A Gene Map of the Human Genome

G. D. Schuler,* M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tomé, A. Aggarwal, E. Bajorek, S. Bentolila, B. B. Birren, A. Butler, A. B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P. J. R. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, C. Edwards, J.-B. Fan, N. Fang, C. Fizames, C. Garrett, L. Green, D. Hadley, M. Harris, P. Harrison, S. Brady, A. Hicks, E. Holloway, L. Hui, S. Hussain, C. Louis-Dit-Sully, J. Ma, A. MacGilvery, C. Mader, A. Maratukulam, T. C. Matise, K. B. McKusick, J. Morissette, A. Mungall, D. Muselet, H. C. Nusbaum, D. C. Page, A. Peck, S. Perkins, M. Piercy, F. Qin, J. Quackenbush, S. Ranby, T. Reif, S. Rozen, C. Sanders, X. She, J. Silva, D. K. Slonim, C. Soderlund, W.-L. Sun, P. Tabar, T. Thangarajah, N. Vega-Czarny, D. Vollrath, S. Voyticky, T. Wilmer, X. Wu, M. D. Adams, C. Auffray, N. A. R. Walter, R. Brandon, A. Dehejia, P. N. Goodfellow, R. Houlgatte, J. R. Hudson Jr., S. E. Ide, K. R. Iorio, W. Y. Lee, N. Seki, T. Nagase, K. Ishikawa, N. Nomura, C. Phillips, M. H. Polymeropoulos, M. Sandusky, K. Schmitt, R. Berry, K. Swanson, R. Torres, J. C. Venter, J. M. Sikela, J. S. Beckmann, J. Weissenbach, R. M. Myers, D. R. Cox, M. R. James, D. Bentley, P. Deloukas, E. S. Lander, T. J. Hudson

The human genome is thought to harbor 50,000 to 100,000 genes, of which about half have been sampled to date in the form of expressed sequence tags. An international consortium was organized to develop and map gene-based sequence tagged site markers on a set of two radiation hybrid panels and a yeast artificial chromosome library. More than 16,000 human genes have been mapped relative to a framework map that contains about 1000 polymorphic genetic markers. The gene map unifies the existing genetic and physical maps with the nucleotide and protein sequence databases in a fashion that should speed the discovery of genes underlying inherited human disease. The integrated resource is available through a site on the World Wide Web at http://www.ncbi.nlm.nih.gov/SCIENCE96/.

Central to the description of an organism's genome is a comprehensive catalog of the sequence and location of all its genes. Gene maps are now available for those organisms whose complete genomic sequence has been determined, including 141 viruses, 51 organelles, two eubacteria, one archeon, and one eukaryote (the yeast, *Saccharomyces cerevisiae*) (*1*). Such a map of the human genome should become available by 2005, as a result of the efforts by the Human Genome Project to determine the complete 3 billion nucleotides of the human DNA sequence and develop suitable computer and laboratory tools for recognizing genes.

In view of the tremendous value of a human gene map for biomedical research, it is not reasonable to wait until the complete sequence is available to begin preparing such a map. There are compelling reasons for constructing a series of increasingly comprehensive gene maps and cross-referencing them to the human genetic map. A key application is the positional cloning (Fig. 1) of disease-causing genes. Genetic mapping of affected families with polymorphic markers that span the genome permits localization of the disease gene to a candidate region, often in the range of 2 to 5 megabases (Mb). Such intervals are physically mapped with overlapping DNA clones, which usually serve as substrates to identify genes ("transcripts") in the region. Subsequently, the genes are scrutinized for the presence of sequence mutations in affected individuals. Regional transcript mapping by current methods, which is difficult and time-consuming, would be supplanted by the availability of a comprehensive, whole-genome gene map. Such a resource would accelerate gene searches for simple Mendelian traits and is essential in the case of complex (polygenic) traits, for which limited genetic resolution will necessitate sifting through multimegabase regions. The availability of an expanding gene inventory for any candidate region is predicted to make the "positional candidate" approach the predominant method for cloning human disease genes

G. D. Schuler and M. S. Boguski, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA. E. A. Stewart, A. Aggarwal, E. Bajorek, A. Chu, S. Cowles, J.-B. Fan, N. Fang, D. Hadley, M. Harris, S. Brady, S. Hussain, C. Mader, A. Maratukulam, K. B. McKusick, S. Perkins, M. Piercy, F. Qin, J. Quackenbush, T. Reif, C. Sanders, X. She, W.-L. Sun, P. Tabar, D. Vollrath, S. Voyticky, R. M. Myers, D. R. Cox, Department of Genetics, Stanford Human Genome Center, Stanford University School of Medicine, Stanford, CA 94305, USA. L. D. Stein, B. B. Birren, A. B. Castle, L. Hui, J. Ma, H. C. Nusbaum, D. C. Page, S. Rozen, J. Silva, D. K. Slonim, X. Wu, Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology Center for Genome Research, 9 Cambridge Center, Cambridge, MA 02142, USA. G. Gyapay, S. Bentolila, N. Chiannilkulchai, N. Drouot, S. Duprat, C. Fizames, D. Muselet, N. Vega-Czarny, J. S. Beckmann, J. Weissenbach, Généthon, CNRS URA 1922, 1 rue de l'Internationale, 91000 Evry, France. K. Rice, A. Butler, C. Clee, T. Dibling, I. Dunham, C. East, C. Edwards, C. Garrett, L. Green, P. Harrison, A. Hicks, E. Holloway, A. MacGilvery, A. Mungall, A. Peck, S. Ranby, C. Soderlund, T. Wilmer, D. Bentley, P. Deloukas, The Sanger Centre, Hinxton Hall, Hinxton, Cambridge CB10 1SA, UK. R. E. White, P. J. R. Day, C. Louis-Dit-Sully, T. Thangarajah, M. R. James, Wellcome Trust Centre for Human Genetics, Nuffield Department of Clinical Medicine, University of Oxford, Windmill Road, Oxford OX3 7BN, UK. P. Rodriguez-Tomé, European Molecular Biology Laboratory Outstation, Hinxton, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. T. C. Matise, Laboratory of Statistical Genetics, The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA. J. Morissette, Centre de Recherche du Centre Hospitalier de l'Universite Laval, 2705 Boulevard Laurier, Ste-Foy, Quebec G1V 4G2, Canada. M. D. Adams, R. Brandon, C. Phillips, M. Sandusky, J. C. Venter, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. C. Auffray and R. Houlgatte, Genexpress, CNRS UPR 420, 7-19 rue, Guy Moquet-Batiment G, 94801 Villejuif, France. N. A. R. Walter, K. R. Iorio, R. Berry, J. M. Sikela, Department of Pharmacology and Molecular Biology Program, University of Colorado Health Sciences Center, 4200 E. Ninth Avenue, Denver, CO 80262, USA. A. Dehejia, S. E. Ide, M. H. Polymeropoulos, R. Torres, Laboratory of Genetic Disease Research, National Center for Human Genome Research, National Institutes of Health, Bethesda, MD 20892, USA. P. N. Goodfellow and K. Schmitt, Department of Genetics, Cambridge University, Tennis Court Road, Cambridge CB2 3EH, UK. J. R. Hudson Jr., W. Y. Lee, K. Swanson, Research Genetics, 2130 S. Memorial Parkway, Huntsville, AL 35801, USA. N. Seki, T. Nagase, K. Ishikawa, N. Nomura, Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292, Japan. E. S. Lander, Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology Center for Genome Research, 9 Cambridge Center, Cambridge, MA 02142, USA, and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. T. J. Hudson, Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology Center for Genome Research, 9 Cambridge Center, Cambridge, MA 02142, USA, Departments of Medicine and Human Genetics and Montreal General Hospital Research Institute, McGill University, Montreal H3G 1A4, Canada.

