# The Human Genome Project

## *And Its Impact On The Study Of Human Disease*


*Eric D. Green, M.D., Ph.D.*

*Chief, Genome Technology Branch*
*Director, NIH Intramural Sequencing Center*
*National Human Genome Research Institute*
*National Institutes of Health*
*Bethesda, Maryland 20892*
*Phone: 301-402-0201*
*FAX: 301-402-4735*
*E-Mail: egreen@nhgri.nih.gov*




Published as Chapter 10 in the 8th edition of
*Metabolic and Molecular Bases of Inherited Disease*
(Scriver CR, Beaudet AL, Sly WS, Valle D,
Childs B, Kinzler KW, & Vogelstein B, eds)

# The Human Genome Project And Its Impact On The Study Of Human Disease

1. For many human diseases, the fundamental defect resides in a simple alteration in the genome—the master "blueprint" of DNA that orchestrates the basic operation of a cell and an organism. Genetic studies often provide the ability to define at a molecular level the nature of such DNA alterations (i.e., mutations). Knowledge of the normal and abnormal forms of genes is invaluable for understanding the basis of many human genetic diseases.

2. The haploid human genome consists of ~3 billion base pairs (bp) of DNA that are distributed among 24 distinct chromosomes (22 autosomes and 2 sex chromosomes). Within this vast array of nucleotides are encoded an estimated 50,000 to 100,000 genes and the necessary elements that control the regulation of their expression.

3. Analyzing a genome involves the construction of various types of maps that reflect different features of the DNA, with the major classes being cytogenetic maps, genetic maps, and physical maps. The highest-resolution physical map is the DNA sequence map, which reflects the precise order of nucleotides along a chromosome. Important technologic advances have produced a number of powerful methods that greatly facilitate the ability to analyze genomes.

4. The Human Genome Project (HGP) is a large, coordinated effort to elucidate the genetic architecture of the human genome and, in parallel, that of several model organisms. The initial phase of this endeavor has mostly involved constructing relatively low-resolution genomic maps and refining the approaches for large-scale DNA sequencing. The next phase of the HGP will focus more on establishing the complete nucleotide sequence of the human and other genomes as well as beginning to decipher the encoded information in a systematic fashion.

5. The products of the HGP are providing a detailed working knowledge about the organization of human DNA and that of several model organisms as well as an infrastructure (in the form of biological, informational, and technological tools) that is already ushering in a spectacular new era of biomedical inquiry. From a clinical viewpoint, this infrastructure is facilitating the identification and characterization of genes that directly and indirectly lead to human disease, which in turn should ultimately improve the ability to diagnose and treat affected individuals.

## GENETICS AS A PARADIGM FOR STUDYING HUMAN BIOLOGY AND DISEASE

Diseases are associated with alterations of normal biological processes and can be caused by infectious agents, environmental influences, genetic anomalies, or combinations of these factors. Human disease is classically studied by comparing affected tissues with their unaffected counterparts. Such studies often reveal biochemical and physiological differences, and this information can, in some cases, be used to formulate appropriate therapies. While this approach has led to the development of a successful treatment for many diseases, it frequently fails to identify the fundamental etiology of the disorder itself. Indeed, the differences encountered in affected tissues are often due to secondary effects rather than consequences of the primary defect. However, in cases where DNA sequence alterations (i.e., mutations) are responsible for the disease, it is possible to identify the fundamental defect by a completely different route, one that uses genetics. Studying diseases by a genetic approach takes advantage of the fact that all humans have an almost identical "DNA blueprint." Alterations at one or a few positions in the DNA sequence itself are often necessary and sufficient to cause the symptoms of a genetic disease. The identification of such causative mutations provides an opportunity to study and understand the basic biological defect responsible for that disease.

In humans, genetic studies often start by identifying traits, usually diseases, that appear to cluster in families. Of course, not all diseases that appear multiple times in the same family are genetic in origin, and possible contributions from nongenetic factors must also be considered. In the case of genetic disorders, the challenge is to identify what is often a single-base-pair alteration among the ~3 billion base pairs (bp) in the haploid human genome. Causative mutations for only a few of the thousands of human genetic diseases have been identified by the use of hints from biochemical or physiological differences between affected and unaffected individuals. Because this approach is difficult to apply to most genetic diseases, an alternative strategy, called "positional cloning," has been developed that allows the identification of a disease gene without relying on knowledge or suppositions about the encoded protein.[1-8] Rather, in this strategy, the disease gene is identified on the basis of its location in the genome (see *Background on Positional Cloning* below).

A positional cloning strategy has been used to isolate the causative genes for numerous genetic diseases

(see http://genome.nhgri.nih.gov/clone), including relatively common ones such as cystic fibrosis,[9,10] Huntington disease,[11] and hereditary hemochromatosis.[12] To date, however, the strategy has been mostly applied to those diseases caused by defects in a single gene. Many common disorders (e.g., cardiovascular, autoimmune, psychiatric) have a genetic etiology, but their inheritance is genetically complex, such that mutations in more than one gene are likely required to produce the phenotype.[13-17] In these cases, the responsible genes are difficult to identify by traditional positional cloning for two reasons. First, the strategy must be successfully implemented for the identification of multiple genes. Second, these diseases are complicated by the frequent absence of a strict correlation between genotype and phenotype; instead, an interplay between genetic and environmental factors is typically encountered. Nonetheless, there is increasing optimism that new technologies can be applied in conjunction with increasingly powerful statistical tools to elucidate the genetic bases of complex diseases[13-17].

All of the steps involved in the isolation of human disease genes are labor-intensive and require the marshaling of extensive resources and specialized skills. A central rationale for the Human Genome Project (HGP)—an intense, international effort to clone, map, and sequence all of the DNA in the human genome—is to simplify the task of identifying human disease genes. This chapter provides an overview of how the information and reagents being generated by the HGP and related efforts are being used to advance our understanding of genetic disease and human biology. Because it is important to be aware of some basic concepts and the language of geneticists and genomicists, the chapter begins with some background on DNA structure and function, the general "anatomy" of the human genome, and information about genes and other relevant sequences. A major emphasis of the chapter is to describe the experimental approaches that are being used to generate maps, to determine DNA sequence, and to identify genes. Also discussed is the strategic plan for the HGP and how this carefully crafted endeavor is affecting the study of genetic diseases and other biological problems. An overview of the important role of studying the genomes of model organisms is also provided. Finally, the potential impact of the HGP on the diagnosis and treatment of genetic disease is outlined, and the important ethical, legal, and social issues that are coming to the forefront as a consequence of the HGP are highlighted.

# STRUCTURE AND ORGANIZATION OF THE HUMAN GENOME

## DNA Basics

DNA is a macromolecule composed of a linear array of deoxyribonucleotides, each of which consists of three components: a nitrogenous base, a sugar (deoxyribose), and phosphate. Each base is linked to adjacent bases on the same strand by the sugar and phosphate groups. The bases in DNA are either purines [adenine (A) and guanine (G)] or pyrimidines [cytosine (C) and thymine (T)], and together these nucleotides constitute the "four-letter alphabet" of DNA that is universal among organisms. In the Watson-Crick helical structure of double-stranded DNA, first reported in 1953,[18] pairing occurs between a purine base on one strand and a pyrimidine base on the opposite strand (i.e., G pairs with C, A pairs with T), thereby making each strand complementary to the other. It is the order of these bases that encodes the genetic information contained within DNA. Physical lengths of DNA are frequently discussed as individual base pairs (bp), thousands of base pairs (kilobase pairs, or kb), or millions of base pairs (megabase pairs, or Mb). An excellent source of background information on DNA biochemistry and recombinant DNA technology is *Recombinant DNA* by Watson et al.[19]

## General Structure of the Human Genome

The human genome contains ~3,000 Mb of DNA[20] (Fig. 10-1) divided among 22 autosomes and 2 sex chromosomes (X and Y), which range in size from ~50 to ~260 Mb,[20,21] as well as the DNA present in mitochondria. Human somatic cells are typically diploid (containing 22 pairs of autosomes and 1 pair of sex chromosomes), whereas germ cells (i.e., sperm and egg) are haploid (containing a single copy of each autosome and one sex chromosome). The physical length of the DNA contained in each human cell is remarkably large, theoretically stretching out for about 1 meter if fully "unpacked" from the associated proteins. The amount of

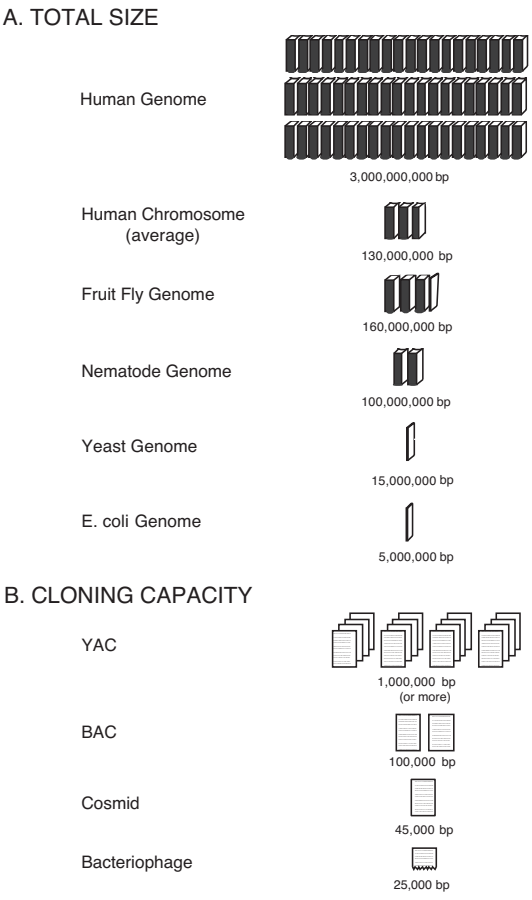


**Fig. 10-1 Relative sizes of genomes, chromosomes, and cloned DNA segments. In panel A, the estimated total sizes of the indicated DNA sources are indicated, along with corresponding schematic representations using books that contain the written DNA sequence. Each book represents ~50,000,000 bp of DNA sequence. In panel B, the total cloning capacities of a YAC, a BAC, a cosmid, and a bacteriophage clone are indicated, along with corresponding schematic representations using pages of books that contain the written DNA sequence. Each page represents ~50,000 bp of DNA sequence, which is roughly 16-times the number of characters on a typical page of this book. (Adapted from Green and Waterston.[474])**

encoded information within the human genome is even more daunting. In fact, a listing of the nucleotide sequence of the ~3,000 Mb in single-letter symbols (G, A, T, C) would fill ~13 sets of the *Encyclopaedia Britannica*,[22] ~750 megabytes of computer disk space, or roughly 1 CD-ROM.[23]

## Subchromosomal Organization of Human DNA

**Structural components of chromosomes.** Human chromosomes are highly organized structures. At the ends of each chromosome are telomeres, which contain sequences thought to stabilize the chromosome, prevent fusion with other DNA, and permit DNA replication without loss of chromosomal material.[24,25] The DNA within a telomere consists of highly reiterated, simple sequence repeats, with the predominant motif in human telomeres being 5′TTAGGG3′.[26] Various human telomeres have been isolated in cloned form,[27-33] allowing more precise dissection of their molecular features and corresponding function. Each chromosome also contains a single centromere, defined functionally by the attachment site of the spindle apparatus during mitosis.[24,34] Human centromeres contain large blocks of repetitive DNA, called "alphoid DNA,"[35-37] which, together with other sequences, span for several megabases. As with telomeres, the cloning and characterization of human centromeric DNA should provide greater insight about the role of specific repetitive sequences in chromosome structure and function.

**Interspersed repetitive DNA sequences.** In contrast to the large and extended blocks of repetitive DNA present in human telomeres and centromeres, most of the remaining regions of human chromosomes contain repetitive sequences that are interspersed among unique segments of DNA.[38,39] The two major classes of interspersed repetitive DNA in the human genome are the short interspersed nucleotide element (SINE) and the long interspersed nucleotide element (LINE). The major SINE is the *Alu* sequence, an ~300-bp segment that is estimated to be present, on average, every 3 to 10 kb (occurring upwards of $10^6$ times in the human genome). The major LINE is the L1 sequence, a segment that spans up to 6.4 kb in length. Often, only a portion of an L1 sequence is present at a particular site, with an estimated $10^4$ to $10^5$ L1 copies (complete or partial) present in the human genome. Together, *Alu* and L1 sequences are thought to account for 10 to 25% of human DNA. In general, copies of the same repeat (*Alu* or L1) present at different sites in the genome are very similar (but typically not identical) in sequence. However, prototypic consensus sequences have been established for the most common human repetitive elements,[40] thereby allowing their identification within stretches of human DNA sequence.

**Coding versus noncoding DNA.** Within the human genome are an estimated 50,000 to 100,000 genes, which can be as small as 100 bp (e.g., the $tRNA^{Tyr}$ gene) to over 2.3 Mb (e.g., the dystrophin gene) in length. However, most human genes are thought to span between 1 and 200 kb of genomic DNA. The amino-acid-encoding portions of genes (i.e., coding DNA) represent a small component of the human genome. In fact, some estimates predict that less than 10% of human DNA reflects coding sequences and their regulatory elements;[41,42] however, a more meaningful assessment of this number awaits more detailed analysis. The remaining noncoding DNA in the human genome consists of repetitive DNA and other sequences whose importance is not completely understood and undoubtedly not fully appreciated.

**Gene structure.** Human genes (also called "transcription units") are complex structures containing several major components. Exons are the segments of DNA in a gene that include the sequences encoding amino acids. Between adjacent exons of a gene are intervening sequences known as introns, which in some cases extend for hundreds of kilobase pairs. Following generation of the corresponding messenger RNA (mRNA) from a gene by transcription, the introns are removed from the mRNA in a series of steps known as splicing. The processed mRNA is then used to direct the sequential and precise addition of amino acids to yield a specific polypeptide chain by a process known as translation. Because different mRNAs can be produced from the same gene by alternate splicing of the primary transcript, there are a larger number of gene products than there are genes—adding to the complex and combinatorial nature of the genome. Interestingly, some introns have been found to contain whole, smaller genes transcribed from the opposite DNA strand.[43,44] Also associated with genes are adjacent regulatory sequences (including promoters, enhancers, inhibitory sequences, and others) that interact with cellular proteins and other components to determine when, where, and to what level transcription (i.e., gene expression) occurs.

**CpG islands.** The dinucleotide CpG (i.e., 5′CG3′) is relatively underrepresented in the human genome, for example, in contrast to the dinucleotide GpC. Among the various enzymes that cleave DNA at precise sequences (called "restriction enzymes") is a class that cuts relatively infrequently within the human genome (called "rare cutters"), most of which contain a CpG dinucleotide within their recognition sequence. Many such rare-cutting restriction enzymes will not cleave the DNA if the nucleotides within the recognition sequence have been modified by the addition of a methyl group (i.e., if they are methylated). Interestingly, at the 5′ end of many human genes are DNA segments that contain an overabundance of unmethylated CpG dinucleotides.[45] These genomic regions are called "CpG islands" (or "HTF islands" for "*Hpa*II tiny fragments," since numerous small DNA fragments are produced from such segments by digestion with the restriction enzyme *Hpa*II). Thus, methylation-sensitive, rare-cutting restriction enzymes can be used to identify undermethylated CpG islands that essentially "mark" the 5′ ends of many (but not all) genes.[45-47]

**Distribution of genes, CpG islands, and repetitive DNA.** The distribution of genes, CpG islands, and interspersed repetitive sequences is not uniform across the human genome. Rather, several interesting patterns are evident that provide some insight about chromosomal organization.[46,48-51] Chromosome preparations can be stained with various agents and examined microscopically; for example, revealing the presence of lighter- and darker-staining regions (or bands) after Giemsa staining (see *Cytogenetic Maps* below). There is evidence that Giemsa-negative (light) bands tend to contain a greater proportion of housekeeping and tissue-specific genes, CpG islands, DNA sequence with a higher GC content, and SINEs. In contrast, Giemsa-positive (dark) bands tend to contain fewer genes and CpG islands, consist of DNA sequence with a lower GC content, and have more LINEs. Another level of chromosomal organization is the presence of DNA blocks that span over 300 kb in length with relatively homogeneous GC compositions (called "isochores"). Interestingly, the composition of genes among different isochores is not uniform, with the highest gene content being associated with the GC-richest isochores.[52-55] Furthermore, there is evidence that many of the gene-richest isochores are located near the ends of human chromosomes (in the subtelomeric regions).[53,56]

## CRITICAL TECHNOLOGIES FOR GENOME ANALYSIS

Central to the HGP has been the development of a number of technologies that are critical for analyzing genomes. A basic understanding of these methods and approaches is necessary to comprehend the experimental bases of most genome mapping and sequencing efforts. Virtually all of the techniques described below represent standard tools in the armamentarium of investigators performing genome research as well as those searching for and characterizing human disease genes. An excellent source of information on the technologies and experimental methods intrinsic to the study of genomes is the four-volume series *Genome Analysis: A Laboratory Manual*.[57]

### Basic Recombinant DNA Techniques

The isolation and characterization of DNA involve the utilization of a fundamental set of techniques that have been refined over the past two decades.[58,59] Most often, the source DNA (e.g., human DNA) is purified and fragmented to yield more manageable-sized pieces. The tools most often used for the latter step are restriction enzymes, each of which cuts double-stranded DNA at a defined sequence of nucleotides. The size of the recognition sequence varies among restriction enzymes (typically from 4 to 8 nucleotides), with those requiring a fewer number of nucleotides cutting more often than those requiring a greater number (the latter being the rare cutters described in *CpG Islands* above). Often, it is necessary to reproduce one or more of the resulting DNA fragments, thereby obtaining sufficient quantities for detailed studies. One way this can be done is by "cloning," whereby foreign DNA is inserted into a rapidly growing organism that is essentially "tricked" into synthesizing the incorporated DNA along with its own. Another way DNA can be reproduced is by the polymerase chain reaction (PCR).

### Polymerase Chain Reaction

Few (if any) experimental techniques have had as dramatic an impact on biomedical research as PCR.[60-62] In the simplest view, PCR involves the in vitro enzymatic synthesis of large amounts of a specific DNA segment. The target DNA is defined by two short (typically 18 to 25 bases each), single-stranded oligonucleotides (primers) that anneal to complementary sequences on opposite strands of the template DNA and initiate (i.e., prime) synthesis back towards one another. The synthesized DNA thus consists of the two oligonucleotides and the sequence between them. Following DNA synthesis, the sample is heated (to greater than 90°C), causing the double-stranded DNA molecules to denature and become single-stranded. Upon cooling, unused oligonucleotides (which are present in excess) anneal to available target DNA molecules, and DNA synthesis is once again allowed to proceed.

The standard cycle (DNA synthesis, denaturation, primer annealing) is repeated 25 to 40 times, with the products of each cycle serving as templates during subsequent cycles. This results in the exponential accumulation (or amplification) of the target DNA sequence defined by the two flanking primers, with the production of as many as a million copies of the target DNA molecule. The size of the DNA segment that can be amplified is typically between 60 and 4000 bp, although segments as large as 10 to 30 kb can be amplified under special conditions. It is important to stress that the critical aspect of PCR is the specificity of the oligonucleotide primers. The ability to amplify a particular DNA sequence often depends on designing an appropriate pair of primers that will uniquely and faithfully anneal to the target DNA under the proper conditions, even when present in a complex mixture such as total genomic DNA.

Like other areas of molecular biology, several important advances have catalyzed the explosive growth of PCR in genome research.[63] These include: (1) the generation of improved thermostable DNA polymerases[64] and PCR-enhancing reagents; (2) the design of more sophisticated instrumentation that improves the efficiency with which PCR assays can be subjected to thermal cycling; and (3) the development of more robust and automated methods for chemically synthesizing large numbers of oligonucleotide primers,[65,66] which in turn has dramatically reduced the cost of synthesizing PCR primers. Together, these and other advances have catapulted PCR to become one of the most widely used experimental methods in research today, including for a wide array of tasks inherent to the study of genomes.[63] As a result, the HGP has benefited tremendously from the use of PCR; ironically, PCR was not yet invented[60,62] when the initial proposals and earliest plans for the HGP were first discussed.[67-69] In retrospect, the thought of embarking on the HGP without PCR is terrifying.

### DNA Cloning Systems

**Standard bacterial cloning systems.** Traditional DNA cloning systems are based in prokaryotic cells, typically the bacterium *Escherichia coli*. For example, plasmids are extrachromosomal DNA molecules that can be engineered to contain relatively small (most often less than 10-kb) pieces of exogenous DNA.[58] A modified form of plasmids, called "cosmids," can accommodate cloned DNA segments upwards of 40 to 45 kb in size[58,70] (Fig. 10-1). Bacteriophages are viruses that infect bacteria, and certain types, such as phage lambda, can be modified to carry up to 25 kb of foreign DNA in cloned form[58] (Fig. 10-1). Because of their capacity for "intermediate" amounts of cloned DNA, cosmid (and to some extent bacteriophage) clones have played important roles in some aspects of genome mapping. For example, strategies for utilizing cosmids to isolate and map large segments of human DNA have been developed;[71-74] however, these efforts have rarely resulted in contiguous cloned coverage extending much beyond 300 to 500 kb. Nonetheless, cosmid and bacteriophage clones serve a valuable supplementary role in the study of human DNA and the isolation of genes of interest.

**Yeast artificial chromosomes.** The yeast artificial chromosome (YAC) cloning system was developed in 1987[75] and provides the ability to isolate DNA segments that are significantly larger than those cloned in traditional bacterial-based systems.[76] In this case, the host is the yeast *Saccharomyces cerevisiae* (a eukaryotic cell), and the cloned DNA is contained within a linear artificial chromosome rather than an extrachromosomal DNA molecule (Fig. 10-2). The cloned DNA contained in YACs can range in size from less than 100 kb to over 1,000 kb, which is roughly 10 to 20 times larger than the capacity of more traditional bacterial cloning systems, such as cosmids (Fig. 10-1). A number of comprehensive YAC libraries have been constructed from human DNA[77-84] and that of numerous other organisms (for review, see Green et al.[76]), and efficient PCR-based strategies for YAC library screening have been developed.[85]

By providing the means to isolate large segments of cloned DNA, YACs have greatly simplified the process of constructing long-range physical maps of DNA. This capability has now been

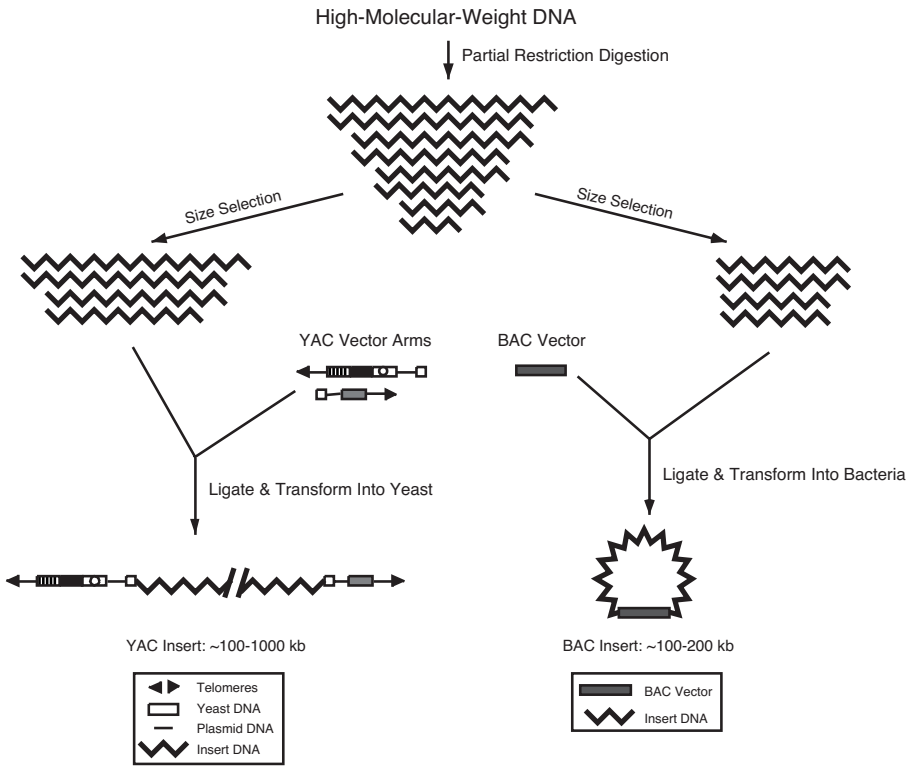


**Fig. 10-2 YAC and BAC cloning systems. The general steps involved in the construction of YAC (left) and BAC (right) clones are summarized. Specifically, high-molecular-weight source DNA (e.g., human DNA) is carefully prepared, partially digested with a restriction enzyme, and size selected to yield large DNA fragments (e.g., typically ~200 to 1000 kb for YACs and ~100 to 300 kb for BACs). Appropriate vector sequences are then ligated to the size-selected, insert DNA. For YACs, this consists of two vector arms that together contain all of the structural elements necessary for the propagation of a chromosome in yeast (see Green et al.[76] for details). For BACs, this consists of a single vector fragment that contains a suitable antibiotic-resistance gene (see Birren et al.[114] for details). The ligated DNA is then transformed into appropriately prepared yeast or bacterial cells, respectively. The systems are set up such that the only cells that grow are those containing the appropriate yeast selectable markers (in the case of YACs) or antibiotic-resistance gene (in the case of BACs). Note that the resulting YACs and BACs are linear and circular DNA molecules, respectively.**

demonstrated by mapping numerous medically relevant regions of the human genome (e.g., those containing the cystic fibrosis gene,[86,87] the dystrophin gene,[88,89] the HLA class I segment,[90] and the Huntington disease gene,[91,92] just to name a few) as well as whole human chromosomes[93-107] (see *Highlights of the Human Genome Project* below). Another novel feature of YAC cloning is the ability to use the yeast host for reconstructing large human genes by the sequential recombination of smaller, overlapping YACs,[76] as has been performed to generate single YACs containing the entire ~200-kb cystic fibrosis gene,[86] the ~230-kb BCL2 protooncogene,[108] and the ~2.3-Mb dystrophin gene.[109]

YAC cloning is not, however, without its associated problems. One disadvantage is the difficulty in purifying large amounts of YAC DNA away from the endogenous yeast DNA. Thus, it often becomes necessary to isolate smaller-insert, bacterial-based clones corresponding to the YAC insert prior to performing manipulations such as DNA sequencing, gene identification, and other routine experimental procedures.[76] A more troubling problem associated with YACs is the frequent presence of two unrelated segments of DNA within a single cloned insert. Such "chimeric" YACs constitute half (or more) of the clones in most libraries made from human genomic DNA.[110] While chimeric clones do not prevent the utilization of YACs for mapping large genomic regions, they can hinder the efficiency and accuracy with which the maps are constructed. A major mechanism by which chimeric YACs form involves recombination between homologous regions (e.g., repetitive DNA) present in unrelated DNA segments.[110] Such yeast-based recombination events likely lead to another problem observed with YACs—the deletion of internal segments within the cloned insert. In the case of chimeric YACs, two different approaches have been successfully used to decrease the problem: the construction of YACs from individual human chromosomes residing within human-rodent hybrid cell lines[76] (see *Somatic Hybrid Cell Lines* below) and the use of recombination-defective yeast strains as hosts.[76] The latter approach has also proven effective at reducing the occurrence of internal deletions in YAC inserts.

**Large-insert bacterial cloning systems.** Since the advent of YAC cloning, several new large-insert bacterial cloning systems have been developed. Among these are the bacteriophage P1 system, with a cloning capacity of roughly 75 to 100 kb,[111,112] as well as the bacterial artificial chromosome (BAC)[113,114] and closely related P1-derived artificial chromosome (PAC)[115] systems, with the latter two providing cloning capacities upwards of 200 to 300 kb.

In particular, BAC cloning has rapidly emerged as a critical tool for genome analysis (Figs. 10-1 and 10-2). Numerous BAC libraries, with clones averaging 100 to 200 kb in size, have been constructed[116,117] (also see http://bacpac.med.buffalo.edu). These can be screened by PCR- and hybridization-based methods;[114,118] the latter has proven particularly robust and involves the use of membranes containing immobilized clone DNA arrayed at very high densities by robotic workstations.[72,119-123] Furthermore, BAC inserts appear to be quite stable during propagation.[113] As a result, BACs are being used for constructing high-resolution maps of human chromosomes[124-126] and numerous other genomes. Such BAC-based physical maps will undoubtedly supplant the first-generation YAC-based maps as well as provide the necessary templates for systematic sequencing of the human and other genomes (see *Genomic Sequencing* below).

## Pulsed-Field Gel Electrophoresis

An important adjunct technology that played a critical role in the development of YACs, BACs, and other large-insert cloning systems is a technique that allows the separation of high-molecular-

weight DNA molecules. Conventional approaches for gel electrophoresis (typically in agarose gels) can resolve DNA molecules that are upwards of ~50 kb in size; however, such methods are incapable of separating significantly larger DNA fragments. To overcome this limitation, techniques have been developed whereby the direction of the electric field applied to the DNA within an agarose gel is periodically alternated.[127-129] This method, called "pulsed-field gel electrophoresis," can be used to separate DNA fragments up to ~10 Mb in size. Numerous refinements and modifications of the basic approaches for pulsed-field gel electrophoresis have been made, making it a routine and straightforward method.[130-132] As a result, pulsed-field gel electrophoresis has been used to study the genomes of model organisms such as yeast,[133,134] to establish long-range restriction maps of human DNA using rare-cutting restriction enzymes,[135-139] to characterize the DNA in large-insert clones such as YACs[75,76,140,141] and BACs,[114] and to detect certain types of mutations causing human genetic diseases.[142]

## Fluorescence *In Situ* Hybridization

A common step in the characterization of a cloned DNA segment is the identification of the approximate site in the genome from which it originated (i.e., its location on a particular chromosome). The most direct route for obtaining such information involves hybridizing the DNA segment to preparations of intact chromosomes from metaphase cells using protocols that allow the structural features of the condensed chromosomes to be preserved. If the DNA probe is labeled appropriately, the position(s) of hybridization can be identified by microscopic examination of the chromosomes, thereby allowing assignment of the DNA segment to a particular subchromosomal region. Previously, radioactive labels were employed, which required lengthy exposure of the chromosomes to film and resulted in poor precision of the chromosomal assignments. Major advances in this technology have occurred in recent years, including the development of protocols for using fluorescent tags to label the DNA probes and fluorescence microscopy to establish the positions of hybridization,[143] a technique referred to as "fluorescence *in situ* hybridization" or FISH (Fig. 10-3).

Continued improvements in the protocols used for performing FISH analysis have greatly enhanced the technology. In general, the basic approaches are now quite efficient and reliable.[144-148] Furthermore, the ability to resolve closely spaced DNA segments has also improved. For example, standard FISH analysis with metaphase-chromosome preparations can be used to discriminate between regions separated by roughly 5 to 10 Mb.[148-150] However, the use of fluorescent tags of different colors in conjunction with specialized methods for preparing the immobilized chromosomes now allows more closely spaced DNA segments to be resolved.[145-148,151-153] Since FISH analysis can establish the position of a cloned DNA segment relative to the source chromosome, this technique has played an important role in genome mapping by spatially organizing clones being used for constructing long-range maps of human chromosomes[154,155] (see *Cytogenetic Maps* below). Variant methods have been developed for making FISH an even more powerful technology, including techniques that allow each human chromosome to be visualized in a different color.[156-158] Such methods offer great promise for more accurate, robust, and potentially automated karyotypic analysis of human chromosomes.

