An STS-Based Map of the Human Genome

Thomas J. Hudson,* Lincoln D. Stein, Sebastian S. Gerety, Junli Ma, Andrew B. Castle, James Silva, Donna K. Slonim, Rafael Baptista, Leonid Kruglyak, Shu-Hua Xu, Xintong Hu, Angela M. E. Colbert, Carl Rosenberg, Mary Pat Reeve-Daly, Steve Rozen, Lester Hui, Xiaoyun Wu, Christina Vestergaard, Kimberly M. Wilson, Jane S. Bae, Shanak Maitra, Soula Ganiatsas, Cheryl A. Evans, Margaret M. DeAngelis, Kimberly A. Ingalls, Robert W. Nahf, Lloyd T. Horton Jr., Michele Oskin Anderson, Alville J. Collymore, Wenjuan Ye, Vardouhie Kouyoumjian, Irena S. Zemsteva, James Tam, Richard Devine, Dorothy F. Courtney, Michelle Turner Renaud, Huy Nguyen, Tara J. O'Connor, Cécile Fizames, Sabine Fauré, Gabor Gyapay, Colette Dib, Jean Morissette, James B. Orlin, Bruce W. Birren, Nathan Goodman, Jean Weissenbach, Trevor L. Hawkins, Simon Foote, David C. Page, Eric S. Lander*

A physical map has been constructed of the human genome containing 15,086 sequence-tagged sites (STSs), with an average spacing of 199 kilobases. The project involved assembly of a radiation hybrid map of the human genome containing 6193 loci and incorporated a genetic linkage map of the human genome containing 5264 loci. This information was combined with the results of STS-content screening of 10,850 loci against a yeast artificial chromosome library to produce an integrated map, anchored by the radiation hybrid and genetic maps. The map provides radiation hybrid coverage of 99 percent and physical coverage of 94 percent of the human genome. The map also represents an early step in an international project to generate a transcript map of the human genome, with more than 3235 expressed sequences localized. The STSs in the map provide a scaffold for initiating large-scale sequencing of the human genome.

A physical map affording ready access to all chromosomal regions is an essential prerequisite for the international effort to sequence the entire human genome. In the shorter term, it is also a key tool for positional cloning of disease genes and for studies of genome organization. Physical maps have evolved over the past decade from their initial conception as a set of overlapping clones (*1*) to the more recent idea of a well-spaced collection of unique landmarks called sequence-tagged sites (STSs), each defined by a polymerase chain reaction (PCR) assay (*2*–*4*). The U.S. Human Genome Project, for example, has set a target of a physical map consisting of 30,000 STSs spaced at intervals of about 100 kb (*5*).

By focusing on STS landmarks, genome researchers sought to insure against the inevitable problems inherent in any given clone library (*2*). The wisdom of this approach was borne out as it emerged that yeast artificial chromosomes (YACs), the best clones for covering large distances, suffer from high rates of chimerism and rearrangement and thus are unsuitable for genomic sequencing (*6*, *7*). STS-based maps sidestep this problem by having a sufficiently high density of landmarks that one can rapidly regenerate physical coverage of any region by PCR-based screening of clones appropriate for sequencing—such as cosmids, bacterial artificial chromosomes, and P1-artificial chromosomes (*8*).

STS-based physical maps with extensive long-range continuity have been constructed for only a handful of human chromosomes: 3, 12, 16, 21, 22, and Y (*3*, *4*, *9*, *10*). These combined maps cover just less than 20% of the genome with about 1600 STSs, and the average spacing on most of these chromosomes is about 250 kb. Projects are also underway for a few additional chromosomes (*11*). An international collaboration among the Centre d'Etude du Polymorphisme Humain (CEPH), Généthon, and Whitehead genome centers has also produced a clone-based physical map estimated to cover up to 75% of the genome in overlapping YAC clones (*7*). The map is clone-based, rather than STS-based, because it was primarily assembled by detecting physical overlaps among the clones themselves (by means of cross-hybridization and fingerprinting methods), with only a sparse set of STS landmarks used as anchors (786 loci fully screened and 1815 loci partially screened on YACs). The map is quite valuable for positional cloning projects, but it does not provide a scaffold for sequencing the human genome: The YAC clones themselves are not suitable for sequencing, and the STS coverage is too sparse to regenerate substantial physical coverage.

Here, we report the construction of an STS-based physical map of the human genome containing more than 15,000 loci, with an average spacing of 199 kb. The map covers the vast majority of the human genome and provides a scaffold for initiating large-scale sequencing.

Basic Strategy

We used three mapping methods to gain information about the proximity of STS loci within the human genome.

1) STS-content mapping. YAC libraries are screened by PCR to identify all clones containing a given locus (*12*). Nearby loci tend to be present in many of the same clones, allowing proximity to be inferred. STS-content linkage can be detected over distances of about 1 Mb, given the average insert size of the YAC library used here.

2) Radiation hybrid (RH) mapping. Hybrid cell lines, each containing many large chromosomal fragments produced by radia-

T. J. Hudson, S. S. Gerety, J. Ma, A. B. Castle, J. Silva, D. K. Slonim, R. Baptista, L. Kruglyak, S.-H. Xu, X. Hu, A. M. E. Colbert, C. Rosenberg, M. P. Reeve-Daly, S. Rozen, L. Hui, X. Wu, C. Vestergaard, K. M. Wilson, J. S. Bae, S. Maitra, S. Ganiatsas, C. A. Evans, M. M. DeAngelis, K. A. Ingalls, R. W. Nahf, L. T. Horton Jr., M. O. Anderson, A. J. Collymore, W. Ye, V. Kouyoumjian, I. S. Zemsteva, J. Tam, R. Devine, D. F. Courtney, M. T. Renaud, H. Nguyen, T. J. O'Connor, B. W. Birren, N. Goodman, T. L. Hawkins, and S. Foote, Whitehead–Massachusetts Institute of Technology (MIT) Center for Genome Research, Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA. L. D. Stein, Whitehead–MIT Center for Genome Research, Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA, and Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115, USA. C. Fizames, S. Fauré, G. Gyapay, C. Dib, and J. Weissenbach, Généthon, CNRS URA 1922, 1 rue de l'Internationale, 91000 Evry, France. J. Morissette, Généthon, CNRS URA 1922, 1 rue de l'Internationale, 91000 Evry, France, and Centre de Recherche du Centre Hospitalier de l'Université Laval, 2705 Boulevard Laurier, Ste-Foy, Québec G1V 4G2, Canada. J. B. Orlin, Sloan School of Management, MIT, Cambridge, MA 02139, USA. D. C. Page, Whitehead–MIT Center for Genome Research, Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA; Howard Hughes Medical Institute, MIT, 9 Cambridge Center, Cambridge, MA 02139, USA; and Department of Biology, MIT, 9 Cambridge Center, Cambridge, MA 02139, USA. E. S. Lander, Whitehead–MIT Center for Genome Research, Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA, and Department of Biology, MIT, 9 Cambridge Center, Cambridge, MA 02139, USA.

*To whom correspondence should be addressed.

tion breakage, are screened by PCR to identify those hybrids that have retained a given locus (*13*). Nearby loci tend to show similar retention patterns, allowing proximity to be inferred. RH linkage can be detected for distances of about 10 Mb, given the average fragment size of the RH panel used here.

3) Genetic mapping. A locus that is polymorphic in the human population can be screened by PCR to determine its inheritance patterns in families (*14, 15*). Nearby loci tend to show similar inheritance patterns, allowing proximity to be inferred. Genetic linkage can be reliably detected over distances of about 30 Mb, given the recombination rate of human chromosomes (*16*).

These three methods were used to produce independent maps and then combined to produce an integrated map. Because RH mapping and genetic mapping can detect linkage over large regions (0.3 to 1% of the genome), comprehensive RH and genetic maps spanning all chromosomes can be assembled with a few thousand loci. The order of loci can be inferred from the extent of correlation in the retention or inheritance patterns, although estimates of fine-structure order are not precise. These methods can thus provide "top-down" information about global position in the genome.

