A Physical Map of 30,000 Human Genes

P. Deloukas,* G. D. Schuler, G. Gyapay, E. M. Beasley, C. Soderlund, P. Rodriguez-Tomé, L. Hui, T. C. Matise, K. B. McKusick, J. S. Beckmann, S. Bentolila, M.-T. Bihoreau, B. B. Birren, J. Browne, A. Butler, A. B. Castle, N. Chiannilkulchai, C. Clee, P. J. R. Day, A. Dehejia, T. Dibling, N. Drouot, S. Duprat, C. Fizames, S. Fox, S. Gelling, L. Green, P. Harrison, R. Hocking, E. Holloway, S. Hunt, S. Keil, P. Lijnzaad, C. Louis-Dit-Sully, J. Ma, A. Mendis, J. Miller, J. Morissette, D. Muselet, H. C. Nusbaum, A. Peck, S. Rozen, D. Simon, D. K. Slonim, R. Staples, L. D. Stein, E. A. Stewart, M. A. Suchard, T. Thangarajah, N. Vega-Czarny, C. Webber, X. Wu, J. C. Auffray, N. Nomura, J. M. Sikela, M. H. Polymeropoulos, M. R. James, E. S. Lander, T. J. Hudson, R. M. Myers, D. R. Cox, J. Weissenbach, M. S. Boguski, D. R. Bentley

A map of 30,181 human gene–based markers was assembled and integrated with the current genetic map by radiation hybrid mapping. The new gene map contains nearly twice as many genes as the previous release, includes most genes that encode proteins of known function, and is twofold to threefold more accurate than the previous version. A redesigned, more informative and functional World Wide Web site (www.ncbi.nlm.nih.gov/genemap) provides the mapping information and associated data and annotations. This resource constitutes an important infrastructure and tool for the study of complex genetic traits, the positional cloning of disease genes, the cross-referencing of mammalian genomes, and validated human transcribed sequences for large-scale studies of gene expression.

The ultimate gene map for an organism is the complete sequence of its genome, annotated with the beginning and ending coordinates of every gene. Construction of such sequence maps has become routine for simpler organisms with relatively small genome sizes (for example, 1 to 20 Mb), and public databases now contain 18 examples of such complete genomic sequences (*1*). For more complex organisms, such as mice and humans, with genome sizes in the 3-Gb range, complete and accurate genome sequences are still 5 to 10 years away (*2*, *3*). However, large quantities of preliminary data ("shotgun assemblies") are already available (*4*) and expected to grow rapidly (*5*). Both of these factors necessitate the construction of gene maps to support basic and applied research in mammalian biology and medicine, as well to aid in the analysis and interpretation of "unfinished" genome sequence data. Extensive libraries of expressed gene sequences (*6*, *7*), combined with physical mapping with radiation hybrid (RH) panels (*8*–*10*), have provided the information, infrastructure, and technology to produce such maps in an efficient and economical manner.

In 1994, an international consortium was formed to construct a human gene map in which cDNA-based sequence-tagged site (STS) markers were physically mapped and then integrated with the genetic map of polymorphic microsatellite markers (*11*). The initial report of this consortium in 1996 described a map of ~16,000 genes (*12*). A new map, reported here, represents a nearly 100% increase in gene density and map accuracy and may contain up to half of all human protein-coding genes. This map should be a valuable resource for the positional candidate cloning of complex (polygenic) disease loci, the construction of complete physical maps of chromosomes for genome sequencing, and comparative analysis of mammalian chromosome structure and evolution. Furthermore, sequence validation that occurs in the process of STS design and mapping creates a quality-assured gene sequence resource for "functional genomics" applications (*13*) such as the design and construction of large-scale gene expression arrays.

This new gene map consists of data from 41,664 STSs (Table 1). As in the previous map (*12*), they are based on 3′ untranslated regions of cDNAs. These STSs represent 30,181 unique genes. Markers were typed on the Genebridge4 (GB4) RH panel (39,886 cDNAs, 1641 microsatellite markers, and 13 telomeric markers), on the G3 RH panel (5013 cDNAs and 2091 microsatellites), or on both panels (1102 microsatellites). All GB4 data (Table 1) were, for the first time, merged into a single map and aligned with the G3 RH map and the genetic map (*11*) with the 1102 microsatellite markers that are common to all three maps. The integrated map is available at www.ncbi.nlm.nih.gov/genemap. In addition, two Web servers [one for each RH panel (*14*)] permit anyone to map a new marker relative to this map.