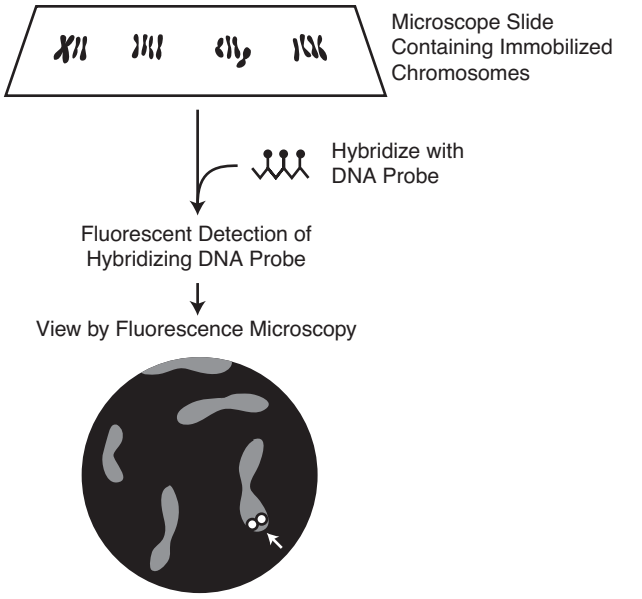


**Fig. 10-3 Establishing the chromosomal position of a DNA segment by fluorescence in situ hybridization (FISH). Intact chromosomes from cells at metaphase are carefully immobilized on a microscope slide. An appropriate DNA probe (e.g., genomic clone) is labeled with a detectable moiety, such as biotin (depicted as dark circles), and hybridized to the immobilized chromosomes. The position(s) of the hybridizing DNA probe is then detected using an appropriate fluorescence-based system (e.g., fluorescently labeled avidin, which binds to biotin). When examined by fluorescence microscopy, the hybridizing probe (indicated by the white arrow) typically appears as two bright yellow spots (one on each chromatid) against an orange background of the chromosome. The approximate chromosomal position of the hybridizing probe can be assessed by parallel examination of the same metaphase chromosomes following appropriate staining.[145-148] (Adapted from Hozier and Davis.[144])**

## Somatic Hybrid Cell Lines

Various somatic cell lines have been constructed that contain the entire genome of the host species (e.g., rodent) along with some amount of foreign DNA from another species (e.g., human). In particularly useful cases, the foreign DNA consists of a single, intact chromosome. For example, such a monochromosomal human-rodent hybrid cell line has been derived for each human chromosome (see http://locus.umdnj.edu/nigms). In some cases, sets of hybrid cell lines, each containing a defined portion (but not all) of a particular human chromosome, are also available. Human-rodent hybrid cell lines thus provide access to more limited parts of the human genome. Of course, the human DNA is not "pure," rather it is mixed within a background of the entire rodent genome (which is roughly the same size as the human genome). The stability of human DNA within human-rodent hybrid cell lines varies widely, especially since there is typically no selective pressure for the rodent cells to retain all of the human DNA. As a result, investigators must be cautious when using such cell lines, which typically should include the routine assessment of the presence and intactness of the human chromosome (or fragment) in the cell line (e.g., by cytogenetic analysis). Human-rodent hybrid cell lines have been used extensively to generate genomic clones (e.g., cosmids, YACs, BACs) from more limited regions of the human genome (e.g., see *Yeast Artificial Chromosomes* above) and as starting material for the development of DNA markers (see *Generation of Sequence-Tagged Sites* below).

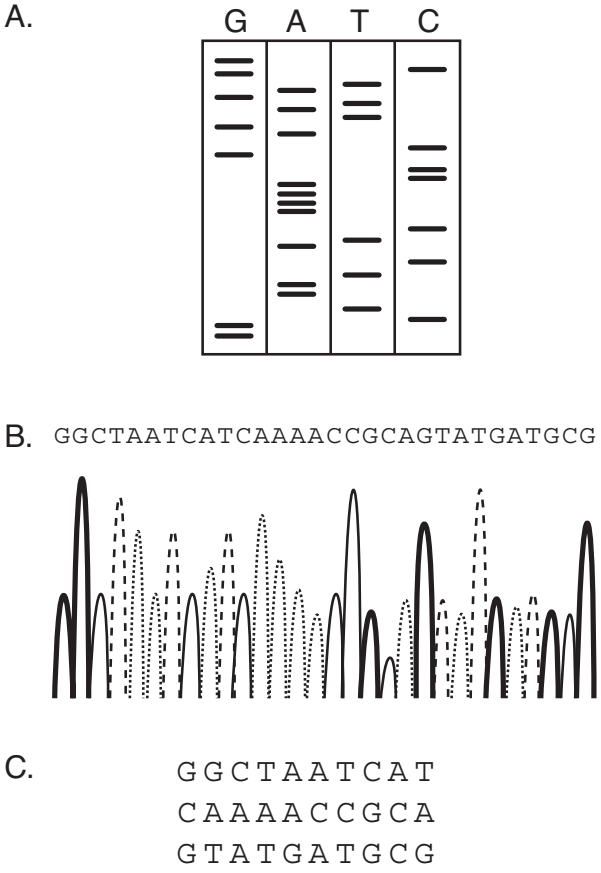


**Fig. 10-4 DNA sequencing by the dideoxy chain-termination method. The dideoxy chain-termination sequencing method yields a population of DNA segments of different lengths, each terminated by a particular type of dideoxy nucleotide (i.e., G, A, T, or C) which "marks" the position of that base in the starting template.[162] The resulting products are then separated by gel electrophoresis. In panel A, the detection of radioactively labeled DNA fragments by autoradiography is shown. In this case, each lane contains the DNA segments generated with a single type of dideoxy nucleotide, with the presence of a band at a particular "rung" position of the sequencing "ladder" reflecting the base at that position in the starting DNA template (the corresponding sequence is shown in panel C). In panel B, the automated detection of fluorescently labeled DNA fragments is illustrated. In this case, a laser is used to detect the migration of the fluorescently labeled, dideoxy-terminated DNA fragments as they are electrophoresed through the gel.[163,166] Each type of peak (indicated by thick, dotted, dashed, and thin lines) reflects the base (G, A, T, and C, respectively) at that position in the starting DNA template (the corresponding sequence is shown in panel C).**

One specialized use of somatic hybrid cell line technology is for constructing radiation hybrid cell lines. Specifically, this involves the recovery of radiation-generated chromosome fragments (e.g., from human chromosomes) within a heterologous cell line (e.g., rodent). The resulting cell lines can be used to construct landmark-based physical maps of large genomic regions[159,160] (see *Radiation Hybrid Mapping* below).

## DNA Sequencing

A handful of techniques have been developed for determining the precise sequence of nucleotides within a stretch of DNA. Some of these methods have been successfully used to generate large amounts of DNA sequence data, while others remain in more developmental stages.[161] The approach described by Sanger and coworkers in 1977,[162] termed "dideoxy chain-termination sequencing," is the most widely utilized sequencing method. This technique involves the in vitro synthesis of DNA molecules in the presence of artificial (dideoxy) nucleotides, which prevent chain extension when incorporated into a growing DNA strand. The resulting population of DNA molecules, which terminate at different nucleotide positions, is then analyzed by gel electrophoresis. The relative migration of the various DNA fragments is used to deduce the sequence of the starting DNA template. Detection of the DNA fragments can be accomplished by the incorporation of radioactive or fluorescent tags[163,164] into the DNA (Fig. 10-4). Following radioactive sequencing, the gels are exposed to X-ray film to allow detection of the DNA fragments. In contrast, fluorescence-based sequencing involves semiautomated, real-time detection of DNA fragments during electrophoresis by laser-based instrumentation.[163,165]

Numerous advances in fluorescence-based sequencing have been made in recent years that make this approach for large-scale DNA sequencing remarkably efficient and robust.[166] These have included the generation of fluorescent-dye-labeled primer[167-169] and fluorescent-dye-labeled terminator[170,171] molecules with improved spectral characteristics and sequencing enzymes that yield longer reads of higher accuracy.[166] Newer sequencing instruments, such as those that employ capillaries rather than slab gels for separating the DNA fragments, are now available as are better sequencing enzymes and associated reagents.

## DNA Chips

A rapidly evolving area of genome research involves the development and utilization of miniaturized DNA microarrays (called "DNA chips") for performing DNA analysis.[172-177d] Essentially, various standard molecular genetics methods (e.g., PCR, DNA-DNA hybridization, fluorescence-based image acquisition) are implemented on a microscale, allowing large numbers of analyses to be performed in a massively parallel fashion. Several distinct, although interrelated, DNA chip technologies have emerged and will undoubtedly evolve into powerful new tools for performing genome analysis.

One major type of DNA chip contains a high-density array of short, single-stranded oligonucleotides (~20 nucleotides long).[178-180a] These are typically immobilized during their synthesis by a process known as photolithography.[178,179] Very high densities of oligonucleotides can be created (e.g., 100,000 to 400,000 oligonucleotides within a 1.28-$cm^2$ area). Target DNA, typically PCR amplified and labeled with a fluorescent tag, is hybridized to the immobilized oligonucleotides. The resulting hybridization pattern is then captured with the aid of a microscope and analyzed with suitable software, allowing the precise sequence of the target DNA to be deduced. Thus, in a simple sense, oligonucleotide-based DNA chips should be regarded as "resequencing" chips. This technology has been developed most extensively by the company Affymetrix. Numerous applications are readily apparent for such an efficient chip-based, resequencing method, with successful implementation already demonstrated for analyzing important sequences within pathogens such as HIV-1[181] and *Mycobacterium*,[182] mapping genomic clones,[183] resequencing human mitochondrial DNA,[184] performing mutation screens for human disease genes,[185,185a] comparing sequences of closely related organisms,[186] simultaneously studying the expression of large numbers of genes,[187,188,188a] and performing large-scale analysis of polymorphisms,[189,189a] such as single nucleotide polymorphisms (SNPs)[190] (see *Genetic Markers* below).

The other major type of DNA chip contains a high-density array of short, double-stranded DNA fragments.[191-196c] Typically, these are first derived by PCR amplification and then immobilized onto glass or nylon surfaces by high-speed robotic workstations. A common implementation scheme for this involves the arraying of cDNA fragments on glass microscope slides at densities of >1000 per 1 $cm^2$. For gene expression studies, hybridization is typically performed with fluorescently labeled mRNA probes, and often includes the use of a two-color fluorescence detection strategy that allows the simultaneous examination of parallel mRNA samples derived from different sources. Once again, the resulting hybridization patterns are captured microscopically, with the resulting data analyzed using appropriate software and assimilated in suitable database systems.[197,197a] Thus, this technology can be used to study the expression of literally thousands of genes in different cells or tissues, all within a single experiment. cDNA microarrays have been used to examine gene expression patterns in cells grown under distinct metabolic conditions[193,195,196] and in human cancer.[194,194a] The true power of such expression profiling technology comes in the ability to monitor very large numbers of genes in parallel, thereby gaining insight about the global and integrated networks regulating gene expression.

Variant types of DNA chips include those that contain highly sophisticated, microfabricated systems for performing standard DNA analyses (e.g., PCR, gel electrophoresis, DNA sequencing, genotyping) on a microscale.[198-205] Here, various technologies have been engineered to allow the handling and analysis of very small DNA samples in a rapid and automated fashion. The development and refinement of such microsystems has tremendous potential to yield important genome analysis tools for use in both research and diagnostic applications.

### Computational Genomics

Among the key genomic technologies are those involving the computational analysis of mapping and sequencing data. The rapidly growing field of computational genomics (or bioinformatics), which encompasses everything from computer-based tools that are required for generating genomic data to those that are essential for using it, now represents a well-respected discipline of biomedical research.[206-210] Most major areas of computational genomics are (in some fashion) heavily dependent on the World Wide Web (Web). It is quite fitting, therefore, that the HGP and a major growth spurt of the Web are occurring at the same time; certainly, the former would be greatly weakened by the absence of the latter.

The generation of genomic maps and sequences relies extensively on the use of ever-improving computational tools.[206-210] These include a wide array of data management and analysis programs that have proven instrumental for performing large-scale genome analysis, particularly for the construction of genomic maps and the generation of genomic sequence data.[211-215] In addition, the availability of Web-based mechanisms for data dissemination allows less refined, evolving data to be made available long before they are at a stage suitable for final publication. The latter has been tremendously important to numerous investigators, who have utilized such preliminary data to accelerate their ongoing studies.

The utilization of massive amounts of accumulating genome mapping and sequencing data is also critically reliant on the availability of powerful and user-friendly computational tools.[206-210] Among these are systems for data storage and retrieval, such as the public databases that house nucleotide and protein sequences.[207,216-223] These databases and associated tools provide a critical service to the biomedical research community by storing, analyzing, cross-referencing, and disseminating all publicly available mapping and sequencing data, including those being produced by the HGP. In addition, there are ever-improving programs for performing sequence comparisons[224-228] and for predicting the presence of genes[229-235] and promoters[236] within genomic sequence. Also available are various Web sites that provide suites of programs for facilitating routine sequence analysis[237] (e.g., see http://gc.bcm.tmc.edu:8088/search-launcher/launcher.html). In fact, numerous sites on the Web are available for accessing and analyzing genomic data; especially good sites include the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov), which operates the premiere genomic database (GenBank), and the National Human Genome Research Institute (http://www.nhgri.nih.gov/Data), which provides a listing of Web sites particularly relevant to genome analysis and the HGP.

Already, biomedical researchers spend sizable amounts of their time in front of computer screens retrieving, analyzing, and manipulating genome mapping and sequencing data. This trend will undoubtedly intensify over time, as increasing amounts of genomic information about numerous organisms are generated at an unprecedented rate.

## GENOME MAPPING AND SEQUENCING: EXPERIMENTAL STRATEGIES

Genomic maps are linear representations of DNA that reflect the organization of landmarks based on some coordinate system. The construction of such maps is critical for attaining a global understanding of genome structure and function. There are three major classes of genomic maps: cytogenetic maps, genetic maps, and physical maps (Fig. 10-5). In each case, the coordinates on which the maps are based reflect the experimental method(s) used to establish the order and intervening distances between landmarks. Importantly, the various mapping methods are associated with characteristic resolution ranges (Fig. 10-6) that dictate the utility of the resulting maps. The development of highly integrated cytogenetic, genetic, and physical maps represents a central activity of the HGP.

### Cytogenetic Maps

A cytogenetic map represents the appearance of a chromosome when properly stained and examined microscopically. Particularly important is the resulting appearance of differentially staining regions (called "bands") that render each chromosome uniquely identifiable (Figs. 10-5 and 10-6). In the case of human chromosomes, individual bands can be specifically discerned and are associated with well-defined names.[238] The most conventional cytogenetic maps depict the 23 chromosomes of the haploid human genome as containing a total of 350 to 500 bands at metaphase (and the amount of DNA split roughly equally between light and dark bands), with each band containing an average of ~5 to 10 Mb of DNA. More sophisticated, higher-resolution methods can be used to detect and represent over 1000 bands in the cytogenetic map of the human genome.[144]

Cytogenetic maps have played a classic role in the diagnosis

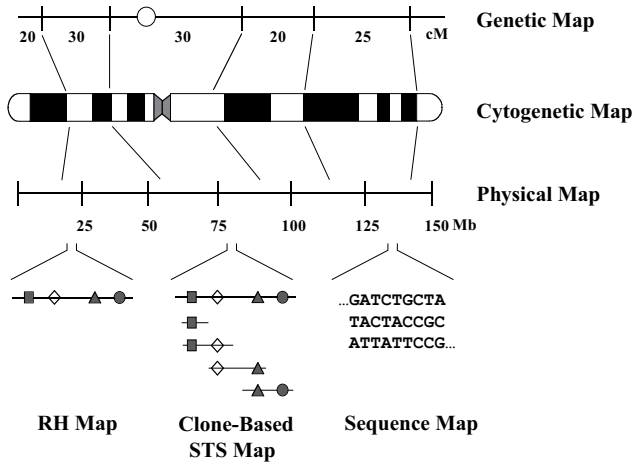


**Fig. 10-5 Schematic representations of the genetic, cytogenetic, and physical maps of a human chromosome. For the genetic map, the positions of several hypothetical genetic markers are indicated, along with the distances in centimorgans (cM) between them. The circle indicates the position of the centromere. For the cytogenetic map, the classic Giemsa-banding pattern of a chromosome is shown. For the physical map, the approximate physical locations of the above genetic markers are indicated, along with the relative distances between them in megabase pairs (Mb). The three types of physical maps [radiation hybrid (RH), clone-based STS, and sequence] depicted along the bottom are discussed in the text. (Adapted from Green and Waterston.[474])**

and study of human genetic diseases. A karyotype, for example, is a visual representation of an individual's cytogenetic map, which may reveal chromosomal deletions, rearrangements, translocations, or other abnormalities. In some cases, the close association between such a cytogenetic abnormality and a particular genetic disorder has served as the starting point for the isolation of the defective gene (e.g., chronic granulomatous disease,[239] Duchenne muscular dystrophy,[240] fragile X syndrome[241]).

On the surface, cytogenetic maps would appear to have a limited role in genome mapping, in that they are relatively low-resolution (e.g., ~5 to 10 Mb) representations of chromosomes (Figs. 10-5 and 10-6) and they are observational in nature, providing neither cloned DNA for additional studies nor significant assistance in obtaining it. However, cytogenetic mapping serves an important adjunct role in the construction of detailed genomic maps. By dividing the human genome into distinguishable units of ~5 to 10 Mb each, cytogenetic maps provide a framework for the construction of other types of maps. For example, cloned DNA segments can be efficiently assigned to specific chromosomal bands by FISH analysis (see *Fluorescence* In Situ *Hybridization* above), thereby providing the ability to coalign (or integrate) other types of maps (e.g., physical and genetic) with the established cytogenetic maps.[154,155] Similarly, FISH analysis can be used to monitor the quality of evolving physical maps. For example, while mapping a particular chromosomal region, newly obtained clones can be analyzed by FISH, with any evidence of hybridization to some other genomic region alerting the investigator to a potential problem. In addition, the orientation of evolving maps with respect to the centromere and telomeres can often be established by FISH analysis using representative clones from each end of the map.[242] Thus, while the basic cytogenetic map of the human genome is essentially established, efforts continue to localize DNA clones and landmarks to precise chromosomal positions by FISH analysis, so as to insure the resulting genomic maps are highly accurate and integrated with one another.

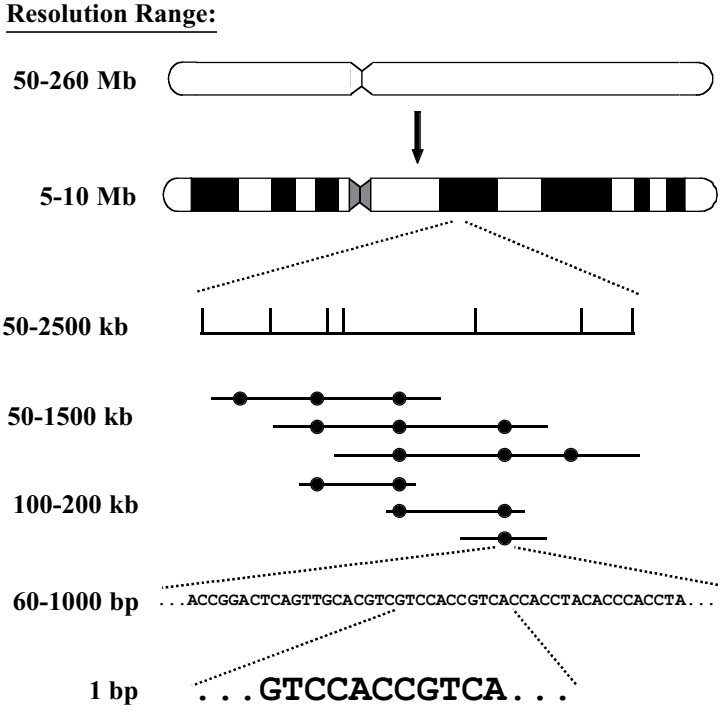


**Fig. 10-6 Characteristic resolution ranges in human genome mapping. Individual human chromosomes range in size from ~50 to 260 Mb. When properly stained and examined microscopically, the characteristic cytogenetic banding pattern gives a unique appearance to each chromosome, with each band containing ~5 to 10 Mb of DNA. Physical mapping techniques, such as long-range restriction mapping by pulsed-field gel electrophoresis, YAC cloning, and BAC cloning are associated with successively decreasing resolution ranges, as indicated. Individual DNA landmarks typically represent much smaller DNA segments (e.g., ~60 to 1000 bp for an STS). The highest level of resolution is the single base pair of DNA sequence. (Adapted from Rossiter and Caskey.[475])**

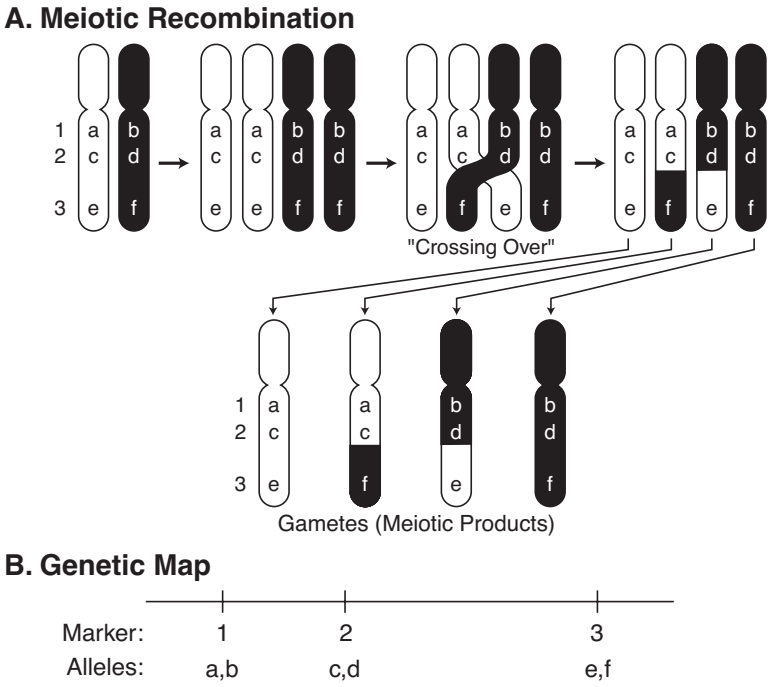


**Fig. 10-7 Fundamental basis of genetic mapping. (Panel A) During meiosis, each chromosome lines up with its homologous partner and is replicated. Paired chromosomes can break and rejoin with each other at one or more points in common (called "crossing over" or "meiotic recombination"), leading to the exchange of DNA. Such a recombination event can thus result in the reassortment of alleles that were previously on the same chromosome. Three hypothetical markers (1, 2, and 3) on the chromosome are indicated, each with two alleles (a/b, c/d, and e/f, respectively). A crossover event is depicted as occurring between markers 2 and 3, yielding two recombinant chromosomes ("a,c,f" and "b,d,e") among the four meiotic products. (Panel B) The depicted genetic map is based on the measured meiotic recombination events, such as that shown in panel A. The distance between markers reflects the frequency with which they are inherited together (i.e., the closer two markers are to one another, the less likely a recombination event will occur between them, and vice versa). (Adapted from Rossiter and Caskey.[475])**

## GENETIC MAPS

Genetic maps (also known as "linkage maps" or "meiotic maps") depict the relative locations of genetic (as opposed to physical) markers across a stretch of DNA. These maps have a more abstract meaning than do physical or cytogenetic maps, since the order and spacing of markers is related to the complex events involved in the transmission of DNA from one generation to the next.

**Theory of genetic mapping.** Most human cells contain two sets of homologous chromosomes, one inherited paternally and one inherited maternally. Thus, for a particular DNA segment (or marker), there can exist two alleles—one on each of the two homologous chromosomes. During the formation of germ cells, the diploid set of chromosomes is divided up during a process known as meiosis, which results in the generation of gametes with only one of each of the pairs of homologous chromosomes (23 chromosomes total). Markers on nonhomologous chromosomes assort randomly during meiosis. Markers on homologous chromosomes tend to be inherited together (i.e., they are "linked"). Often, a recombination event occurs between two homologous chromosomes (i.e., there is an exchange of chromosomal material between the homologues inherited from the mother and the father), resulting in two new "hybrid" chromosomes, each containing portions of the starting homologous chromosomes (Fig. 10-7). Following such a recombination event, some previously linked markers may no longer cosegregate.

Genetic mapping is simply the process of measuring the probability that two closely spaced markers on a chromosome will remain together during meiosis. This is accomplished by analyzing multiple members of known families and measuring the frequency of recombination between markers. The greater the frequency of recombination observed between two markers, the larger the genetic distance separating them, and vice versa (Fig. 10-7). The resulting genetic map depicts the genetic distance between different markers and, therefore, their relative order (Figs. 10-5 and 10-7). The unit of measure in human genetic maps is the centimorgan (cM), with 1 cM corresponding to a probability of 1% that a recombination event will occur in a single meiosis (i.e., 1 recombination event, on average, every 100 meioses). The human genome consists of ~3,300 cM in genetic distance. The correlation between genetic distance and physical distance varies throughout the genome, since some regions are more susceptible to meiotic recombination than others. As a very rough guide, 1 cM in genetic disease, on average, corresponds to ~1 Mb in physical distance.

**Genetic markers.** Genetic markers serve the function of discriminating between homologous chromosomes, thereby allowing recombination events to be detected. To be useful, a genetic marker must display variance among different copies of the same chromosome (i.e., it must be polymorphic), thereby allowing it to be followed during passage from one generation to the next. The informativeness of a genetic marker reflects the actual likelihood that it will be different on two homologous chromosomes (i.e., the likelihood that each chromosome will contain a distinct allele). A number of different types of genetic markers have been used for constructing genetic maps. One type of genetic marker is an inherited disease itself (e.g., Huntington disease, neurofibromatosis, cystic fibrosis, sickle cell anemia). Other biological features can also serve as genetic markers (e.g., blood cell surface antigens, serum proteins, tissue markers). However, most markers used for genetic mapping are DNA sequence differences that are neutral with respect to the phenotypic status of the organism.

There are several major classes of DNA-sequence-based genetic markers. Restriction fragment length polymorphisms (RFLPs) reflect sequence variations that result in DNA fragments of different sizes following restriction digestion, Southern

blotting,[243] and hybridization with an appropriate probe[244-247] (Fig. 10-8). Since most RFLPs reflect two or very few distinct variants, these markers are typically not very informative, thereby limiting their usefulness. Another class of genetic markers consists of tandemly repeated DNA segments. Included among these markers are those containing a variable number of tandem repeats[248] (VNTRs) or "minisatellites," with each repeat unit containing ~11 to 60 bp (Fig. 10-8). The latter polymorphisms are most often detected by agarose gel electrophoresis, Southern blotting,[243] and hybridization with an appropriate probe. Since a greater number of alleles are typically encountered (reflecting different numbers of repeated units), these markers are generally more informative than RFLPs.

In the case of human genetic maps, RFLPs and VNTRs have been largely supplanted by a newer type of genetic marker, termed the "short tandem repeat"[249-251] (STR) or "microsatellite." These polymorphisms are based on differences in the lengths of DNA tracts composed of tandemly repeated di-, tri-, or tetranucleotides (typically repeated a total of ~5 to 30 times) (Fig. 10-8). STRs are particularly useful for genetic mapping because they tend to be more informative than other types of genetic markers. For detecting an STR, PCR primers that flank the tandem repeat are used to amplify the entire segment, and the size of the variable fragment is measured by gel electrophoresis[252] (Figs. 10-8 and 10-9). The most commonly encountered STR in the human genome consists of the dinucleotide CA (or GT on the opposite strand). STRs are now widely used for constructing genetic maps due to their frequent occurrence in the human genome (e.g., estimated every 30 to 60 kb for CA repeats), informativeness, and efficient detection by PCR-based analyses.

The final class of genetic marker is the single-nucleotide polymorphism (SNP). With these markers, a single base is variant at a particular site, typically being one of two possible nucleotides (i.e., biallelic; Fig. 10-8). SNPs can be readily detected by PCR amplification of the surrounding DNA followed by analysis of the resulting PCR product by one of many possible methods, ranging from direct DNA sequencing to various non-gel-based approaches.[190,252-254] In the human genome, SNPs occur relatively frequently, roughly every 500 to 1000 bp of genomic DNA (slightly less often in coding sequences), and appear to be distributed in a relatively uniform fashion. However, because each SNP is typically associated with only two alleles, these genetic markers are not as informative as STRs. Thus, to generate the same amount of data required for performing genetic mapping studies, a larger number of SNPs are required compared to STRs.[253,255] However, the ability to use automated, non-gel-based detection methods, [252-254] such as DNA chips[190] (see *DNA Chips* above), will almost certainly make SNPs the genetic markers of choice for performing large-scale human genetic mapping studies in the future.[255-258] Indeed, high-throughput, SNP-based genetic mapping methods are envisioned as being critical for unraveling the underlying bases of genetically complex diseases.[13-17]

**Construction of genetic maps.** The process of constructing genetic maps of human DNA is a relatively complicated and tedious process[259] because many of the desirable features found with other experimental organisms (e.g., large numbers of offspring, controlled matings, relatively rapid generation times) are not available. Nonetheless, by utilizing multigeneration families containing large sibships along with living parents and grandparents, sufficient data can be generated to build high-quality genetic maps. A French research group [CEPH (Centre d'Etude du Polymorphisme Humain)] was established to facilitate the distribution of cells and DNA from carefully collected and documented families.[260] Thus, different investigators around the world can use the same families for constructing genetic maps, thereby allowing different data sets to be more readily compared and integrated.

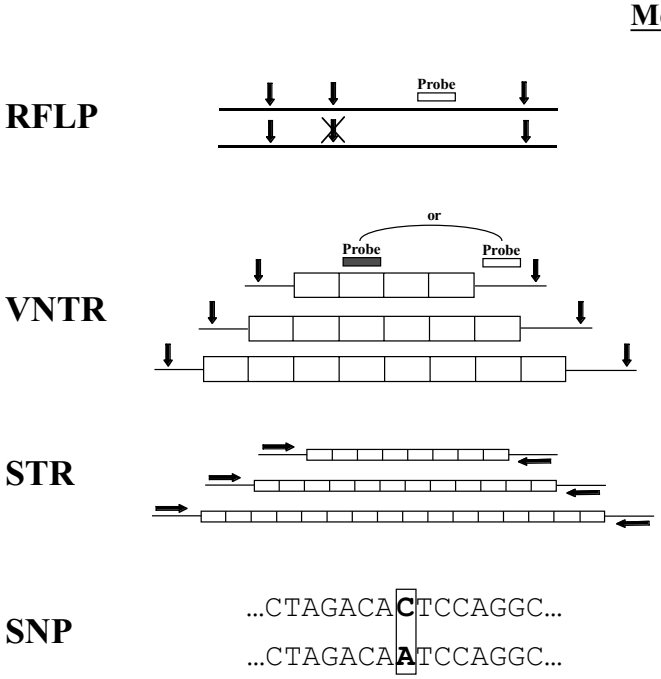


**Fig. 10-8 Genetic markers and their detection. Four major types of genetic markers used for genetic mapping are depicted, along with the experimental method used for their detection. RFLPs are typically associated with the variable presence of a restriction site(s) (indicated by vertical arrows) in a stretch of DNA and are detected by restriction digestion, Southern blotting, and hybridization with a suitable probe. VNTRs reflect variable sizes of DNA tracts harboring a repeated DNA segment, again typically detected by digestion with a restriction enzyme, Southern blotting, and hybridization with a probe. STRs consist of smaller simple-sequence (e.g., di-, tri-, or tetranucleotide) repeats, which can be detected by PCR with primers (depicted as horizontal arrows) that flank the repeated region (see Fig. 10-9). SNPs reflect differences of a single nucleotide at a defined position (shown here as a C vs. A) and are typically detected by PCR amplification of the DNA segment harboring the polymorphism followed by an analytical step that allows discrimination between the different sequences. (Adapted from Rossiter and Caskey.[475])**

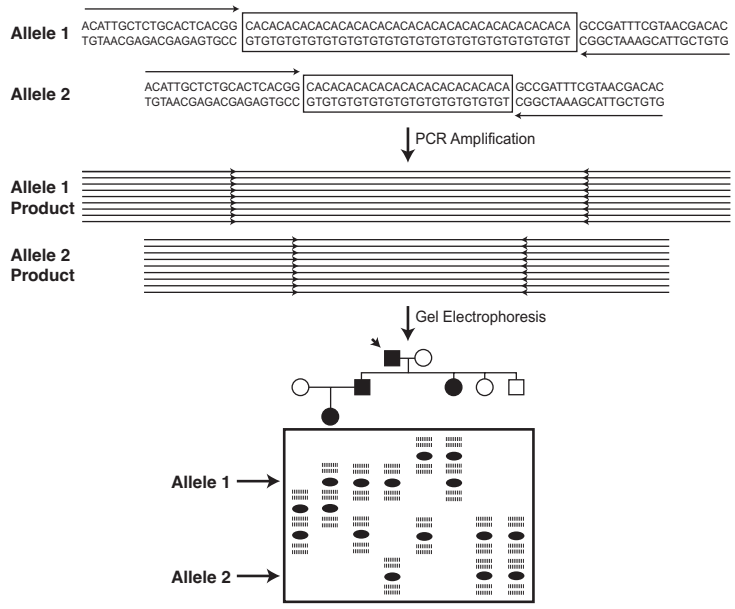


**Fig. 10-9 A short tandem repeat (STR) genetic marker reflecting a variable number of CA repeats. Regions of the human genome that contain stretches of reiterated CA dinucleotides are often polymorphic with respect to the number of repeat units.[249-251] One hypothetical segment is depicted, along with PCR primers (indicated by horizontal arrows) that can be used to amplify the CA-repeat unit. The two alleles (1 and 2) present in one of the family members (indicated with an arrow) are shown. PCR amplification of the STR marker in that individual yields two products of different sizes, which can be resolved on a high-resolution polyacrylamide gel. The results of analyzing the entire family for this genetic marker reveal the presence of multiple alleles, which is characteristic of STRs. A common finding with PCR amplification of CA repeats is the presence of shadow (or stutter) bands (depicted as light lines) smaller and/or larger than the expected product. (Adapted from Germino and Somlo.[600])**

Provided the availability of genetic markers and DNA from large families, the process of constructing genetic maps is now well established. Each DNA sample is analyzed with an appropriate set of genetic markers using a suitable detection method(s) (see Fig. 10-8), and the resulting data are carefully collected and recorded. Sophisticated computational tools have been developed for assessing the inheritance patterns of the markers and for deducing the resulting genetic maps.[261] A first-generation genetic map of the human genome was reported in 1987 and consisted of markers (predominantly RFLP-based) that were spaced, on average, every 10 cM.[262] As part of the HGP, second-generation, higher-resolution genetic maps consisting of STR-type markers have now been constructed for the human genome[263-268] (see *Highlights of the Human Genome Project* below).

