Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 10841–10848, November 1995

Review

How is the Human Genome Project doing, and what have we learned so far?

Mark S. Guyer and Francis S. Collins

National Center for Human Genome Research, National Institutes of Health, Bethesda, MD 20892

ABSTRACT **In this paper, we describe the accomplishments of the initial phase of the Human Genome Project, with particular attention to the progress made toward achieving the defined goals for constructing genetic and physical maps of the human genome and determining the sequence of human DNA, identifying the complete set of human genes, and analyzing the need for adequate policies for using the information about human genetics in ways that maximize the benefits for individuals and society.**

The purpose of the Human Genome Project (HGP) is to generate a set of information, material, and technology resources that will be readily available to the entire scientific community. This "scientific infrastructure" will vastly improve the ability of investigators from a variety of fields to apply molecular approaches to the study of a wide range of biological problems as we enter the 21st century. Specifically, the HGP is intended to develop, efficiently and cost-effectively, detailed genetic and physical maps of the human genome, determine the complete nucleotide sequence of human DNA, and develop the new technology necessary to achieve these ambitious goals. This information will lead, in turn, to the locations of the estimated 50,000–100,000 genes within the human genome. The goals of the HGP also include the construction of detailed genetic and physical maps and acquisition of DNA sequence information characterizing the genomes of several nonhuman organisms used extensively in research laboratories as model systems. This review focuses on progress on the characterization of the genome of *Homo sapiens*; progress on the characterization of the genomes of several model organisms is discussed in the accompanying set of papers.

Among the many issues raised in the early discussions of the genome project in the late 1980s was its feasibility. There was considerable skepticism that technology adequate to the challenge of building dense maps and sequencing billions of base pairs of DNA was available or could be developed in a reasonable period of time. Thus, perhaps the most important lesson that has been learned from the HGP is that it can be done.

Indeed, from the beginning, progress in genomics research has been remarkably rapid (Fig. 1). In part, this has been due to the strategic planning that has been an integral part of the HGP. In 1990, the U.S. agencies involved [the National Institutes of Health (NIH) National Center for Human Genome Research (NCHGR) and the Department of Energy (DOE) Office of Health and Environmental Research] publicly presented a plan that described the scientific and other goals for the first phase of the project (2). Even in 1990, the gathering pace of genome research was beginning to be evident, and the NIH/DOE plan noted that the general plan that had been described just two years earlier by the National Research Council (3) was "still appropriate, but some of the details must be changed as improvements in the technology have occurred in the past two years." As for the 1990 plan, it was refined and extended just three years later (4) "because a much more sophisticated and detailed understanding of what needs to be done and how to do it is now available." And in 1995, the signs are rapidly emerging that the 1993 plan is similarly being superseded by scientific developments, so that a new plan will likely be needed by 1997.

Genetic Mapping

In late 1994, the human genetic linkage map became the first of the major goals of the HGP to be reached, when an international group of investigators, representing contributions from >100 laboratories, published a comprehensive human linkage map (5). The map contains 5826 loci covering 4000 centimorgans (cM) on a sex-averaged map, representing an average marker density of 0.7 cM. This is well beyond the initial HGP goal of a 2- to 5-cM map. The rapid success in genetic linkage mapping was the result, in large part, of the introduction of microsatellite-based genetic markers (6, 7) and the development of large-scale, semiautomated methods for marker isolation, typing, and analysis (8, 9).

This map is much more detailed than any previous human linkage map; yet, it is still far from ideal. Only 908 of the markers on it are ordered with high confidence (odds of >1000:1) (5). These constitute a "framework map" of about 4-cM resolution. The additional 4500 markers are localized with respect to the markers of the framework map with odds for order between 10:1 and 1000:1. There also are certain technical problems with microsatellites that make them more difficult to use than one would like, especially for highly automated approaches to genotyping (10). Continued improvement in the technology for genetic mapping and genotyping, giving more reliable order information, improved markers, and better methods for genotyping on a large scale would make the human genetic linkage map even more useful.

Physical Mapping

Considerable progress has also been made in constructing physical maps of the human genome. The current 5-year HGP goal for physical mapping calls for completion of maps based on sequence-tagged site (STS)* markers, with an average spacing of 100 kb. STS-based physical maps were taken as the HGP goal because STSs provide a "common mapping language"; since they can be used as markers in a variety of mapping techniques, STSs can be used to integrate and compare physical maps constructed by different techniques (11). For the same reason, STSs can also be used to integrate genetic and physical maps. STSs have rapidly gained acceptance as the markers of choice for map construction (Fig. 1*B*; most of the PCR probes in the GDB are STSs), and the goal of a physical map of the human genome with a resolution of 100 kb—i.e., 30,000 STSs—is well within sight. Already, more than 23,000 STSs (mapped at least to a chromosome) have been deposited in GDB† (as of June 1, 1995).

By using a variety of mapping strategies, long-range clone contiguity has been achieved for several individual chromo-

Abbreviations: HGP, Human Genome Project; NCHGR, National Center for Human Genome Research; cM, centimorgan(s); STS, sequence-tagged site; ELSI, ethics, legal, and social implications; CF, cystic fibrosis; GDB, Genome Data Base; EST, expressed sequence tag.

*STSs, or sequenced-tagged sites, are "short tracts of single-copy DNA sequence that can easily be recovered by PCR as the landmarks that define the physical map" (11).

†Information about access and use of GDB may be obtained by e-mail at help@gdb.org.

