Computational comparison of human genomic sequence assemblies for a region of chromosome 4

Computational Comparison of Human Genomic Sequence Assemblies for a Region of Chromosome 4

Colin A.M. Semple,[1,2] Stewart W. Morris, David J. Porteous, and Kathryn L. Evans

Medical Genetics Section, Department of Medical Sciences, The University of Edinburgh, Molecular Medicine Centre, Western General Hospital, Edinburgh EH4 2XU, United Kingdom

Much of the available human genomic sequence data exist in a fragmentary draft state following the completion of the initial high-volume sequencing performed by the International Human Genome Sequencing Consortium (IHGSC) and Celera Genomics (CG). We compared six draft genome assemblies over a region of chromosome 4p (D4S394–D4S403), two consecutive releases by the IHGSC at University of California, Santa Cruz (UCSC), two consecutive releases from the National Centre for Biotechnology Information (NCBI), the public release from CG, and a hybrid assembly we have produced using IHGSC and CG sequence data. This region presents particular problems for genomic sequence assembly algorithms as it contains a large tandem repeat and is sparsely covered by draft sequences. The six assemblies differed both in terms of their relative coverage of sequence data from the region and in their estimated rates of misassembly. The CG assembly method attained the lowest level of misassembly, whereas NCBI and UCSC assemblies had the highest levels of coverage. All assemblies examined included <60% of the publicly available sequence from the region. At least 6% of the sequence data within the CG assembly for the D4S394–D4S403 region was not present in publicly available sequence data. We also show that even in a problematic region, existing software tools can be used with high-quality mapping data to produce genomic sequence contigs with a low rate of rearrangements.

[All sequence accessions for the genomic sequence assemblies analyzed and the data sets used to assess coverage and rates of misassembly are available from http://www.ed.ac.uk/~csemple.]

The human genome sequence is expected to remain in draft form until the year 2003 (Roach et al. 1999). Nevertheless, preliminary draft genome assemblies of the unfinished data offer a wealth of information. There have been three major efforts to produce such assemblies, the freely available Human Genome Project Working Draft (http://genome.ucsc.edu/) at the University of California, Santa Cruz (UCSC) described by the International Human Genome Sequencing Consortium (IHGSC) (2001); the freely available National Centre for Biotechnology Information (NCBI) assembly (http://www.ncbi.nlm.nih.gov/); and the Celera Genomics (CG) assembly (http://public.celera.com/) described in Venter et al. (2001). The relative merits of such hybrid assemblies are of particular interest as both the IHGSC and CG data sets contain unique sequences (Aach et al. 2001). Unfortunately, CG sequence data are only available publicly in the form of an assembly that also includes IHGSC data, which complicates the construction of hybrid assemblies. Given the restrictions placed on whole genome analysis of the CG data, a large-scale comparison of the available assemblies is difficult. Aach et al. (2001) did perform some analyses on this scale, but they compared the CG assembly with an NCBI assembly and did not examine an assembly produced at UCSC. They also omitted any general assessment of the degree of misassembly (sequences assembled in the wrong order and/or orientation).

Here, we compare the quality of six draft genome assemblies over the 4p15.3–4p16.1 region between markers D4S394 and D4S403. This region was found previously to be linked (maximum multipoint LOD score =4.8) to affective disorder by Blackwood et al. (1996), and a 6. 9-Mb contig encompassing it was recently constructed (Evans et al. 2001). The D4S394–D4S403 region itself was estimated at 5.8 Mb. This region provides an instructive comparison of these assemblies for the following three reasons: (1) it contains a well-established, dense coverage of marker sequences in known order (Evans et al. 2001) that allow an assessment of the degree of misassembly; (2) it contains a large tandem repeat (Kogi et al. 1997); (3) it contains subregions that are underrepresented in available clone libraries and are, therefore, sparsely covered by draft sequence data (Evans et al. 2001). Both (1) and (3) represent worst-case scenarios for assembly algorithms. The assemblies compared (see Table 1) were two consecutive releases from UCSC (referred to as assemblies UCSC1 and UCSC2), two consecutive releases from NCBI (referred to as NCBI1 and NCBI2), the public release from CG (referred to as `CELERA`), and a hybrid assembly (referred to as `HYBRID`), which we have produced using IHGSC and CG sequence data along with the physical mapping data of Evans et al. (2001).