**Uses of genetic maps.** Genetic maps facilitate the search for genes associated with human disease. In conventional positional cloning strategies, the disease gene itself can serve as one genetic marker whose linkage to other genetic markers is assessed by the analysis of affected families. In the ideal cases, more closely linked genetic markers are identified that allow the genomic region containing the disease gene to be limited to an interval that can be readily studied by physical mapping methods (see *Background on Positional Cloning* below). Such an approach is particularly well-suited for single-gene disorders. In the future, human genetics studies will increasingly focus on the analysis of more genetically complex diseases. This will be accompanied by a shift away from the use of individual families and towards the use of large-scale association studies, whereby common genetic variants will be correlated with specific traits by analyzing large numbers of individuals (as opposed to families).[14,256] Such efforts will depend on the availability of large sets of SNP markers, an associated catalog of known human sequence variants (see http://www.ncbi.nlm.nih.gov/SNP), and more robust technologies for large-scale SNP detection[255-258] (see *DNA Chips* above).

Genetic maps are also important for at least two other applications. First, they are valuable as a framework for assembling physical maps. While genetic and physical maps provide different information about the corresponding DNA, they are colinear with respect to the order of markers. Thus, genetic mapping can complement physical mapping by providing information about the order of physical landmarks based on their genetic positions. For example, this can be accomplished by simply localizing mapped genetic markers on a physical map[154] (see *Generation of Sequence-Tagged Sites* below). Second, genetic maps are valuable for studying subtle aspects of inheritance. For example, there is no uniform relationship between genetic distance (i.e., recombination frequency) and physical distance; in the case of human DNA, the rough correlation of 1 cM to 1 Mb represents an estimated average, with considerable variation occurring throughout the genome. Interestingly, the relative recombination frequencies per physical length of DNA are higher near the ends of chromosomes (telomeres) and lower near the centromeric regions. The availability of higher-resolution genetic maps in conjunction with more complete physical maps is allowing the molecular bases for these observations to be studied in a more rigorous fashion.[106,107]

## Physical Maps

Physical maps depict the relative locations of physical landmarks across a stretch of DNA, much like a travel map indicates the locations of cities along a highway. The construction of physical maps is performed either with the total genomic DNA of an organism or with smaller pieces derived from that genome. The conventional approach for the latter involves fragmenting and then cloning the DNA, so as to purify and analyze individual DNA fragments. However, by fragmenting the DNA, the order of the DNA segments is lost, leaving the challenge of correctly putting the pieces back together to create an accurate map. Hence, the analogy is often made to a jigsaw puzzle; however, there are some important differences. Instead of just one copy of every DNA fragment, many copies are present within the collection of clones. In addition, the same DNA segment may be present on a number of different-sized, overlapping clones. In assembling such a clone-based physical map, individual clones are analyzed for the presence of landmarks and then compared with other clones. When

two clones are found to have one or more landmarks in common, they can be conceptually overlaid (or overlapped), and because such clones are typically notidentical (only overlapping), a slightly larger segment is reconstructed. A collection of ordered, overlapping clones is called a "contig," since the clones together contain a contiguous segment of DNA. Thus, a clone-based physical map (i.e., a contig map) reflects both a collection of overlapping clones and an ordered set of DNA landmarks.

**Restriction sites as landmarks.** Various types of landmarks can be used to construct physical maps of DNA. A simple one is the restriction site, which reflects the site of cleavage by a specific restriction enzyme(s). Since a large number of restriction enzymes are available, each recognizing and cutting DNA at a defined sequence, detailed restriction maps can be theoretically constructed by using several different enzymes. Long-range, low-resolution restriction maps of uncloned genomic DNA can be constructed by using rare-cutting restriction enzymes and pulsed-field gel electrophoresis;[135-137,139] however, the resulting maps are usually not very detailed and provide no direct access to the DNA itself. More typical is the detection of restriction sites within cloned DNA. Here, individual clones are analyzed, and the physical distances between restriction sites established. This information provides a restriction-site-based "fingerprint" for each clone, which in turn can be compared with other clones to deduce overlap relationships. Such an approach was particularly effective for constructing physical maps of several model organism genomes[269-271] and is now being adapted for use in constructing high-resolution physical maps en route to systematic sequencing of the human and other genomes[124-126,272] (see *Genomic Sequencing* below).

**Sequence-tagged sites as landmarks.** Early in the HGP and shortly after the advent of PCR, a new type of DNA landmark was envisioned: the "sequence-tagged site" (STS).[273] STSs are short stretches (e.g., ~60 to 1000 bp) of unique DNA sequence that can be specifically detected by a PCR assay (Fig. 10-10); in essence, STSs are the physical landmarks and PCR is the experimental method used to detect them. STS maps simply represent the relative positions of STSs across a stretch of DNA[86,274] (Figs. 10-5 and 10-10). STSs provide several key advantages over other landmarks (e.g., restriction sites, hybridization probes) that make them well-suited for use in physical mapping. First, the use of PCR as the front-line analytical tool for detecting DNA landmarks is highly desirable because of its high sensitivity, specificity, and potential for automation. Second, all of the relevant information about an STS (e.g., the sequence of the two oligonucleotide primers, reaction components, temperature-cycling parameters) can be stored and accessed electronically (such as in the dbSTS database; see http://www.ncbi.nlm.nih.gov/dbSTS), thereby making that STS experimentally accessible to any laboratory. As a result, the DNA segment corresponding to a particular STS can be generated simply by synthesizing the appropriate oligonucleotide primers and performing PCR under the described conditions.[275] This facilitates the assimilation and comparison of STS maps constructed in different laboratories and/or by different methods. Similarly, the availability of a PCR assay specific for each STS provides the means to isolate readily a clone containing the corresponding genomic region; this offers a desirable element of flexibility with respect to the ability to utilize different cloning systems or new, improved clone libraries. Finally, since STSs are, by definition, sequence-based, physical maps assembled with STSs as the landmarks can be readily integrated with evolving sequence maps by simple computer analysis. Because of these features, STSs rapidly ascended to become the dominant landmark for constructing physical maps of mammalian chromosomes. STSs are most commonly mapped using large-insert clones, such as YACs (see *Yeast Artificial Chromosome-Based Sequence-Tagged Site-Content Mapping* below), or radiation hybrid cell lines (see *Radiation Hybrid Mapping* below).

**Generation of sequence-tagged sites.** To construct detailed physical maps of large genomic regions (such as human chromosomes), many thousands of STSs must be developed. The generation of an STS involves determining a small amount of DNA sequence (~100 to 400 bp), developing a PCR assay that will specifically amplify the corresponding DNA segment (including the design[276] and synthesis of the two oligonucleotide primers), and demonstrating that the site is functionally unique within the genome.[275,277,278] A key step in this process is the generation of DNA sequence, which can either be obtained from a totally random piece of genomic DNA or from a more targeted DNA

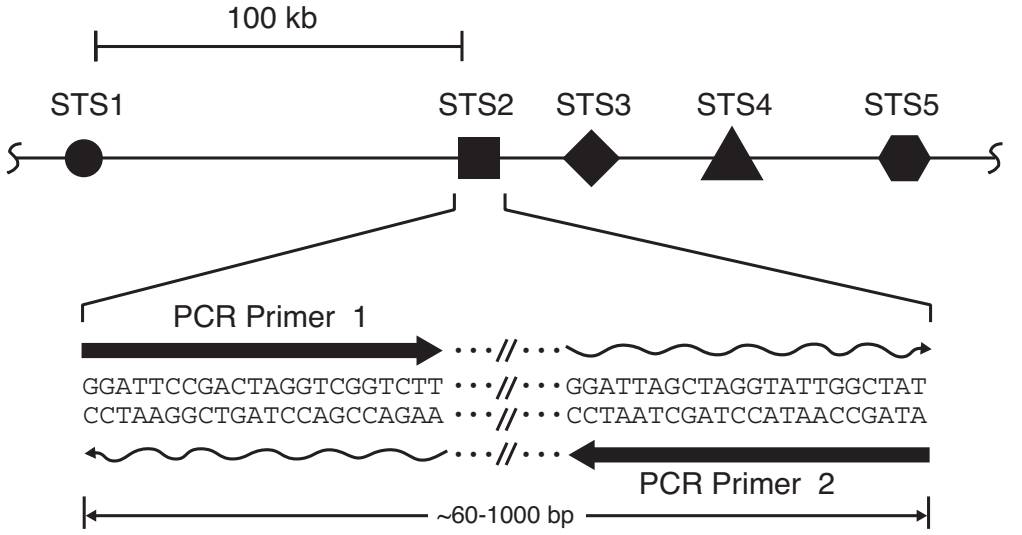


**Fig. 10-10 STSs and STS maps. An STS is a unique DNA sequence in the size range of ~60 to 1000 bp that can specifically be detected by a PCR assay employing two oligonucleotide primers (indicated by horizontal arrows). A physical map of a human chromosome can be represented by the relative positions of STSs (depicted as unique symbols), each of which is associated with a specific PCR assay. (Adapted from Green and Olson[86] and Green and Green.[274])**

segment, with the latter being technically more demanding.

Several methods have been established for generating DNA sequence from targeted genomic regions, such as individual human chromosomes. Many of these approaches involve the use of human-rodent hybrid cell lines containing a single human chromosome or portion thereof (see *Somatic Hybrid Cell Lines* above). Several strategies can be used to isolate the human DNA away from the rodent DNA background. For example, lambda or cosmid clones derived from such hybrid cell lines can be screened for the presence of human-specific repetitive sequences. Since most segments of human DNA present in lambda (15 to 25 kb) or cosmid (35 to 45 kb) clones should harbor at least one repetitive sequence (e.g., *Alu*), clones containing human DNA can be identified by hybridization with radioactively labeled human-DNA probes.[275,277] Alternatively, segments of human DNA within the hybrid cell lines can be amplified by using a variant type of PCR, called "*Alu*-PCR"[279-281] (Fig. 10-11). In this method, total DNA from the hybrid cell line is used as template in PCR assays employing oligonucleotide primers specific to consensus *Alu* sequences.[40] The PCR products generated are enriched for DNA segments residing between adjacent *Alu* repeats.[279-281] *Alu*-PCR can thus be used to amplify the human but not the rodent DNA from a hybrid cell line, with the resulting DNA segments then utilized for developing STSs.[277,282] A second approach for isolating targeted genomic regions uses the technique of flow-sorting,[283-285] which separates individual chromosomes based on quantitation of the laser-induced fluorescence following staining. Typically, flow-sorting yields very small quantities of DNA, which must then be subcloned prior to analysis.[283-285] While flow-sorted DNA has been successfully used for developing STSs from individual human chromosomes,[275,277,286,287] the presence of irrelevant DNA segments has in some cases hindered the utility of this approach.[275,277] Other strategies for targeted STS generation include the use of 3′ untranslated regions of cDNAs[288] and sequences derived from the ends of large-insert clones, such as YACs.[289,290]

Efficient strategies for developing STSs have been established[98,274,275,277,278,286,287] and utilized for many large-scale physical mapping projects. In the case of the human genome, large collections of STSs have been generated in both genome-wide[98,291-293] and targeted fashions (see http://www.ncbi.nlm.nih.gov/dbSTS). It is also important to note that the large collections of developed PCR-based human genetic markers[263-268,294] (specifically STRs; see *Genetic Markers* above) are also STSs; that is, each is a unique sequence associated with a PCR assay. These markers can thus serve as landmarks on both the genetic map and the physical map, with their presence on both serving to integrate the two maps together.

**Yeast artificial chromosome-based sequence-tagged site-content mapping.** The major approach employed for constructing the first-generation clone-based physical maps of the human genome has involved the use of STSs as the landmarks and YACs as the source of the cloned DNA.[86,274] Because of their large insert size, YACs simplify the process of assembling contigs covering large genomic regions (much as jigsaw puzzles made of larger pieces are easier to assemble than those made of smaller pieces). In this strategy, YAC clones are isolated from appropriate libraries (typically by using an STS-specific PCR assay[85]) and analyzed for the presence of additional STSs (Fig. 10-12). An "STS content" is thus established for each YAC. Two clones are assumed to overlap if they have one or more STSs in common. By establishing the overlaps among a group of YACs, contig maps can be deduced that reflect both the relationships among the clones as well as the relative order of the STSs (Fig. 10-12). This general strategy, called "STS-content mapping,"[86,274] has been used to construct large YAC contigs corresponding to a number of regions of the human genome, including segments encompassing important disease genes[86-89,91,92] (e.g., see Fig. 10-13) and whole human

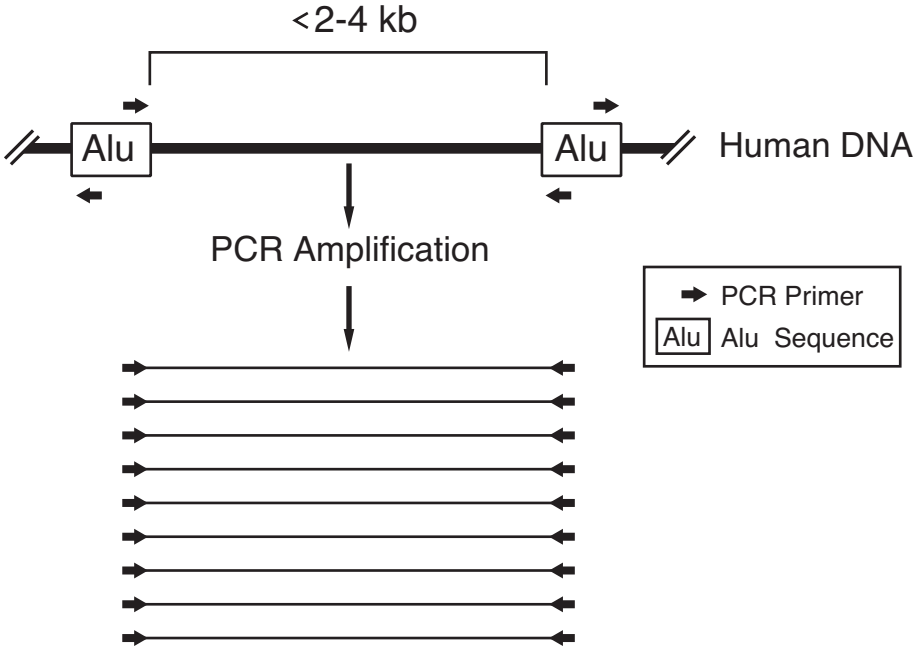


**Fig. 10-11 Amplification of DNA by Alu-PCR. Interspersed repetitive sequences in human DNA, such as Alu repeats, are often closely spaced to one another (due to their high frequency in the genome). These repeated sequences represent potential annealing sites for PCR primers (depicted as horizontal arrows), in particular those that are complementary to known consensus Alu sequences and point outward from the repeat units. When two Alu repeats are in close enough proximity to one another (e.g., less than 2 to 4 kb apart), the region between them can be amplified by Alu-PCR.[279-281] The resulting PCR product consists of the outermost portions of the Alu repeat together with the intervening DNA segment. Typically, Alu-PCR is performed on samples of human DNA that contain numerous Alu repeats, resulting in the amplification of a heterogeneous collection of different PCR products.**

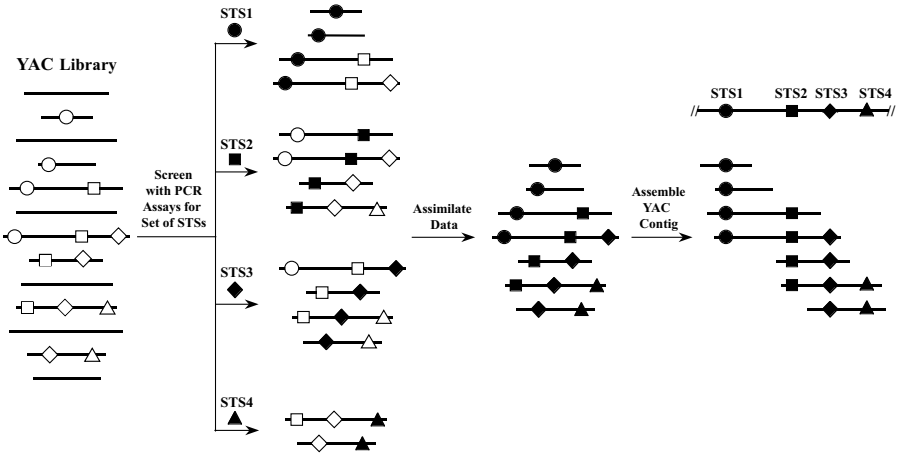


**Fig. 10-12 General strategy for constructing YAC-based STS-content maps. A YAC library, consisting of clones of various sizes and unknown compositions of STSs (depicted as open symbols), is screened for a set of STSs using a PCR-based screening method.[85] For each group of isolated clones, only the STS whose corresponding PCR assay was used to perform the library screen is known to be present in the positive YACs (depicted as filled-in symbols). However, data about which YACs contain which STSs can then be assimilated and the size of each YAC measured by pulsed-field gel electrophoresis. The resulting information can be used to assemble a YAC contig, reflecting both the overlap relationships among the clones and the relative order of the STSs. (Adapted from Green and Olson[86] and Green and Green.[274])**

chromosomes.[93-95,98,100,101,103-107] A global view of the YAC-based STS-content map constructed for human chromosome 7 is provided in Fig. 10-14.

**Radiation hybrid mapping.** Other, non-clone-based approaches are available for establishing the relative order of landmarks (such as STSs) across a stretch of DNA. One technique, called "radiation hybrid mapping,"[159,160,295,296] exploits the ability to recover chromosome fragments in rodent cells. In one application of this method, human chromosomes in cultured cells are fragmented by irradiation, and individual pieces are recovered by fusion of the irradiated cells with a rodent cell line. Each of the resulting cell lines typically contains numerous, disjointed fragments of the starting chromosome(s). A set of such radiation hybrid cell lines (typically ~90), each containing a different assortment of human chromosomal fragments, are then isolated and analyzed for the presence or absence of DNA landmarks. An efficient approach for the latter involves the use of PCR for detecting STSs within the cell lines.[160] The relative spacing between two STSs can be deduced based on analyzing the data about their coexistence within the collection of cell lines. The fundamental principle is that the frequency of radiation-induced chromosomal breaks between STSs is proportional to the distance between them; thus, closely spaced STSs tend to coexist in a larger fraction of cell lines compared to those that are far apart on a chromosome or that are on separate chromosomes. Statistical analyses are used to order the STSs and to estimate the relative distances between them,[159,295,297-299] resulting in the assembly of a radiation hybrid map.

Radiation hybrid mapping represents an important adjunct approach for physical mapping. For example, this method facilitated the assembly of a YAC-based physical map of the human genome.[98] In these studies, the resulting STS-based radiation hybrid map was used to confirm the STS order deduced by YAC-based STS-content mapping, to order individual YAC contigs, and to orient YAC contigs relative to the centromere and telomeres. Thus, the combined use of independent methods for ordering STSs helped in the construction of an overall more consistent map. Radiation hybrid mapping has also been used extensively to map genes in the human genome[292,293] (see http://www.ncbi.nlm.nih.gov/genemap98) (see *cDNA Sequencing* below). Furthermore, STS-based radiation hybrid maps of increasing resolution are actively being constructed for the human[291,300] (also see http://shgc-www.stanford.edu) and mouse[301] genomes. Finally, radiation hybrid mapping provides an approach for the comparative mapping of genomes from different animal species.[302]

**Sequence maps.** The highest-resolution physical map is the DNA sequence map (Figs. 10-5 and 10-6), which depicts the precise order of nucleotides across a stretch of DNA. There are two major types of data generated for constructing sequence maps: that derived from cDNA (copies of mRNA molecules; i.e., expressed sequences) and that from genomic DNA. Each type of sequence has its relative value and limitations, with both types frequently generated for analyzing genomes.

**cDNA sequencing.** Sequence data generated from cDNA provide an enriched source of information about the small part of the genome that is expressed and encodes protein. Refinements in the methods for large-scale DNA sequencing[166] coupled with improved protocols for generating "normalized" cDNA libraries (where individual transcripts are more equally represented[303-307]) have resulted in the efficient generation of large numbers of sequences from randomly selected cDNA clones. Such sequences are called "expressed-sequence tags" (ESTs), since they correspond to partial tags of expressed sequences.[308] Literally hundreds of thousands of human ESTs have been generated[309,310] and are available in the public databases[311-314] (e.g., dbEST, see http://www.ncbi.nlm.nih.gov/dbEST). Together, these ESTs likely represent a major part of the human gene catalogue. Furthermore, just under 4,000,000 ESTs (from numerous organisms) were present in dbEST as of April, 2000.

ESTs can be used as a vehicle for mapping the corresponding genes in the genome. Most often, an STS is generated from the EST sequence[98,288,292,293,300,315,316] and mapped by YAC-based STS-content mapping[98] and/or radiation hybrid mapping.[98,292,293,300] The latter strategy has been applied in a large-scale fashion to construct a gene map (also called a "transcript map") of the human genome[292,293] (see http://www.ncbi.nlm.nih.gov/genemap98).

The generation and mapping of large numbers of human ESTs enhances the ability to perform various types of studies, such as acquiring an initial exposure to the gene repertoire of the human genome,[317,318] facilitating the isolation of human disease genes by positional cloning (see *Isolation of Disease Genes by Positional Cloning* below), and studying the differential expression of genes in various organisms, tissues, and disease

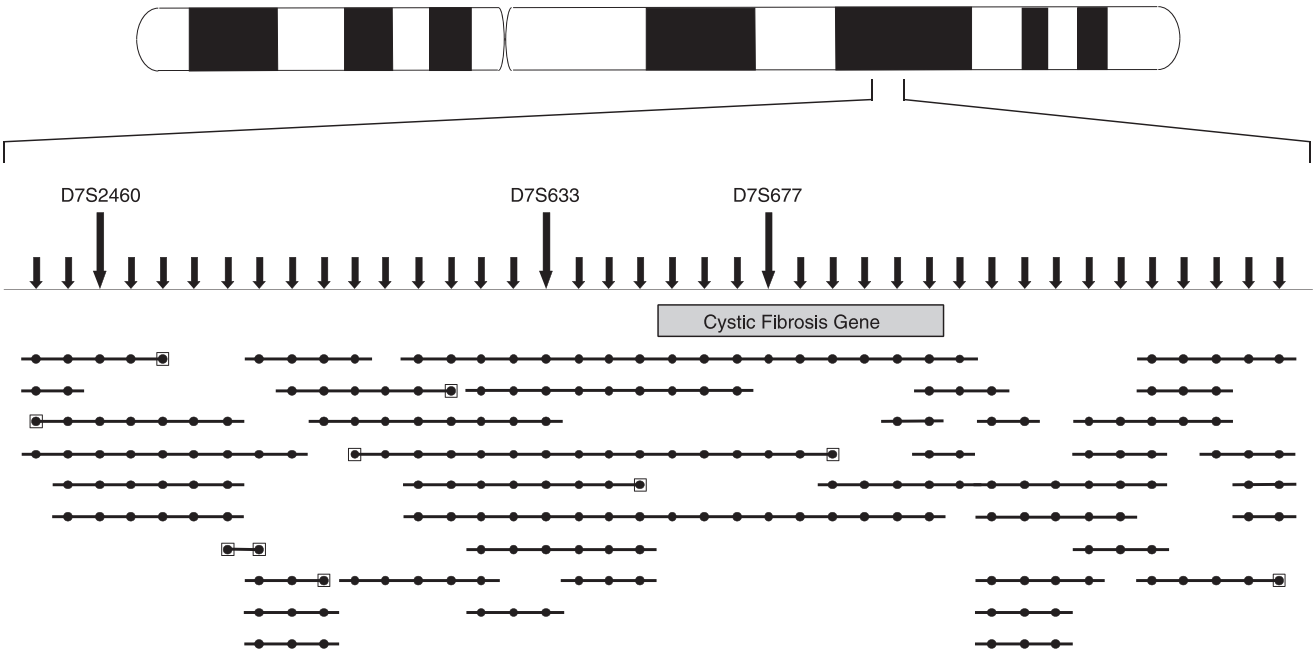


**Fig. 10-13 A YAC-based STS-content map of the human genomic region containing the cystic fibrosis gene. A YAC contig map of the region of human chromosome 7 containing the cystic fibrosis gene is depicted. The vertical arrows along the top indicate the relative positions of the STSs, while the horizontal bars represent YACs. For simplicity, the STSs are spaced in an equidistant fashion and only a representative subset of mapped YACs is shown. The presence of an STS in a YAC is indicated by a closed circle at the appropriate position on the corresponding horizontal bar. When an STS corresponds to the insert end of a YAC, an open square is placed around that circle at the end of the YAC from which it was derived. The indicated overlap relationships among YACs were established by the presence of one or more common STSs. The contig spans across ~2.0 Mb of DNA, with the approximate position of the ~200-kb cystic fibrosis gene indicated. The relative position of this physical map on the chromosome 7 cytogenetic map was established by performing FISH analysis with several YACs from the contig.[154] Similarly, the localization of several genetic markers (D7S2460, D7S633, and D7S677) on the YAC contig allows integration of the genetic and physical maps at these positions.[154] Additional details about this physical map have been reported[86,107] (also see http://genome.nhgri.nih.gov/CHR7).**

states.[319,320] ESTs are not, however, without their limitations. First, since most ESTs reflect single-pass sequence reads, the corresponding sequence is generally of lower accuracy and less contiguity compared to high-quality genomic sequence (see *Genomic Sequencing* below). Second, the data associated with an EST provide little to no information about the structure or regulation of the corresponding gene. Similarly, there is typically no insight about complex situations such as the production of multiple mRNAs from the same gene or the production of the same mRNA from multiple genes. Finally, ESTs are limited to those mRNA molecules that are present during the construction of a cDNA library, with rare transcripts or those not expressed in the tissue at the time of harvesting being difficult to identify. Thus, while large-scale EST generation provides an important source of biologically relevant DNA sequence, it should be regarded as a supplement (rather than a substitute) to complete genomic sequencing.

A variant form of EST analysis, called "serial analysis of gene expression" (SAGE),[321,322] involves the construction of libraries with clones containing concatemerized short sequence tags (e.g., 9 to 10 bp) derived from cDNA. Large numbers of sequence reads are then generated from individual clones, and the frequency of different short tags assessed. The resulting data thus allow a quantitative cataloging and comparison of expressed genes from a defined biological source (e.g., tissue, organism, or developmental stage). This methodology, which in essence is a sequence-based approach for monitoring gene expression levels, offers great promise for numerous applications.

**Genomic sequencing.** The ability to establish the complete sequence of large genomes has advanced tremendously in recent years as a result of the experience gained in sequencing the genomes of model organisms, such as the yeast *S. cerevisiae*[323-329] (e.g., see http://genome-www.stanford.edu/Saccharomyces), the bacterium *E. coli*,[330] and the nematode *C. elegans*[331-335a] (see http://www.sanger.ac.uk and http://genome.wustl.edu/gsc; also see *Studying the Genomes of Model Organisms* below). These accomplishments have been associated with a steady accumulation of incremental improvements in the approaches used for performing large-scale sequencing by fluorescence-based methods, rather than any individual revolutionary advance. As a result, the HGP is now firmly focused on the important task of sequencing the human genome[334-339] (see *Sequencing the Human Genome* below), with plans to then sequence the mouse genome and inevitably many others thereafter.

The general strategy that is currently being used for systematic sequencing of large genomes (e.g., the human genome) can be broadly divided into several major steps.[342-342a] First, suitable clones must be selected for sequencing. For a variety of reasons, YACs are not the clones of choice for use in genomic sequencing, despite their central role in the construction of long-range physical maps. Rather, BACs (and similarly PACs; note that only BACs are further mentioned below) are associated with a number of features that make them well suited as sequencing templates (see *Large-Insert*

*Bacterial Cloning Systems* above). Most often, available STSs (previously ordered utilizing YACs or radiation hybrid cell lines) are used to isolate BACs, which in turn are analyzed to establish overlap relationships among the clones and to assemble BAC contigs (Fig. 10-15). The latter typically involves some type of restriction-enzyme-based fingerprint analysis.[124-126,272] From the resulting BAC contig maps, minimally overlapping sets of clones that together span a genomic region of interest are selected for sequencing.[342a] Such high-resolution, bacterial-clone-based maps, which are usually highly redundant with respect to clone coverage and provide the ability to select clones that share very small overlapping regions, are commonly referred to as "sequence-ready maps" (Fig. 10-15).

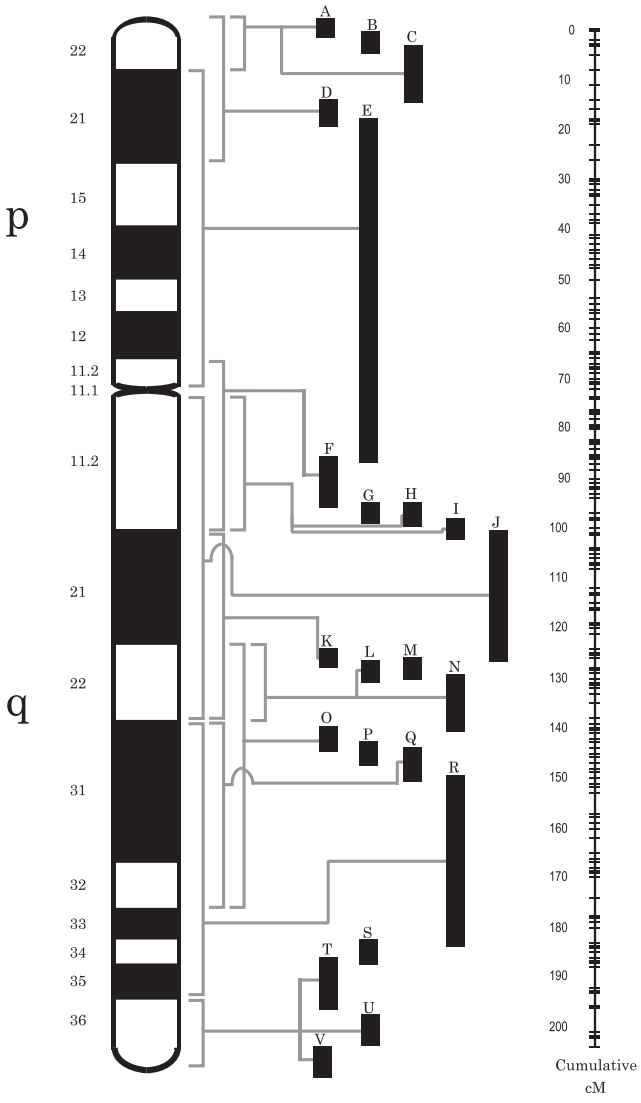


**Fig. 10-14 Physical map of a human chromosome. A global overview of a first-generation physical map of a human chromosome constructed by the HGP is depicted. Specifically, a YAC-based STS-content map of human chromosome 7 is aligned relative to the cytogenetic map (left) and genetic map (right).[107] Each of the vertical bars in the middle (labeled A-V) corresponds to an individual YAC contig (see Fig. 10-13 for a more detailed view of a very small portion of contig O). Some of the contigs are quite large (e.g., contig E spans ~50 Mb or ~30% of the chromosome). Note that the positions of virtually all of the YAC contigs on the cytogenetic and genetic maps have been established. For additional details, see Bouffard et al.[107] and http://genome.nhgri.nih.gov/CHR7.**

In the second step of this process, selected clones are individually subjected to complete, high-accuracy sequencing. The most commonly used strategy for this step is called "shotgun sequencing"[343] (Fig. 10-16) (although a small number of groups are using a variant approach that employs transposons[344]). Shotgun sequencing begins with the construction of a subclone library, consisting of subclones that each contain a small (e.g., ~1 to 3 kb), random piece of the starting template (e.g., BAC). Sequence reads are then obtained from one or both ends of a large number of subclones. Sufficient sequence data are generated such that each base position of the starting template is read, on average, 5 to 10 times (sometimes referred to as "5X to 10X coverage"). Computational tools are then used to analyze the resulting sequences, so as to identify those that overlap to form sequence contigs (each consisting of an assembled consensus sequence). This process has been greatly facilitated by powerful new software that assesses the quality of the data associated with each nucleotide of sequence and provides a user-friendly set of tools for reviewing and editing the resulting sequence assemblies.[211-213] With this software, an increasing amount of the data production and monitoring process can be more automated, thereby increasing the overall efficiency of the shotgun sequencing strategy. The initial shotgun sequencing data typically result in the assembly of a small number of sequence contigs, the order and orientation of which may or may not be known. The next phase involves generating sequence data in a highly directed fashion, so as to fill in the remaining gaps and merge the sequence contigs together. This "finishing" process (as it is often called) uses a number of specialized computational and experimental tools, requires highly trained technical personnel, and often involves generating sequence reads from particularly difficult sequences (e.g., repeated structures).[343]

The last step of genomic sequencing involves a final review of the entire assembled sequence, which includes both checking for any ambiguities or problem areas and analyzing the sequence for features of interest, such as the presence of genes, ESTs, repetitive elements, and other matching sequences. This latter activity is referred to as "sequence annotation." Finally, the complete sequence and its associated annotations are submitted to a public database, such as GenBank.