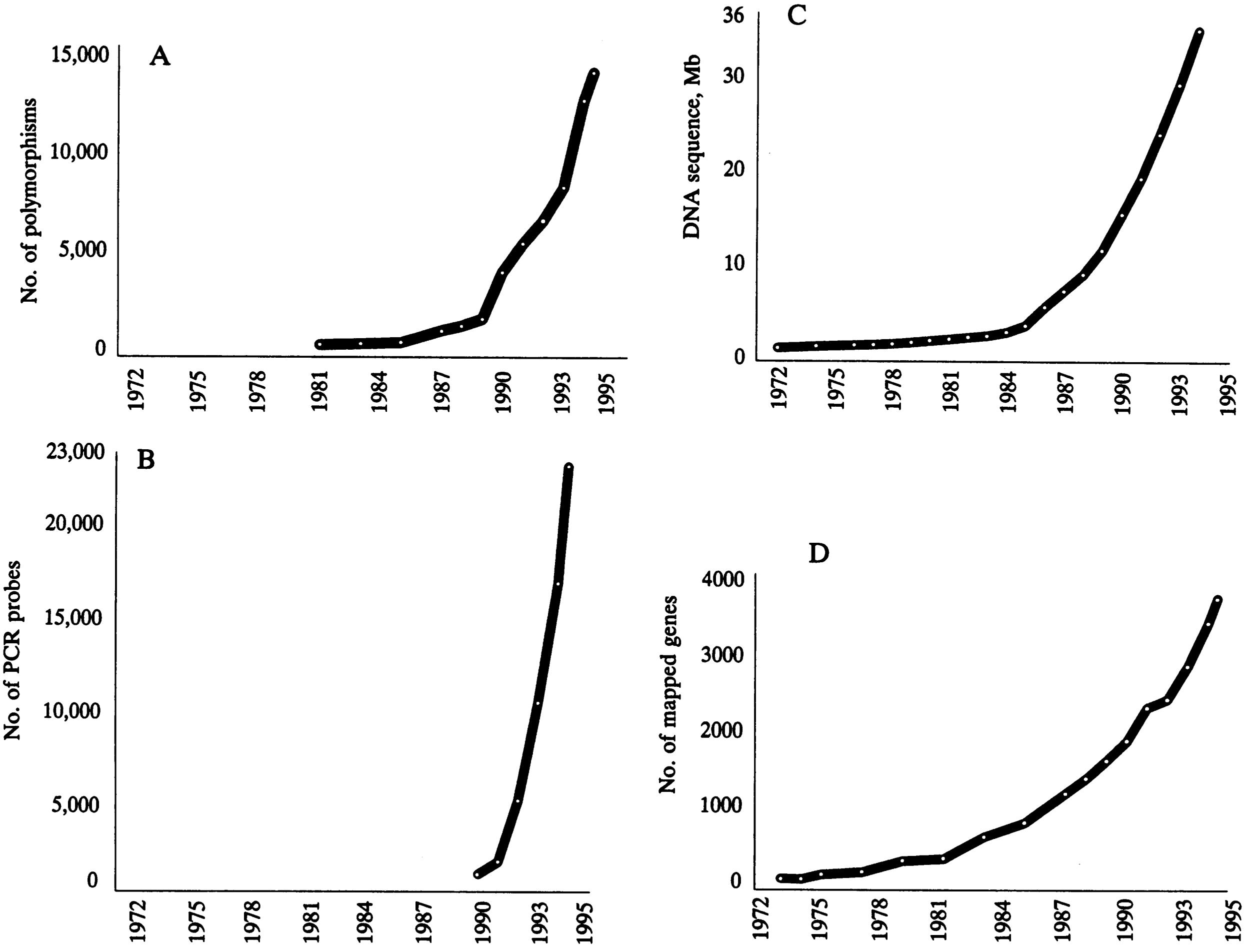

FIG. 1. Genomic information in public data bases, 1972–1995. In the cases of polymorphisms (*A*), PCR probes (*B*), and mapped genes (*D*), the number of each data type shown is the number in the Genomic Data base (GDB) as of September 30 of each year from 1990 to 1994; for 1995, the number is as of March 31. For the years 1972–1989, data were taken from the report of the DNA Committee at Human Gene Mapping Workshop 11 (1); for these years, the number of polymorphisms was approximated by the number of polymorphic D-segments, while the number of mapped genes was approximated by the number of genes. In the case of DNA sequences (*C*), the number shown is the amount of sequence in GenBank and is based on the earliest date of a published reference for each entry.

somes. Clone/STS maps of the entire euchromatic regions of chromosomes 21 (12) and Y (13) were published in 1992, and maps for several other chromosomes, including 3, 4, 7, 11, 12, 16, 19, 22, and X[‡] are nearing completion. On this set of chromosomes, between 50% and >100% of the STSs needed to achieve 100-kb average resolution have been isolated, from 5% to >75% of those STSs have been ordered, and between 70% and 80% of each of the chromosomes have been isolated in overlapping clone sets. Some very large contigs have been constructed: 50 Mb on the X chromosome, 42 Mb on chromosome 4, 34 Mb on chromosome 3, >28 Mb on the Y chromosome, 24 Mb on chromosome 11, and >20 Mb on chromosome 22 (14).

As an alternative to the chromosome-specific approaches that have led to these individual maps, a few groups have pursued genome-wide assembly strategies for constructing physical maps. In 1994, French scientists published a low-resolution physical map of the entire human genome (15) which was based on data from STS screening (STS content mapping), restriction digestion of individual yeast artificial chromosome (YAC) clones, and *Alu*-PCR fingerprinting to assemble contigs. Although much of this map needs additional confirmation, many of the smaller contigs appear to be reliable. These data have been useful in building the large, more highly characterized contigs of several of the chromosomes reported above. The genome-wide STS content mapping strategy has also been employed by a U.S. group that has recently achieved significant increases in the efficiency of STS typing and has already released data on more than 10,000 STSs.[§]

Radiation hybrid (RH) mapping is a different approach to physical mapping (16) that determines marker order and distance by a statistical analysis of marker distribution in a series of chromosomal fragments generated from one or more chromosomes by x-irradiation. Recently, the feasibility of a "whole-genome" approach to RH mapping has been demonstrated by using a set of radiation hybrids made from the entire human genome to construct maps of 100 markers on chromosome 4 and 300 markers on chromosome 7 (D. R. Cox, D. Vollrath, and R. M. Myers, personal communication).

None of the first-generation maps is error free. Errors in physical maps come from at least two sources. In most clone libraries, a fraction of the individual clones are rear-

[‡] Of these, only the chromosome 22 map has been published (14). The others are available electronically: chromosome 3, http://mars.uthscsa.edu/; chromosome 4, http://shgc.stanford.edu; chromosome 7, http://www.nchgr.nih.gov; chromosome 11, http://mcdermott.swmed.edu; chromosome 12, http://paella.med.yale.edu/chr12/Home.html; chromosome 16, http://www-ls.lanl.gov/masterhgp.html; chromosome 19, http://www-bio.llnl.gov/bbrp/genome/chrommap.html; chromosome X, http://ibc.wustl.edu:70/1/GCM.

[§] Data may be obtained by anonymous file transport protocol (ftp) (genome.wi.mit.edu), Internet e-mail (genome_database@genome.mit.edu), or World-Wide Web (http://www-genome.wi.mit.edu).

ranged relative to the genome from which the clones were derived. Furthermore, current map assembly procedures do not always produce the correct order. Some of the problems with the initial physical maps will resolve themselves; as the density of markers on the physical maps increases, many of the internal inconsistencies will become evident and will be corrected upon reexamination of the data. This has already happened in the case of one well-mapped chromosome: chromosome 21 (17).

Quality control in mapping will also be aided by the use of multiple, independent mapping methods. Comparison of independently generated maps which are based on different ordering algorithms, such as radiation hybrid maps and STS content maps, will allow differences in what should be the same map to be readily identified. In most cases, further analysis will allow the differences to be resolved and the correct map to be determined. The use of a common subset of markers in the different mapping efforts is necessary for such a comparison. Criteria for the assessment and reporting of map quality and mapping progress have been proposed by two groups (18, 19).

With approximately 20,000 of the 30,000 STSs needed to reach the original goal already in hand, an STS-based physical map of the human genome, with some regions mapped in more detail than others and an average interval of approximately 150 kb between markers, is expected to be available in the next year. A more detailed map, with STSs at an average resolution of 100 kb, was originally expected no later than the end of 1998 (4) but may be achieved sooner.

In spite of these impressive advances, further improvements in mapping technology will be essential to facilitate cost-effective construction of maps of other genomes and maps that comprise sequencing substrates. New vectors and better methods for library construction will be necessary to take full advantage of the power of genomic approaches. It is likely that physical maps of greater than 100-kb resolution (20–23) will ultimately be needed, both as substrates for DNA sequencing and for use in other types of biological research. However, it is also probable that different uses (in particular, different sequencing strategies) will have different requirements in terms of the nature and degree of detail of the very-high-resolution physical maps. Thus, it will be important for the development of new strategies and reagents for very-high-resolution physical mapping to be well integrated with the proposed use(s) of such maps.

DNA Sequencing

From the beginning, the sequencing component of the HGP seemed the most daunting. Of all the areas of genomic research in which new technology was needed, DNA sequencing was probably the one in which the greatest advances were needed. In many ways that is still true. During the initial phase of the HGP, there has been (by budget necessity) a relative underinvestment in sequencing technology in favor of map development. Nevertheless, during the past several years, substantial improvements in gel-based methods have led to a significant increase in the capacity to sequence large contiguous regions of genomic DNA.

