructed, which is publicly accessible at `www.yhrd.org` (Figure 17.2). The YHRD contains over 45,000 profiles as well as clusters that represent regional and ethnic populations (Roewer et al. 2005).

Currently, eight large metapopulations are searchable: (1) Eurasian, (2) East Asian, (3) Australian Aboriginal, (4) African,

Metapopulation	# Haplotypes	# distinct Haplotypes	# Populations	Avg. # haplotypes per population
Worldwide	46,720	19,587	386	121.03627
Eurasian MP	**33,008**	**12,852**	**267**	**123.625465**
Eurasian MP / European MP	27,312	10,372	194	140.78351
Eurasian MP / Altaic MP	2,477	1,540	31	79.90323
Eurasian MP / Caucasian MP	502	395	13	38.615383
Eurasian MP / Uralic-Yukaghir MP	450	162	3	150.0
Eurasian MP / Indo Iranian MP	1,215	644	20	60.75
Eurasian MP / Indian MP	1,052	748	6	175.33333
East Asian MP	**7,359**	**4,536**	**51**	**144.29411**
East Asian MP / Korean MP	384	336	3	128.0
East Asian MP / Japanese MP	545	449	5	109.0
East Asian MP / Sino Tibetan MP	4,601	2,590	25	184.04
East Asian MP / Austroasiatic MP	313	246	4	78.25
East Asian MP / Thai MP	112	109	1	112.0
East Asian MP / Austronesian MP	1,143	918	10	114.3
East Asian MP / Indo Pacific MP	29	26	1	29.0
East Asian MP / Dravidian MP	232	187	2	116.0
Australian Aboriginal MP	**0**	**0**	**0**	**NaN**
African MP	**2,819**	**1,549**	**25**	**112.76**
African MP / Subsaharan MP	1,112	633	9	123.55556
African MP / Afro-American MP	1,310	881	14	93.57143
African MP / Afro-Caribbean MP	397	328	2	198.5
Amerindian MP	**283**	**215**	**10**	**28.3**
Eskimo Aleut MP	**70**	**46**	**1**	**70.0**
Admixed MP	**1,974**	**1,234**	**20**	**98.7**
Afroeurasian MP	**1,207**	**812**	**12**	**100.583336**
Afroeurasian MP / Semitic MP	976	731	10	97.6
Afroeurasian MP / Berber MP	30	20	1	30.0
Afroeurasian MP / Cushitic MP	201	87	1	201.0

Release "20" from 2006-12-27 13:03:15

FIGURE 17.2. Haplotype distributions in the Y-chromosome haplotype reference database.

(5) Amerindian, (6) Eskimo Aleut, (7) admixed, and (8) AfroEurasian. Some of these metapopulations are further divided into subpopulations:

1. Eurasian
   1.1. European
      1.1.1. Western European
      1.1.2. Eastern European
      1.1.3. Southeastern European

1.2. Altaic
1.3. Caucasian
1.4. Uralic
1.5. Indo-Iranian
1.6. Indian

2. East Asian
2.1. Korean
2.2. Japanese
2.3. Sino-Tibetan
2.4. Austroasiatic
2.5. Thai
2.6. Austronesian
2.7. Indo-Pacific
2.8. Dravidian

4. African
4.1. Sub-Saharan
4.2. African-American
4.3. AfroCaribbean.

8. AfroEurasian
8.1. Semitic
8.2. Berber
8.3. Cushitic

Y-STR testing can be accomplished on a cheek swab DNA sample by any qualified forensic laboratory, and matching to populations in the YHRD is free and can be done by anyone with a basic knowledge of population genetics and an Internet connection.

For $100, the National Geographic Society provides a service called the **Genographic Project** that matches mitochondrial and Y-chromosome DNA markers to haplotype groups associated with specific geographic regions and human migrations. The

Genographic Project Public Participation Kit involves a cheek swab to acquire a DNA sample that is mailed back to the laboratory. This haplotype information links people to known anthropological subgroups that divided humans 30,000–60,000 years ago. Participants receive an analysis of where and when their haplogroup originated and how they lived, including a map that shows migrations across the planet. As the project progresses, the database of samples will expand, and it will become possible to identify other genealogical associations among groups of people.

The GeneBase **DNA Ancestry Project** is a private company that is pursuing a similar goal, but without the academic connections of the National Geographic Society. They offer tools to connect modern geneology and family history information with Y-chromosome haplotype information. People can build public profiles of their ancestries and discover links to other people with similar DNA markers.

From a medical perspective, genetic genealogy is essentially harmless and valueless. The genome-wide SNP, mitochondrial, and Y-chromosome markers that are assessed in these tests are intentionally chosen to be neutral markers (noncoding polymorphisms) that have no apparent effects on health. The tests themselves appear to be accurate (markers are correctly genotyped), but do not provide any health-related information. Customers of these services may discover surprising or unwanted information about their ethnic makeup (or paternity), but privacy controls on the data seem to be adequate. Since the markers have no health implications, there is little concern that the data, if accidentally released, could have negative implications for health insurance or employment. The only potential positive medical impact of a genetic genealogy test might be if a person discovered that he or she shared ancestry to an unexpected ethnic group, and if that group had an elevated frequency of some medically important alleles (e.g., sickle cell trait in Africans, BRCA2 and Tay–Sachs

in Ashkenazi Jews). In that case, one might be inspired to have additional genetic tests to investigate whether one carries these alleles that would not be considered a risk for people who are not members of these groups.

NUTRIGENOMICS

A number of companies are now marketing DNA "health" tests directly to consumers. These tests are based on taking DNA samples (cheek swab or mouthwash rinse) self-submitted by consumers and genotyping several common SNPs of "disease associated" genes, culled from the scientific literature and public databases. These tests claim to provide predictive information about health risks, which may be remediated by special diets (nutrigenomics), vitamins and/or herbal supplements, or cosmetic products. Companies claim to identify genetic susceptibility to conditions including obesity, heart disease, cholesterol metabolism, addiction, and some cancers. Several companies have made claims that genetic profiles can be used to design personalized nutritional programs. The leading company offering nutrigenomics tests is Sciona, a Colorado-based company that provides a 19-gene test: "Sciona is a privately held, international company that provides personalized health and nutrition recommendations based on an individual's diet, lifestyle and unique genetic profile." The genes screened in the Sciona test are

- APOC3, CETP, LPL and eNOS—involved in lipid metabolism
- MTHFR, MS-MTRR, and CBS—vitamin B genes involved in maintenance and repair, especially of cardiovascular and nervous systems
- MnSOD and SOD3—antioxidant genes
- IL-6 and TNFα—markers of inflammation

- ACE, PPAR, γ-2, and VDR—involved in glucose and insulin regulation
- VDR and COL1A1—involved in bone health

Another company, Genova Diagnostics (formerly Great Smokies Diagnostics; Ashville, NC) sells a set of SNP-based health tests:

- *CardioGenomicPlus® Profile*. Evaluates genetic variations (SNPs) that modulate blood pressure regulation, lipid balance, nutrient metabolism, inflammation, and oxidative stress.
- *OsteoGenomic® Profile*. Evaluates SNPs that modulate bone formation (collagen synthesis), bone breakdown (resorption), and inflammation, including key regulatory mechanisms affecting calcium and vitamin D_3 metabolism.
- *DetoxiGenomic® Profile*. Evaluates SNPs associated with increased risk of impaired detoxification capacity, especially when exposed to environmental toxins. It also identifies individuals potentially susceptible to adverse drug reactions.
- *ImmunoGenomic™ Profile*. Evaluates SNPs that modulate immune and inflammatory activity. Polymorphisms affect the levels and activity of the cytokines. These variations can affect balance between cell (TH-1) and humoral (TH-2) immunity, reveal potential defects in immune system defense, and stimulate mechanisms leading to chronic, overactive inflammatory responses.
- *NeuroGenomic™ Profile*. Evaluates genetic variations (SNPs) that modulate estrogen metabolism, coagulation, cardiovascular disease, and osteoporosis.

While studies have implicated each of these genes in the relevant medical areas, there is no clinical evidence that a person who has specific alleles of any of these genes has a significantly

elevated chance of disease. Furthermore, there is no clinical evidence that dietary modifications, vitamins, or herbal supplements have any direct remedial effect on the pathways affected by these genes, or that people with the "at risk" alleles have different responses to these interventions than people with "normal" alleles.

The US Government Accountability Office (GAO 2006) has determined that these tests provide unreliable and misleading advice to consumers about disease risk. The tests may "instill unwarranted fear about impeding illnesses without evidence." The GAO report emphasized that these tests are completely unregulated by the government and not certified or supervised by any medical organization. The American College of Medical Genetics opposes direct-to-consumer sales of genetic tests because they may harm health (ACMG 2004).

From a clinical perspective, there are three considerations for evaluating the usefulness of a genetic test: analytic validity, clinical validity, and clinical utility. **Analytic validity** is simply the technical reliability of the test—that it accurately identifies alleles of the genes that it claims to evaluate. Given the high quality of genotyping equipment available from many different manufacturers, simply enforcing sample handling and laboratory standards should ensure good analytic validity for SNP-based assays. One important point to keep in mind is that these commercial genetic tests are simple screens for specific SNP alleles in the target genes that have been found to exist at appreciable frequencies in the population. If a person tests negative for the common variant allele, that does not mean that the person has a normal version of that gene. The person may harbor a different mutation that affects the function of that gene—within either the coding sequence, a regulatory mutation, or a mutation in another gene that affects the expression of the target gene. The analytic validity of the specific SNP genotype test does not equate to an accurate assessment of the phenotype for the target gene.

Clinical validity refers to the likelihood that a patient with a given result for a gene test (i.e., a positive test result for an allele that has been identified as correlated with disease) will actually develop the disease. Also, that a person with a negative test result will not develop the disease. This is equivalent to the concept of "penetrance" in classical genetics—some alleles have a moderate or negligible effect on a quantitative phenotype, and the effects of alleles of some genes are completely masked by other components of the patient's genetic makeup. Many alleles interact with environmental factors to yield variable levels of expression or phenotypic effect in different people with the same genotype, so a person might have a "risk factor" allele, yet never develop the associated disease. For example, the mutation that causes Huntington's disease has nearly 100% penetrance. Everyone with the allele will eventually develop the disease. Very few people who test negative for known disease-causing alleles will develop Huntington's disease. The breast cancer susceptibility gene BRCA2 is estimated to have 40–60% penetrance—a woman with a mutant allele will have a 40–60% chance of developing breast cancer in her lifetime. However, there are many different mutations in the BRCA2 gene (450 mutations have been identified) with variable levels of associated disease risk, so a SNP-based screening for the most common mutant alleles will not give an absolute prediction of cancer risk.

Clinical utility is the value of a test in making medical decisions. If there is no known intervention for a genetic condition, or if the medical treatment of a patient will be the same regardless of the outcome of a test, then its clinical utility is zero. Many of the genes assessed by the direct-to-consumer tests have low clinical utility since the "targeted" nutritional recommendations based on these tests are essentially the same commonsense advice that nutritionists give to everyone. How is a person likely to react to the information that they have a 40% elevated risk of heart disease compared to the general population? The concept of clinical

utility also includes a cost–benefit estimate. For example, 0.4% of the general population is estimated to carry mutations in the BRCA2 gene that carry an elevated risk of breast cancer (Malone et al. 2006), so widespread screening would not be cost-effective. However, screening for BRCA2 mutations is advised for people with a family history of cancer.

Mixed in with the pseudoscience are some potentially valid "nutrigenomic" assessments. Mutant alleles of the MTHFR gene reduce the production of folate, which is essential in removing homocysteine from the blood. High levels of homocysteine are associated with blood clots, rapid bone loss, and mental retardation in children. Other symptoms of homocystinuria may include, eye-lens dislocation, seizures, violent behavior, and vascular damage such as premature atherosclerosis and thromboembolism. One allele of MTHFR that leads to reduced folate levels is present in 30–40% of the US population, and 10–15% of Americans are homozygous for this allele. There is evidence that people homozygous for this MTHFR variant can avoid homocystenemia by taking supplements of folic acid. For pregnant women who were homozygous for the MTHFR variant, supplements of folate reduced the chance of neural tube defects (spina bifida and anencephaly) in the fetus (Scholl and Johnson 2000). Of course, folic acid supplements are recommended for all pregnant women, regardless of their genetic profiles, so the clinical utility of the MTHFR test is negligible.

Aside from the lack of clinical utility, companies selling nutrigenomic tests have the potential to abuse consumer trust in genetic science in order to market costly products. Since the alleles tested in nutrigenomic screens are, by definition, present at high frequencies in the general population, many people will have at least a few of the mutant or "of concern" alleles. Without violating the analytical validity of their tests, companies who provide these tests have the opportunity to market "targeted" products to a large segment of the population. If these products

contain drugs, "nutritional supplements," or custom-formulated foods, the potential for exploitative pricing is obvious. There is a further social cost by creating the false impression that *everyone* needs preventive medicines to control their genetic risk factors rather than standard good nutritional practices.

Another concern for the medical practitioner is that people will purchase these consumer genetic tests, and then ask for advice from their physicians once they receive the results. That puts the physician in a very difficult position. It is difficult to obtain detailed information about the genes tested in these proprietary tests, as well as the nature of the tests (what SNPs were evaluated in each gene), and the physicians must be aware of all the concerns discussed above with regard to the clinical validity and clinical utility of a predictive genetic test for complex disease. It is not possible to dismiss the results of these tests as meaningless; nor is it possible to give fully informed medical advice on the basis of the results. Even if the physician is fully informed about the nature of the genes being tested and has detailed information about clinical and molecular genetic studies on each of these genes, it is still not possible to give patients medically sound advice about how to modify their diets, lifestyles, or medications on the basis of this genetic information.

Privacy Concerns

While the clinical utility of direct-to-consumer genetic health tests is low, there are risks that consumers may not anticipate. The alleles assessed by these tests do have medical implications. In combination with family history and clinical data, a patient's genotype for these alleles may have predictive or diagnostic power. There is also the potential that information obtained in a genetic test for one trait may provide information about the risks of other diseases (pleiotropic effects). For example, tests for variants of the APOE gene are included in many of the consumer

nutrigenomic and "cardio" screens. However, variations in APOE have also been associated with increased risk of Alzheimer's disease (van der Flier et al. 2006). This may be information that the consumer does not want, and it may have negative effects if it became known to insurers or employers. The information generated by the companies providing consumer genetic testing is stored by these companies, and it may be released for commercial purposes or by accidental data loss. It is possible that these data could be obtained by insurance companies or employers and used to discriminate against individuals, regardless of the actual ability of this information to predict clinical outcomes. Some consumer genetic testing companies have considered storing the actual DNA samples and patient identifying information, so that additional genetic tests could be performed in the future.

References

American College of Medical Genetics (ACMG). 2004. ACMG statement on direct-to-consumer genetic testing. *Genet Med* **6**(1):60.

Frudakis T, Kondragunta V, Nachimuthu P et al. 2002. A classifier for SNP-based racial inference. *J Forens Sci*.

Government Accountability Office (GAO). 2006. *Nutrigenetic Testing: Tests Purchased from Four Web Sites Mislead Consumers* (`http://www.gao.gov/new.items/d06977t.pdf`; accessed 7/27/2006).