## STUDYING THE GENOMES OF MODEL ORGANISMS

Mapping and sequencing the human genome will, in principle, reveal all the information needed for the biological development of a human being. However, the ability to interpret most of this information will be heavily dependent on parallel studies of non-human organisms used extensively in research laboratories as model systems. Experimentation in humans is rightly limited by ethical considerations, not to mention many practical factors. In contrast, model organisms such as bacteria, yeasts, worms, flies, and mice can be easily manipulated, especially genetically. Importantly, the knowledge gained from mapping and sequencing the genomes of these other organisms is directly relevant to understanding the human genome and many aspects of human biology, both in normal and abnormal states[319,345] (see *Comparative Study of the Biology of Humans and Other Organisms* below).

Model organisms have another important feature: most have much smaller genomes than that of humans. This feature makes

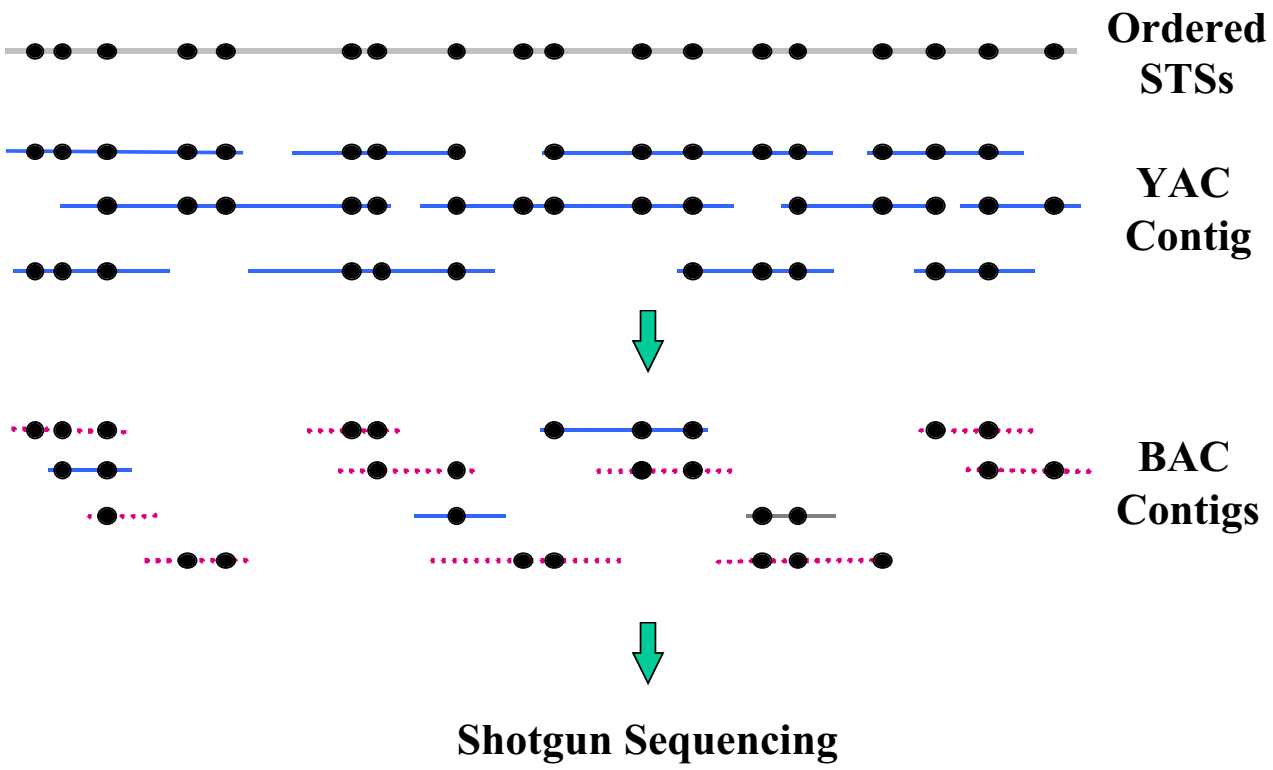


**Fig. 10-15 Construction of clone-based physical maps suitable for sequencing the human genome. The first-generation physical maps of human chromosomes constructed in the HGP are predominantly YAC-based STS-content maps, which provide both collections of overlapping YAC clones and ordered sets of STSs. For various technical reasons, YACs are not ideal templates for DNA sequencing. Instead, the mapped STSs can be used to isolate smaller-insert BAC (or PAC) clones, which in turn can be analyzed[124,125,272] and assembled into contigs. Typically, the resulting nascent BAC contigs are smaller than the corresponding YAC contig. However, the gaps between BAC contigs can be filled by various methods (e.g., by developing new STSs from appropriate BAC ends and then isolating additional overlapping clones). From the resulting BAC contigs, minimally overlapping BACs (shown as dotted lines) can be selected and individually subjected to shotgun sequencing (see text and Fig. 10-16).**

comprehensive genome analysis more straightforward. As illustrated in Fig. 10-1, the genomes of *E. coli*, yeast, nematode, and fruit fly are roughly 1/600, 1/200, 1/30, and 1/20 the size of the human genome, respectively. With smaller genomes, newly developed approaches can be tested more readily. Indeed the experiences gained mapping the yeast[269,346] and nematode[270] genomes have heavily influenced the strategies adopted for analyzing the human and mouse genomes. The smaller genomes of *E. coli*, yeast, nematode, and fruit fly are densely packed with genes, so that the relative amount of information derived from systematic DNA sequencing is high. Importantly, these smaller genomes actually contain many of the same genes that are found in the human genome[319,345] (*see Comparative Study of the Biology of Humans and Other Organisms* below), making their study of particular interest. For the early phases of the HGP, a limited set of model organisms was designated for priority study, and each of these is detailed below.

From the earliest days of molecular biology, studies with the simple prokaryote *E. coli* have revealed many of the fundamental processes of life. A physical map of the single, circular chromosome in the form of overlapping bacteriophage clones was completed in 1987,[271] and systematic efforts have now produced the complete sequence for the ~5 Mb of DNA in the *E. coli* genome.[330] This DNA sequence is proving valuable for identifying many of the essential functions needed to maintain independent life as well as for comparing to other bacterial genome sequences (see *Highlights of the Human Genome Project* below).

Eukaryotic organisms, with their DNA compartmentalized in nuclei, are evolutionarily very distant from *E. coli*. The yeast *S. cerevisiae*, with a genome size of about ~15 Mb, is the simplest model organism with a nucleus. Comparison between it and *E. coli* is revealing the additional basic functions that distinguish eukaryotes from prokaryotes. The 16 chromosomes of *S. cerevisiae* can be separated by pulsed-field gel electrophoresis,[133] a set of overlapping bacteriophage clones representing most of the genome have been assembled,[269] and both long-range[134] and high-resolution[346] restriction maps have been constructed for each chromosome. In turn, the latter physical map proved extremely valuable to many yeast researchers, including those involved in the sequencing of the entire yeast genome. In 1996, the sequence of the yeast genome was completed[323-329] (e.g., see http://genome-www.stanford.edu/Saccharomyces). Yeast geneticists are now devising myriad approaches for exploiting a complete genomic sequence for studying complex genetic problems and for probing subtle aspects of yeast biology.[347-349]

Next in terms of complexity are the nematode *C. elegans* and the fruit fly *D. melanogaster*, with genomes of ~100 Mb (in 6 chromosomes) and ~160 Mb (in 4 chromosomes), respectively. As multicellular animals, these organisms share with mammals specialized cell types such as nerve, muscle, intestine, and gonad. With the combination of molecular and mutational capabilities in the context of detailed biological information, these two

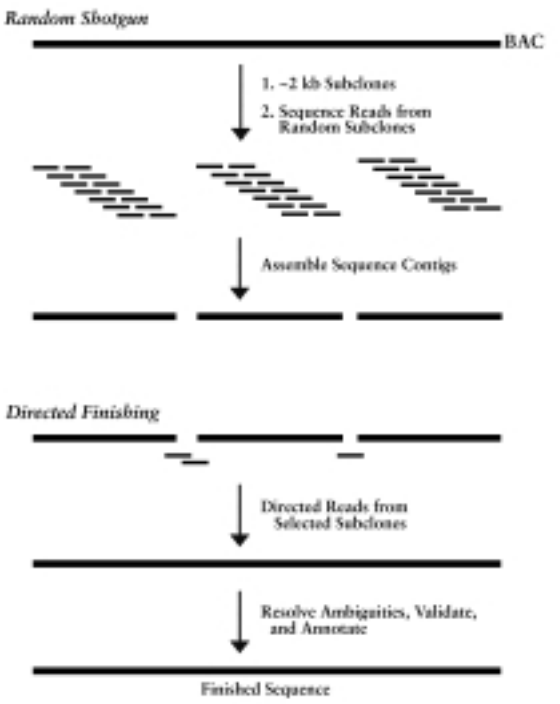


**Fig. 10-16 Shotgun DNA sequencing. The dominant genomic sequencing strategy being used in the HGP is shotgun sequencing, which consists of two major phases. In the first "random shotgun" phase, a genomic clone (e.g., BAC) is subcloned into ~2-kb fragments. Sequence reads are then obtained from one or both ends of a large number of randomly selected subclones. Sufficient sequence data are generated such that each nucleotide of the starting clone is read, on average, about 5 to 10 times. This redundant sequence data is then analyzed with various computational tools, allowing the assembly of sequence contigs. Typically, only a handful of gaps between sequence contigs remain at this stage. In the second "directed finishing" phase, additional data are generated to complete the sequence, most often by obtaining directed reads from strategically selected subclones that provide sequence into or across the remaining gaps. This typically allows the merger of the remaining sequence contigs to yield a final, contiguous (i.e., finished) sequence, which is then analyzed to resolve ambiguities, validated by various tests, and annotated with respect to the location of known genes, ESTs, repetitive sequences, and other sequence features of interest.**

organisms have become powerful systems for studying the role of numerous genes in development and behavior. For *C. elegans*, a detailed physical map and corresponding clone set for virtually the entire genome have been assembled with cosmids[270] and YACs.[350,351] These clones provided the necessary templates for systematic genomic sequencing[331-335a] (see http://www.sanger.ac.uk; http://genome.wustl.edu/gsc), which by the end of 1998 yielded the first complete genomic sequence of a multicellular organism. This sequence is profoundly enhancing studies with *C. elegans* in numerous ways.[352,353] For *D. melanogaster*, a well-established cytogenetic map has been available since the 1930's—in the form of the famous polytene chromosomes of the salivary gland. These structures represent a thousand or more aligned copies of each chromosome in an extended conformation, so that segments as close as 20 kb can be resolved, thereby providing a framework for organizing other mapping information.[354,355] Physical mapping of the Drosophila genome is well underway[356,357] and is being enhanced by efforts to perform systematic gene disruptions using *P* transposable elements.[358] The sequencing of the Drosophila genome is now largely complete[354,355] (see http://flybase.harvard.edu:7081).

Among the model organisms, the mouse (as a mammal) is the most closely related to humans in terms of developmental program and biological complexity. This animal provides the closely related features of mammalian development and physiology, but in a system that is powerful in terms of its potential for genetic manipulation. The availability of fully inbred strains and a generation time of a few months make classical genetic studies feasible.[359-361] Furthermore, the ability to manipulate the mouse germline, including the capacity to add or delete specific genes en route to the creation of transgenic mice,[362-367] is providing the means for establishing the function of many genes present in mouse and man[361,368,369]. Also important is the close evolutionary homology in the arrangement of gene segments along the chromosomes.[360,361,370-376a] These features have prompted a systematic effort to construct genetic and physical maps of the mouse genome[359,377-381] (e.g., see http://www-genome.wi.mit.edu), which is roughly the same size as the human genome, as well as to develop increasingly robust computational tools for accessing and analyzing mouse genomic information.[382,383] The long-term goal is to establish the complete sequence of the mouse genome which, by comparison with the human genome sequence, will enable more precise identification of genes and their regulatory elements.[359,360,384-387] In fact, many believe that the rigorous interpretation of the human genome sequence will only be possible by comparison with the mouse genome sequence (and perhaps the genomic sequences of other more distantly related organisms). Such comparisons should reveal important evolutionarily conserved sequences, such as those that encode key structural features of proteins and those that control the regulation of gene expression. Already, significant insight has been gained from pilot-scale comparative analyses of human and mouse cDNA[388,389] and genomic[390-393] sequences.

While the organisms detailed above are included under the official umbrella of the currently planned HGP, the successes in genome mapping and sequencing have catalyzed genome analysis initiatives for a number of other organisms, including the fission yeast (*Schizosaccharomyces pombe*),[394,395] the laboratory rat (*Rattus novegicus*),[396-398] the zebrafish (*Danio rerio*),[399-403] the pufferfish (*Fugu rubripes*),[404-409] various plant species[410] (in particular, *Arabidopsis thaliana*[411-414]), and numerous bacteria and other microbes (see *Highlights of the Human Genome Project* below). Such studies aim to exploit the unique and specialized aspects of the particular organisms for various research applications, to provide a broader collection of organisms from different points in evolution for comparative analyses, and to strengthen the ability to perform genetic-based studies in a larger set of organisms.

## THE HUMAN GENOME PROJECT

### HISTORY OF THE HUMAN GENOME PROJECT

The official beginning of the HGP in the United States was heralded on October 1, 1990. However, the intellectual and administrative processes responsible for the initiation of the project had already been operating for a number of years before this.[415,416] Detailed historical accounts of the HGP have been compiled.[415-424a] In brief, the origins of the HGP are thought by most to date back to a meeting in Alta, Utah in 1984, where the discussion focused on the analysis of DNA for the purpose of

detecting mutations among atomic bomb survivors.[67] Shortly after this meeting, the concept of a comprehensive program of genome study was entertained by two groups. First, a 1985 conference in Santa Cruz, California was convened to examine the feasibility of sequencing the human genome.[68] Second, Charles DeLisi initiated discussions within the Department of Energy about the merits of genome-wide sequencing.[425] Because of their interest in the health effects of radiation and other types of environmental hazards, the Department of Energy viewed establishing the sequence of the human genome as critically important for programs aimed at monitoring DNA sequence changes. Furthermore, DeLisi contended that the Department of Energy, with its expertise in a diversity of complementary fields (e.g., analytical chemistry, applied physics, engineering, computer science) and experience at directing large-scale projects, would be a strong participant for such an endeavor.[425] Additional support for the HGP came independently from Renato Dulbecco, who argued in 1986 that sequencing the human genome and identifying all the encoded genes would be an efficient way to expedite cancer research.[69] Importantly, he stressed that it would be more desirable to elucidate all of this information at once, rather than obtaining it piecemeal over an extended period of time.

Two highly influential reports published in 1988 guided the development of the structure and scope of the early phases of the HGP in the United States—one by the National Research Council Committee on Mapping and Sequencing the Human Genome[426] and the other by a committee operating under the auspices of the United States Congress Office of Technology Assessment.[427] Together, these reports called for a systematic effort of genome mapping and sequencing, provided recommendations about the scope and goals of the effort, outlined the roles for both the National Institutes of Health and the Department of Energy in administering the project in the United States, and recommended funding levels for the endeavor. Of note, the general program outlined by these reports has remained fundamentally unchanged, despite numerous advances in the technologies available for genome analysis. Amidst all these discussions was a significant amount of intense debate within the scientific community as to the merits of the HGP,[42,428-438] although virtually all of the negative aspects of this have since waned. The Department of Energy's Office of Health and Environmental Research initiated their formal program in 1987. The Office for Human Genome Research at the National Institutes of Health was created in 1988; later that year, this office became the National Center for Human Genome Research [NCHGR; note that in 1997, NCHGR became an NIH Institute, the National Human Genome Research Institute (NHGRI)]. Appropriations for both the Department of Energy and National Institutes of Health programs were initiated in 1988 (although the Department of Energy's program started the previous year using funds diverted from other sources). The first set of formal goals for the project in the United States were established in 1990,[439] at which time the project officially began.

While the historic roots of the HGP are largely based in the United States, the project is international in structure and spirit. In addition to the National Institutes of Health and Department of Energy orchestrating the HGP in the United States, there are analogous agencies coordinating efforts underway in other countries, particularly England, France, Germany, Italy, Canada, Australia, Japan, and China. In this regard, the project is truly international in terms of collaboration and coordination.

As an aside, the inclusion of "human" in the name "Human Genome Project" is, of course, a misnomer, since it does not accurately reflect the breadth of the overall initiative. Rather, from the beginning, parallel mapping and sequencing of non-human model organisms have been central components of the HGP (see *Studying the Genomes of Model Organisms* above and *Comparative Study of the Biology of Humans and Other Organisms* below).

## Scientific Plan of the Human Genome Project

In the United States, the currently planned HGP has a 15-year timetable. Since its inception, the project has been associated with carefully crafted milestone-oriented goals that reflect current and realistic near-term capabilities, including an initial set established in 1990[420-422,439] and two successive sets of five-year goals starting in 1993[440] and 1998.[441] The key elements of these goals have focused on establishing infrastructure, developing requisite technologies, and generating the necessary inventories of data. At the same time, the goals have attempted to be visionary, flexible, and integrated with the ongoing planning process but openly acknowledged as transient in nature, due to the continual advances in the technologies for genome analysis.[442] Finally, the formulation of these goals has been accompanied by critical discussions of the evolving plans for the project by some of the key participants.[256,336,342,442]

A summary of the 1998-2003 goals for the United States HGP[441] is provided in Table 10-1. A number of important points about these goals should be emphasized. First, completing the human genome sequence by 2003 represents the highest-priority goal (see *Sequencing the Human Genome* below). In fact, the initiation of this effort occurred two years earlier than originally anticipated. Second, associated with sequencing the human genome are plans to improve further DNA sequencing technology, so as to make the sequencing of other genomes as well as the resequencing of human DNA as efficient and cost-effective as possible. Third, plans to catalog common human sequence variants are included within the HGP for the first time. Studies of human sequence variation represent a critical and rapidly evolving growth area in human genetics.[256] New initiatives will include the construction of a third-generation, SNP-based genetic map of the human genome (with the first two generation genetic maps being composed of RFLPs/VNTRs and STRs, respectively; see *Construction of Genetic Maps* above; also see http://www.ncbi.nlm.nih.gov/SNP). Fourth, also emphasized within the HGP goals are a series of activities falling under the general category of "functional genomics,"[443] an area of genome research that broadly deals with the development and implementation of technologies for exploiting complete genomic sequence. For example, this includes techniques for examining gene expression on a genome-wide scale (e.g., see *DNA Chips* above). At the forefront of many areas of functional genomics are yeast, *C. elegans,* and *D. melanogaster* geneticists, who are fortunate to have complete genomic sequence for their organisms of study. Included among the numerous initiatives in functional genomics are efforts to generate improved (e.g., full-length) cDNA libraries and derive complete cDNA sequences for comprehensive sets of human and other organisms' genes. Fifth, the HGP will continue to emphasize comparative genome analysis of model organisms, for the first time including explicit plans for obtaining the complete mouse genomic sequence. Finally, the HGP goals include efforts to support a range of associated activities, including those aiming to foster the development of improved computational genomics tools, those addressing the important ethical, legal, and social issues relating to genome mapping and sequencing (see *Ethical, Legal, and Social Implications of the Human Genome Project* below), and those supporting the training of individuals in genome research. Of note, the latter will intentionally include the recruitment of investigators with expertise outside of biology (e.g., engineering, chemistry, physics) into the field.[444]

**Table 10-1**
**Summary of the 1998-2003 Goals for the United States Human Genome Project**

**Human Genome Sequence:** Complete the human genome sequence by the end of 2003 while emphasizing the establishment of a "working draft" version for at least 90% of the genome by 2000, the development of a sustainable capacity for large-scale sequencing, the generation of large contiguous stretches of high-quality sequence, and the provision of ready access to the data.

**Sequencing Technology:** Continue incremental improvements in current sequencing methods so as to increase the throughput and reduce the cost of sequencing, with emphasis on automation, miniaturization, and process integration. In parallel, support interdisciplinary research for developing novel sequencing technologies and the means for implementing such technologies into established sequence-producing operations.

**Human Sequence Variation:** Develop the technology for rapid, large-scale identification and scoring of SNPs, with the aims of identifying and cataloguing the common variants in the coding regions of the majority of human genes and creating a human SNP map of at least 100,000 markers. In addition, establish the intellectual foundations and requisite public resources of DNA samples and cell lines for studying human variation.

**Technology for Functional Genomics:** Generate complete sets of full-length cDNA clones for humans and model organisms, develop the technology for defining the spatial and temporal patterns of gene expression, design new strategies for the global study of noncoding sequences, design new approaches for systematic mutagenesis of genes, and advance the understanding of protein function on a genome-wide basis.

**Comparative Genomics:** Complete the sequence of the C. elegans genome by 1998 and the Drosophila genome by 2002. For mouse genome analysis, develop more detailed physical and genetic maps, construct additional cDNA resources, and, by 2005, complete the genomic sequence. Identify and initiate studies on other model organisms that will markedly contribute to the understanding of the human genome.

**Ethical, Legal, and Social Implications (ELSI):** Examine the issues surrounding the completion of the human genome sequence and the study of human genetic variation, study the issues raised by the integration of genetic technologies and information into health care, public health activities, and nonclinical settings, and explore how the new genetic information will influence various societal issues related to genetics.

**Bioinformatics and Computational Biology:** Develop better tools for data generation and capture, improve the content and utility of databases, create mechanisms for sharing and disseminating exportable software, and construct appropriate tools and databases for dealing with comprehensive studies of gene expression and function as well as with sequence homology and variation.

**Training and Manpower:** Facilitate the training of new scientific specialists with expertise in genomics research (including the recruitment of non-biological scientists from fields such as computer science, engineering, mathematics, physics, and chemistry) and aid in the establishment of academic career paths for genome scientists. Increase the number of scholars who are knowledgeable both in genetics and in ethics, law, and social sciences.

**Source: From Collins et al.[441]**

## HIGHLIGHTS OF THE HUMAN GENOME PROJECT

To date, the HGP has achieved virtually all of its well-formulated goals. Various reviews have charted this progress.[23,424,445-447] While numerous individual accomplishments can be cited, several major areas of highlights should be emphasized. These are best appreciated when considered within the context of the planned timetable for the HGP with respect to the construction of genetic, physical, and sequence maps of the human genome (Fig. 10-17).

With respect to genetic map construction, a high-resolution human genetic map consisting of PCR-based, STR-type markers has been assembled[263-267] (e.g., see http://www.genethon.fr/genethon_en.html), with the total number of markers generated to date far exceeding the number proposed for the HGP. Specifically, the original goal for the HGP was the development of ~1500 microsatellite-based genetic markers (thus providing a genetic marker, on average, every 2 to 5 cM); to date over 20,000 such markers have been generated.

With respect to physical map construction, the initial goals of the HGP included establishing complete clone coverage of the human genome and mapping an STS, on average, every ~100 kb across all human chromosomes. Targeted efforts to construct YAC-based physical maps (mostly by STS-content mapping) have been completed for a handful of individual human chromosomes.[93-95,99-107] In addition to these studies have been analogous genome-wide mapping efforts[95,96,98] (e.g., see http://www-genome.wi.mit.edu and http://www.cephb.fr/bio/ceph-genethon-map.html). Attainment of better than 100-kb average STS spacing has been reported for two human chromosomes, X[106] and 7.[107] Supplementing the YAC-based physical map of the human genome is an evolving STS-based radiation hybrid map[291,300] (also see http://shgc-www.stanford.edu).

In parallel to human genome mapping have been efforts to construct genetic and physical maps of the mouse genome. In particular, markedly improved STR-based genetic maps of the mouse genome have been assembled.[377-380] Similarly, a first-generation YAC-based physical map of the mouse genome has been constructed (e.g., see http://www-genome.wi.mit.edu).

With respect to DNA sequencing, the initial emphasis of the HGP was to improve the efficiency of existing methods, to develop new technologies, and to begin systematic sequencing of model organisms. These efforts produced the complete DNA sequences of the *S. cerevisiae*[323-329] (e.g., see http://genome-www.stanford.edu/Saccharomyces), *E. coli,*[330] *C. elegans*[331-335a] (see http://www.sanger.ac.uk and http://genome.wustl.edu/gsc), and *D. melanogaster* genomes.

The refinement of strategies for performing large-scale DNA sequencing within the HGP has led to another major set of accomplishments: the complete genomic sequencing of numerous microorganisms.[448-456] In fact, elucidating the sequence of a whole microbial genome is now considered a relatively straightforward endeavor. Among the sequenced microbes are bacteria of major medical importance, such as *E. coli*,[330] *H. influenzae*,[457] *M.*

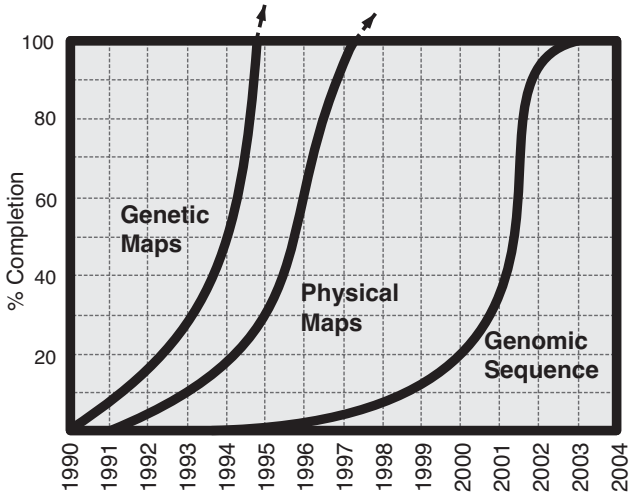


**Fig. 10-17 Timetable for human genome analysis in the HGP. The approximate schedule for completing the human genetic, physical, and sequence maps in the HGP is depicted.**

*tuberculosis*,[458] *H. pylori*,[459] *B. burgdorferi*,[460] *M. genitalium*,[461] *T. pallidum*,[462] *C. trachomatis*,[463] *R. prowazekii*,[463a] and many others. Similar efforts to sequence the genomes of medically important parasitic pathogens, such as *T. brucei* and *P. falciparum*,[463a] are also being performed. An updated listing of sequenced microbial and parasitic genomes is available on the Web (http://www.tigr.org/tdb). In addition to providing insight about the genetic bases of microbial physiology, evolution, and virulence, studying the complete genomic sequence of important infectious pathogens should facilitate the development of more robust diagnostic tests, the design of improved antimicrobial agents, and the identification of candidate vaccine targets.[449,451,453]

In the case of sequencing human DNA, the early emphasis was on the generation of ESTs (see *cDNA Sequencing* above). Large collections of human ESTs have been established[309,310] (see http://www.ncbi.nlm.nih.gov/dbEST), which in turn have been used to construct a radiation-hybrid-mapping-based transcript map of the human genome[292,293] (see http://www.ncbi.nlm.nih.gov/genemap98) that may already contain upwards of half the human genes. Perhaps the most striking highlight of the HGP related to DNA sequencing, however, is the realization that the human genome will be sequenced shortly after we enter the next century[336,338,342] (see *Sequencing the Human Genome* below).

A final highlight of the HGP worth noting is the continual and consistent emphasis on generating high-quality mapping and sequencing data. For example, criteria for monitoring progress and completeness of physical maps have been established.[464,465] Similarly, the maintenance of rigorous accuracy standards for DNA sequencing, in particular for the generated human genome sequence, is viewed as a high priority.[466]

### Sequencing the Human Genome

The early phases of the HGP have brought major advances in the approaches for performing large-scale DNA sequencing. Numerous factors have contributed to this, including subtle improvements in instrumentation, optimized experimental methods, and refined operation of large production groups (see *Genomic Sequencing* above). These developments, in conjunction with the successful construction of physical maps of human chromosomes, resulted in the earlier-than-anticipated launching of efforts to sequence the human genome,[336-342] with the aim of completing the first-generation sequence by 2003 (Fig. 10-17). The year 2003 is particularly significant, since it will mark the 50th anniversary of the discovery of the double helix structure of DNA by James Watson and Francis Crick.[18]

Within the HGP, sequencing of the human genome is being performed using a clone-by-clone approach, whereby individual mapped clones (BACs or PACs) are sequenced, most often by a shotgun sequencing strategy[342, 342a, 343] (see Figs. 10-15 and 10-16). However, other options have been proposed. For example, some have advocated applying a shotgun sequencing strategy to the entire human genome en masse (a "whole genome shotgun").[467] While strong arguments against such a strategy have been made,[468] at least one private company is pursuing this general plan using a recently developed sequencing instrument.[469]

Several additional points about the ongoing efforts to sequence the human genome deserve mention. First, this activity is being carefully coordinated at an international level,[470] so as to complete the sequence as rapidly as possible, avoid unnecessary duplication, and share technical advances and expertise. Second, all sequence data being generated by the publicly funded HGP are made available via the Web on a regular basis (typically nightly). Note that this includes both the final, finished sequence of individual clones as well as the evolving, preliminary sequence data of clones whose analysis is still in progress (see Fig. 10-16). Such a policy of "immediate data release" has been widely supported by the participating groups,[471,472] although this has not been without some debate about its desirability.[473] Third, en route to completing the human genome sequence, a major effort is being made to generate as much preliminary sequence data as rapidly as possible, with the aim of producing a "working draft" version of the sequence for at least 90% of the human genome by the end of 2001.[441] Fourth, care has been taken to protect the individuals whose DNA is being sequenced by the HGP; specifically, several new BAC libraries designated for use in sequencing the human genome have been constructed from the DNA of anonymous individuals. In this regard, it is worth noting that the first-generation sequence of the human genome will consist of a patchwork of sequences from multiple individuals. Finally, the HGP aims to generate a highly accurate sequence of the human genome, with an error rate of $10^{-4}$ (or less) accepted as the standard for finished sequence.[466] Available data indicate that such an accuracy level is being achieved.

## Impact of the Human Genome Project on the Study of Human Disease

The HGP promises to provide a number of interrelated benefits to biology and clinical medicine. These include an improved ability to isolate, characterize, and manipulate the genes involved in normal physiology and human disease. Numerous reviews have detailed how the HGP will impact various areas of clinical medicine.[474-486a]

### Impact of the Human Genome Project on the Positional Cloning of Human Disease Genes

**Background on positional cloning**. Thousands of genes are known to cause disease when present in a mutated form (see http://www.ncbi.nlm.nih.gov/omim). Of course, the number of genes that influence human diseases in a more indirect fashion is undoubtedly much higher. A major effort of modern molecular genetics is to identify the genes that are in some way associated with human disease.