Three paths have been taken to increase the capacity for genomic DNA sequencing. The first has attempted to maximize the capacity of current sequencing technology. A second has focused on the "evolutionary" development of new methods for high-throughput, electrophoresis-based sequencing, while a third has been directed to the development of entirely new, "revolutionary," technologies for nonelectrophoresis-based sequencing.

Improvement in the throughput of current sequencing methods has been steady over the past several years and has led to a rather striking increase in sequencing capacity. For example, over the last 2 years, the throughput of automated sequencing instruments has improved by about 2- to 3-fold. This improvement has come from expansion in the number of electrophoresis lanes per gel, lengthening of the sequencing gel, and increases in the number of daily runs of which such instruments are capable.

New sequencing strategies have also been developed. The most commonly employed strategy for large-scale sequencing remains the "shotgun" approach, in which each member of a set of randomly-generated subclones of the target region is completely sequenced and the sequence of the target region is reconstructed by a computer-based assembly process that compares the sequences of the subclones and finds the overlapping regions. However, for statistical reasons, a pure shotgun approach requires a very large amount of random sequencing (to at least a 10-fold redundancy) to obtain the complete target sequence and is, therefore, quite expensive. To reduce this cost, most sequencing efforts have employed modifications of the basic shotgun approach, involving random sequencing to a lower redundancy, partial sequence assembly, and completion by a more directed strategy. Other approaches to reducing redundancy (24, 25) promise to lead to further increases in overall efficiency.

Alternatives involving completely directed strategies for large-scale sequencing have also been developed. Transposon-based sequencing strategies (refs. 26 and 27; P. Cartwright, R. Gesteland, and R. Weiss, personal communication) which involve using mapped transposon insertions within the region to be sequenced as sites for initiating sequencing reactions promise to reduce both the amount of sequence redundancy needed to achieve closure and the difficulty of sequence assembly. New approaches to primer walking, involving the use of short oligonucleotides to construct sequencing primers *in situ* (28, 29) or inexpensive oligonucleotide synthesis (T. Brennan, D. Lashkari & R. W. Davis, personal communication) are also being tested. Yet another sequencing strategy is "multiplex" sequencing (30), which employs a different approach to sample preparation that theoretically offers improvement in efficiency in combination with either shotgun or directed approaches.

Combined with increased attention to issues of laboratory organization and management, these developments have allowed throughput of as many as several megabases of sequence per year to be achieved in a small number of laboratories. The highest output to date has come from the collaborative effort of the laboratories of R. Waterston and J. Sulston to sequence the 100-Mb genome of the nematode *Caenorhabditis elegans* (see accompanying article). Together, these two groups have already finished, and submitted to the sequence databases more than 16.5 Mb (sequence submission in GenBank) of sequence data. Other laboratories have generated the sequence of megabase regions of the DNA of other organisms, including *Haemophilus influenzae, Escherichia coli, Myobacterium leprae*, and *Drosophila melanogaster* (sequence submissions are in GenBank), while three labs have each sequenced at least 1 Mb of mammalian DNA. L. Hood's group has completed the sequence of >1 Mb of the T-cell receptor region from human and mouse DNA (ref. 31; L. Rowen, personal communication). In B. Roe's laboratory, the sequence of >1.5 Mb of human DNA from chromosomes 9 and 22 has been determined (ref. 32; B. Roe, personal communication). Finally, the sequencing group at the Sanger Center has determined the sequence of ≈1.5 Mb of human DNA, primarily in the region of the Huntington disease (HD) gene on chromosome 4.[¶]

Order-of-magnitude improvements in DNA sequencing capability are promised by the next generation of electrophoresis-based automated sequencing instruments now under development. The expectation is based on "evolutionary" advances, including new geometries for electrophoresis (such as ultrathin gels and capillaries), new matrices for DNA-fragment resolution, improved detection technologies, and improvements in process automation, all of which will lead to increased through-

[¶] A total of 0.5 Mb of sequence from the HD region and 0.1 Mb from Xq28 have been submitted to the public nucleic acid sequence data bases; additional data, representing sequence that is largely finished, is available by ftp from ftp.sanger.ac.uk pub/human/sequences or by mosaic from http://www.sanger.ac.uk.

put by increased parallelization of sample processing, increased electrophoresis speed, and decreased sample size.

In particular, the miniaturization of electrophoresis-based sequencing holds great promise. Recent work in ultrathin (capillary or slab) gel electrophoresis (33–35) has shown that the advantageous heat-transfer properties afford by the ultrathin geometry allows the greatly increased separation speed predicted for this technique. Ultrathin gels also have the advantage of allowing greatly reduced sample size. Further reductions in scale and increases in throughput can be anticipated through the application of technologies for microfabrication (36) and microelectronic mechanical systems to DNA sequencing. Considerable technological challenges remain before miniaturized systems for DNA sequencing become routinely available, but their potential is enormous.

Also under development are a number of different approaches to DNA sequencing that are "revolutionary" in that they do not rely on the electrophoresis of sets ("ladders") of DNA fragments. Mass spectrometric methods for DNA sequence analysis are being investigated in several laboratories. Although the application of this technology to large-scale genomic DNA sequence analysis is considered to be at least several years away, recent progress suggests that these approaches have a chance for ultimate success in sequencing and/or mapping applications. For example, the recent identification of a matrix, 3-hydroxypicolinic acid, that allows very reproducible volatilization of a nucleic acid sample without fragmentation (37) increases prospects for the application of matrix-assisted laser desorption/time of flight mass spectrometry in DNA sequencing. Single-base resolution of oligonucleotides in the length range of 40–50 nt has been achieved (C. Becker, personal communication); recently, results have been obtained that demonstrate resolution of the products of Sanger sequencing reactions up to 35 nt in length (C. Becker, personal communication).

Sequencing by hybridization (SBH) is another nonelectrophoresis-based approach in which the sequence of a DNA fragment is determined by hybridizing the DNA to a known set of oligomers arrayed on a surface (38, 39). In principle, the pattern of annealing to the oligonucleotide array allows the sequence of the target DNA to be determined. Challenges in developing this technique include efficient synthesis of high density arrays, better understanding of hybridization reaction kinetics, and improved detection schemes and methods for data analysis. The utility of this approach for detection of specific gene sequences, as well as single-base mutations, has recently been demonstrated (40). A particular challenge for the SBH approach is the determination of the sequence of large regions of DNA containing repeated sequences. The combination of the sequence-checking capabilities of SBH with rapid sequencing by other methods is potentially a very powerful approach to *de novo* sequencing.

Over the past 3 years, progress in sequencing technology development has been steady. Current sequencing technology appears to be capable of accomplishing the complete sequencing of the genomes of several model organisms before the end of the century. Recently, it has been argued that current technology is even capable of being used to generate a first-pass sequence of the human genome in a reasonable period of time and for a reasonable cost (41). This presents a crucial strategic question to the HGP. Should the complete genomic sequencing of human DNA be begun with current technology (and whatever technological improvements occur in the course of doing so) or should "production sequencing" be delayed for some time in favor of continued emphasis on development of further improvements in sequencing technology?