In contrast, STS-content mapping provides "bottom-up" information. It reveals tight physical linkage among loci but is useful only over short distances and does not provide extensive long-range connectivity across chromosomes (*17*). Two STSs are said to be singly linked if they share at least one YAC in common and doubly linked if they share at least two YACs (*17*). Single linkage is an inadequate criterion for declaring adjacency of STSs, because of the high rate of YAC chimerism (about 50%) and the possibility of laboratory error. Double linkage, however, turns out to be a reliable indication, because two genomic regions are unlikely to be juxtaposed in multiple independent YACs. Accordingly, a three-step procedure was used. (i) STSs were assembled into doubly linked contigs (groups of STSs connected by double linkage). (ii) The doubly linked contigs were localized within the genome on the basis of RH and genetic map information about loci in the contig. (iii) Single linkage was then used to join contigs localized to the same small genomic region. The overall strategy is illustrated in Fig. 1. We now describe the data generation, map construction, and map analysis in greater detail.

Data Generation

Marker development. Over the course of the project, we tested 20,795 distinct PCR assays. These candidate STSs were initially characterized to see whether they were likely to detect a unique genomic locus (*18*) and whether they consistently yielded correct results on control samples under uniform production conditions. A total of 16,239 STSs met these stringent criteria and were used for mapping. The STSs fell into one of the following four categories.

1) Random loci. We generated 3027 working STSs by sequencing random human genomic clones and discarding those that appeared to contain repetitive sequences (*19*).

2) Expressed sequences. We developed 921 STSs from complete complementary DNA (cDNA) sequences in GenBank, taken from the Unigene collection (*20*). Another 3349 STSs were developed from expressed sequence tags (ESTs). Of these, 71% came from the dbEST database (*21*), 13% from the laboratory of Jim Sikela, 9% from the Institute for Genomic Research, and 7% from various other sources (*22*). We found that the success rate for STSs derived from the last 200 base pairs (bp) of 3′-untranslated regions (UTRs) of cDNAs was similar to that for STSs derived from random genomic DNA, consistent with the idea that introns rarely occur near the ends of 3′-UTRs (*23*). The results indicate that PCR assays can be readily derived for the vast majority of cDNAs.

3) Genetic markers. A total of 6986 loci were used, consisting of 5264 polymorphic loci developed at Généthon (primarily CA repeats) (*24*) and 1722 loci developed by the Cooperative Human Linkage Center (CHLC) (primarily tri- and tetranucleotide repeats) (*15*).

4) Other loci. A total of 1956 STSs were developed from various sources. These included 1091 CA-repeat loci developed at Généthon that were not sufficiently polymorphic to be useful for genetic mapping, as well as 865 loci from chromosome 22–specific and chromosome Y–specific libraries and gifts from other laboratories (*3, 25, 26*).

A total of 15,086 STSs appear in the final maps. The number of markers of each type appearing in the final STS-content, RH, and genetic maps is shown in Table 1.

STS-content mapping: Methodology. STSs were screened against 25,344 clones from plates 709 to 972 of the CEPH mega-YAC library (*7*), estimated to have an average insert size of 1001 kb and to provide roughly 8.4-fold coverage of the genome. To facilitate screening, we used a hierarchical pooling system. The library was divided into 33 "blocks," each corresponding to eight microtiter plates or roughly 0.25 genome equivalent. For each block, we prepared one "superpool" containing DNA from all the clones and 28 "subpools" by using a three-dimensional pooling system based on the row, plate, and column address of each clone. Specifically, there were 8, 8, and 12 subpools consisting of YACs in the same plate, row, and column, respectively. There was thus a total of 957 super- and subpools.

For blocks with a single positive YAC, the row, column, and plate subpools should specify the precise address of the YAC ("definite addresses"). If a block contained two or more positive YACs or if one of the three subpool dimensions did not yield a positive, partial information was obtained ("incomplete addresses") (*27*). Such incomplete addresses could consist of up to 12 possible addresses (for example, in the case that a column address was missing). Incomplete addresses were not used in initial map assembly but were used at the final stages to detect connections between nearby loci. Definite addresses composed 88% of the total hits.

Half of the markers were screened by a two-level procedure, in which we first identified the positive superpools and then tested only the corresponding subpools. The other half were screened by a one-level procedure, bypassing the superpools and directly screening all subpools. Although the latter procedure involves more reactions,

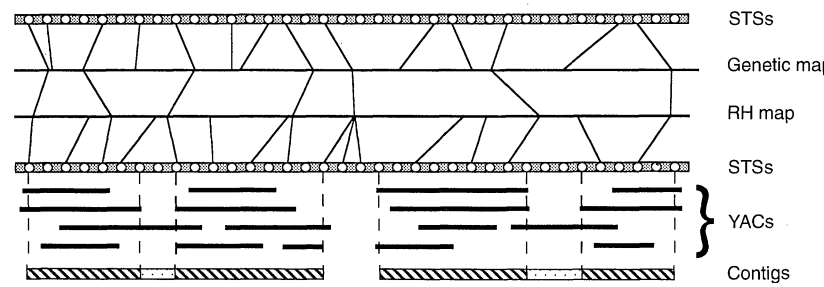


Fig. 1. Schematic diagram of the STS-based map. STSs are shown as circles on the first and fourth line. Loci that are genetically mapped or RH mapped are connected to the appropriate position on these maps, with connections between these maps in the cases of loci present in both maps. YACs containing STSs are shown below. The STSs fall into two singly linked contigs (stippled rectangles) and four doubly linked contigs (striped rectangles). Single linkage is not reliable for connecting arbitrary doubly linked contigs, but it is reliable in the case of anchored doubly linked contigs known to be adjacent on the genetic or RH map, as in the figure.

each locus is treated in an identical manner, which offers advantages for automation. In both procedures, we identified the positive pools by spotting the PCR reactions on membranes, hybridizing them to a chemiluminescent probe specific for each STS, capturing the resulting signal directly by a charge-coupled device (CCD) camera, and up-loading the results into our database (*28*); this approach proved to be much more efficient than the traditional detection procedure of gel electrophoresis.

Because the project involved processing more than 15 million reactions, laboratory automation was essential. We collaborated with an engineering firm, Intelligent Automation Systems, Incorporated, (IAS) of Cambridge, Massachusetts, to design and build various special-purpose machines to accelerate STS-based mapping.

Table 1. Overview of mapped STSs.

STSs on final map	No. of loci
STS-content map	10,850
RH map	6,193
Genetic map	5,264
Intersection of	
STS-content and RH maps	4,036
STS-content and genetic maps	3,106
RH and genetic maps	887
All three maps	807
Total loci	15,086

The two-level screening procedure was carried out with a large robotic liquid-pipetting workstation and two custom-designed thermocyclers (Fig. 2). A laboratory information management system used the superpool results to automatically program the robotic workstation to set up the appropriate subpool screens. The system has a maximal throughput of 6144 PCR reactions per run.

The one-level screening procedure was made feasible by the development of a massively parallel factory-style automation system nicknamed the Genomatron (Fig. 2). The Genomatron was also developed in collaboration with IAS and consists of three stations. The first station assembles PCR reactions in custom-fabricated 1536-well microtiter "cards" and seals the wells by welding a thin plastic film across the card. The second station thermocycles the reactions by transporting the cards over three chambers that force temperature-controlled water to flow uniformly between the cards. The third station transfers the reactions from one microtiter card onto a hybridization membrane affixed to the bottom of a second microtiter card by piercing the first card with a bed of 1536 hypodermic needles and sucking the reactions downward with a vacuum plenum. These "filter cards" were then manually hybridized with a chemiluminescent probe and read by the CCD camera. The stations were computer controlled, and the microtiter cards were assigned a bar code to facilitate sample tracking. Each station was designed to process 96 microtiter cards, providing a throughput of nearly 150,000 reactions per run.

STS-content mapping: Results. A total of 11,750 STSs yielded from 1 to 15 definite YAC addresses and were considered successfully screened (*29*); typical loci yielded approximately one additional incomplete address. STSs having more than 15 definite hits were excluded as likely to detect multiple genomic loci (*30*).