*To whom correspondence should be addressed.

(2). Gene maps are also valuable because they shed light on genome organization, including clustering of related genes and conservation of gene order among species.

Constructing a human gene map requires two tools: a large database of genes and an efficient mapping methodology. Both have become available in the past few years through a convergence of high-throughput mapping and sequencing technologies. An international consortium of groups in North America, Europe, and Japan was organized to coordinate a mapping effort (3). This article is the first report from this consortium.

Human Gene Catalog: The UniGene Set

The human genome has been estimated to contain 50,000 to 100,000 genes, on the basis of a variety of indirect techniques (4). Yet, the number of genes actually identified was less than 2000 as recently as 5 years ago (5). Scientists such as Brenner (6) called for large-scale complementary DNA (cDNA) sequencing efforts as a component of the Human Genome Project. The idea was taken up most vigorously by Venter and colleagues, who focused on generating short cDNA fragments, which they called expressed sequence tags (ESTs) (7). A number of other laboratories followed suit (8–10), and since that time, particularly in the past 2 years, the public cDNA collection has swelled to more than 600,000 sequences (about 450,000 of which are human), representing 65% of the entries in the GenBank database (Fig. 2). The EST collection includes portions from 50 to 70% of genes discovered by other means, suggesting that the current EST databases may represent more than half of all human genes (10). [This may be an overestimate, inasmuch as both EST collections and known genes may be biased against rare messenger RNAs (mRNAs).]

To create a human gene catalog, it was necessary to cluster these sequence fragments into groups representing distinct genes. A gene may be represented by multiple ESTs, which may correspond to different portions of a transcript or various alternatively spliced transcripts (Fig. 3). To illustrate the importance of this task, consider that a single gene product, serum albumin, is represented by more than 1300 EST sequences in GenBank. To make mapping efficient and cost-effective, it was necessary to select a single representative sequence from each unique gene. This was accomplished by focusing on 3′ untranslated regions (3′ UTRs) of mRNAs, whose sequences can be efficiently converted to gene-specific sequence tagged sites (STSs) (11) for mapping, as originally proposed by Sikela and co-workers (12).

We developed an information resource called UniGene (Table 1) that is the result of large-scale DNA sequence comparisons among 163,215 3′ ESTs and 8516 3′ ends of known genes selected from GenBank. These sequences were subjected to an optimal alignment procedure to identify sequence pairs with at least 97% identity (13). Sequences were grouped into 49,625 clusters, which is a reasonable estimate of the number of human genes sampled so far. Of these, 4563 (9%) correspond to known genes, with the remainder represented only by ESTs. Other efforts have resulted in similar gene catalogs (14).

Global Mapping Methodologies: RH and YAC Mapping

A variety of techniques have been used for mapping genes. In genetic mapping, genes are localized by analysis of transmission of polymorphic loci. The concept of a "transcript map" has existed for more than 30 years since Jacob and Monod coined the term "messenger RNA," localized the β-galactosidase gene to a genetically defined bin on the *Escherichia coli* chromosome, and postulated a discrete starting point for transcription (15). The first "whole genome" transcript maps were constructed in the mid-to-late 1970s by analysis of mRNA-DNA hybridization on viral or organelle

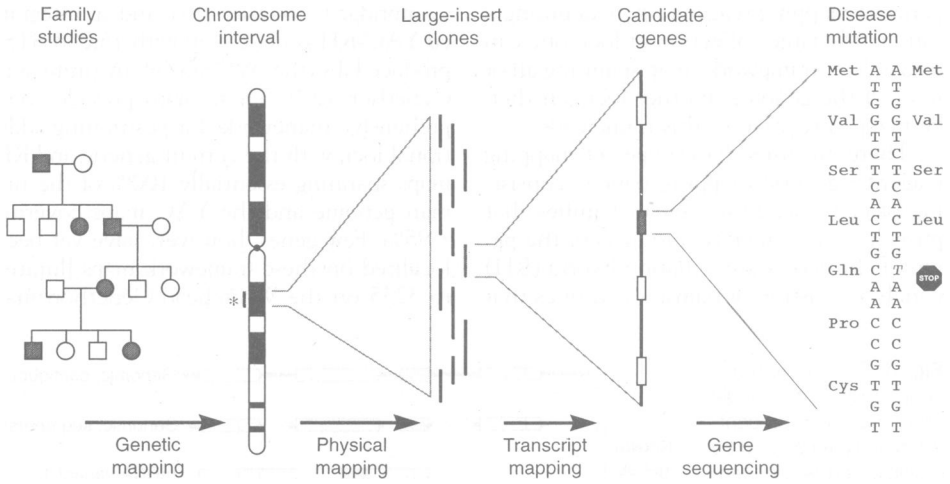

Fig. 1. Steps in positional cloning. Positioning of disease loci to chromosomal regions with genetic markers has become increasingly straightforward, particularly given the recent release of the Généthon genetic map containing 5264 markers (17). However, identification and evaluation of the genes within the implicated region remains a major stumbling block.

