

# A High-Resolution Radiation Hybrid Map of the Human Genome Draft Sequence

Michael Olivier,<sup>1</sup> Amita Aggarwal,<sup>1</sup> Jennifer Allen,<sup>1</sup>  
 Annalisa A. Almendras,<sup>1</sup> Eva S. Bajorek,<sup>1</sup> Ellen M. Beasley,<sup>1\*</sup>  
 Shannon D. Brady,<sup>1</sup> Jannette M. Bushard,<sup>1</sup> Valerie I. Bustos,<sup>1</sup>  
 Angela Chu,<sup>1</sup> Tisha R. Chung,<sup>1</sup> Anniek De Witte,<sup>1</sup>  
 Mirian E. Denys,<sup>1</sup> Rakly Dominguez,<sup>1</sup> Nicole Y. Fang,<sup>1</sup>  
 Brian D. Foster,<sup>1</sup> Robert W. Freudenberg,<sup>1</sup> David Hadley,<sup>1</sup>  
 Libby R. Hamilton,<sup>1</sup> Tonya J. Jeffrey,<sup>1</sup> Libusha Kelly,<sup>1</sup>  
 Laura Lazzeroni,<sup>1</sup> Michelle R. Levy,<sup>1</sup> Saskia C. Lewis,<sup>1</sup> Xia Liu,<sup>1</sup>  
 Frederick J. Lopez,<sup>1</sup> Brent Louie,<sup>1</sup> Joseph P. Marquis,<sup>1</sup>  
 Robert A. Martinez,<sup>1</sup> Margaret K. Matsuura,<sup>1</sup>  
 Nedda S. Misherghi,<sup>1</sup> Jolanna A. Norton,<sup>1</sup> Adam Olshen,<sup>1†</sup>  
 Shanti M. Perkins,<sup>1</sup> Amy J. Perou,<sup>1</sup> Chris Piercy,<sup>1</sup> Mark Piercy,<sup>1</sup>  
 Fawn Qin,<sup>1</sup> Tim Reif,<sup>1</sup> Kelly Sheppard,<sup>1</sup> Vida Shokoohi,<sup>1</sup>  
 Geoff A. Smick,<sup>1</sup> Wei-Lin Sun,<sup>1</sup> Elizabeth A. Stewart,<sup>1‡</sup>  
 J. Fernando Tejada,<sup>1</sup> Nguyet M. Tran,<sup>1</sup> Tonatiuh Trejo,<sup>1</sup>  
 Nu T. Vo,<sup>1</sup> Simon C. M. Yan,<sup>1</sup> Deborah L. Zierden,<sup>1</sup>  
 Shaying Zhao,<sup>2</sup> Ravi Sachidanandam,<sup>3</sup> Barbara J. Trask,<sup>4</sup>  
 Richard M. Myers,<sup>1</sup> David R. Cox<sup>1§</sup>

We have constructed a physical map of the human genome by using a panel of 90 whole-genome radiation hybrids (the TNG panel) in conjunction with 40,322 sequence-tagged sites (STSs) derived from random genomic sequences as well as expressed sequences. Of 36,678 STSs on the TNG radiation hybrid map, only 3604 (9.8%) were absent from the unassembled draft sequence of the human genome. Of 20,030 STSs ordered on the TNG map as well as the assembled human genome draft sequence and the Celera assembled human genome sequence, 36% of the STSs had a discrepant order between the working draft sequence and the Celera sequence. The TNG map order was identical to one of the two sequence orders in 60% of these discrepant cases.

The ultimate map of any organism is the complete sequence of its genome. Over the past several years, an international consortium of scientists has taken advantage of technical improvements in DNA sequencing and mapping technologies to generate a working draft of the human genome sequence (1). The first step in this effort involved the

construction of bacterial artificial chromosomes (BACs), each containing a stable segment of approximately 160 kilobase pairs (kbp) of the human genome (2). Collections of human BACs estimated to represent more than a 10-fold redundancy of the human genome were used to generate BAC maps (3, 4). This process resulted in a minimally redundant set of BACs, assembled into physically separate contigs, representing the majority of the human genome and serving as the substrate for large-scale DNA sequencing (3). The human DNA segment of each BAC was fragmented into a redundant collection of smaller clones, which were each sequenced and assembled into a limited number of sequence contigs per BAC. Finally, all DNA sequence within each physical contig consisting of multiple BACs was assembled. Given that the BAC contigs used as substrates for sequencing are separated by physical gaps of unknown size, and given that the multiple sequence contigs representing each single BAC insert are often unordered and unori-

ented with respect to each other, the sequence product is referred to as the "working draft" to distinguish it from completed genomic sequence in which the size of all gaps and the order and orientation of all sequence is known.

