ARTICLES

The Human Genome Project

THOMAS D. YAGER,[†] THOMAS E. ZEWERT,[‡] AND LEROY E. HOOD[*,§]

NSF Center for Molecular Biotechnology, Division of Biology (139-74), California Institute of Technology, Pasadena, California 91125

Received August 3, 1993

I. Introduction

The Human Genome Project (HGP) is a coordinated worldwide effort to precisely map the human genome and the genomes of selected model organisms.[1,2] The first explicit proposal for this project dates from 1985 although its foundations (both conceptual and technological) can be traced back many years in genetics, molecular biology, and biotechnology.[3,4] The HGP has matured rapidly and is producing results of great significance.

II. Genome Maps

The HGP's immediate goal is to create dense and cross-referenced genetic, physical, and nucleotide-sequence maps of selected genomes. Genetic and nucleotide-sequence maps relate to fundamentally different aspects of a living species. A genetic map describes the biological expression of information encoded in its chromosomes. A nucleotide-sequence map describes the physical basis of how this information is stored and represented. (Physical maps are derivative, because they can be generated from knowledge of the complete nucleotide sequence.)

These maps will contribute greatly to the foundation of knowledge in biology and medicine. Figure 1 presents an example of these maps, spanning the CFTR (cystic fibrosis) gene on human chromosome 7. Highly detailed maps will be constructed for every human chromosome, and also for the chromosomes of certain model species, and will be accessible in computer databases.

A. Genetic Map. In higher organisms such as vertebrates, most cells contain two copies of each chromosome, one from the mother and the other from the father. During *meiosis* (the special cell division event that creates the sex cells, or eggs and sperm) the maternal and paternal copies of each chromosome pair together in precise alignment and then exchange segments. This process is called *genetic recombination.* The sites of exchange between paired chromosomes appear to be determined largely at random. Thus the sex cells generated by different meioses in an individual are all genetically unique. After recombination, the paired chromosomes separate and partition into the sex cells, which thus obtain only one copy of each chromosome. During conception, egg and sperm fuse to regenerate the normal two-copy number of each chromosome in the fertilized zygote.

The probability of recombination between two identifiable sites (markers) on a chromosome increases with their physical separation distance (Figure 2). The discovery of this fact led to the idea that a *genetic map* (an ordering relation between markers) could be defined using recombination probability as a measure of genetic distance.[5,6] A 1-*centimorgan* distance corresponds to 1% probability that genetic recombination will occur between two markers during a single meiosis. The ratio between genetic and physical distances on a chromosome varies between species, between sexes, and with the rare presence of "hot spots" or "cold spots" where recombination happens at anomalously high or low rates.[7]

For many years, only rudimentary genetic maps could be constructed for the human (and most other species) because it was hard to find informative genetic markers that could easily be scored. This changed with the realization that *DNA polymorphisms* could be used as

T. D. Yager received an M.S. degree in developmental biology (University of Denver), a Ph.D. in biophysics (Oregon State University), and postdoctoral training at the University of Oregon, the Weizmann Institute (Rehovot, Israel), and Kings College (London, U.K.). He then joined Leroy Hood's group at the Center for Molecular Biotechnology at Caltech. Now at the Hospital for Sick Children (Toronto, Canada), Dr. Yager is using a combination of biophysical, genetic, and microinjection technologies to study developmental control events in a model vertebrate, the zebra fish.

Thomas E. Zewert obtained B.S. and M.S. degrees from Yale and a Ph.D. from Caltech in chemistry. After a year in industry he spent three years as a research scientist in the NSF Molecular Biotechnology Center at Caltech. His main research interests during this period were the design, synthesis, and evaluation of new polymers for the electrophoretic separation of biomolecules and their sensitive detection. He is currently studying medicine at Harvard School of Medicine.

Leroy Hood is Gates Professor and Chairman of the newly created Department of Molecular Biotechnology at the University of Washington School of Medicine. His research interests include molecular immunology, biotechnology, and the Human Genome Project. His laboratory played a major role in developing DNA and protein synthesizers and sequencers. He also played a major role in deciphering the mysteries of antibody diversity. Recently his laboratory finished the 610-kilobase sequence of the human T cell receptor β locus, the first complex multigene family analyzed in higher organisms.

[†] Present address: Division of Nephrology, Hospital of Sick Children, 555 University Ave., Toronto, Canada M5G 1X8.

[‡] Present address: Vanderbilt Hall, 107 Ave. Louis Pasteur, P.O. Box 65, Boston, MA 02115.

[§] Present address: University of Washington School of Medicine, Department of Molecular Biotechnology, FJ-20, Seattle, WA 98195.

[1] Alberts, B. M.; et al. *Mapping and Sequencing the Human Genome;* National Academy Press: Washington, DC, 1988.

[2] Olson, M. V. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 4338-4344.

[3] Sinsheimer, R. L. *Genomics* **1989**, *5*, 954-956.

[4] Watson, J. D.; Cook-Deegan, R. M. *FASEB J.* **1991**, *5*, 8-11.

[5] Sturtevant, A. H. *J. Exp. Zool.* **1913**, *14*, 43-59.

[6] Haldane, J. B. S. *J. Genet.* **1919**, *8*, 299-309.

[7] Chakravarti, A. *Genomics* **1991**, *11*, 219-222.