The successfully screened loci produced an average of 6.4 YACs per STS, considering only definite addresses. A total of 18,879 YACs were hit by at least one STS. For these YACs, the average hit rate was 3.8 STSs per YAC. The average size of the YACs hit by the STSs was about 1.1 Mb (~10% greater than for the library as a

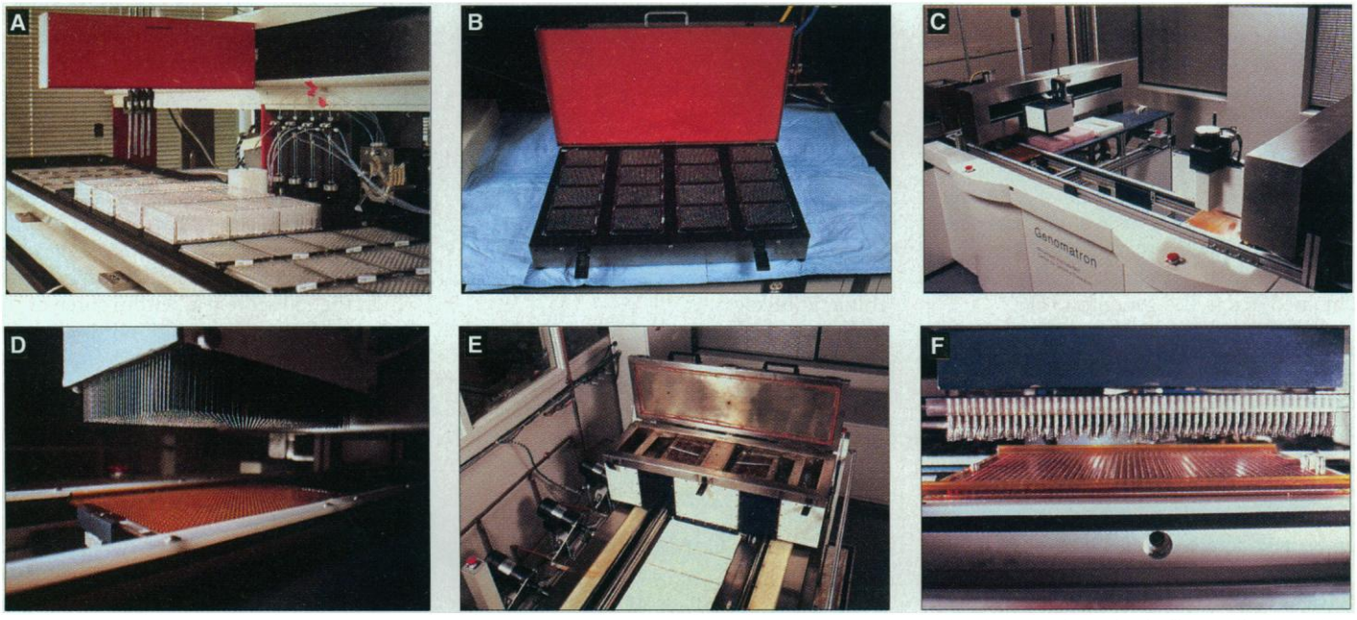


Fig. 2. The first automated system developed for the project was (**A**) a robotic station to set up PCR reactions and (**B**) custom-built "waffle iron" thermocyclers accomodating 16 192-well microtiter plates; the system has a capacity of 6144 PCR reactions per run. The second automated system was the Genomatron, which consists of three robotic stations. PCR reactions are set up in 1536-well microtiter cards (consisting of 15 cm by 24 cm injection molded plastic cards with 1536 holes, to the bottom of which a plastic film is heat-sealed to create wells). The first station (**C**) assembles the PCR reactions. Each run can process up to 96 cards per run, providing a capacity of nearly 150,000 wells. Cards are dispensed by a coining mechanism and travel along a conveyor belt to substations containing a bar code reader; a 1536-head pipettor (**D**) that dispenses template DNAs to be screened; a 48-head pipettor that dispenses PCR primer mixes, including polymerase; a plate sealer that heat-seals a plastic film on the top of the card to create separate reaction chambers; and a refrigerated storage station. The second station is a thermocycler (**E**) that uses three large waterbaths. Up to 96 sealed cards containing PCR reactions are placed in a chamber that travels over the water baths, which pump water at the appropriate denaturing, annealing, and extension temperature. The third station is a parallel "spotting" device that transfers PCR reactions from a card to a nylon filter affixed to the bottom of a second card. After the two cards are aligned, a bed of 1536 hypodermic needles (**F**) pierces a sealed card containing the reactions while a vacuum manifold draws the reaction mixtures down onto the membrane on the second. The filter cards are manually hybridized and subjected to a chemiluminescent detection protocol. Light signals are recorded with a cooled CCD camera.

whole), corresponding to 6.9-fold coverage of the genome. Some 78% of the STSs showed double-linkage to at least one other STS.

The false positive rate was investigated by regrowing and testing individual YACs. Several thousand addresses were tested, and 95% could be directly confirmed, with the remainder constituting actual false positives, deletions during regrowth, or technical failures during retesting. The false positive rate is thus at most 5% of definite addresses, and the chance of any particular YAC occurring as a false positive in a given screen is about 1.5×10^{-5}. False positive addresses thus will rarely create false links among STSs known to lie in the same genomic region. The false negative rate cannot be computed directly, but the fact that an average of 6.4 hits was seen in 8.4 genome equivalents suggests a rate of about 20%. False negatives pose a less serious problem than false positives (which join incorrect genomic regions), but they can lead to incorrect local ordering of STSs. The false positive and negative rates were reinvestigated once the maps were constructed, as discussed below.

Radiation hybrid mapping. STSs were screened against the GeneBridge 4 whole-genome radiation hybrid panel, consisting of 91 human-on-hamster somatic hybrid cell lines. Each line retains about one-third of the human genome in fragments of about 10 Mb in size. The GeneBridge 4 panel (Research Genetics, Huntsville, Alabama) was developed in the laboratory of P. Goodfellow and distributed to the scientific community as a resource for the mapping of expressed sequences. As part of a separate project, the panel has been characterized for more than 500 well-spaced genetic markers to confirm that substantial linkage can be obtained across the genome (*31*).

RH mapping was performed with essentially the same protocol as for the YAC screening: PCR reactions were set up either by the Genomatron (with each 1536-well microtiter card containing reactions for eight loci) or by the robotic workstation (by using 192-well microtiter plates), spotted on membranes, hybridized to a chemiluminescent probe, and detected by a CCD camera (*32*).

Scoring results from RH panels requires considerable caution. Human chromosomal fragments are present at various molarities among the hybrid cell lines; thus, the ability to detect their presence may vary with the sensitivity of each PCR assay. As a result, STSs that are immediately adjacent in the genome could conceivably give somewhat different retention patterns, which would limit the ability to determine fine-structure order. To minimize discrepancies due to assays near the limit of detection, we performed all assays in duplicate. Hybrids were scored if the two duplicates gave concordant positive or negative results but were recorded as "discrepant" if the duplicates were discordant. The mean discrepancy rate was 1.2%; loci with a discrepancy rate exceeding 4.5% were eliminated as unreliable.

A total of 6469 STSs were successfully screened on the GeneBridge 4 RH panel. The overall retention rate of the panel was 32% (or about 18% per haploid genome from the diploid donor cell).

Genetic mapping. Genetic linkage information was used from the recent Généthon linkage map of the human genome, containing 5264 polymorphic markers (*24*). Genetic linkage information was not incorporated for the 1722 CHLC genetic markers studied.

Chromosomal assignment. Before undertaking map construction, we attempted to assign all loci to specific chromosomes by multiple, independent methods. Most STSs were screened against the NIGMS 1 polychromosomal hybrid panel (*33*), resulting in unambiguous chromosomal assignment in about 75% of the cases (with the remainder having high background from the host genome or poor signal). STSs defining genetic markers typically had chromosomal assignments on the basis of linkage analysis. STSs were also assigned to chromosomes if they were tightly linked by RH screening or doubly linked by YAC screening to chromosomally assigned loci (*34*).

Some 96% of the loci could be chromosomally assigned, with the majority of these being assigned by at least two independent methods. Conflicting assignments were noted in a small proportion of cases (2%); these were subjected to intense scrutiny and resolved in the majority of cases (*35*). Loci that could not be reliably assigned to a chromosome were omitted from map construction, to avoid problems associated with chimeric linkages.