This is not a question of whether continued development of improved technology for DNA sequencing is important. Vast reductions in DNA sequencing costs and increases in DNA sequencing rates will be required to elucidate the enormous amount of biological information about individual traits that will come from making correlations between the variation in particular traits among many individuals and differences among those individuals in the sequence of relevant, large (megabase) regions of the genome. Similarly, an untold amount of useful information will be obtained by knowing the sequences of the genomes of many other organisms, ranging from infectious microorganisms to agriculturally important plants. Technology far beyond today's will be required to achieve the necessary sequencing capability. Therefore, even if revolutionary new sequencing technologies were unlikely to be sufficiently developed in the next few years to contribute to the first human genome sequence, their continued development will continue to be an important activity of the HGP. Rather, the questions currently demanding answers are how best to stimulate the development of new sequencing technology and whether acquisition of large amounts of human DNA sequence sooner rather than later will be a stimulus or deterrent to further technology development.

Another dilemma facing the HGP involves attracting new groups to large-scale sequencing. It is important to involve as many groups as possible in the development of production-sequencing capability to provide competition and innovative approaches to this problem. At the same time, doing so has a cost, the cost of the learning curve. It has been the experience in large genome centers that efficiency increases with experience. Among the groups that have focused their attention on production, the efficiency of data generation has increased in the course of confronting and solving the problems that have arisen as their production capacity (for either mapping or sequencing) has increased. There remains a question about how efficiency scales with size. Will efficiency continue to increase as the group size increases, or is there an optimal group size in terms of efficiency (Fig. 2)? At present, we do not have a good sense of what the trade-offs and costs will be in paying the costs for new groups to move up the learning curve compared with increasing the size of the few existing groups.

Gene Identification

Starting from contigs of large-insert clones, a number of gene isolation and localization techniques, such as cDNA selection and enrichment methods (42) and exon amplification methods (43–45), have been used in the positional cloning of a number of human genes. So far, however, none of the methods has been shown to be sufficiently robust to support large-scale gene identification; in Fig. 1*D*, for example, it can be seen that the rate of identification of new genes is still relatively low. Thus, there is not yet confidence that the cataloging and mapping of the complete collection of human genes can be done with the kind of completeness, efficiency, and cost-effectiveness demanded by the HGP (46). An optimal technology or combination of technologies for identifying genes in large cloned regions remains to be formulated.

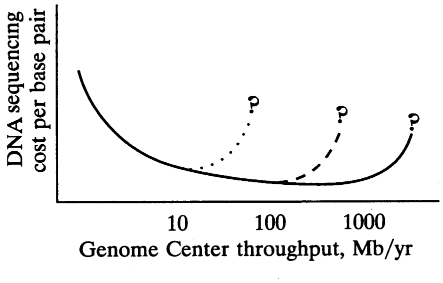

FIG. 2. Hypothetical curves of cost-efficiency in DNA sequencing. Current experience indicates that economies of scale reduce the cost per base pair of sequencing up to a production level of 10 Mb per year; however, no group has thus far pushed beyond this level, so we do not know at what point such economies become overwhelmed by the managerial problems of running a very large-scale sequencing center. Three possible curves are shown, without any information at present to determine which will apply to large-scale mammalian DNA sequencing. The ultimate division of labor in accomplishing the sequencing component of the HGP will depend critically on which of these curves applies, as that is likely to play a major role in determining the optimal number of centers to complete the sequence for the most reasonable cost.

The term "expressed sequence tag" (EST) has been used to refer to a short DNA sequence derived from a cDNA clone and used to identify a genomic coding sequence (47). A considerable number of ESTs have been collected in a public repository, dbEST (188,122 as of June 15, 1995; ref. 48), and a much larger number will become publicly available by the end of 1995 through an effort supported by the pharmaceutical company Merck; over 100,000 ESTs from this effort have been put into dbEST in the first few months of operation. STSs derived from these ESTs will rapidly be incorporated into physical maps through the efforts of an international mapping consortium (49). These projects lend confidence that the slope of the curve in Fig. 1*D* will soon increase significantly and that half or more of the genes in the human genome will be localized on physical maps within the next few years. However, the actual number of genes that can be identified and mapped by this approach cannot be predicted, nor can the proportion of the total number of human genes they will represent. The number of genes that can be identified in this way depends entirely on the quality of the cDNA libraries from which the ESTs are obtained, and cDNA libraries that represent the complete set of human genes have not been constructed.

Medical Applications

The term "gene identification" is also used more broadly to describe the isolation of the gene(s) underlying a particular phenotype or trait, such as a disease. In this sense, gene identification is a biological problem with considerable medical significance that HGP resources are already helping to solve. It is one measure of the success of the HGP to date that, through the use of these new tools, the speed with which disease genes are being found is rapidly increasing. New disease genes are being found and reported at a rate of several per month, compared with a few per year not so long ago (50).

During the past 2 years, many genes for human diseases have been isolated by using the powerful methods of positional cloning (51), where no prior knowledge of the gene's function is available. In contrast to positional cloning, functional cloning or candidate-gene strategies make use of information that is known about the gene product and/or the function of the gene of interest that suggest it as a particularly good candidate for the affected gene in a particular disease. For example, keratins are the major class of proteins found in skin cells. The effects of mutations on the properties of certain keratin proteins and the observation that mice in which such mutant genes are expressed exhibit features of epidermolysis bullosa simplex (EBS) led to the discovery that mutations in the human K5 and K14 keratin genes are responsible for the severest forms of EBS in humans (52).

The candidate-gene methods have also taken advantage of the improvement in genomic maps, resources, and technologies, and, as genomic maps improve and become more populated with gene sequences, a new strategy is emerging. Known as "positional candidate" cloning (50, 53), this approach begins with mapping the disease gene to a small chromosomal interval. All genes located in that interval can then be tested for their involvement, starting with any whose product functions in a way that suggests possible involvement. In other words, a gene becomes a candidate not only by virtue of the properties of its protein product but by its map location as well.

As more human genes have been identified and cloned, new insights are already being gained into human biology and disease. Some of these might not be surprising in that they show that many aspects of human molecular genetics are familiar and similar to those seen in more well-studied organisms. For instance, as elaborated below, different phenotypes have now been found to arise from different mutations in a single human gene, and similar phenotypes have been found to arise from mutations in different (related) genes. Yet, this similarity is itself important because it confirms once more that many of the lessons we have learned about the biology of nonhuman organisms will be directly applicable to our understanding of human biology.

Several human genes have been analyzed sufficiently thoroughly that analysis of mutational spectra has been possible. Mutations in the receptor tyrosine kinase gene *RET* can give rise to any of four different syndromes: familial medullary thyroid carcinoma (54), multiple endocrine neoplasia types 2A (54, 55) and 2B (56), and Hirschsprung disease (57, 58). There is no reported cancer association with the latter condition. Similarly, while a large number of mutations in the *CFTR* gene cause cystic fibrosis (CF) (of varying degrees of severity) (59), mutations in *CFTR* have also been found in patients who do not exhibit the standard pulmonary symptoms of CF. A significant fraction of otherwise healthy males who are infertile due to a congenital bilateral absence of the vas deferens (CBAVD) carry one or two known *CFTR* mutations (60), although recent evidence indicates that the CBAVD condition itself is genetically heterogeneous (61).

Genetic analyses of inherited conditions are also beginning to give the anticipated leads into better understanding of human biology. A number of examples illustrate that mutations in members of a gene family can give rise to a set of related pathologies. For instance, several rare and often diagnostically confusing craniosynostosis syndromes [Crouzon syndrome (62–64), Jackson-Weiss syndrome (63), Pfeiffer syndrome (65–68), Apert syndrome (64), and achondroplasia (69)] have been found to be caused by mutations in one of several fibroblast growth factor receptor genes. Similarly, as noted above, several blistering skin diseases are caused by mutations in different keratin genes. Other skin diseases, however, affecting different cell types within the epidermis, are caused by mutations in genes representing other gene families (70, 71).