As members of the effort to produce a working draft sequence of the highest quality possible, we have used the technique of radiation hybrid (RH) mapping, in conjunction with more than 40,000 unique human STSs, to identify those segments of the human genome that are absent from the working draft sequence, to provide independent estimates of the size and location of missing sequence in relation to the existing sequence, and to provide order information for more than 15,000 of the clones that were used to create the human working draft.

The construction of human genome topography has been greatly facilitated in recent years by the development of STSs as genomic landmarks (5). Each STS is defined by a short segment of 100 to 500 base pairs (bp) of human DNA sequence and is assayed by means of polymerase chain reaction (PCR) (6). Common sets of such sequence-based markers can be easily screened and therefore can be used to integrate maps constructed by different mapping methods. The development of STSs from short DNA sequences derived from a variety of clone sources provides a method to obtain markers from regions of the human genome that may be difficult to clone by any single vector system. Recent experience suggests that up to 10% of certain gene-rich regions of human chromosome 21 are composed of such "hard-to-clone" DNA (7). We have developed a large set of STSs using DNA sequence derived from a diverse set of sources in an effort to maximize the coverage of the human genome (8). Our strategy involved an initial electronic analysis of genomic DNA sequence to eliminate repetitive DNA sequences, followed by an automated selection of oligonucleotide primers to generate PCR products 90 to 350 bp in length under a single set of reaction conditions, as described (9). PCR products were assayed by ethidium bromide staining after agarose gel electrophoresis. An STS was judged successful when the primers produced a distinct PCR product of the expected size from total human DNA and failed to produce a product of this size from either hamster or mouse genomic DNA. We generated a total of 41,234 human STSs that met these criteria. Of these STSs, 14,953 were scored on rodent-human hybrid somatic cell mapping panels to determine their chromosomal location (10, 11). A total of 14,041 of these 14,953 STSs (94%) could be assigned to a unique human chromosome. These 14,041 chromosome-specific STSs, as well as the

<sup>1</sup>Stanford Human Genome Center, Stanford University School of Medicine, 975 California Avenue, Palo Alto, CA 94304, USA. <sup>2</sup>The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. <sup>3</sup>Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA. <sup>4</sup>Division of Human Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109, USA.

\*Present address: Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA.

†Present address: UCSF Comprehensive Cancer Center, 2340 Sutter Street, San Francisco, CA 94143, USA.

‡Present address: Incyte Genomics Inc., 3160 Porter Drive, Palo Alto, CA 94304, USA.

§To whom correspondence should be addressed. E-mail: cox@shgc.stanford.edu

## ANALYSIS OF GENOMIC INFORMATION

remaining 26,281 STSs not scored on the chromosomal mapping panel, were used to construct a high-resolution RH map of the human genome as described below.

The observation that random fragments of human chromosomes are retained in somatic cell hybrids between irradiated diploid human cells and nonirradiated hamster cells in the absence of selection for human chromosomal material provides the basis for RH mapping. DNA isolated from a panel of 80 to 100 such independent RH clones serves as a mapping reagent for ordering STSs and for determining the distances between them in the human genome. In this approach, the frequency of irradiation-induced breakage between two human markers is used as a measure of distance, and marker order is determined in a manner analogous to meiotic linkage mapping (12). As in the case of meiotic linkage mapping, the relative confidence in alternative marker orders in RH maps can be assessed by comparing LOD scores (logarithm of the odds ratio of linkage versus no linkage) using appropriate maximum likelihood statistical methods (13, 14). An important advantage of RH mapping is that hybrid panels for constructing maps at very different levels of resolution can be generated by experimentally manipulating the dose of irradiation. RH maps of the human genome published to date have used either the GB4 RH panel, which was constructed by using 3000 rad of x-rays, or the G3 RH panel, which was constructed by using 10,000 rad of x-rays (15–18). Both of these RH panels provide RH maps with good long-range continuity. However, STSs separated by less than 1000 kbp in the genome are not ordered routinely with high confidence when these mapping panels are used, due to the relatively small number of chromosome breaks. In contrast, a panel of 90 independent RH clones constructed at the Stanford Human Genome Center (SHGC) with 50,000 rad of x-rays (the TNG panel) allows STSs separated by less than 100 kb in the genome to be ordered with high confidence (19). The price of this increased resolution is that a large number of STSs need to be scored on the TNG panel to produce RH maps with good long-range contiguity. Here, we describe the use of the TNG panel in conjunction with the Stanford G3 panel to produce a high-resolution contiguous RH map of the human genome. Our map does not incorporate mapping information from other sources, except where specifically indicated below.