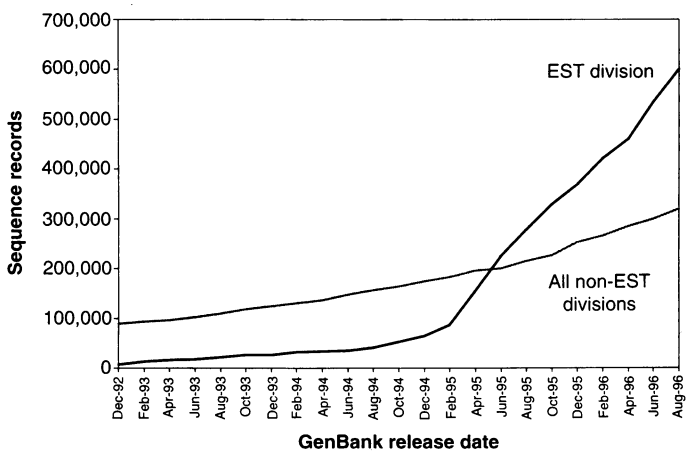

Fig. 2. Availability of genes for mapping. A large-scale gene map of any genome requires the availability of a large set of genes. However, before 1991, less than 2000 unique human gene sequences had accumulated since the advent of rapid DNA sequencing in the mid-1970s. After the development of a new strategy to rapidly obtain sequence samples of large numbers of the transcribed portions of genes, a special division of GenBank devoted to these EST sequences was created in 1992. For several years the growth of the EST division was comparable with other GenBank divisions, but a sharp increase in this rate was seen at the beginning of 1995, which corresponded to the appearance of the first data from the Washington University–Merck & Company EST project (10). In the most recent release (15 August 1996), 65% of all GenBank entries were ESTs. Considering only those sequences of human origin, there were 447,642 sequences in the EST division compared with 52,667 in the PRI (primate) division. These human ESTs provide an abundant, but redundant, source of mapping candidates.

genomes by means of S1 nuclease mapping or electron microscopy (*16*). Because of the small size of these genomes, these techniques allowed researchers to characterize temporal aspects of gene expression as well as gene locations. These methods, however, are not suitable for high-throughput construction of a human gene map.

In the past few years, genome mapping has converged around a unified approach in which the presence or absence of loci in a panel of mapping reagents is scored. Loci are typically defined by short stretches of unique sequence (STSs) and are tested by means of a polymerase chain reaction (PCR) assay (*11*).

Two STS loci are determined to lie nearby one another in the genome if they yield similar patterns of presence or absence in a panel of mapping reagents. By examining a sufficiently large collection of loci, one can assemble a "framework" map spanning all or most of the genome. Further loci can then be mapped relative to this framework.

There are three basic types of mapping reagents: genetic mapping panels, consisting of cell lines from human families that provide various meiotic products of the parental chromosomes; radiation hybrid (RH) panels, consisting of hamster cell lines that contain many large fragments of human DNA produced by radiation breakage; and yeast artificial chromosome (YAC) libraries, consisting of yeast cells that contain individual fragments of human DNA. Genetic mapping can only be performed on polymorphic STSs (those showing variant forms that make it possible to distinguish presence or absence against the background of a complete human genome), whereas RH and YAC mapping are suitable for use with any unique human STS.

Various genome-wide STS-based human maps were completed in 1995, including a genetic map with 5264 genetic markers produced by Généthon (*17*), a YAC map with 2601 STSs by Centre d'Etude du Polymorphisme Humain (CEPH) (*18*), a RH map with 850 STSs produced by Généthon and Cambridge University (*19*), and an integrated YAC-RH genetic map with 15,086 STSs produced by the Whitehead Institute and Généthon (*20*). These maps provide comprehensive frameworks for positioning additional loci, with the current genetic and RH maps spanning essentially 100% of the human genome and the YAC maps covering ~95%. Few genes, however, have yet been localized on these framework maps [limited to 3235 on the Whitehead-Généthon map (*20*) and 318 mapped by researchers at the University of Colorado (*21*)].

The goal of the consortium was to develop and map a large collection of gene-based STSs relative to RH and YAC panels. We used two RH panels and one YAC panel. The Genebridge4 RH panel was produced by Goodfellow and colleagues and consists of 93 hamster cell lines, each retaining ~32% of the human genome in random fragments of ~10 Mb (*19*). The G3 RH panel was produced by Cox and colleagues and consists of 83 hamster cell lines, each retaining ~15% of the human genome in random fragments of ~4 Mb (*22*). The YAC library produced at CEPH contains 32,000 clones with inserts of ~1 Mb, providing roughly eightfold coverage of the human genome. STSs were mapped against one or more of these panels and were then localized relative to a common framework map. The use of different mapping resources to create an integrated map minimizes the effect of any artifacts or deficiencies of particular reagents. Also, features of different reagents often complement one another. For the common framework, we selected a set of 1000 well-spaced genetic markers from the Généthon genetic map. Somewhat different framework subsets were used in each mapping panel, but at least 70% of the markers were common to all three mapping panels (*23*).

To coordinate mapping efforts, the groups in the consortium selected nonoverlapping sets of UniGene entries (*24*), developed primer pairs, and registered these mapping candidates in RHalloc, a database used by the consortium to track which sequences were being mapped and to flag potential duplications. Raw RH mapping results were deposited in the database RHdb, a freely available public database (Table 1). Although the groups used different mapping protocols, some tests were duplicated to monitor and control error rates.

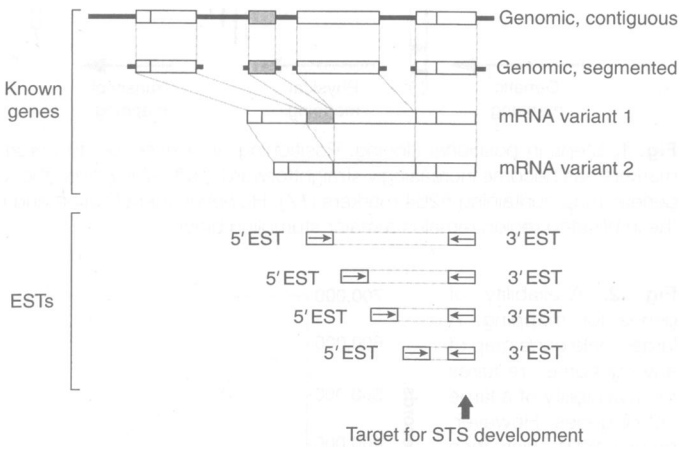

Fig. 3. One gene, many sequences. A single gene may be represented multiple times and by multiple forms in GenBank, which makes selection of unique mapping candidates difficult. Among ESTs, redundancy is a consequence of multiple observations of the same gene, particularly for highly expressed genes and despite procedures designed to minimize resampling (*45*). For known genes, human sequences from the primate (PRI) division of GenBank were used as a starting point and their 3′ UTRs were extracted with various combinations of features of type "CDS," "mRNA," "prim_transcript," "premsg," and "exon."