Kallmann syndrome is a condition characterized by an unusual constellation of seemingly unrelated features: hypogonadism resulting from an endocrinological problem (deficiency of hypothalamic gonadotropin-releasing hormone) and anosmia (inability to smell). Isolation of the *KAL* gene (72, 73) led to the recognition that the predicted *KAL* gene product was related to neural-adhesion molecules. This is consistent with a role of the gene product in embryonic neural migration and with the observation of a common developmental pathway of two types of neurons that originate in the olfactory placode: the olfactory neurons and the neurons that produce gonadotropin-releasing hormone.

Analysis of other genetic diseases in humans has led to the identification of a previously unobserved mutational mechanism. Fragile X syndrome (74), spinal and bulbar muscular atropy (75), myotonic dystrophy (76–78), Huntington disease (79), and several others have been found to be caused by trinucleotide-expansion mutations in which the number of copies of a trinucleotide repeat sequence is significantly increased in affected compared with nonaffected individuals.

Charcot–Marie–Tooth disease type 1A (CMT1A) represents another unique (to date) genetic situation, in that it most frequently results from duplication of the region of chromosome 17 containing the *PMP22* gene (80). Hypotheses suggesting that this is a gene dosage effect are being tested. Point mutations in *PMP22* have also been found in patients with CMT1A (80) and, as in the case of *RET* and *CFTR*, other mutations in *PMP22* have been found to be responsible for other clinical conditions: Dejerine–Sottas syndrome (81) and, possibly, hereditary neuropathy with liability to pressure palsies (82).

Informatics

With respect to genomics, informatics issues range from the algorithmic and data-management needs of mapping and sequencing projects to the dissemination of data to the scientific community, and the development of adequate informatics approaches to this variety of issues has been an important area of research since the inception of the HGP. Data analysis and data management are actually open-ended problems in that today's solutions will, in general, have a limited lifetime as new approaches to genomic analysis and more data are developed. Nonetheless, much progress has been made in genome informatics in the past 5 years. A considerable number of individual laboratories have developed and are using computer-based systems for automating the acquisition, management, analysis, and distribution of experimental data. Similarly, a number of new data bases have been created for the purpose of allowing the rapid distribution of the findings of genome research.

In fact, the number of data sources and programs of interest is too large to summarize in this article, but information about many may be obtained electronically from http://www.nchgr.nih.gov. The current state of genome informatics is driven by the rapidly expanding capabilities afforded by the increased capacity of computers and networks. Through the use of ftp sites and the World Wide Web, for example, many genome centers provide electronic access to much larger amounts of data and less refined data than could be published. While this system has the advantage of allowing those who generate genome data to take the responsibility for its rapid provision to the research community, it can sometimes appear to be overly anarchic. More seriously, it can pose some serious challenges to the users of the data as they attempt to locate diverse information sources and integrate data in diverse formats. Recently, both the central data bases and independent software developers have made significant progress on data-access systems and tools.

Ethical, Legal, and Social Implications (ELSI)

It was recognized almost from the outset of the HGP that the increased knowledge of human biology and human disease that would come from application of genomic information would raise a number of substantive ethical and policy issues for individuals and for society. Accordingly, an ELSI research program has been an integral element of genome programs around the world. In the U.S., the ELSI research program has focused on identifying and addressing a few high priority areas raised by the most immediate potential applications or consequences of genome research. These are ethical issues surrounding genetics research, responsible clinical integration of new genetic technologies, privacy and fair use of genetic information, and professional and public education about these issues.

Research and education projects have been funded in each of these priority areas. In pursuing these goals, there has been an emphasis on the integration of sound scientific knowledge with an understanding of the many historical, ethnocultural, social, and psychological factors that influence policy development and service delivery in human genetics. The first of these projects are now beginning to be completed, and the research results are leading to the identification of policy options intended to ensure that genetic information is used for the benefit of individuals and society.

One of the first examples is a coordinated set of studies that examined issues surrounding genetic testing and counseling of patients for CF mutations. The investigators participating in these studies were organized into a consortium to conduct the studies more efficiently and productively, to allow a broader range of issues to be addressed, and to pursue a coordinated approach to the development of policy recommendations related to CF testing. These studies found that interest in the general population for CF testing was much lower than anticipated. The demand for CF testing was found to be influenced by a number of variables, including the cost of the test, the timing, the setting, and even the manner in which the testing was offered. It was also found that, although people can be educated about CF testing (with the goal of allowing them to make informed choices about testing), doing so is difficult because of the complexity of the information to be communicated, the lack of incentives on the part of some providers to teach people about the information, and the lack of motivation on the part of consumers to learn it. The results of this research have also led to proposals about optimal methods to provide CF testing to those who desire it. These and other results of the studies have been discussed in professional society meetings, and it is anticipated that clinical policy recommendations will emerge from these organizations.

A second major effort in the area of introduction of genetic tests was initiated in the autumn of 1994 and consists of a set of projects to examine issues surrounding testing and counseling for heritable breast, ovarian, and colon cancer risks. Questions to be addressed in these studies include the interest and demand for testing, the impact of testing, and alternative ways to provide such services. As with the CF studies, the investigators involved in these projects will form a consortium to pool resources, reduce duplication of effort, and increase coordination of some aspects of the studies. By pooling initial findings where possible and following the results of this set of research projects, it may be feasible to identify emerging themes and develop some policy recommendations sooner than if the projects were conducted and assessed independently.

In another approach to addressing issues associated with the use and regulation of new genetic tests, the joint NIH/DOE ELSI Working Group[‖] has created a Genetic Testing Task Force that will review the "state of the art" of genetic testing, examine the strengths and weaknesses of current practices and policies relating to testing, and, if needed, recommend changes or policy options that will ensure that the public is protected, such that only the appropriate tests are done, and these by qualified laboratories. The Task Force will report its findings through the ELSI Working Group and the National Advisory Council for Human Genome Research to the Directors of the NCHGR and the NIH and eventually to the Secretary of the Department of Health and Human Services.

Finally, in 1994, the Institute of Medicine published a study of professional policy issues in the clinical integration of new genetic tests (83). This report offers a number of recommendations for the laboratory quality control of DNA diagnostics and the provision of genetic testing in the clinical setting.

With the advent of new technologies to identify more genetic information about individuals, concerns arise about the privacy and fair use of the information. One of the initial products of the discovery of a variant gene that underlies a particular condition is information that can be useful in predicting the likelihood that an individual will develop that condition. This can potentially serve the individual well by opening the door to preventative interventions, including more informed presymptomatic screening or lifestyle changes involving diet, exercise patterns, or environment. At the same time, however, this information may also have unwelcome effects, such as increased anxiety, altered family relationships, stigmatization, and discrimination on the basis of genotype. Concerns about stigmatization and discrimination are cited as particularly troubling, especially in relation to employability and insurability. The Task Force on Genetic Information and Insurance, established by the joint NIH/DOE ELSI Working Group, assessed the potential impact of advances in human genetics on the current system of health-care coverage in the U.S. The Task Force report,[**] issued in 1993, contained recommendations for managing that impact within a reformed health-care system. Model legislation dealing with genetic privacy has been drafted by G. Annas, L. Glantz, and P. Roche (G. Annas, personal

[‖] The joint NIH/DOE ELSI Working Group was established to explore and propose options for the development of sound professional and public policies related to human genome research and its applications.

[**] *Genetic Information and Health Insurance: Report of the Task Force on Genetic Information and Insurance*, is available from the ELSI Branch, NCHGR, Building 38A, Room 613, NIH, Bethesda, MD 20817.

communication)[††] and is being reviewed at both the federal and state levels.