0001-4842/94/0127-0094$04.50/0 © 1994 American Chemical Society

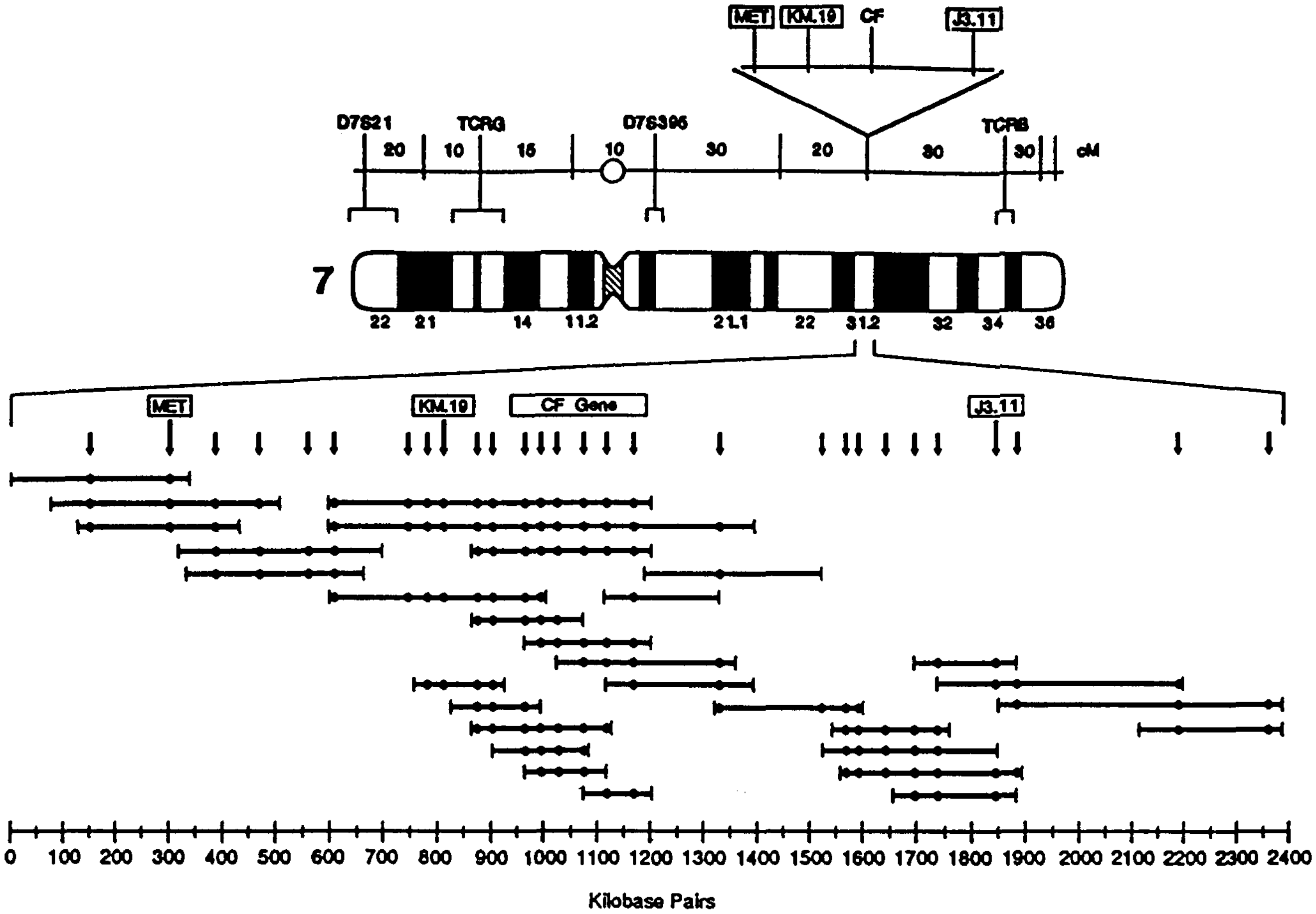

Figure 1. Genetic, physical, and nucleotide sequence maps around the CFTR (cystic fibrosis) gene on human chromosome 7. This chromosome represents ~5.1% of the human genome (~1.5 × 10^7 base pairs). In the figure, the object with alternating light and dark bands represents the chromosome in its condensed state, as seen in cytogenetic staining procedures. Above the chromosome is a crude genetic map with a spacing of 10–30 centimorgans between adjacent markers. Below the chromosome is an expansion of the region around the cystic fibrosis gene. Arrows indicate the positions of STSs. The set of horizontal bars represents a contig of YAC clones. Dots indicate STSs that were used to establish overlap relations between individual YAC clones in the contig. The scale at bottom is calibrated in kilobase pairs of nucleotide sequence. Reprinted with permission from Green, E. D.; Waterston, R. H. *JAMA, J. Am. Med. Assoc.* **1991**, *266* (14), 1966–1975 (Oct 9). Copyright 1991 American Medical Association.

genetic markers.[8] A DNA polymorphism is a segment of DNA that occurs at a unique chromosomal location, and that varies in nucleotide sequence between members of the same species. DNA polymorphisms arise from the accumulation of random, often harmless mutations in a genome during evolution.[9] There are many types of natural DNA polymorphism, such as single nucleotide sites where either of two bases can be found ("single-nucleotide dialleles"), and sites containing *v*ariable *n*umbers of *t*andem *r*epeats of short sequence elements (VNTRs).[10,11] These different types of polymorphism have characteristic genome frequencies, ranging in the human from one per 500 base pairs for single-nucleotide dialleles, to one per 50 000 base pairs for VNTRs.[10,12]

A gene that influences an observable trait or "phenotype" can be localized on a genetic map by *genetic linkage analysis*.[13] This is a maximum-likelihood method which assigns genetic map locations to a set of markers, to produce greatest consistency with the observed inheritance of a phenotype in a family pedigree (see Figure 2). Placement of a gene on a genetic map is the first step toward physically isolating it as DNA by "positional cloning".[14,15] Genetic linkage analysis is computationally intense because of the combinatorial effect that N linked markers can be arranged in $N!/2$ alternative orders. Efficient ordering algorithms and specialized "computing engines" (parallel array processors or dedicated computer chips) may be required to analyze genetic linkage between the hundreds of markers which now comprise a typical genetic map.[16,17] Classical genetic linkage analysis can be augmented by a *physical* method in which the linear order of some of the markers is directly established.[14] One physical method uses X-rays to fragment a chromosome. This

[8] Botstein, D.; White, R. L.; Skolnick, M.; Davis, R. W. *Am. J. Hum. Genet.* **1980**, *32*, 314–331.