RESULTS

Draft sequence assemblies were compared with a framework set of 107 sequences accurately ordered across the region to

[1]Present address: Bioinformatics, MRC Human Genetics Unit, Edinburgh EH4 2XU, UK.
[2]Corresponding author.
E-MAIL Colin.Semple@ed.ac.uk; FAX 44-131-343-2620.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.207902. Article published online before print in February 2002.

12:424–429 ©2002 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

Table 1. The Draft Genome Assemblies Examined

Assembly	Version	Origin
NCBI1	9/2/00	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Genome
NCBI2	16/4/01	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Genome
UCSC1	9/1/01	http://genome.ucsc.edu/goldenPath/hgTracks.html
UCSC2	5/4/01	http://genome.ucsc.edu/goldenPath/hgTracks.html
CELERA	public	http://public.celera.com/
HYBRID	NA	see text

assess the degree of misassembly and to a nonredundant set of genomic sequences from the region (NR) to assess coverage. As expected, across assembly methods coverage increased with the length of assemblies (Spearman's rank correlation coefficient $r_s = 0.943$, $p < 0.05$), whereas the number of gaps within the contigs decreases ($r_s = 0.943$, $p < 0.05$). This suggests that our measurements of coverage are reasonably accurate. Table 2 shows the results achieved by the different assembly methods. The CELERA assembly contained fewer misassemblies than any other examined, but also had the lowest coverage and the highest number of contigs. The NCBI assemblies contained the highest number of misassemblies, including the inclusion in NCBI2 of >15 Mb of sequence that does not appear to map to the D4S394–D4S403 region (see sequence retrieval section of Methods). With such a large amount of sequence from outside of the D4S394–D4S403 region, an accurate assessment of the number of NCBI2 misassemblies could not be made, consequently, the rate of misassembly for NCBI2 in Table 2 is likely to be an underestimate. The UCSC assemblies combined relatively high coverage with rates of misassembly that, although higher than in CELERA, are lower than in the NCBI assemblies. Broadly, it would appear that across methods, the number of misassemblies tends to rise as assembly coverage increases, but in the absence of reliable data for HYBRID and NCBI2, this correlation is not significant ($r_s = 0.80$, $p > 0.1$). It is also notable that all assemblies included <60% of the available EMBL HTG sequence from the region, and all contained some degree of misassembly. A subset of 34 framework markers was found to be present within completed BAC sequences (see Methods). In every assembly, other than CELERA, the misassemblies observed included markers from this subset, in orders different from those seen in completed BAC sequences.

Considerable variability was seen between subsequent releases of assemblies produced by the same method (Table 2). Although the number of contigs spanning the region remained similar between NCBI1 and NCBI2, the coverage of the region increased, but only at the cost of catastrophic misassemblies (see sequence retrieval section of Methods). Between UCSC1 and UCSC2, a modest decrease in the frequency of misassembly appears to have been achieved, but at the expense of coverage of the region, which has decreased in common with the overall length of the assembly. Thus, with an increase in sequence data within the region of 575,894 bp, the NCBI and UCSC assembly methods have produced, respectively, an increase (5%) and a decrease (11%) in coverage.

Table 2. Comparison of Draft Sequence Assemblies Across D4S394-D4S403 Region

Assembly	NCBI1	NCBI2[a]	UCSC1	UCSC2	CELERA	HYBRID
Version	9/2/00	16/4/01	9/1/01	5/4/01	public	NA
Length[b] (bp)	4,220,059	7,982,790	6,597,859	5,725,683	3,359,224	3,510,128
Contigs	9	10	3	4	81	37
Gaps in ctgs	373	466	420	325	197	0
Gaps (bp)	37,300	46,600	590,800	631,400	472,294	0
Framework[c] comparison						
Duplications[d]	3	3	1	3	0	NA
Deletions[e]	5	8	10	10	6	23
Rearrangements[f]	12	13	11	5	1	NA
Misassemblies/Mb[g]	4.74	3.01	3.33	3.14	2.08	NA
NR[h] coverage						
NR fragments	3961	4427	4383	3516	1768	2322
NR (bp)	1,446,441	1,597,804	1,588,701	1,292,726	619,743	834,492
Coverage[i]	0.54	0.59	0.59	0.48	0.23	0.31
Annotation						
Repetitive sequence (bp)	2,395,019	3,438,941	2,689,302	1,989,867	2,418,685	1,386,794
PRS447 (bp)	0	0	25,707	25,707	15,679	0