Personnel. The project was carried out during a period of 2.5 years by a team at Whitehead having an average of 16 people involved in mapping, three people involved in sequencing, and five people involved in data management and computational analysis.

Map Construction

Top-down maps. The genetic and RH maps are top-down maps, which provide a global framework and offer many tests of internal consistency. The first step in constructing an RH linkage map was to make high-quality "framework" maps across each chromosome. For this purpose, we included only loci with independent chromosomal assignments and with retention rates in the range of 10 to 60% (unusually high or low retention rates can produce spurious linkage). We wrote a computer package, RHMAPPER, that implements RH mapping for hybrids constructed from diploid sources and incorporates probabilistic error detection and error correction (*36*). Using this program, we generated a framework map—that is, an ordered set of markers such that each consecutive pair was linked with a lod score > 10 (lod score is the logarithm of the likelihood ratio for linkage), and the order was better than all local alternatives by a lod score > 2.5. The framework map included 1339 loci and provided complete connectivity across each chromosome arm with no gaps over 30 centiRays (cR) (cR is a measure of distance that is analogous to centimorgans but depends on the radiation dose). There were, however, large intervals across most centromeres (*37*), a phenomenon that has been previously seen for chromosome 11 (*38*). The total length of the map is 11,042 cR (omitting the centromeric intervals), corresponding to a fairly uniform average of about 300 kb/cR across most chromosomes.

We then localized the remaining markers relative to the framework map. These loci could not be uniquely ordered, either because of close proximity to a framework marker (loci with identical retention patterns cannot be ordered with respect to one another) or because of potentially erroneous typing results (that cause apparent "double-breaks" regardless of the interval in which the marker is placed). RHMAPPER allowed for the possibility of false positive and false negative typings and flagged probable errors (about two-thirds of which were found to be real errors in cases that were subsequently retested). The nonframework markers were estimated by the computer analysis to have an average residual error rate of just less than 1%. To reflect the uncertainty in order, each locus was assigned to the collection of intervals for which the lod score was within three of the optimal position. Loci were not included if they mapped more than 15 cR from a framework marker (that is, past the end of the map or in a large centromeric gap), because such positions could result from a high proportion of errors. In all, 6193 of 6469 loci tested were placed in the RH map.

Together, the two top-down maps contained a total of 10,572 loci. The reliability of the maps can be assessed by studying the loci in common. For loci present in both the genetic map and the framework RH map, there were only four conflicts in order; the loci involved were separated by 1 centimorgan (cM) in three cases and 3 cM in one case. The close agreement between the maps suggests that they correctly reflect the global order of loci in the genome.

Bottom-up map. Using the STS-content data, we assembled doubly linked contigs and checked that they did not connect loci known to map in different chromosomal regions. We then noted information about

single linkages among loci, which could provide connections between nearby doubly linked contigs in the course of integrating the top-down and bottom-up maps. Of the 11,750 STSs successfully screened against the YAC library, 10,850 (92%) showed single linkage to other STSs on the same chromosome. The remaining 8% were not included in the STS-content map.

Integrated map. We next sought to construct an integrated map by combining the STS-content, RH, and genetic linkage information. Each chromosome was treated separately: Only loci that had been assigned to the chromosome were used. Possible orders for the loci were compared by means of a linear scoring function, with the following three components: (i) continuity of STS content, reflecting whether the loci were present in the same YACs; (ii) continuity of RH linkage, reflecting whether the loci were present in the same RH hybrids; and (iii) consistency with top-down maps, incorporating a modest penalty for each violation of the genetic order or RH framework order. The specific parameters were chosen on the basis of the expected chance of concordance and discordance for nearby loci, so that the overall scoring function approximated a logarithm-likelihood for the order (*39*). The "optimal" order for the loci was found by combinatorial search through simulated annealing. Once the basic orders were established, incomplete addresses were used to identify additional links between nearby loci. The orders were then subjected to local optimization, manual inspection, and refinements where appropriate.

Gap closure. Loci fell into contigs of consecutive STSs connected by YACs and separated by gaps with no apparent YAC connection. Many of these apparent gaps are likely to be undetected overlaps; theoretical considerations would suggest that most gaps should actually be closed (*17*). We attempted to close these gaps by using non–STS-based information from the recent CEPH physical mapping project (*7*), inferring YAC overlaps on the basis of fingerprint analysis and Alu-PCR hybridization. Because the Alu-PCR hybridization data have a high false positive rate, gaps were closed only when there were at least seven hybridization links between adjacent contigs. Such closures should usually be correct, because only 3% of pairs of distant contigs meet this criterion. The data indicate overlap for about 50% of adjacent contigs. These gaps were declared tentatively closed, pending direct evaluation.

Description of the Map

The final map contains 15,086 loci, distributed across the 22 autosomes and two sex chromosomes (Tables 2, 3, and 4). The 10,850 loci mapped on YACs fall into 653 contigs connecting an average of 17 STSs each before gap closure and 377 contigs with an average of 29 STSs after gap closure. We examined the local density of YAC hits and contigs across the length of each chromosome. The results were relatively similar across the genome, with the notable exception of the chromosomes 1p36, 19, 22, and X. The map has less continuity in these regions, apparently because of systematic underrepresentation in the CEPH Mega-YAC library (see YAC density in Table 4), a problem that has been previously noted (*7*). Chromosome X is underrepresented because the library was made from a male cell line. The autosomal deficits could reflect cloning biases of the yeast host, inasmuch as these are all regions of high GC content (*40*).

The physical map contains a wealth of information, which is ill-suited for presentation in traditional printed form. The complete physical map—including the STS sequences, RH retention patterns, YAC addresses, and order of loci—would require more than 900 journal pages to display. A compressed view of chromosome 14 is shown in Fig. 3, to illustrate the general nature of the map. The complete data for the map can be freely accessed through a World Wide Web server at the Whitehead Institute (http://www-genome.wi.mit.edu/), which includes various tools for analysis.

Table 2. Types of STSs. Chr., chromosome.

Chr.	Total STSs	Random STS*	Genes		Genetic markers		Other loci	ESTs (obs/exp)†
			ESTs	GenBank	Généthon	CHLC		
1	1,374	252	275	106	460	153	128	1.4
2	1,275	307	181	67	452	146	122	0.8
3	1,097	269	181	64	353	134	96	0.9
4	919	210	112	45	281	121	150	0.7
5	858	196	125	30	312	97	98	0.8
6	858	181	114	39	312	108	104	0.8
7	781	168	141	39	272	83	78	1.1
8	739	183	104	35	248	104	65	0.7
9	577	132	106	30	188	68	53	1.1
10	719	154	131	26	281	60	67	1.1
11	706	122	140	42	272	64	66	1.5
12	707	132	104	64	250	91	66	1.0
13	418	102	48	13	164	54	37	0.6
14	489	106	95	27	163	53	45	1.2
15	428	97	97	22	145	30	37	1.3
16	435	87	79	18	180	32	39	1.2
17	447	66	97	39	186	34	25	1.9
18	403	91	46	18	136	64	48	0.7
19	246	23	45	20	121	15	22	2.6
20	386	84	68	26	144	32	32	1.1
21	156	28	18	12	61	13	24	0.8
22	274	19	38	17	67	12	122	2.6
X	587	145	63	28	216	28	107	0.6
Y‡	207	0	0	0	0	0	207	
Total	15,086	3,154	2,408	827	5,264	1,595	1,838	1.0