The identification and isolation of human disease genes has largely occurred by one of two basic strategies: functional cloning and positional cloning[1-3] (Fig. 10-18). With functional cloning, the disease gene is isolated as a result of preexisting knowledge of the fundamental physiological defect, which provides sufficient insight about the function of the protein encoded by the defective gene. Often, the cloning of the gene is preceded by the purification of its protein product or by the acquisition of sufficient information about the protein's function. Thus, in functional cloning, mapping the gene follows its isolation (Fig. 10-18). Classic examples of disease genes identified by a functional cloning strategy include β-thalassemia,[487] phenylketonuria,[488] and glucose-6-phosphate dehydrogenase (G6PD) deficiency.[489]

For most of the myriad genetic disorders (including the probable thousands not yet uncovered), there is little to no advanced insight about the function of the defective gene. In the great majority of cases, biochemical studies fail to provide any meaningful clues. For studying these more typical genetic diseases,

the strategy of positional cloning has been refined in recent years[1-8] and successfully employed for isolating numerous disease genes (see http://genome.nhgri.nih.gov/clone). With positional cloning, isolation of the gene follows the establishment of its position within the genome by genetic and/or physical mapping techniques. In most cases, these efforts proceed with limited knowledge of the gene's function or the nature of the underlying pathological process. Thus, in this strategy, mapping precedes cloning (moving in the opposite direction as functional cloning), and gene function is defined only after the gene has been isolated and characterized (Fig. 10-18). While this process was originally called "reverse genetics,"[490-493] this somewhat incorrect terminology has now been abandoned.[1]

**Isolation of disease genes by positional cloning.** The identification and isolation of disease genes by a positional cloning strategy can be conceptually divided into the series of steps depicted in Fig. 10-19, each of which is discussed below.

The starting point for positional cloning is the collection of families with multiple affected individuals, preferably in several generations. Of critical importance is the establishment of the correct phenotype for as many family members as possible and the isolation of DNA from these individuals. Discrepancies between an individual's phenotype and genotype can be due to: (1) an incorrect diagnosis (i.e., phenotype assignment); (2) genetic heterogeneity (defects in more than one gene being associated with the phenotype); (3) incomplete penetrance (some individuals inheriting the defective gene do not express the phenotype, or at least not at the time of evaluation); or (4) a mix-up of the DNA sample(s). Thus, significant effort must be invested during the collection of family resources to minimize errors, especially those caused by preventable mistakes.

In the next stage of positional cloning, the general region of the genome containing the defective gene is identified. In the ideal case, the disease is closely associated with a cytogenetic abnormality(s), which immediately defines the critical region harboring the gene (since such a cytogenetic alteration is likely to have interrupted the gene or its regulatory element). Examples of diseases associated with cytogenetic abnormalities include genetic disorders (e.g., Duchenne muscular dystrophy,[240] neurofibromatosis,[44,494] fragile X syndrome,[241] Lowe syndrome[495]) and various types of cancer[496] (see http://www.ncbi.nlm.nih.gov/CCAP/mitelsum.cgi).

The great majority of genetic diseases are not, however, associated with cytogenetic abnormalities. For these disorders, genetic mapping is used as the front-line tool for identifying the general genomic region containing the gene. At first, DNA samples from multiple family members are analyzed for a set of genetic markers spread across the genome, searching for one (or a few) that shows linkage (coinheritance) with the disease. If successful, this typically assigns the defective gene to a rough chromosomal location, usually spanning tens of millions of base pairs. More refined localization is then accomplished using additional genetic markers that map nearby. The extent to which the region can be delimited by genetic mapping depends on a number of factors, such as the quality of the family resources (e.g., family size, availability of DNA) as well as the number, distribution, and informativeness of the genetic markers in the region. A critical component of this type of genetic mapping is the use of sophisticated computational tools for calculating the various parameters to determine if sufficient data have been collected to allow confident establishment of linkage using the available pedigrees.[261]

In the best cases, genetic mapping allows the critical region to be limited to ~1 cM; however, the corresponding physical size of such a region can vary widely due to regional and sex-specific differences in recombination rates. Typically, ~1 cM in genetic distance corresponds to roughly 1 to 3 Mb, and such a DNA interval can then usually be analyzed by physical mapping methods. Most often, the closest mapping genetic markers are used as the starting points for clone isolation (e.g., YACs, BACs). If necessary, new markers can be derived from the ends of isolated clones and then used to identify new, overlapping clones. The iterative isolation of adjacent, overlapping clones by such a strategy is called "chromosome walking." Additional markers can also be derived from the targeted region by other strategies (see *Generation of STSs* above) and used for clone isolation. In the ideal cases, the entire genomic interval between the flanking genetic markers is isolated in cloned form. The resulting clones can be

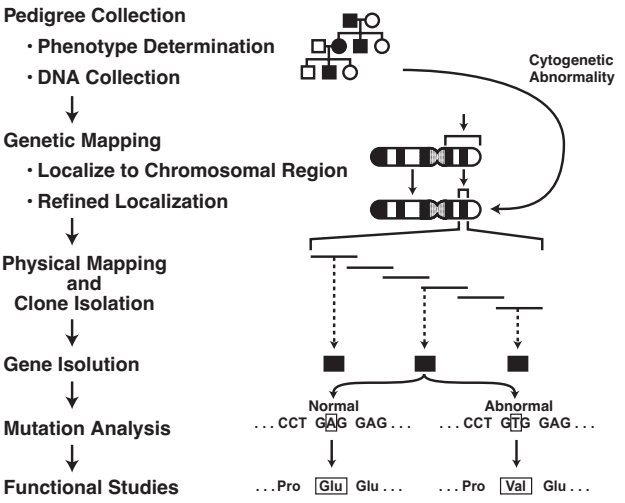


**Fig. 10-19 Positional cloning of human disease genes. The major steps involved in the isolation of human disease genes by a positional cloning strategy are depicted (additional details are provided in the text). In the hypothetical example shown, a single base change (A → T) results in an amino acid change (Glu → Val) in the encoded protein, which in turn causes the disease. Remarkably, such a single-base-pair change in the ~3-billion-base-pair human genome can be lethal.**

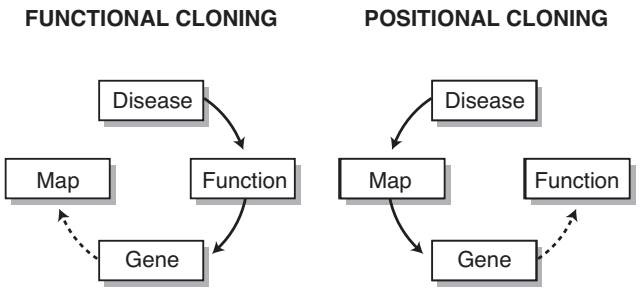


**Fig. 10-18 Functional versus positional cloning of disease genes. In functional cloning, the study of gene function precedes gene identification. In positional cloning, gene mapping precedes gene identification. The last step in each case (gene mapping and defining gene function, respectively) is not critical for the isolation of the disease gene itself. (Adapted from Collins.[1])**

used for developing additional genetic markers (to reduce further the critical region by genetic mapping) and, importantly, for gene isolation.

A particularly challenging step in a positional cloning approach is the identification of encoded genes within the critical region.[7,8,497-501] A number of different strategies have been developed for detecting genes in large genomic clones (e.g., YACs, BACs, PACs, P1s, cosmids) (Table 10-2). In general, each of these methods is relatively labor intensive, suffers from a number of inherent limitations, and is rarely used alone during the search for a disease gene.

One major class of techniques for identifying coding regions within genomic clones employs DNA-DNA hybridization-based methods (Table 10-2). For example, small DNA fragments derived from a starting genomic clone can be used to probe DNA from a variety of different animal species (immobilized on a membrane referred to as a "zoo blot"). Detection of cross-hybridizing (i.e., presumed homologous) DNA suggests the presence of sequences that may have been evolutionarily conserved and therefore are likely to be essential for biological function. The same probes can also be hybridized to mRNA derived from various human tissues (immobilized on a membrane referred to as a "northern blot"[502]) to detect the presence of expressed sequences. A wide survey of tissues is advisable, especially since the expression pattern of the gene is generally unknown at this stage. Finally, the probes can be used in conjunction with rare-cutting restriction enzymes to search for the presence of CpG islands, which often (but not always) mark the 5′ ends of genes.[45-47,503]

Alternatively, genomic clones can be used either to probe directly cDNA libraries[503] (made from the mRNA of particular tissues) or to capture cDNA sequences by a method known as "direct cDNA selection."[504-509] In the latter technique, DNA from the genomic clone is used as an "affinity matrix" to capture complementary sequences present in mRNA mixtures or cDNA libraries. There are a number of important issues regarding the source of mRNA or cDNA libraries for gene isolation. First, a particular gene will only be identified if it is expressed in the tissue from which the mRNA or cDNA library was derived. Second, not all mRNA molecules are expressed at equal levels in a particular tissue. In fact, some genes are expressed at exceedingly low levels, and these can be particularly difficult to isolate. One route for overcoming this general problem is to use "normalized" libraries that contain a more equal representation of different cDNA sequences, regardless of the initial levels of the corresponding mRNA molecules.[303-307]

A second class of strategies for gene isolation exploits particular aspects of gene structure or function (Table 10-2). In one approach, the presence of sequences necessary for proper removal of introns (called "splice junctions") allows the isolation of the adjacent exon by a technique called "exon amplification" or "exon trapping."[510-515] Similar approaches for trapping the most 3′-terminal exon in a gene (using the associated polyA tract as the signal) have also been developed.[515,516] Alternatively, some genes can be isolated by transferring the cloned genomic DNA into an appropriate mammalian cell and selecting for the function of the gene. Of course, the latter approach requires some prior information about the likely function of the gene itself.

A final class of gene identification strategies involves the analysis of DNA sequence derived from the critical region (Table 10-2). Approaches for examining both partial (incomplete) or finished (complete) sequence data are available. For example, one strategy for gene identification involves generating random sequences from a clone(s) of interest such that each nucleotide is read, on average, 1 to 2 times (similar to the first phase of a shotgun sequencing project; see Fig. 10-16 and *Genomic Sequencing* above). All resulting sequence data, which mostly consist of unassembled individual sequence reads, are then compared to available databases (e.g., dbEST, see http://www.ncbi.nlm.nih.gov/dbEST; see *cDNA Sequencing* above) using various computational tools.[225,226,517] Resulting sequence matches are suggestive of genes and require more careful analysis and follow-up studies. This general strategy, called "sample sequencing," has proven effective for the identification of important human disease genes.[518,519] Completed genomic sequence is, of course, even more amenable to computational analyses for identifying genes, with increasingly sophisticated tools becoming available for predicting the presence of genes[229-235] and performing complex sequence comparisons[224-228] (see *Computational Genomics* above).

**Table 10-2**
**Strategies for Identifying Genes in Large Genomic Regions**

**DNA-DNA Hybridization Based**
- **Using Small DNA Fragments**
  - ❖ Analysis of DNA from Other Organisms ("Zoo Blots")
  - ❖ Analysis of RNA from Individual Tissues by Northern Blotting
  - ❖ Identification of CpG Islands
- **Using Large Genomic Clones (e.g., YACs, BACs, PACs, P1s, Cosmids)**
  - ❖ Hybridization-Based Screening of cDNA Libraries
  - ❖ "Direct Selection" of cDNAs Using Immobilized DNA

**Function Based**
- ❖ Exon Trapping/Amplification
- ❖ PolyA Signal Trapping
- ❖ Gene Transfer and Transcript Identification

**DNA Sequence Based**
- ❖ Comparison to Sequence Databases (ESTs, Known Genes)
- ❖ Detection of Open Reading Frames (ORFs)
- ❖ Prediction of Coding Sequences
- ❖ Comparison to Sequences from Other Organisms

**Source: Adapted from Collins.[1]**

All genes identified within a critical region become candidate genes for that disease. Proof that a particular gene is the correct one requires demonstration that the disease is associated with mutations in that gene. Thus, the next stage of analysis involves the difficult task of searching for mutations within candidate genes[520,521] and demonstrating that such mutations show the proper inheritance (e.g., recessive, dominant) relative to the disease. These genetic alterations can range from single-base-pair changes to more gross aberrations (e.g., large deletions,[522] expanded tracts of trinucleotide repeats[523-525]). While mutations can occur anywhere within a gene (including its regulatory elements and introns), the majority of mutations reported to date have been within coding regions.

Often, the initial screening for mutations in affected individuals involves analyzing the gene for gross rearrangements by conventional and pulsed-field gel electrophoresis. Most often, this fails to demonstrate a mutation, and the effort then shifts to searching for more subtle DNA alterations involving one or a few nucleotides.[520,521] A number of techniques have been developed for this purpose, including denaturing gradient gel electrophoresis (DGGE),[526-528] RNAse[529,530] or chemical[531] cleavage of mismatches, single-strand conformation polymorphism (SSCP) analysis,[532] and direct DNA sequencing. Ultimately, DNA sequencing must be used to establish the precise nature of any mutation. With improved methods now available, direct DNA sequencing is increasingly

being used for mutation detection rather than these other methods.

An issue that must be continually addressed during the mutation-detection stage of a positional cloning project is the discrimination between innocent sequence polymorphisms (which may simply be linked to the nearby disease gene) and actual mutations (which cause the disease). Insight about the potential effect of a given mutation (e.g., changing an amino acid at a predicted key site in the protein) can often provide strong supportive evidence for its role in the disease. However, the ultimate proof that a candidate gene is the correct one often requires evidence that the normal form of the gene can correct the abnormal phenotype and/or that the mutant form of the gene can cause the abnormal phenotype.

**Impact of the human genome project on positional cloning.** The HGP is dramatically simplifying the process of positional cloning by improving virtually every one of its steps. First, the availability of higher-resolution genetic maps, better sets of informative STR-type markers, and more efficient methods for genotype analysis is allowing human disease genes to be assigned to more precise genetic locations. Similarly, the construction of comprehensive physical maps of each chromosome, including the assembly of associated clone sets and the localization of genes and ESTs, is dramatically reducing the time it takes to isolate and characterize critical regions for genes of interest. Finally, advances in DNA sequencing are making sequence-based gene discovery and mutation detection more routine and robust.

With the increased mapping of genes and ESTs throughout the human genome, a variant form of positional cloning, termed "positional candidate cloning,[2]" has come to the forefront. In this strategy, a genetically defined critical region for a disease gene is identified and then evaluated for the presence of already mapped, viable candidate genes. Should such an available candidate gene turn out to be the disease gene of interest, then the steps involving the cloning of the region and the isolation of genes can be skipped entirely. Thus, in this approach, information about both the physical position and the likely candidacy of a gene is used to implicate it as the cause of a genetic disease. As a result of the increased number of mapped and characterized genes, it is now more common to identify a human disease gene by a positional candidate approach than by a pure positional cloning approach.

Overall, there is a strong correlation between the amount of mapping and sequencing data generated by the HGP and the rate at which human disease genes have been identified by a positional cloning or positional candidate approach (e.g., compare Fig. 10-17 and http://genome.nhgri.nih.gov/clone). In many cases, the contribution by the HGP to the identification of the gene has been subtle, while in other cases it has been critical.[533-538]

Important future studies in human genetics will involve unraveling the genetic bases of diseases that are particularly complicated to study, such as those that are rare (and have limited family resources available), are caused by defects in more than one gene (polygenic diseases), or are a consequence of combined genetic and nongenetic factors (multifactorial)[13-17]. Perhaps the greatest ultimate impact of the HGP will be to improve the capacity for defining the genetic alterations associated with such medically important, genetically complex diseases. For example, studying polygenic disorders is inherently difficult in humans, in part because of the small pedigree sizes and lack of controlled matings. More detailed genomic maps [including a dense SNP map (see http://www.ncbi.nlm.nih.gov/SNP) and, eventually a whole-genome sequence map] should help to overcome such limitations, allowing more precise correlation between sequence variation and heritable phenotypes to be made.[13-17] In addition, newer technology may eventually allow specific steps in a standard positional cloning strategy to be bypassed. One such technique is genome mismatch scanning,[539,540] which is a sophisticated method that allows the regions that are identical between different genomes (such as those concordant for a particular trait) to be isolated.[539,541,542] This technique, which can be combined with DNA chip detection schemes to identify corresponding clones containing the DNA of interest,[543] may allow the genetic-mapping stage of positional cloning (i.e., marker-by-marker genotyping) to be skipped. Thus, such a method may allow the rapid isolation of those genomic regions containing genes for genetically complex diseases. Together, the improved genomic infrastructure provided by the HGP coupled with more sophisticated technologies should greatly help to define the polygenic factors underlying human disease susceptibilities and predispositions.

## Comparative Study of the Biology of Humans and Other Organisms

From the onset, a major component of the HGP has been to map and sequence the genomes of model organisms whose biological properties have been examined for decades by geneticists, biochemists, and physiologists (see *Studying the Genomes of Model Organisms* above). In fact, analyzing the genomes of model organisms has played a critical role in developing the strategies, technologies, and infrastructure needed for studying human DNA. However, the actual mapping and sequencing of the human genome will almost certainly be more straightforward than the difficult and challenging task of determining the functions of genes and the bases of human genetic disease. In this regard, the knowledge gained from research on model organisms will provide a framework for utilizing the reagents and information produced from human genome studies. The strong emphasis on studying model organisms within the HGP is based on the fundamental feature of biology that all organisms are related through a common evolutionary tree and share the same general type of DNA blueprint, with a tremendous degree of conservation of gene structure and function existing across a diverse array of organisms.

The accomplishments of the HGP have catalyzed several major efforts aiming to make connections between gene structure and function in model organisms and that in man. For example, the availability of the complete yeast sequence[323-329] (e.g., see http://genome-www.stanford.edu/Saccharomyces) is allowing the systematic cross-referencing of yeast genes with those in the human genome,[544-546] with particular emphasis on identifying, cataloging, and studying those genes associated with human disease.[547-551] A similar effort is underway for comparing fruit fly (Drosophila) and human genes.[552]

As a result of these and various other studies, there are now numerous examples where the study of a gene in bacteria, yeast, worms, fruit flies, and/or mice has provided important insight about the function of a particular human gene.[319,345] Remarkably, it is often the case that the human gene can functionally substitute for its counterpart, even in the distantly related yeast *S. cerevisiae*. A very small but illustrative sampling of cases where sequence homology and/or cross-species functional studies have proven valuable includes:

(1) The yeast *STE6* gene, which encodes a protein required for secreting a peptide pheromone factor involved in yeast mating,[553] is highly homologous to the human *MDR1* (multidrug resistance) gene, which encodes a protein that renders tumor cells resistant to a number of chemotherapeutic agents.[554] This strong sequence similarity motivated researchers to transfer the mouse homologue of the *MDR1* gene into a mutant yeast strain defective in the *STE6* gene; remarkably, the mouse gene was able to correct the mutant phenotype in the yeast.[555]

(2) The gene mutated in neurofibromatosis type 1, a common autosomal dominant disease associated with a constellation of symptoms (including characteristic neurofibromas), was cloned[44,494] and found to be highly homologous to the mammalian RAS GTPase activating gene[556,557] and two yeast genes called *IRA1* and *IRA2*.[557,558] Gene transfer experiments demonstrated that a segment of the mammalian neurofibromatosis type 1 gene can complement (i.e., correct) the function of defective *IRA* genes in yeast.[557]

(3) The human *ERCC-3* gene encodes a presumed DNA helicase involved in repairing specific types of DNA damage, and defects in this gene are responsible for two rare genetic diseases: xeroderma pigmentosum and Cockayne syndrome.[559] A Drosophila gene called *haywire* appears to be the fruit fly equivalent of *ERCC-3*, and mutations in *haywire* result in some of the same effects as those seen in xeroderma pigmentosum.[560] However, not all *haywire* alleles are associated with the identical phenotype. This information may help to explain the variability in symptoms seen in different xeroderma pigmentosum and Cockayne patients.

(4) The yeast *SGS1* gene, which encodes another DNA helicase, is closely related to the human Werner's syndrome gene.[561] Mutations in these genes result in a premature aging phenotype in yeast and humans, respectively, revealing a conserved mechanism of cellular aging.

(5) There are now numerous examples where the close evolutionary relationship between mice and humans has been exploited to understand the function of particular human genes. In fact, such studies provide a key rationale for the comparative mapping[374-376a] and sequencing[384-386] of the human and mouse genomes.[359,360,387] Increasingly, gene-characterization studies in either human or mouse quickly broaden to include examining the gene in the other species. This scenario has been particularly valuable for the study of human disease genes[361]. Myriad examples can be cited where important biological insight was gained by combined human and mouse genetic analysis; however, a strikingly dramatic example is the study of mouse obesity mutants.[562] Here, the complicated task of unraveling the genetic bases of human obesity[563] was aided by the identification of the mouse *ob* gene,[564] which encodes the protein leptin. This discovery has revealed a new physiological pathway for weight control and catalyzed numerous new experimental initiatives aimed at examining the role of leptin in human obesity, identifying the other critical components in the leptin pathway, and developing possible pharmacological agents for therapeutic weight control.[565]

In short, it is now clear that comparative genome analysis is a powerful approach for characterizing gene structure and function. In this regard, knowledge of the complete sequence of the human genome and that of an increasing number of other organisms should be regarded as the ultimate framework for deciphering the functional information encoded in DNA.

## Advances in Molecular Diagnostics Resulting from the Human Genome Project

Molecular diagnostics can be broadly defined as the testing of DNA (or RNA) within a clinical context, and this medical discipline is rapidly growing in scope and importance.[566-571] The applications of molecular diagnostics span a wide range of human disorders, including tests for hereditary, neoplastic, and infectious diseases.

The HGP will accelerate the growth of molecular diagnostics in two respects. First, by facilitating the identification of disease genes (see *Impact of the Human Genome Project on Positional Cloning* above), an increasing number of clinically relevant, human mutations will be uncovered. With this growing insight about the genetic bases of disease will come increased opportunities to make diagnostic and prognostic assessments based on examination of an individual's DNA. Second, many of the same methods and instruments being developed to construct genetic, physical, and sequence maps of the human genome are finding immediate utility for testing DNA in a clinical setting.[566-571] A prelude of this phenomenon is already evident with PCR, which is already being used extensively for clinical testing. A typical molecular diagnostic laboratory in the future will likely perform hundreds if not thousands of PCR assays per day, with any refinements made by the HGP for high-throughput PCR testing being of immediate utility. However, it is likely that such a laboratory will not have benches filled with PCR machines, gels, and power supplies; rather, there will inevitably be numerous advances with respect to automation and miniaturization, such as the implementation of various types of DNA chips (see *DNA Chips* above). Thus, the insight gained about the genetic bases of human disease in conjunction with the continued developments in experimental genetic technologies should dramatically enhance the ability to perform diagnostic DNA testing.

## Prospects for Therapeutic Benefits from the Human Genome Project

The HGP promises to transform the ability to understand human genetic diseases by providing a unique interplay between genetics and clinical medicine. For all of the reasons discussed above, physicians and scientists should gain new insights about the genetic components that contribute to disease and acquire better means for establishing whether a patient has inherited a genetic defect. However, the full impact of the HGP should extend beyond these areas, and, in the long run, enhance the ability to treat patients with genetic abnormalities.

A number of interrelated aspects of the HGP offer the potential for having a positive impact on patient care. First, for some genetic disorders, presymptomatic knowledge of an inherited defect can provide meaningful opportunities for the use of preventive measures (e.g., lifestyle alterations, increased surveillance to aid early diagnosis, targeted intervention) that may serve to minimize morbidity. As the HGP progresses, an increasing number of genes for diseases in this category are being discovered. Second, the improved capacity to define the precise molecular defects causing genetic diseases should aid efforts to elucidate the underlying pathophysiology. Such knowledge should facilitate the design of more rational treatments for genetic diseases, which could include the development of better pharmacologic agents, the exogenous synthesis and delivery of a missing gene product, or the introduction of the normal form of a gene into an affected patient (i.e., gene therapy[572-574]). Of note,

the latter might eventually include the use of "mammalian artificial chromosome" (MAC) vectors [575-579] for gene therapy; some of these vectors are being adapted from cloning systems used for genome mapping (e.g., YACs).

One notable and rapidly evolving area of therapeutics that directly relates to the HGP involves unraveling the genetic basis of drug responsiveness. Specifically, this discipline (called "pharmacogenomics" or "pharmacogenetics") focuses on elucidating the genetic determinants that affect drug action, with the long-term goal of establishing diagnostic tests and customized therapeutic regimens that will allow drugs to be prescribed in safer and more effective fashions[580,581]. This involves correlating drug responsiveness with genetic variation, which often will be subtle and complex.[582,583] However, the products of the HGP should make this task more approachable. Successes in pharmacogenomics may help to remove the empiric nature associated with many aspects of drug therapy and provide more rational approaches for predicting how individuals will respond to particular therapeutic modalities.

## ETHICAL, LEGAL, AND SOCIAL IMPLICATIONS OF THE HUMAN GENOME PROJECT

With the fruits of the HGP, including both the mapping and sequencing data and the improved technologies for studying DNA, has come the identification of numerous substantive ethical and policy issues. For example, one of the early anticipated benefits of the HGP—the ability to identify and isolate genes that play important roles in human disease—has indeed become a reality. However, in most cases, cloning a human disease gene is only the first step in the long process towards developing a rational therapy. Since the latter almost always lags behind the generation of new diagnostic tests, the identification of disease genes typically provides the ability to identify individuals at risk for disorders that are associated with limited therapeutic options. Furthermore, with increasing emphasis being placed on the study of human variation and the development of improved technologies for identifying such variation (see Table 10-1), the potential availability of genetic information about individuals will undoubtedly increase. How such information will or should be used by patients, physicians, and society raises a number of issues that require immediate, thoughtful consideration.

To address these concerns, the architects of the United States HGP established the Ethical, Legal, and Social Implications (ELSI) Program as an integral component of the HGP.[584] This program provides a novel approach to the study of ethical, legal, and social issues by carefully integrating its agenda with that of the ongoing genome mapping and sequencing efforts[446,585] (see http://www.nhgri.nih.gov/ELSI). By design, the ELSI Program has brought together individuals with diverse areas of expertise (e.g., medical geneticists, ethicists, historians, theologians, legal scholars, policy analysts, sociologists). The program's mission includes both research studies and education projects for addressing ELSI issues as well as policy analysis and development efforts for translating the empirical research findings into pragmatic policy and programmatic recommendations. The NIH committed initially 3% and later 5% of its total HGP budget to the ELSI Program.The United States Department of Energy also established an ELSI Program (see http:/www.ornl.gov/TechResources Human Genome/resource/elsi.html).

The central mission of the ELSI Program has been to identify and address the key ethical, legal, and social issues relating to the HGP (and other associated genomic and genetic research activities) and to facilitate the establishment and institution of appropriate safeguards[446,585] (see http://www.nhgri.nih.gov/ELSI). The program has focused on four high-priority areas of study: (1) privacy and fair use of genetic information; (2) safe and effective integration of new genetic technologies into clinical practice; (3) issues surrounding genetics research; and (4) public and professional education.

The increased availability of genetic data raises numerous issues about who should have access to this potentially powerful information. Several surveys have shown that this is a real concern to people.[586-588] In the past, genetic information has indeed been used to discriminate against individuals (e.g., sickle-cell anemia carriers). Of particular concern is the fear of losing or being denied health insurance because of a genetic predisposition to a disease. Ironically, the substantive issues surrounding health insurance discrimination threaten both the potential usage of new genetic technologies to improve human health and the ability to conduct the very research needed to understand, treat, and prevent genetic disease. In 1995, the National Action Plan on Breast Cancer (NAPBC) in conjunction with the ELSI Program developed detailed policy recommendations to prohibit genetic discrimination in health insurance.[589] As a result of these recommendations, the Health Insurance Portability and Accountability Act of 1996 [104 Public Law No. 104-191, 701, 110 STAT, 1936 (1996)], which includes prohibitions on the use of genetic information in the group health insurance market, became the first United States Federal law enacted to protect against the misuse of genetic information. Efforts are ongoing to pass additional Federal legislation that would broaden this protection to those with individual health insurance coverage and would in a more general way prohibit health insurers from asking for or using genetic information. In the United States, 30 states have now enacted laws to prevent the use of genetic information by health insurers[590] [Barbara Fuller (NHGRI), personal communication]. Finally, initial policy recommendations to address concerns about the use of genetic information in the workplace have been formulated[591] (see also http://www.nhgri.nih.gov/HGP/Reports/genetics_workplace.html).

Genetic testing is increasingly becoming an integral component of health care. As a result, the ELSI Program has examined the key issues surrounding the introduction of new genetic tests into clinical practice. An initial set of studies explored the testing and counseling for cystic fibrosis mutations, with the aim of examining alternative approaches for genetic education, testing, and counseling.[592] Based on these studies, a 1997 NIH Consensus Conference recommended optimal practices for performing cystic fibrosis genetic testing (see http://odp.od.nih.gov/consensus/statements/cdc/106/106_stmt.html). Similarly, in anticipation of the discovery of cancer predisposing genes, the ELSI Program sponsored studies to examine the psychosocial and clinical impact of genetic testing in families with heritable forms of breast, ovarian, and colon cancer. These studies have resulted in the development of valuable experience-based guidance for implementing genetic tests for cancer susceptibility.[593-596] Similar investigations relating to genetic testing for Alzheimer disease[597] and hemochromatosis[598] have also been performed. In 1994, the NIH and the Department of Energy created the Task Force on Genetic Testing to evaluate genetic testing in the United States and to make recommendations for ensuring that such tests are safe and effective. The resulting report contains recommendations for federal agencies, testing laboratories, and health professionals[599] (see http://www.nhgri.nih.gov/ELSI/TFGT_final).

As genetic testing increases, the use and interpretation of those tests will become the responsibility of a wider array of health

professionals beyond those considered to be "genetics specialists," including primary care physicians, nurses, physician assistants, nurse practitioners, psychologists, and social workers. In addition, public policy makers and the general public will increasingly be called upon to consider critical issues relating to genetic testing. In anticipation of these situations, the ELSI Program has initiated various educational efforts that aim to train health care professionals in interpreting new DNA-based diagnostic tests, to increase public genetic literacy through both the schools and the media, to encourage public discussion about genetic issues, and to provide genetic education for appropriate public policy makers. One such initiative is the establishment of the National Coalition for Health Professional Education in Genetics (NCHPEG), a coordinated effort to promote health professional education about advances in human genetics. While health professionals are the primary target audience for NCHPEG, the program plans to include efforts that aim to provide appropriate materials and guidance for educating policy makers and, through health care providers, patients, families, and the general public.

Adequate consideration of the ethical, legal, and social issues concerning human genome research and the HGP is critical for the successful introduction of genetic information into the mainstream of medical practice and society. The ELSI Program has established a solid foundation for addressing these issues and will undoubtedly continue to provide critical leadership in this area.

## CONCLUSION

The HGP is one of the most important projects—if not *the* most important project—ever undertaken in biomedical research. It is fundamentally an endeavor aiming to develop tools for the study of biology and medicine. These tools reflect both an information resource in the form of genetic, physical, and sequence maps of the human genome and of several model organisms, as well as an ever-increasing number of experimental technologies that are becoming standard techniques in the armamentarium of biomedical researchers. In this regard, an exciting new "genomic revolution" has started and it is permanently changing the way research is performed. The new and powerful foundation of genetic information is empowering investigators to tackle complex problems relating to disease, development, and evolution that were previously unapproachable. The direct impact of the HGP on clinical medicine is already being realized, with both a dramatic acceleration in the identification of human disease genes and the continual development of new approaches for analyzing patient DNA. Ultimately, though, the true legacy of the HGP will be to provide future generations of scientists and clinicians an unprecedented resource—the human "genetic blueprint"—that will allow them to define better the genetic bases of disease and to use that information for designing more effective therapies.


## ACKNOWLEDGMENTS

I wish to thank Francis Collins, David Schlessinger, Kathy Hudson, Maynard Olson, Andy Baxevanis, Mark Boguski, Gerry Bouffard, Dana Jordan, Tyrone Summers, and Gabriela Adelt Green for helpful suggestions, Rick Myers and David Cox for contributions to the earlier version of this chapter, Renee Gamborg and Rob Lazott for administrative assistance, and Maggie Green for her dedicated and loyal support.