A recent gratifying development is the ruling by the U.S. Equal Employment Opportunities Commission (EEOC) that genetic discrimination in employment decisions is illegal. The EEOC has formally ruled that the Americans with Disabilities Act (ADA), passed in 1991, does extend coverage to individuals who are at risk for future illness and who are discriminated against in employment decisions on that basis. The argument is that these individuals are 'regarded as' having a disability by their employer, even though they are currently well, and therefore the protections of the ADA should be extended to them. The need to provide legal protections against genetic discrimination in hiring practices was considered a high priority by the ELSI Working Group from its very first deliberations, and the resolution of this problem by the EEOC is seen by many as a landmark development. Its significance even extends beyond employment discrimination because it establishes the general principle that discrimination against individuals on the basis of their genetic inheritance is unjust.

While much research in human genetics is focused on the causes of human disease and disability, genes and genetic markers that appear to be associated with other human characteristics are being reported on an increasingly frequent basis. Reports that associate genes with human traits that exhibit a wide range of variation in the general population, from stature, weight, and metabolism to learning ability, behavior, and sexual orientation, raise very different and potentially controversial social issues (84, 85). Typically, such genetic studies provide only introductory and incomplete clues about the interplay of biological, psychological, and sociocultural factors that influence the development and expression of these complex human traits. However, the results of such research can be misinterpreted in two important ways. First, they can be interpreted to imply that such traits can be reduced to the expression of particular genes; this has the effect of deemphasizing the important role of psychosocial and other environmental factors. Second, the results can also be interpreted in a way that narrows the range of variation considered to be "normal" or "healthy." Such overly deterministic interpretations can, in turn, be misused to undercut the respect owed to individuals as responsible moral agents or to inappropriately label individuals as sick or abnormal. Both forms of misinterpretation can have untoward consequences, such as devaluing human genetic diversity or fostering social discrimination on the basis of genotype. As it proceeds, the HGP will need to foster a better understanding of the meaning of human genetic variation among members of the public and the professions and expand its efforts to propose policy initiatives designed to prevent genetic stigmatization, discrimination, and other misinterpretation or misuses of genetic information.

Conclusion

The beginning phase of the HGP has been remarkably successful. The amount of genome data describing human DNA and the DNA of other organisms that is available in public data bases has increased enormously, and the information is being used at an increasing rate. The contributions that the HGP has already made to advance the study of inherited disease and other biological phenomena are, by now, widely recognized in the scientific community. The community is no longer arguing whether the HGP is a good idea but is now debating the most effective ways to reap its rewards.

These achievements are attributable to several decisions that were made early in the life of the program. One was to base the implementation of the program on a visionary, flexible, and on-going planning process. Over the years, the program has benefitted from advice from a large number of members of the scientific community, both from those directly involved in genome research and from those who are not but who will be the ultimate users of the products of the HGP. Second, the HGP has been highly dependent upon the introduction of new technology for mapping and sequencing, the development of which has been emphasized as part of the research program from the beginning. Further technology development will continue to be critical, not only to achieving the long-range goals of the HGP (which are, in fact, rather circumscribed), but also to the ability to take advantage of the (virtually unlimited) promise offered by the sequence-based approach to biological research that will be opened up by the HGP. Third, large research centers, both publicly and privately financed, have moved rapidly along the "learning curve" and have achieved significant efficiencies of scale that have allowed them to generate large amounts of data in relatively short periods of time. Fourth, the consensus that genomic information should be publicly available has led to policies on the part of the funding agencies and practices on the part of the genome research community that have resulted in an exemplary record for rapid publication of the data in accessible formats with the result that this valuable information is available to the scientific community much more rapidly than through conventional publication means. Finally, the international aspects of the HGP have been a demonstration of the power of an international collaboration in achieving a very ambitious set of goals.

We are rapidly approaching the time when we will have the initial products of the HGP, including maps, genomic DNA sequences, and the improved technology for genomic analysis, in hand. That will represent the true point of initiation for the era of sequence-based biological investigation.

Note Added in Proof. After this manuscript was submitted, several significant publications appeared addressing issues discussed in this paper. These included YAC contig maps of human chromosomes 3 (86), 12 (87), and 22 (88); an integrated physical map of chromosome 16(89); and an improved version (90) of the whole genome map originally published by the Centre d'Étude du Polymorphisme Humain–Genethon group (15). Investigators from The Institute of Genome Research (TIGR) and Human Genome Sciences, Inc., reported a set of 174,472 ESTs which, when combined with other 188,406 ESTs from the public data base dbEST (48), allowed the definition of 87,983 distinct sequences (91), consisting of 29,599 merged sequences and 58,384 nonoverlapping ESTs. Finally, investigators from TIGR also published the complete DNA sequence of *Haemophilus influenzae* (92).

The ideas discussed in this paper have benefitted from the long and productive collaboration that we have had with our colleagues at the NCHGR. We particularly want to thank Elke Jordan, David Benton, Carol Dahl, Bettie Graham, Jane Peterson, and Bob Strausberg for their helpful critiques and suggestions during preparation of the manuscript.

1. Williamson, R., Bowcock, A., Kidd, K., Pearson, P., Schmidtke, J., Ceverha, P., Chipperfield, M., Cooper, D. N., Coutelle, C., Hewitt, J., Klinger, K., Beckmann, J., Tolley, M. & Maidak, B. (1991) *Cytogenet. Cell Genet.* **58,** 1190–1832.
2. U.S. Department of Health and Human Services and U.S. Department of Energy. (1990) *Understanding Our Genetic Inheritance: The U.S. Human Genome Project: The First Five Years* (National Institutes of Health, Bethesda), NIH Publ. No. 90-1590.
3. Commission of Life Sciences, National Research Council (1988) *Mapping and Sequencing the Human Genome* (Natl. Acad. Press, Washington, DC).
4. Collins, F. & Galas, D. (1993) *Science* **262,** 43–46.
5. Cooperative Human Linkage Center, Genethon, University of Utah, Yale University, and Centre d'Étude du Polymorphisme Humain (1994) *Science* **265,** 2049–2054.
6. Weber, J. L. & May, P. E. (1989) *Am. J. Hum. Genet.* **44,** 388–396.
7. Litt, M. & Luty, J. A. (1989) *Am. J. Hum. Genet.* **44,** 397–401.
8. Gyapay, G., Morisette, J., Vignal, A., Dib, C., Fizames, C., Millasseau, P., Marc, S., Bernardi, G., Lathrop, M. & Weissenbach, J. (1994) *Nat. Genet.* **7,** 246–339.
9. Buetow, K. H., Weber, J. L., Ludwigsen, S., Scherpbier-Heddema, T., Duyk, G. M., Sheffield, V. C., Wang, Z. & Murray, J. C. (1994) *Nat. Genet.* **6,** 391–393.
10. Hauge, X. Y. & Litt, M. (1993) *Hum. Mol. Genet.* **2,** 411–415.
11. Olson, M., Hood, L., Cantor, C. & Botstein, D. (1989) *Science* **245,** 1434–1435.
12. Chumakov, I., Rigault, P., Guillou, S., Ougen, P., Billaut, A., *et al.* (1992) *Nature (London)* **359,** 380–387.

[††] Available from http://www-busph.bu.edu/Depts/HealthLaw/.