This new map is twofold to threefold more accurate than the 1996 gene map by several criteria. Some markers were mapped in duplicate to make it possible to detect discrepancies between independent experimental results. The error rate in assignment of the same marker to different chromosomes was 0.52% (compared with 1% in the 1996 map). The error rate in chromosome assignment was also assessed, with the e-PCR program (*15*), by matching STSs to 122 Mb of human genomic sequence data present in GenBank as of April 1998. Twenty-three of the 2134 STSs tested matched a genomic sequence from a different chromosome from that determined by the RH mapping, corresponding to an error rate of 1.08% (3.76% for the 1996 map). To assess the accuracy of marker placements along the chromosomes, we converted positions of markers on the GB4 and G3 maps (in cR3000 and cR10000 coordinates, respectively) to centimorgan (cM) coordinates on the genetic map for direct comparison. Only 1.35% of markers were discrepant by 10 cM or more (2.5% for the 1996 map), and 1.78% of markers had positions differing by 5 cM. This substantial improvement in quality of the new map is due primarily to retyping or removal (or both) of markers suspected of being in error on the basis of analysis of the previous map.

The chromosomal distribution of 30,075 distinct gene-based markers (excluding those with conflicting chromosome assignments) is given in Table 2. The ratio of observed versus expected genes per chromosome [based on the physical length of each chromosome (*16*)] indicates a significantly higher gene density for chromosomes 1, 11, 17, 19, and 22 and a significantly lower gene density for chromosomes 4, 5, 8, 13, 18, and X.

The total number of human genes has been estimated at 60,000 to 70,000 (*17*). Therefore, this map contains transcript markers approaching half of all human genes. The map includes 18,703 of the 46,045 entries in UniGene (*7*) and 4684 (78%) of the about 6000 human genes of known function (*18*). Work is continuing not only to map all remaining unmapped cDNAs but also to redevelop and retype markers for cDNAs that failed initial mapping attempts. New efforts by the community to convert expressed sequence tags (ESTs) into more accurate and complete cDNA clones and sequences (*3*) will aid this process enormously.

The main practical value of having a dense and integrated genetic-physical map of genes is to accelerate the discovery, by positional and positional candidate cloning (*19*), of human disease genes. In the calendar year after publication of the 16,000-gene map (*12*), isolation of

Table 1. Number of markers in the current and previous gene maps by contributor.

Contributor	1996 gene map	1998 gene map
Sanger Centre	2,554	12,710
Whitehead Institute/MIT Center for Genome Research	8,116	9,977
Généthon	2,629	9,178
Stanford Human Genome Center (G3 Rh panel)	2,875	5,021
Wellcome Trust Centre for Human Genetics	2,068	3,585
National Human Genome Research Institute	165	766
University of Colorado Health Sciences Center	127	310
Kazusa DNA Research Institute	113	117
Total gene-based markers	18,647	41,664

Table 2. Chromosome distribution of distinct gene-based STS markers.

Chromosome	Observed	Expected	Observed/ expected ratio	χ^2
1	3114	2507	1.24	147.0*
2	2257	2431	0.93	12.5
3	2015	2040	0.99	0.3
4	1478	1935	0.76	107.9*
5	1529	1849	0.83	55.4*
6	1893	1744	1.08	12.7
7	1594	1630	0.98	0.8
8	1206	1478	0.82	50.1*
9	1248	1382	0.90	13.0
10	1371	1373	1.00	0.0
11	1755	1373	1.28	106.3*
12	1585	1363	1.16	36.2
13	703	934	0.75	57.1*
14	1047	886	1.18	29.3
15	1029	848	1.21	38.6
16	849	934	0.91	7.7
17	1263	877	1.44	169.9*
18	523	810	0.65	101.7*
19	1114	638	1.74	355.1*
20	758	686	1.10	7.6
21	305	371	0.82	11.7
22	565	410	1.38	58.6*
X	874	1563	0.56	303.7*

*Statistically significant at $P < 0.001$.