Table 1. Uniform resource locators (URLs) for gene map information on the World Wide Web.

Information source	URL
Integrated gene map	http://www.ncbi.nlm.nih.gov/SCIENCE96/
Mapping laboratory sites	
Généthon	http://www.genethon.fr/
The Sanger Centre	http://www.sanger.ac.uk/
Stanford Human Genome Center	http://www-shgc.stanford.edu/
Wellcome Trust Centre for Human Genetics	http://www.well.ox.ac.uk/
Whitehead Institute/MIT	http://www.genome.wi.mit.edu/
Allied resources	
RHdb, European Bioinformatics Institute	http://www.ebi.ac.uk/RHdb/
UniGene, National Center for Biotechnology Information	http://www.ncbi.nlm.nih.gov/UniGene/

Construction of the Gene Map

A total of 20,104 gene-based STSs were mapped (Table 2), from about 30,000 STS assays attempted (*25*). Because some genes were independently mapped at more than one laboratory (thereby facilitating comparison and quality assessment), these gene-based STSs correspond to 16,354 distinct loci. Nearly 19,000 gene-based STSs were successfully screened on at least one of the RH panels. An additional 1090 were mapped on the Whitehead Institute STS content YAC map. The contribution of each mapping method and mapping group is summarized in Table 2.

The integrated map is shown in the accompanying chart. Use of selected Généthon polymorphic markers as a mapping

framework allowed the positions of gene-based markers to be resolved to specific intervals measured in centimorgans (cM). The distributions of gene-based STSs along the Généthon maps for each chromosome were plotted as histograms, with the height of the bars proportional to the number of distinct loci per centimorgan (*26*). Genes localized telomeric to the most distal genetic framework markers are in separate bins above and below the histograms. To correlate these data with earlier work, cross-references to the standard ideograms were plotted for those genetic markers that have been localized by fluorescent in situ hybridization (FISH) to banded metaphase chromosomes (*27*). The genetic (*17*) and estimated physical (*28*) lengths of the chromosomes are given below the maps.

As a counterpart to the chart, a World Wide Web site (Table 1) has been developed for more effective use of the map. For example, given a single marker or pair of markers, it is possible to retrieve an inventory of genes mapping to the specified genetic interval (*29*). Furthermore, technical mapping details are presented along with links to maps from individual laboratories (Table 1); these maps represent subsets of the genes in the integrated map that were often placed at higher resolution and accompanied by scores describing the confidence of map placement. Another function of the electronic version of this map is to connect gene-based STSs to nucleotide and protein sequences (see below). These associations, in turn, serve as links to a larger information space consisting primarily of the biomedical literature represented in MEDLINE, but also databases of three-dimensional structures (*30*). Such information should be of value in prioritizing disease gene candidates.

Quality of the Maps

The reliability of the maps can be assessed by examining the 3114 out of 20,104 loci mapped by two different laboratories. In 98% of these cases, the two laboratories assigned the locus to the same chromosome. From the 2% discordance rate, one can estimate that there is an overall error rate of 1% of loci placed on the wrong chromosome. There are many explanations for such conflicts, including laboratory errors, map construction errors, data management errors, and assays that detect loci present at multiple locations in the genome.

To study these problems, the Whitehead group tested a subset of 78 loci that appeared to be discordant with results from other groups. Using an independent mapping method (involving testing loci on the NIGMS1 polychromosomal hybrid mapping panel), the Whitehead group confirmed its own chromosomal assignment in 32 cases, confirmed the conflicting assignment in 28 cases, and found instances consistent with an assay detecting multiple loci in 13 cases (*31*). In a separate test, the Sanger and Whitehead groups examined five discrepancies in STSs derived from different sequences within a single UniGene cluster. In three cases, the different sequences mapped unambiguously to distinct locations (*32*).

Conflicts in the localization of genes along a chromosome were also examined. Of the 3049 loci independently mapped to the same chromosome by more than one group, 92% mapped to either the same, overlapping, or adjacent genetic intervals. Fewer than 2.5% of markers were assigned to intervals that differed by more than 10 cM. Potential sources of such errors include those mentioned above as well as errors in the typing of nearby framework markers (*33*). For cases in which assignments could not be resolved, both positions were listed in the gene map.

Broadly speaking, quality assessment demonstrates that 99% of the loci are placed on the correct chromosome and 95% are mapped with relatively high precision to the correct subchromosomal location; however, the data contain a low frequency of erroneous results, often due to repeated loci or other technical complications.

Table 2. Numbers of cDNAs localized with different mapping resources.

Contributor	Mapping resource			Total
	G3	GB4	YAC	
Whitehead Institute/MIT Center for Genome Research	–	8,116	1,090	9,206
Sanger Centre	349	2,554	–	2,903
Stanford Human Genome Center	2,875	–	–	2,875
Généthon	–	2,629	–	2,629
Wellcome Trust Centre for Human Genetics	–	2,068	–	2,068
National Center for Human Genome Research	–	165	–	165
University of Colorado Health Sciences Center	–	127	–	127
Kazusa DNA Research Institute	10	113	–	123
Total mapped cDNAs	3,234	15,804	1,090	20,128
Unique mapped cDNAs	3,102	13,767	1,070	16,354

Distribution of Human Genes

The distribution of gene-based markers, relative to the Généthon genetic map, is shown for each chromosome on the accompanying chart. To examine the distribution of genes across the genome, we focused only on loci identified from random ESTs, ignoring the 3091 loci derived from full-length genes in GenBank or chromosome-specific mapping projects, as there are systematic biases in the chromosomal distribution of these genes. By comparing the number of mapped ESTs to the cytogenetic length of the chromosome, we observed a significant excess of genes on chromosomes 1, 17, and 19 and a significant deficit on chromosomes 4, 13, 18, 21, and X (Table 3). The findings were consistent with conclusions based on the study of 3300 genes on the Whitehead map. The only substantial difference was that the previously reported excess of genes on chromosome 22 (*20*) was not seen, a conclusion that may have resulted from the small number of ESTs and random STSs used to calculate the relative densities.

The distribution of genes across individual chromosomes appears to show striking fluctuations (see histograms on the chart). Preliminary impressions suggest a higher gene density occurring in lightly staining chromosomal bands, as previously proposed on the basis of smaller samples of mapped genes (*34*). However, firm conclusions cannot yet be drawn because of fundamental uncertainties in the map. First, the genes are shown with respect to framework genetic

Table 3. Expected (Exp) and observed (Obs) chromosomal distributions of mapped cDNAs.