[a]Modified from original version—see Methods.
[b]Length: total length of all contigs including gaps within contigs.
[c]Framework: a set of 107 sequences accurately ordered across the region.
[d]Duplications: observations of the additional appearance of a marker relative to the framework set.
[e]Deletions: observations of the absence of marker or a series of contiguous markers relative to the framework set.
[f]Rearrangements: observations of marker orders differing from the framework set that are not the result of duplications or deletions.
[g]Miassemblies/Mb: total number of duplications, deletions and rearrangements per Mb.
[h]NR: a nonredundant set of genomic sequence data from the region.
[i]Coverage: the proportion of sequence from the nonredundant genomic sequence data set (NR) present in assembly.

The amount of sequence unique to the CELERA assembly was assessed in the following way. All CELERA contigs were masked for repeats and then divided into fragments of 100 bp or less, giving 60,047 fragments in total. These fragments were then searched against EMBL HTG. These searches showed that 3564 fragments, originating from 48 of the 81 CELERA contigs (see Table 2) and containing 322,854 bp of unmasked sequence, failed to generate matches (≥95% identical over ≥50 bp) to publicly available sequence. It would therefore appear that CELERA contains at least 322,854 bp (equivalent to ~10% of CELERA and ~6% of the D4S394–D4S403 region, assuming a size of 5.8 Mb) not present in public databases. However, in the absence of any mapping data independent of the CELERA assembly, we cannot exclude the possibility that some or all of this 322,854 bp originates from outside of the D4S394–D4S403 region.

Each assembly was assessed with respect to the estimated amounts of repetitive sequence. Total repeat content was comparatively high in CELERA, which contained a higher proportion of repetitive sequence than any other assembly, despite being the shortest assembly with lowest coverage of the region. According to Kogi et al. (1997), the region contains at least 6 copies of the pRS447 repeat, totalling 28,512 bp. However, the BAC AC022770 from the region contains 75,357 bp of DNA matching pRS447 (BLASTN matches with $E \leq 1 \times e^{-10}$ and ≥98% identity), which corresponds to almost 16 copies of the repeat. Table 2 shows that CELERA and both UCSC assemblies incorporate pRS447 sequence, but none of these contained enough pRS447 sequence to represent the complete tandem repeat region, consisting of at least 16 copies. The NR data set produced for the region is ~26% repetitive sequence, whereas 40% of all IHGSC sequence from the region [according to the December 12, 2000 release of the high-resolution physical map of the genome produced by the International Human Genome Mapping Consortium (IHGMC) (2001)] is repetitive. Because the former data should under-represent and the latter data over-represent some sequences in the region, the actual sequence of the region should be composed of between 26% and 40% repetitive sequence. Only the repeat content of UCSC2 falls within these limits.

DISCUSSION

A number of generalizations can be made across all draft human genome assemblies for this region of chromosome 4p. As one might expect, coverage increased with assembly length, whereas the number of gaps within contigs decreased. All assemblies examined contained <60% of the available sequence from the region, and all contained some degree of misassembly, as measured by deviations from both our framework marker order and the order observed in completed BAC sequences. One might assume that as genomic sequence data accumulates and coverage of a region rises, there might be a decrease in the number of misassemblies, as new sequence data closes gaps and reduces ambiguity. However, in this region, there is no evidence that as coverage increases, the rate of misassembly drops, which suggests that current assembly methods have not optimally incorporated new sequence data in the region under study. All assemblies under-represent the region containing the pRS447 tandem repeat unit, which is the predicted consequence of encountering large duplicated segments during assembly (Eichler 2001). A recent study found that duplicated segments that are 90%–98% identical and in excess of 1 Kb constitute 3.6% of all human genomic sequence, and suggested that such segments may cause significant problems in accurate human genome assembly (Bailey et al. 2001). The data presented here support these suggestions.