All 40,322 STSs described above were scored in duplicate on the each of the 90 clones of the TNG panel, as previously described (16). In addition, a subset of 10,227 of these STSs were also scored in duplicate

on the 83 clones of the G3 RH panel. We used these TNG and G3 RH data in conjunction with the chromosomal assignments of the 14,041 STSs described above, to assign an additional 25,486 STSs to unique human chromosomes (20). All 39,527 STSs that could be assigned to a specific chromosome were used for RH map construction. Chromosomal assignment of STSs before map building allowed us to construct the RH map for each chromosome independently, rather than using all data simultaneously. This approach greatly reduces the number of pairwise comparisons required in map construction and results in improved power to determine marker order.

Initially, G3 RH data for a given chromosome were used to construct a contiguous RH map using a variation of our previously described mapping method (21). In total, 8351 of the 10,227 STSs scored on the G3 panel (82%) were ordered with respect to each other on this G3 RH map of the human genome, with an average physical distance between ordered markers of 381 kbp. Only 275 of the 8328 LOD scores between adjacent STSs on the map (3%) were less than 3.0. The segments of the genome flanked by two low LOD scores were ordered and oriented by means of STSs on the G3 map that were also included in the Genethon meiotic linkage map

and which provided order information based on that map (22). As noted above, markers separated by less than a physical distance of 1000 kbp were not ordered routinely with high confidence. An additional 1466 unique STSs, each with an RH vector identical to an ordered marker, were assigned the same map position as the ordered STS, placing a total of 9817 STSs (96% of all 10,227 STSs scored on the G3 panel) on the G3 map. The complete G3 map can be accessed at Web table 1 and <http://shgc.stanford.edu> (23).

Next, we used the TNG data to build chromosome-specific maps as described above. In total, 33,627 STSs of the 39,527 STSs used in TNG map construction (85%) were ordered with respect to each other on the high-resolution TNG map of the human genome, with an average physical distance between ordered markers of 94 kbp (Table 1). Only 3676 of the 33,604 LOD scores between adjacent STSs on the TNG map (11%) were less than LOD 3.0. We defined an STS contig as a segment of the TNG map flanked by two LOD scores less than 3.0. G3 map position was used to help order and orient the STS contigs on the TNG map relative to one another, because 8961 of the STSs with map locations on the G3 map were also placed on the TNG map. An additional 3051 unique STSs, each with an RH vector

**Table 1.** Summary of TNG RH map. Mbp, megabase pairs.

Chromosome number	Physical length (Mbp)*	Mapped STSs†	Ordered STSs	Average density of ordered STSs (kbp)	STS contigs (LOD score >3.0)	Mapped STSs with no draft sequence accession hit	Mapped STSs in draft sequence (%)
1	263	4,343	3,709	71	309	477	89
2	255	2,898	2,693	95	410	296	90
3	214	3,071	2,758	78	259	310	90
4	203	4,133	3,701	55	295	511	88
5	194	2,095	1,967	99	227	250	88
6	183	1,969	1,867	98	224	195	90
7	171	1,646	1,539	111	202	120	93
8	155	1,733	1,619	96	180	234	86
9	145	1,364	1,290	112	123	89	93
10	144	1,634	1,525	94	151	138	92
11	144	1,592	1,496	96	167	138	91
12	143	1,494	1,379	104	183	147	90
13	98	987	927	106	112	83	92
14	93	1,040	972	96	120	35	97
15	89	1,018	936	95	97	120	88
16	98	1,086	991	99	107	161	85
17	92	892	831	111	104	93	90
18	85	857	822	103	71	87	90
19	67	457	445	151	62	23	95
20	72	544	523	138	77	12	98
21	34	971	837	41	23	15	99
22	34	265	258	132	41	15	94
X	164	473	432	380	118	45	90
Y	35	116	110	318	14	10	91
Total	3175	36,678	33,627	94	3676	3604	90