Ivanov PL, Wadhams MJ, Roby RK, Holland MM, Weedn VW, Parsons TJ. 1996. Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. *Nat Genet* **12**:417–420.

Malone KE, Daling JR, Ostrander EA et al. 2006. Prevalence and Predictors of BRCA1 and BRCA2 mutations in a population-based study of breast cancer in white and black American women ages 35 to 64 Years. *Cancer Res* **66**(16):8297–8308.

Pascali VL, Dobosz M, Brinkmann B. 1998. Coordinating Y-chromosomal STR research for the courts. *Int J Legal Med* **112**(1):1.

Roewer L, Croucher PJ, Willuweit S, Lu TT, Kayser M, Lessig R, de Knijff P, Jobling MA, Tyler-Smith C, Krawczak M on behalf of the Forensic

Y chromosome User Group. 2005. Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. *Hum Genet* **116**(4):279–291.

Scholl TO, Johnson WG. 2000. Folic acid: Influence on the outcome of pregnancy. *Am J Clin Nutr* **71**(5):1295s–1303s.

Sykes, B. 2002. *The Seven Daughters of Eve*. Norton.

van der Flier WM, Schoonenboom SN, Pijnenburg YA, Fox NC, Scheltens P. 2006. The effect of APOE genotype on clinical phenotype in Alzheimer disease. *Neurology* **67**(3):526–527.

CHAPTER

18

The Ethics of Medical Genomics

The US government spends approximately $30 billion per year on medical research, primarily through the National Institutes of Health (NIH). This huge commitment of resources represents a tacit contract between the public, who pay for the research with their taxes and support it via their elected representatives, and the scientific establishment, that this medical research should lead to improved healthcare for Americans. Part of the rationale for the government support of the Human Genome Project from 1990 to the present includes setting aside a full 5% of the funding for investigations into the ethical, legal, and social implications (ELSIs) of this technology. In order to develop genomics technologies from pure discovery to medical applications, diagnostics and therapies must be evaluated in clinical trials on volunteer patients. Patients must consent to participate in clinical trials, which requires a high level of trust between potential research subjects and scientists. Once these genomic diagnostics and therapies are validated and made available to medical practitioners, patients must trust that they are safe and effective, and that the genetic

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

information revealed will not be used against their interests. Trust is built on a foundation of ethics with solid guarantees and believable enforcement.

Efforts to establish strong ethical rules for genomic research are well founded on historical precedent. In order for modern genomic medicine to be accepted by the public, it must overcome the checkered history of human genetics that includes many examples of misapplication of genetics to social policy, particularly in the pseudoscience of eugenics. Since current public attitudes toward medical genetics and genomics are shaped by this history, it is necessary to examine it in some detail before addressing the current debate about the ethics of medical genomics and its impact on the medical professional.

Eugenics

In the late nineteenth and early twentieth centuries, Sir Francis Galton (1909) and Charles Davenport (1910) developed the concepts of eugenics (see Figure 18.1) as a social policy to improve the human race by encouraging the most genetically fit people to have more children and to prevent reproduction of people deemed "unfit." Eugenics was enthusiastically adopted by many respected scientists and by the US government, leading to the establishment of the Eugenics Record Office at Cold Spring Harbor Laboratory from 1910 to 1939.

The eugenics policies that were implemented in the United States ranged from restrictions on immigration to the involuntary sterilization of jailed criminals or persons institutionalized for reasons of "insanity or feeblemindedness." Harry Laughlin of the Eugenics Record Office published a Model Eugenical Sterilization Law in 1914 (Laughlin 1914), which became the basis for laws in 33 states. For example, the State of Virginia enacted a "Eugenical Sterilization Act" in 1924, based on Laughlin's model law, which stated that "heredity plays an important part in the

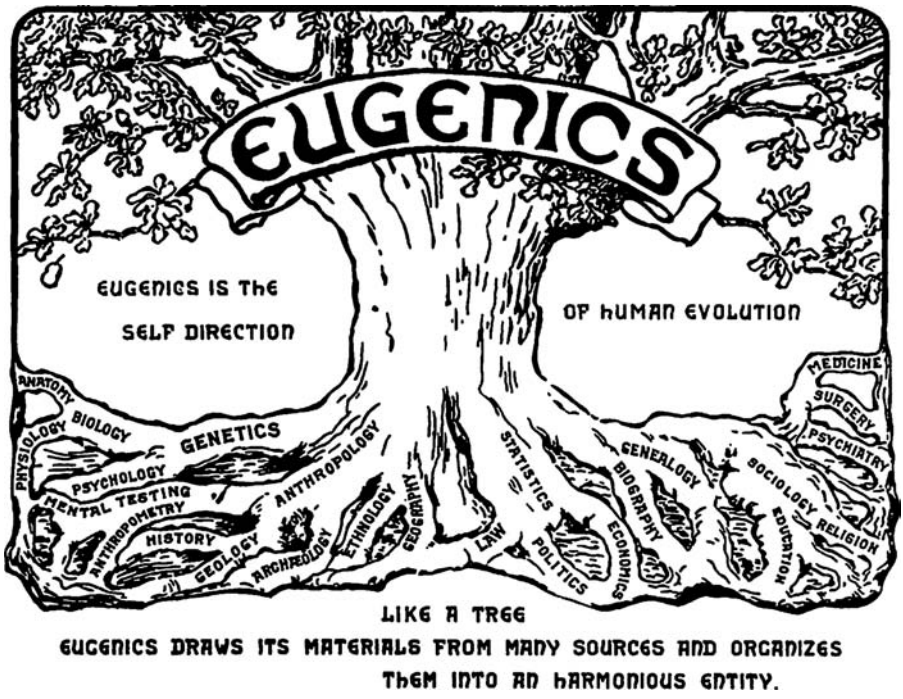


FIGURE 18.1. Logo of the Second International Congress of Eugenics, 1921. Reprinted with permission, American Philosophical Society (Laughlin 1921).

transmission of insanity, idiocy, imbecility, epilepsy and crime. . ." The law provided for the sterilization of persons who were "probable potential parents of socially inadequate offspring." These involuntary sterilizations were upheld by the US Supreme Court in 1927 (*Buck v Bell*). Chief Justice Oliver Wendell Holmes wrote in his opinion: "society can prevent those who are manifestly unfit from continuing their kind. . . Three generations of imbeciles are enough" (Lombardo 1985).

Over 60,000 people in the United States were subjected to eugenic sterilizations before the last of the laws were repealed in the 1970s (CSHL Eugenics Archive). Eugenics concepts were also incorporated in the "racial hygiene" social policies of the German Third Reich government. In fact, the German government directly

adopted text from Laughlin's model law as the basis for a 1933 "Law for the Prevention of Defective Progeny," which was used as the legal basis for the sterilization of more than 350,000 people. The Nazi policy of racial cleansing led to mass exterminations of millions of Jews, Gypsies, homosexuals, and other disfavored groups.

Eugenics was broadly accepted throughout American society during the 1920s and 1930s. There were eugenics exhibits at state fairs, contests for "fitter families," films, public lectures, religious sermons, and eugenics chapters in biology textbooks. In the scientific establishment, there were university courses, eugenics foundations, societies, newsletters, conferences, and scholarly journals.

This history of eugenics seems misguided from our current perspective since our present understanding of genetics contradicts the vast majority of these eugenics claims. It is now clear that even during the heyday of eugenics (1910–1935), there was no body of carefully collected and peer reviewed data to back up these eugenics policies and statements attributed to well-respected scientific "authorities." Data collected and reported by eugenics researchers suffered from many deficits: lack of clearly defined traits (e.g., "feeblemindedness"), bias in data collection (lower IQ test scores of non-English-speaking immigrants), and outright falsification of data. Even the data that had some objective validity, such as family histories of mental illness or the ethnic makeup of prison populations, did not account for the impact of social and economic factors that might lead people from disadvantaged families or social groups to be imprisoned or diagnosed as mentally ill at greater rates than people from more privileged social groups.

These eugenics scientists, some of whom made meaningful contributions in other fields, did not apply rigorous scientific principles to the social aspects of genetics. There is also ample evidence that eugenics laws were implemented with little regard for justice or due process—falsified evidence, collusion between

defense and prosecution attorneys, and bogus expert testimony were common in these cases. It seems that enthusiasm for eugenics as the social application of "modern genetics science" was fueled by prejudices against unpopular racial and ethnic groups and disabled people so that it overwhelmed the usual safeguards of both science and the law.

The history of scientifically unsound eugenic social policy is not confined to the early twentieth century. Laws against intermarriage between "black" and "white" people (racial categories without clear definition or any genetic basis) remained in force until the late 1990s in some US states. Pseudogenetic arguments for the intellectual inferiority of "blacks" (or other groups targeted by racists) to "whites" still can be found in public discourse—such as the 1994 book *The Bell Curve* by Herrnstein and Murray. Other simplistic explanations of single genes responsible for human behaviors such as violence, alcoholism, or homosexuality continue to surface in the scientific literature and the mainstream press. It is important to keep in mind that the evidence for simple genetic controls of intelligence, violence, criminality, or any other form of human behavior is no more convincing today than it was in the 1920s. However, as more human genes are mapped and their functions investigated, these simplistic explanations are likely to increase. Simplistic explanations are inevitable as the media reduce complex scientific discoveries to 30-second items on the evening news: "Gene for alcoholism found; story at 11." Thus there is a need for genomic scientists to pay special attention to the press coverage as well as the social implications of their work.

In 1970, the US government initiated a program of mandatory screening for nonsymptomatic carriers of the sickle cell anemia trait. This screening was implemented primarily among children entering the public school system. Some insurance companies denied coverage or charged higher premiums for people who carried this trait (overwhelmingly African Americans), even though possession of this trait carries no health risk to the individual

(in fact, it provides improved resistance to malaria). Many African-Americans believed that this screening program was the first step in a policy of genocide by preventing sickle cell carriers from marrying and having children.

This historical perspective of eugenics may create an atmosphere of popular distrust surrounding the claims of currently respected scientists when they speak about the social applications of genomic technologies. If the now-absurd claims of eugenics were approved by the scientific and legal authorities of the early twentieth century, what certainties can be provided to the public that medical genomics technologies are not also influenced by the social and economic biases of our own generation? What seems like objective, evidence-based medicine right now, may appear laughably naive a few generations in the future.

Human Genome Diversity Project and Population Genetics

One aspect of the Human Genome Project that was known as the "Human Genome Diversity Project" (HGDP) was designed to collect and analyze DNA samples from a very wide range of ethnic groups and "genetically distinct" or isolated populations in order to gain a greater understanding of human genetic diversity and evolution. The history of the HGDP provides a modern example of "ethics in action" as applied to genomics technology. One major motivation for the HGDP was the belief among anthropologists and population geneticists that the Human Genome Project would sequence DNA samples from primarily Anglo-European people, thus neglecting important diversity in the world's other ethnic populations. The members of many minority groups and "ethnically distinct" populations met the HGDP proposal with a level of suspicion similar to the sickle cell screening program. They characterized the HGDP as the "vampire project," which would take their blood, but return no benefits to them—a form

of "molecular colonialism." In fact, the prevailing belief among members of these groups was that any genetic information discovered about distinct populations would be used to discriminate against them.

Scientists involved in the HGDP were bewildered by the outpouring of negative feelings by members of indigenous groups, since their own work on the project was motivated by a genuine interest in the evolutionary history of humanity and a belief that knowledge of human genetic diversity would lead to medical benefits for all people, particularly the members of these groups. It is not that the HGDP lacked proper "informed consent" protocols for its sampling methods, but that all Western scientists lack credibility among many members of indigenous groups. (Cunningham and Scharper 1996).

While the HGDP was not directly related to any initiatives to patent genes or cell lines, the NIH was correctly identified as a major sponsor of the HGDP and also as the applicant for a number of gene and cell line patents. In March 1995, the NIH obtained a patent on a cell line infected with a leukemia-associated virus from a man from the Hagahai people of Papua New Guinea (PNG 1996). The NIH and the Centers for Disease Control also pursued patents on cell lines derived from blood donated by a woman from the Guaymi population of Panama and from a person from the Solomon Islands. These patent applications led to considerable negative publicity for the NIH and for the concept of genome "prospecting" by Western scientists among people from less developed countries.

The NIH did not pursue these patents in order to profit from selling drugs or from contracts with drug companies. Rather, US government policy encourages the NIH to transfer information, materials, and intellectual property rights to private companies who can develop and manufacture the drugs and medical devices needed to care for the sick. In the current legal climate, this technology transfer can best be done if the NIH holds a patent,

so that a corporate partner can receive a clear and unencumbered right to develop a drug based on NIH discoveries. This approach makes sense from the perspective of US science policy administrators, but it is counter to the feelings of many advocates for the rights of indigenous peoples. Ruth Liloqula, Director of Agricultural Research for the Ministry of Agriculture and Fisheries of the Solomon Islands, stated: "In our culture, genes are not inventions" (Kreeger 1996). The underlying social contract in American society is that we all support biomedical research—through tax dollars, participation in clinical trials, or through our work as scientists and medical professionals—in return for improvements in healthcare. The reality in indigenous communities in developing countries is that they will benefit much less than the average American from these improvements in healthcare technology, especially those developed by large pharmaceutical companies. In fact, opposition to the HGDP specifically, and against population genetic research in general is quite widespread throughout the world. Examples of official Declarations against the HGDP include the Karioca Declaration (June 1992, Brazil); the Mataatua Declaration (June 1983, Aotearoa, New Zealand); the UN Working Group on Indigenous Populations, 10th Session (July 1993, Geneva); Maori Congress (1993, Aotearoa, New Zealand); National Congress of American Indians (December 3, 1993, Resolution NV-93-118); Maori Congress Indigenous Peoples Roundtable (June 1994); Guaymi General Congress (1994, Panama); Geneva IPR Workshop (August 1994); Latin and South American Consultation on Indigenous Peoples Knowledge, Santa Cruz de la Sierra, Bolivia (September 1994); Asian Consultation on the Protection and Conservation of Indigenous Peoples Knowledge, Sabbah, Malaysia (February 1995), Pan American Health Organisation Resolution (April 15, 1995); and Pacific Consultation on the Protection and Conservation of Indigenous Peoples Knowledge, Suva Statement (May 1995).

The NIH scientists involved in the Papua New Guinea cell line project were working in good faith, following all available guidelines for the fair treatment of human subjects. In fact, they made substantial efforts to be culturally sensitive. The project originated as part of a long-term study of the origin of human retroviruses throughout the world. NIH scientists collaborated with Dr. Carol Jenkins of the PNG Institute of Medical Research, who had been involved with many medical projects with the Hagahai people. Blood samples were collected by Dr. Jenkins from the Hagahai with the prior consent of local Hagahai leaders and the PNG government. Dr. Jenkins was listed as a coinventor of the patent, and she pledged that her share of any patent royalties would be given to the Hagahai people.

Publicity surrounding the NIH patent resulted in a number of public accusations being made against Dr. Jenkins by various organizations working for the rights of indigenous peoples. Rural Advancement Foundation International (RAFI) claimed in a press release that "The United States Government has issued a patent on . . . an indigenous man of the Hagahai people" (Taubes 1995). The PNG Foreign Affairs Secretary later issued a statement exonerating Dr. Jenkins on behalf of the PNG government: "It is clear that this research has been done with the full consent of the Hagahai people as well as approval from the PNG Medical Research Institute and that the benefit of this research, when fully realized, will be shared among all concerned."