There is wide variation in most of the measurable characteristics of the publicly available draft genome assemblies for the region. CELERA has a relatively low level of misassembly, particularly given the relatively high proportion of repetitive sequence within it, but has the lowest coverage. Because the estimates of coverage were made with sequence available to all assemblies, including CELERA, this low coverage is not a consequence of sequence availability. Rather, the CG assembly method must have excluded more publicly available sequence than the other methods. The comparatively low rate of misassembly in CELERA may be due to the superiority of the assembly algorithm used. Alternatively, it may reflect the use of additional, high-resolution mapping data to order and orientate sequence contigs. This additional data is a product of the CG-sequencing strategy and takes the form of paired sequencing reads in known relative orientation and separated by a known distance. NCBI assemblies for this region appear to have a high incidence of misassembly, although they included a smaller proportion of repetitive sequence in general, and no pRS447 sequence. The UCSC assembly method combined more than double the coverage of CELERA with only around one misassembly per Mb more than CELERA, despite including more pRS447 sequence than the other assemblies.

Olivier et al. (2001) compared orders of 20,874 TNG radiation hybrid map STSs (at an average density of 1 marker per ~150 Kb) in the UCSC1 and CELERA assemblies. They found widespread differences between these assemblies, such that 36% of TNG STS pairs were present in orders that differed between UCSC1 and CELERA. The TNG order was consistent with the CELERA assembly order slightly more often than with the UCSC1 order. As Olivier et al. (2001) used different methods to match markers to UCSC1 and CELERA, it was not possible to come to any conclusions regarding the relative coverage of the assemblies. The accuracy of the UCSC assembly method has been tested using artificial data sets (derived by fragmenting large regions of finished sequence), in which the actual order and orientation of sequence fragments is known and can be compared with that produced by the UCSC algorithm (see http://genome.ucsc.edu/). In such tests, the algorithm was found on average to assign ~10% of fragments the wrong orientation and to place ~15% of fragments in the wrong order. Similarly, omitting deleted or duplicated markers (Table 2), in our data we observed 19% (20/107) and 11% (12/107) of markers in orders that differed from the framework set in the UCSC1 and UCSC2 assemblies, respectively. In agreement with these observations, Katsanis et al. (2001) examined various UCSC consecutive draft genome assembly releases and reported that 10%–15% of EST sequences identified within them appeared to be on wrongly assembled genomic sequences.

Aach et al. (2001) compared the complete CELERA draft genome assembly with an NCBI genome assembly produced prior to the NCBI assemblies examined here, and reported that ~0.14% of sequence in either assembly was unique. Aach et al. (2001) used an indirect method, generating every possible stretch of 15 nucleotides (15-mers), and then determining the number of times it occurred in each assembly. The 15-mers found only once in either assembly were referred to as candidate unique 15-mers (cu-15s), and for each assembly

11% of cu-15s were not found in the other. Because of the substantial rate of error in identifying cu-15s (estimated at >9%), their final estimate of the actual amount of unique sequence present in each assembly was much lower (0.14%). The observation that ~7000 (~26%) of CG-annotated genes have no BLASTN similarity matches to the Ensembl set, which is based only upon public sequence data, suggests that there may be a higher number of unique sequences (Gaasterland and Oprea 2001). The Aach et al. (2001) estimate may be inaccurate as a consequence of restricting their analyses to cu-15s, because sequences may occur more than once in one assembly and yet be absent from the other. For instance, in the D4S394–D4S403 region, all 15-mer sequences within the pRS447 tandem repeat unit occurred multiple times within the CELERA assembly, but are also absent from both the NCBI1 and NCBI2 assemblies. The data presented here suggest that the proportion of unique sequences present in the CELERA assembly varies widely across the genome, such that 10% of the CELERA assembly for the D4S394–D4S403 region was not present in the publicly available sequence data. This is close to the estimate made by Venter et al. (2001) of 240 Mb (~8%) of unique sequence in the CELERA assembly. Thus, in spite of the relatively low coverage of the CELERA assembly, it contains unique sequences that can be included in hybrid assemblies incorporating both public and CELERA sequence data.