## REFERENCES


1. Collins FS: Positional cloning: Let's not call it reverse anymore. *Nat Genet* **1**:3, 1992.
2. Ballabio A: The rise and fall of positional cloning? *Nat Genet* **3**:277, 1993.
3. Collins FS: Positional cloning moves from perditional to traditional. *Nat Genet* **9**:347, 1995.
4. Nelson DL: Positional cloning reaches maturity. *Curr Opin Genet Dev* **5**:298, 1995.
5. Foote S: Genetic analysis of disease susceptibility (disease susceptibility). *Aust NZ J Med* **25**:757, 1995.
6. Wolff RK: Positional cloning: A review and perspective. *Drug Development Research* **41**:129, 1997.
7. Boehm T: Positional cloning and gene identification. *Methods* **14**:152, 1998.
8. Ballabio A, Brown S, Fischer E: Strategies for gene discovery in mammalian systems, in Birren B, Green ED, Klapholz S, Myers RM, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 2 Detecting Genes*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1998, p 1.
9. Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, Rozmahel R, et al.: Identification of the cystic fibrosis gene: Chromosome walking and jumping. *Science* **245**:1059, 1989.
10. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, et al.: Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA. *Science* **245**:1066, 1989.
11. The Huntington's Disease Collaborative Research Group: A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**:971, 1993.
12. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, et al.: A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat Genet* **13**:399, 1996.
13. Bell JI: Polygenic disease. *Curr Opin Genet Develop* **3**:466, 1993.
14. Lander ES, Schork NJ: Genetic dissection of complex traits. *Science* **265**:2037, 1994.
15. Kruglyak L, Lander ES: High-resolution genetic mapping of complex traits. *Am J Hum Genet* **56**:1212, 1995.
16. Lander E, Kruglyak L: Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nat Genet* **11**:241, 1995.
17. Zhang H, Zhao H, Merikangas K: Strategies to identify genes for complex diseases. *Ann Med* **29**:493, 1997.
18. Watson JD, Crick FHC: Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171**:737, 1953.
19. Watson JD, Gilman M, Witkowski J, Zoller M: *Recombinant DNA,* New York, W.H. Freeman and Company, 1992.
20. Morton NE: Parameters of the human genome. *Proc Natl Acad Sci USA* **88**:7474, 1991.
21. Trask B, van den Engh G, Mayall B, Gray JW: Chromosome heteromorphism quantified by high-resolution bivariate flow karyotyping. *Am J Hum Genet* **45**:739, 1989.
22. McKusick VA: Mapping and sequencing the human genome. *N Engl J Med* **320**:910, 1989.
23. Olson MV: The human genome project. *Proc Natl Acad Sci USA* **90**:4338, 1993.
24. Tyler-Smith C, Willard HF: Mammalian chromosome structure. *Curr Opin Genet Develop* **3**:390, 1993.
25. Biessmann H, Mason JM: Telomeric repeat sequences. *Chromosoma* **103**:154, 1994.
26. Moyzis RK, Buckingham JM, Cram LS, Dani M, Deaven LL, Jones MD, Meyne J, et al.: A highly conserved repetitive DNA sequence, $(TTAGGG)_n$, present at the telomeres of human chromosomes. *Proc Natl Acad Sci USA* **85**:6622, 1988.
27. Riethman HC, Moyzis RK, Meyne J, Burke DT, Olson MV: Cloning human telomeric DNA fragments into *Saccharomyces cerevisiae* using a yeast-artificial chromosome vector. *Proc Natl Acad Sci USA* **86**:6240, 1989.
28. Brown WRA: Molecular cloning of human telomeres in yeast. *Nature* **338**:774, 1989.
29. Cross SH, Allshire RC, McKay SJ, McGill NI, Cooke HJ: Cloning of human telomeres by complementation in yeast. *Nature* **338**:771, 1989.
30. Cheng J-F, Smith CL, Cantor CR: Isolation and characterization of a human telomere. *Nucleic Acids Res* **17**:6109, 1989.
31. Dobson MJ, Brown WRA: Cloning human telomeres in yeast artificial chromosomes, in Anand R (ed.): *Techniques for the Analysis of Complex Genomes*, London, Academic Press, 1992, p 81.
32. National Institutes of Health and Institute of Molecular Medicine Collaboration: A complete set of human telomeric probes and their clinical application. *Nat Genet* **14**:86, 1996.
33. Riethman H: Closing in on telomeric closure. *Genome Res* **7**:853, 1997.
34. Murphy TD, Karpen GH: Centromeres take flight: Alpha satellite and the quest for the human centromere. *Cell* **93**:317, 1998.
35. Devilee P, Slagboom P, Cornelisse CJ, Pearson PL: Sequence heterogeneity within the human alphoid repetitive DNA family. *Nucleic Acids Res* **14**:2059, 1986.
36. Waye JS, Willard HF: Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: A survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res* **15**:7549, 1987.
37. Choo KH, Vissel B, Nagy A, Earle E, Kalitsis P: A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. *Nucleic Acids Res* **19**:1179, 1991.
38. Jurka J, Pethiyagoda C: Simple repetitive DNA sequences from primates: Compilation and analysis. *J Mol Evol* **40**:120, 1995.
39. Jurka J: Repeats in genomic DNA: Mining and meaning. *Curr Opin Struct Biol* **8**:333, 1998.
40. Jurka J, Walichiewicz J, Milosavljevic A: Prototypic sequences for human repetitive DNA. *J Mol Evol* **35**:286, 1992.
41. Ohno S: An argument for the genetic simplicity of man and other mammals. *J Hum Evol* **1**:651, 1972.
42. Gall JG: Human genome sequencing. *Science* **233**:1367, 1986.
43. Levinson B, Kenwrick S, Lakich D, Hammonds Jr. G, Gitschier J: A transcribed gene in an intron of the human factor VIII gene. *Genomics* **7**:1, 1990.
44. Wallace MR, Marchuk DA, Andersen LB, Letcher R, Odeh HM, Saulino AM, Fountain JW, et al.: Type 1 neurofibromatosis gene: Identification of a large transcript disrupted in three NF1 patients. *Science* **249**:181, 1990.
45. Bird AP: CpG-rich islands and the function of DNA methylation. *Nature* **321**:209, 1986.
46. Cross SH, Bird AP: CpG islands and genes. *Curr Opin Genet Dev* **5**:309, 1995.
47. John RM, Cross SH: Gene detection by the identification of CpG islands, in Birren B, Green ED, Klapholz S, Myers RM, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 2 Detecting Genes*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1998, p 217.
48. Korenberg JR, Rykowski MC: Human genome organization: Alu, Lines, and the molecular structure of metaphase chromosome bands. *Cell* **53**:391, 1988.
49. Bickmore WA, Sumner AT: Mammalian chromosome banding-An expression of genome organization. *Trends Genet* **5**:143, 1989.
50. Craig JM, Bickmore WA: The distribution of CpG islands in mammalian chromosomes. *Nat Genet* **7**:376, 1994.
51. Gardiner K: Human genome organization. *Curr Opin Genet Dev* **5**:315, 1995.
52. Bernardi G: The vertebrate genome: Isochores and evolution. *Mol Biol Evol* **10**:186, 1993.
53. Bernardi G: The human genome: Organization and evolutionary history. *Ann Rev Genetics* **29**:445, 1995.
54. Saccone S, Caccio S, Kusuda J, Andreozzi L, Bernardi G: Identification of the gene-richest bands in human chromosomes. *Gene* **174**:85, 1996.
55. Zoubak S, Clay O, Bernardi G: The gene distribution of the human genome. *Gene* **174**:95, 1996.
56. Saccone S, De Sario A, Della Valle G, Bernardi G: The highest gene concentrations in the human genome are in telomeric bands of metaphase chromosomes. *Proc Natl Acad Sci USA* **89**:4913, 1992.
57. *Genome Analysis: A Laboratory Manual. Vols. 1-4*: Green, ED, Birren, B, Klapholz, S, Myers, RM, and Hieter, P (eds). Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1997.
58. Sambrook J, Fritsch EF, Maniatis T: *Molecular Cloning: A Laboratory Manual, 2nd edition,* Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1989.
59. Wolff R, Gemmill R: Purifying and analyzing genomic DNA, in Birren B, Green ED, Klapholz S, Myers RM, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 1 Analyzing DNA*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1997, p 1.
60. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H: Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* **LI**:263, 1986.
61. Mullis KB, Faloona FA: Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction, in Wu R (ed.): *Methods in Enzymology (Volume 155)*, San Diego, Academic Press, Inc., 1987, p 335.
62. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N: Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**:1350, 1985.
63. Fanning S, Gibbs RA: PCR in genome analysis, in Birren B, Green ED, Klapholz S, Myers RM, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 1 Analyzing DNA*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1997, p 249.
64. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, et al.: Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**:487, 1988.

65. Lashkari DA, Hunicke-Smith SP, Norgren RM, Davis RW, Brennan T: An automated multiplex oligonucleotide synthesizer: Development of high-throughput, low-cost DNA synthesis. *Proc Natl Acad Sci USA* **92**:7912, 1995.
66. Rayner S, Brignac S, Bumeister R, Belosludtsev Y, Ward T, Grant O, O'Brien K, et al.: MerMade: An oligodeoxyribonucleotide synthesizer for high throughput oligonucleotide production in dual 96-well plates. *Genome Res* **8**:741, 1998.
67. Cook-Deegan RM: The Alta Summit, December 1984. *Genomics* **5**:661, 1989.
68. Sinsheimer RL: The Santa Cruz Workshop-May 1985. *Genomics* **5**:954, 1989.
69. Dulbecco R: A turning point in cancer research: Sequencing the human genome. *Science* **231**:1055, 1986.
70. Evans GA: Cosmids, in Birren B, Green ED, Klapholz S, Myers RM, Riethman H, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 3 Cloning Systems*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1998, p 87.
71. Evans GA, Lewis KA: Physical mapping of complex genomes by cosmid multiplex analysis. *Proc Natl Acad Sci USA* **86**:5030, 1989.
72. Nizetic D, Zehetner G, Monaco AP, Gellen L, Young BD, Lehrach H: Construction, arraying, and high-density screening of large insert libraries of human chromosomes X and 21: Their potential use as reference libraries. *Proc Natl Acad Sci USA* **88**:3233, 1991.
73. Stallings RL, Torney DC, Hildebrand CE, Longmire JL, Deaven LL, Jett JH, Doggett NA, et al.: Physical mapping of human chromosomes by repetitive sequence fingerprinting. *Proc Natl Acad Sci USA* **87**:6218, 1990.
74. Ivens AC, King TF, Little PFR: Cosmid clones and generation of detailed DNA maps of large chromosomal regions, in Adolph KW (ed.): *Methods in Molecular Genetics (Gene and Chromosome Analysis, Part A)*, San Diego, Academic Press, 1993, p 173.
75. Burke DT, Carle GF, Olson MV: Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236**:806, 1987.
76. Green ED, Hieter P, Spencer FA: Yeast artificial chromosomes, in Birren B, Green ED, Klapholz S, Myers RM, Riethman H, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 3 Cloning Systems*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1998, p 297.
77. Burke DT, Olson MV: Preparation of clone libraries in yeast artificial-chromosome vectors. *Methods Enzymol* **194**:251, 1991.
78. Brownstein BH, Silverman GA, Little RD, Burke DT, Korsmeyer SJ, Schlessinger D, Olson MV: Isolation of single-copy human genes from a library of yeast artificial chromosome clones. *Science* **244**:1348, 1989.
79. Imai T, Olson MV: Second-generation approach to the construction of yeast artificial-chromosome libraries. *Genomics* **8**:297, 1990.
80. Albertsen HM, Abderrahim H, Cann HM, Dausset J, Le Paslier D, Cohen D: Construction and characterization of a yeast artificial chromosome library containing seven haploid genome equivalents. *Proc Natl Acad Sci USA* **87**:4256, 1990.
81. Dausset J, Ougen P, Abderrahim H, Billault A, Sambucy J-L, Cohen D, Le Paslier D: The CEPH YAC library. *Behring Inst Mitt* **91**:13, 1992.
82. Anand R, Riley JH, Butler R, Smith JC, Markham AF: A 3.5 genome equivalent multi access YAC library: Construction, characterisation, screening and storage. *Nucleic Acids Res* **18**:1951, 1990.
83. Anand R, Villasante A, Tyler-Smith C: Construction of yeast artificial chromosome libraries with large inserts using fractionation by pulsed-field gel electrophoresis. *Nucleic Acids Res* **17**:3425, 1989.
84. Larin Z, Monaco AP, Lehrach H: Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. *Proc Natl Acad Sci USA* **88**:4123, 1991.
85. Green ED, Olson MV: Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. *Proc Natl Acad Sci USA* **87**:1213, 1990.
86. Green ED, Olson MV: Chromosomal region of the cystic fibrosis gene in yeast artificial chromosomes: A model for human genome mapping. *Science* **250**:94, 1990.
87. Anand R, Ogilvie DJ, Butler R, Riley JH, Finniear RS, Powell SJ, Smith JC, et al.: A yeast artificial chromosome contig encompassing the cystic fibrosis locus. *Genomics* **9**:124, 1991.
88. Coffey AJ, Roberts RG, Green ED, Cole CG, Butler R, Anand R, Giannelli F, et al.: Construction of a 2.6-Mb contig in yeast artificial chromosomes spanning the human dystrophin gene using an STS-based approach. *Genomics* **12**:474, 1992.
89. Monaco AP, Walker AP, Millwood I, Larin Z, Lehrach H: A yeast artificial chromosome contig containing the complete Duchenne muscular dystrophy gene. *Genomics* **12**:465, 1992.
90. Geraghty DE, Pei J, Lipsky B, Hansen JA, Taillon-Miller P, Bronson SK, Chaplin DD: Cloning and physical mapping of the *HLA* class I region spanning the *HLA-E*-to-*HLA-F* interval by using yeast artificial chromosomes. *Proc Natl Acad Sci USA* **89**:2669, 1992.
91. Zuo J, Robbins C, Taillon-Miller P, Cox DR, Myers RM: Cloning of the Huntington disease region in yeast artificial chromosomes. *Hum Mol Genet* **1**:149, 1992.
92. Bates GP, Valdes J, Hummerich H, Baxendale S, Le Paslier DL, Monaco AP, Tagle D, et al.: Characterization of a yeast artificial chromosome contig spanning the Huntington's disease gene candidate region. *Nat Genet* **1**:180, 1992.
93. Foote S, Vollrath D, Hilton A, Page DC: The human Y chromosome: Overlapping DNA clones spanning the euchromatic region. *Science* **258**:60, 1992.
94. Vollrath D, Foote S, Hilton A, Brown LG, Beer-Romero P, Bogan JS, Page DC: The human Y chromosome: A 43-interval map based on naturally occurring deletions. *Science* **258**:52, 1992.
95. Chumakov IM, Rigault P, Guillou S, Ougen P, Billaut A, Guasconi G, Gervy P, et al.: Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* **359**:380, 1992.
96. Cohen D, Chumakov I, Weissenbach J: A first-generation physical map of the human genome. *Nature* **366**:698, 1993.
97. Chumakov IM, Rigault P, Le Gall I, Bellanne-Chantelot C, Billault A, Guillou S, Soularue P, et al.: A YAC contig map of the human genome. *Nature* **377**:175, 1995.
98. Hudson TJ, Stein LD, Gerety SS, Ma J, Castle AB, Silva J, Slonim DK, et al.: An STS-based map of the human genome. *Science* **270**:1945, 1995.
99. Gemmill RM, Chumakov I, Scott P, Waggoner B, Rigault P, Cypser J, Chen Q, et al.: A second-generation YAC contig map of human chromosome 3. *Nature* **377**:299, 1995.
100. Krauter K, Montgomery K, Yoon S-J, LeBlanc-Straceski J, Renault B, Marondel I, Herdman V, et al.: A second-generation YAC contig map of human chromosome 12. *Nature* **377**:321, 1995.
101. Collins JE, Cole CG, Smink LJ, Garrett CL, Leversha MA, Soderlund CA, Maslen GL, et al.: A high-density YAC contig map of human chromosome 22. *Nature* **377**:367, 1995.
102. Korenberg JR, Chen X-N, Mitchell S, Fannin S, Gerwehr S, Cohen D, Chumakov I: A high-fidelity physical map of human chromosome 21q in yeast artificial chromosomes. *Genome Res* **5**:427, 1995.
103. Bell CJ, Budarf ML, Nieuwenhuijsen BW, Barnoski BL, Buetow KH, Campbell K, Colbert AME, et al.: Integration of physical, breakpoint and genetic maps of chromosome 22: Localization of 587 yeast artificial chromosomes with 238 mapped markers. *Hum Mol Genet* **4**:59, 1995.
104. Quackenbush J, Davies C, Bailis JM, Khristich JV, Diggle K, Marchuck Y, Tobin J, et al.: An STS content map of human chromosome 11: Localization of 910 YAC clones and 109 islands. *Genomics* **29**:512, 1995.
105. Qin S, Nowak NJ, Zhang J, Sait SNJ, Mayers PG, Higgins MJ, Cheng Y-J, et al.: A high-resolution physical map of human chromosome 11. *Proc Natl Acad Sci USA* **93**:3149, 1996.
106. Nagaraja R, MacMillan S, Kere J, Jones C, Griffin S, Schmatz M, Terrell J, et al.: X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. *Genome Res* **7**:210, 1997.
107. Bouffard GG, Idol JR, Braden VV, Iyer LM, Cunningham AF, Weintraub LA, Touchman JW, et al.: A physical map of human chromosome 7: An integrated YAC contig map with average STS spacing of 79 kb. *Genome Res* **7**:673, 1997.
108. Silverman GA, Green ED, Young RL, Jockel JI, Domer PH, Korsmeyer SJ: Meiotic recombination between yeast artificial chromosomes yields a single clone containing the entire *BCL2* protooncogene. *Proc Natl Acad Sci USA* **87**:9913, 1990.
109. Den Dunnen JT, Grootscholten PM, Dauwerse JG, Walker AP, Monaco AP, Butler R, Anand R, et al.: Reconstruction of the 2.4 Mb human DMD-gene by homologous YAC recombination. *Hum Mol Genet* **1**:19, 1992.
110. Green ED, Riethman HC, Dutchik JE, Olson MV: Detection and characterization of chimeric yeast artificial-chromosome clones. *Genomics* **11**:658, 1991.
111. Sternberg N: Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. *Proc Natl Acad Sci USA* **87**:103, 1990.
112. Sternberg N: Cloning into bacteriophage P1 vectors, in Birren B, Green ED, Klapholz S, Myers RM, Riethman H, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 3 Cloning Systems*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1998, p 203.
113. Shizuya H, Birren B, Kim U-J, Mancino V, Slepak T, Tachiiri Y, Simon M: Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc Natl Acad Sci USA* **89**:8794, 1992.
114. Birren B, Mancino V, Shizuya H: Bacterial artificial chromosomes, in Birren B, Green ED, Klapholz S, Myers RM, Riethman H, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 3 Cloning Systems*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1998, p 241.
115. Ioannou PA, Amemiya CT, Garnes J, Kroisel PM, Shizuya H, Chen C, Batzer MA, et al.: A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nat Genet* **6**:84, 1994.
116. Kim U-J, Birren BW, Slepak T, Mancino V, Boysen C, Kang H-L, Simon MI, et al.: Construction and characterization of a human bacterial artificial chromosome library. *Genomics* **34**:213, 1996.

117. Asakawa S, Abe I, Kudoh Y, Kishi N, Wang Y, Kubota R, Kudoh J, et al.: Human BAC library: Construction and rapid screening. *Gene* **191**:69, 1997.
118. Dunham I, Dewar K, Kim U-J, Ross M: Bacterial cloning systems, in Birren B, Green ED, Klapholz S, Myers RM, Riethman H, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 3 Cloning Systems*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1998, p 1.
119. Bentley DR, Todd C, Collins J, Holland J, Dunham I, Hassock S, Bankier A, et al.: The development and application of automated gridding for efficient screening of yeast and bacterial ordered libraries. *Genomics* **12**:534, 1992.
120. Ross MT, Hoheisel JD, Monaco AP, Larin Z, Zehetner G, Lehrach H: High-density gridded YAC filters: Their potential as genome mapping tools, in Anand R (ed.): *Techniques for the Analysis of Complex Genomes*, London, Academic Press, 1992, p 137.
121. Cox RD, Meier-Ewert S, Ross M, Larin Z, Monaco AP, Lehrach H: Genome mapping and cloning of mutations using yeast artificial chromosomes. *Methods Enzymol* **225**:637, 1993.
122. Olsen AS, Combs J, Garcia E, Elliott J, Amemiya C, de Jong P, Threadgill G: Automated production of high density cosmid and YAC colony filters using a robotic workstation. *BioTechniques* **14**:116, 1993.
123. Copeland A, Lennon G: Rapid arrayed filter production using the 'ORCA' robot. *Nature* **369**:421, 1994.
124. Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, et al.: High throughput fingerprint analysis of large-insert clones. *Genome Res* **7**:1072, 1997.
125. Gregory SG, Howell GR, Bentley DR: Genome mapping by fluorescent fingerprinting. *Genome Res* **7**:1162, 1997.
126. McPherson JD: Sequence ready-or not? *Genome Res* **7**:1111, 1997.
127. Schwartz DC, Cantor CR: Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**:67, 1984.
128. Carle GF, Frank M, Olson MV: Electrophoretic separations of large DNA molecules by periodic inversion of the electric field. *Science* **232**:65, 1986.
129. Chu G, Vollrath D, Davis RW: Separation of large DNA molecules by contour-clamped homogenous electric fields. *Science* **234**:1582, 1986.
130. Birren B, Lai E: *Pulsed field gel electrophoresis: A practical guide,* San Diego, Academic Press, 1993.
131. Wrestler JC, Lipes BD, Birren BW, Lai E: Pulsed-field gel electrophoresis. *Methods Enzymol* **270**:255, 1996.
132. Riethman H, Birren B, Gnirke A: Preparation, manipulation, and mapping of HMW DNA, in Birren B, Green ED, Klapholz S, Myers RM, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 1 Analyzing DNA*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1997, p 83.
133. Carle GF, Olson MV: An electrophoretic karyotype for yeast. *Proc Natl Acad Sci USA* **82**:3756, 1985.
134. Link AJ, Olson MV: Physical map of the *Saccharomyces cerevisiae* genome at 110-kilobase resolution. *Genetics* **127**:681, 1991.
135. Smith DR: Genomic long-range restriction mapping. *Methods: A Companion to Methods in Enzymology* **1**:195, 1990.
136. Poustka A: Physical mapping by PFGE. *Methods: A Companion to Methods in Enzymology* **1**:204, 1990.
137. Evans GA: Physical mapping of the human genome by pulsed field gel analysis. *Curr Opin Genet Develop* **1**:75, 1991.
138. Burmeister M: Strategies for mapping large regions of mammalian genomes, in Burmeister M, Ulanovsky L (eds.): *Methods in Molecular Biology (Volume 12: Pulsed-Field Gel Electrophoresis)*, Totowa, The Humana Press, Inc., 1992, p 259.
139. Bickmore W: Analysis of genomic DNAs by pulsed-field gel electrophoresis, in Anand R (ed.): *Techniques for the Analysis of Complex Genomes*, London, Academic Press, Inc., 1992, p 19.
140. Chandrasekharappa SC, Marchuk DA, Collins FS: Analysis of yeast artificial chromosome clones, in Burmeister M, Ulanovsky L (eds.): *Methods in Molecular Biology (Volume 12: Pulsed-Field Gel Electrophoresis)*, Totowa, NJ, The Human Press, Inc., 1992, p 235.
141. Bentley DR: The analysis of YAC clones, in Anand R (ed.): *Techniques for the Analysis of Complex Genomes*, London, Academic Press Limited, 1992, p 113.
142. Den Dunnen JT, van Ommen G-JB: Application of pulsed-field gel electrophoresis to genetic diagnosis, in Mathew C (ed.): *Methods in Molecular Biology Volume 9: Protocols in Human Molecular Genetics*, Clifton, The Humana Press, Inc., 1991, p 313.
143. Pinkel D, Straume T, Gray JW: Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proc Natl Acad Sci USA* **83**:2934, 1986.
144. Hozier JC, Davis LM: Cytogenetic approaches to genome mapping. *Anal Biochem* **200**:205, 1992.
145. Buckle VJ, Kearney L: New methods in cytogenetics. *Curr Opin Genet Dev* **4**:374, 1994.
146. van Ommen G-JB, Breuning MH, Raap AK: FISH in genome research and molecular diagnostics. *Curr Opin Genet Dev* **5**:304, 1995.
147. Heng HH, Spyropoulos B, Moens PB: FISH technology in chromosome and genome research. *BioEssays* **19**:75, 1997.
148. Trask B: Fluorescence in situ hybridization, in Birren B, Green ED, Hieter P, Klapholz S, Myers RM, Riethman H, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 4 Mapping Genomes*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1999, p 303.
149. Cherif D, Julier C, Delattre O, Derre J, Lathrop GM, Berger R: Simultaneous localization of cosmids and chromosome R-banding by fluorescence microscopy: Application to regional mapping of human chromosome 11. *Proc Natl Acad Sci USA* **87**:6639, 1990.
150. Lichter P, Chang Tang C-J, Call K, Hermanson G, Evans GA, Housman D, Ward DC: High-resolution mapping of human chromosome 11 by *in situ* hybridization with cosmid clones. *Science* **247**:64, 1990.
151. Palotie A, Heiskanen M, Laan M, Horelli-Kuitunen N: High-resolution fluorescence *in situ* hybridization: A new approach in genome mapping. *Finnish Medical Society DUODECIM* **28**:101, 1996.
152. Heiskanen M, Peltonen L, Palotif A: Visual mapping by high resolution FISH. *Trends Genet* **12**:379, 1996.
153. Michalet X, Ekong R, Fougerousse F, Rousseaux S, Schurra C, Hornigold N, van Slegtenhorst M, et al.: Dynamic molecular combing: Stretching the whole human genome for high-resolution studies. *Science* **277**:1518, 1997.
154. Green ED, Idol JR, Mohr-Tidwell RM, Braden VV, Peluso DC, Fulton RS, Massa HF, et al.: Integration of physical, genetic and cytogenetic maps of human chromosome 7: Isolation and analysis of yeast artificial chromosome clones for 117 mapped genetic markers. *Hum Mol Genet* **3**:489, 1994.
155. Bray-Ward P, Menninger J, Lieman J, Desai T, Mokady N, Banks A, Ward DC: Integration of the cytogenetic, genetic, and physical maps of the human genome by FISH mapping of CEPH YAC clones. *Genomics* **32**:1, 1996.
156. Schrock E, du Manoir S, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA, Ning Y, et al.: Multicolor spectral karyotyping of human chromosomes. *Science* **273**:494, 1996.
157. Speicher MR, Gwyn Ballard S, Ward DC: Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nat Genet* **12**:368, 1996.
158. Ried T, Schrock E, Ning Y, Wienberg J: Chromosome painting: A useful art. *Hum Mol Genet* **7**:1619, 1998.
159. Cox DR, Burmeister M, Price ER, Kim S, Myers RM: Radiation hybrid mapping: A somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250**:245, 1990.
160. Matise TC, Wasmuth JJ, Myers RM, McPherson JD: Somatic cell genetics and radiation hybrid mapping, in Birren B, Green ED, Hieter P, Klapholz S, Myers RM, Riethman H, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 4 Mapping Genomes*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1999, p 259.
161. *Automated DNA sequencing and analysis*: Adams, MD, Fields, C, and Venter, JC (eds). San Diego, Academic Press, Inc., 1994.
162. Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**:5463, 1977.
163. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, et al.: Fluorescence detection in automated DNA sequence analysis. *Nature* **321**:674, 1986.
164. Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, et al.: A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**:336, 1987.
165. Hunkapiller T, Kaiser RJ, Koop BF, Hood L: Large-scale and automated DNA sequence determination. *Science* **254**:59, 1991.
166. Wilson RK, Mardis ER: Fluorescence-based DNA sequencing, in Birren B, Green ED, Klapholz S, Myers RM, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 1 Analyzing DNA*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1997, p 301.
167. Ju J, Ruan C, Fuller CW, Glazer AN, Mathies RA: Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis. *Proc Natl Acad Sci USA* **92**:4347, 1995.
168. Ju J, Glazer AN, Mathies RA: Energy transfer primers: A new fluorescence labeling paradigm for DNA sequencing and analysis. *Nat Med* **2**:246, 1996.
169. Glazer AN, Mathies RA: Energy-transfer fluorescent reagents for DNA analyses. *Curr Opin Biotechnol* **8**:94, 1997.
170. Rosenblum BB, Lee LG, Spurgeon SL, Khan SH, Menchen SM, Heiner CR, Chen SM: New dye-labeled terminators for improved DNA sequencing patterns. *Nucleic Acids Res* **25**:4500, 1997.
171. Heiner CR, Hunkapiller KL, Chen S-M, Glass JI, Chen EY: Sequencing multimegabase-template DNA with BigDye terminator chemistry. *Genome Res* **8**:557, 1998.
172. Johnston M: Gene chips: Array of hope for understanding gene regulation. *Curr Biol* **8**:R171, 1998.
173. Ramsay G: DNA chips: State-of-the art. *Nat Biotechnol* **16**:40, 1998.
174. Marshall A, Hodgson J: DNA chips: An array of possibilities. *Nat Biotechnol* **16**:27, 1998.