P. Deloukas, C. Soderlund, A. Butler, C. Clee, T. Dibling, S. Gelling, L. Green, P. Harrison, R. Hocking, E. Holloway, S. Hunt, A. Mendis, A. Peck, D. Simon, R. Staples, D. R. Bentley, Sanger Centre, Hinxton Hall, Hinxton, Cambridge CB10 1SA UK. 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. G. Gyapay, S. Bentolila, N. Chiannilkulchai, N. Drouot, S. Duprat, C. Fizames, D. Muselet, N. Vega-Czarny, J. Weissenbach, Généthon, CNRS URA 1922, 1 rue de l'Internationale, 91000 Evry, France, and Genoscope Centre National de Sequencage, 2 rue Gaston Cremieux, 91000 Evry, France. E. M. Beasley, K. B. McKusick, E. A. Stewart, R. M. Myers, D. R. Cox, Department of Genetics, Stanford Human Genome Center, Stanford University School of Medicine, Stanford, CA 94305, USA. P. Rodriguez-Tomé and P. Lijnzaad, European Molecular Biology Laboratory Outstation, Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. L. Hui, B. B. Birren, A. B. Castle, J. Ma, H. C. Nusbaum, S. Rozen, 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. T. C. Matise, Laboratory of Statistical Genetics, Rockefeller University, 1230 York Avenue, New York, NY 10021, USA. J. S. Beckmann, Généthon, CNRS URA 1922, 1 rue de l'Internationale, 91000 Evry, France. M.-T. Bihoreau, J. Browne, P. J. R. Day, S. Fox, S. Keil, C. Louis-Dit-Sully, J. Miller, T. Thangarajah, C. Webber, M. R. James, Wellcome Trust Centre for Human Genetics, Nuffield Department of Clinical Medicine, University of Oxford, Windmill Road, Oxford OX3 7BN, UK. A. Dehejia and M. H. Polymeropoulos, Laboratory of Genetic Disease Research, National Center for Human Genome Research, National Institutes of Health, Bethesda, MD 20892 USA. J. Morissette, Centre de Recherche du Centre Hospitalier de l'Université Laval, 2705 Boulevard Laurier, Ste-Foy, Quebec G1V 4G2, Canada. L. D. Stein, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. M. A. Suchard, Department of Biomathematics, University of California, Los Angeles School of Medicine, Los Angeles, CA 90095, USA. J. Hudson, Research Genetics, 2130 Memorial Parkway S.W., Huntsville, AL 35801, USA. C. Auffray, Genexpress, CNRS UPR 420, 7-19 rue Guy Moquet–Batiment G, 94801 Villejuif, France. N. Nomura, Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292, Japan. J. M. Sikela, Department of Pharmacology and Molecular Biology Program, University of Colorado Health Sciences Center, 4200 East Ninth Avenue, Denver, CO 80262, USA. 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, and Montreal General Hospital Research Institute, McGill University, Montreal, Canada.

*To whom correspondence should be addressed.

16 genes by positional approaches was reported (*20*). Retrospective analysis shows that 44% (7 out of 16) of these genes had already been isolated as ESTs and mapped at the time of their cloning. This fraction increases to 69% (11 out of 16) when the data from the current map are considered.

Comparative analysis has a long and fruitful history in biology, and detailed comparative maps of mammalian genomes have shed light on chromosome evolution. The identification and cross-referencing of genes allow insights into similarities and differences of physiology and development as well as candidates for transgenesis and gene knockout experiments. Thus, it was of interest to determine the extent to which genes on the current human map could be related to orthologous genes in other mammals. Makalowski and Boguski (*21*) have assembled a set of 1880 human genes along with their rat or mouse (or both) orthologs. When these genes were analyzed for overlap with the 30,181 mapped human genes in the current study, we found that 70% of these human genes with rodent counterparts are present. This data set therefore provides an excellent index for cross-referencing the human map with emerging gene-based physical maps of the mouse and rat genomes (*22*).