The construction of the HYBRID assembly shows that even in a problematic region, existing software tools can be used with high-quality mapping data to produce genomic sequence contigs with a low rate of rearrangements. However, the rather stringent thresholds associated with this low rate also increased the number of short contigs produced. Subsequently, this reduced assembly coverage when many of these contigs could not be ordered relative to one another, on the basis of the framework marker set. This problem is reflected in the relatively high number of contigs and low coverage of the HYBRID assembly. It is notable that even with the relatively conservative approach to contig building of phrap (using default settings) and with the small number of contigs found to contain framework markers, a misassembled contig was still observed. In a comparison with publicly available assembly algorithms, phrap was found to be most successful in generating genomic sequence contigs (Chen and Skiena 2000). phrap was found to combine the production of a relatively small number of large contigs with comparatively low rates of error. Thus, although valuable additional information could be gained from the production of hybrid assemblies, caution should be observed in using the resulting contigs, which may contain misassemblies, even when using the best available tools. In the absence of independent mapping data, these errors may be difficult to detect, thus, our analysis emphasizes the continuing value of accurate, high-resolution physical maps.

Major efforts are underway to define the coding regions (http://www.ensembl.org/) and variation (Altshuler et al. 2000) present in the draft sequence, and it is hoped that these data will accelerate the positional cloning of disease genes. To identify and fully investigate such genes, the order and relative orientation of features (genes, regulatory sequences, markers, and repeats) must be known within the region of interest. The differences between assemblies in this region were found to result in differences in annotation; assemblies varied in terms of the proportion of repetitive sequence and the number of pRS447 tandem repeat units. Annotation differences were also identified by Hogenesch et al. (2001) between the CELERA assembly and a UCSC assembly, which predates those examined here. They found large differences between the genes found in these assemblies, such that more than a third of the genes identified in one assembly were not found in the other, and conclude that differences between the underlying draft sequence assemblies may cause these differences in annotation. Problems with assembly quality and coverage in regions such as D4S394–D4S403 suggest that, at least for the purposes of gene finding, approaches that are not dependent upon building a definitive assembly may be useful. For example, Semple et al. (2001) used all available sequence and mapping information for a region of chromosome 11 to identify novel genes. Because of deficiencies in coverage and misassemblies in this region of chromosome 11, it was not possible to obtain the full sequences for these genes from publicly available assemblies. For the next generation of common, complex disease-mapping studies, there will be a heavy reliance upon linkage disequilibrium (LD) mapping. The extent of LD now appears to vary dramatically across the genome and optimizing LD mapping strategies will be dependent upon accurate, high-resolution maps of regions of interest. Such a detailed description of a region cannot be derived from an inaccurate assembly of the available genomic sequence.

METHODS

Sequence Retrieval

Searches of the genomic sequence annotation in the Genomes section of NCBI Entrez (Wheeler et al. 2001) identified nine NCBI contigs (Build 21, February 9, 2001 release) spanning the D4S394–D4S403 region (NT_006362.2, NT_023040.2, NT_006307.2, NT_022870.2, NT_022855.2, NT_006342.2, NT_022808.2, NT_016407.2, NT_006335.2). These nine contigs are collectively referred to as assembly NCBI1 in this study. Analogous searches using the UCSC Human Genome Browser identified the three UCSC contigs (January 9, 2001 update of October 7, 2000 Freeze data set) spanning the same region and named, following International Human Genome Mapping Consortium (2001) annotation, ctg15735, ctg15968, and ctg13685. These three contigs make up the assembly referred to here as assembly UCSC1. The presence of D4S394 and D4S403 in the assemblies retrieved was verified using BLASTN (Altschul et al. 1997). These searches, for NCBI and UCSC contigs in the region, were repeated at a later date to obtain the equivalent updated sequence assemblies. These were the Build 22, April 16, 2001 release at the NCBI and the April 5, 2001 release of December 12, 2000 freeze data set at the UCSC, referred to here as assemblies NCBI2 and UCSC2, respectively. According to the GenBank annotation of BAC sequence entries from the region, an additional 575,894 bp of sequence had become available for construction of these later assemblies. The contigs reported by Venter et al. (2001) were BLAST searched for the positions of D4S394 and D4S403, and then intervening sequences were identified using the SRS (Etzold and Argos 1993) search utilities at the CG Publication Site (http://public.celera.com/). The resulting 81 sequences encompassing the region are here referred to as the CELERA assembly. Table 1 summarizes the genomic sequence assemblies that were examined in this study.