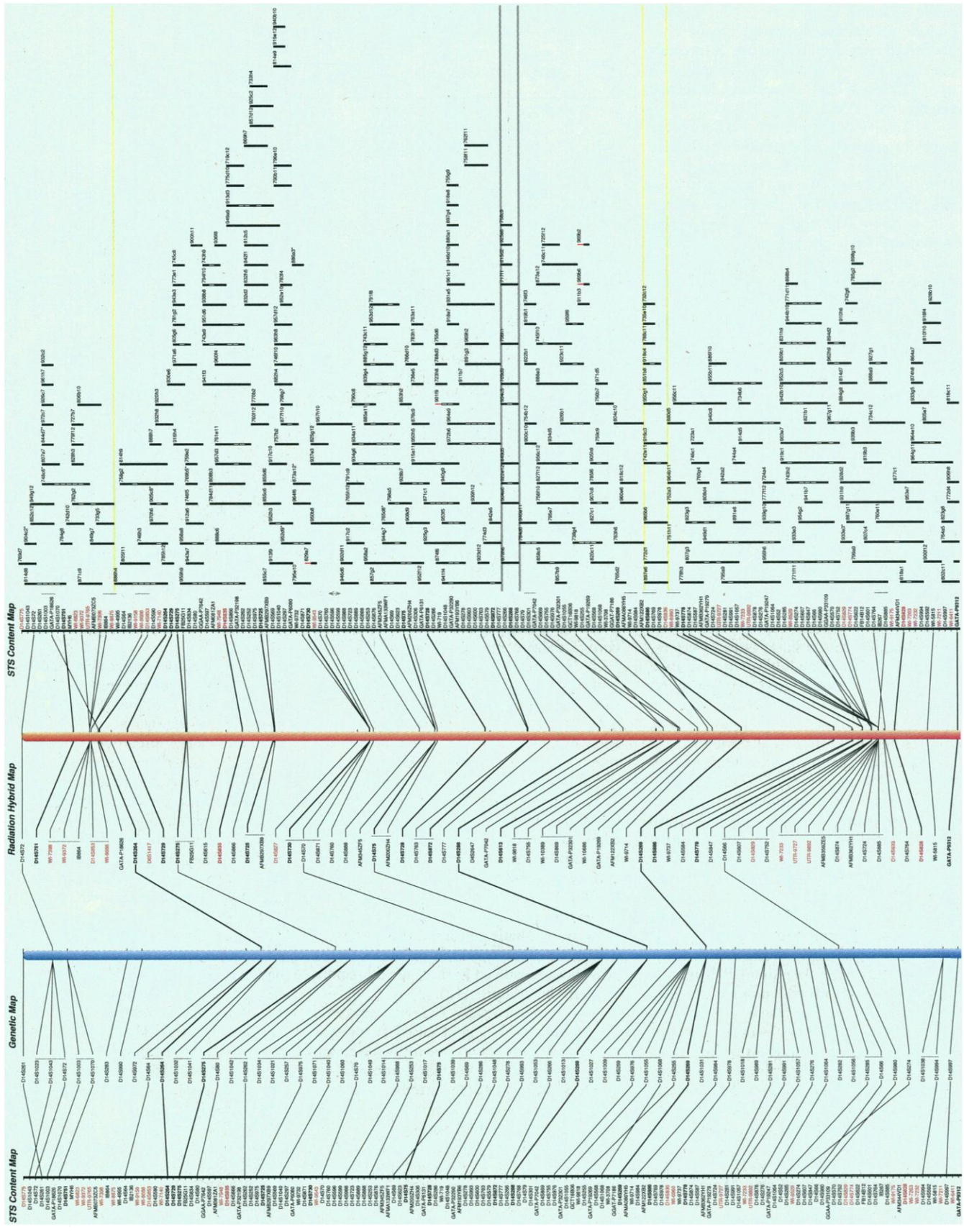
*Unbiased STSs, generated by sequencing from a random genomic library. †Ratio of observed (obs) number of ESTs divided by expected (exp) number, assuming that ESTs follow the same distribution as random STSs. ‡From our previously reported work (*3*).

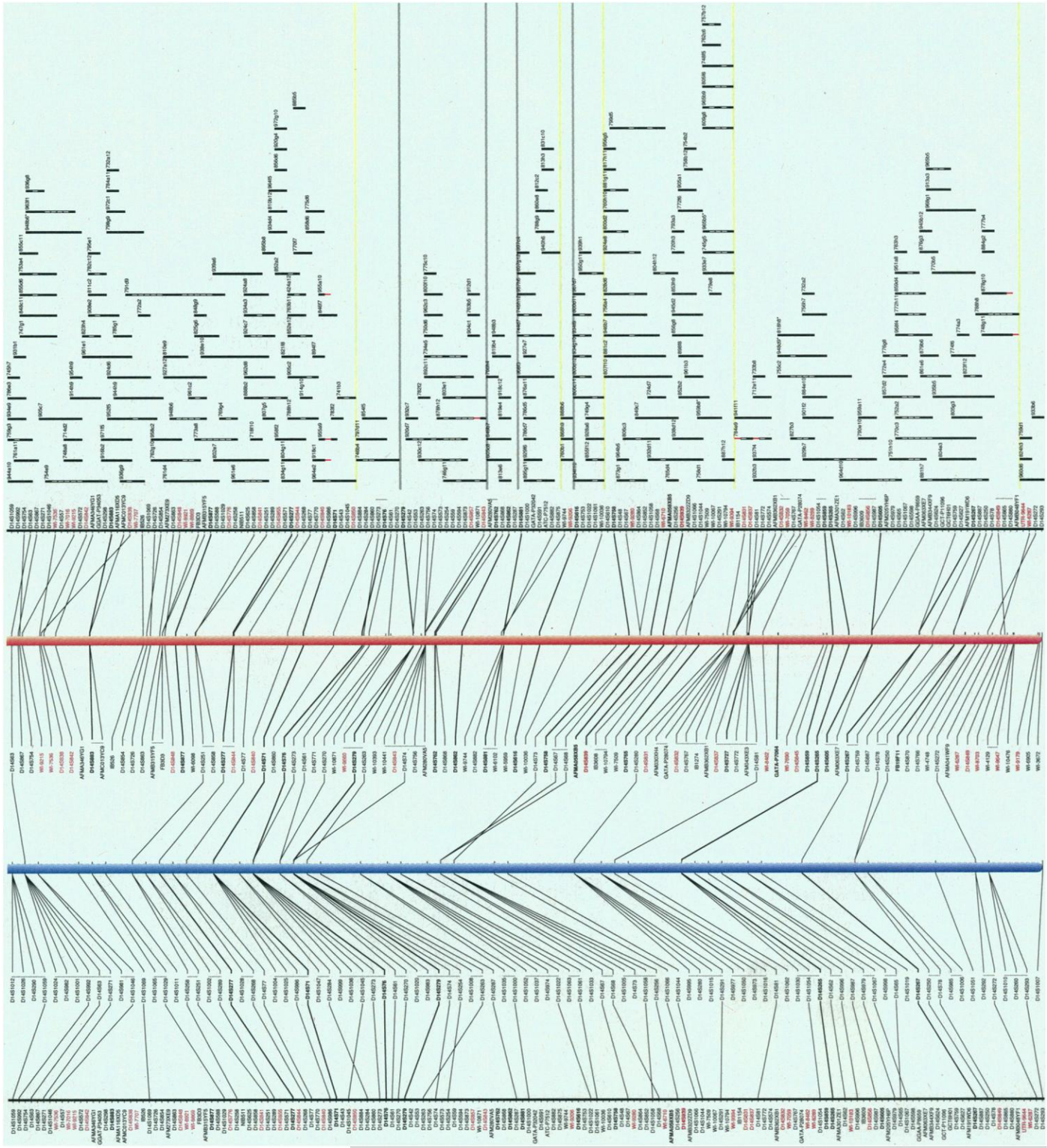
Coverage

We sought to determine how much of the human genome is covered by the physical map. For this purpose, we derived a new collection of random STSs—by sequencing random clones from an M13 library, selecting PCR primers, and retaining those loci that gave consistent amplification of a single fragment in control experiments. The first 100 STSs produced in this fashion were then screened against the NIGMS 1 hybrid panel, the RH panel, and the YAC library. Because the goal was to obtain an unbiased assessment of coverage, special efforts were made to obtain complete data for each locus.

RH data was obtained for all 100 STSs. (In six cases, it was necessary to resort to acrylamide gel electrophoresis of radioactively labeled products to circumvent problems posed by rodent background.) All 100 loci could be positioned on the RH map with a lod ≥ 8, on the correct chromosome as determined by the polychromosomal hybrid panel (*41*). The RH map thus appears to cover the vast majority of the human genome.