[9] Nei, M. *Molecular Evolutionary Genetics*; Columbia University Press: New York, 1987.

[10] Jeffreys, A. J.; Wilson, V.; Neumann, R.; Thein, S. L. *Nature* **1985**, *314*, 67–73.

[11] Nakamura, Y.; Leppert, M.; O'Connell, P.; Wolff, R.; Holm, T.; Culver, M.; Martin, C.; Fujimoto, E.; Hoff, M.; Kumlin, E.; White, R. *Science* **1987**, *235*, 1616–1622.

[12] Weber, J. L. *Genetic and Physical Mapping, Vol. 1. Genome Analysis*; Davies, K. E.; Tilghman, S. M., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, 1990; pp 159–181.

[13] Ott, J. *Analysis of Human Genetic Linkage*, 2nd ed.; Johns Hopkins University Press: Baltimore, 1992.

[14] Collins, F. S. *Nature Genetics* **1992**, *1*, 3–6.

[15] Ballabio, A. *Nature Genetics* **1993**, *3*, 277–279.

[16] Lander, E. S.; Green, P. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 2363–2367.

[17] Miller, P. L.; Nadkarni, P. M.; Bercovitz, P. A. *Comput. Appl. Biosci.* **1992**, *8*, 141–147.

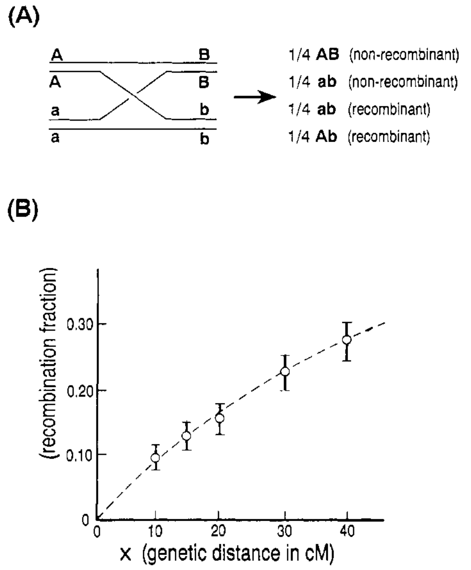

Figure 2. The basis of genetic mapping. (A) Schematic of paternally- and maternally-derived copies of a chromosome, paired at meiosis, after DNA replication but before separation of replicated copies. Each line represents a double-stranded DNA molecule. Two genetic marker loci (A and B) are considered. Locus A is represented originally by the allele (sequence variant) "A" on the paternal chromosome and by allele "a" on the maternal chromosome. Similarly, marker B is represented originally by allele "B" on the paternal chromosome and by allele "b" on the maternal chromosome. In this diagram, a single genetic recombination event is occurring between the markers A and B. An odd number of genetic recombination events will cause a topological change in linkage between alleles at loci A and B, to create the following genotypes in the daughter cells (with frequencies f): AB (nonrecombinant, $f \approx 0.25$), ab (nonrecombinant, $f \approx 0.25$), aB (recombinant, $f \approx 0.25$), Ab (recombinant, $f \approx 0.25$). (B) A "genetic mapping function" relates the observed fraction of recombinant daughter cells (Θ) to the genetic distance (x) between loci.[13] A simple mapping function ($\Theta = (1 - e^{-2x})/2$) can be derived analytically by assuming that recombination is a Poisson process.[6] The standard deviation of the sampling distribution of Θ decreases with increasing number (N) of independent meioses and has been calculated for $N = 100$ in this figure (error bars).

allows one to determine which pairs of markers reside next to each other on a given chromosome fragment.[18]

A genetic map's *resolution* can be defined as the average distance between its least ambiguously ordered markers (framework markers). The HGP has placed a high value on improving the resolution of the genetic maps of many species, especially of the human and the mouse, which have quite similar biologies. A low-resolution (5–10-centimorgan) human genetic map has been constructed using DNA polymorphisms as markers.[19,20] A major goal of the HGP is to refine this map to <2-centimorgan resolution. With a genetic map of sufficient resolution (perhaps 1 cM), multigene traits can be resolved into underlying single-gene components. In addition, the genetic and environmental components of complex diseases such as cancer can be separated and thus understood more clearly.

B. Physical Map. The physical map of a genome is defined in units of linear physical distance such as DNA base pairs. (In B-form DNA, one base pair corresponds to ~3.6-Å length.) The physical map is not irreducible elementary knowledge, because it can be derived from the nucleotide sequence. It is perhaps best viewed as a means for obtaining the nucleotide sequence, and for facilitating biological experiments that involve transfer of genes into organisms.

Traditionally, physical genome maps have been constructed by a combination of "top–down" and "bottom–up" strategies.[21] In a top–down strategy, the genome is divided into chromosomes by flow-sorting.[22] Then the chromosomes are further divided, by site-specific cleavage, into smaller fragments which are cloned. In a bottom–up strategy, short genomic fragments are cloned at random, and a representation of each chromosome is built up as a *contig*, or set of contiguous overlapping clones. (*Cloning* is the process by which a DNA fragment of interest is covalently joined to a carrier DNA molecule, or *vector*, that can replicate in a host bacterium or yeast. A collection of different clones that is derived from a particular genomic region, or from a pool of tissue-specific messenger RNA molecules, is called a *library*.)