Despite the best intentions of the researchers, there is clearly an imbalance in the value proposition when genetic samples are collected from isolated populations. Communities of indigenous people are economically disadvantaged and their way of life is under attack from a variety of social and environmental forces. By participating in genetic research projects, they give up a unique resource, yet receive very little of tangible value in return. Western scientists gain valuable information, which can be used to

advance their personal research interests, their grant applications, and their careers. The information can then move on, via patents, publication, or inclusion in computer databases, so that it can be used by Western corporations for drug development research. The accusation is often true that the researchers care more about the data than about the people from whom it is collected.

Groups that advocate for indigenous peoples argue that despite statements and rules to the contrary, the Diversity Project will be caught up in the current trend to commercialize genes. HGDP scientists, they fear, could simply become agents of the commercial interests of pharmaceutical companies. While a particular research project, such as the HGDP, may not be directly related to the patenting of a certain gene, it is clear that the net result of Western government-sponsored genomics research will be commercial products such as genetic tests and new drugs, and that large corporations will profit from these products—and populations of indigenous people will not. In fact, the specific information produced by the HGDP will be entered into publicly accessible Internet databases—which will be much more valuable to international pharmaceutical corporations than it to the indigenous peoples from whom it was derived. In fact, it is entirely likely that some of the individual scientists involved in the HGDP have relationships with biotech corporations that lead to personal financial benefits or research sponsorship derived in some way from their work with samples from indigenous groups.

There are examples of DNA-based information obtained from isolated populations or indigenous peoples that have been commercialized for substantial profit. Scientists from Sequana Therapeutics (a California-based genomics company) collected samples from the people of Tristan de Cunha, a tiny island of just under 300 inhabitants located halfway between Brazil and South Africa. The inhabitants, who are all descendants of the island's original seven families, exhibit one of the world's highest

incidences of asthma (30% of the population suffer from asthma, and 20% are carriers). Sequana sold the licensing rights to a diagnostic test for asthma to the German pharmaceutical giant Boehringer Ingelheim for $70 million. Although scientists and fieldworkers extracting the samples from indigenous populations follow informed prior consent guidelines, these guidelines often do not mention how their DNA or a product derived from it may become a marketable commodity that could potentially benefit private companies.

The appropriate remedy for these complaints of gene exploitation is both *obvious* and *impossible* under the current system of worldwide patents and intellectual property. The system allows patents for products of scientific work such as "discovered" natural chemicals, DNA sequences, or cell lines. This system has long benefited "chemical prospectors" who travel to remote locations, interview native people about their use of medicinal plants and other natural products, then bring samples of those products back to corporate laboratories to be extracted and chemically characterized. Drugs or other chemical products discovered in the samples can then be sold by the companies without any compensation to the people from whom they were taken. It is clear that fear of the same practices underlie public outcry against the use of human samples from the indigenous peoples for genomic research.

The scientific members of the HGDP have created a rather noble (if practically unworkable) set of principles based on the concept of returning value to the indigenous people who contribute samples to the project:

> The HGDP will not profit from the samples and it will do its best to make sure that financial profits, if any, return to the sampled populations. The best ways to implement these commitments are not yet entirely clear. Implementation depends on some complex issues of patent and contract law that have not been entirely resolved, as well as on some decisions by the sampled populations or their representatives on how best to proceed.

Representatives of groups advocating for the rights of indigenous peoples have argued instead for a worldwide ban on the patenting of any living organism or DNA sequences derived from them. They argue that there is an inevitable connection between the patenting of life forms and the capitalistic expropriation of biological resources. The pharmaceutical industry would counterargue that there is a necessary link between patents and the investment in research that would develop a usable drug or DNA-based diagnostic product. Without patent protection, there can be no practical applications of medical genomics.

There is an important lesson to be learned here. Accusations of genetic exploitation were raised against scientists. After careful examination of the specific cases, all of the scientists involved were completely exonerated. Some additional changes were made in the laws and charters governing the actions of institutions such as the NIH and the HGDP to make them more sensitive to the ethics of working with human genetic material. Yet the overall lasting impression in the public mind remains one of suspicion and distrust. The HGDP is still widely characterized as the "vampire project." Accusations of exploitation are remembered by the public much better than official denials of wrongdoing. Social activists have more public credibility on these issues than NIH scientists. Plans to "educate the public" in order to increase the acceptance of genomic technologies in medicine are likely to run into the same obstacles. During the PNG patent dispute, Temple University anthropologist and former NSF director Jonathan Friedlander noted a "widespread public distrust of the scientific technological enterprise and a willingness to believe the worst of people with scientific knowledge."

Genetic Discrimination

One of the primary public concerns in America and Europe about genomic technologies—particularly about genetic testing—is

the potential for **genetic discrimination**. This is discrimination against an individual or against members of that individual's family (or ethnic group) solely because of differences in DNA sequence that are not associated with any presently observable disease symptoms. This discrimination is based on the notion that people with certain genetic characteristics are at increased risk of disease in the future, and thus would be charged higher insurance rates, denied certain jobs, and so on. Most Americans believe that this information should not be available to employers and insurance companies. Fortunately, governments have been heeding this public sentiment and laws are being put into place to ensure the confidentiality of genetic information and to ban the use of genetic information in employment and health insurance. However, it will be difficult to build up public trust in these laws as adequate protection against genetic discrimination. Clearly the physician will be called on to convince patients as to the adequacy of the laws designed to prevent genetic discrimination when it becomes necessary to obtain consent for a genetic test.

One somewhat naive view being promoted by science policy advisory groups is that the public will accept new genetic technologies once they are better informed about the goals and benefits of the project and better educated about the underlying scientific principles. These scientists and public officials dramatically underestimate the fundamental distrust that many segments of the public hold for institutions such as the government, insurance companies, and the healthcare system, when it comes to genetics. The crucial issue is not that the public does not understand the technology (although it is true that they do not), it is that they trust neither the source of the information nor the motives of the institutions sponsoring the projects.

The case for the social risks of genetic tests is not getting any easier since a few companies will inevitably abuse the technology. In one court case, the Burlington Northern Santa Fe Railroad admitted to the use of genetic tests in an attempt to disprove

claims for compensation from employees with carpal-tunnel syndrome. Employees who filed an injury claim were forced to provide a blood sample. The company submitted the samples for a genetic test for hereditary neuropathy (HNPP) without the consent of the employees. However, since carpal-tunnel syndrome does not have a genetic basis, the illegal testing had no basis in science—more likely it was an ill-conceived attempt to confuse the issue in court cases fighting the payment of workers compensation. The forced testing was stopped by order of the Federal Equal Employment Opportunity Commission (EEOC) in February 2001, and in March the company apologized to the employees who were subjected to testing and agreed to destroy all blood samples and records of the test results, and to pay damages and legal fees. However the net result of this well-publicized case is not to instill faith in the vigilance of federal agencies, but rather to confirm public suspicions that genetic tests will be made in secret and that there will be conspiracies to use the results against people.

Ironically, at the present time there seems to be little evidence that health insurance companies are using genetic information or that they plan to use such information in the future. Among health insurance and health maintenance organization (HMO) companies, the prevailing opinion is that the results of genetic tests have too little predictive value about the *short-term* health prospects of a person. Most Americans change heath insurance plans quite frequently, so long-term health risks are not particularly important in making coverage decisions. Furthermore, the vast majority of Americans have health insurance plans provided by the government (Medicare and Medicaid) or from their employers, where coverage is automatic (cannot be disallowed by current health or preexisting conditions) and premiums are set at the same level for everyone. The Federal Health Insurance Portability and Accountability Act (HIPAA) of 1996 has significantly

diminished the threat of genetic discrimination in group health insurance. HIPAA forbids insurance plans from labeling otherwise healthy people with a genetic condition as having a "preexisting condition."

There is considerably more interest in genetic information among life insurance providers. Many genetic factors can influence a person's life expectancy, and thus the profit or loss that might be expected from a life insurance policy purchased by that person. Insurers have claimed that knowledge of these factors would enable them to more accurately evaluate risk and more fairly set premiums for each person. However, it seems unlikely as a social policy, in America or Europe, that life insurance companies will be allowed to set premiums or deny coverage on the basis of genetic tests. If insurers are prevented from using genetic data, there is considerable concern that individuals may discover through genetic testing that they have a substantial risk of contracting some serious condition, and therefore purchase a large amount of life insurance. If the insurer does not have access to this genetic information, then they could lose money on these policies purchased where a person knows more about her or his own life expectancy than the insurance company does. It is quite difficult to design an equitable social policy that can address this imbalance of information.

Impact on Physicians and Researchers

Many of the ethical issues being raised in regard to medical genomics are not new: safety and efficacy of new treatments, privacy, discrimination, and informed consent. It has long been the case that many kinds of medical information (not just genetic) that resides in computer databases (or paper files) can be damaging to a person if they were to get into the wrong hands—an employer; an insurance company, or the press.

These are some of the key arguments that have been raised about genetic data, all of which also apply to other kinds of medical data:

- *It is predictive*—but so is testing for any kind of infection or condition with a long latency for the development of symptoms such as routine cancer screening or cholesterol levels.
- *Privacy/confidentiality is important*—but obviously this applies for all medical information, including such routine information as family medical history and sexually transmitted diseases.
- *It involves risk factors and probability*—but ordinary testing for cholesterol level implies a risk for heart disease.
- *It has social/family/insurance/discrimination impact*—but so does psychiatric disease, AIDS, and sexually transmitted disease.

However, there are important concerns about informed consent for medical genomics that do require special attention from medical professionals. This is particularly true when genetic tests can reveal potential risk factors that do not indicate that a person currently has a disease, and in fact may never develop it; or a test that may indicate the presence of a disease for which there is no treatment. In fact, it may be quite sensible for a person to choose not to have a genetic test that might predict a future disease for which there is no cure and no preventive measures. The key issue here revolves around providing enough information to the patients so that they can make a truly informed decision about whether to undergo a genetic test (informed consent). There is also a strong requirement for careful communication of test results to the patients so that they can fully understand the implications and make informed decisions about their healthcare, about communicating this information to family members, and so on.

The burden of communication placed on the physician in order to achieve informed consent is indeed a heavy one. Not only must the patient be educated in the relevant molecular biology in order to understand the nature of a deleterious allele and the mechanics of the test by which it is identified; they must also be led to understand the nature of risk and probability. There is a substantial body of social science studies showing that people tend to misinterpret statistical information and make demonstrably poor choices that work against their own interests (e.g., public lotteries and other games of chance). It is not at all clear whether a physician (or genetic counselor) can ever provide enough information to patients to allow them to make a truly informed choice for their own best interests in regard to a genetic test. Furthermore, no healthcare professional ever enters a discussion with a patient without some subtle bias as to what course of action would be best for that patient. Again, the bias of the professional often strongly influences the decision of the client.

One concern that is particularly enhanced in the era of genetic testing is the persistence of genetic data and the ease by which many tests can be applied to a tiny sample, or that genetic information gathered for one purpose can later be reanalyzed to reveal other information about a person. In the era of medical genomics, a routine blood or tissue sample taken for some innocuous purpose (such as a pharmacogenomic test in order to prescribe a drug) can provide a complete genetic profile of a person. Even if the sample is destroyed, the DNA sequence information obtained in one test may reveal other sensitive information—for instance, a gene form that causes susceptibility to a particular drug side effect may also be linked to a higher risk of some form of cancer. Clearly there is a need for strong laws backed up by well-executed policies and procedures to prevent unauthorized genetic testing of people or of access to their genetic information, wherever it may exist, in patient records, computer databases, and so on.

This concern has been addressed by legislation at the state, federal, and international levels. The Genetic Privacy Act was proposed as federal legislation in 1995, and finally voted into law by Congress as the Genetic Information Nondiscrimination Act (GINA) in May of 2008. President Clinton signed an executive order in February 2000, prohibiting federal agencies from using genetic information to discriminate against employees. The "Genetic Nondiscrimination in Health Insurance and Employment Act" (S 318/HR 602) was debated in both the US Senate and House of Representatives in 2001. The bill passed the Senate in 2005, but not the House. The bill prohibits insurers from rejecting anyone or adjusting fees on the basis of genetic information; make illegal genetic discrimination in all areas of employment, including hiring and compensation; and forbid insurers and employers from requiring genetic testing. The American Civil Liberties Union is supporting this law.

The Health Insurance Portability and Accountability Act, often referred to as "HIPAA," was passed by Congress and signed into law by President Clinton in August 1996. Among many other provisions, the law prohibits health insurance plans from considering genetic information in determining eligibility for coverage or in setting premiums. Note that HIPAA applies only to group health insurance, not to private insurance purchased directly by an individual, and not to disability, long-term care, or life insurance.

As of mid-2006, approximately 43 states have enacted their own legislation providing some forms of protection for the confidentiality and use of genetic test results in health insurance; however, the specifics of these laws vary greatly from state to state. These state laws may have prohibitions against genetic discrimination that extend to employment and certain commercial transactions, health insurance, disability, long-term care, and life insurance. These laws place heavy burdens on the physician

to understand and implement these protections. In particular, physicians will need to do the following:

- Identify genetic tests and protected "genetic information" under the law's confidentiality provisions.
- Obtain the required written consents both to conduct a genetic test and to release the test results to anyone other than the patient.
- The physician must provide information to the patient about the reliability of the test and the availability of follow-up genetic counseling.
- Educate office staff and institute office procedures to assure appropriate handling of genetic information.

On learning of all of these new legal obligations, a physician might prefer not to engage in any genetic testing, just to avoid new hassles. However, it turns out that a substantial amount of information that is already in patient charts such as family histories and enzyme tests actually contains potentially protected genetic information under these new laws. Genetic testing in its simplest form takes place in physicians' offices, clinics, and hospitals everyday. Talking to a healthcare provider about your family history can reveal genetic information about your current health and predisposition to disease; and this information becomes part of your permanent medical record. According to Massachusetts Medical Society Vice President Charles A. Welch, M.D., "Since genetic information is ubiquitous in patient records, the requirement that physicians separate genetic information from the medical record is, in many cases, a requirement to do the impossible" (Green and Nicastro 2001). Ironically, the new laws being created because of fear among the general public about the misuse of new genetic testing data may end up improving the overall privacy of medical data.

Clinical Research

Another ramification of new medical data privacy laws, especially HIPAA, is that biomedical researchers will have more difficulty in obtaining research samples. Traditionally, any tissue that was removed from a patient during surgical procedures in a hospital was considered fair game for medical research studies. Often the samples were "anonymized" in some way so that there was no way of identifying the person who "donated" the sample. Many of the new medical privacy laws clearly specify that all tissue samples removed from a person are the personal property of that person, and the hospital can use those samples only in ways for which the person specifically provides **informed consent**. This complicates the research process in many ways. First, it is not always known beforehand what research studies will be conducted with a given sample. For example, tissue may be frozen or stored in other ways for future use, transferred to distant laboratories for one study, and then another study may be added to the project. Even once physical samples of patient tissue are destroyed, the genomic information collected for a consented study may be reanalyzed for another purpose, or incorporated in a database for use in conjunction with other data. How can the patient consent in advance for a study that was not planned at the time the sample was collected? Alternately, it would be very difficult to recontact every patient to consent for some later study on stored samples or information. Also, if documents must be maintained proving that informed consent has been given for the use of each sample in each research study, then those documents themselves become a security risk, since the identity of the patient is linked to the sample.