175. Castellino AM: When the chips are down. *Genome Res* **7**:943, 1997.
176. Strachan T, Abitbol M, Davidson D, Beckmann JS: A new dimension for the human genome project: Towards comprehensive expression maps. *Nat Genet* **16**:126, 1997.
177. Lillie J: Probing the genome for new drugs and targets with DNA arrays. *Drug Development Research* **41**:160, 1997.
177a. Lander ES: Array of hope. *Nat Genet* **21** (**Suppl.**):3, 1999.
177b. Southern E, Mir K, Shchepinov M: Molecular interactions on microarrays. *Nat Genet* **21** (**Suppl.**):5, 1999.
177c. Debouck C, Goodfellow PN: DNA microarrays in drug discovery and development. *Nat Genet* **21** (**Suppl.**):48, 1999.
177d. Brown PO, Botstein D: Exploring the new world of the genome with DNA microarrays. *Nat Genet* **21** (**Suppl.**):33, 1999.
178. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D: Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**:767, 1991.
179. Fodor SPA: Massively parallel genomics. *Science* **277**:393, 1997.
180. Southern EM: DNA chips: Analysing sequence by hybridization to oligonucleotides on a large scale. *Trends Genet* **12**:110, 1996.
180a. Lipshutz RJ, Fodor SPA, Gingeras TR, Lockhart DJ: High density synthetic oligonucleotide arrays. *Nat Genet* **21** (**Suppl.**):20, 1999.
181. Kozal MJ, Shah N, Shen N, Yang R, Fucini R, Merigan TC, Richman DD, et al.: Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nat Med* **2**:753, 1996.
182. Gingeras TR, Ghandour G, Wang E, Berno A, Small PM, Drobniewski F, Alland D, et al.: Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic *Mycobacterium* DNA arrays. *Genome Res* **8**:435, 1998.
183. Sapolsky RJ, Lipshutz RJ: Mapping genomic library clones using oligonucleotide arrays. *Genomics* **33**:445, 1996.
184. Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, et al.: Accessing genetic information with high-density DNA arrays. *Science* **274**:610, 1996.
185. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS: Detection of heterozygous mutations in *BRCA1* using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat Genet* **14**:441, 1996.
185a. Hacia JG: Resequencing and mutational analysis using oligonucleotide microarrays. *Nat Genet* **21** (**Suppl.**):42, 1999.
186. Hacia JG, Makalowski W, Edgemon K, Erdos MR, Robbins CM, Fodor SPA, Brody LC, et al.: Evolutionary sequence comparisons using high-density oligonucleotide arrays. *Nat Genet* **18**:155, 1998.
187. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, et al.: Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**:1675, 1996.
188. Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ: Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat Biotechnol* **15**:1359, 1997.
188a. Holstege FCP, Jennings EG, Wyrick JJ, Ihn Lee T, Hengartner CJ, Green MR, Golub TR, et al.: Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**:717, 1998.
189. Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, et al.: Direct allelic variation scanning the yeast genome. *Science* **281**:1194, 1998.
189a. Chakravarti A: Population genetics— making sense out of sequence. *Nat Genet* **21** (**Suppl.**):56, 1999.
190. Wang DG, Fan J-B, Siao C-J, Berno A, Young P, Sapolsky R, Ghandour G, et al.: Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**:1077, 1998.
191. Schena M, Shalon D, Davis RW, Brown PO: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**:467, 1995.
192. Schena M: Genome analysis with gene expression microarrays. *BioEssays* **18**:427, 1996.
193. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW: Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* **93**:10614, 1996.
194. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, et al.: Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* **14**:457, 1996.
194a. Cole KA, Krizman DB, Emmert-Buck MR: The genetics of cancer— a 3D model. *Nat Genet* **21** (**Suppl.**):38, 1999.
195. DeRisi JL, Iyer VR, Brown PO: Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**:680, 1997.
196. Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, et al.: Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci USA* **94**:13057, 1997.
196a. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: Expression profiling using cDNA microarrays. *Nat Genet* **21** (**Suppl.**):10, 1999.
196b. Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G: Making and reading microarrays. *Nat Genet* **21** (**Suppl.**):15, 1999.
196c. Bowtell DDL: Options available— from start to finish— for obtaining expression data by microarray. *Nat Genet* **21** (**Suppl.**):25, 1999.
197. Ermolaeva O, Rastogi M, Pruitt KD, Schuler GD, Bittner ML, Chen Y, Simon R, et al.: Data management and analysis for gene expression arrays. *Nat Genet* **20**:19, 1998.
197a. Bassett Jr DE, Eisen MB, Boguski MS: Gene expression informatics— it's all in your mine. *Nat Genet* **21** (**Suppl.**):51, 1999.
198. Woolley AT, Mathies RA: Ultra-high-speed DNA fragment separations using microfabricated capillary array electrophoresis chips. *Proc Natl Acad Sci USA* **91**:11348, 1994.
199. Woolley AT, Hadley D, Landre P, deMello AJ, Mathies RA, Northrup MA: Functional integration of PCR amplification and capillary electrophoresis in a microfabricated DNA analysis device. *Anal Chem* **68**:4081, 1996.
200. Woolley AT, Sensabaugh GF, Mathies RA: High-speed DNA genotyping using microfabricated capillary array electrophoresis chips. *Anal Chem* **69**:2181, 1997.
201. Burns MA, Mastrangelo CH, Sammarco TS, Man FP, Webster JR, Johnson BN, Foerster B, et al.: Microfabricated structures for integrated DNA analysis. *Proc Natl Acad Sci USA* **93**:5556, 1996.
202. Burke DT, Burns MA, Mastrangelo C: Microfabrication technologies for integrated nucleic acid analysis. *Genome Res* **7**:189, 1997.
203. Simpson PC, Roach D, Woolley AT, Thorsen T, Johnston R, Sensabaugh GF, Mathies RA: High-throughput genetic analysis using microfabricated 96-sample capillary array electrophoresis microplates. *Proc Natl Acad Sci USA* **95**:2256, 1998.
204. Kopp MU, de Mello AJ, Manz A: Chemical amplification: Continuous-flow PCR on a chip. *Science* **280**:1046, 1998.
205. Burns MA, Johnson BN, Brahmasandra SN, Handique K, Webster JR, Krishnan M, Sammarco TS, et al.: An integrated nanoliter DNA analysis device. *Science* **282**:484, 1998.
206. *Methods in Enzymology Volume 266: Computer Methods for Macromolecular Sequence Analysis*.: Doolittle, RF (ed). San Diego, Academic Press, Inc., 1996.
207. Fischer C, Schweigert S, Spreckelsen C, Vogel F: Programs, databases, and expert systems for human geneticists - a survey. *Hum Genet* **97**:129, 1996.
208. Baxevanis AD, Boguski MS, Ouellette BFF: Computational analysis of DNA and protein sequences, in Birren B, Green ED, Klapholz S, Myers RM, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 1 Analyzing DNA*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1997, p 533.
209. *Bioinformatics: A practical guide to the analysis of genes and proteins*: Baxevanis, AD and Ouellette, BFF (eds). New York, John Wiley and Sons, 1998.
210. *Guide to Human Genome Computing*: Bishop, M (ed). London, Academic Press, 1998.
211. Ewing B, Hillier L, Wendl MC, Green P: Base-calling of automated sequencer traces using *Phred*. I. accuracy assessment. *Genome Res* **8**:175, 1998.
212. Ewing B, Green P: Base-calling of automated sequencer traces using *Phred*. II. error probabilities. *Genome Res* **8**:186, 1998.
213. Gordon D, Abajian C, Green P: *Consed*: A graphical tool for sequence finishing. *Genome Res* **8**:195, 1998.
214. Bonfield JK, Smith KF, Staden R: The new DNA sequence assembly program. *Nucleic Acids Res* **23**:4992, 1995.
215. Dear S, Durbin R, Hillier L, Marth G, Thierry-Mieg J, Mott R: Sequence assembly with CAFTOOLS. *Genome Res* **8**:260, 1998.
216. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BFF: GenBank. *Nucleic Acids Res* **26**:1, 1998.
217. Stoesser G, Moseley MA, Sleep J, McGowran M, Garcia-Pastor M, Sterk P: The EMBL nucleotide sequence database. *Nucleic Acids Res* **26**:8, 1998.
218. Tateno Y, Fukami-Kobayashi K, Miyazaki S, Sugawara H, Gojobori T: DNA data bank of Japan at work on genome sequence data. *Nucleic Acids Res* **26**:16, 1998.
219. Ouellette BFF, Boguski MS: Database divisions and homology search files: A guide for the perplexed. *Genome Res* **7**:952, 1997.
220. Barker WC, Garavelli JS, Haft DH, Hunt LT, Marzec CR, Orcutt BC, Srinivasarao GY, et al.: The PIR-international protein sequence database. *Nucleic Acids Res* **26**:27, 1998.
221. Bairoch A, Apweiler R: The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Res* **26**:38, 1998.
222. Borsani G, Ballabio A, Banfi S: A practical guide to orient yourself in the labyrinth of genome databases. *Hum Mol Genet* **7**:1641, 1998.
223. Gelbart WM: Databases in genomic research. *Science* **282**:659, 1998.
224. Pearson WR, Lipman DJ: Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85**:2444, 1988.
225. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* **215**:403, 1990.
226. Altschul SF, Boguski MS, Gish W, Wootton JC: Issues in searching molecular sequence databases. *Nat Genet* **6**:119, 1994.
227. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* **25**:3389, 1997.
228. Zhang J, Madden TL: PowerBLAST: A new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res* **7**:649, 1997.

229. Uberbacher EC, Mural RJ: Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc Natl Acad Sci USA* **88**:11261, 1991.
230. Hutchinson GB, Hayden MR: The prediction of exons through an analysis of spliceable open reading frames. *Nucleic Acids Res* **20**:3453, 1992.
231. Xu Y, Mural R, Shah M, Uberbacher E: Recognizing exons in genomic sequence using GRAIL II. *Genet Eng* **16**:241, 1994.
232. Solovyev VV, Salamov AA, Lawrence CB: Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res* **22**:5156, 1994.
233. Burge C, Karlin S: Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**:78, 1997.
234. Claverie J-M: Computational methods for the identification of genes in vertebrate genomic sequences. *Hum Mol Genet* **6**:1735, 1997.
235. Burge CB, Karlin S: Finding the genes in genomic DNA. *Curr Opin Struct Biol* **8**:346, 1998.
236. Fickett JW, Hatzigeorgiou AG: Eukaryotic promoter recognition. *Genome Res* **7**:861, 1997.
237. Smith RF, Wiese BA, Wojzynski MK, Davison DB, Worley KC: BCM search launcher-An integrated interface to molecular biology data base search and analysis services available on the World Wide Web. *Genome Res* **6**:454, 1996.
238. *ISCN: An international system for human cytogenetic nomenclature. Birth Defects: Original Article Series (Volume 21, No. 1)*: Harnden, DG and Klinger, HP (eds). New York, Karger, 1985.
239. Royer-Pokora B, Kunkel LM, Monaco AP, Goff SC, Newburger PE, Baehner RL, Cole FS, et al.: Cloning the gene for an inherited human disorder- chronic granulomatous disease- on the basis of its chromosomal location. *Nature* **322**:32, 1986.
240. Monaco AP, Neve RL, Colletti-Feener C, Bertelson CJ, Kurnit DM, Kunkel LM: Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. *Nature* **323**:646, 1986.
241. Verkerk AJMH, Pieretti M, Sutcliffe JS, Fu Y-H, Kuhl DPA, Pizzuti A, Reiner O, et al.: Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**:905, 1991.
242. Trask BJ, Massa H, Kenwrick S, Gitschier J: Mapping of human chromosome Xq28 by two-color fluorescence in situ hybridization of DNA sequences to interphase cell nuclei. *Am J Hum Genet* **48**:1, 1991.
243. Southern EM: Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* **98**:503, 1975.
244. Botstein D, White RL, Skolnick M, Davis RW: Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**:314, 1980.
245. White R, Leppert M, Bishop DT, Barker D, Berkowitz J, Brown C, Callahan P, et al.: Construction of linkage maps with DNA markers for human chromosomes. *Nature* **313**:101, 1985.
246. Donis-Keller H, Barker DF, Knowlton RG, Schumm JW, Braman JC, Green P: Highly polymorphic RFLP probes as diagnostic tools. *Cold Spring Harb Symp Quant Biol* **LI**:317, 1986.
247. Lander ES, Botstein D: Mapping complex genetic traits in humans: New method using a complete RFLP linkage map. *Cold Spring Harb Symp Quant Biol* **LI**:49, 1986.
248. Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, et al.: Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235**:1616, 1987.
249. Weber JL, May PE: Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet* **44**:388, 1989.
250. Litt M, Luty JA: A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet* **44**:397, 1989.
251. Weber JL: Human DNA polymorphisms based on length variations in simple-sequence tandem repeats, in Davies KE, Tilghman SM (eds.): *Genome Analysis (Volume 1: Genetic and Physical Mapping)*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1990, p 159.
252. Dietrich WF, Weber JL, Nickerson DA, Kwok P-Y: Identification and analysis of DNA polymorphisms, in Birren B, Green ED, Hieter P, Klapholz S, Myers RM, Riethman H, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 4 Mapping Genomes*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1999, p 135.
253. Nickerson DA, Whitehurst C, Boysen C, Charmley P, Kaiser R, Hood L: Identification of clusters of biallelic polymorphic sequence-tagged sites (pSTSs) that generate highly informative and automatable markers for genetic linkage mapping. *Genomics* **12**:377, 1992.
254. Landegren U, Nilsson M, Kwok P-Y: Reading bits of genetic information: Methods for single-nucleotide polymorphism analysis. *Genome Res* **8**:769, 1998.
255. Kruglyak L: The use of a genetic map of biallelic markers in linkage studies. *Nat Genet* **17**:21, 1997.
256. Collins FS, Guyer MS, Chakravarti A: Variations on a theme: Cataloging human DNA sequence variation. *Science* **278**:1580, 1997.
257. Schafer AJ, Hawkins JR: DNA variation and the future of human genetics. *Nat Biotechnol* **16**:33, 1998.
258. Weiss KM: In search of human variation. *Genome Res* **8**:691, 1998.
259. Chakravarti A, Lynn A: Meiotic mapping in humans, in Birren B, Green ED, Hieter P, Klapholz S, Myers RM, Riethman H, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 4 Mapping Genomes*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1999, p 1.
260. Dausset J, Cann H, Cohen D, Lathrop M, Lalouel J-M, White R: Centre d'Etude du Polymorphisme Humain (CEPH): Collaborative genetic mapping of the human genome. *Genomics* **6**:575, 1990.
261. Terwilliger JD, Ott J: *Handbook of Human Genetic Linkage,* Baltimore, Johns Hopkins University Press, 1994.
262. Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, et al.: A genetic linkage map of the human genome. *Cell* **51**:319, 1987.
263. Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, et al.: A second-generation linkage map of the human genome. *Nature* **359**:794, 1992.
264. Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, Millasseau P, Marc S, et al.: The 1993-94 Genethon human genetic linkage map. *Nat Genet* **7**:246, 1994.
265. Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, et al.: A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**:152, 1996.
266. Murray JC, Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Heddema T, Manion F, Quillen J, et al.: A comprehensive human linkage map with centimorgan density. *Science* **265**:2049, 1994.
267. Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Heddema T, Duyk GM, Sheffield VC, Wang Z, et al.: Integrated human genome-wide maps constructed using the CEPH reference panel. *Nat Genet* **6**:391, 1994.
268. Utah Marker Development Group: A collection of ordered tetranucleotide-repeat markers from the human genome. *Am J Hum Genet* **57**:619, 1995.
269. Olson MV, Dutchik JE, Graham MY, Brodeur GM, Helms C, Frank M, MacCollin M, et al.: Random-clone strategy for genomic restriction mapping in yeast. *Proc Natl Acad Sci USA* **83**:7826, 1986.
270. Coulson A, Sulston J, Brenner S, Karn J: Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc Natl Acad Sci USA* **83**:7821, 1986.
271. Kohara Y, Akiyama K, Isono K: The physical map of the whole E. coli chromosome: Application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50**:495, 1987.
272. Wong GK-S, Yu J, Thayer EC, Olson MV: Multiple-complete-digest restriction fragment mapping: Generating sequence-ready maps for large-scale DNA sequencing. *Proc Natl Acad Sci USA* **94**:5225, 1997.
273. Olson M, Hood L, Cantor C, Botstein D: A common language for physical mapping of the human genome. *Science* **245**:1434, 1989.
274. Green ED, Green P: Sequence-tagged site (STS) content mapping of human chromosomes: theoretical considerations and early experiences. *PCR Methods Appl* **1**:77, 1991.
275. Green ED, Mohr RM, Idol JR, Jones M, Buckingham JM, Deaven LL, Moyzis RK, et al.: Systematic generation of sequence-tagged sites for physical mapping of human chromosomes: Application to the mapping of human chromosome 7 using yeast artificial chromosomes. *Genomics* **11**:548, 1991.
276. Hillier L, Green P: OSP: A computer program for choosing PCR and DNA sequencing primers. *PCR Methods Appl* **1**:124, 1991.
277. Green ED: Physical mapping of human chromosomes: Generation of chromosome-specific sequence-tagged sites, in Adolph KW (ed.): *Methods in Molecular Genetics. Vol. 1 Gene and Chromosome Analysis (Part A)*, San Diego, Academic Press, 1993, p 192.
278. Vollrath D: DNA markers for physical mapping, in Birren B, Green ED, Hieter P, Klapholz S, Myers RM, Riethman H, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 4 Mapping genomes*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1999, p 187.
279. Nelson DL, Ledbetter SA, Corbo L, Victoria MF, Ramirez-Solis R, Webster TD, Ledbetter DH, et al.: *Alu* polymerase chain reaction: A method for rapid isolation of human-specific sequences from complex DNA sources. *Proc Natl Acad Sci USA* **86**:6686, 1989.
280. Nelson DL: Interspersed repetitive sequence polymerase chain reaction (IRS PCR) for generation of human DNA fragments from complex sources. *Methods* **2**:60, 1991.
281. Kass DH, Batzer MA: Inter-*Alu* polymerase chain reaction: Advancements and applications. *Anal Biochem* **228**:185, 1995.
282. Cole CG, Goodfellow PN, Bobrow M, Bentley DR: Generation of novel sequence tagged sites (STSs) from discrete chromosomal regions using *Alu*-PCR. *Genomics* **10**:816, 1991.
283. Deaven LL, Van Dilla MA, Bartholdi MF, Carrano AV, Cram LS, Fuscoe JC, Gray JW, et al.: Construction of human chromosome-specific DNA libraries from flow-sorted chromosomes. *Cold Spring Harb Symp Quant Biol* **LI**:159, 1986.
284. Collins C, Kuo WL, Segraves R, Fuscoe J, Pinkel D, Gray JW: Construction and characterization of plasmid libraries enriched in sequences from single human chromosomes. *Genomics* **11**:997, 1991.

285. Shimizu N, Minoshima S: Gene mapping and fine-structure analysis of the human genome using flow-sorted chromosomes, in Adolph KW (ed.): *Advanced Techniques in Chromosome Research*, New York, Marcel Dekker, Inc., 1991, p 135.
286. Goold RD, diSibio GL, Xu H, Lang DB, Dadgar J, Magrane GG, Dugaiczyk A, et al.: The development of sequence-tagged sites for human chromosome 4. *Hum Mol Genet* **2**:1271, 1993.
287. Smith MW, Clark SP, Hutchinson JS, Wei YH, Churukian AC, Daniels LB, Diggle KL, et al.: A sequence-tagged site map of human chromosome 11. *Genomics* **17**:699, 1993.
288. Wilcox AS, Khan AS, Hopkins JA, Sikela JM: Use of 3¢ untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: Implications for an expression map of the genome. *Nucleic Acids Res* **19**:1837, 1991.
289. Kere J, Nagaraja R, Mumm S, Ciccodicola A, D'Urso M, Schlessinger D: Mapping human chromosomes by walking with sequence-tagged sites from end fragments of yeast artificial chromosome inserts. *Genomics* **14**:241, 1992.
290. Bouffard GG, Iyer LM, Idol JR, Braden VV, Cunningham AF, Weintraub LA, Mohr-Tidwell RM, et al.: A collection of 1814 human chromosome 7-specific STSs. *Genome Res* **7**:59, 1997.
291. Stewart EA, McKusick KB, Aggarwal A, Bajorek E, Brady S, Chu A, Fang N, et al.: An STS-based radiation hybrid map of the human genome. *Genome Res* **7**:422, 1997.
292. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, et al.: A gene map of the human genome. *Science* **274**:540, 1996.
293. Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, Hui L, et al.: A physical map of 30,000 human genes. *Science* **282**:744, 1998.
294. Sheffield VC, Weber JL, Buetow KH, Murray JC, Even DA, Wiles K, Gastier JM, et al.: A collection of tri- and tetranucleotide repeat markers used to generate high quality, high resolution human genome-wide linkage maps. *Hum Mol Genet* **4**:1837, 1995.
295. Walter MA, Spillett DJ, Thomas P, Weissenbach J, Goodfellow PN: A method for constructing radiation hybrid maps of whole genomes. *Nat Genet* **7**:22, 1994.
296. McCarthy LC: Whole genome radiation hybrid mapping. *Trends Genet* **12**:491, 1996.
297. Lunetta KL, Boehnke M: Multipoint radiation hybrid mapping: Comparison of methods, sample size requirements, and optimal study characteristics. *Genomics* **21**:92, 1994.
298. Lange K, Boehnke M, Cox DR, Lunetta KL: Statistical methods for polyploid radiation hybrid mapping. *Genome Res* **5**:136, 1995.
299. Lunetta KL, Boehnke M, Lange K, Cox DR: Experimental design and error detection for polyploid radiation hybrid mapping. *Genome Res* **5**:151, 1995.
300. Gyapay G, Schmitt K, Fizames C, Jones H, Vega-Czarny N, Spillett D, Muselet D, et al.: A radiation hybrid map of the human genome. *Hum Mol Genet* **5**:339, 1996.
301. McCarthy LC, Terrett J, Davis ME, Knights CJ, Smith AL, Critcher R, Schmitt K, et al.: A first-generation whole genome-radiation hybrid map spanning the mouse genome. *Genome Res* **7**:1153, 1997.
302. Yang Y-P, Womack JE: Parallel radiation hybrid mapping: A powerful tool for high-resolution genomic comparison. *Genome Res* **8**:731, 1998.
303. Travis GH, Sutcliffe JG: Phenol emulsion-enhanced DNA-driven subtractive cDNA cloning: Isolation of low-abundance monkey cortex-specific mRNAs. *Proc Natl Acad Sci USA* **85**:1696, 1988.
304. Patanjali SR, Parimoo S, Weissman SM: Construction of a uniform-abundance (normalized) cDNA library. *Proc Natl Acad Sci USA* **88**:1943, 1991.
305. Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, Efstratiadis A: Construction and characterization of a normalized cDNA library. *Proc Natl Acad Sci USA* **91**:9228, 1994.
306. Bonaldo MF, Lennon G, Soares MB: Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res* **6**:791, 1996.
307. Soares MB, Bonaldo MF: Constructing and screening normalized cDNA libraries, in Birren B, Green ED, Klapholz S, Myers RM, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 2 Detecting Genes*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1998, p 49.
308. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, et al.: Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252**:1651, 1991.
309. Hillier L, Lennon G, Becker M, Bonaldo M, Chiapelli B, Chissoe S, Dietrich N, et al.: Generation and analysis of 280,000 human expressed sequence tags. *Genome Res* **6**:807, 1996.
310. Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, et al.: Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377**:3, 1995.
311. Boguski MS, Lowe TM, Tolstoshev CM: dbEST-Database for "expressed sequence tags." *Nat Genet* **4**:332, 1993.
312. Boguski MS, Schuler GD: ESTablishing a human transcript map. *Nat Genet* **10**:369, 1995.
313. Boguski MS: The turning point in genome research. *Trends Biochem Sci* **20**:295, 1995.
314. Schuler GD: Pieces of the puzzle: Expressed sequence tags and the catalog of human genes. *J Mol Med* **75**:694, 1997.
315. Khan AS, Wilcox AS, Polymeropoulos MH, Hopkins JA, Stevens TJ, Robinson M, Orpana AK, et al.: Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nat Genet* **2**:180, 1992.
316. Berry R, Stevens TJ, Walter NAR, Wilcox AS, Rubano T, Hopkins JA, Weber J, et al.: Gene-based sequence-tagged-sites (STSs) as the basis for a human gene map. *Nat Genet* **10**:415, 1995.
317. Gerhold D, Caskey CT: It's the genes! EST access to human genome content. *BioEssays* **18**:973, 1996.
318. Okubo K, Matsubara K: Complementary DNA sequence (EST) collections and the expression information of the human genome. *FEBS Lett* **403**:225, 1997.
319. Tilghman SM: Lessons learned, promises kept: A biologist's eye view of the genome project. *Genome Res* **6**:773, 1996.
320. Marra MA, Hillier L, Waterston RH: Expressed sequence tags-ESTablishing bridges between genomes. *Trends Genet* **14**:4, 1998.
321. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: Serial analysis of gene expression. *Science* **270**:484, 1995.
322. Adams MD: Serial analysis of gene expression: ESTs get smaller. *BioEssays* **18**:261, 1996.
323. The Yeast Genome Directory. *Nature* **387S**:1, 1997.
324. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, et al.: Life with 6000 genes. *Science* **274**:546, 1996.
325. Hieter P, Bassett Jr. DE, Valle D: The yeast genome-A common currency. *Nat Genet* **13**:253, 1996.
326. Johnston M: Genome sequencing: The complete code for a eukaryotic cell. *Curr Biol* **6**:500, 1996.
327. Dujon B: The yeast genome project: What did we learn? *Trends Genet* **12**:263, 1996.
328. Walsh S, Barrell B: The *Saccharomyces cerevisiae* genome on the World Wide Web. *Trends Genet* **12**:276, 1996.
329. Thomas K: Yeasties and beasties: 7 years of genome sequencing. *FEBS Lett* **396**:1, 1996.
330. Blattner FR, Plunkett III G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, et al.: The complete genome sequence of *Escherichia coli* K-12. *Science* **277**:1453, 1997.
331. Sulston J, Du Z, Thomas K, Wilson R, Hillier L, Staden R, Halloran N, et al.: The *C. elegans* genome sequencing project: A beginning. *Nature* **356**:37, 1992.
332. Wilson R, Ainscough R, Anderson K, Baynes C, Berks M, Bonfield J, Burton J, et al.: 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature* **368**:32, 1994.
333. Hodgkin J, Plasterk RHA, Waterston RH: The nematode *Caenorhabditis elegans* and its genome. *Science* **270**:410, 1995.
334. Waterston R, Sulston J: The genome of *Caenorhabditis elegans*. *Proc Natl Acad Sci USA* **92**:10836, 1995.
335. Berks M and the *C. elegans* Genome Mapping and Sequencing Consortium: Around the genomes: The *C. elegans* genome sequencing project. *Genome Res* **5**:99, 1995.
335a. The *C. elegans* Sequencing Consortium: Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**:2012, 1998.
336. Olson MV: A time to sequence. *Science* **270**:394, 1995.
337. Gibbs RA: Pressing ahead with human genome sequencing. *Nat Genet* **11**:121, 1995.
338. Boguski M, Chakravarti A, Gibbs R, Green E, Myers RM: The end of the beginning: The race to begin human genome sequencing. *Genome Res* **6**:771, 1996.
339. Gibbs RA: Hares and tortoises in the race to sequence the human genome: Expectations and realities. *Trends Genet* **13**:381, 1997.
340. Rowen L, Mahairas G, Hood L: Sequencing the human genome. *Science* **278**:605, 1997.
341. Beck S, Sterk P: Genome-scale DNA sequencing: Where are we? *Curr Opin Biotechnol* **9**:116, 1998.
342. Waterston R, Sulston JE: The Human Genome Project: Reaching the finish line. *Science* **282**:53, 1998.
342a. The Sanger Centre, The Washington University Genome Sequencing Center: Toward a complete human genome sequence. *Genome Res* **8**:1097, 1998.
343. Wilson RK, Mardis ER: Shotgun sequencing, in Birren B, Green ED, Klapholz S, Myers RM, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 1 Analyzing DNA*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1997, p 397.
344. Kimmel BE, Palazzolo MJ, Martin CH, Boeke JD, Devine SE: Transposon-mediated DNA sequencing, in Birren B, Green ED, Klapholz S, Myers RM, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 1 Analyzing DNA*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1997, p 455.
345. Miklos GLG, Rubin GM: The role of the genome project in determining gene function: Insights from model organisms. *Cell* **86**:521, 1996.

346. Riles L, Dutchik JE, Baktha A, McCauley BK, Thayer EC, Leckie MP, Braden VV, et al.: Physical maps of the six smallest chromosomes of *Saccharomyces cerevisiae* at a resolution of 2.6 kilobase pairs. *Genetics* **134**:81, 1993.
347. Bassett Jr. DE, Basrai MA, Connelly C, Hyland KM, Kitagawa K, Mayer ML, Morrow DM, et al.: Exploiting the complete yeast genome sequence. *Curr Opin Genet Dev* **6**:763, 1996.
348. Botstein D, Chervitz SA, Cherry JM: Yeast as a model organism. *Science* **277**:1259, 1997.
349. Hudson Jr. JR, Dawson EP, Rushing KL, Jackson CH, Lockshon D, Conover D, Lanciault C, et al.: The complete set of predicted genes from *Saccharomyces cerevisiae* in a readily usable form. *Genome Res* **7**:1169, 1997.
350. Coulson A, Waterston R, Kiff J, Sulston J, Kohara Y: Genome linking with yeast artificial chromosomes. *Nature* **335**:184, 1988.
351. Coulson A, Kozono Y, Lutterbach B, Shownkeen R, Sulston J, Waterston R: YACs and the *C. elegans* genome. *BioEssays* **13**:413, 1991.
352. Hodgkin J, Herman RK: Changing styles in *C. elegans* genetics. *Trends Genet* **14**:352, 1998.
353. Walhout M, Endoh H, Thierry-Mieg N, Wong W, Vidal M: A model of elegance. *Am J Hum Genet* **63**:955, 1998.
354. Rubin GM: Around the genomes: The *Drosophila* genome project. *Genome Res* **6**:71, 1996.
355. Rubin GM: The *Drosophila* genome project: A progress report. *Trends Genet* **14**:340, 1998.
356. Hartl DL: Genome map of *Drosophila melanogaster* based on yeast artificial chromosomes, in Davies KE, Tilghman SM (eds.): *Genome Analysis (Volume 4: Strategies for Physical Mapping)*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1992, p 39.
357. Kimmerly W, Stultz K, Lewis S, Lewis K, Lustre V, Romero R, Benke J, et al.: A P1-based physical map of the *Drosophila* euchromatic genome. *Genome Res* **6**:414, 1996.
358. Spradling AC, Stern DM, Kiss I, Roote J, Laverty T, Rubin GM: Gene disruptions using *P* transposable elements: An integral component of the Drosophila genome project. *Proc Natl Acad Sci USA* **92**:10824, 1995.
359. Dietrich WF, Copeland NG, Gilbert DJ, Miller JC, Jenkins NA, Lander ES: Mapping the mouse genome: Current status and future prospects. *Proc Natl Acad Sci USA* **92**:10849, 1995.
360. Meisler MH: The role of the laboratory mouse in the human genome project. *Am J Hum Genet* **59**:764, 1996.
361. Fisher EM: The contribution of the mouse to advances in human genetics. *Advances in Genetics* **35**:155, 1997.
362. Jaenisch R: Transgenic animals. *Science* **240**:1468, 1988.
363. Capecchi MR: Altering the genome by homologous recombination. *Science* **244**:1288, 1989.
364. Robbins J: Gene targeting. The precise manipulation of the mammalian genome. *Circ Res* **73**:3, 1993.
365. Wynshaw-Boris A: Model mice and human disease. *Nat Genet* **13**:259, 1996.
366. Thomas KR, Capecchi MR: Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. *Cell* **51**:503, 1987.
367. Brown SDM, Nolan PM: Mouse mutagenesis-Systematic studies of mammalian gene function. *Hum Mol Genet* **7**:1627, 1998.
368. Rubin EM, Barsh GS: Biological insights through genomics: Mouse to man. *J Clin Invest* **97**:275, 1996.
369. Rubin EM, Smith DJ: Optimizing the mouse to sift sequence for function. *Trends Genet* **13**:423, 1997.
370. Nadeau JH: Maps of linkage and synteny homologies between mouse and man. *Trends Genet* **5**:82, 1989.
371. O'Brien SJ: Mammalian genome mapping: Lessons and prospects. *Curr Opin Genet Develop* **1**:105, 1990.
372. Cox RD, Lehrach H: Genome mapping: PCR based meiotic and somatic cell hybrid analysis. *BioEssays* **13**:193, 1991.
373. O'Brien SJ, Womack JE, Lyons LA, Moore KJ, Jenkins NA, Copeland NG: Anchored reference loci for comparative genome mapping in mammals. *Nat Genet* **3**:103, 1993.
374. Eppig JT, Nadeau JH: Comparative maps: The mammalian jigsaw puzzle. *Curr Opin Genet Dev* **5**:709, 1995.
375. DeBry RW, Seldin MF: Human/mouse homology relationships. *Genomics* **33**:337, 1996.
376. Carver EA, Stubbs L: Zooming in on the human-mouse comparative map: Genome conservation re-examined on a high-resolution scale. *Genome Res* **7**:1123, 1997.
376a. Nadeau JH, Sankoff D: Counting on comparative maps. *Trends Genet* **14**:495, 1998.
377. Dietrich W, Katz H, Lincoln SE, Shin H-S, Friedman J, Dracopoli NC, Lander ES: A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics* **131**:423, 1992.
378. Dietrich WF, Miller JC, Steen RG, Merchant M, Damron D, Nahf R, Gross A, et al.: A genetic map of the mouse with 4,006 simple sequence length polymorphisms. *Nat Genet* **7**:220, 1994.
379. Dietrich WF, Miller J, Steen R, Merchant MA, Damron-Boles D, Husain Z, Dredge R, et al.: A comprehensive genetic map of the mouse genome. *Nature* **380**:149, 1996.
380. Rhodes M, Straw R, Fernando S, Evans A, Lacey T, Dearlove A, Greystrong J, et al.: A high-resolution microsatellite map of the mouse genome. *Genome Res* **8**:531, 1998.
381. Herman GE: Physical mapping of the mouse genome. *Methods* **14**:135, 1998.
382. Eppig JT, Blake JA, Davisson MT, Richardson JE: Informatics for mouse genetics and genome mapping. *Methods* **14**:179, 1998.
383. Seldin MF: Genome surfing: Using internet-based informatic tools toward functional genetic studies in mouse and humans. *Methods* **13**:445, 1997.
384. Hood L, Koop B, Goverman J, Hunkapiller T: Model genomes: The benefits of analysing homologous human and mouse sequences. *Trends Biotechnol* **10**:19, 1992.
385. Koop BF: Human and rodent DNA sequence comparisons: A mosaic model of genomic evolution. *Trends Genet* **11**:367, 1995.
386. Mallon A-M, Strivens M: DNA sequence analysis and comparative sequencing. *Methods* **14**:160, 1998.
387. Eisen JA: Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* **8**:163, 1998.
388. Makalowski W, Zhang J, Boguski MS: Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res* **6**:846, 1996.
389. Makalowski W, Boguski MS: Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc Natl Acad Sci USA* **95**:9407, 1998.
390. Koop BF, Hood L: Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat Genet* **7**:48, 1994.
391. Galili N, Baldwin HS, Lund J, Reeves R, Gong W, Wang Z, Roe BA, et al.: A region of mouse chromosome 16 is syntenic to the DiGeorge, velocardiofacial syndrome minimal critical region. *Genome Res* **7**:17, 1997.
392. Oeltjen JC, Malley TM, Muzny DM, Miller W, Gibbs RA, Belmont JW: Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res* **7**:315, 1997.
393. Ansari-Lari MA, Oeltjen JC, Schwartz S, Zhang Z, Muzny DM, Lu J, Gorrell JH, et al.: Comparative sequence analysis of a gene-rich cluster on human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res* **8**:29, 1998.
394. Mizukami T, Chang WI, Garkavtsev I, Kaplan N, Lombardi D, Matsumoto T, Niwa O, et al.: A 13 kb resolution cosmid map of the 14 Mb fission yeast genome by nonrandom sequence-tagged site mapping. *Cell* **73**:121, 1993.
395. Hoheisel JD, Maier E, Mott R, McCarthy L, Grigoriev AV, Schalkwyk LC, Nizetic D, et al.: High resolution cosmid and P1 maps spanning the 14 Mb genome of the fission yeast S. pombe. *Cell* **73**:109, 1993.
396. Jacob HJ, Brown DM, Bunker RK, Daly MJ, Dzau VJ, Goodman A, Koike G, et al.: A genetic linkage map of the laboratory rat, *Rattus norvegicus*. *Nat Genet* **9**:63, 1995.
397. James MR, Lindpaintner K: Why map the rat? *Trends Genet* **13**:171, 1997.
398. Brown DM, Matise TC, Koike G, Simon JS, Winer ES, Zangen S, McLaughlin MG, et al.: An integrated genetic linkage map of the laboratory rat. *Mamm Genome* **9**:521, 1998.
399. Felsenfeld AL: Defining the boundaries of zebrafish developmental genetics. *Nat Genet* **14**:258, 1996.
400. Postlethwait JH, Talbot WS: Zebrafish genomics: From mutants to genes. *Trends Genet* **13**:183, 1997.
401. Knapik EW, Goodman A, Ekker M, Chevrette M, Delgado J, Neuhauss S, Shimoda N, et al.: A microsatellite genetic linkage map for zebrafish (*Danio rerio*). *Nat Genet* **18**:338, 1998.
402. Postlethwait JH, Yan Y-L, Gates MA, Horne S, Amores A, Brownlie A, Donovan A, et al.: Vertebrate genome evolution and the zebrafish gene map. *Nat Genet* **18**:345, 1998.
403. Beier DR: Zebrafish: Genomics on the fast track. *Genome Res* **8**:9, 1998.
404. Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S: Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* **366**:265, 1993.
405. Mileham P, Brown SDM: The pufferfish genome: Small is beautiful? *BioEssays* **16**:153, 1994.
406. Koop BF, Nadeau JH: Pufferfish and a new paradigm for comparative genome analysis. *Proc Natl Acad Sci USA* **93**:1363, 1996.
407. Elgar G: Quality not quantity: The pufferfish genome. *Hum Mol Genet* **5**:1437, 1996.
408. Elgar G, Sandford R, Aparicio S, Macrae A, Venkatesh B, Brenner S: Small is beautiful: comparative genomics with the pufferfish (*Fugu rubripes*). *Trends Genet* **12**:145, 1996.
409. Angrist M: Less is more: Compact genomes pay dividends. *Genome Res* **8**:683, 1998.
410. Dean C, Schmidt R: Plant genomes: A current molecular description. *Annu Rev Plant Physiol Plant Mol Biol* **46**:395, 1995.
411. Goodman HM, Ecker JR, Dean C: The genome of *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* **92**:10831, 1995.