YAC screening data was also obtained for all 100 STSs. Two STSs detected no YACs in the library, consistent with previous observations that about 2% of DNA sequences appear to be absent from the CEPH Mega-YAC library (*7*). Four STSs detected YAC hits, but none with links to another STS in the correct chromosomal region; these loci


STS Content Map
Radiation Hybrid Map
Genetic Map
STS Content Map



could thus not be localized on the STS-content map (*42*). These four STSs appear to be in regions of low YAC coverage, inasmuch as they hit one, one, one, and two YACs, respectively. The remaining 94 STSs could all be localized on the STS content map [with 91 being doubly linked and three being singly linked to existing contigs anchored in the correct chromosomal region in the top-down map (*43*)]. The 100 loci detected an average of 6.5 YACs.

The map covers the vast majority of the human genome. We estimate that 99% of random STSs can be readily positioned on the RH map, and 94% can be positioned on the STS-content map relative to YAC clones.

The physical map thus fills a major need in human genetics, providing a general method by which an investigator can map a locus in the human genome by screening readily available RH or YAC pools and comparing the resulting pattern with the map. To make this information easily accessible to the scientific community, we have written a "map server." The server reports the likely position of an STS given information about its YACs, RH pattern, and chromosomal assignment. The server is freely available via our World Wide Web site.

Fig. 3 (previous pages). Integrated map of human chromosome 14q. Long vertical lines represent the STS-content map (first and fourth lines, in black), genetic map (second line, in blue), and RH map (third line, in orange), in the same fashion as the diagram in Fig. 1. All three maps are drawn to equal length. The four columns of STS names correspond to the four lines. For the STS-content map, intermarker distance is not known and loci are displayed as equally spaced. For genetic and RH maps, loci are indicated at positions spaced proportionally along the map according to the respective metrics. Loci in common between two maps are connected by black lines. Loci belonging to the RH framework map (in which the relative ordering is supported by lod > 2.5) are shown in bold type and with thicker connecting lines. Loci derived from expressed sequences are shown in purple. YACs are displayed as black rectangles, to the right of the STSs that were found to be contained in the clone. YAC names are shown to the top right. Unfilled portions of YACs represent assays that were negative. Thin red lines in some YACs represent incomplete addresses that were resolved by virtue of overlap with addresses from a nearby locus. Gaps between contigs are shown as horizontal lines separating groups of YACs. Gaps that were likely to be undetected overlaps based on Alu-PCR hybridization or fingerprint information (see text) are shown in yellow; gaps for which there is no evidence of overlap are shown in gray. Vertical dotted gray bars indicate STSs with identical data for given mapping method. YACs detected by only a single STS were omitted from this display. These YAC addresses can be obtained from the Whitehead Institute–MIT Center for Genome Research World Wide Web server at URL http://www-genome.wi.mit.edu/. Figure represents slightly earlier version of the map, from the 14,000-marker stage.

Accuracy

Although the long-range order of the map is reliable because of top-down anchoring, precise local orders must be regarded as only approximate. Local ordering depends on the position of loci with respect to individual breakpoints, that is, the ends of YAC or RH fragments. The accuracy of such inference is limited by the presence of false positives and false negatives in our data, as well as by the presence of internal deletions in YACs. Whereas the long-range order tends to be over-determined in genomic maps, several alternative local orders may be reasonably compatible with the data. The "best" order may change with the alteration of a few data points.

We used three approaches to evaluate the accuracy of the data and the map.

1) Rescreening of loci on chromosome 14. Chromosome 14 was divided into 16 regions and regional YAC panels were defined, consisting of all clones hit by one or more loci in the region. For each regional YAC panel, individual DNAs were prepared from each clone. We tested 112 STSs against their corresponding panels to directly compare the results from high-throughput screening of pools with the screening of individual clones. We found a false positive rate of 5.5% and a false negative rate of 19.5% in our high-throughput screening data, both of which were consistent with our earlier indirect estimates. We constructed a new STS map of the chromosome using these more complete data; the new map showed about six instances of local reorderings involving two to five loci.

2) Comparison with an independently constructed map of chromosome 12. We compared our map with a recently reported map of this chromosome (*10*) containing enough loci in common to provide a meaningful test. Of 171 loci in common, there were about a dozen instances of small local inversions involving two to three adjacent markers. A substantial difference in position was seen for only a single marker, AFM263WH1. Our map shows tight STS-content linkage of this locus to genetic markers at 91 cM on the Généthon map, whereas the other map places it near genetic markers at 105 cM. In fact, the position on our map agrees well with the reported genetic map location for this locus (at 93 cM), so we believe it to be correct. In any case, the two maps showed relatively few conflicts.

3) Internal consistency checking. We looked for instances in which pairs of loci occurred in an order on the final STS map that was strongly disfavored by the RH or

Table 3. Genetic and RH maps. Dashes indicate not applicable.

Chr.	Physical length (Mb)*	Genetic map			RH map			
		No. of loci	Length (cM)	Genetic vs. physical (cM/Mb)	Framework map No. of loci	Total RH map	RH length (cR)†	RH vs. Physical (cR/Mb)
1	248	461	293	1.2	107	559	743	3.0
2	240	452	277	1.2	119	532	977	4.1
3	202	353	233	1.2	95	475	801	4.0
4	191	280	212	1.1	80	370	552	2.9
5	183	312	198	1.1	60	339	508	2.8
6	173	311	201	1.2	97	374	739	4.3
7	161	272	184	1.1	63	360	591	3.7
8	146	249	166	1.1	77	264	711	4.9
9	137	189	166	1.2	75	260	440	3.2
10	136	281	182	1.3	71	297	599	4.4
11	136	273	156	1.1	66	302	515	3.8
12	135	249	169	1.3	58	294	565	4.2
13	92	164	117	1.3	46	169	309	3.3
14	88	162	129	1.5	38	210	319	3.6
15	84	145	110	1.3	41	185	342	4.1
16	92	180	131	1.4	33	186	235	2.5
17	87	186	129	1.5	34	156	347	4.0
18	80	136	124	1.5	52	175	450	5.6
19	63	121	110	1.7	21	107	221	3.5
20	68	144	96	1.4	30	157	265	3.9
21	37	61	60	1.6	15	61	151	4.1
22	41	67	58	1.4	15	89	141	3.5
X	155	216	198	1.3	46	272	521	3.4
Y	26	–	–	–	–	–	–	–
Total	3,000	5,264	3,699	1.2	1,339	6,193	11,042	3.7

*Physical lengths were calculated on the basis of a genome of 3000 Mb, with proportional lengths of chromosomes as reported (*46*). †Total length of the RH framework map, omitting the large interval at the centromere.

genetic maps. By measuring the frequency of such occurrences as a function of the distance between the loci in the STS map, we estimated that about 0.5% of the loci may be significantly misplaced in the maps.

In summary, the local order in the map must be regarded as uncertain. There will surely be many errors requiring attention and correction. The effective resolution of the map is certainly lower than the average spacing between loci and may be about 1 Mb. To improve the local accuracy of the maps, investigators interested in particular regions would be well advised to retest the STSs against an RH panel with higher resolution [such as the G3 panel developed by D. Cox and R. Myers, in which the fragments are about 1/10 those in the GeneBridge 4 panel] and against regional YAC panels, as described above. In this fashion, the map provides the tools for its own refinement.

Finally, we note that direct comparison of our STS-based map with the recently reported YAC-based map (*7*) is difficult, because of the very different natures of the maps. For example, it is not meaningful to compare the STS orders in the two maps: The YAC-based map almost exclusively involved genetic markers and provided no independent information about locus order, but instead simply adopted the genetic order. It is also problematic to compare the specific YACs identified, because the YAC-based map involved only partial screening of most STSs and did not fully distinguish clones actually belonging to paths through a region from those representing false positive hybridization. At the grossest level, it is possible to compare the coverage of the maps: The current map appears to cover about 95% of the genome (the precise amount depends on the type of mapping information used), whereas the other map was reported to cover about 75%. More detailed comparison would be worthwhile, as it would likely lead to improvements in both maps.