Just as with routine genetic testing, volunteer participation as a research subject in a clinical trial that involves genomics involves some risk of genetic discrimination. There is a high likelihood that research tissues and data will become incorporated

in permanent biobanks and databases, and there is always the possibility that negative genetic information may be associated back to a particular research participant. If research subjects are fully informed about this possibility, then it may be difficult to obtain their consent to participate in the research study. A high level of trust must be established between patient and scientist in order to provide believable assurances of privacy for relevant medical data.

Another issue in genomics research that was mentioned previously in connection with the Human Genome Diversity Project is the potential for the creation of economically valuable products. The ultimate goal of medical genomics research is to produce drugs and diagnostic tests that will be put into widespread use in healthcare. These products are produced by private companies and sold at a profit, often a very substantial profit. Volunteer participants in clinical trials are essential for the development of these products, yet they do not receive any of the economic benefits. Scientists, hospitals, and universities conducting clinical research often do maintain intellectual property interests in the products of their research. So there is a fundamental ethical question—why should patients participate in clinical research with some possibility of harm to themselves, but no chance for economic benefit? There is no sensible solution available for this problem. It would create an entire new set of ethical dilemmas to pay patients to participate in research, since some patients might be unduly influenced to participate by economic need. If research subjects were compensated directly from the profits derived from commercially successful research, why should these patients receive more benefits than the much larger number of volunteers who participated, at some personal risk, in the many other studies that did not lead directly to a commercial product?

It is obviously counterproductive for privacy and intellectual property laws to block the basic research that is needed to develop the genetic tests, whose results are the subject of this debate.

There would be no debate over genetic testing if there were no clear benefits and widespread public demand for this testing. We all want the benefits of genetic screening for treatable diseases and drugs tailored to our genetic profile without the risks of genetic discrimination. There is no reason why carefully crafted laws cannot be created to reach this goal. However, the legacy of past unethical uses of genetics creates an emotionally polarized debate and deep-seated distrust of scientists, government, and drug, medical, and insurance companies and agencies around these issues.

Despite all of the recommendations of advisory committees and task forces and the adoption of statements of principles by scientific bodies, religious organizations, and governments, the public remains convinced that genetic information can and will be used against them. This is a well-founded fear—and one that is shared by many health care professionals when it comes to their own personal medical records. Clearly, medical professionals are going to face a significant barrier of public mistrust before the benefits of routine genetic testing can be realized.

References

Cunningham H, Scharper S. 1996. *Human Genome Project Patenting Indigenous People. Third World Network Features*, (`http://www.dartmouth.edu/~cbbc/courses/bio4/bio4-1996/HumanGenome3rdWorld.html`; accessed 2/23/96).

Davenport CB. 1910. *Eugenics: The Science of Human Improvement through Better Breeding.* New York, Henry Holt and Co.

ELSI Research Planning and Evluation Group. 2000. *A Review and Analysis of the Ethical, Legal, and Social Implications (ELSI) Research Programs at the National Institutes of Health and the Department of Energy: Final Report of the ELSI Research Planning and Evaluation Group*, Feb 10, 2000.

Galton F. 1909. *Essays in Eugenics*. Eugenics Education Society, London.

Green MJ, Nicastro DP. 2001. *New State Genetic Privacy Act. Vital Signs*. Massachusetts Medical Society.

Herrnstein RJ and Murray C. 1994. *The Bell Curve: Intelligence and Class Structure in American Life.* New York, Free Press.

Image Archive on the American Eugenics Movement (CSHL Eugenics Archive). `http://vector.cshl.org/eugenics`, Dolan DNA Learning Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Kreeger KY. 1996. Proposed human genome diversity project still plagued by controversy and questions. *Scientist* **10** (20):1.

Laughlin HH. 1914. *Report of the Committee to Study and to Report on the Best Practical Means of Cutting off the Defective Germ Plasm in the American Population*. Eugenics Record Office Bulletin, N 10; pp 1–64.

Laughlin HH. 1921. The Second International Exhibition of Eugenics held September 22 to October 22, 1921 in connection with the Second International Congress of Eugenics in the American Museum of Natural History, New York. Baltimore, William and Wilkins Co., 1923.

Lombardo P. A. 1985. Three generations, no imbeciles: New light on buck v. bell. *NY Uni Law Rev* **60**: 30–62.

PNG. 1996. Papua New Guinea Secretary for Foreign Affairs and Trade, press release, March 1996.

Taubes G. 1995. Scientists attacked for patenting pacific tribe. *Science* **270** (5239): 1112.

Genetic Testing: Scientific Background for Policymakers

Amanda K. Sarata
Analyst in Genetics, Domestic Policy Divison

Summary

In the 109th Congress, several pieces of legislation were introduced that related to genetic and genomic technology and testing, including the Genetic Information Nondiscrimination Act of 2005 (S. 306/H.R. 1227), the Genomics and Personalized Medicine Act of 2006 (S. 3822), and the Prenatally Diagnosed Condition Awareness Act (S. 609). Although none of these bills passed, they signal the growing importance of the public policy issues surrounding the clinical and public health implications of new genetic technology. As genetic technologies proliferate and are increasingly used to guide clinical treatment, these public policy issues are likely to continue to garner considerable attention. Understanding the basic scientific concepts underlying genetics and genetic testing

Prepared for Members and Committees of Congress, Congressional Research Service, January 26, 2007.

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

may help facilitate the development of more effective public policy in this area.

Most diseases have a genetic component. Some diseases such as Huntington's disease are caused by a specific gene. Other diseases, such as heart disease and cancer, are caused by a complex combination of genetic and environmental factors. For this reason, the public health burden of genetic disease is substantial, as is its clinical significance. Experts note that society has recently entered a transition period in which specific genetic knowledge is becoming critical to the delivery of effective health care for everyone. Therefore, the value of and role for genetic testing in clinical medicine is likely to increase significantly in the future.

CONTENTS

INTRODUCTION

Virtually all disease has a genetic component.[1] The term "genetic disease" has traditionally been used to refer to rare monogenic (caused by a single gene) inherited disease, for example, cystic fibrosis. However, we now know that all complex diseases, including common chronic conditions such as cancer, heart disease, and diabetes, are the product of some combination of genetic and environmental factors. For this reason, they could all be said to be "genetic diseases." Considering this broader definition of genetic disease, the public health burden of genetic disease can be seen to be substantial. In addition, an individual patient's genetic makeup, and the genetic makeup of his or her disease, will help guide clinical decisionmaking. Experts note that "we have recently entered a transition period in which specific genetic knowledge is becoming critical to the delivery of effective health care for everyone."[2] For this reason, the value of and role for genetic testing in clinical medicine is likely to increase significantly in the future. As the role of genetics in clinical medicine and public health continues to grow, so will the importance of public policy issues raised by genetic technologies.

Science is only beginning to unlock the complex nature of the interaction between genes and the environment in common disease, and their respective contributions to the disease process. The

[1]Collins FS, McCusick VA. 2001. Implications of the Human Genome Project for medical science. *JAMA* **285**:540–544.

[2]Guttmacher AE, Collins FS. 2002. Genomic medicine—a primer. *New Engl J Med* **347**(19):1512–1520.

information gleaned from the Human Genome Project will help, and is currently helping, scientists and clinicians identify common genetic variation that contributes to disease. In addition, research conducted utilizing large population databases that collect health, genetic, and environmental information about entire populations will likely provide more information about the genetic and environmental underpinnings of common diseases. Many countries have established such databases, including Iceland, the United Kingdom, and Estonia. The knowledge of the potential relevance of genetic information to the clinical management of nearly all patients coupled with the lack of complete information about the genetic and environmental factors underlying disease creates a challenging climate for public policymaking.

In many cases, the results of genetic testing may be used to guide clinical management of patients. For example, more frequent screening may be recommended for individuals at increased risk of certain diseases by virtue of their genetic makeup, such as colorectal and breast cancer. In some cases, prophylactic surgery may even be indicated. Decisions about courses of treatment and dosing may also be guided by genetic testing, as might reproductive decisions (both clinical and personal). However, many diseases do not have any treatment available (for example, Huntington's disease). In these cases, the benefits of genetic testing lie largely in the information they provide an individual about his or her risk of future disease or current disease status. The value of genetic information in these cases is personal to individuals, who may choose to utilize this information to help guide medical and other life decisions for themselves and their families. The information can affect decisions about reproduction, the types or amount of health, life, or disability insurance to purchase, or career and education choices. As genetic research continues to advance rapidly, it will often be the case that genetic testing may be able to provide information about the probability

of a health outcome without an accompanying treatment option. This situation again creates unique public policy challenges, for example, in terms of the financing of genetic testing services and education about the value of testing (see S. 609, 109th Congress, for example).

Concerns about privacy and the use and misuse of genetic information, as well as issues of **genetic exceptionalism**[3] and **genetic determinism**,[4] may need to be balanced with the potential of genetics and genetic technology to change how care is delivered and to personalize medical care and treatment of disease.

This report will summarize basic scientific concepts in genetics and will provide an overview of genetic tests, their main characteristics, and the key policy issues they raise.

FUNDAMENTAL CONCEPTS IN GENETICS

The following section explains key concepts in genetics that are essential for understanding genetic testing and issues associated with testing that are of interest to Congress.

CELLS CONTAIN CHROMOSOMES

Humans have 23 pairs of chromosomes in the nucleus of most cells in their bodies. These include 22 pairs of autosomal chromosomes (numbered 1 through 22) and one pair of sex chromosomes (X and Y). One copy of each autosomal chromosome is inherited

[3]**Genetic exceptionalism** is the concept that genetic information is inherently unique, should receive special consideration, and should be treated differently from other medical information.

[4]**Genetic determinism** is the concept that our genes are our destiny and that they solely determine our behavioral and physical characteristics. This concept has mostly fallen out of favor as the substantial role of the environment in determining characteristics has been recognized.

from the mother and from the father, and each parent contributes one sex chromosome.

Many syndromes involving abnormal human development result from abnormal numbers of chromosomes (such as Down syndrome). Other diseases, such as leukemia, can be caused by breaks in or rearrangements of chromosome pieces.

Chromosomes Contain DNA

Chromosomes are composed of deoxyribonucleic acid (DNA) and protein. DNA is comprised of complex chemical substances called bases. Strands made up of combinations of the four bases [adenine (A), guanine (G), cytosine (C), and thymine (T)] twist together to form a double helix (like a spiral staircase). Chromosomes contain almost 3 billion base pairs of DNA that code for about 20,000–25,000 genes (this is a current estimate, although it may change and has changed several times since the publication of the human genome sequence).[5]

DNA Codes for Protein

Proteins are fundamental components of all living cells. They include enzymes, structural elements, and hormones. Each protein is made up of a specific sequence of amino acids. This sequence of amino acids is determined by the specific order of bases in a section of DNA. A gene is the section of DNA that contains the sequence that corresponds to a specific protein. Changes to the DNA sequence, called **mutations**, can change the amino acid sequence. Thus, variations in DNA sequence can manifest as

[5]National Research Council. 2006. *Reaping the Benefits of Genomic and Proteomic Research: Intellectual Property Rights, Innovation, and Public Health,* p. 19. National Academies Press, Washington, DC.

variations in the protein which may affect the function of the protein. This may result in, or contribute to the development of, a genetic disease.

Genotype Influences Phenotype

Although most of the genome is very similar between individuals, there can be significant variation in physical appearance or function between individuals. In other words, although we share most of the genetic material we have, we can see that there are significant differences in our physical appearance (height, weight, eye color, etc.). Humans inherit one copy (or **allele**) of most genes from each parent. The specific alleles that are present on a chromosome pair constitute a person's **genotype**. The actual observable physical trait is known as the **phenotype**. For example, having two brown-eye color alleles would be an example of a genotype and having brown eyes would be the phenotype.

Many complex factors affect how a genotype (DNA) translates to a phenotype (observable trait) in ways that are not yet clear for many traits or conditions. Study of a person's genotype may determine if a person has a mutation associated with a disease, but only observation of the phenotype can determine if that person actually has physical characteristics or symptoms of the disease. Generally, the risk of developing a disease caused by a single mutation can be more easily predicted than the risk of developing a complex disease caused by multiple mutations in multiple genes and environmental factors. Complex diseases, such as heart disease, cancer, immune disorders, or mental illness, for example, have both inherited and environmental components that are very difficult to separate. Thus, it can be difficult to determine whether an individual will develop symptoms, how severe the symptoms may be, or when they may appear.

Genetic Tests

What Is a Genetic Test?

Scientifically, a genetic test is defined as:

> An analysis performed on human DNA, RNA, genes, and/or chromosomes to detect heritable or acquired genotypes, mutations, phenotypes, or karyotypes that cause or are likely to cause a specific disease or condition. A genetic test also is the analysis of human proteins and certain metabolites, which are predominantly used to detect heritable or acquired genotypes, mutations, or phenotypes.[6]

Once the sequence of a gene is known, looking for specific changes is relatively straightforward using the modern techniques of molecular biology. In fact, these methods have become so advanced that hundreds or thousands of genetic variations can be detected simultaneously using a technology called a microarray.

Policy Issues The way genetic test is defined is extremely important to the development of genetics-related public policy. For example, the above scientific definition is very broad and inclusive, but this may not be the best way to achieve certain policy goals. It may sometimes be desirable to limit the definition only to predictive, and not diagnostic, genetic testing (see "What Are the Different Types of Genetic Tests?"). In other cases, it may be desirable to limit the definition to only analysis of specific material, such as DNA, RNA, and chromosomes, but not metabolites or proteins. Considerable variation in the definition of genetic test may be found in the many state genetic nondiscrimination laws. Policies extending protection against discrimination may aim to

[6]Report of the Secretary's Advisory Committee on Genetic Testing (SACGT), *Enhancing the Oversight of Genetic Tests: Recommendations of the SACGT,* July 2000 (`http://www4.od.nih.gov/oba/sacgt/reports/oversight_report.pdf`; accessed 1/05/07).

be as broad as possible, whereas policies addressing coverage of genetic tests may aim to be more limited.

How Many Genetic Tests Are Available?

As of January 5, 2007, genetic tests were available for 1343 diseases. Of those tests, 1046 were available for clinical diagnosis, while 297 were available for research only.[7] The majority of these tests are for single-gene rare diseases. Asked about the realistic promise of genetic technology, Francis Collins, the Director of the National Human Genome Research Institute predicted,

> I think we can count on the availability within the next decade of a panel of genetic tests that are going to be offered to all of us to determine our risk of common illnesses, focused particularly on those diseases for which there is some intervention available for those found to be at high risk.[8]

What Are the Different Types of Genetic Tests?

Most clinical genetic tests are for rare disorders, but increasingly, tests are becoming available to determine susceptibility to common, complex diseases and to predict response to medication.

With respect to health-related tests (i.e., excluding tests used for forensic purposes, such as "DNA fingerprinting"), there are two general types of genetic testing: **diagnostic** and **predictive**. Genetic tests can be utilized to identify the presence or absence of a disease (diagnostic). Predictive genetic tests can be used to predict if an individual will definitely get a disease in the future (predictive–presymptomatic) or to predict the risk of an individual getting a disease in the future (predictive–predispositional).

