A physical map of the human genome

The International Human Genome Mapping Consortium*

* *A partial list of authors appears at the end of this paper. A full list is available as Supplementary Information.*

The human genome is by far the largest genome to be sequenced, and its size and complexity present many challenges for sequence assembly. The International Human Genome Sequencing Consortium constructed a map of the whole genome to enable the selection of clones for sequencing and for the accurate assembly of the genome sequence. Here we report the construction of the whole-genome bacterial artificial chromosome (BAC) map and its integration with previous landmark maps and information from mapping efforts focused on specific chromosomal regions. We also describe the integration of sequence data with the map.

The International Human Genome Sequencing Consortium (IHGSC) used a hierarchical mapping and sequencing strategy to construct the working draft of the human genome. This clone-based approach involves generating an overlapping series of clones that covers the entire genome. Each clone is 'fingerprinted' on the basis of the pattern of fragments generated by restriction enzyme digestion[1,2]. Clones are then selected for shotgun sequencing and the whole genome sequence is reconstructed by map-guided assembly of overlapping clone sequences[3].

The availability of the whole-genome clone-based map assisted the sequencing of the human genome in many respects. The fingerprinted BAC map made it possible to select clones for sequencing that would ensure comprehensive coverage of the genome and reduce sequencing redundancy. In addition, the challenge of sequence assembly was minimized by restricting random shotgun sequencing to individual clones. Furthermore, the clone-based map also enabled the identification of large segments of the genome that are repeated, thereby simplifying the assembly. Many IHGSC centres had developed chromosomal maps and resources that were not integrated, so it was essential to have a unifying genome map to enable localization of clones, with respect to previously sequenced clones, before they were sequenced. The accurate fingerprinting and sizing of each clone enabled us to verify the accuracy of shotgun sequence[4] assembly of each clone.

The human genome presented unique challenges for the development of a clone-based physical map. Its size of 3.2 gigabases (Gb), which is 25 times as large as any previously mapped genome, meant that proportionately more clones had to be analysed. Its greater complexity also made it more difficult to distinguish true overlaps, which was further complicated by the repeat-rich nature of the genome. Early efforts to construct clone-based regional and even chromosomal physical maps of the human genome using cosmid libraries derived from isolated human chromosomes met with limited success[5,6]. By contrast, maps based on sequence-tagged site (STS) landmarks provided greater coverage of the genome[7–9], as did genetic maps based on variations in simple sequence repeats in STS landmarks[10,11]. The development of P1-artificial chromosome (PAC)[12] and bacterial artificial chromosome (BAC)[13] cloning systems was pivotal to the success of the whole-genome map. They provided larger inserts, more stable clones and better coverage of the genome.

Clone-based maps similar to that described here have been important in the sequencing of most large genomes, including those of *Saccharomyces cerevisiae*[1], *Caenorhabditis elegans*[2] and *Arabidopsis thaliana*[14]. A clone-based map also contributed to the sequencing of the *Drosophila melanogaster* genome[15,16] and a combined mapping and sequencing strategy is being applied to the mouse genome[17,18]. This work illustrates the benefit of using the clone-based map in the assembly of the human genome sequence.

Construction of the whole-genome BAC map

The pilot phase of the sequencing project began in 1995, at which time efforts were renewed to develop clone-based maps covering specific regions of the genome. To construct these regional maps, we screened PAC and BAC clones for STS markers, fingerprinted the positive clones, integrated them into the existing maps, and selected the largest, intact clones with minimal overlap for sequencing.

To keep pace with the ramping up of the sequencing effort in 1998, the ongoing efforts to construct the whole-genome BAC map were increased approximately tenfold. The whole-genome BAC map was constructed in several steps. First we collected fingerprint data for a large sample of random clones from a genome-wide BAC library. We then assembled the BAC map, first by using the fingerprint data to cluster highly related clones automatically, then by further refining them manually, and last by merging contigs with related clones at their ends. Finally, in parallel with construction of the BAC map, we mapped the chromosomal positions of individual clones on the basis of landmarks from existing landmark maps.

Fingerprinting the BAC clones

In October 1998, we began fingerprinting 300,000 BACs from the RPCI-11 library[19] (http://www.chori.org/bacpac/). Redundancy of sampling was vital to achieve high continuity in the final map[14]. Assuming an average BAC insert size of 150,000 base pairs (bp) and a genome size of 3.2 Gb, this level of fingerprinting would provide roughly 15-fold coverage of the genome. The library was derived from male DNA, providing full coverage of all 24 human chromosomes but with half as much coverage of the sex chromosomes as of the autosomes. Our experience with the library found it to be of high quality with uniformly large-insert clones, few non-recombinant clones and little cross-contamination of source plates. The RPCI-11 library was one of the first libraries to meet the informed consent criteria in accordance with the NHGRI policy for the Use of Human Subjects in Large Scale Sequencing (http://www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/Human_subjects.html).