A contig is defined by deducing all the overlap relations between its constituent clones (see Figure 1). Two methods are commonly used to deduce these overlaps.[23] In *chromosome-walking*,[21] the ends of a cloned DNA sequence are used as "probes" to detect overlap with other clones in a library. Detection is based on the principle of hybridization, which is the formation of a complex between the probe and its complementary nucleotide sequence in the target DNA. Chromosome-walking will fail if either end of the insert contains a repetitive DNA element that can hybridize to many sites in the genome. In *fingerprint-matching*,[24,25] clones that overlap are identified on the basis of displaying identical features (fingerprints) in some assay, such as restriction-endonuclease digestion.

Because the information content of a fingerprint can be low, the traditional method of physical mapping may lead to the assignment of false overlap relations.[26] The consequence is a definition of contigs at the *conceptual* level that do not accurately represent the *true* structure of the genome. Traditional physical mapping also depends on cloning, which raises the problem that a clonally-unstable sequence may produce a gap in the physical map.

The "polymerase chain reaction" (PCR) suggests a new approach for physical mapping which does not suffer from the above shortcomings. PCR[27–29] is an *in vitro* method for amplifying virtually any DNA sequence (shorter than about 2000 base pairs) that can be bounded on the ends by unique oligonucleotide primers. PCR is a true chain-reaction process, in which the amount of product increases exponentially with the

[18] Cox, D. R.; Burmeister, M.; Price, E. R.; Kim, S.; Myers, R. M. *Science* **1990**, *250*, 245–250.

[19] Donis-Keller, H.; et al. *Cell* **1987**, *51*, 319–337.

[20] Weissenbach, J.; Gyapay, G.; Dib, C.; Vignal, A.; Morissette, J.; Millasseau, P.; Vaysseix, G.; Lathrop, M. *Nature* **1992**, *359*, 794–801.

[21] Evans, G. A.; Lewis, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 5030–5034.

[22] Gray, J. W.; Dean, P. N.; Fuscoe, J. C.; Peters, D. C.; Trask, B. J.; van den Engh, G. J.; van Dilla, M. A. *Science* **1987**, *238*, 323–328.

[23] Evans, G. A. *BioEssays* **1991**, *13*, 39–44.

[24] Coulson, A.; Sulston, J.; Brenner, S.; Karn, J. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 7821–7825.

[25] Craig, A. G.; Nizetic, D.; Hoheisel, J. D.; Zehetner, G.; Lehrach, H. *Nucleic Acids Res.* **1990**, *18*, 2653–2660.

[26] Branscomb, E.; Slezak, T.; Pae, R.; Galas, D.; Carrano, A. V.; Waterman, M. *Genomics* **1990**, *8*, 351–366.

[27] Gibbs, R. A. *Anal. Chem.* **1990**, *62*, 1202–1214.

[28] Erlich, H. A.; Gelfand, D.; Sninsky, J. J. *Science* **1991**, *252*, 1643–1651.

[29] Bloch, W. *Biochemistry* **1991**, *30*, 2735–2747.

number of reaction cycles. A "*sequence tagged site*" (STS) is defined as a unique DNA sequence that can be amplified with high specificity from genomic DNA by PCR.[30] A physical genome map can be defined unambiguously in terms of an ordered set of STSs (Figure 1).[30]

Four advantages are gained by defining a physical genome map with STSs. (1) The need for centralized clone banks is eliminated. (2) Clones can be correctly overlaid to build up contigs by matching their content of STSs (Figure 1). When STSs of known sequence are used, a more definitive match will be obtained than is possible with fingerprints. This is because STSs of known sequence contain more information than fingerprints. (3) Clone-rearrangement artifacts will be readily detected, if they are larger than the spacing between adjacent STSs. (However, if a change in the sequence of a clone falls entirely between adjacent STSs, then it will be missed.) (4) Polymorphic STSs can serve as correspondence points between genetic and physical maps.[31-33]

It is simple to generate new STSs. Random genomic DNA fragments are sequenced, and this sequence information is used to design PCR primers. The chromosomal locations of the new STSs are then determined, using the following techniques. (1) An STS can be used in PCR, to screen a panel of *somatic cell hybrids*.[34] This is a set of nonhuman cell lines, each of which carries a different human chromosome or chromosome fragment, in addition to its own genome. PCR will yield a positive result on the cell line that carries the matching human chromosome fragment. The resolution of this method depends on the distribution of sizes of human chromosome fragments in the deletion panel and currently is ~5% of a chromosome's length. (2) An STS can be used as a probe for sequence-specific hybridization to chromosomes that are immobilized on a solid surface. The unique position of a bound probe, relative to the ends of the chromosome, often can be determined with a resolution of about ±1% of a chromosome's length.[35] (3) If the STS contains a DNA polymorphism, then it can serve as a genetic marker. Placement of an STS on the genetic map will help in determining its physical map location.

A large number of STSs can be defined over a small region of a genome, to produce a high-resolution physical map. To achieve the current HGP goal of one STS "landmark" every 100 000 base pairs, at least 30 000 STSs must be designed and validated.[36] Several pilot projects are underway to assess the feasibility of this goal.[37]

C. Nucleotide Sequence Map. A gene is exactly specified by its nucleotide sequence, and there is a collinear and unique relationship between a gene's nucleotide sequence and the sequence of amino acids in the protein that it encodes. It is important to learn the amino acid sequence of a protein because conserved functional motifs may be revealed. Ultimately, application of protein-folding rules to the amino acid sequence may allow the protein's three-dimensional structure to be predicted.[38,39]