[7]Gene tests (`http://www.genetests.org`; accessed 1/05/07).

[8]Rabinowitz E. 2003. Genetics in medicine: Hype or real promise? *Health Plan* (Jan/Feb 2003).

For example, testing for mutations in the BRCA1 and/or BRCA2 genes provides probabilistic information about how likely an individual is to develop breast cancer in his or her lifetime (predispositional). The genetic test for Huntington's disease provides genetic information that is predictive in that it allows a physician to predict with certainty whether an individual will develop the disease, but does not allow the physician to determine when the onset of symptoms will actually occur (presymptomatic). In both of these examples, the individual does not have the clinical disease at the time of genetic testing, as they would with diagnostic genetic testing.

Within this broader framework of diagnostic and predictive genetic tests, several distinct types of genetic testing can be considered. Reproductive genetic testing can identify carriers of genetic disorders, establish prenatal diagnoses or prognoses, or identify genetic variation in embryos before they are used in *in vitro* fertilization. Reproductive testing, such as prenatal testing, may be either diagnostic or predictive in nature. Newborn screening is a type of genetic testing that identifies newborns with certain metabolic or inherited conditions (although not all newborn screening tests are genetic tests). Traditionally, most newborn screening has been diagnostic, but recently several states have added some predictive genetic testing to their panels of newborn screening (for example, Maryland includes testing for cystic fibrosis).[9] Finally, pharmacogenomic testing, or testing to determine a patient's likely response to a medication, may be considered either diagnostic or predictive, depending on the context in which it is being utilized.

POLICY ISSUES Generally, predictive genetic testing (both presymptomatic and predispositional), rather than diagnostic

[9]Newborn Screening Programs, Family Health Administration, Maryland Department of Health and Mental Hygiene (`http://www.fha.state.md.us/genetics/html/nbs_ndx.html`; accessed on 1/05/2007).

testing, raises more complex ethical, legal and social issues. For example, issues surrounding insurance coverage and reimbursement for this type of test, especially if no treatment is available, are far more complex than with diagnostic genetic testing. A private insurer may feel that paying for a test that predicts the onset of a disease with no treatment is not cost-effective. Even more complicated are cases where the test only shows an increased probability of getting a disease. In addition, Medicare's screening exclusion means that this type of test generally will not be covered for the elderly population.[10]

Another issue is the oversight of genetic tests. Strong oversight of genetic tests may be more important where the information is probabilistic rather than diagnostic and when a treatment is not available. Finally, issues of genetic discrimination may be different with predictive testing than they are with diagnostic testing. Genetic discrimination may be defined as differential treatment in either health insurance coverage or employment based on an individual's genotype. Discriminatory action based on the possibility of something happening in the future, or even the certainty of it happening in the future, might raise more concern than would action taken based on diagnostic information. With probabilistic genetic information, the health outcome of concern may never manifest, or if it is certain to, may not manifest for decades into the future.

The Genetic Test Result

Genetic tests can provide information about both inherited genetic variations, that is, the individual's genes that were inherited from their mother and father, as well as about acquired genetic

[10]Secretary's Advisory Committee on Genetics, Health, and Society. 2006. *Coverage and Reimbursement of Genetic Tests and Services* (Feb 2006) (`http://www4.od.nih.gov/oba/sacghs/reports/CR_report.pdf`, accessed on 1/05/2007). CMS has interpreted the Medicare statute to exclude coverage of preventive care unless specifically authorized by Congress.

variations, such as those that cause some tumors. Acquired mutations are not inherited, but rather are acquired in DNA due to replication errors or exposure to mutagenic chemicals and radiation (e.g., UV rays).

DNA-based testing of inherited genetic variants differs from other medical testing in important ways: it can have exceptionally long-range predictive powers over the lifespan of an individual; it can predict disease or increased risk for disease in the absence of clinical signs or symptoms; it can reveal the sharing of genetic variants within families at precise and calculable rates; and, at least theoretically, it has the potential to generate a unique identifier profile for individuals. Also, unlike most other medical tests, the stability of DNA means that most genetic tests can be performed on material from a body and continue to provide information after the individual has died.

Genetic changes to inherited genes can be acquired throughout a person's life. Tests that are performed for acquired genetic markers that occur with a disease have implications only for individuals with the disease, and not family members. Tests for acquired genetic changes are also usually diagnostic rather than predictive, since these tests are generally performed after symptoms present.

Pharmacogenomic testing may be used to determine acquired genetic variations in disease tissue (i.e., acquired variations in a tumor) or may be used to determine inherited variations in an individual's drug metabolizing enzymes. For example, with respect to determining acquired variation in disease tissue, a tumor may have acquired genetic changes that make it different from normal tissue that may also render that tumor susceptible or resistant to chemotherapy. With respect to inherited variation in drug metabolizing enzymes, an individual may be found to be a slow metabolizer of a certain type of drug (statins, for example) and this information can be used to guide both drug choice and dosing.

Policy Issues In some cases, people feel differently about their genetic information than they do about other medical information, a sentiment embodied by the concept of **genetic exceptionalism**. This may be based on the stated differences between genetic testing and other medical testing, but also may be based on personal belief that genetic information is powerful and different than other medical information. For this reason, public policies around genetic discrimination in health insurance, employment, and sometimes life insurance proliferated at the state level in the 1990s, and genetic nondiscrimination legislation has been considered by Congress for nearly a decade. Whether genetic information is somehow different from other medical information; whether it can be separated logically from other medical information; and whether it deserves special protection are all important public policy issues.

Pharmacogenomic testing is important because it will help provide the foundation for personalized medicine. Personalized medicine is healthcare based on individualized diagnosis and treatment for each patient determined by information at the genomic level. Many public policy issues are associated with personalized medicine. For example, there is some uncertainty currently as to how health insurers will assess and choose to cover pharmacogenomic testing as it becomes available. In addition, there are issues surrounding the regulation of pharmacogenomic testing. The Genomics and Personalized Medicine Act of 2006 (S. 3822, 109th Congress) considered many of these issues.

Characteristics of Genetic Tests

Genetic tests function in two environments: the laboratory and the clinic. Genetic tests are evaluated based primarily on three characteristics: analytical validity, clinical validity, and clinical utility.

Analytical Validity Analytical validity is defined as the ability of a test to detect or measure the analyte it is intended to detect or measure.[11] This characteristic is critical for all clinical laboratory testing, not only genetic testing, as it provides information about the ability of the test to perform reliably at its most basic level. This characteristic is relevant to how well a test performs in a laboratory.

Clinical Validity The clinical validity of a genetic test is its ability to accurately diagnose or predict the risk of a particular clinical outcome. A genetic test's clinical validity relies on an established connection between the DNA variant being tested for and a specific health outcome. Clinical validity is a measure of how well a test performs in a clinical rather than laboratory setting. Many measures are used to assess clinical validity, but the two of key importance are clinical sensitivity and positive predictive value. Genetic tests can be either diagnostic or predictive and, therefore, the measures used to assess the clinical validity of a genetic test must take this into consideration. For the purposes of a genetic test, positive predictive value can be defined as the probability that a person with a positive test result (i.e., the DNA variant tested for is present) either has or will develop the disease that the test is designed to detect. Positive predictive value is the test measure most commonly used by physicians to gauge the usefulness of a test to clinical management of patients. Determining the positive predictive value of a predictive genetic test may be difficult because there are many different DNA variants and environmental modifiers that may affect the development of a disease. In other words, a DNA variant may have a known association with a specific health outcome, but it may not always be causal.

[11] An **analyte** is defined as a substance or chemical constituent undergoing analysis.

Clinical sensitivity may be defined as the probability that people who have, or will develop a disease, are detected by the test.

Clinical Utility Clinical utility takes into account the impact and usefulness of the test results to the individual and family and primarily considers the implications that the test results have for health outcomes (for example, is treatment or preventive care available for the disease). It also includes the utility of the test more broadly for society, and can encompass considerations of the psychological, social, and economic consequences of testing.

Policy Issues These three characteristics of genetic tests have important ties to public policy issues. For example, although the analytical validity of genetic tests is regulated by the Centers for Medicare and Medicaid Services (CMS) through the Clinical Laboratory Improvement Amendments (CLIA) of 1988 (P.L. 100-578), the clinical validity of the majority of genetic tests is not regulated at all. This has raised concerns about direct-to-consumer marketing of genetic tests where the connection between a DNA variant and a clinical outcome has not been clearly established. Marketing of such tests to consumers directly may mislead consumers into believing that the advice given them based on the results of such tests could improve their health status/outcomes when in fact there is no scientific basis underlying such an assertion. This issue was the subject of a July 2006 hearing by the Senate Special Committee on Aging. In addition, clinical utility and clinical validity both figure prominently into coverage decisions by payers, but a lack of data often hinders coverage decisions, leaving patients without coverage for these expensive tests.

Coverage by Health Insurers

Health insurers are playing an increasingly large role in determining which medical tests are available by deciding which tests they

will pay for as part of patient benefit packages. Many aspects of genetic tests, including their clinical validity and utility, may complicate the coverage decisionmaking process for insurers. While insurers require that a test be approved by the Food and Drug Administration (when required), they also want evidence that it is "medically necessary," that is, evidence demonstrating that a test will affect a patient's health outcome in a positive way. This additional requirement of evidence of improved health outcomes underscores the importance of patient participation in long-term research in genetic medicine. Particularly for genetic tests, data on health outcomes may take a very long time to collect.

POLICY ISSUES Decisions by insurers to cover new genetic tests have a significant impact on the utilization of such tests and their eventual integration into the healthcare system. The integration of personalized medicine into the health care system will be significantly affected by coverage decisions. Although insurers are beginning to cover pharmacogenomic tests and treatments, the high cost of such tests and treatments often means that insurers require very stringent evidence that they will improve health outcomes. In addition, the fact that Medicare does not routinely cover preventive services (unless authorized specifically by Congress) means that coverage for many genetic tests and services, which may be considered preventive, may not be granted under Medicare. As Medicare coverage decisions are often looked to by private insurers as a guide for their own coverage decisions, it is difficult to predict what effect this might have on the uptake and utilization of genetic tests more broadly.

REGULATION OF GENETIC TESTS BY THE FEDERAL GOVERNMENT

Genetic tests are regulated by the FDA and CMS through CLIA. FDA regulates genetic tests that are manufactured by industry

and sold for clinical diagnostic use. These test kits usually come prepackaged with all of the reagents and instructions that a laboratory needs to perform the test and are considered to be products by the FDA. FDA requires manufacturers of the kits to make sure that the test detects what they say it will, in the patient population in which they intend the test to be used. With respect to the characteristics of a genetic test, this process requires manufacturers to prove that their test is clinically valid. Depending on the perceived risk associated with the intended use promoted by the manufacturer, genetic tests must either prove that they are safe and effective, or that they are substantially equivalent to something that is already on the market that has the same intended use.

Most genetic tests are performed, not with test kits, but rather as laboratory testing services (or "homebrew" tests), meaning that clinical laboratories themselves perform the test in-house and make most or all of the reagents used in the tests. Homebrew tests are not currently regulated by the FDA in the way kits are and, therefore, the clinical validity of the vast majority of genetic tests is not regulated. The FDA does regulate certain components used in homebrew tests, known as **analyte-specific reagents** (ASRs), if the ASR is commercially available. If the ASR is made in-house by a laboratory performing a homebrew test, the test is not regulated at all by the FDA. This type of test is known as a "homebrew–homebrew" test.

Any clinical test that is performed with results returned to the patient must be performed in a CLIA-certified laboratory. CLIA is primarily administered by CMS in conjunction with the Centers for Disease Control and Prevention (CDC) and the FDA.[12] FDA determines the category of complexity of the test so that laboratories know which parts of CLIA they must follow. As

[12]See `http://www.cms.hhs.gov/CLIA/08_Waived_PPMP_Laboratory_Project.asp`, accessed 10/16/06.

previously noted, CLIA regulates the analytical validity of a clinical laboratory test only. It generally establishes requirements for laboratory processes, such as personnel training and quality control/quality assurance programs. CLIA requires laboratories to prove that their tests work properly, to maintain the appropriate documentation, and to show that tests are interpreted by laboratory professionals with the appropriate training. However, CLIA does not require that tests made by laboratories undergo any review by an outside agency to see if they work properly. Proponents of CLIA argue that regulation of the testing process gives the laboratories optimal flexibility to modify tests as new information becomes available. Critics argue that CLIA does not go far enough to assure the accuracy of genetic tests since it only addresses analytical validity and not clinical validity.

Glossary

accession number A unique number assigned to a nucleotide, protein, structure, or genome record by a sequence database builder.

algorithm A step-by-step method for solving a computational problem.

alignment A one-to-one matching of two sequences so that each character in a pair of sequences is associated with a single character of the other sequence or with a gap. Alignments are often displayed as two rows with a third row in between indicating levels of similarity. For example

```
GCT---GTCTGAACCCAACCAGACGGAGAATGA
:::   ::: :: : :   :::  :::::: ::
GCTCCTGTCGGACCTCCTGCAGGGGGAGAACGA
```

allele Alternate forms of a gene that occur at the same locus (see **polymorphism**). All of the variant forms of a gene that are found in a population.

α-helix (alphahelix) The most common three-dimensonal secondary structure for polypeptide chains (proteins),

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

determined by Linus Pauling in 1951. It resembles a spiral staircase in which the steps are formed by individual amino acids spaced at intervals of 1.5 Å, with 3.6 amino acids per turn. The helix is held together by hydrogen bonds between the carbonyl group (COOH) of one amino acid residue and the imino group (NH) of the residue four positions further down the chain.

alternative splicing Variations in the process of removing introns from the primary transcript of a gene that lead to different mature mRNAs.

annotation The descriptive text that accompanies a sequence in a database record.

anticodon The 3 bases of a tRNA molecule that form a complementary match to an mRNA codon and thus allow the tRNA to perform the key translation step in the process of information transfer from nucleic acid to protein.

assembly The process of aligning and building a consensus (contig) from overlapping short sequence reads determined by DNA sequencing.

autosomes Chromosomes that are not involved in sex determination.

β-pleated sheet (β-sheet, betasheet) A protein secondary structure in which two or more extended polypeptide chains line up in parallel to form a planar array that is held together by interchain hydrogen bonds. The pleats are formed by the angles of bonds between amino acids in the polypeptide chains.

BAC (bacterial artificial chromosome) A cloning vector based on the naturally occurring F-factor plasmid from *E. coli* that can contain from 100,000 to over 300,000 bases of inserted DNA.

base pairs Hydrogen-bonded pairs of DNA nucleotides. Adenine always pairs with thymidine and guanine bonds with cytosine (A–T and G–C base pairs).

bioinformatics The use of computers for the acquisition, management, and analysis of biological information.

BLAST (basic local alignment search tool) A fast heuristic database similarity search tool developed by Altschul, Gish, Miller, Myers, and Lipman at the NCBI that allows the entire world to search query sequences against the GenBank database over the Web. BLAST is able to detect relationships among sequences that share only isolated regions of similarity. BLAST software and source code is also available for UNIX computers for free from the NCBI. Variants of the BLAST program include blastn (DNA query vs. DNA database), blastp (protein query vs. protein database), blastx (translated DNA query vs. protein database), tblastn (protein query vs. translated DNA database), and tblastx (translated DNA query vs. translated DNA database).