To meet the goal of fingerprinting 300,000 BAC clones in one year, we devised a tandem 121-lane agarose gel format, allowing the simultaneous electrophoresis of 50 standard 'marker' DNA lanes and of 192 BAC restriction digests (Fig. 1), thereby reducing the number of gels, without loss of restriction fragment size accuracy or fidelity of clone tracking (see Supplementary Information). With these and other improvements in the fingerprinting technology and resources, we increased throughput tenfold to process more than 20,000 fingerprints (which equates to approximately onefold clone coverage of the human genome) each week. We also sampled clones from the RPCI-13 and CT-C/D1 BAC libraries, which were constructed using a different restriction enzyme (Table 1). This provided differential sampling of the genome, given the different distribution of the restriction enzyme sites within the genome. In

addition, the RPCI-13 library is derived from female DNA, which improves the representation of the X chromosome in the whole-genome BAC map.

Assembling the BAC map

By the end of 1999, with the fingerprint data on the BAC clones entered into an FPC database[20–23] (http://www.sanger.ac.uk/Software/fpc/), we were ready to construct the initial fingerprint assembly that would form the basis for further work on the map. We experimented with various strategies for automated assembly that would be as complete and as consistent as possible (see Supplementary Information).

First, we edited the fingerprint data itself. In early tests of assembly, we found that the variability in the mobility of small fragments (< 600 bp) led to artefactually low estimates of overlaps between clones. We therefore removed fragments smaller than 600 bp before assembly. Similarly, variability in estimating band numbers in 'multiplets' (instances where more than one fragment is located at nearly the same position on the gel) also caused problems. To reduce the variability between the number of bands called in these multiplet situations and thus increase the reliability with which related clones are correctly overlapped, these fragments were collapsed to only a single band in the resulting fingerprint. This 'sanitizing' process resulted in clusters of increased reliability.

Second, we evaluated the impact of varying the threshold for the 'overlap statistic', which is a measure of clone similarity, and the tolerance for accepting two bands from different clones as the same. We compared the clusters obtained for consistency with known regions and with other mapping data for the fingerprinted clones (primarily radiation hybrid chromosomal localization data from the Stanford Human Genome Center (SHGC)). The parameters finally used (overlap statistic of 3×10^{-12} or about 75% clone overlap and 0.7 mm tolerance) balanced the total number of clusters (which decreased with less stringent parameters) and the number of chimaeric clusters (which decreased with more stringent parameters). The automated assembly of 283,287 BAC clones resulted in 7,133 clusters containing 93% of all fingerprints in the database (Table 2). The remaining unincorporated clones (singletons) were excluded, as they contained too few bands to be included by automated assembly under these conditions or simply had no closely related clones. These latter clones included artefacts such as clones that had rearranged or had poor quality data, as well as rare clones representing poorly sampled portions of the genome.

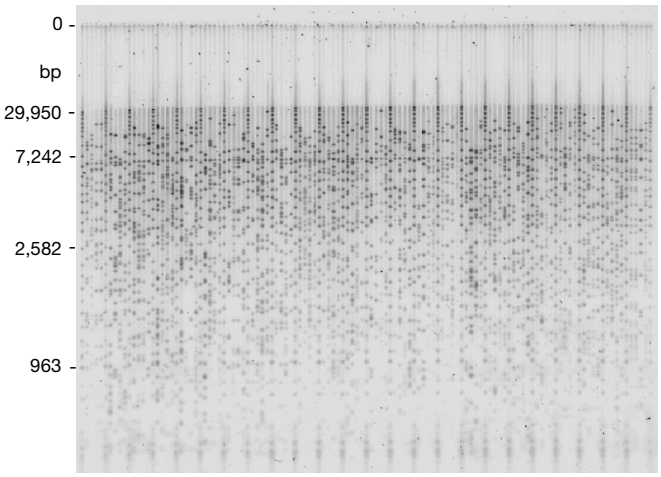

Figure 1 Example of the improved high-throughput fingerprint gel. BAC DNAs are digested with *Hin*dIII and visualized on a SYBR-green-stained 1% agarose gel. Every fifth lane contains a mixture of marker DNAs; the sizes of selected marker fragments are indicated. 0, origin of fragment migration.