The HGP became feasible only after rapid DNA-sequencing was invented (Nobel Prize in chemistry, 1980). Currently two DNA-sequencing methods are used, which are based on similar principles. In *chemical sequencing* a single-stranded DNA molecule is labeled at one end, and then it is cleaved at random with four different base-specific reagents in four separate reactions.[40] In *enzymatic sequencing*, a DNA polymerase enzyme is used to synthesize the complementary strand of a DNA template, and a nonextendable deoxynucleotide analogue is incorporated at random positions in each of four separate base-specific reactions.[41] Then gel electrophoresis of the four reaction mixtures is used to resolve the DNA fragments that are generated by base-specific cleavage in chemical sequencing, or by termination after A, C, G, or T residues in enzymatic sequencing. From the pattern or "ladder" of electrophoretic bands on the gel, the DNA sequence can be deduced. The chemical and enzymatic sequencing methods are complementary. The chemical method can yield information about the structure of DNA (as well as its sequence), because unusual DNA conformations may display altered reactivities toward the chemical sequencing reagents. The enzymatic method, on the other hand, is more rapid and can be partially automated. These two methods continue to be refined, with a clear trend toward greater automation.[42]

Current technology allows the complete nucleotide sequence of a substantial genetic region to be determined. Many reports are now appearing which describe the determination of sequences of >100 000 contiguous base pairs from the human genome and the genomes of various model organisms. The record to date is the sequence of an *entire* yeast chromosome which spans ~300 000 contiguous base pairs.[43] In spite of these achievements, several problems persist in bringing rapid DNA-sequencing from the small to large scale, which we discuss below.

D. Integration of Maps. The genetic, physical, and nucleotide maps of a genome should each be continuous, and many evenly-spaced correspondence points should be defined between them.[1] For the human genome, technical limitations and insufficient data have conspired to make this a challenging goal.[1] (1) Many human genes have imprecise genetic map locations. This is due to a paucity of highly informative, tightly-linked markers. It is also due to a lack of large human families which carry detectable phenotypes (such as inherited diseases) that are caused by defects in the genes in question. Such families provide the raw

[30] Olson, M.; Hood, L.; Cantor, C.; Botstein, D. *Science* **1989**, *245*, 1434–1435.

[31] Levitt, R. C. *Genomics* **1991**, *11*, 484–489.

[32] Nickerson, D. A.; Whitehurst, C.; Bosyen, C.; Charmley, P.; Kaiser, R.; Hood, L. *Genomics* **1992**, *12*, 377–387.

[33] Avramopoulos, D.; Chakravarti, A.; Antonarakis, S. E. *Genomics* **1993**, *15*, 98–102.

[34] Abbott, C.; Povey, S. In *Molecular Biology in Medicine*; Mathew, C., Ed.; Humana Press: Totowa, NJ, 1990.

[35] Lichter, P.; Tang, C. C.; Call, K.; Hermanson, G.; Evans, G. A.; Housman, D.; Ward, D. C. *Science* **1990**, *247*, 64–69.

[36] Barillot, E.; Dausset, J.; Cohen, D. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 3917–3921.

[37] Green, E. D.; Green, P. *PCR Methods Appl.* **1991**, *1*, 77–90.

[38] Head-Gordon, T.; Stillinger, F. H.; Arrecis, J. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 11076–11080.

[39] Zwanzig, R.; Szabo, A.; Bagchi, B. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 20–22.

[40] Maxam, A. M.; Gilbert, W. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 560–564.

[41] Sanger, F.; Nicklen S.; Coulson A. R. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 5463–5467.

[42] Hunkapiller, T.; Kaiser, R. J.; Koop, B. F.; Hood, L. *Science* **1991**, *254*, 59–67.

[43] Oliver, S. G.; et al. *Nature* **1992**, *357*, 38–46.

material for genetic linkage analysis. (2) Genes may have unknown physical map locations, and the mutations in DNA sequence which cause phenotypic changes may be unidentified. (3) Ambiguities and errors in the determination of a long nucleotide sequence are hard to eliminate completely.[44]

III. Challenges Faced by the HGP

The HGP has made a bold and significant start toward its goals. However, the sheer magnitude of these goals raises a number of technical problems.

A. Clone Instabilities. Yeast artificial chromosomes ("YACs") are useful cloning vectors because they can carry large inserts (between 10^5 and 10^6 base pairs in length).[45,46] This large insert size should, in principle, allow contigs to be built efficiently. However, many human genomic YAC libraries are found to contain "chimeric" clones, in which noncontiguous genomic DNA fragments are falsely juxtaposed.[47,49] This probably is due to recombination between human-specific repetitive sequence elements, during propagation of the YACs in yeast.[47] The spontaneous generation of recombinant YACs introduces a bottleneck in physical mapping. All the YACs in each contig must be tested for integrity, to make sure they do not contain fragments from different (noncontiguous) chromosomal regions. Recombination-deficient yeast strains, in which human-derived YACs are more stable, may alleviate this problem.[50]

Repetitive elements are not the only cause of clone instability. Some vector/host combinations are prone to spontaneous deletion or rearrangement of genetically-unstable sequences such as long sequence-repeats, homopolymer tracts, or biological regulatory elements.[51–53] Specialized hosts and vectors will help to minimize these instabilities.[53–54,56]

As long as cloning is an integral step in genome research, the existence of clonally-unstable sequences will create problems. It is possible that a technique such as PCR-mediated gene synthesis[57] may ultimately make conventional cloning obsolete.