13. Foote, S., Vollrath, D., Hilton, A. & Page, D. (1992) *Science* **258,** 60–66.
14. Bell, C. J., Budarf, M. L., Nieuwenhuijsen, B., Barnoski, B., Buetow, K., *et al.* (1995) *Hum. Mol. Genet.* **4,** 59–69.
15. Cohen, D., Chumakov, I. & Weissenbach, J. (1993) *Nature (London)* **359,** 698–701.
16. Cox, D. R., Burmeister, M., Price, E. R., Kim, S. & Myers, R. M. (1990) *Science* **250,** 245–250.
17. Delabar, J.-M., Créau, N., Sinet, P.-M., Ritter, O., Antonarakis, S., Burmeister, M., Chakravarti, A., Nizetic, D., Ohki, M., Patterson, D., Petersen, M., Reeves, R. & Van Broeckhoven, C. (1993) *Genomics* **18,** 735–745.
18. Olson, M. V. & Green, P. (1993) *Cold Spring Harbor Symp. Quant. Biol.* **53,** 349–355.
19. Cox, D. R., Green, E. D., Lander, E. S., Cohen, D. & Myers, R. M. (1994) *Science* **265,** 2031–2032.
20. Smith, M. W., Holmsen, A. L., Wei, Y. H., Peterson, M. & Evans, G. A. (1994) *Nat. Genet.* **7,** 40–47.
21. Fischer, S. G., Cayanis, E., Russo, J. J., Sunjevaric, I., Boukhgalter, B., Zhang, P., Yu, M.-T., Rothstein, R., Warburton, D., Edelman, I. & Efstradiatis, A. (1994) *Genomics* **21,** 525–537.
22. Rouquier, S., Batzer, M. A. & Giorgi, D. (1994) *Anal. Biochem.* **217,** 205–209.
23. Meng, X., Benson, K., Chada, K., Huff, E. J. & Schwartz, D. C. (1995) *Nat. Genet.* **9,** 432–438.
24. Chen, E. Y., Schlessinger, D. & Kere, J. (1993) *Genomics* **17,** 651–656.
25. Roach, J. C., Boysen, C., Wang, K. & Hood, L. (1995) *Genomics* **26,** 345–353.
26. Yoshida, K., Strathmann, M. P., Mayeda, C. A., Martin, C. H. & Palazzolo, M. J. (1993) *Nucleic Acids Res.* **21,** 3553–3562.
27. Cherry, J. L., Young, H., DiSera, L. J., Ferguson, F. M., Kimball, A. W., Dunn, D. M., Gesteland, R. F. & Weiss, R. B. (1994) *Genomics* **20,** 68–74.
28. Kaczorowski, T. & Szybalski, W. (1993) *Gene* **135,** 286–290.
29. Kieleczawa, J., Dunn, J. J. & Studier, F. W. (1992) *Science* **258,** 1787–1791.
30. Church, G. M. & Kiefer-Higgins, S. (1988) *Science* **240,** 185–188.
31. Koop, B. F., Rowen, L., Wang, K., Kuo, C. L., Seto, D., Lenstra, J. A., Howard, S., Shan, W., Deshpande, P. & Hood, L. (1994) *Genomics* **19,** 478–493.
32. Chissoe, S. L., Bodenteich, A., Wang, Y.-F., Wang, Y.-P., Burian, D., *et al.* (1995) *Genomics,* **27,** 67–82.
33. Khrapko, K., Hanekamp, J. S., Thilly, W. G., Belenkii, A., Foret, F. & Karger, B. (1994) *Nucleic Acids Res.* **22,** 364–369.
34. Kostichka, A. J., Marchbanks, M., Brumley, R. L., Drossman, H. & Smith, L. M., (1992) *BioTechnology* **10,** 78–81.
35. Brumley, R. L., Jr., & Smith, L. M. (1991) *Nucleic Acids Res.* **19,** 4121–4126.
36. Woolley, A. T. & Mathies, R. A. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 11348–11352.
37. Wu, K. J., Shaler, T. A. & Becker, C. H. (1994) *Anal. Chem.* **66,** 1637–1645.
38. Khrapko, K. R., Lysov, Y. P., Khorlyn, A. A., Florentiev, V. L. & Mirzabekov, A. D. (1989) *FEBS Lett.* **256,** 118–122.
39. Drmanac, R., Labat, L., Brukner, I. & Crkvenjakov, R. (1989) *Genomics* **4,** 114–128.
40. Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P. & Fodor, S. P. A. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 5022–5026.
41. Marshall, E. (1995) *Science* **267,** 783–784.
42. Lovett, M. (1994) *Trends Genet.* **10,** 352–357.
43. Duyk, G. M., Kim, S. W., Myers, R. M. & Cox, D. R. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 8995–8999.
44. Church, D. M., Stotler, C. J., Rutter, J. L., Murrell, J. R., Trofatter, J. A. & Buckler, A. J. (1994) *Nat. Genet.* **6,** 98–105.
45. Krizman, D. & Berget, S. (1994) *Nucleic Acids Res.* **21,** 5198–5202.
46. Brennan, M. B. & Hochgeschwender, U. (1995) *Hum. Mol. Genet.* **4,** 153–156.
47. Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R. & Venter, J. C. (1991) *Science* **252,** 1651–1656.
48. Boguski, M. S., Lowe, T. M. J. & Tolstoshev, C. (1993) *Nat. Genet.* **4,** 332–333.
49. Dickson, D. (1994) *Nature (London)* **371,** 551.
50. Collins, F. S. (1995) *Nat. Genet.* **9,** 347–350.
51. Collins, F. S. (1992) *Nat. Genet.* **1,** 3–6.
52. Fuchs, E. & Coulombe, P. A. (1992) *Cell* **69,** 899–902.
53. Ballabio, A. (1993) *Nat. Genet.* **3,** 277–279.
54. Donis-Keller, H., Dou, S., Chi, D., Carlson, K., Toshima, K., Lairmore, T., Howe, J., Moley, J., Goodfellow, P. & Well, S., Jr. (1993) *Hum. Mol. Genet.* **2,** 851–856.
55. Mulligan, L. M., Kwok, J., Healey, C., Elsdon, M., Eng, C., Gardner, E., Love, D., Mole, S., Moore, J., Papi, L., Ponder, M., Telenius, H., Tunnacliffe, A. & Ponder, B. (1993) *Nature (London)* **363,** 458–460.
56. Hofstra, R., Landsvater, R., Ceccherini, I., Stulp, R., Stelwagen, T., Luo, Y., Pasini, B., Hoppener, J., Ploos van Amstel, H., Romeo, G., Lips, C. & Buys, C. (1994) *Nature (London)* **367,** 375–376.
57. Romeo, G., Ronchetto, P., Luo, Y., Barone, V., Seri, M., Ceccherini, I., Pasini, B., Bocciardi, R., Lerone, M., Kaariainen, H. & Martucciello, C. (1994) *Nature (London)* **367,** 377–378.
58. Edery, P., Lyonnet, S., Mulligan, L., Pelet, A., Dow, E., Abel, L., Holder, S., Nihoul-Fekete, C., Ponder, B. & Munnich, A. (1994) *Nature (London)* **367,** 378–380.
59. Tsui, L.-C. (1992) *Trends Genet.* **8,** 392–398.
60. Anguiano, A., Oates, R. D., Amos, J. A., Dean, M., Gerrard, B., Stewart, C., Maher, T. A., White, M. B. & Milunsky, A. (1992) *J. Am. Med. Assoc.* **267,** 1794–1797.