BLASTN searches of the HTG (high-throughput genomic) section of the EMBL sequence database revealed that D4S394 and D4S403 lie in the unfinished, chromosome 4 BAC sequences AC004555 and AC007126, respectively. The region between these two BACs is spanned by 56 other BAC se-

quences (of which 11 are complete) according to the high-resolution physical map of the genome (December 12, 2000 release) produced by the International Human Genome Mapping Consortium (2001). The mapping data of Evans et al (2001) also supported the presence of these 58 BAC sequences in the region. In addition, all BACs were found to contain chromosome 4 STS sequences from the region using e-PCR (Schuler 1997).

It should be noted that our NCBI2 (4/16/01) data set is amended from the version we obtained from the NCBI. Upon examination, the NCBI version was found to consist of 42 NCBI contigs (totalling 23,322,979 bp) and, of these, only 10 were found to match the 58 BAC sequences known to be from the region. The remaining 32 contigs (totalling 15,340,189 bp) were found not to match any of these BAC sequences. BLAST matches to HTG EMBL BAC sequences produced by these 32 contigs combined with the mapping data (International Human Genome Mapping Consortium 2001) for these matching BACs, shows that much of this sequence comes from other regions of chromosome 4, outside of the D4S394–D4S403 region. Six of these thirty-two contigs are annotated (within their GenBank sequence entries) as coming from chromosomes other than 4. It would appear that there have been at least four large insertions of NCBI contigs into the D4S394–D4S403 region of the NCBI2 assembly from elsewhere in the genome. Because we could not accurately assess the degree of misassembly or coverage within the 32 extra contigs, only the 10 NCBI contigs that matched BAC sequences from the D4S394–D4S403 region were retained, in the original NCBI order, to represent the NCBI2 assembly.

NonRedundant Genomic Sequence Data Set Construction

A nonredundant genomic sequence data set (NR) was constructed to test the relative amounts of EMBL HTG genomic sequence from the region included by the assemblies. Because the CELERA assembly used an older (September 1, 2000) version of the IHGSC sequence data than the other assemblies, all sequence versions of BAC sequences used to make NR were those available before September 1, 2000. In three cases, the BAC sequences did not exist before this date and these were omitted (BAC sequences AC080003, AC084048, and AC079899), leaving 55 BAC sequences from the region. All BAC sequences were separated into their component contigs and were then further fragmented into sequences ≤500-bp long and masked for repetitive sequence using RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html; A.F.A. Smit and P. Green, unpubl.). All masked fragments were BLASTN searched against all others, and all BLASTN matches with E values $\leq 1 \times e^{-10}$ and ≥98% identity were deemed significant. Many fragments had significant matches to more than one other fragment. To eliminate this redundancy, all sequences that matched one or more other sequences were clustered into groups on a single-link basis. For example, if sequence A significantly matched sequences B and C, then all three were included in the same group, regardless of whether there was a match between B and C. Only one 500-bp sequence was taken from each group, together with all fragments with no BLASTN matches, and these sequences were screened for high numbers of masked nucleotides. All sequences with <150 bp of unmasked nucleotides were removed, giving a total of 5808 nonredundant fragments across the region, containing 2,694,848 bp of genomic sequence. All assemblies were then assessed for the proportion of sequence included from this set (NR), by use of a BLAST threshold of ≥98% identity over ≥100 bp.