412. Settles AM, Byrne M: Opportunities and challenges grow from *Arabidopsis* genome sequencing. *Genome Res* **8**:83, 1998.
413. Schmidt R: Physical mapping of the *Arabidopsis thaliana* genome. *Plant Physiol Biochem* **36**:1, 1998.
414. Bevan M, Bancroft I, Bent E, Love K, Goodman H, Dean C, Bergkamp R, et al.: Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. *Nature* **391**:485, 1998.
415. Watson JD, Jordon E: The human genome program at the National Institutes of Health. *Genomics* **5**:654, 1989.
416. Watson JD: The human genome project: Past, present, and future. *Science* **248**:44, 1990.
417. Cantor CR: Orchestrating the human genome project. *Science* **248**:49, 1990.
418. Cook-Deegan RM: The genesis of the human genome project, in Friedmann T (ed.): *Molecular Genetic Medicine (Volume 1)*, San Diego, Academic Press, Inc., 1991, p 1.
419. Watson JD, Cook-Deegan RM: Origins of the human genome project. *FASEB J* **5**:8, 1991.
420. Jordan E: Organization and long-range plan. *Anal Chem* **63**:420A, 1991.
421. Jordan E: Invited editorial: The human genome project: Where did it come from, where is it going? *Am J Hum Genet* **51**:1, 1992.
422. Engel LW: The human genome project: History, goals, and progress to date. *Arch Pathol Lab Med* **117**:459, 1993.
423. Haq MM: Medical genetics and the human genome project: A historical review. *Tex Med* **89**:68, 1993.
424. McKusick VA: Genomics: Structural and functional studies of genomes. *Genomics* **45**:244, 1997.
424a. Burris J, Cook-Deegan R, Alberts B: The Human Genome Project after a decade: Policy issues. *Nat Genet* **20**:333, 1998.
425. DeLisi C: The human genome project. *Am Sci* **76**:488, 1988.
426. National Research Council (Committee on Mapping and Sequencing the Human Genome): *Mapping and Sequencing the Human Genome,* Washington, D.C., National Academy Press, 1988.
427. Office of Technology Assessment: *Mapping Our Genes-Genome Projects: How Big? How Fast?,* Washington, D.C., 61-OTA-BA-373, 1988.
428. Watson JD: The human genome initiative: A statement of need. *Hosp Pract* **26**:69, 1991.
429. Gilbert W: Towards a paradigm shift in biology. *Nature* **349**:99, 1991.
430. Berg P: All our collective ingenuity will be needed. *FASEB J* **5**:75, 1991.
431. Yager TD, Nickerson DA, Hood LE: The human genome project: Creating an infrastructure for biology and medicine. *Trends Biochem Sci* **16**:454, 1991.
432. Martin RG: We gnomes find the project an atlas but no treasure. *New Biol* **2**:385, 1990.
433. Rechsteiner MC: The human genome project: misguided science policy. *Trends Biochem Sci* **16**:455, 1991.
434. Davis BD, Colleagues: The human genome and other initiatives. *Science* **249**:342, 1990.
435. Tauber AI, Sarkar S: The human genome project: Has blind reductionism gone too far? *Perspect Biol Med* **35**:221, 1992.
436. Rosenberg LE: The human genome project. *Bull N Y Acad Med* **68**:113, 1992.
437. Richardson WC: Summary, conference on the human genome project: An agenda for science & society. *Bull N Y Acad Med* **68**:162, 1992.
438. Koshland DE: Sequences and consequences of the human genome. *Science* **246**:189, 1989.
439. U.S. Department of Health and Human Services, U.S. Department of Energy: *Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years, FY 1991-1995 (DOE/ER-0452P),* Springfield, National Technical Information Service, 1990.
440. Collins F, Galas D: A new five-year plan for the U.S. human genome project. *Science* **262**:43, 1993.
441. Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L, Members of the DOE and NIH Planning Groups: New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282**:682, 1998.
442. Lander ES: The new genomics: Global views of biology. *Science* **274**:536, 1996.
443. Hieter P, Boguski M: Functional genomics: It's all how you read it. *Science* **278**:601, 1997.
444. Olson MV: A tale of two cities. *Anal Chem* **63**:416A, 1991.
445. Collins FS: Ahead of schedule and under budget: The Genome Project passes its fifth birthday. *Proc Natl Acad Sci USA* **92**:10821, 1995.
446. Guyer MS, Collins FS: How is the Human Genome Project doing, and what have we learned so far? *Proc Natl Acad Sci USA* **92**:10841, 1995.
447. Uddhav K, Ketan S: Advances in the human genome project. *Molecular Biology Reports* **25**:27, 1998.
448. Koonin EV, Mushegian AR, Rudd KE: Sequencing and analysis of bacterial genomes. *Curr Biol* **6**:404, 1996.
449. Smith DR: Microbial pathogen genomes-New strategies for identifying therapeutics and vaccine targets. *Trends Biotechnol* **14**:290, 1996.
450. Koonin EV: Big time for small genomes. *Genome Res* **7**:418, 1997.
451. Fraser CM, Fleischmann RD: Strategies for whole microbial genome sequencing and analysis. *Electrophoresis* **18**:1207, 1997.
452. Koonin EV, Galperin MY: Prokaryotic genomes: The emerging paradigm of genome-based microbiology. *Curr Opin Genet Dev* **7**:757, 1997.
453. Jenks PJ: Sequencing microbial genomes-What will it do for microbiology? *J Med Microbiol* **47**:375, 1998.
454. Doolittle RF: Microbial genomes opened up. *Nature* **392**:339, 1998.
455. Koonin EV, Tatusov RL, Galperin MY: Beyond complete genomes: From sequence to structure and function. *Curr Opin Struct Biol* **8**:355, 1998.
456. Saier Jr. MH: Genome sequencing and informatics: New tools for biochemical discoveries. *Plant Physiol* **117**:1129, 1998.
457. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, et al.: Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**:496, 1995.
458. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, et al.: Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**:537, 1998.
459. Tomb J-F, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, et al.: The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**:539, 1997.
460. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, et al.: Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**:580, 1997.
461. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, et al.: The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**:397, 1995.
462. Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, et al.: Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**:375, 1998.
463. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, et al.: Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**:754, 1998.
463a. Andersson SGE, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UCM, Podowski RM, Naslund AK, et al.: The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**:133, 1998.
463b. Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Aravind L, Koonin EV, Shallom S, et al.: Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**:1126, 1998.
464. Olson MV, Green P: Criterion for the completeness of large-scale physical maps of DNA. *Cold Spring Harb Symp Quant Biol* **LVIII**:349, 1993.
465. Cox DR, Green ED, Lander ES, Cohen D, Myers RM: Assessing mapping progress in the human genome project. *Science* **265**:2031, 1994.
466. Olson M, Green P: A "quality-first" credo for the human genome project. *Genome Res* **8**:414, 1998.
467. Weber JL, Myers EW: Human whole-genome shotgun sequencing. *Genome Res* **7**:401, 1997.
468. Green P: Against a whole-genome shotgun. *Genome Res* **7**:410, 1997.
469. Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M: Shotgun sequencing of the human genome. *Science* **280**:1540, 1998.
470. Bentley DR, Pruitt KD, Deloukas P, Schuler GD, Ostell J: Coordination of human genome sequencing via a consensus framework map. *Trends Genet* **14**:381, 1998.
471. Statement on the Rapid Release of Genomic DNA Sequence. *Genome Res* **8**:413, 1998.
472. Bentley DR: Genomic sequence information should be released immediately and freely in the public domain. *Science* **274**:533, 1996.
473. Adams MD, Venter JC: Should non-peer-reviewed raw DNA sequence data release be forced on the scientific community? *Science* **274**:534, 1996.
474. Green ED, Waterston RH: The human genome project: Prospects and implications for clinical medicine. *JAMA* **266**:1966, 1991.
475. Rossiter BJF, Caskey CT: The human genome project and clinical medicine. *Oncology* **6**:61, 1992.
476. Caskey CT: DNA-based medicine: Prevention and therapy, in Kevles DJ, Hood L (eds.): *The Code of Codes: Scientific and Social Issues in the Human Genome Project*, Cambridge, Harvard University Press, 1992, p 112.
477. Hood L: Biology and medicine in the twenty-first century, in Kevles DL, Hood L (eds.): *The Code of Codes: Scientific and Social Issues in the Human Genome Project*, Cambridge, Harvard University Press, 1992, p 136.
478. Whittaker LA: The implications of the human genome project for family practice. *J Fam Pract* **35**:294, 1992.
479. Guyer MS, Collins FS: The human genome project and the future of medicine. *AJDC* **147**:1145, 1993.
480. Sawicki MP, Samara G, Hurwitz M, Passaro Jr. E: Human genome project. *Am J Surg* **165**:258, 1993.
481. Charo RA: Effect of the human genome initiative on women's rights and reproductive decisions. *Fetal Diagn Ther* **8**:148, 1993.

482. Sachs BP, Korf B: The human genome project: Implications for the practicing obstetrician. *Obstet Gynecol* **81**:458, 1993.
483. Burn J: Relevance of the Human Genome Project to inherited metabolic disease. *J Inher Metabl Dis* **17**:421, 1994.
484. Cui K-H: Genome project and human reproduction. *Molecular Human Reproduction* **10**:1275, 1995.
485. Williams JK, Lessick M: Genome research: Implications for children. *Pediatr Nurs* **22**:40, 1996.
486. Ellsworth DL, Hallman DM, Boerwinkle E: Impact of the human genome project on epidemiologic research. *Epidemiol Rev* **19**:3, 1997.
486a. Evans GA: The Human Genome Project: Applications in the diagnosis and treatment of neurologic disease. *Arch Neurol* **55**:1287, 1998.
487. Orkin SH, Nathan DG: The molecular genetics of thalassemia, in Harris H, Hirschhorn K (eds.): *Advances in Human Genetics (Volume 11)*, New York, Plenum Press, 1981, p 233.
488. Kwok SCM, Ledley FD, DiLella AG, Robson KJH, Woo SLC: Nucleotide sequence of a full-length complementary DNA clone and amino acid sequence of human phenylalanine hydroxylase. *Biochem* **24**:556, 1985.
489. Persico MG, Viglietto G, Martini G, Toniolo D, Paonessa G, Moscatelli C, Dono R, et al.: Isolation of human glucose-6-phosphate dehydrogenase (G6PD) cDNA clones: Primary structure of the protein and unusual 5¢ non-coding region. *Nucleic Acids Res* **14**:2511, 1986.
490. Orkin SH: Reverse genetics and human disease. *Cell* **47**:845, 1986.
491. Ruddle FH: The William Allan Memorial Award address: Reverse genetics and beyond. *Am J Hum Genet* **36**:944, 1984.
492. Orkin SH: "Forward" and "reverse" genetics of inherited human disorders: The thalassemia syndromes and chronic granulomatous disease. *Harvey Lect* **83**:57, 1989.
493. Friedmann T: Opinion: the human genome project-Some implications of extensive "reverse genetic" medicine. *Am J Hum Genet* **46**:407, 1990.
494. Cawthon RM, Weiss R, Xu G, Viskochil D, Culver M, Stevens J, Robertson M, et al.: A major segment of the neurofibromatosis type 1 gene: cDNA sequence, genomic structure, and point mutations. *Cell* **62**:193, 1990.
495. Attree O, Olivos IM, Okabe I, Bailey LC, Nelson DL, Lewis RA, McInnes RR, et al.: The Lowe's oculocerebrorenal syndrome gene encodes a protein highly homologous to inositol polyphosphate-5-phosphatase. *Nature* **358**:239, 1992.
496. Mitelman F, Mertens F, Johansson B: A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nat Genet* **155**:417, 1997.
497. Parrish JE, Nelson DL: Methods for finding genes: A major rate-limiting step in positional cloning. *Genet Anal Tech Appl* **10**:29, 1993.
498. Chen E, d'Urso M, Schlessinger D: Functional mapping of the human genome by cDNA localization *versus* sequencing. *BioEssays* **16**:693, 1994.
499. Brennan MB, Hochgeschwender U: So many needles, so much hay. *Hum Mol Genet* **4**:153, 1995.
500. Gardiner K, Mural RJ: Getting the message: Identifying transcribed sequences. *Trends Genet* **11**:77, 1995.
501. Monaco AP: Isolation of genes from cloned DNA. *Curr Opin Genet Dev* **4**:360, 1994.
502. Thomas PS: Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. *Proc Natl Acad Sci USA* **77**:5201, 1980.
503. Elvin P, Butler R, Hedge PJ: Transcribed sequences within YACs: HTF island cloning and cDNA library screening, in Anand R (ed.): *Techniques for the Analysis of Complex Genomes*, London, Academic Press Limited, 1992, p 155.
504. Lovett M, Kere J, Hinton LM: Direct selection: A method for the isolation of cDNAs encoded by large genomic regions. *Proc Natl Acad Sci USA* **88**:9628, 1991.
505. Parimoo S, Patanjali SR, Shukla H, Chaplin DD, Weissman SM: cDNA selection: Efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc Natl Acad Sci USA* **88**:9623, 1991.
506. Tagle DA, Swaroop M, Lovett M, Collins FS: Magnetic bead capture of expressed sequences encoded within large genomic segments. *Nature* **361**:751, 1993.
507. Lovett M: Fishing for complements: Finding genes by direct selection. *Trends Genet* **10**:352, 1994.
508. Parimoo S, Patanjali SR, Kolluri R, Xu H, Wei H, Weissman SM: cDNA selection and other approaches in positional cloning. *Anal Biochem* **228**:1, 1995.
509. Peterson AS: Direct cDNA selection, in Birren B, Green ED, Klapholz S, Myers RM, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 2 Detecting Genes*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1998, p 159.
510. Duyk GM, Kim S, Myers RM, Cox DR: Exon trapping: A genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA. *Proc Natl Acad Sci USA* **87**:8995, 1990.
511. Buckler AJ, Chang DD, Graw SL, Brook JD, Haber DA, Sharp PA, Housman DE: Exon amplification: A strategy to isolate mammalian genes based on RNA splicing. *Proc Natl Acad Sci USA* **88**:4005, 1991.
512. Hamaguchi M, Sakamoto H, Tsuruta H, Sasaki H, Muto T, Sugimura T, Terada M: Establishment of a highly sensitive and specific exon-trapping system. *Proc Natl Acad Sci USA* **89**:9779, 1992.
513. Andreadis A, Nisson PE, Kosik KS, Watkins PC: The exon trapping assay partly discriminates against alternatively spliced exons. *Nucleic Acids Res* **21**:2217, 1993.
514. Church DM, Stotler CJ, Rutter JL, Murrell JR, Trofatter JA, Buckler AJ: Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nat Genet* **6**:98, 1994.
515. Krizman DB: Exon trapping, in Birren B, Green ED, Klapholz S, Myers RM, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 2 Detecting Genes*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1998, p 191.
516. Krizman DB, Berget SM: Efficient selection of 3¢-terminal exons from vertebrate DNA. *Nucleic Acids Res* **21**:5198, 1993.
517. Claverie J-M: A streamlined random sequencing strategy for finding coding exons. *Genomics* **23**:575, 1994.
518. Chandrasekharappa SC, Guru SC, Manickam P, Olufemi S-E, Collins FS, Emmert-Buck MR, Debelenko LV, et al.: Positional cloning of the gene for multiple endocrine neoplasia-type 1. *Science* **276**:404, 1997.
519. The International FMF Consortium: Ancient missense mutations in a new member of the RoRet gene family are likely to cause familial Mediterranean fever. *Cell* **90**:797, 1997.
520. Cotton RG: Slowly but surely towards better scanning for mutations. *Trends Genet* **13**:43, 1997.
521. Myers RM, Hedrick Ellenson L, Hayashi K: Detection of DNA variation, in Birren B, Green ED, Klapholz S, Myers RM, Roskams J (eds.): *Genome Analysis: A Laboratory Manual. Vol. 2 Detecting Genes*, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, 1998, p 287.
522. Lupski JR: Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14**:417, 1998.
523. Caskey CT, Pizzuti A, Fu Y-H, Fenwick Jr. RG, Nelson DL: Triplet repeat mutations in human disease. *Science* **256**:784, 1992.
524. Richards RI, Sutherland GR: Heritable unstable DNA sequences. *Nat Genet* **1**:7, 1992.
525. Nelson DL, Warren ST: Trinucleotide repeat instability: When and where? *Nat Genet* **4**:107, 1993.
526. Fischer SG, Lerman LS: DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: Correspondence with melting theory. *Proc Natl Acad Sci USA* **80**:1579, 1983.
527. Myers RM, Lumelsky N, Lerman LS, Maniatis T: Detection of single base substitutions in total genomic DNA. *Nature* **313**:495, 1985.
528. Myers RM, Maniatis T, Lerman LS: Detection and localization of single base changes by denaturing gradient gel electrophoresis, in Wu R (ed.): *Methods in Enzymology (Volume 155)*, San Diego, Academic Press, Inc., 1987, p 501.
529. Winter E, Yamamoto F, Almoguera C, Perucho M: A method to detect and characterize point mutations in transcribed genes: Amplification and overexpression of the mutant c-Ki-*ras* allele in human tumor cells. *Proc Natl Acad Sci USA* **82**:7575, 1985.
530. Myers RM, Larin Z, Maniatis T: Detection of single base substitutions by ribonuclease cleavage at mismatches in RNA:DNA duplexes. *Science* **230**:1242, 1985.
531. Cotton RGH, Rodrigues NR, Campbell RD: Reactivity of cytosine and thymine in single-base-pair mismatches with hydroxylamine and osmium tetroxide and its application to the study of mutations. *Proc Natl Acad Sci USA* **85**:4397, 1988.
532. Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T: Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc Natl Acad Sci USA* **86**:2766, 1989.
533. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, et al.: Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**:789, 1995.
534. Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, Dutra A, Pike B, et al.: Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* **276**:2045, 1997.
535. Everett LA, Glaser B, Beck JC, Idol JR, Buchs A, Heyman M, Adawi F, et al.: Pendred syndrome is caused by mutations in a putative sulphate transporter gene (*PDS*). *Nat Genet* **17**:411, 1997.
536. Yamagata K, Furuta H, Oda N, Kaisaki PJ, Menzel S, Cox NJ, Fajans SS, et al.: Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young. *Nature* **384**:458, 1996.
537. Coffey AJ, Brooksbank RA, Brandau O, Oohashi T, Howell GR, Bye JM, Cahn AP, et al.: Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene. *Nat Genet* **20**:129, 1998.
538. Van Laer L, Huizing EH, Verstreken M, van Zuijlen D, Wauters JG, Bossuyt PJ, Van de Heyning P, et al.: Nonsyndromic hearing impairment is associated with a mutation in *DFNA5*. *Nat Genet* **20**:194, 1998.
539. Nelson SF, McCusker JH, Sander MA, Kee Y, Modrich P, Brown PO: Genomic mismatch scanning: A new approach to genetic linkage mapping. *Nat Genet* **4**:11, 1993.
540. Brown PO: Genome scanning methods. *Curr Opin Genet Dev* **4**:366, 1994.

541. Cheung VG, Nelson SF: Genomic mismatch scanning identifies human genomic DNA shared identical by descent. *Genomics* **47**:1, 1998.
542. McAllister L, Penland L, Brown PO: Enrichment for loci identical-by-descent between pairs of mouse or human genomes by genomic mismatch scanning. *Genomics* **47**:7, 1998.
543. Cheung VG, Gregg JP, Gogolin-Ewens KJ, Bandong J, Stanley CA, Baker L, Higgins MJ, et al.: Linkage-disequilibrium mapping without genotyping. *Nat Genet* **18**:225, 1998.
544. Tugendreich S, Boguski MS, Seldin MS, Hieter P: Linking yeast genetics to mammalian genomes: Identification and mapping of the human homolog of CDC27 via the expressed sequence tag (EST) data base. *Proc Natl Acad Sci USA* **90**:10031, 1993.
545. Tugendreich S, Bassett Jr. DE, McKusick VA, Boguski MS, Hieter P: Genes conserved in yeast and humans. *Hum Mol Genet* **3**:1509, 1994.
546. Bassett Jr. DE, Boguski MS, Spencer F, Reeves R, Goebl M, Hieter P: Comparative genomics, genome cross-referencing and XREFdb. *Trends Genet* **11**:372, 1995.
547. Bassett Jr. DE, Boguski MS, Hieter P: Yeast genes and human disease. *Nature* **379**:589, 1996.
548. Mushegian AR, Bassett Jr. DE, Boguski MS, Bork P, Koonin EV: Positionally cloned human disease genes: Patterns of evolutionary conservation and functional motifs. *Proc Natl Acad Sci USA* **94**:5831, 1997.
549. Bassett Jr. DE, Boguski MS, Spencer F, Reeves R, Kim S, Weaver T, Hieter P: Genome cross-referencing and XREFdb: Implications for the identification and analysis of genes mutated in human disease. *Nat Genet* **15**:339, 1997.
550. Foury F: Human genetic diseases: A cross-talk between man and yeast. *Gene* **195**:1, 1997.
551. Andrade MA, Sander C, Valencia A: Updated catalogue of homologues to human disease-related proteins in the yeast genome. *FEBS Lett* **426**:7, 1998.
552. Banfi S, Borsani G, Rossi E, Bernard L, Guffanti A, Rubboli F, Marchitiello A, et al.: Identification and mapping of human cDNAs homologous to *Drosophila* mutant genes through EST database searching. *Nat Genet* **13**:167, 1996.
553. Kuchler K, Sterne RE, Thorner J: *Saccharomyces cerevisiae STE6* gene product: A novel pathway for protein export in eukaryotic cells. *EMBO J* **8**:3973, 1989.
554. McGrath JP, Varshavsky A: The yeast *STE6* gene encodes a homologue of the mammalian multidrug resistance P-glycoprotein. *Nature* **340**:400, 1989.
555. Raymond M, Gros P, Whiteway M, Thomas DY: Functional complementation of yeast *ste6* by a mammalian multidrug resistance *mdr* gene. *Science* **256**:232, 1992.
556. Bollag G, McCormick F: Differential regulation of rasGAP and neurofibromatosis gene product activities. *Nature* **351**:576, 1991.
557. Ballester R, Marchuk D, Boguski M, Saulino A, Letcher R, Wigler M, Collins F: The *NF1* locus encodes a protein functionally related to mammalian GAP and yeast *IRA* proteins. *Cell* **63**:851, 1990.
558. Han J-W, McCormick F, Macara IG: Regulation of Ras-GAP and the neurofibromatosis-1 gene product by eicosanoids. *Science* **252**:576, 1991.
559. Weeda G, van Ham RCA, Vermeulen W, Bootsma D, van der Eb AJ, Hoeijmakers JHJ: A presumed DNA helicase encoded by *ERCC-3* is involved in the human repair disorders xeroderma pigmentosum and Cockayne's syndrome. *Cell* **62**:777, 1990.
560. Mounkes LC, Jones RS, Liang B-C, Gelbart W, Fuller MT: A Drosophila model for xeroderma pigmentosum and Cockayne's syndrome: *haywire* encodes the fly homolog of *ERCC3,* a human excision repair gene. *Cell* **71**:925, 1992.
561. Sinclair DA, Mills K, Guarente L: Accelerated aging and nucleolar fragmentation in yeast *sgs1* mutants. *Science* **277**:1313, 1997.
562. Fisler JS, Warden CH: Mapping of mouse obesity genes: A generic approach to a complex trait. *J Nutr* **127**:1909S, 1997.
563. Comuzzie AG, Allison DB: The search for human obesity genes. *Science* **280**:1374, 1998.
564. Zhang Y, Proenca R, Maffei M, Barone M, Leopold L, Friedman JM: Positional cloning of the mouse *obese* gene and its human homologue. *Nature* **372**:425, 1994.
565. Friedman JM, Halaas JL: Leptin and the regulation of body weight in mammals. *Nature* **395**:763, 1998.
566. Korf B: Molecular diagnosis (1). *N Engl J Med* **332**:1218, 1995.
567. Korf B: Molecular diagnosis (2). *N Engl J Med* **332**:1499, 1995.
568. Shikata H, Utsumi N, Kuivaniemi H, Tromp G: DNA-based diagnostics in the study of heritable and acquired disorders. *J Lab Clin Med* **125**:421, 1995.
569. Korf BR: Advances in molecular diagnosis. *Curr Opin Obstet Gynecol* **8**:130, 1996.
570. Kiechle FL: Diagnostic molecular pathology in the twenty-first century. *Clin Lab Med* **16**:213, 1996.
571. Wagener C: Molecular diagnostics. *J Mol Med* **75**:728, 1997.
572. Miller AD: Human gene therapy comes of age. *Nature* **357**:455, 1992.
573. Mulligan RC: The basic science of gene therapy. *Science* **260**:926, 1993.
574. Anderson WF: Human gene therapy. *Nature* **392**:25, 1998.
575. Ascenzioni F, Donini P, Lipps HJ: Mammalian artificial chromosomes-Vectors for somatic gene therapy. *Cancer Lett* **118**:135, 1997.
576. Huxley C: Mammalian artificial chromosomes and chromosome transgenics. *Trends Genet* **13**:345, 1997.
577. Willard HF: Human artificial chromosomes coming into focus. *Nat Biotechnol* **16**:415, 1998.
578. Vos J-MH: Mammalian artificial chromosomes as tools for gene therapy. *Curr Opin Genet Dev* **8**:351, 1998.
579. Grimes B, Cooke H: Engineering mammalian chromosomes. *Hum Mol Genet* **7**:1635, 1998.
580. Housman D, Ledley FD: Why pharmacogenomics? Why now? *Nat Biotechnol* **16**:492, 1998.
581. Marshall A: Getting the right drug into the right patient. *Nat Biotechnol* **15**:1249, 1997.
582. Martinez FD, Graves PE, Baldini M, Solomon S, Erickson R: Association between genetic polymorphisms of the beta2-adrenoceptor and response to albuterol in children with and without a history of wheezing. *J Clin Invest* **100**:3184, 1997.
583. Kleyn PW, Vesell ES: Genetic variation as a guide to drug development. *Science* **281**:1820, 1998.
584. Cook-Deegan R: *The Gene Wars: Science, Politics, and the Human Genome,* New York, W.W. Norton & Company, 1994.
585. Meslin EM, Thomson EJ, Boyer JT: The ethical, legal, and social implications research program at the National Human Genome Research Institute. *Kennedy Inst Ethics J* **7**:291, 1997.
586. Billings PR, Kohn MA, de Cuevas M, Beckwith J, Alper JS, Natowicz MR: Discrimination as a consequence of genetic testing. *Am J Hum Genet* **50**:476, 1992.
587. Alper JS, Geller LN, Barash CI, Billings PR, Laden V, Natowicz MR: Genetic discrimination and screening for hemochromatosis. *J Public Health Policy* **15**:345, 1994.
588. Lapham EV, Kozma C, Weiss JO: Genetic discrimination: Perspectives of consumers. *Science* **274**:621, 1996.
589. Hudson KL, Rothenberg KH, Andrews LB, Kahn MJ, Collins FS: Genetic discrimination and health insurance: An urgent need for reform. *Science* **270**:391, 1995.
590. Rothenberg K: Genetic information and health insurance: State legislative approaches. *J Law Med Ethics* **23**:312, 1995.
591. Rothenberg K, Fuller B, Rothstein M, Duster T, Ellis Kahn MJ, Cunningham R, Fine B, et al.: Genetic information and the workplace: Legislative approaches and policy changes. *Science* **275**:1755, 1997.
592. Wilfond BS, Nolan K: National policy development for the clinical application of genetic diagnostic technologies. Lessons from cystic fibrosis. *JAMA* **270**:2948, 1993.
593. Wilfond B, Rothenberg K, Thomson E, Lerman C: Cancer genetic susceptibility testing: Ethical and policy implications for future research and clinical practice. *J Law Med Ethics* **25**:243, 1997.
594. Burke W, Petersen G, Lynch P, Botkin J, Daly M, Garber J, Kahn MJ, et al.: Recommendations for follow-up care of individuals with an inherited predisposition to cancer. I. Hereditary nonpolyposis colon cancer. Cancer Genetics Studies Consortium. *JAMA* **277**:915, 1997.
595. Burke W, Daly M, Garber J, Botkin J, Kahn MJ, Lynch P, McTiernan A, et al.: Recommendations for follow-up care of individuals with an inherited predisposition to cancer. II. BRCA1 and BRCA2. Cancer Genetic Studies Consortium. *JAMA* **277**:997, 1997.
596. Geller G, Botkin JR, Green MJ, Press N, Biesecker BB, Wilfond B, Grana G, et al.: Genetic testing for susceptibility to adult-onset cancer. The process and content of informed consent. *JAMA* **277**:1467, 1997.
597. Post SG, Whitehouse PJ, Binstock RH, Bird TD, Eckert SK, Farrer LA, Fleck LM, et al.: The clinical introduction of genetic testing for Alzheimer disease. An ethical perspective. *JAMA* **277**:832, 1997.
598. Burke W, Thomson E, Khoury MJ, McDonnell SM, Press N, Adams PC, Barton JC, et al.: Hereditary hemochromatosis: Gene discovery and its implications for population-based screening. *JAMA* **280**:172, 1998.
599. Holtzman NA, Murphy PD, Watson MS, Barr PA: Predictive genetic testing: From basic research to clinical practice. *Science* **278**:602, 1997.
600. Germino GG, Somlo S: A positional cloning approach to inherited renal disease. *Semin Nephrol* **12**:541, 1992.


## ABBREVIATIONS

| | |
|---|---|
| BAC | BACTERIAL ARTIFICIAL CHROMOSOME |
| BP | BASE PAIR |
| EST | EXPRESSED-SEQUENCE TAG |
| HGP | HUMAN GENOME PROJECT |
| KB | KILOBASE PAIR |
| LINE | LONG INTERSPERSED NUCLEOTIDE ELEMENT |
| MB | MAGABASE PAIR |
| PAC | P1-DERIVED ARTIFICIAL CHROMOSOME |
| SNP | SINGLE NUCLEOTIDE POLYMORPHISM |
| STS | SEQUENCE-TAGGED SITE |
| SINE | SHORT INTERSPERSED NUCLEOTIDE ELEMENT |
| WEB | WORLD WIDE WEB |