Genome-scale expression monitoring or profiling (*23*), a rapidly expanding area of functional genomics, relies on the availability of large catalogs of cDNA sequences or arrays of clones (or both). The problems posed by sequence redundancy and inaccuracy are as critical for gene expression applications as they have been for transcript mapping. Furthermore, additional problems in these catalogs have become apparent, necessitating the authentication of sequences and clone reagents. Our collection of nearly 42,000 successfully mapped, gene-based STSs, representing ~30,000 unique human transcripts, provides a large, validated set of human sequences that can be used to design gene-specific oligonucleotides or select cDNA-derived polymerase chain reaction products for populating gene expression arrays (or both). Use of this set could lead to a very useful confluence of mapping and expression information for human genes.

We have produced a map containing perhaps half of all human genes. In the future, this map and subsequent versions will ultimately be replaced by the complete sequence of the human genome. Until then, this reference resource should contribute substantially to the advancement of structural and functional genomics, to comparative biology, and to the isolation of human disease genes, particularly those underlying complex traits.

References and Notes

1. Entrez Genomes at www.ncbi.nlm.nih.gov/Entrez/Genome/org.html
2. At the time of writing, somewhat more than 5% of the human genome sequence is completed (J. Zhang, O. Pickeral, M. S. Boguski, G. D. Schuler, unpublished observations).
3. F. Collins *et al.*, *Science* **282**, XXX (1998).
4. These data are available through the High Throughput Genomic Sequences division of GenBank at www.ncbi.nlm.nih.gov/HTGS [B. F. F. Ouellette and M. S. Boguski, *Genome Res.* **7**, 952 (1997)] and many sequencing center Web sites (K. D. Pruitt, *ibid.*, p. 1038).
5. J. C. Venter *et al.*, *Science* **280**, 1540 (1998).
6. A. R. Williamson, K. O. Elliston, J. L. Sturchio, *J. NIH Res.* **7**, 63 (1995); M. D. Adams *et al.*, *Nature* **377** (suppl.), 3 (1995); L. D. Hillier *et al.*, *Genome Res.* **6**, 807 (1996); R. Houlgatte *et al.*, *ibid.* **5**, 272 (1995); B. A. Eckman *et al.*, *Bioinformatics* **14**, 2 (1998).
7. M. S. Boguski and G. D. Schuler, *Nature Genet.* **10**, 369 (1995); G. D. Schuler, *J. Mol. Med.* **75**, 694 (1997).
8. T. J. Hudson *et al.*, *Science* **270**, 1945 (1995).
9. G. Gyapay *et al.*, *Hum. Mol. Genet.* **5**, 339 (1996). The GB4 RH panel consists of 93 hamster cell lines, each retaining about 32% of the human genome in fragments averaging ~25 Mb.
10. E. A. Stewart *et al.*, *Genome Res.* **7**, 422 (1997). The Stanford G3 RH panel consists of 83 hamster lines with a 16% retention frequency of fragments of ~2.4 Mb. Alignment of the maps demonstrates excellent agreement in the order of RH framework markers and markers on the genetic map (*11*).
11. C. Dib *et al.*, *Nature* **380**, 152 (1996).
12. G. D. Schuler *et al.*, *Science* **274**, 540 (1996).
13. P. Hieter and M. Boguski, *ibid.* **278**, 601 (1997).
14. The server for the G3 RH panel is available at www-shgc.stanford.edu/RH/rhserver_form2.html, and the server for the GeneBridge4 (GB4) RH panel is available at www.sanger.ac.uk/RHserver
15. G. Schuler, *Genome Res.* **7**, 541 (1997).
16. N. E. Morton, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 7474 (1991). For the acrocentric chromosomes (13, 14, 15, 21, and 22), only the lengths of the long arms were used to calculate the expected gene numbers and χ^2 values (Table 2).
17. C. Fields, M. D. Adams, O. White, J. C. Venter, *Nature Genet.* **7**, 345 (1994).
18. These numbers reflect only those STSs whose primer sequences could be unequivocally matched to cDNA in UniGene clusters or to mRNA sequence records in GenBank. A portion of the STSs used in this study were designed from genomic sequences surrounding the genes [N. Kenmochi *et al.*, *Genome Res.* **8**, 509 (1998)] and thus do not match sequences in the transcript. In some other cases, matches to ESTs were observed, but these ESTs did not meet the conservative minimum quality criteria for inclusion in UniGene.
19. F. S. Collins, *Nature Genet.* **9**, 347 (1995).
20. M. van Slegtenhorst *et al.*, *Science* **277**, 805 (1997); N. A. Quaderi *et al.*, *Nature Genet.* **17**, 285 (1997); L. J. Ozelius *et al.*, *ibid.*, p. 40; J. Ortego, J. Escribano, M. Coca-Prados, *FEBS Lett.* **413**, 349 (1997); K. Nagamine *et al.*, *Nature Genet.* **17**, 393 (1997); E. D. Lynch *et al.*, *Science* **278**, 1315 (1997); Q. Y. Li *et al.*, *Nature Genet.* **15**, 21 (1997); M. Gebbia *et al.*, *ibid.* **17**, 305 (1997); L. A. Everett *et al.*, *ibid.*, p. 411; The International FMF Consortium, *Cell* **90**, 797 (1997); The Finnish-German APECED Consortium, *Nature Genet.* **17**, 399 (1997); S. C. Chandrasekharappa *et al.*, *Science* **276**, 404 (1997); E. D. Carstea *et al.*, *ibid.* **277**, 228 (1997); C. T. Basson *et al.*, *Nature Genet.* **15**, 30 (1997); R. Allikmets *et al.*, *ibid.*, p. 236; E. M. Stone *et al.*, *Science* **275**, 668 (1997); T. Kishino, M. Lalande, J. Wagstaff, *Nature Genet.* **15**, 70 (1997); T. Oda *et al.*, *ibid.* **16**, 235 (1997).
21. W. Makalowski and M. S. Boguski, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9407 (1998).
22. M. R. James and K. Lindpaintner, *Trends Genet.* **13**, 171 (1997); S. A. Camper and M. H. Meisler, *Mamm. Genome* **8**, 461 (1997).
23. C. Dahl, Comprehensive Transcript Profiling Analysis (1998). Available at www.ncbi.nlm.nih.gov/ncicgap/expression_tech_info.html