B. Large-Scale DNA Sequencing. Large-scale genomic DNA-sequencing is complicated by logistic concerns (bottlenecks, quality control, etc.) and also by several problems of scale which are not important in small projects. (1) A long nucleotide-sequence determination will contain, as a natural feature, many DNA polymorphisms. Therefore the sequence of a large genetic region is not unique and must be defined as a consensus or ensemble average.[58] (2) A long sequence determination will contain many errors and possibly even gaps. The level of accuracy that is required in the determination of the nucleotide sequence of a genetic region will vary with the perceived importance of the region and with the intended use of the sequence information.[44] (3) The cost of assembling a final genomic sequence is high, because only ~500 base pairs of DNA sequence can be determined reliably in a single assay, while a genetically interesting region could easily span >1 000 000 base pairs.[59] The current technology for DNA-sequencing (based on separating nested sets of DNA fragments on gels) may be limited to a resolution of ~1000 nucleotides because of the physics underlying the gel-separation process.[60]

Many approaches have been devised for assembling a long complete sequence from short fragments. These may be ranked according to their degree of "random" or "directed" character.[42] In a purely *random* strategy, DNA from the region of interest is fragmented and subcloned. Subclones are sequenced at random, and the sequences are then correctly superimposed by an exhaustive pairwise comparison. In the most common *directed* strategy, one end of the genetic region of interest is sequenced first. This allows a new sequencing primer to be designed, to read further into the interior of the region. This "primer-walking" operation is continued until the other end of the region is reached. Neither the random nor the directed strategy is perfect. A purely random approach requires that each base position, on average, be sequenced many times (to eliminate all gaps), while a purely directed approach requires many sequencing primers.[61,62] It may prove best to combine random and directed strategies.[42] Alternatively, a radically different DNA-sequencing technology, based for example on scanning tunneling microscopy, may eventually circumvent this problem by allowing a direct "readout" of long contiguous DNA sequences.[63]

Recent proposals suggest that dedicated cDNA-sequencing should serve as a complement to large-scale genomic sequencing.[64,65] cDNAs represent the portion of a genome that is expressed in mRNA and thus translated into protein. When the sequence from a cDNA clone is used to define an STS, then a functional gene becomes localized on the physical genome map. If the STS contains a DNA polymorphism, then the gene may also be localized on the genetic map.[31,33] This proposal is superficially attractive because cDNA-sequencing rapidly produces significant new information and can be done with current technology. It also has a high tolerance to sequencing errors since a high fraction of nucleotides in a cDNA sequence will encode

[44] States, D. J. *Trends Genet.* **1992**, *8*, 52–55.

[45] Burke, D. T.; Carle, G. F.; Olson, M. V. *Science* **1987**, *236*, 806–812.

[46] Larin, Z.; Monaco, A. P.; Lehrach, H. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 4123–4127.

[47] Green, E. D.; Riethman, H. C.; Dutchik, J. E.; Olson, M. V. *Genomics* **1991**, *11*, 658–669.

[48] Selleri, L.; Eubanks, J. H.; Giovannini, M.; Hermanson, G. G.; Romo, A.; Djabali, M.; Maurer, S.; McElligott, D. L.; Smith, M. W.; Evans, G. A. *Genomics* **1992**, *14*, 536–541.

[49] Anderson, C. *Science* **1993**, *259*, 1684–1687.

[50] Moir, D. T. *Human Genome: 1991–1992 Program Report*; U. S. Dept. of Energy: Washington, DC, 1992; p 90.

[51] Leach, D. R. F.; Stahl, F. W. *Nature* **1983**, *305*, 448–451.

[52] Williams, W. L.; Muller, U. R. *J. Mol. Biol.* **1987**, *196*, 743–755.

[53] Gentz, R.; Langner, A.; Chang, A. C. Y.; Cohen, S. N.; Bujard, H. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 4936–4940.

[54] Wyman, A. R.; Wolfe, L. B.; Botstein, D. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 2880–2884.

[55] Gibson, T. J.; Coulson, A. R.; Sulston, J. E.; Little, P. F. R. *Gene* **1987**, *53*, 275–281.

[56] Kim, U.-J.; Shizuya, H.; de Jong, P. J.; Birren, B.; Simon, M. I. *Nucleic Acids Res.* **1992**, *20*, 1083–1085.

[57] Jayaraman, K.; Fingar, S. A.; Shah, J.; Fyles, J. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 4084–4088.

[58] Cavalli-Sforza, L. L. *Am. J. Hum. Genet.* **1990**, *46*, 649–651.

[59] Monaco, A. P.; Kunkel, L. M. *Trends Genet.* **1987**, *3*, 33–37.

[60] Grossman, P. D.; Menchen, S.; Hershey, D. *Genet. Anal.: Tech. Appl.* **1992**, *9*, 9–16.

[61] Kaiser, R. J.; MacKellar, S. L.; Vinayak, R. S.; Sanders, J. Z.; Saavedra, R. A.; Hood L. E. *Nucleic Acids Res.* **1989**, *17*, 6087–6102.

[62] Hawkins, T. L.; Sulston, J. E. *Technique* **1990**, *2*, 307–310.

[63] Lindsay, S. M.; Philipp, M. *Genet. Anal.: Tech. Appl.* **1991**, *8*, 8–13.

[64] Sikela, J. M.; Auffray, C. *Nature Genetics* **1993**, *3*, 189–191.

[65] Adams, M. D.; Kerlavage, A. R.; Fields, C.; Venter, J. C. *Nature Genetics* **1993**, *4*, 256–267.

protein.[44,66] Unfortunately, however, it relies on imperfectly representative cDNA libraries as a source of sequencing templates and, thus, will miss many genes that are rarely expressed. It will also miss most control sequences, which are not transcribed into mRNA. A final criticism is that a typical cDNA-sequencing strategy will produce two discontinuous sequence fragments from the two ends of a cDNA clone. A redundantly-determined sequence over the entire coding region is much more informative but is correspondingly harder and more expensive to obtain.

We believe that there is no easy substitute for the difficult task of complete genomic sequencing. Only this approach can guarantee that nothing significant will be missed.[42,67] If resources allow, it should be even better to sequence the homologous regions of human and model genomes.[68] Of course, the information-poor regions of a genome (such as tandemly-repeated short DNA sequences around the centromeres of chromosomes) are not worth sequencing completely.