\*These estimates of physical length do not include the short arms of chromosomes 13, 14, 15, 21, and 22, or the repetitive portion of the long arm of the Y chromosome. †Mapped STSs include all 33,627 ordered STSs on the TNG map, as well as an additional 3051 unique STSs with RH vectors identical to ordered TNG STSs.

identical to an ordered TNG marker, were assigned the same map position as the ordered STS, giving a total of 36,678 STSs on the TNG map out of a total of 39,527 STSs used in map construction (93%) (24). The complete TNG map can be accessed at Web table 2 and <http://shgc.stanford.edu>.

In addition to providing LOD scores between adjacent mapped markers as a measure of confidence of marker order, RH mapping provides distance measures between adjacent markers in the map. This distance, based on the frequency of breakage between two markers in the radiation hybrid clones, is measured in units called centirays (cR). Previous work has demonstrated a direct correlation between cR units and physical distance in kb, which is fairly constant across the genome for any given RH panel (16). We compared cR distances between STSs on the TNG and G3 maps with actual physical distances between these STSs based on DNA sequence. We determined that 1 cR on the TNG map corresponds to an average of 2 kbp of physical distance, whereas 1 cR on the G3 map corresponds to a physical distance of 13 kbp. A comparison of 1778 pairs of STSs, where each member of a pair had an identical RH vector and thus a distance of 0 cR, revealed that in 86% of cases, the STSs defining the pair were separated by less than 20 kbp of physical distance. Distances less than 10 cR were found to overestimate the true physical distance, while distances greater than 70 cR were found to underestimate the true physical distance. Of the 33,604 cR distances between adjacent STSs on the TNG map, 30,420 (91%) are less than 100 cR, equal to a physical distance of less than 200 kbp. In 512 adjacent STS intervals, the STSs appeared to be completely unlinked, and we were unable to calculate a cR distance. In these instances, we picked STSs flanking the gaps that had also been ordered on the G3 map and used the G3 map distance information to estimate the physical distance of the gap. In this way, the physical distance between all STSs ordered on the TNG map was estimated.

The availability of a working draft sequence of the human genome has had a dramatic impact on the way genomic research is performed. Before the availability of a large amount of draft sequence, the only method for determining if a specific STS marker mapped to a specific DNA clone was to carry out a PCR reaction with DNA from the clone of interest. Although simple in principle, determining which of tens of thousands of STSs are present in each of tens of thousands of clones is a formidable and expensive task. In contrast, with the availability of the draft sequence,

one can map "in silico" (through computer analysis) by electronically comparing the DNA sequence of an STS of interest to all of the available draft sequences. In this way, it is possible to identify which clones contain the STS of interest and determine the precise location of the STS on each clone (25). We used an in silico mapping approach to determine which of the STSs present in the TNG RH map are also present in the working draft sequence. Because one of our primary goals in this work was to determine which, if any, STSs in the RH map were absent from the draft sequence, we needed a method with high sensitivity. Although "electronic PCR" can be performed with only oligonucleotide primer information and the known size of the STS sequence (25), we have found that this method identifies only 80% of STSs known to be present in a given sequence contig, even allowing for single-base differences between the primer sequence and the draft sequence. Comparing the entire sequence of the STS with the draft sequence is a more sensitive method of in silico mapping. However, this approach is complicated by repeated sequences that are present internal to the primer sequences of an STS, which pose no problem for specificity when performing a PCR reaction, but which can lead to significant false positive hits when sequence alignment algorithms are used. Although it is difficult to rigorously determine the ideal in silico mapping parameters, we used the BLAST algorithm (26), requiring an alignment of 100 bases or greater with an identity match of 90% or greater and an E value of less than  $1.0E^{-45}$ , to compare our STS sequences with the working draft sequence. This provides a good compromise between specificity and sensitivity for this electronic mapping approach. We found that only 15 of the 971 STSs (1%) ordered on the TNG map of chromosome 21 failed to hit the finished DNA sequence of chromosome 21 when these in silico conditions were used (Table 1) (27).