Ordered Marker Set Construction

An ordered marker set (the framework set) was constructed from a subset of the STS sequences reported to lie between D4S394 and D4S403 according to a detailed physical map of the region (Evans et al. 2001) that was constructed using the SAM contig assembly program (Soderlund and Dunham 1995). Any STSs showing ambiguity in ordering were excluded from the framework set. Because all STSs were designed from larger sequences (e.g., BAC end or coding sequences), these larger sequences were substituted for the shorter STS sequences in the framework set. No STSs included in the framework set contained sequence from the 4752-bp tandem repeat unit pRS447 (sequence accession D38378) (Kogi et al. 1997). The result was a framework set of 107 sequences accurately ordered across the region with an average marker density of 1 marker per 54.21 Kb, on the basis of an estimated size of 5.8 Mb. The order of markers given in the framework set was compared with that seen in the 11 completely sequenced BAC clones from the region. All framework markers were searched against the complete BAC sequences (11 BACs with a total length of 1,987,413 bp), and were deemed to be present in a BAC clone when they generated a BLASTN alignment ≥150 bp long and ≥98% identical. A total of 34 (32%) framework sequences were found in these complete BAC sequences, an average marker density of 1 marker per 58.45 Kb. In all cases, the ordering of framework sequences seen in the completed sequences was the same as in the framework set, with no duplications or omissions of any markers. Thus, as far as we can ascertain, the framework ordering of sequences accurately reflects the real order present in the genome. Olivier et al. (2001) used a BLAST threshold of 90% identity over 100 bp to examine inconsistencies in marker orders between STSs from the TNG radiation hybrid map and one of the UCSC assemblies used here (1/9/01 release of 10/7/00 version). Here, the framework marker set was compared with all assemblies by use of a more stringent BLAST threshold of ≥98% identity over ≥100 bp.

Hybrid Sequence Assembly

All 58 BAC sequences retrieved from EMBL HTG were fragmented into their constituent contigs according to the annotation in each sequence entry. All contigs from the CELERA assembly were fragmented on the basis of the incidence of runs of >2 nucleotides marked as N. These CELERA fragments contain CG and IHGSC sequence data. It would have been preferable to obtain fragments containing only CG sequence, but unassembled CG sequence data are not publicly available. The HTG EMBL and CELERA derived data gave a total of 1390 sequence fragments to form the basis for the HYBRID assembly. The CONSED sequence assembly tools (Gordon et al. 1998), incorporating the .longreads modified versions of the phrap and crossmatch programs, were used to assemble the 1390 fragments. Assembly was preceded by the production of artificial sequence chromatograms and sequence quality files for all sequences using the mktrace program (distributed with CONSED); this enables the user to view the contigs produced. The HYBRID assembly produced consisted of 382 contigs. Poor sequence quality at the end of reads and sequencing artefacts such as simple repeat expansions probably prevented the detection of some overlaps, causing the number of contigs to be artificially high. The BLASTCLUST algorithm (distributed by the NCBI with BLAST) was used to determine a minimum number of contigs, clustering CONSED contigs together on the basis of overlaps showing ≥98% identity over ≥0.1 of the length of at least one contig. However, even with this arguably liberal definition of overlap, the minimum number of contigs was found to be 298. Of the 382 original CONSED contigs produced, 50 were found to match sequences in the framework marker set. A subset of 12 of these 50 contigs were removed because they contained framework markers found in other contigs, and one other contig was removed because it contained a misassembly (i.e., it contained a marker order

that disagreed with the framework order). The remaining 37 contigs were ordered according to their framework marker content and constitute the HYBRID assembly.

Repetitive Sequence Content

Some basic measures of the repeat content of the various assemblies were made. Each assembly was masked for the presence of repetitive sequences using RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html.; A.F.A. Smit and P. Green, unpubl.) with default settings. The masked assemblies were then searched (BLASTN matches with $E \leq 1 \times e^{-10}$ and ≥98% identity) for the presence of copies of the 4752-bp pRS447 tandem repeat unit that has been reported in the region (Kogi et al. 1997).