Chromosome	Obs	Exp	Obs/Exp	χ^2
1	1378	1088	1.27	77.29**
2	1106	1053	1.05	2.66
3	954	886	1.08	5.21
4	640	838	0.76	46.78**
5	696	803	0.87	14.25
6	720	759	0.95	2.00
7	730	706	1.03	0.81
8	573	640	0.89	7.01
9	594	601	0.99	0.08
10	591	596	0.99	0.04
11	691	596	1.16	15.14
12	574	592	0.97	0.54
13	256	404	0.63	54.21**
14	434	386	1.12	5.96
15	416	368	1.13	6.26
16	412	404	1.02	0.15
17	548	382	1.44	72.13**
18	261	351	0.74	23.07**
19	446	276	1.61	104.7**
20	368	298	1.23	16.44
21	105	162	0.65	20.05*
22	186	180	1.03	0.2
X	369	680	0.54	142.2**

*Statistically significant at $P < 0.0005$. **Statistically significant at $P < 0.0001$.

markers, and genetic distances are known not to be directly proportional to physical distances. Second, some gene clustering may be due to errors in the underlying framework map that may exclude loci from certain regions. Third, the correspondence between the genetic framework and cytogenetic map is indirect and incomplete.

Many multigene families cluster in the same physical region of the genome. With the gene map one can examine such clustering by searching for regions containing cDNAs showing sequence similarity to related proteins recognized by common keywords in Swiss-Prot database entries. To demonstrate this, we sorted marker sequence sets by keywords and then assigned them to 50-cM bins on the basis of their map locations. Several clustered multigene families were identified without prior knowledge of their localization. For example, marker sets with sequence matches to Swiss-Prot entries containing the keyword "keratin" were tightly clustered in two regions: one on chromosome 17 (5/13 sets; $P < 0.0001$), and another on chromosome 12 (7/13 sets; $P < 0.0001$). This correlates well with the known locations of the type I and type II cytokeratin genes (35). Similarly, gene sets matching the keyword "MHC" were tightly clustered to chromosome 6 (18/25 sets; $P < 0.0001$), corresponding to the known location of the major histocompatibility gene family (36). The "serpin" family of serine protease inhibitors clustered on chromosome 18 (4/18 sets; $P < 0.005$) and chromosome 14 (6/18 sets; $P < 0.0001$), corresponding to previously reported locations (37). These findings suggest that the map is sufficiently dense to identify clustered multigene families that have not been previously described.

Comparative Genomics

Evolutionary conservation of homologous genes from different organisms is of theoretical and practical interest. Often, the putative function of a newly isolated human disease gene is revealed by its sequence similarity to a well-studied gene in another organism. Notable examples include homology between the Alzheimer's disease gene *AD3* and a protein encoded by the genome of the nematode C. *elegans*, and the similarity between the *DPC4* gene involved in pancreatic carcinoma and a *Drosophila* gene implicated in the transforming growth factor–β pathway (38). There is a wealth of examples in which yeast genes have shed light on human disease (39). It was of interest, therefore, to analyze our data set of mapped human genes with respect to potential homologs in other organisms, particularly because more than 90% of our markers derive from ESTs corresponding to proteins of unknown function rather than from characterized genes. Information on similarities with better understood genes in other organisms serves as a form of sequence annotation and might provide clues to possible functions. We compared protein translations of all of the cDNAs mapped in this study (including the corresponding 5′ ends of mapped 3′ ESTs) to all of the protein sequences in the Swiss-Prot database (5). For each of the 15 most highly represented organisms in Swiss-Prot, species-specific protein subsets were generated and compared with the human sequences by using the BLASTX program and scoring systems optimized for the different evolutionary distances (40). Human genes were thus reciprocally cross-referenced to the most significant matching sequences in each of the 15 selected organisms, plus the one best match to an organism outside of this group (labeled "Other organisms" in Table 4). Altogether, 21% of the 16,354 mapped genes have products with significant ($P < 10^{-6}$) similarity to at least one known protein. This level of similarity may seem low, but many ESTs consist only of 3′ UTRs, which are not protein-encoding, and thus no significant BLASTX matches (40) would be detected, even for related genes. Thus, the values in Table 4 are conservative estimates of the extent of cross-referencing possible between mapped human genes and genes in other organisms.

The results of these comparisons are shown in Table 4. For example, when the 2131 mouse (*Mus musculus*) protein sequences in Swiss-Prot were compared against the 16,354 mapped UniGenes, 1098 (52%) of the mouse proteins matched 1767 UniGene open reading frames (ORFs) with a chance probability of $P < 10^{-6}$. Notably, for those organisms whose genomes have been entirely (*S. cerevisiae*) or extensively (*C. elegans*) sequenced, the number of matching proteins is a smaller fraction of the total for these organisms than is observed for most other eukaryotes. One explanation for this observation is that whole-genome sequencing is systematic and thorough and thus has generated large numbers of novel genes, only some of which have been conserved or observed to date in humans. Another explanation is bias in the database caused by technical aspects of "functional" cloning or the fact that, after initial cloning of a particular gene, its homologs are often systematically cloned from selected other species.

In the analysis of putative new genes from whole-genome sequencing projects, it is common practice to describe or annotate potential gene products as "hypothetical proteins" or simply "ORFs." Over 400 cases of mapped human genes or ESTs aligned with high significance to yeast hypothetical proteins or ORFs, indicating that both the yeast and human sequences represent authentic genes maintained over a vast evolutionary distance. Similarly, we observed more than 200 cases of significant matches between human genes or ESTs and hypothetical proteins encoded in the nematode genome. Such cross-phylum sequence conservation implies that these gene products are important for some as-yet-undiscovered biological functions.

Table 4. Reciprocal cross-referencing of the products of mapped human genes to known proteins found in humans and other organisms. Species-specific subsets of the Swiss-Prot database (5) were constructed as described in the text. Reciprocal similarity searches were conducted with the TBLASTN and BLASTX programs (40) with the Swiss-Prot proteins and the UniGene cDNAs, respectively, as query sequences. Any sequence matches with a probability of chance occurrence of $P < 10^{-6}$ were considered significant cross-references. Scoring was based on PAM matrices (40) that were chosen to maximize the detectability of moderately conserved proteins known experimentally to be homologs among various species.