24. We thank A. Aggarwal, E. Bajorek, S. Brady, M. Denys, S. Lewis, B. Louie, F. Lopez, J. Marquis, J. Norton, T. Odineca, A. Perou, M. Piercy, N. Vo, V. Shokoohi, and W.-L. Sun from the Stanford Human Genome Center and M. O. Anderson, A. J. Collymore, R. Devine, D. Gray, L. T. Horton Jr., R. Kouyoumjian, J. Tam, Y. Wu, and W. Ye from the Whitehead Institute for technical assistance and R. Berry, N. Walter, and K. Iorio from the University of Colorado for mapping contributions. We gratefully acknowledge the support of the Wellcome Trust to the Sanger Centre and the Wellcome Trust Centre for Human Genetics, Oxford, and the support of the NIH to the Whitehead Institute for Biomedical Research and Stanford Human Genome Center. The Sanger Centre, Généthon, and Oxford also received support from the European Union (grant BMH4-CT95-1565). T.J.H. is a recipient of a Clinician Scientist Award from the Medical Research Council of Canada.

26 May 1998; accepted 21 August 1998

Ordering of the Numerosities 1 to 9 by Monkeys

Elizabeth M. Brannon and Herbert S. Terrace

A fundamental question in cognitive science is whether animals can represent numerosity (a property of a stimulus that is defined by the number of discriminable elements it contains) and use numerical representations computationally. Here, it was shown that rhesus monkeys represent the numerosity of visual stimuli and detect their ordinal disparity. Two monkeys were first trained to respond to exemplars of the numerosities 1 to 4 in an ascending numerical order (1 → 2 → 3 → 4). As a control for non-numerical cues, exemplars were varied with respect to size, shape, and color. The monkeys were later tested, without reward, on their ability to order stimulus pairs composed of the novel numerosities 5 to 9. Both monkeys responded in an ascending order to the novel numerosities. These results show that rhesus monkeys represent the numerosities 1 to 9 on an ordinal scale.

Many animal taxa can discriminate stimuli differing in numerosity (*1*). The importance of this capacity has evoked considerable controversy. Some have argued that animals have a natural ability to discriminate numerosity (*2, 3*); others maintain that animals attend to numerosity as a "last resort," that is, only if all other bases for discrimination are eliminated (for example, the shape, color, brightness, size, frequency, or duration of a stimulus) (*4*).