Distribution of Genes

The map also sheds light on the organization of the human genome. By comparing the chromosomal distribution of the expressed sequences to the chromosomal distribution of the random single-copy sequences (both determined in the same manner), one can draw inferences about the density of genes on different chromosomes. We compared the observed number on each chromosome to the expected number, assuming that expressed sequences have the same distribution as random STSs (Table 2). Chromosomes 1, 11, 17, 19, and 22 showed a statistically significant excess of expressed sequences ($P = 0.001$ after correction for multiple testing). Chromosomes 17, 19, and 22, which showed the greatest excess, have been previously suggested to have a high density of genes on the basis of indirect evidence (*40*). Chromosome X was the only chromosome to show a statistically significant deficit of expressed sequences—only about half as many as expected. This would suggest that there is a low gene density on this sex chromosome, although alternative explanations are possible (*44*). We also analyzed the raw data from two recent papers reporting chromosomal assignment of expressed sequence tags (ESTs) (*21, 22*) and found a similar deficit of X-linked loci.

A Scaffold for Sequencing the Genome

As genetic and physical maps approach their intended goals, attention is turning to the challenge of sequencing the entire human genome. A key issue is how to obtain the required sequence-ready clones. STS-based maps provide a general solution by making it possible to generate extensive physical coverage of a region by screening a single high-quality human genomic library.

One could, for example, proceed as follows: Screen the STSs in a region against a bacterial artificial chromosome (BAC) library having 150-kb inserts and 10-fold coverage, use a simple fingerprinting scheme to detect overlaps among adjacent clones, and select a minimally overlapping set for sequencing. Given a physical map containing 30,000 ordered STSs, one would screen about 100 STSs and fingerprint about 520 BACs to cover a 10-Mb region; this task could be readily accomplished in a few days with modest automation and would not contribute significantly to the cost of sequencing. The resulting BACs would be expected to cover about 95% of the region in ordered sequence contigs (*17*). The region could then be closed by straightforward walking—that is, serially screening the BAC library with STSs derived from sequences at the ends of each contig.

The current map falls short in terms of marker density and local order, but neither shortcoming poses a serious obstacle for initiating large-scale sequencing now. With the 15,000 STSs currently available, one should cover about 75% by direct screening, 90% by one round of walking, and more than 95% with two rounds (*17*). The desired map with 30,000 STSs will likely be available within the next 2 years, through current projects underway at several centers including our own. Uncertainties about locus order can be overcome simply by scoring the STSs from

Table 4. STS-content mapping of YACs.

Chr.	STS-content mapped loci	STS spacing (kb/STS)	No. of YACs*	Contigs			YAC hits per STS‖	STSs per contig¶
				Before gap closure†	After gap closure‡	Avg. size (Mb)§		
1	1,048	237	1,393	49	34	7.3	6.7	30.8
2	933	258	1,469	56	20	12.0	7.3	46.7
3	791	255	1,192	46	30	6.7	7.5	26.4
4	718	267	1,272	42	11	17.4	8.2	65.3
5	651	281	1,163	35	19	9.6	7.9	34.3
6	641	269	1,091	40	24	7.2	7.8	26.7
7	559	288	942	39	13	12.4	7.8	43.0
8	552	265	945	23	11	13.3	8.0	50.2
9	394	347	675	28	12	11.4	8.0	32.8
10	519	262	750	36	26	5.2	6.8	20.0
11	490	277	696	23	14	9.7	7.2	35.0
12	509	265	842	29	16	8.4	7.4	31.8
13	300	308	556	12	5	18.5	8.0	60.0
14	352	249	593	9	6	14.6	8.0	58.7
15	301	279	439	16	10	8.4	7.0	30.1
16	255	362	308	26	16	5.8	6.0	15.9
17	267	325	330	27	17	5.1	5.8	15.7
18	315	254	478	16	8	10.0	7.6	39.4
19	79	800	76	17	15	4.2	4.7	5.3
20	266	255	328	15	10	6.8	6.6	26.6
21	113	325	182	4	2	18.4	8.0	56.5
22	182	223	134	11	11	3.7	5.4	16.5
X	408	379	406	53	46	3.4	4.7	8.9
Y#	207	128	234	1	1	26.4	4.1	207.0
Total	10,850	276	16,494	653	377	8.0	7.3	28.8

*Number of YACs hit by at least two STSs on the chromosome. YACs hit by only one STSs are omitted. †Number of contigs based only on STS-content data. ‡Number of contigs after gap closure by Alu-PCR hybridization and fingerprint data. §Average contig size estimated by assuming that 94% of the chromosome length is covered and dividing by the number of contigs. ‖Includes all YAC hits in the library screen (not limited to YACs having multiple STS hits on the chromosome) and thus reflects coverage by the library. ¶Reflects contigs after gap closure. #From our previously reported work (*3*).

the relevant region on a high-resolution RH panel in parallel with screening them on the BAC library. As a simple test, we scored the STSs from a 3-Mb region on chromosome 6 on the G3 RH panel and were able to readily infer the fine-structure order of nearly all the loci with high confidence (*45*).

The use of STS-based maps as a scaffold for large-scale sequencing has several advantages: It can be initiated now with the existing STS-based map; it automatically anchors sequences in the genome; it does not require chromosome-specific libraries, which involve specialized preparation procedures and often have cryptic biases; it allows improved libraries to be substituted as they become available; and it promotes decentralization by allowing sequencing efforts to focus on regions of any given size, in contrast to entire chromosomes.

In summary, the physical map must still be refined but is already adequate to allow initiation of the international project to sequence the entire human genome—a landmark effort that will set the stage for the biology of the next century.