ACKNOWLEDGMENTS

This work was supported in part by the UK Medical Research Council.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Aach, J., Bulyk, M.L., Church, G.M., Comander, J., Derti, A., and Shendure, J. 2001. Computational comparison of two draft sequences of the human genome. *Nature* **409:** 856–859.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L., and Lander, E.S. 2000. A SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407:** 513–516.

Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J., and Eichler, E.E. 2001. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11:** 1005–1017.

Blackwood, D.H., He, L., Morris, S.W., McLean, A., Whitton, C., Thomson, M., Walker, M.T., Woodburn, K., Sharp, C.M., Wright, A.F., et al. 1996. A locus for bipolar affective disorder on chromosome 4p. *Nat. Genet.* **12:** 427–430.

Chen, T. and Skiena, S.S. 2000. A case study in genome-level fragment assembly. *Bioinformatics* **16:** 494–500.

Eichler, E.E. 2001. Segmental duplications: What's missing, misassigned, and misassembled—and should we care? *Genome Res.* **11:** 653–656.

Etzold, T. and Argos, P. 1993. Transforming a set of biological flat file libraries to a fast access network. *Comput. Appl. Biosci.* **9:** 59–64.

Evans, K.L., Le Hellard, S., Morris, S.W., Lawson, D., Whitton, C., Semple, C.A., Fantes, J.A., Torrance, H.S., Malloy, M.P., Maule, J.C., et al. 2001. A 6.9-mb high-resolution bac/pac contig of human 4p15.3-p16.1, a candidate region for bipolar affective disorder. *Genomics* **71:** 315–323.

Gaasterland, T. and Oprea, M. 2001. Whole-genome analysis: Annotations and updates. *Curr. Opin. Struct. Biol.* **11:** 377–381.

Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* **8:** 195–202.

Hogenesch, J.B., Ching, K.A., Batalov, S., Su, A.I., Walker, J.R., Zhou, Y., Kay, S.A., Schultz, P.G., and Cooke, M.P. 2001. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106:** 413–415.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

International Human Genome Mapping Consortium. 2001. A physical map of the human genome. *Nature* **409:** 934–941.

Katsanis, N., Worley, K.C., and Lupski, J.R. 2001. An evaluation of the draft human genome sequence. *Nat. Genet.* **29:** 88–91.

Kogi, M., Fukushige, S., Lefevre, C., Hadano, S., and Ikeda, J.E. 1997. A novel tandem repeat sequence located on human chromosome 4p: Isolation and characterization. *Genomics* **42:** 278–283.

Olivier, M., Aggarwal, A., Allen, J., Almendras, A.A., Bajorek, E.S., Beasley, E.M., Brady, S.D., Bushard, J.M., Bustos, V.I., Chu, A., et al. 2001. A high-resolution radiation HYBRID map of the human genome draft sequence. *Science* **291:** 1298–1302.

Roach, J.C., Siegel, A.F., van den Engh, G., Trask, B., and Hood, L. 1999. Gaps in the human genome project. *Nature* **401:** 843–845.

Schuler, G.D. 1997. Sequence mapping by electronic PCR. *Genome Res.* **7:** 541–550.

Semple, C.A.M., Devon, R.S., Le Hellard, S., and Porteous, D.J. 2001. Identification of genes from a schizophrenia-linked translocation breakpoint region. *Genomics* **73:** 123–126.

Soderlund, C. and Dunham, I. 1995. SAM: A system for iteratively building marker maps. *Comput. Appl. Biosci.* **11:** 645–655.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., and Holt, R.A. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L., et al. 2001. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **29:** 11–16.

WEB SITE REFERENCES

http://www.ensembl.org/; the coding regions present in the draft sequence.

http://ftp.genome.washington.edu/RM/RepeatMasker.html; RepeatMasker.

http://genome.ucsc.edu/; the freely available Human Genome Project Working Draft at the University of California, Santa Cruz (UCSC) described by the International Human Genome Sequencing Consortium (2001).

http://public.celera.com/; the freely available Celera Genomics (CG) assembly described in Venter et al. (2001).

http://www.ncbi.nlm.nih.gov/; the freely available National Centre for Biotechnology Information (NCBI) assembly.

Received July 31, 2001; accepted in revised form December 14, 2001.