We compared all 36,678 STSs on the TNG RH map with the 5 September 2000 unassembled GenBank human sequence data release, using the in silico mapping parameters described above. Only 3604 of these STSs (9.8%) fail to hit a GenBank sequence, suggesting that the publicly available sequence covers more than 90% of the human genome (Table 1 and Web table 2).

As expected, because of the "finished" state of the chromosome sequence, the chromosome with the lowest percentage of nonhit STSs is chromosome 21 with 1% missing. Next lowest are chromosomes 20 and 14, with only 2% and 3% of STSs missing, respectively. Chromosomes 16

and 8 have the highest percentage of STSs with no sequence hits, with 15% and 14% missing, respectively. Overall, the sequence coverage of the genome appears to be fairly uniform (Table 1 and Web table 2). Cases in which STSs exist that have no counterparts in the sequence database serve as valuable reagents for completing the sequence of the human genome. STSs that fail to hit a match in the presently available sequence can be used to screen a variety of human DNA libraries cloned in different vectors, providing an efficient method for closing existing clone gaps in the human draft sequence.

By assigning each STS that hit a sequence in the working draft to a specific sequence (as shown by a unique accession number in GenBank), the draft sequence can be linked to the RH map. This provides an independent measure of order for the clones that were used to generate the draft sequence of the genome and an estimate of the physical length of gaps between non-overlapping clones. On numerous occasions, a single STS was found to hit multiple sequence accessions with a variety of different E values, and in each case, meeting the minimal criteria for a hit described above. In such cases, we considered only those hits with the most significant E value for further evaluation. Even with this criterion imposed, the 33,078 individual STSs on the TNG map with at least one hit in the draft sequence produced a total of 54,225 hits, giving an average of 1.7 hits per STS. For every STS with more than one accession hit, we chose the accession with the greatest number of hits by other STSs mapping to the same chromosome and linked the STS to this unique accession. In this way, we created a minimal set of 15,718 accessions that contained a single hit for each of the 33,078 TNG map STSs. Of this minimal accession set, 7957 (51%) contained more than one STS hit, providing an opportunity to determine how often multiple STSs hitting a single clone had discrepant chromosome assignments. We found that 475 of these 7957 accessions (6%), contained STSs that mapped to two or more chromosomes. At this time, it is unclear if these discrepancies are due to errors in chromosomal assignment, if they represent false positive electronic mapping hits, or if they represent chimeric clones that actually contain human DNA from more than one human chromosome. Further studies of the 475 accessions in question are required to distinguish among these possibilities.

The primary substrate used to generate the draft sequence of the human genome was a minimally redundant set of BAC clones assembled into physically separate clone contigs by restriction fingerprint

## ANALYSIS OF GENOMIC INFORMATION

**Table 2.** Comparison of the order of 8975 STSs on the TNG RH map and the BAC fingerprint map with their order in the draft sequence.

Pairs of adjacent STSs	Positions separating STSs in sequence	Cumulative fraction of STS pairs
<i>TNG RH map</i>		
6663	1	0.74
790	2	0.83
291	3	0.86
191	4	0.88
131	5	0.9
286	6–10	0.93
623	>10	1
Total: 8975		
<i>BAC fingerprint map</i>		
6641	1	0.74
1263	2	0.88
323	3	0.92
141	4	0.93
54	5	0.94
180	6–10	0.96
373	>10	1
Total: 8975		

mapping (3). Although fingerprint mapping has the advantage of mapping the clones that are used for DNA sequencing, it has the disadvantage that the restriction fragments used to build restriction maps are not sequence-based like the STS markers used in RH mapping. We compared the order of sequence accessions in the TNG map with the 5 September 2000 data release of ordered accessions, which is posted on the Washington University Genome Sequencing Center website ([http://genome.wustl.edu:8021/pub/gsc1/fpc\\_files/freeze\\_2000\\_09\\_05/MAP](http://genome.wustl.edu:8021/pub/gsc1/fpc_files/freeze_2000_09_05/MAP)). This data set consists of 31,102 ordered sequence accessions grouped by fingerprint mapping into 1001 physically distinct clone contigs. We determined that 14,363 of the 15,718 sequence accessions on the TNG map (92%) are included in the fingerprint data set and provide a basis for comparison of the two maps. A total of 773 of these 14,363 accessions (5%) were assigned to a differ-

ent chromosome in the TNG map as compared to the fingerprint map. Of the 1001 fingerprint contigs, 882 (88%) are represented by one or more accessions on the TNG map (28).