REFERENCES AND NOTES

1. M. Olson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7826 (1986); A. Coulson *et al.*, *ibid.*, p. 7821; Y. Kohara *et al.*, *Cell* **50**, 495 (1987).
2. M. Olson, L. Hood, C. Cantor, D. Botstein, *Science* **245**, 1434 (1989).
3. S. Foote, D. Vollrath, A. Hilton, D. C. Page, *ibid.* **258**, 60 (1992).
4. I. Chumakov *et al.*, *Nature* **359**, 380 (1992).
5. F. Collins and D. Galas, *Science* **262**, 43 (1993).
6. L. Selleri *et al.*, *Genomics* **14**, 536 (1992); M. Haldi *et al.*, *ibid.* **24**, 478 (1994).
7. I. Chumakov *et al.*, *Nature* **377**, S175 (1995).
8. H. Shizuya *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8794 (1992); P. A. Ioannous *et al.*, *Nature Genet.* **6**, 84 (1994).
9. R. M. Gemmill *et al.*, *Nature* **377**, S299 (1995); N. A. Doggett *et al.*, *ibid.*, p. S335; J. E. Collins *et al.*, *ibid.*, p. S367.
10. K. Krauter *et al.*, *ibid.*, p. S321.
11. E. D. Green *et al.*, *Hum. Mol. Genet.* **3**, 489 (1994); J. Quackenbush *et al.*, *Genomics* **29**, 512 (1995); T. Alitalo *et al.*, *ibid.* **25**, 691 (1995).
12. E. D. Green and M. V. Olson, *Science* **250**, 94 (1990).
13. D. R. Cox, M. Burmeister, E. Roydon Price, S. Kim, R. M. Myers, *ibid.*, p. 245; M. A. Walter *et al.*, *Nature Genet.* **7**, 22 (1994).
14. D. Botstein *et al.*, *Am. J. Hum. Genet.* **32**, 314 (1980); J. L. Weber and P. E. May, *ibid.* **44**, 388 (1989).
15. J. Murray *et al.*, *Science* **265**, 2049 (1994).
16. J. Ott, *Analysis of Human Genetic Linkage* (Johns Hopkins Press, Baltimore, MD, 1991).
17. R. Arratia *et al.*, *Genomics* **11**, 806 (1991).
18. Loci were used only if they produced a single clear band visible by ethidium bromide staining, except for genetic markers, which were used even when they produced more than one band on an agarose gel (in view of their value in providing top-down orientation). Assays meeting this criterion are more likely, although not certain, to represent single unique loci in the genome.
19. Sequence data were analyzed with the Whitehead/MIT STS Pipeline software, which removes vector sequences, identifies duplicate sequences, and uses sequence similarity programs (FASTA and BLASTN) to eliminate known repeat sequences. Primers were chosen with PRIMER (M. J. Daly, S. Lincoln, E. S. Lander, Whitehead Institute) having the desired T_m (temperature at which 50% of double-stranded DNA is denatured) for primers set at 58°C.
20. M. Boguski and G. Schuler, *Nature Genet.* **10**, 369 (1995).
21. Nonredundant ESTs were part of the UniEST set prepared by M. Boguski and G. Schuler at the National Center for Biotechnology Information, derived from the dbEST database. These ESTs were taken from the Washington University–Merck EST project and the GenExpress project; R. Houlgatte *et al.*, *Genome Res.* **5**, 272 (1995).
22. R. Berry *et al.*, *Nature Genet.* **10**, 415 (1995); M. D. Adams *et al.*, *Nature* **377**, S3 (1995).
23. To select primers from ESTs, we modified the STS pipeline (*19*) to select shorter PCR products of 100 to 150 bp near the end of the 3′-UTRs {but 20 bp away from the polyadenylate [poly(A)] tract} in a region of high sequence quality.
24. J. Weissenbach *et al.*, *Nature* **359**, 794 (1992); G. Gyapay *et al.*, *Nature Genet.* **7**, 247 (1994); C. Dib *et al.*, *Nature*, in press.
25. C. J. Bell *et al.*, *Hum. Mol. Genet.* **4**, 59 (1995).
26. STSs were kindly shared by D. Cox and R. Myers, Stanford University, Stanford, CA, and J. Gastier, Harvard University, Cambridge, MA.
27. Some incomplete addresses could be resolved by a simple band-matching test with complete addresses by using CEPH fingerprint data, as described (*25*). Others could be resolved by virtue of comparison with complete addresses for nearby STSs.
28. Dot-blotted PCR products were initially detected by using ECL kits (Amersham), as described (*25*). We later switched to overnight hybridization with a biotinylated oligonucleotide probe to an internal sequence, followed by chemiluminescent detection with a peroxidase catalyzed luminol reaction, as described [R. P. M. Gijlswijk *et al.*, *Mol. Cell. Probes* **6**, 223 (1992)]. STSs known to contain an internal repeat sequence such as CA or AGAT were probed with an oligonucleotide for the repeat. Other STSs were probed with a specific internal oligonucleotide, having a T_m of 58°C. Computer images of each hybridization were obtained with a CCD camera. VIEW software (C. Rosenberg; Whitehead Institute) was used to locate and determine the intensity of positive dots. A small proportion of STSs were screened by standard agarose gel stained with ethidium bromide.
29. It is not possible to draw conclusions about library coverage from the overall number of STSs with no definite addresses, because many of these represented weak PCR assays that sometimes worked on human control DNAs but failed on YAC pools.
30. The probability that a unique sequence would occur more than 15 times in a random library with 8.4-fold coverage is about 1%. Some of these STSs may thus be unique loci, but they were excluded to guard against repeats.
31. G. Gyapay *et al.*, *Hum. Mol. Genet.*, in press.
32. Dot-blotted PCR products were detected as for the YAC screening, except that STSs containing CA repeats were screened with oligonucleotides containing unique internal sequence rather than $(CA)_n$, because the latter produced high background.
33. H. L. Drwinga *et al.*, *Genomics* **16**, 311 (1993); T. J. Hudson *et al.*, *ibid.* **13**, 622 (1992). In 300 cases, ambiguous or conflicting assignments were resolved by using the NIGMS Human/Rodent Somatic Cell Hybrid Panel #2, described in B. L. Dubois and S. Naylor *et al.*, *ibid.* **16**, 315 (1993).
34. In 151 cases, STSs were chromosomally assigned by virtue of having at least three single links to other markers on a chromosome and no links to any loci on any other chromosome.
35. About 200 such conflicts were resolved. Half were resolved by repeating the typing of the somatic cell hybrid panel. For the other half, an STS was demonstrated to amplify products from more than one chromosome. Such STSs were discarded.
36. RHMAPPER (L. Kruglyak, D. K. Slonim, L. D. Stein, E. S. Lander, unpublished data) uses a hidden Markov model to account for breaks in diploid DNA and for false positives and negatives, as in E. S. Lander and P. Green, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 2363 (1987); and E. S. Lander and S. Lincoln, *Genomics* **14**, 604 (1992). Framework maps were initiated and extended by a greedy algorithm and then subjected to local permutation tests, thereby allowing for efficient exploration of a vast space of possible orders.
37. In most cases, frameworks for the two chromosomal arms were constructed separately and then oriented and joined by using information from the genetic map. There is significant pairwise RH linkage (at lod > 5.0) between framework markers on opposite sides of the centromere on nine of the final framework maps, but not on the other 14.
38. M. R. James *et al.*, *Nature Genet.* **8**, 70 (1994).
39. For each pair of consecutive STSs, a positive score a_{++} was added for each YAC containing both, a negative score a_{+-} for each YAC containing one but not the other, a positive score b_{++} for each hybrid containing both, and a negative score b_{+-} for each hybrid containing one but not the other. Letting x_{11}, x_{10}, and x_{00} denote the probabilities that two STSs separated by about 500 kb would be observed to be both present, both absent, or one present and one absent in a randomly chosen YAC, we set $a_{++} = \log(x_{11}/x_{00})$ and $a_{+-} = \log(x_{10}/x_{00})$. The constants b_{++} and b_{+-} were defined similarly with respect to individual RH cell lines. The various probabilities were calculated on the basis of the distribution of fragment sizes and the inferred false positive and false negative rates. These four parameters, a_{++}, a_{+-}, b_{++}, and b_{+-}, were thus determined directly from the data and were not optimized. The two weighting parameters for conflicts with the genetic and framework RH maps were chosen by optimization in test cases.
40. S. Saccone *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 4913 (1992).
41. Three markers mapped into large centromeric intervals on the correct chromosome; they had high lod scores but were about 30 cR away from the closest marker. All were confirmed by double-linkage with YACs. For three other markers, chromosomal assignment could not be obtained from polychromosomal hybrid panels because of rodent background.
42. For one of these four loci, there was a (presumably chimeric) single YAC link to a marker on the same chromosome but located 70 cR from the correct location.
43. Three of the loci belonged to doubly linked contigs that were anchored by virtue of a CHLC genetic marker.
44. If gene promoters on chromosome X have the same average expression level as on autosomes, then the fact that only one X chromosome is active (due to hemizygosity in males and X inactivation in females) would cause transcripts from X-linked genes to be half as abundant. Because half of the cDNAs came from nonnormalized libraries and half from normalized libraries, the occurrence of ESTs in the relatively small set examined will partly reflect abundance. This issue will recede when enough ESTs have been isolated to overcome issues related to message levels. Underrepresentation of chromosome X could also conceivably represent some other systematic bias of which we are not aware.
45. T. J. Hudson *et al.*, data not shown.
46. J. S. Heslop-Harrison *et al.*, *Hum. Genet.* **84**, 27 (1989).

47. We thank A. Kaufman, O. Merport, and J. Spencer for technical assistance; L. Bennett for computer system administration; G. Rogers and M. Foley for assistance with media preparation and glass washing; S. Gordon, A. Christopher, P. Mansfield, and others at Intelligent Automation Systems for assistance in design, construction, and maintenance of automation equipment; D. Cox, R. Myers, J. Sikela, M. Adams, J. Murray, and K. Buetow for sharing data including sequences and markers; M. Boguski and G. Schuler for assistance in analysis of EST sequences; and D. Cohen and I. Chumakov for sharing the CEPH YAC library in 1992 and for public distribution of their STS-content, Alu-PCR, and fingerprint data. Supported by NIH award HG00098 to E.S.L. and by the Whitehead Institute for Biomedical Research. T.J.H. was a recipient of a Clinician Scientist Award from the Medical Research Council of Canada. L.K. is a recipient of a Special Emphasis Research Career Award (HG00017) from the National Center for Human Genome Research.

17 October 1995; accepted 8 November 1995