Table 1 Sources of clones used

Library	Clones in current whole-genome map	Type	Vector	Enzyme	Average insert size (kb)
RPCI-4, -5*	568	PAC	pCYPAC2	*Mbo*I	116
RPCI-11*	272,027	BAC	pBACe3.6	*Eco*RI	174
			pTARBAC1†	*Mbo*I	196
RPCI-13*	59,051	BAC	pBACe3.6	*Mbo*I or *Dpn*II	149
CT-A, -B‡	228	BAC	pBeloBAC11	*Hin*dIII	120
CT-C, -D1‡	52,725	BAC	pBeloBAC11	*Hin*dIII	125
CT-D2‡	559	BAC	pBeloBAC11	*Eco*RI	190
Other§	10,231				

*http://www.chori.org/bacpac/
†RPCI-11, segment 5
‡http://informa.bio.caltech.edu/Bac_info.html
§Various clones from multiple libraries sent by collaborating centres.

As fingerprints from new clones were added after the initial assembly, there was a disproportionate increase in the number of singletons (Table 2). These new data were only incorporated into existing clusters or contigs if they added needed depth or helped to join contigs. We noted a further increase in singletons as new libraries were sampled (particularly from the RPCI-13 and CT-C/D1 libraries). One possible explanation is that these new libraries encompass regions of the genome not represented in the initial RPCI-11 library.

Most clones (97.5%) in the current whole-genome BAC map are derived from RPCI-11 (272,027/69.2%), RPCI-13 (59,051/14.9%) and CT-C/D1(52,725/13.3%) (Table 1). Although only about two-thirds of the fingerprint data are derived from DNA from a single individual, we did not experience any problems in assembly arising from polymorphisms between the individuals from whom the DNA was obtained.

Achieving map continuity

The goals of the manual editing were to refine the ordering of the clones within clusters to create contigs, to disassemble larger chimaeric contigs (representing clusters of two or more sets of non-overlapping clones) and to join contigs. This process involved first editing the fingerprint assemblies (using the tools encapsulated in FPC) to ensure that every clone within a contig was properly situated with respect to its most highly related neighbours, defined by fingerprint similarity[14] (see Supplementary Information). About 600 chimaeric clusters were identified and disassembled. To identify potential joins, we then used clones at the extreme ends of each contig to query the FPC database at a lower required fingerprint overlap stringency (overlap statistic of 1×10^{-8} or about 50% clone overlap) than was used during initial assembly. Joins were incorporated into the map if the fingerprinting data was logically consistent with the proposed map order (Fig. 2).

The most notable effect of the intensive editing was the greater than fivefold reduction in total contigs, from a high of 7,700 contigs after chimaeric contigs had been disassembled, to 1,246 by the 7 October 2000 data freeze of the draft genome sequence[3] (Table 2). The longest contig in this set encompasses more than 60 Mb of draft genome sequence and the mean contig size is estimated to be

Table 2 Status of FPC database after automated assembly and manual editing

	Automated assembly	Manually edited database
Date	December 1999	September 2000
BAC clones in FPC	283,287	372,264
Number of contigs	7,133	1,447
Clones in contigs	264,555	295,828
Number of singletons*	18,732	76,436
Contigs containing:		
>25 clones	3,012	912
9–25 clones	1,844	260
3–9 clones	1,957	204
2 clones	887	71

*Clones not incorporated into any contig; see text.

around 2.9 Mb. At the time of writing, the number of contigs had fallen further, to just 965 contigs.

As the contigs became accurately positioned and oriented with respect to one another (see below) and with the emergence of the draft sequence, end clones of adjacent contigs with overlapping sequence were recognized. After inspection of the sequence overlap to rule out shared sequence resulting from internal repeated segments, about half of the candidate joins were well supported by the fingerprint data and were integrated into the map. Another 62 had unconvincing evidence of overlap based on fingerprints but were tagged as overlapping on the basis of sequence alone.

The contigs appeared to be appropriately distributed among the chromosomes on the basis of the expected size of the chromosomes. The number of contigs per chromosome varies with the size of the chromosomes and the efforts made at closure (Table 3). Chromosomes 6, 7, 13, 14, 15, 20 and Y have relatively few remaining gaps, with 21, 29, 15, 21, 19, 10 and 8 contigs, respectively.