Boolean search terms The logical terms AND, OR, and NOT, which are used to make database searches more precise.

bottleneck A severe reduction in the number of individuals in a population, leading to a reduction in the genetic diversity of that population in later generations.

"Central Dogma" of molecular biology DNA is transcribed into RNA, which is translated into protein (proposed by Francis Crick in 1957).

cDNA Complementary DNA — a piece of DNA copied *in vitro* from mRNA by a reverse transcriptase enzyme.

chiasma The physical crossover point between pairs of homologous chromosomes in the process of recombination that can be observed during the diplotene and diakinetic stages of prophase 1 and during metaphase 1 of meiosis.

chimeria A hybrid, particularly a synthetic DNA molecule, that is the result of ligation of DNA fragments that come from different organisms.

chromosome A complete DNA molecule that carries a set of genes in a linear array. The basic unit of heredity.

class prediction A diagnostic method that reliably categorizes a sample into one of a defined set of classes on the basis of an assay (e.g., acute myeloid leukemia vs. normal).

cloning The process of growing a group of genetically identical cells (or organisms) from a single ancestor. Also, the process of producing many identical copies of a segment of DNA or a gene using recombinant DNA technology.

cloning vector A DNA construct such as a plasmid, modified viral genome, or artificial chromosome that can be used to carry a gene or fragment of DNA for purposes of cloning.

coding sequence The portion of a gene that is transcribed into mRNA.

codon A linear group of three nucleotides on a DNA segment that codes for one of the 20 amino acids (see **genetic code**).

conserved sequence A base sequence in a DNA molecule (or an amino acid sequence in a protein) that has remained essentially unchanged throughout evolution.

contig A consensus sequence generated from a set of overlapping sequence fragments that represent a large piece of DNA, usually a genomic region from a particular chromosome.

diploid A genome (the DNA contained in each cell) that consists of two homologous copies of each chromosome.

divergence The gradual acquisition of dissimilar characters by related organisms over time as two taxa move away from a common point of origin (see **sequence divergence**).

diversity The number of base differences between two genomes divided by the genome size.

domain A discrete portion of a protein with its own function. The combination of domains in a single protein determines its overall function.

dominant An allele (or the trait encoded by that allele) that produces its characteristic phenotype when present in the heterozygous condition.

DNA Deoxyribonucleic acid, the information containing part of chromosomes that is responsible for both the transmission of hereditary traits and the moment by moment control of cellular physiology.

DNA sequencing The laboratory method of determining the nucleotide sequence of a piece of DNA, usually using the process of interrupted replication and gel electrophoresis developed by Fred Sanger.

EMBL (European Molecular Biology Laboratory) The European branch of the three part International Nucleotide Sequence Database Collaboration (together with GenBank and DDBJ), which maintains the EMBL Data Library (a repository of all public DNA and protein sequence data). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. However, database files obtained from EMBL are in a different format than those obtained from GenBank. The EMBL, established in 1974, is supported by 14 European countries and Israel. Like the NCBI, the EMBL also provides extensive bioinformatics tools.

enhancer A regulatory DNA sequence that increases transcription of a gene. An enhancer can function in either orientation and may be located up to several thousand base pairs upstream or downstream from the gene that it regulates.

Entrez Entrez is the online search and retrieval system that integrates information from databases at NCBI. These databases include nucleotide sequences, protein sequences, macromolecular structures, whole genomes, and MEDLINE, through PubMed.

EST Expressed sequence tag—a partial sequence of a cDNA clone created by collecting single sequencing runs from the 3′ and 5′ ends of a cDNA clone.

***E* score (expected value)** The expected value (*E*) is a parameter that describes the number of hits that one can "expect" to see just by chance when searching a database of a particular size. An *E* value of 1 is equivalent to a match that would occur by chance once in a search of that database.

exon A segment of an interrupted gene (i.e., a gene that contains introns) that is represented in the mature mRNA product—the portions of an mRNA that is left after all introns are spliced out, which serves as a template for protein synthesis.

FASTA A fast heuristic sequence similarity search program developed by Pearson and Lipman. Searches for local regions of similarity between sequences, tolerant of gaps. The related program TFASTA compares a protein query sequence to a DNA databank translated in all six reading frames.

FASTA format A simple universal text format for storing DNA and protein sequences. The sequence begins with a > character followed by a single-line description (or header), followed by lines of sequence data.

founder effect Differences in the allele frequencies of a specific subpopulation as compared with the rest of the species due to random differences in the small number of alleles carried by the individuals who were the founders of the subpopulation.

functional genomics The study of the function of every gene and protein in the genome, including roles in metabolism, physiology, development, regulatory networks, and so on.

gap A space inserted into a sequence to improve its alignment with another sequence.

gap creation penalty The cost of inserting a new gap in a sequence when creating an alignment and calculating its score.

gap extension penalty The cost of extending an existing gap by one residue in an alignment.

GenBank A repository of all public DNA and protein sequence data. GenBank is the US branch of the three-part International Nucleotide Sequence Database Collaboration (together with EMBL and DDBJ). GenBank is currently administered by the National Center for Biotechnology Information, National Library of Medicine, in Bethesda, Maryland, a division of the US National Institutes of Health.

gene A segment of DNA sequence (a locus on a chromosome) that is involved in producing a protein. It includes regions that precede and follow the coding region as well as all introns and exons. The exact boundaries of a gene are often ill-defined since many promoter and enhancer regions dispersed over many kilobases may influence transcription.

gene expression The process by which a gene provides the information for the synthesis of protein, specifically, transcription into mRNA followed by translation into protein.

gene expression profile A pattern of changes in the expression of a specific set of genes that is characteristic of a particular disease or treatment (e.g., cancerous vs. normal cells). The detection of this pattern may be limited to a particular type of gene expression measurement technology.

gene family A group of closely related genes that make similar protein products.

gene regulatory network A map of the relationships between a number of different genes and gene products (proteins), regulatory molecules, and so on that define the regulatory response of a cell with respect to a particular physiological function.

genetic code The correspondence between 3-base DNA codons and amino acids that directs the translation of mRNA into protein. There is one "standard" genetic code for all eukaryotes, but some prokaryotes and subcellular organelles use variant codes.

genetic determinism The unsubstantiated theory that genetic factors determine a person's health, behavior, intelligence, or other complex attributes.

genetic engineering See **recombinant DNA**.

genome All of the genetic material in a cell or an organism.

Genome Ontology (GO) A standard set of consistent naming conventions that can be used to describe gene and protein functions in all organisms based on molecular function, biological process, and cellular location.

genome project The research and technology development effort aimed at mapping and sequencing the entire genome of human beings and other organisms.

genomics The use of high-throughput molecular biology technologies to study large numbers of genes and gene products all at once in whole cells, whole tissues, or whole organisms.

GenPept A comprehensive protein database that contains all of the translated coding regions of GenBank sequences.

global alignment A complete end-to-end alignment of two sequences. This can often be misleading if the two sequences are of different length or only share a limited region of similarity.

haplotype A specific set of linked alleles from a group of adjacent genes that are inherited together over a number of generations.

HapMap A database that systematically calculates linkage between SNP markers across the whole genome in human populations.

helix–turn–helix A protein secondary structure found in many DNA-binding proteins. Two adjacent α-helixes are oriented at right angles two each other.

heterozygosity The presence of different alleles of a gene in one individual or in a population. A measure of genetic diversity.

heterozygous An organism (or cell) with two different alleles for a particular gene.

heuristic A computational method based on a process of successive approximations. Heuristic methods are much faster, but may miss some solutions to a problem that would be found using more laborious rigorous computational methods.

HMM (hidden Markov model) A statistical model of the consensus sequence of a sequence family (i.e., protein domain). HMMs are based on probability theory—they are "trained" using a set of sequences that are known to be part of a family

(a multiple alignment), and can then be applied on a large scale to search databases for other members of the family.

homologs Sequences that are similar because of their evolution from a common ancestor.

homology Similarity between two sequences because of their evolution from a common ancestor.

homozygous An organism (or cell) with two identical copies of the same allele for a particular gene.

HSP (high-scoring segment pair) An alignment of two sequence regions where no gaps have been inserted and with a similarity score higher than a threshold value.

identity See **sequence identity.**

informatics The study of the application of computer and statistical techniques to the management of information. In genome projects, informatics includes the development of methods to search databases quickly, to analyze DNA sequence information, and to predict protein sequence and structure from DNA sequence data.

intron (intervening sequence) A segment of DNA that is transcribed, but removed from the mRNA by a splicing reaction before translation into protein occurs.

in vitro (Latin) Literally "in glass," meaning outside the organism in the laboratory. Usually in a tissue culture.

in vivo (Latin) Literally "in life," meaning within a living organism.

Ligase An enzyme that can use ATP to create phosphate bonds between the ends of two DNA fragments, effectively splicing two DNA molecules into one.

linkage A relationship between two genes located nearby on a single chromosome where the combination of alleles found in each parent appear together in the progeny more frequently than would be expected by chance.

linkage analysis The process of locating genes on the chromosome by measuring recombination rates between phenotypic and genetic markers (or finding markers that do not recombine away from a phenotype).

linkage disequilibrium A set of alleles that remain more tightly linked than would be expected by chance among the members of a population.

locus A specific spot on a chromosome—the location of a gene, a mutation, or other genetic marker. A given locus can be found on any pair of homologous chromosomes.

MEDLINE (PubMed) The US National Library of Medicine's bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, and the biological sciences. The MEDLINE file contains bibliographic citations and author abstracts from approximately 3900 current biomedical journals published in the United States and 70 foreign countries. PubMed is a Web-based search tool for MEDLINE.

meiosis The process of double cell division in a reproductive cell that produces haploid gametes.

microsatellite A form of repetitive or low-complexity DNA that is composed of a short sequence [1–15 base pairs (bp) in length] that is tandemly repeated many times. This is often a hotspot for mutations.

minisatellites repetitive DNA sequence composed of tandemly repeating units of 10–100 bp.

mismatch In an alignment, two corresponding symbols that are not the same.

mitosis The process of cell division that produces a pair of daughter cells that are genetically identical to each other and to the parent cell.

motif A region within a group of related protein or DNA sequences that is evolutionarily conserved—presumably because of its functional importance.

mRNA (messenger RNA) RNA molecules that are synthesized from a DNA template in the nucleus (a gene) and transported to ribosomes in the cytoplasm, where they serve as a template for the synthesis of protein (**translation**).

multiple alignment The alignment of three or more sequences—usually done by the progressive pairwise method—which yields an approximate rather than an optimal answer.

mutation A change in DNA sequence.

neutral mutations A change in DNA sequence that has no phenotypic effect (or has no effect on fitness).

NCBI (National Center for Biotechnology Information) A branch of the US National Library of Medicine, which is part of the NIH. The NCBI is the home of GenBank, BLAST, MedLine/PubMed, and Entrez.

noncoding sequence A region of DNA that is not translated into protein. Some noncoding sequences are regulatory portions of genes, others may serve structural purposes (telomeres, centromeres), while others have no known function.

OMIM (Online Mendelian Inheritance in Man) An online database of human genes and genetic disorders authored and edited by Dr. Victor A. McKusick. The database contains textual information, pictures, and reference information. It also contains copious links to NCBI's Entrez database of MEDLINE articles and sequence information.

ORF (open reading frame) A region of DNA that begins with a translation "start" codon (ATG) and continues until a "stop" codon is reached—this is usually understood to imply a protein-coding region of DNA.

orthologs Similar genes (homologs) that perform identical functions in different species—identical genes from different species.

paralogs Similar genes (homologs) that perform different (but related) functions either within a species or in different species—members of a gene family. The line between orthologs and paralogs grows less distinct when proteins are compared between distantly related organisms—is a bacterial protein an ortholog of a human protein that performs an identical function if the two share only 15% sequence identity?

PCR (polymerase chain reaction) A method of repeatedly copying segments of DNA using short oligonucleotide primers (10–30 bases long) and heat-stable polymerase enzymes in a cycle of heating an cooling so as to produce an exponential increase in the number of target fragments.

Pfam An online database of protein families, multiple sequence alignments, and hidden Markov models covering many common protein domains, created by Sonnhammer ELL, Eddy SR, Birney E, Bateman A, and Durbin R. Pfam is a semiautomatic protein family database, which aims to be comprehensive as well as accurate.

pharmacogenomics The use of associations between the effects of drugs and genetic markers to develop genetic tests that can be used to fine-tune patient diagnosis and treatment.

phylogenetics Field of biology that deals with the relationships between organisms. It includes the discovery of these relationships, and the study of the causes behind this pattern.

phylogeny The evolutionary history of an organism as it is traced back connecting through shared ancestors to lineages of other organisms.

plasmid A circular DNA molecule that can autonomously replicate within a host cell (usually a bacterium).

polymorphism A difference in DNA sequence at a particular locus.

position-specific scoring matrix A table of amino acid frequencies at each position in a sequence calculated from a multiple alignment of similar sequences.

posttranscriptional regulation Regulation of gene expression that acts on the mRNA (i.e., after transcription). This includes regulation of alternative intron splicing, polyadenylation, microRNA binding, 5′ capping, mRNA stability, and rates of translation.

posttranslational regulation Regulation of gene expression that acts at the protein level (i.e., after translation). This includes differential rates of protein degradation, intracellular localization and/or excretion, internal crosslinking, protease cleavage, the formation of dimers or multiprotein complexes, phosphorylation, and other biochemical modifications.

primer A short single-straded DNA (or RNA) fragment that can anneal to a single-stranded template DNA to form a starting point for DNA polymerase to extend a new DNA strand complementary to the template, forming a duplex DNA molecule.

ProDom An online protein domain database created by an automatic compilation of homologous domains from all known protein sequences (`SWISS-PROT` + `TREMBL` + `TREMBL` updates) using recursive PSI-BLAST searches.

profile analysis A similarity search method based on an alignment of several conserved sequences, such as a protein motif. The frequency of each amino acid is computed for each position in the alignment; then this matrix of position-specific scores is used to search a database.

progressive pairwise alignment A multiple alignment algorithm that first ranks a set of sequences by their overall similarity, then aligns the two most similar, creates a consensus sequence, aligns the consensus with the next sequence, makes an new consensus, and repeats until all of the sequences are aligned.

promoter A region of DNA that extends 150–300 bp upstream from the transcription start site of a gene that contains binding sites for RNA polymerase and regulatory DNA binding proteins.

ProSite ProSite is the most authoritative database of protein families and domains. It consists of biologically significant sites, patterns, and profiles, complied by expert biologists. Created and maintained by Amos Bairoch and colleagues at the Swiss Institute of Bioinformatics.

protein family Most proteins can be grouped, on the basis of similarities in their sequences, into a limited number of families. Proteins or protein domains belonging to a particular family generally share functional attributes and are derived from a common ancestor.

proteome All of the proteins present in a cell or tissue (or organism).

proteomics The simultaneous investigation of all of the proteins in a cell or organism.