Organism	PAM	Swiss-Prot proteins			Mapped genes	
		Sequences	Proteins matched	% of proteins	Genes matched	% of genes
Homo sapiens	20	3480	1877	54%	2640	17%
Mus musculus	20	2131	1098	52%	1767	11%
Rattus norvegicus	20	1857	1044	56%	1714	11%
Bos taurus	20	815	482	59%	921	6%
Gallus gallus	40	637	377	59%	769	5%
Xenopus laevis	40	508	266	52%	522	3%
Drosophila melanogaster	60	818	399	49%	885	6%
Caenorhabditis elegans	80	1006	322	32%	709	5%
Arabidopsis thaliana	120	499	179	36%	167	1%
Saccharomyces cerevisiae	140	3676	741	20%	756	5%
Schizosaccharomyces pombe	160	640	223	36%	250	2%
Escherichia coli	220	3480	173	5%	174	1%
Hemophilus influenzae	220	1578	121	8%	127	1%
Bacillus subtilis	220	1397	103	7%	118	1%
Salmonella typhimurium	220	603	43	7%	44	0%
Other organisms	120	29905	1184	4%	1761	11%

Conclusions and Future Directions

The value of a human gene map has become increasingly clear in recent years. In some notable cases, disease-gene hunts have been dramatically accelerated by combining approximate-linkage information with partial inventories of candidate genes in the region. In the most favorable case, the region contains a "positional candidate" whose known or inferred function relates to the pathophysiology of the disease. Examples include the identification of *APOE* in late-onset Alzheimer's disease, *MLH1* in hereditary nonpolyposis colon cancer, *FGFR3* in achrondroplasia, and *RET* as the gene responsible for both multiple endocrine neoplasia type 2A and Hirshsprung's disease (*41*). Even in the absence of such a "smoking gun," regional gene catalogs accelerate the search by providing a wealth of markers and transcripts. A comprehensive gene map would ideally allow investigators to proceed immediately to gene characterization (*42*).

The work reported here has greatly increased the number of mapped human genes. At the end of 1994 (about the time this project began), there were 5131 human genes described as mapped in the Genome Data Base (*43*); however, the technical approaches used to map these genes were variable, as were the levels of accuracy and resolution. Many of these previously mapped genes were remapped by this project to provide sequence-based markers on a common and consistent framework. Thus, the number of mapped human genes has more than tripled compared with what was available 22 months ago, and the 16,354 genes on the current map may represent one-fifth of all protein-coding genes in our genome. Furthermore, this new map has sufficient accuracy and resolution to localize genes to within a few megabases, which corresponds well with the regions typically encountered in disease-gene hunts.

This article represents the fruits of the first 18 months of an international collaboration. There is no fundamental barrier to extending this effort toward the goal of localizing the majority of human genes, while recognizing that the gene map will never be truly complete until the entire sequence is in hand. To achieve this, it will be necessary to extend the gene diversity in the public EST database [perhaps through the sequencing of cDNA libraries made by subtractive cloning to diminish resampling of known genes (*44*)] and to continue EST mapping [by generating additional STS assays from genes that were not successfully mapped initially (*45*), and from newly identified genes]. With continued efforts in the years ahead, disease-gene hunts should be transformed into the systematic interrogation of suspects, with revolutionary consequences for our approach to understanding genetic susceptibilities to disease.