61. Rave-Harel, N., Madgar, I., Goshen, R., Nissim-Rafinia, M., Ziadni, A., Rahat, A., Chiba, O., Kalman, Y. M., Brautbar, C., Levinson, D., Augarten, A., Kerem, A. & Kerem, B. (1995) *Am. J. Hum. Genet.* **56,** 1359–1366.
62. Reardon, W., Winter, R. M., Rutland, P., Pulleyn, L. J., Jones, B. M. & Malcolm, S. (1994) *Nat. Genet.* **8,** 98–103.
63. Jabs, E. W., Li, X., Scott, A. F., Chen, W., Eccles, M., Mao, J.-i., Charnas, L. R., Jackson, C. E. & Jaye, M. (1994) *Nat. Genet.* **8,** 275–279.
64. Wilkie, A. O. M., Slaney, S. F., Oldridge, M., Poole, M. D., Ashworth, G. J., Hockely, A. D., Hayward, R. D., David, D. J., Pulleyn, L. J., Rutland, P., Malcolm, S., Winter, R. M. & Reardon, W. (1995) *Nat. Genet.* **9,** 165–171.
65. Rutland, P., Pulleyn, L. J., Reardon, W., Baraitser, M., Hayward, R., Jones, B., Malcolm, S., Winter, R. M., Oldridge, M., Slaney, S. F., Poole, M. D. & Wilkie, A. O. M. (1995) *Nat. Genet.* **9,** 173–176.
66. Muenke, M., Schell, U., Hehr, A., Robin, N. H., Losken, H. W., Schinzel, A., Pulleyn, L. J., Rutland, P., Reardon, W., Malcolm, S. & Winter, R. (1994) *Nat. Genet.* **8,** 269–274.
67. Schell, U., Hehr, A., Feldman, G. J., Robin, M. H., Zackai, E. H., de Die-Smulders, C., Viskochil, D. H., Stewart, J. M., Wolff, G., Ohashi, H., Price, R. A., Cohen, M. M., Jr., & Muenke, M. (1995) *Hum. Mol. Genet.* **4,** 323–328.
68. Lajeunie, E., Ma, H. W., Bonaventure, J., Munnich, A., Le Merrer, M. & Renier, D. (1995) *Nat. Genet.* **9,** 108.
69. Shiang, R., Thompson, L., Zhu, Y.-Z., Church, D. M., Fielder, T. J., Bocian, M., Winokur, S. T. & Wasmuth, J. J. (1994) *Cell* **78,** 335–342.
70. Carroll, J. M. & Goldsmith, L. A. (1995) *Mol. Med.* **1,** 123–126.
71. Russell, L. J., DiGiovanna, J. J., Rogers, G. R., Steinert, P. M., Hashem, N., Compton, J. G. & Bale, S. J. (1995) *Nat. Genet.* **9,** 279–283.
72. Legouis, R., Hardelin, J.-P., Levilliers, J., Claverie, J.-M., Compain, S., Wunderle, V., Millasseau, P., Le Pasilier, D., Cohen, D., Caterina, D., Bougueleret, L., Delemarre-Van de Waal, H., Lutfalla, G., Weissenbach, J. & Petit, C. (1991) *Cell* **67,** 423–435.
73. Franco, B., Guoli, S., Pragliola, A., Incerti, B., Bardoni, B., Tonlorenzi, R., Carrozzo, R., Maestrini, E., Pieretti, M., Taillon-Miller, P., Brown, C. J., Willard, H. F., Lawrence, C., Persico, M. G., Camerino, G. & Ballabio, A. (1991) *Nature (London)* **353,** 529–536.
74. Verkerk, A. J. M. H., Pieretti, M., Sutcliffe, J. S., Fu, Y.-H., Kuhl, D. P. A., Pizzuti, A., Reiner, O., Richards, S., Victoria, M. F., Zhang, F., Eussen, B. E., van Ommen, G.-J. B., Bionden, L. A. J., Riggins, G. J., Chastain, J. L., Kunst, C. B., Galjaard, H., Caskey, C. T., Nelson, D. L., Oostra, B. A. & Warren, S. T. (1991) *Cell* **65,** 905–914.
75. LaSpada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E. & Fischbeck, K. H. (1991) *Nature (London)* **352,** 77–79.
76. Brook, J. D., McCurrach, M. E., Harley, H. G., Buckler, A. J., Church, D., Aburatani, H., Hunter, K., Stanton, V. P., Thirion, J.-P., Hudson, T., Sohn, R., Zemelman, B., Snell, R. G., Rundle, S. A., Crow, S., Davies, J., Shelbourne, P., Buxton, J., Jones, C., Juvonen, J., Johnson, K., Harper, P. S., Shaw, D. J. & Housman, D. E. (1992) *Cell* **68,** 799–808.
77. Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barcelo, J., O'Hoy, K., Leblond, S., Earle-MacDonald, J., de Jong, P. J., Wierenga, B. & Korneluck, R. G. (1992) *Science* **255,** 1253–1255.
78. Fu, Y.-H., Pizzuti, A., Fenwick, R. G., Jr., King, J., Rajnarayan, S., Dunne, P. W., Dubel, J., Nasser, G. A., Ashizawa, T., de Jong, P., Wieringa, B., Korneluk, R., Perryman, M. B., Epstein, H. F. & Caskey, C. T. (1992) *Science* **255,** 1256–1258.
79. The Huntington's Disease Collaborative Research Group (1993) *Cell* **72,** 971–983.
80. Patel, P. I. & Lupski, J. R. (1994) *Trends Genet.* **10,** 128–133.
81. Roa, B. B., Dyck, P. J., Marks, H. G., Chance, P. F. & Lupski, J. R. (1993) *Nat. Genet.* **5,** 269–273.
82. Chance, P. F., Anderson, M. K., Leppig, K. A., Lensch, M. W., Matsunami, N., Smith, B., Swanson, P. D., Odelberg, S. J., Disteche, C. M. & Bird, T. D. (1993) *Cell* **72,** 143–151.
83. Committee on Assessing Genetic Risks, Division of Health Sciences Policy, Institute of Medicine (1994) *Assessing Genetic Risks: Implications for Health Policy* (Natl. Acad. Press, Washington, DC).
84. Hamer, D. H., Hu, S., Magnuson, V. L., Hu, N. & Pattatucci, A. M. L. (1993) *Science* **261,** 321–327.
85. Brunner, H. G., Nelen, M., Breakefield, X. O., Ropers, H. H. & van Oost, B. A. (1993) *Science* **262,** 578–580.
86. Gemmill, R. M., Chumakov, I., Scott, P., Waggoner, B., Rigault, P, *et al.* (1995) *Nature (London)* **377,** Suppl., 299–319.
87. Krauter, K., Montgormery, K., Yoon, S.-L., LeBlanc-Straceski, J., Renault, B., *et al.* (1995) *Nature (London)* **377,** Suppl., 321–333.
88. Collins, J. E., Cole, C. G., Smink, L. J., Garrett, C. L., Leversha, M. A., *et al.* (1995) *Nature (London)* **377,** Suppl., 367–379.
89. Doggett, N. A., Goodwin, L. A., Tesmer, J. G., Meincke, L. J., Bruce, D. C., *et al.* (1995) *Nature (London)* **377,** Suppl., 335–365.
90. Chumakov, I. M., Rigault, P., Le Gall, I., Bellanne-Chantelot, C., Billaut, A., *et al.* (1995) *Nature (London)* **377,** Suppl., 175–297.
91. Adams, M. D., Kerlavage, A. R., Fleischmann, R. D., Fuldner, R. A., Bult, C. J., *et al.* (1995) *Nature (London)* **377,** Suppl., 3–174.
92. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al.* (1995) *Science* **296,** 496–512.