C. Detecting New Genes. The human genome is estimated to contain, on average, one gene per 30 000 base pairs of DNA.[69] If the average gene specifies a protein of molecular weight 50 000 Daltons, then only 5% of its nominal 30 000 base pair length will directly encode the protein. (Some of the remaining DNA must specify other essential biological functions such as control of transcription. However, some DNA may also be without function, persisting in the genome simply because there is no positive selective pressure for its removal.[9])

Direct sequencing is the only approach that is virtually guaranteed to reveal all the coding sequence in a genetic region of interest. After a sequence is determined, it can be analyzed in two complementary ways. (1) It can be compared against the entire database of known genes and conserved sequence elements.[70–72] (2) It can be analyzed by pattern-recognition algorithms which detect regions of sequence that have high protein-coding potential.[73–75] Currently the direct-sequencing method appears cost-effective only for information-rich regions of a genome, although this will change as sequencing techniques become more efficient.

Selective "gene hunting" approaches have also been developed, which involve minimal DNA-sequencing and clearly work well in some cases:[76,77]

(1) **Genetic traps** are specialized cloning vectors that force a biological selection, in the host, for recovery of certain genetic elements. Genetic traps have been successfully used, for example, to detect promoters and enhancers, which are elements that direct RNA polymerase to initiate gene transcription.[76–78]

(2) **Motif screening** uses evolutionarily conserved sequence motifs to detect homologies in large blocks of uncharacterized DNA, either by hybridization or by PCR.[79–81] This technique is limited to the detection of motifs that are already known. Also, it may fail if the motifs of interest have not been highly conserved during evolution.

(3) **Detection of CG Islands.** Through evolution, the 5′-ends of vertebrate genes have come to contain a high frequency of CG dinucleotides and, accordingly, are known as "CG islands".[82] These can readily be detected by digestion with CG-specific restriction enzymes[83] or by hybridization using short CG-specific probes.[84]

D. Identifying Mutations That Affect Phenotype. Certain genetic mutations are found empirically to produce detectable alterations in phenotype. This causal relation has been of great value in elucidating the biology of a species. It is also of particular medical interest when mutations in key human genes produce severe inherited diseases. Currently several techniques are preferred for identifying mutations that affect biological phenotype:

(1) **Direct Sequencing.** If a phenotype-altering mutation falls within a small genomic region (<2000 base pairs), then PCR may be used to amplify this region from the DNA of normal and affected individuals, for direct sequence comparison.[85–87]

(2) **Detection of Single-Stranded Conformational Polymorphisms (SSCPs).** If a PCR-amplified DNA molecule is separated into single strands by heating and then quickly jumped to low temperature, the separated DNA strands can be trapped as internally-stacked and hydrogen-bonded structures, in preference to repairing with each other.[88] Normal and mutant variants of a DNA molecule will often produce different single-stranded conformational isomers, which may be resolved on a gel. Two drawbacks are that the optimum gel conditions must be found by trial and error, and certain polymorphisms may not produce distinct conformational isomers.

(3) **Denaturant-Gradient Gel Electrophoresis (DGGE).** In this technique,[89] a DNA molecule is electrophoresed at a temperature slightly below the equilibrium melting temperature (T_m) of its least-stable domain. If a gradient of denaturant exists along the direction of electrophoresis, then this domain will denature at a characteristic and measurable position

[66] States, D. J.; Botstein, D. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 5518–5522.

[67] Hood, L.; Smith L. *Issues Sci. Technol.* **1987**, *3*, 36–46.

[68] Hood, L.; Koop, B.; Goverman, J.; Hunkapiller, T. *Trends Biotechnol.* **1992**, *10*, 19–22.

[69] Ohno, S. *Trends Genet.* **1986**, *2*, 8.

[70] Lipman, D. J.; Pearson, W. R. *Science* **1985**, *227*, 1435–1441.

[71] Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215* , 403–410.

[72] Gish, W.; States, D. J. *Nature Genetics* **1993**, *3*, 266–272.

[73] Nakata, K.; Kanehisa, M.; DeLisi, C. *Nucleic Acids Res.* **1985**, *13*, 5327–5340.

[74] Gelfand, M. S. *Nucleic Acids Res.* **1990**, *18*, 5865–5869.

[75] Uberbacher, E. C.; Mural, R. J. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 11261–11265.

[76] Hochgeschwender, U.; Brennan, M. B. *BioEssays* **1991**, *13*, 139–144.

[77] Hochgeschwender, U. *Trends Genet.* **1992**, *8*, 41–44.

[78] Freeman, M. *Curr. Biol.* **1991**, *1*, 378–381.

[79] Lathe, R. *J. Mol. Biol.* **1985**, *183*, 1–12.

[80] Lee, C. C.; Wu, X.; Gibbs, R. A.; Cook, R. G.; Munzy, D. M.; Caskey, C. T. *Science* **1988**, *239*, 1288–1291.

[81] Mazzarella, R.; Montanero, V.; Kere, J.; Reinbold, R.; Ciccodicola, A.; D'Urso, M.; Schlessinger, D. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 3681–3685.

[82] Bird, A. P. *Trends Genet.* **1987**, *3*, 342–347.

[83] Hermanson, G. G.; Lichter, P.; Selleri, L.; Ward, D. C.; Evans, G. A. *Genomics* **1992**, *13*, 134–143.

[84] Melmer, G.; Buchwald, M. *DNA Cell Biol.* **1990**, *9*, 377–385.

[85] Levedakou, E. N.; Landegren, U.; Hood, L. E. *BioTechniques* **1989**, *7*, 438–442.

[86] Gyllensten, U. B. *BioTechniques* **1989**, *7*, 700–708.

[87] Bevan, I. S.; Rapley, R.; Walker, M. R. *PCR Methods Appl.* **1992**, *1*, 222–228.

[88] Orita, M.; Suzuki, Y.; Sekiya, T.; Hayashi, K. *Genomics* **1989**, *5*, 874–879.