REFERENCES AND NOTES

1. Gene maps are available from GenBank Genomes Division at http://www.ncbi.nlm.nih.gov/
2. F. S. Collins, *Nature Genet.* **9**, 347 (1995).
3. The consortium consisted of genome mapping centers or groups at the Whitehead Institute for Biomedical Research, the Sanger Centre, Généthon, Stanford University, Oxford University, the University of Colorado Health Sciences Center, and informatics groups at the National Center for Biotechnology Information and the European Bioinformatics Institute. Additional laboratories having contributed to this mapping effort are the National Center for Human Genome Research and Kazusa DNA Research Institute.
4. F. Antequera and A. Bird, *Nature Genet.* **8**, 114 (1994); C. Fields *et al.*, *ibid.* **7**, 345 (1994); R. Nowak, *Science* **263**, 608 (1994).
5. Swiss-Prot [A. Bairoch and R. Apweiler, *Nucleic Acids Res.* **24**, 21 (1996)], a minimally redundant protein database, contained only 1790 human sequences in the release of 19 August 1991 (A. Bairoch, personal communication).
6. S. Brenner, *Ciba Found. Symp.* **149**, 6 (1990).
7. M. D. Adams *et al.*, *Science* **252**, 1651 (1991).
8. A. S. Khan *et al.*, *Nature Genet.* **2**, 180 (1992); K. Okubo *et al.*, *ibid.*, p. 173; J. M. Sikela and C. Auffray, *ibid.* **3**, 189 (1993).
9. R. Houlgatte *et al.*, *Genome Res.* **5**, 272 (1995).
10. L. Hillier *et al.*, *ibid.* **6**, 807 (1996).
11. M. Olson *et al.*, *Science* **245**, 1434 (1989).
12. A. S. Wilcox *et al.*, *Nucleic Acids Res.* **19**, 1837 (1991). The majority of ESTs have been derived from cDNAs that were directionally cloned from their 3′ ends {using an oligo(dT) primer that anneals to the polyadenylate [poly(A)] tail found at the end of most human mRNAs}. About half of all EST sequences represent putative 3′ ends. Two advantages of using the 3′ UTRs are that they rarely contain introns and they usually display less sequence conservation than do coding regions [W. Makalowski *et al.*, *Genome Res.* **6**, 846 (1996)]. The former feature leads to PCR product sizes that are small enough to amplify; the latter feature makes it easier to discriminate among gene family members that are very similar in their coding regions.
13. For efficiency, a hashing scheme was used to rapidly identify pairs of sequences having a potential relation before subjecting them to full optimal alignment. Any pair of sequences sharing at least two 13-base words separated by no more then two bases was considered an initial candidate. These pairs were analyzed with a variant of the basic Smith-Waterman algorithm [K.-M. Chao *et al.*, *Comput. Appl. Biosci.* **8**, 481 (1992)] in which the search for the optimal alignment was constrained to a band encompassing all observed word hits plus an additional 10 diagonals to each side. Alignment scores were calculated by summing +1 for a match, −2 for a mismatch, −1 for a gap position, and zero for an ambiguous position. The alignment quality was the score divided by the alignment length and was required to be at least 0.91 to be accepted. By this measure, a 100-base alignment with 97 matches and 3 mismatches would just meet the cutoff. An additional constraint was that the observed alignment must extend to within 35 bases of the edge of the search space.
14. The Genexpress Index: R. Houlgatte *et al.*, *Genome Res.* **5**, 272 (1995); the Merck Gene Index: J. S. Aaronson *et al.*, *ibid.* **6**, 829 (1996); and the THC (TIGR human cDNA) collection: M. D. Adams *et al.*, *Nature* **377**, 3 (1995), available at http://www.tigr.org/tdb/hummap/hummap.html
15. F. Jacob *et al.*, *J. Mol. Biol.* **3**, 318 (1961); F. Jacob *et al.*, *C. R. Acad. Sci. Paris* **258**, 3125 (1964).
16. For example, P. A. Sharp, A. J. Berk, S. M. Berget, *Methods Enzymol.* **65**, 750 (1980); J. Battey and D. A. Clayton, *Cell* **14**, 143 (1978).
17. C. Dib *et al.*, *Nature* **380**, 152 (1996).
18. I. M. Chumakov *et al.*, *ibid.* **377**, 175 (1995).
19. G. Gyapay *et al.*, *Hum. Molec. Genet.* **5**, 339 (1996).
20. T. J. Hudson *et al.*, *Science* **270**, 1945 (1995).
21. R. Berry *et al.*, *Nature Genet.* **10**, 415 (1995).
22. E. A. Stewart, personal communication.
23. Genbridge4 Framework: analysis of all 1549 Généthon markers mapped on the GB4 panel shows that 873 markers can be ordered with high confidence (>1000:1) on the basis of only the RH retention patterns. The MultiMap [T. C. Matise *et al.*, *Nature Genet.* **6**, 384 (1994)] implementation of RADMAP (T. Matise *et al.*, available at http://linkage.rockefeller.edu/multimap) was used in the framework map construction by using a maximum-interval-theta of 0.5, a minimum-interval-theta of 0.05, and an odds threshold of 3 for adding markers to the map. In cases where the local order of markers on the RH map did not support the genetic map order with odds of at least 1000:1, markers were removed until such disagreements were resolved. This criterion was relaxed in 19 cases to odds as low as 10:1 to allow spanning of large gaps or at the telomeres. Additional reference markers that do not fulfill the 1000:1 confidence order were included by some mapping groups as reference markers for EST binning. Inclusion criteria used to select these additional markers included (i) markers separated by at least 2 cM and (ii) retention rates between 10 and 60%. This defined a scaffold map of 1038 reference markers. Genetic order was enforced for these reference markers. G3 Framework: reference set of 1000 Généthon markers typed in duplicate and incorporated into the Stanford G3 RH maps. All the framework markers are at 1000:1 odds on the Généthon map, and 707 of the 1000 are at 1000:1 odds on the Stanford map. YAC-Map Framework: genetic positions for the 1090 gene-based markers derived from the genetically anchored double-linked YAC contigs on the STS content YAC map generated at the Whitehead Institute/MIT Genome Center. This map, which has previously been described (*17*), has been expanded to include 566 additional Généthon markers, for a total of 4082 markers.
24. Some mapping candidates were initially selected from the Genexpress, THC, and Kazusa [N. Nomura *et al.*, *DNA Res.* **1**, 27 (1994); N. Nomura *et al.*, *ibid.*, p. 223; T. Nagase *et al.*, *ibid.* **2**, 167 (1995); T. Nagase, N. Seki, K. Ishikawa, A. Tanaka, N. Nomura, *ibid.* **3**, 17 (1996)] collections and retrospectively cross-referenced to UniGene entries.
25. Assays were considered unsuccessful if they consistently yielded no product or multiple products in human DNA, interfering bands in hamster DNA (about 5%), abnormally low (<10%) or high (>60%) retention rates in RH panels, or more than four discrepants between duplicate tests. An additional 10% of assays meeting these criteria fail to map relative to the reference set of genetic markers, possibly because of their being placed past the end of the maps or because of a high proportion of errors.
26. The majority of gene-based STSs were localized to an interval defined by two genetic framework markers or mapped at zero distance from (that is, were nonrecombinant with) a single reference marker, which allowed their positions to be resolved to centimorgan coordinates. In a small number of cases, markers were placed by two-point analysis, so that only the nearest framework marker is known but not an interval. For the purposes of drawing the histogram, a virtual interval was defined that extended half the distance to the nearest framework markers on each side. Sometimes markers mapped between a reference marker and the telomere, and these were plotted in separate bins above and below the maps. A uniform 1.5-cM bin size was used in plotting the histogram. If the interval determined for a marker spanned several of these bins, its contribution was split evenly among them. Duplicate mappings were counted only once, so the heights of the bars are proportional to distinct loci per centimorgan. In the case of mapping conflicts, a partial contribution was made at each of the possible locations.
27. Most of the data were derived from a recent study [P. Bray-Ward *et al.*, *Genomics* **32**, 1 (1996)] in which a large number of CEPH YACs were mapped