PubMed A Web-based search tool for MEDLINE at the NCBI Website.

query A word or number used as the basis for a database search.

recessive An allele (or the trait encoded by that allele) that does not produce its characteristic phenotype when present in the heterozygous condition. The recessive phenotype is hidden in the F_1 generation, but emerges in 25% of the progeny from an F_2 self-cross. Most genetic diseases are the result of gene defects that are present as recessive traits at low to moderate

frequencies in the population, but emerge in progeny when two parents both carry the same recessive allele.

recombination The crossing over of alleles between homologous chromosome pairs during meiosis that allows for new (nonparental) combinations of alleles to appear among genes on the same chromosome.

recombinant DNA cloning The use of molecular biology techniques such as restriction enzymes, ligation, and cloning to transfer genes among organisms (also known as **genetic engineering**).

replication The process of synthesizing new DNA by copying an existing strand, using it as a template for the addition of complementary bases, catalyzed by a DNA polymerase enzyme.

restriction enzyme A protein, manufactured by a species of bacteria, which recognizes a specific short DNA sequence within a long double-stranded DNA molecule, and cuts both strands of the DNA at that spot.

scoring matrix (substitution matrix) A table that assigns a value to every possible amino acid (or nucleotide) pair. This table is used when calculating alignment scores.

segregation The separation of chromosomes (and the alleles that they carry) during meiosis. Alleles on different chromosomes segregate randomly among the gametes (and the progeny).

sequence identity The percentage of residues identical between two aligned sequences.

sequence similarity The percentage of amino acid residues similar between two aligned protein sequences. Usually calculated by setting a threshold score from a scoring matrix to distinguish similar from unsimilar and counting the percentage of residues that are above this threshold.

shotgun method A sequencing method that involves randomly sequencing tiny cloned pieces of the genome, with no foreknowledge of where on a chromosome the piece originally came from.

signal sequence A 16–30 amino acid sequence located at the amino terminal (*N*-terminal) end of a secreted polypeptide, that serves as a routing label to direct the protein to the appropriate subcellular compartment. The signal sequence is removed during posttranslational processing.

significance A statistical term used to define the likelihood of a particular result being produced by chance. Significance values for sequence similarity searches are expressed as probabilities (*p* values or *e* values) so that value of 0.05 represents 1 chance in 20 that a given result is due to chance.

similarity See **sequence similarity**.

sister chromatids A pair of homologous chromosomes aligned during meiosis.

Smith–Waterman algorithm A rigorous dynamic programming method for deriving the optimal local alignment between the best matching regions of two sequences. It can be used to compare a single sequence to all of the sequences in an entire database to determine the best matches, but this is a very slow (but sensitive) method of similarity searching.

SNPs Single-nucleotide polymorphisms; single-base-pair mutations that appear at frequencies above 1% in the population.

somatic All of the cells in the body that are not gametes (sex cells).

structural proteomics The study of three-dimensional protein structures on all proteins in a cell, tissue, or organism as a guide to gene/protein function.

SwissProt A curated protein sequence database that provides a high level of annotations, a minimal level of redundancy, and high level of integration with other databases. SwissProt contains only those protein sequences that have been experimentally verified in some way—none of these "hypothetical proteins" are of unknown function.

synteny A large group of genes that appear in the same order on the chromosomes of two different species.

systematics The process of classification of organisms into a formal hierarchical system of groups (taxa). This is done through a process of reconstructing a single phylogenetic tree for all forms of life that uncovers the historical pattern of events that led to the current distribution and diversity of life.

taxa A named group of related organisms identified by systematics.

threading A method of computing the three-dimensional structure of a protein from its sequence by comparison with a homologous protein of known structure.

transcription Synthesis of RNA on a DNA template by RNA polymerase enzyme.

transcription factor A protein that binds DNA at specific sequences and regulates the transcription of specific genes.

transduction The transfer of new DNA into a cell by a virus (and stable integration into the cell's genome).

transfection The process of inserting new DNA into a eukaryotic cell (and stable integration into the cell's genome).

transformation The introduction of foreign DNA into a cell and expression of genes from the introduced DNA (this does not necessarily include stable integration into the host cell genome).

translation Synthesis of protein on an mRNA template by the ribosome complex.

TrEMBL (translations of EMBL) A database supplement to Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated into Swiss-Prot.

UniGene An online database (at NCBI) of clustered GenBank and EST sequences for human, mouse, and rat. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and the map location.

Index

Essentials of Medical Genomics, Second Edition By Stuart M. Brown

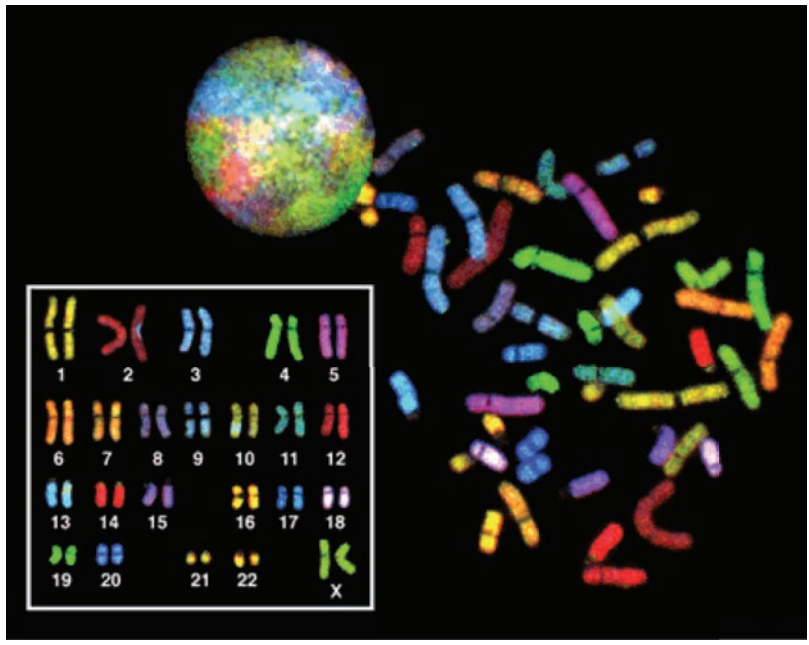


FIGURE 1-1. Human karyotype—SKY image: available at `http://www.accessexcellence.org/AB/GG/sky.gif`; credit to Chroma Technology Inc.

FIGURE 1-4. Anaphase chromosomes in a dividing lily cell.

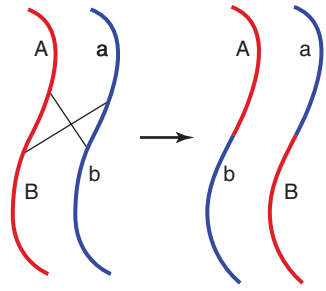


FIGURE 1-6. Schematic diagram of a single crossover between a chromosome with A–B alleles and a chromosome with a–b alleles to form A–b and a–B recombinant chromosomes.

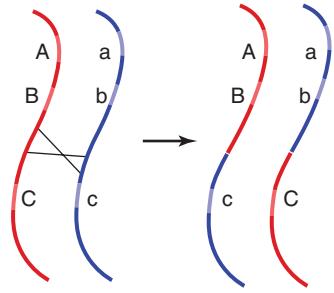


FIGURE 1-7. Genes A and B are tightly linked so that they are not separated by recombination, but gene C is farther away. After recombination occurs in some meiotic cells, gametes are produced with allele combinations ABC, abc, ABc, and abC.

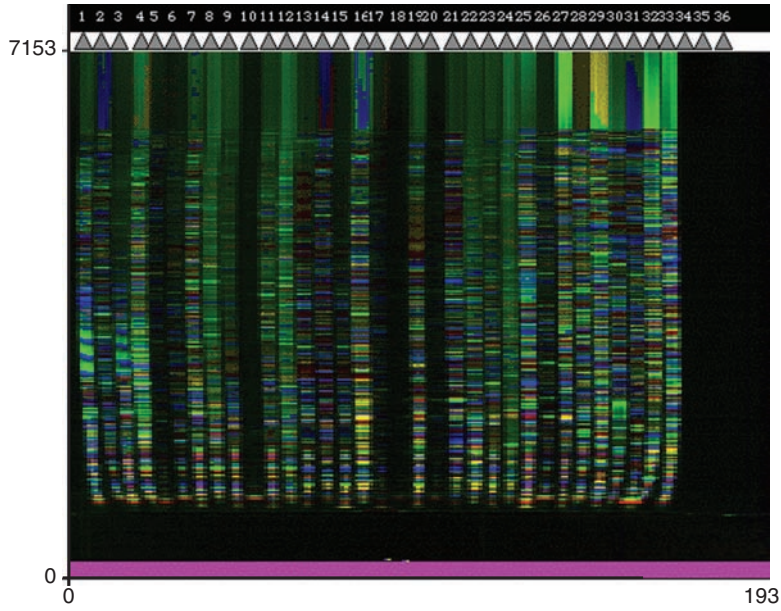


FIGURE 2-10. A fluorescent sequencing gel produced on an ABI automated sequencer. Each lane contains all 4 bases (in different colors).

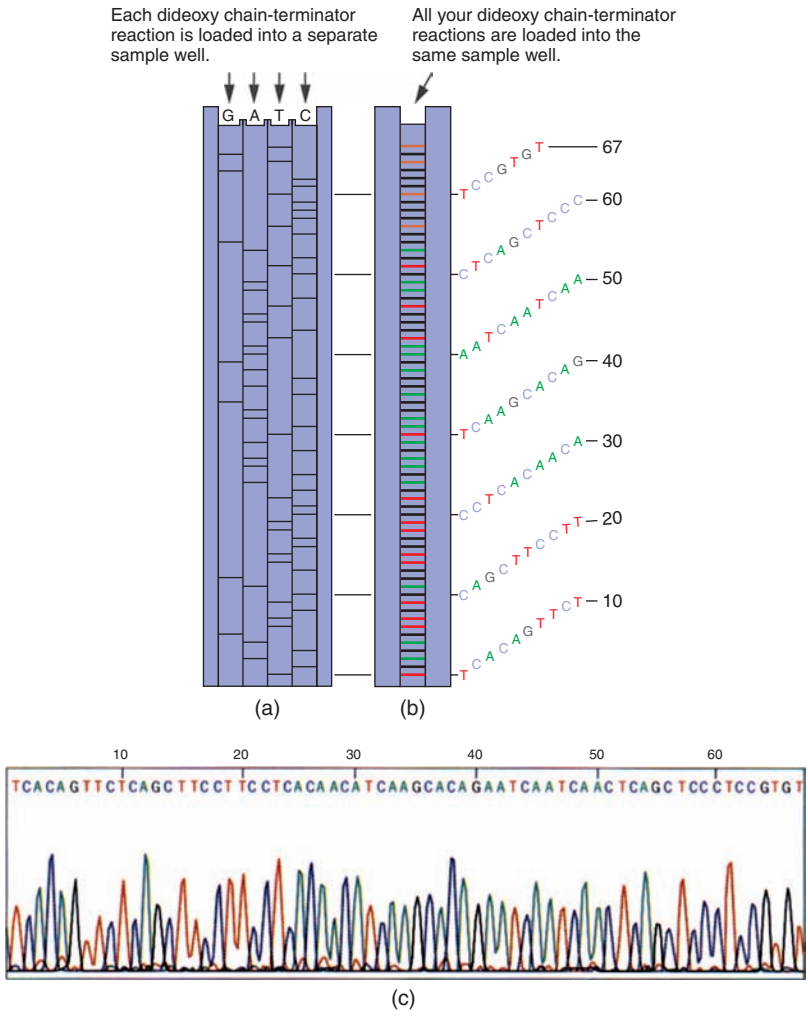


FIGURE 2-11. ABI fluorescent sequencers allow all 4 bases to be sequenced in a single gel lane and feature automated data collection.

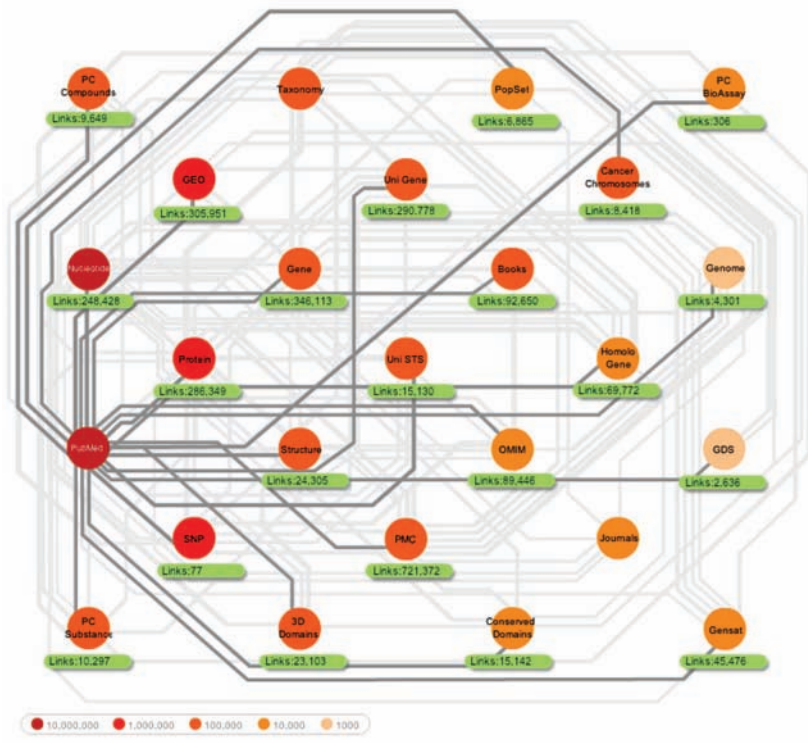


FIGURE 3-2. Links between databases in the Entrez system.

FIGURE 4-1. With computers, it's easy to find patterns, even if they are not really there. These letters can be found in butterfly wings. (Kjell B. Sandred, Butterfly Alphabet, Inc.,Washington, DC.)

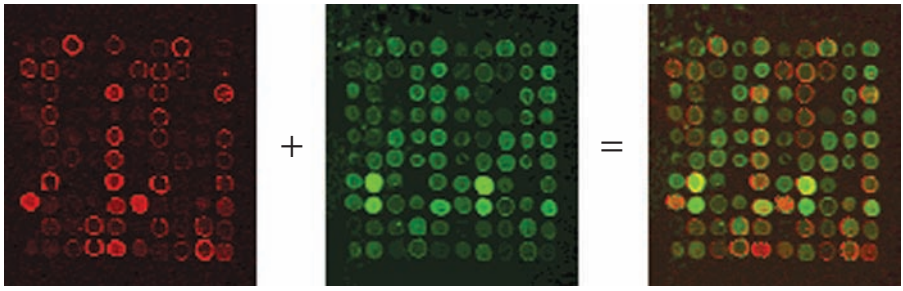


FIGURE 8-2. Two fluorescent images of a microarray with red and green false colors combined to show relative gene expression in two samples.