We compared the order of 8976 sequence accessions with the same chromosome assignments in the fingerprint and TNG maps in the following manner: The accessions for each chromosome were ordered on the basis of the fingerprint data and numbered sequentially, beginning with 1, providing each accession on the chromosome with a unique fingerprint order number. All sequence accessions on the chromosome then received a TNG order number, reflecting its position in the TNG chromosome map. The list was then sorted by fingerprint order number, and the pattern of TNG order numbers was analyzed. Moving from top to bottom on the sorted list, we calculated the absolute difference in TNG order number for each pair of adjacent STSs, and then determined the number of all intervals differing by one position, two positions, three positions, and so forth. If the fingerprint order and the TNG order were completely consistent, all intervals defined by adjacent STSs would differ by one position. This method of comparing the orders of shared markers in two different maps has high practical value, because it provides information regarding the number of order discrepancies as well as the magnitude of those discrepancies. The results of the comparison between the TNG map and fingerprint map revealed that of 8975 total intervals, 7722 (86%) differed by five positions or fewer, whereas 856 intervals (10%) differed by more than 10 positions. Although this analysis provides a direct comparison of sequence accession order in the TNG map versus the fingerprint map, it does not provide information regarding which map, if any, most often reflects the true order of sequence in the human genome. To address this question, we carried out a similar analysis comparing the TNG

map and the fingerprint map to the assembled draft sequence of the human genome [<http://genome.ucsc.edu>, final 7 October freeze assembly from 9 January 2001 (1)]. Sequence information derived from all STSs on the TNG map was used by David Haussler and colleagues to perform electronic PCR with the assembled draft sequence and determine the order of the STSs in this assembly. Given that both the fingerprint map and the TNG map were used in conjunction with other mapping information to assemble the draft sequence, comparison of the TNG and fingerprint maps with the draft sequence does not provide truly independent validation. Nevertheless, such a comparison provides a useful means of examining each map in relation to the final draft sequence product. Of 8975 total intervals, 7744 (86%) differed by three or fewer positions when the TNG order was compared to the draft sequence order. In contrast, 8227 of the same 8975 total intervals (92%) differed by three or fewer positions when the fingerprint map order was compared to the draft sequence order (Table 2). Overall, these results indicate that although both maps have inconsistencies with the assembled draft sequence, the discrepancies between the draft sequence and map order are often at different positions in the TNG and fingerprint maps. Thus, using the order information from both of these maps as well as others to produce the final assembly of the draft sequence has resulted in a much better quality product than would have been produced by using either map alone (1).

Another important physical map, which has provided order information for the draft sequence, is the fluorescence in situ hybridization (FISH) map (29). The TNG map contains 886 unique STSs derived from clones, which have been mapped by FISH. Of these 886 markers, 28 (3%) have a different chromosomal assignment on the FISH map compared to the TNG map. An analysis of the order of the 858 markers

**Table 3.** Comparison of STS order in the draft sequence and the Celera sequence with the order in the TNG map.

Comparison of STS order	Pairs of adjacent STSs	Fraction of STS pairs
<i>Draft order</i>		
Draft order concordant with Celera order	12,791	0.64
Draft order discordant with Celera order, TNG order concordant with draft order	1,750	0.09
Draft order discordant with Celera order, TNG order concordant with Celera order	2,408	0.12
Draft order discordant with Celera order, TNG order discordant with draft and Celera orders	3,057	0.15
Total adjacent STS pairs compared	20,006	
<i>Celera order</i>		
Celera order concordant with draft order	12,791	0.64
Celera order discordant with draft order, TNG order concordant with Celera order	2,600	0.13
Celera order discordant with draft order, TNG order concordant with draft order	1,711	0.09
Celera order discordant with draft order, TNG order discordant with Celera and draft orders	2,904	0.14
Total adjacent STS pairs compared	20,006	

that have a consistent chromosome assignment on both maps reveals that 836 (97%) have a consistent order on both the TNG and FISH maps (29). The basis for the discrepancies remains to be determined. The consistent STSs provide valuable reference points that anchor the TNG map to the cytogenetic map (Web table 2).