by FISH across the whole genome. A subset of the results were selected in which there was no evidence of the YAC having been chimeric, and the position was resolved to a cytogenetic interval (as opposed to fractional length only). A Généthon marker was given for each of these YACs, which allowed the position in centimorgans to be determined. Because this study contained no data for chromosomes 19, 21, and 22, an alternative strategy was used: Genes that are nonrecombinant with respect to Généthon markers [table 5 in C. Dib *et al.*, *Nature* **380** (suppl.), iii (1996)] and have well-known cytogenetic locations [Online Mendelian Inheritance in Man (OMIM) at http://www.ncbi.nlm.nih.gov/] were used to establish the cross-reference.
28. N. E. Morton, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 7474 (1991).
29. Detailed instructions and examples are provided on the Web site.
30. G. D. Schuler, J. A. Epstein, H. Ohkawa, J. A. Kans, *Methods Enzymol.* **266**, 141 (1996).
31. Five additional assays yielded ambiguous results as a result of the presence of an interfering mouse band or poor amplification.
32. This could be the result of either a trivial primer tube labeling error or a sequence clustering error in which two separate genes were erroneously assigned to the same UniGene entry.
33. Typing errors in a framework marker would allow a close EST to map with significant lod scores (logarithm of the odds ratio for linkage) to a correct chromosome location but would tend to localize the marker in a distant bin, in order to minimize "double-breaks" caused by the erroneous typings of the framework marker.
34. J. M. Craig and W. A. Bickmore, *Bioessays* **15**, 349 (1993).
35. V. Romano *et al.*, *Cytogenet. Cell Genet.* **48**, 148 (1988).
36. S. Bahram, M. Bresnahan, D. E. Geraghty, T. Spies, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 6259 (1994).
37. G. D. Billingsley *et al.*, *Am. J. Hum. Genet.* **52**, 343 (1993); S. S. Schneider *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3147 (1995).
38. R. Sherrington *et al.*, *Nature* **375**, 754 (1995); D. Levitan and I. Greenwald, *ibid.* **377**, 351 (1995); S. A. Hahn *et al.*, *Science* **271**, 350 (1996).
39. S. Tugendreich *et al.*, *Hum. Mol. Genet.* **3**, 1509 (1994); D. E. Bassett Jr., M. S. Boguski, P. Hieter, *Nature* **379**, 589 (1996); P. Hieter, D. E. Bassett Jr., D. Valle, *Nature Genet.* **13**, 253 (1996).
40. The scoring systems were amino acid substitution matrices based on the PAM (point accepted mutation) model of evolutionary distance. Pam matrices may be generated for any number of PAMs by extrapolation of observed mutation frequencies [M. O. Dayhoff *et al.*, in *Atlas of Protein Sequence and Structure,* M. O. Dayhoff, Ed. (National Biomedical Research Foundation, Washington, DC, 1978), vol. 5, suppl. 3, pp. 345–352; S. F. Altschul, *J. Mol. Evol.* **36**, 290 (1993)]. PAM matrices were customized for scoring matches between sequences in each of the 15 species pairs; for the pool of remaining proteins ("Other organisms" in Table 4), the PAM120 matrix was used because it has been shown to be good for general-purpose searching [S. F. Altschul, *J. Mol. Biol.* **219**, 555 (1991)]. The BLASTX program [W. Gish and D. J. States, *Nature Genet.* **3**, 266 (1993)] takes a nucleotide sequence query (EST or gene), translates it into all six conceptual ORFs, and then compares these with protein sequences in the database; TBLASTN performs a similar function but instead searches a protein query sequence against six-frame translations of each entry in a nucleotide sequence database. Searches were performed with *E* = 1e-6 and *E2* = 1e-5 as the primary and secondary expectation parameters.
41. M. A. Pericak-Vance *et al.*, *Am. J. Hum. Genet.* **48**, 1034 (1991); W. J. Strittmatter *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 1977 (1993); R. Fishel *et al.*, *Cell* **75**, 1027 (1993); N. Papadopoulos *et al.*, *Science* **263**, 1625 (1994); R. Shiang *et al.*, *Cell* **78**, 335 (1994); L. M. Mulligan *et al.*, *Nature* **363**, 458 (1993); P. Edery *et al.*, *ibid.* **367**, 378 (1994).
42. The potential effect of a comprehensive gene map is well illustrated by the fact that 82% of genes that have been positionally cloned to date are represented by one or more ESTs in GenBank (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_genes/).
43. K. H. Fasman, A. J. Cuticchia, D. T. Kingsbury, *Nucleic Acids Res.* **22**, 3462 (1994).
44. M. Bonaldo *et al.*, *Genome Res.* **6**, 791 (1996).
45. Consensus sequences from at least three overlapping ESTs were generated from 294 UniGene clusters. Of the 230 (78%) redesigned primer pairs, 188 (82%) of these yielded successful PCR assays.
46. We thank M. O. Anderson, A. J. Collymore, D. F. Courtney, R. Devine, D. Gray, L. T. Horton Jr., V. Kouyoumjian, J. Tam, W. Ye, and I. S. Zemsteva from the Whitehead Institute for technical assistance. We thank W. Miller, E. Myers, D. J. Lipman, and A. Schaffer for essential contributions toward the development of UniGene. Supported by NIH awards HG00098 to E.S.L., HG00206 to R.M.M., HG00835 to J.M.S., and HG00151 to T.C.M., and by the Whitehead Institute for Biomedical Research and the Wellcome Trust. T.J.H. is a recipient of a Clinician Scientist Award from the Medical Research Council of Canada. D.C.P. is an assistant investigator of the Howard Hughes Medical Institute. The Stanford Human Genome Center and the Whitehead Institute–MIT Genome Center are thankful for the oligonucleotides purchased with funds donated by Sandoz Pharmaceuticals. Généthon is supported by the Association Francaise contre les Myopathies and the Groupement d'Etudes sur le Genome. The Sanger Centre, Généthon, and Oxford are grateful for support from the European Union EVRHEST programme. The Human Genome Organization and the Wellcome Trust sponsored a series of meetings from October 1994 to November 1995 without which this collaboration would not have been possible.

Life with 6000 Genes

A. Goffeau,* B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, S. G. Oliver

The genome of the yeast *Saccharomyces cerevisiae* has been completely sequenced through a worldwide collaboration. The sequence of 12,068 kilobases defines 5885 potential protein-encoding genes, approximately 140 genes specifying ribosomal RNA, 40 genes for small nuclear RNA molecules, and 275 transfer RNA genes. In addition, the complete sequence provides information about the higher order organization of yeast's 16 chromosomes and allows some insight into their evolutionary history. The genome shows a considerable amount of apparent genetic redundancy, and one of the major problems to be tackled during the next stage of the yeast genome project is to elucidate the biological functions of all of these genes.

The genome of the yeast *Saccharomyces cerevisiae* has been completely sequenced through an international effort involving some 600 scientists in Europe, North America, and Japan. It is the largest genome to be completely sequenced so far (a record that we hope will soon be bettered) and is the first complete genome sequence of a eukaryote. A number of public data libraries compiling the mapping information and nucleotide and protein sequence data from each of the 16 yeast chromosomes (*1–16*) have been established (Table 1).

The position of *S. cerevisiae* as a model eukaryote owes much to its intrinsic advantages as an experimental system. It is a unicellular organism that (unlike many more complex eukaryotes) can be grown on defined media, which gives the experimenter complete control over its chemical and physical environment. *S. cerevisiae* has a life cycle that is ideally suited to classical genetic analysis, and this has permitted construction of a detailed genetic map that defines the haploid set of 16 chromosomes. Moreover, very efficient techniques have been developed that permit any of the 6000 genes to be replaced with a mutant allele, or completely deleted from the genome, with absolute accuracy (*17–19*). The combination of a large number of chromosomes and a small genome size meant that it was possible to divide sequencing responsibilities conveniently among the different international groups involved in the project.

Old Questions and New Answers

The genome. At the beginning of the sequencing project, perhaps 1000 genes encoding either RNA or protein products had been defined by genetic analysis (*20*). The complete genome sequence defines some 5885 open reading frames (ORFs) that are likely to specify protein products in the yeast cell. This means that a protein-encoding gene is found for every 2 kb of the yeast genome, with almost 70% of the total sequence consisting of ORFs (*21*). The yeast genome is much more compact than those of its more complex relatives in the eukary-