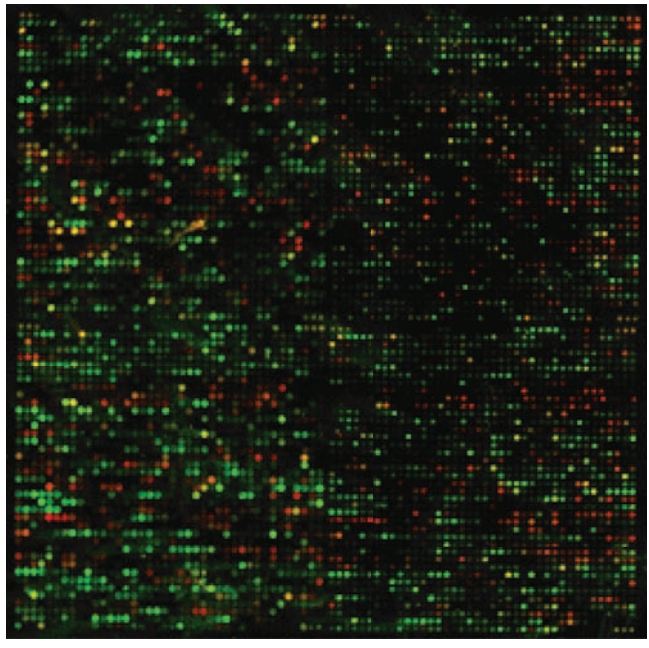


FIGURE 8-7. An example of a spotted cDNA array hybridized with a mixture of two probes with different fluorescent labels and visualized as a red-green false-color image.

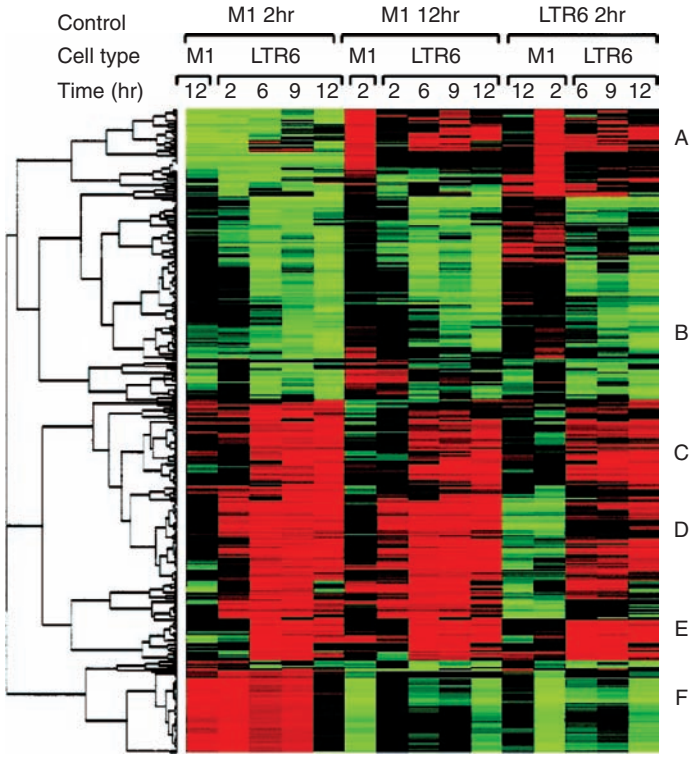


FIGURE 9-5. A two-way clustered heatmap ofmicroarray data as produced by Cluster and TreeView software by Mike Eisen.

FIGURE 12-1. Petunia flower with variegated pattern caused by cosuppression of chalcone synthase (pigment) gene by RNA interference (photo by R. Jorgensen, reprinted with permission).

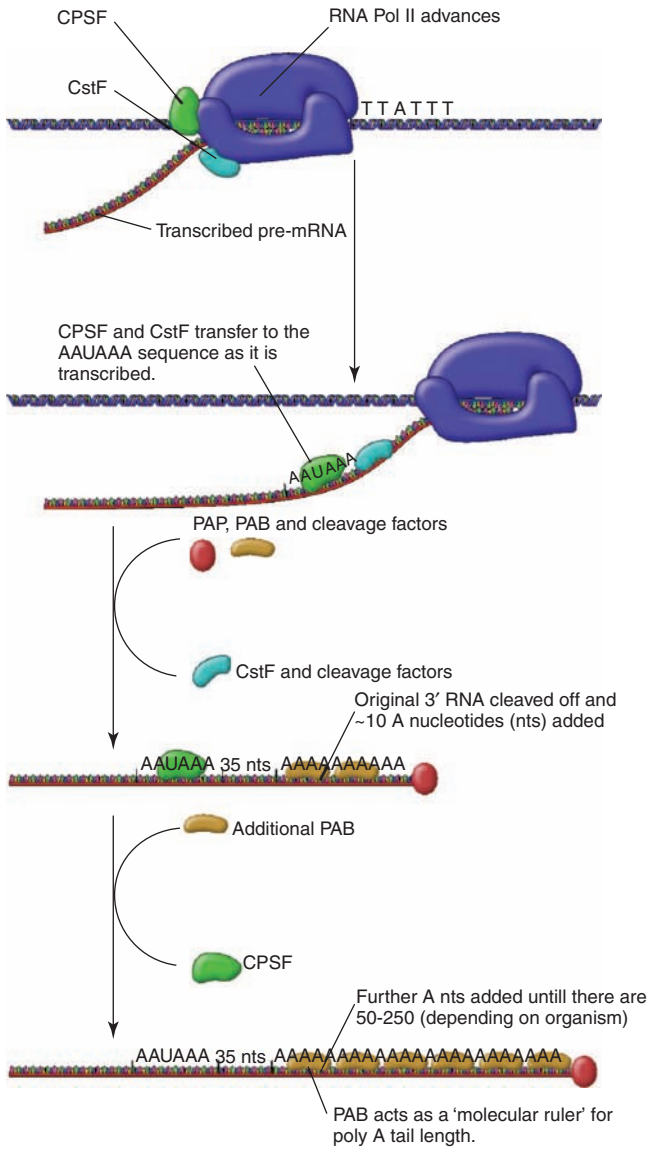


FIGURE 13-2. Addition of the polyA tail is initiated by recognition of AAUAAA signal sequence by cleavage factors. Then polyadenylate polymerase (PAP) adds adenosine residues to form a polyA tail 50–250 bases long.

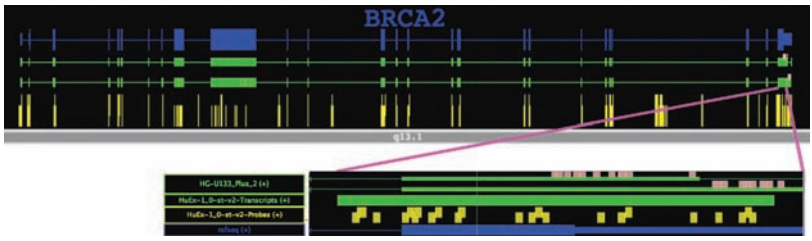


FIGURE 13-8. Affymetrix expression probes (`HG-U133_Plus_2`) and Exon probes (`HuEx-1.0`) for the human BRCA2 gene. The diagram also shows three different predictions for the transcription start site and the RefSeq gene model.

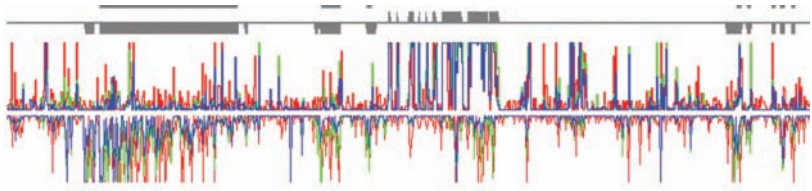


FIGURE 14-3. A portion of the data display for a whole-genome tiling array.

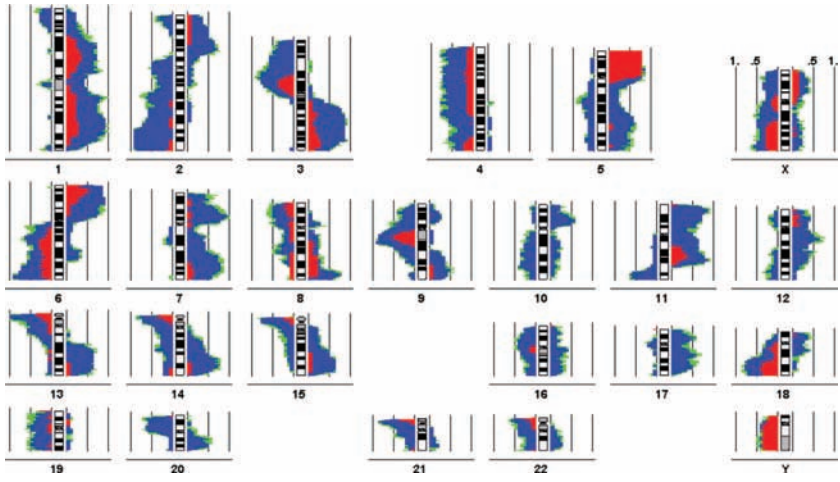


FIGURE 14-5. A display of ArrayCGH data spanning the entire human genome.

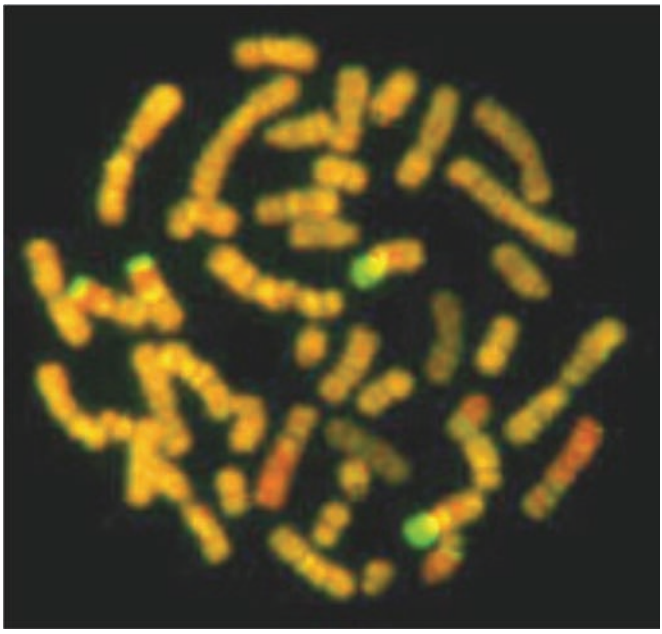


FIGURE 15-1. Hybridization of fluorescent tagged genomic DNA from normal (green) and tumor (red) cells to a chromosome squash from a normal cell. Green regions represent deletions and red regions, amplifications.

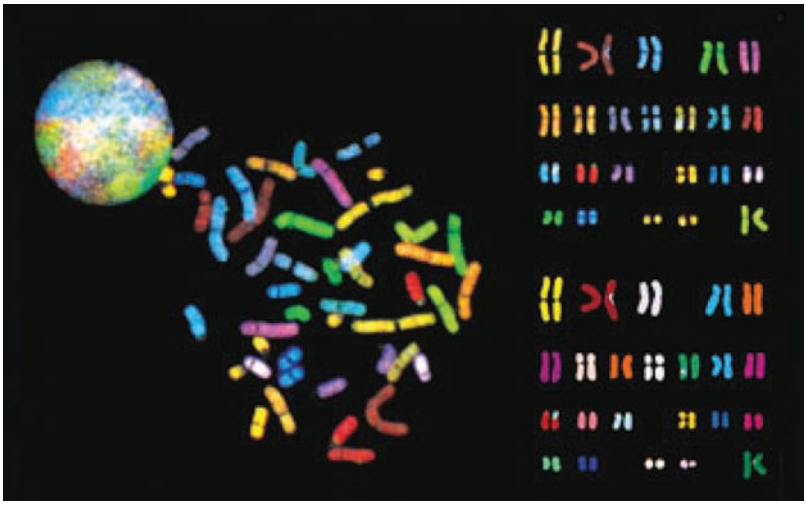


FIGURE 15-3. SKY spectral karyotyping allows for the identification of translocations between different chromosomes.

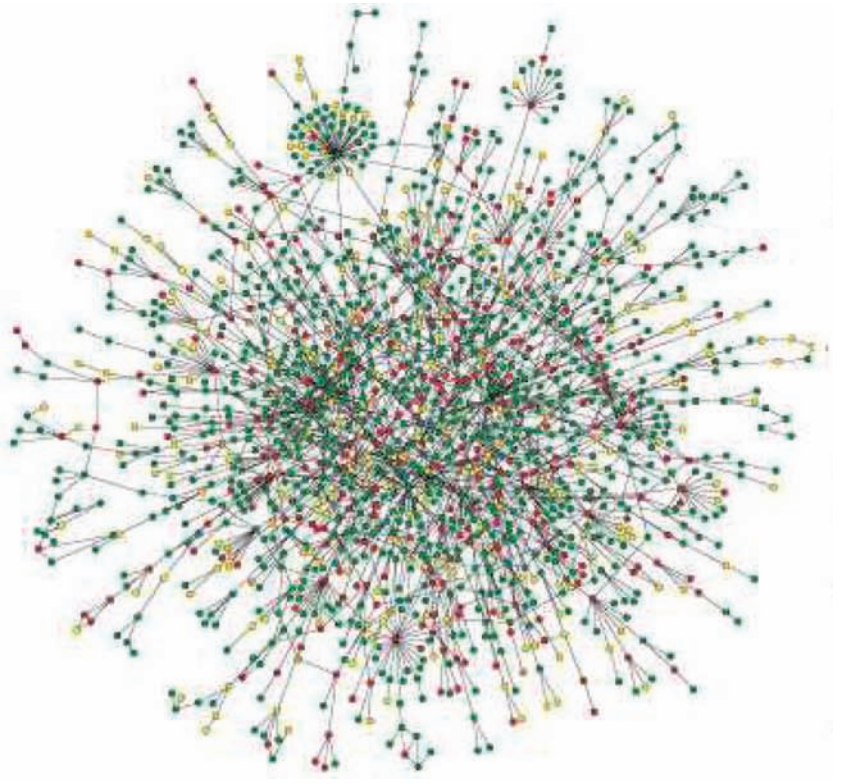


FIGURE 16-2. A map of protein–protein interactions for 1870 yeast proteins (Jeong et al. 2001).

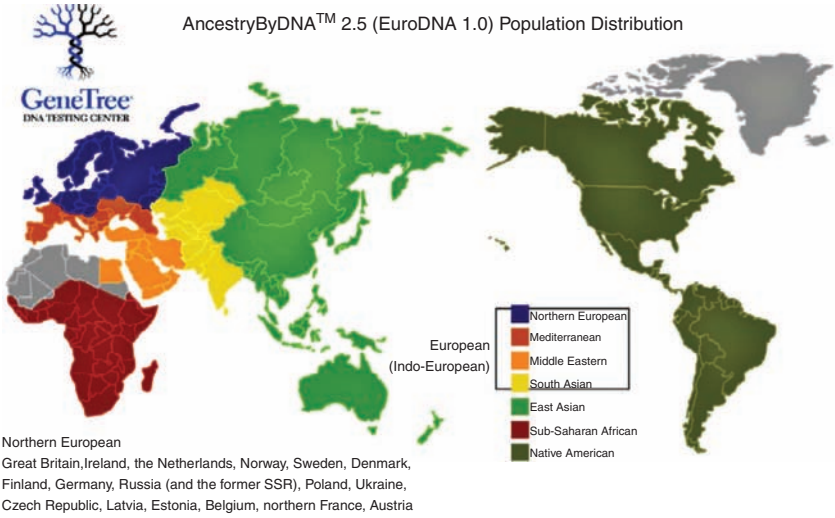


FIGURE 17-1. AncestryByDNA map of human haplotype population distribution.