[89] Myers, R. M.; Maniatis, T.; Lerman, L. S. *Methods Enzymol.* **1987**, *155*, 501–527.

down the gel. This gel position will be altered if the least-stable DNA domain contains a polymorphism.

(4) **Oligonucleotide Hybridization.** A long DNA molecule should, in principle, give a unique hybridization pattern (fingerprint) when used as a probe against a gridded array of all unique-sequence oligonucleotides of a certain length (N). Thus long DNA molecules might be rapidly screened for differences in polymorphism content by comparing their hybridization patterns against a gridded oligonucleotide array.[90] Currently, the chief difficulty with this proposal is the apparent requirement that all possible oligonucleotide N-mers must be synthesized in array format on a filter. 'Because there are four nucleotide bases in DNA, the number of required N-mers will equal 4^N, which is an exponentially-increasing function of N. Also, filter hybridization with short oligonucleotides is technically difficult.

IV. Immediate Uses of Partial Genome Maps

It may prove slow and expensive to achieve complete continuity and correspondence of all three types of genome maps.[1] However, even incomplete maps can provide valuable information.

A. Positional Cloning. A disease-causing gene can be isolated on the basis of its chromosomal location, instead of its biological function (which may be unknown). This "*positional cloning*" approach can succeed even when genetic, physical, and nucleotide sequence maps are incomplete.[14,15] A positional cloning project begins by identifying human families in which a disease shows a clear pattern of inheritance according to Mendel's laws of genetics.[91] The disease-causing gene is then localized on the genetic map by linkage analysis, using DNA sequence polymorphisms as genetic markers.[13] A disease-causing gene can usually be localized between flanking genetic markers with a precision of ±1 centimorgan, which corresponds roughly to ±1 000 000 base pairs in the human genome. At this point, all potential genes in the region must be identified, and a systematic process of elimination must be used to identify the gene that causes the disease. It sometimes happens that an individual can be found who displays both the disease phenotype and also a small chromosomal deletion that spans the disease-causing gene and can be seen in the microscope.[14] In such a case, a "subtractive" genomic library can be made, which is greatly enriched in those DNA sequences which are present in a normal individual but are deleted from the diseased individual.[92] Such a subtractive library will greatly facilitate the isolation of the disease-causing gene. Positional cloning has been stunningly successful and has identified the genes that cause many devastating human genetic diseases including cystic fibrosis, muscular dystrophy, Huntington's chorea, and specific cancers such as retinoblastoma.[14,15]

B. Functional Studies. The HGP is rapidly generating clones and partial sequence information for many genes. This will lead to functional studies of two kinds. (1) Specific genes can be inactivated in the genome of a living organism by "gene-targeting" methods.[93] (2) New (or modified) genes can be stably introduced into the genome, to create so-called "transgenic" organisms.[94] Together, these two experimental approaches allow the *functional* importance of particular genes to be assessed during embryonic development. In principle, any gene can be studied by these methods once its sequence is known.

V. Ultimate Consequences of the HGP

Some of the HGP's biological and medical consequences are easy to grasp. (1) Great advances will be made in diagnosing many human diseases that are inherited as single-gene traits.[95] Eventually, multigenic traits will be resolved into single-gene components, and the environmental and genetic factors in complex diseases will be disentangled. (2) Sequence comparison will allow a much deeper understanding of the evolution of genes and regulatory elements. (3) Once all human genes are identified, mapped, and sequenced, it will be possible to define the "transcription map" of an organism.[77] Such a map will describe not only the chromosomal locations and sequences of the genes but also their patterns of expression and functional importance during the development of an organism.[77] (4) The techniques of gene-targeting and transgenic animal creation (above) will provide the technological basis for gene-replacement therapy in humans.[96] This has the promise of correcting, at a fundamental level, many currently untreatable genetically-based diseases.

The HGP will also have major impacts on society that are much harder to predict.[97,98] One obvious topic of concern is the confidentiality and potential misuse of genetic screening information. Major scientific advances often have been the catalysts of social change,[99] and in this respect the HGP is not unusual. However, the HGP is unique in that 3–5% of its funding is earmarked for study of the social implications of the project.[100] If the extent of social change is proportionate to the magnitude of the underlying scientific advance, then a new age of biology, medicine, and society could result from the Human Genome Project.

[90] Drmanac, R.; Drmanac, S.; Strezoska, Z.; Paunesku, T.; Labat, I.; Zeremski, M.; Snoddy, J.; Funkhouser, W. K.; Koop, B.; Hood, L.; Crkvenjakov, R. *Science* **1993**, *260*, 1649–1652.

[91] McKusick, V. A. *Mendelian Inheritance in Man: Catalogs of Autosomal Dominant, Autosomal Recessive, and X-Linked Phenotypes*, 10th ed.; Johns Hopkins University Press: Baltimore, 1992.

[92] Kunkel, L. M.; Monaco, A. P.; Middlesworth, W.; Ochs, H. D.; Latt, S. A. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 4778–4782.

[93] Sedivy, J. M.; Joyner, A. L. *Gene Targeting*; W. H. Freeman: New York, 1992.

[94] Hanahan, D. *Science* **1989**, *246*, 1265–1275.

[95] Caskey, C. T. *Science* **1987**, *236*, 1223–1228.

[96] Vega, M. A. *Human Genet.* **1991**, *87*, 245–253.

[97] Nelkin, D; Tancredi, L. *Dangerous Diagnostics: The Social Power of Biological Information;* Basic Books: New York, 1989.

[98] Keveles, D., Hood L., Eds. *The Code of Codes*; Harvard University Press: Cambridge, 1992.

[99] Kuhn, T. S. *The Structure of Scientific Revolutions*, 2nd ed.; University of Chicago Press: Chicago, 1970.

[100] McGourty, C. *Nature* **1989**, *342*, 603.