The draft sequence of the human genome is a tremendous achievement, which will enable scientific discovery in ways not previously possible. However, far greater scientific value will be gained from the completed sequence of the human genome. In addition to DNA sequencing efforts of the International Consortium, (1), Celera Genomics has used a whole-genome shotgun approach to produce an assembled sequence of the human genome (30). In an effort to determine how the Celera sequence adds information to the working draft sequence and vice versa, Mark Adams and Peter Li at Celera have used sequence alignment methods to assign 34,725 SHGC STS sequences to unique positions in the Celera sequence and provided us with the ordered list of STSs. We compared this list with the 33,627 SHGC STSs ordered on the TNG map and with the 27,471 SHGC STSs assigned to unique locations in the 7 October freeze final assembled draft sequence by David Haussler and colleagues (<http://genome.ucsc.edu>). A total of 20,874 STSs are shared in common in the three lists. Because each group used different methods to identify STSs in the sequence, it is not appropriate to use the STS list to draw any conclusions regarding coverage of the genome. However, the 20,874 STSs, spaced at an average distance of ~150 kbp in the genome, provide valid information for assessing the consistency of STS chromosome assignment and order in the draft sequence, the Celera sequence, and the TNG map. Of the 20,874 STSs, only 766 (3.7%) revealed a chromosome discrepancy. In 220 of these cases (29%), the TNG chromosome assignment was different from that shared in both the draft and Celera sequences. In 331 cases (43%), the Celera chromosome assignment differed from that shared by the draft sequence and the TNG map. In 69 cases (9%), the chromosome assignment in the draft sequence differed from that shared by the Celera sequence and the TNG map. Finally, in 146 cases (19%), the draft sequence, the Celera sequence, and the TNG map each had a different chromosome assignment for the STS. Given the caveat that the data sets are not completely independent, those chromosome assignments supported by two of the three data sets are most likely the correct assignments. These results illustrate the utility of comparing multiple data sets to

identify potential errors. A comparison of the order of the 20,030 STSs that had a consistent chromosome assignment in the three data sets is presented in Web table 3 and summarized in Table 3. For each STS, we identified the STS that was immediately distal in the draft sequence order and determined if the two STS were adjacent in the same orientation in the Celera sequence. Of 20,006 comparisons, 12,791 (64%) revealed an identical order and 8183 (36%) revealed a discrepant order. As shown in Table 3, the result is the same if we begin with the STSs in the Celera sequence order rather than the draft sequence order. Of the discrepancies, 51% represent pairs of markers that are inverted in the draft sequence versus the Celera sequence, and 86% of all discrepancies are separated by five positions or fewer in the draft sequence versus the Celera sequence. Comparison of the order of discrepant STSs with their order in the TNG map revealed that the TNG order was consistent with one of the two sequence orders ~60% of the time. In these cases, the TNG order was consistent with the Celera order slightly more often than the draft sequence order. However, the most striking observation is that neither sequence data set predominated. For any given discrepancy, if the TNG map was found to be consistent with one of the two sequence orders, it was nearly as likely to be the draft sequence order as the Celera sequence order. These results again illustrate the utility of using multiple data sets in an effort to resolve discrepancies. Resolution of discrepancies will play an increasingly important role in ongoing efforts to complete the sequence of the human genome. Although the path to produce finished sequence of the human inserts in each BAC clone is well defined, the approach for identifying and isolating those sequences missing from the available sequence is less clear. The 3604 STSs ordered on the TNG map, which are not present in the draft sequence, serve as very useful reagents for isolating the missing portions of the human genome. We note that of these 3604 STSs, 2472 (69%) are present on the list of SHGC STSs identified in the Celera sequence. Given that difficult-to-clone regions are often in gene-rich areas of the human genome (7), isolation of the segments of the genome which are presently absent from the draft sequence becomes a biologically relevant task, and not simply an exercise in completing this large-scale project. The TNG radiation hybrid map of the human genome draft sequence is a valuable tool that will help the international scientific community reach the goal of a finished sequence of the human genome as rapidly as possible.

#### References and Notes

1. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
2. K. Osoegawa *et al.*, *Genomics* **52**, 1 (1998).

3. J. D. McPherson *et al.*, *Nature* **409**, 934 (2001).
4. M. A. Marra *et al.*, *Genome Res.* **7**, 1072 (1997).
5. M. Olson, L. Hood, C. Cantor, D. Botstein, *Science* **245**, 1434 (1989).
6. E. D. Green, M. V. Olson, *Science* **250**, 94 (1990).
7. M. Hattori *et al.*, *Nature* **405**, 311 (2000).
8. A total of 3237 STSs are shared in common with the Genethon meiotic linkage map (22), 12,544 STSs are derived from BAC end DNA sequence, 14,539 STSs are derived from cDNA and/or expressed sequenced tag sequence, and 8667 STSs are derived from short fragments of genomic DNA that contain single-nucleotide polymorphisms (SNPs) identified by members of the SNP Consortium.
9. E. M. Beasley, R. M. Myers, D. R. Cox, L. C. Lazzeroni, in *PCR Applications* (Academic Press, San Diego, CA, 1999), pp. 55–71.
10. H. L. Drwings, L. H. Toji, C. H. Kim, A. E. Greene, R. A. Mulivor, *Genomics* **16**, 311 (1993).
11. B. L. DuBois, S. L. Naylor, *Genomics* **16**, 315 (1993).
12. D. R. Cox, M. Burmeister, E. R. Price, S. Kim, R. M. Myers, *Science* **250**, 245 (1990).
13. M. Boehnke, K. Lange, D. R. Cox, *Am. J. Hum. Genet.* **49**, 1174 (1991).
14. K. L. Lunetta, M. Boehnke, *Genomics* **21**, 92 (1994).
15. G. Gyapay *et al.*, *Hum. Mol. Genet.* **5**, 339 (1996).
16. E. A. Stewart *et al.*, *Genome Res.* **7**, 422 (1997).
17. T. J. Hudson *et al.*, *Science* **270**, 1945 (1995).
18. P. Deloukas *et al.*, *Science* **282**, 744 (1998).
19. K. L. Lunetta, M. Boehnke, K. Lange, D. R. Cox, *Am. J. Hum. Genet.* **59**, 717 (1996).
20. RH vectors were used to construct two-point LOD scores for each of the 26,281 STSs unassigned to a unique chromosome and all 14,041 STSs described above with a unique chromosome assignment. Unassigned STSs were given the same chromosomal assignment as a previously assigned STS if the two-point LOD score between the two markers was greater than or equal to 6.0. Some 301 STSs had no two-point LOD score greater than or equal to 6.0, and 494 STSs had high two-point LOD scores to multiple STSs assigned to more than one chromosome. These 795 STSs (3% of the 26,281 previously unassigned STSs) which could not be assigned to a specific chromosome were not used in RH map construction.
21. Both the G3 and the TNG RH maps were constructed by using a variation of a previously described mapping algorithm (16).
22. C. Dib *et al.*, *Nature* **380**, 152 (1996).
23. Supplementary data are available at [www.science.org/cgi/content/full/291/5507/1298/DC1](http://www.science.org/cgi/content/full/291/5507/1298/DC1)
24. A total of 716 of the STSs that failed to map had no LOD score greater than 3.0 with any other STS, whereas another 2133 STSs mapped to regions of the genome flanked by low LOD scores and that contained no G3 order information.
25. G. D. Schulz, *Trends Biotechnol.* **16**, 456 (1998).
26. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
27. Although the fact that the draft sequence of individual clones is often in unordered and unoriented fragments might reduce the hit rate somewhat below that of finished sequence, the effect should be minimal given that we require an alignment of only 100 base pairs or greater to call a positive hit.
28. Some 1313 sequence accessions hit by unmapped SHGC STSs are represented in 46 additional fingerprint contigs, giving a total of 928 out of 1001 fingerprint contigs (93%) hit by SHGC STSs.
29. V. G. Cheung *et al.*, *Nature* **409**, 953 (2001).
30. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
31. We thank D. Haussler, S. Rogic, T. Furey, and J. Kent for providing the order of SHGC STSs in the draft sequence, P. Li and M. Adams for providing the order of SHGC STSs in the Celera sequence, K. Frazer and J. Sheehan for assistance with BLAST searches, K. Frazer and N. Patil for helpful discussions, and all members of the Stanford Human Genome Center, past and present, for outstanding technical support. This work was supported by grants from NIH and from the SNP Consortium.

7 November 2000; accepted 19 January 2001