Integration of other map data

To increase the utility of the whole-genome BAC map, we incorporated various map data to anchor the contigs along the 24 chromosomes. Using selected markers from the CEPH Généthon genetic map[10], the GeneMap'99 genome-wide radiation hybrid map (http://www.ncbi.nlm.nih.gov/genemap/)[24–26] and from plasmid library sequences prepared from flow-sorted chromosomes (Sanger Centre, unpublished data), we hybridized 13,695 markers against colony filter replicas of the RPCI-11 library. This enabled us to position 96,283 different BAC clones as genome anchor points for the contigs.

In addition, because the RPCI-11 library was used for other genome initiatives, much additional marker information was available from other laboratories. Importantly, 9,018 STSs derived from BAC end sequences were assigned to chromosomes (D. R. Cox, unpublished data), with many of these selected deliberately because they came from clones in unlocalized contigs. Although over 15% of the available BAC end sequences of the RPCI-11 library are reported to be apparently mislabelled with respect to the microtitre well address from which they originated[27], two or more independently derived BAC end positions reliably yielded the correct chromosomal assignment of many contigs. In addition, chromosomal assignment and integration of cytogenetic map positions were achieved by utilizing 3,412 BACs mapped by fluorescence *in situ* hybridization (FISH) data[28].

As the working draft sequence accumulated, known markers within the sequence were readily identified by electronic PCR (ePCR), a program that searches sequence for STSs by identifying the associated primer sequences in the correct orientation and with correct spacing[29]. These data were incorporated into the FPC

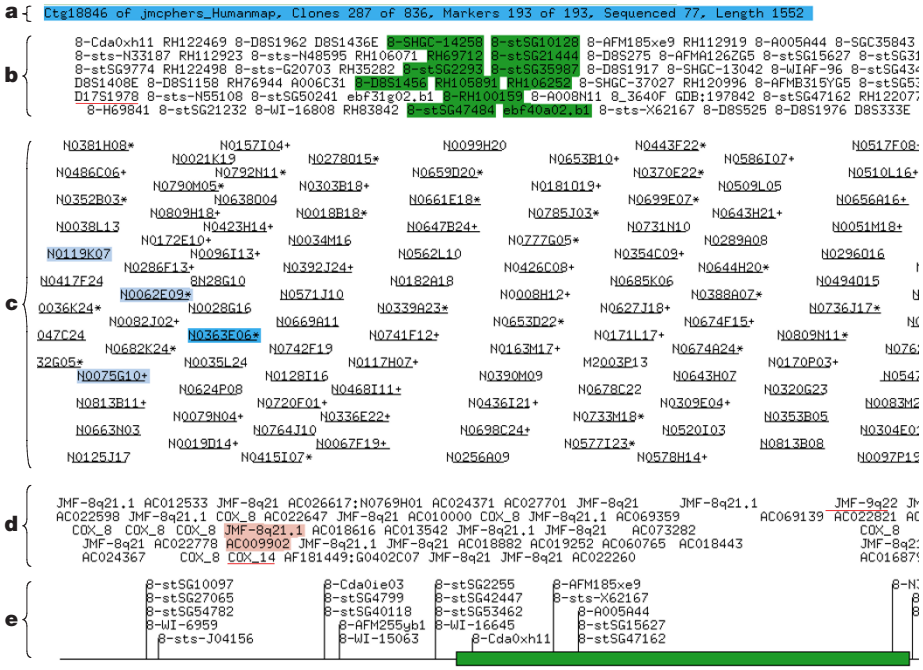

Figure 2 Example contig from the whole-genome BAC map. Portion of contig shown is localized to chromosomal region 8q21, composed of 836 BAC clones ordered by restriction fingerprint mapping. **a**, Contig summary information. Only 287 of the 836 clones are displayed. Redundant clones are 'buried' in their parent clones as indicated by the + and * clone name suffices (see (**c**)). The contig contains 193 markers; 77 clones have been selected for sequencing. Contig length: 1,552 unique restriction fragments (~ 6 Mb). **b**, Markers associated with clones in the display. Green: specifically associated with clone N0363E06 (aqua in **c**). There are 69,507 markers currently in the database associated with clones, largely by ePCR. Only one marker of the 62 shown is inconsistent with the 8q21 localization of this contig (D17S978, red underline). This is probably not a unique marker in the genome as the clone with which it is associated also contains several chromosome 8 markers. **c**, Partial display of the contig, showing 112 of the 287 clones visible with this view. Blue, example clones selected for sequencing. These clones were believed to overlap as they shared several restriction fragments; overlaps have been confirmed by working draft sequence. **d**, Data associated with the clones in **c**. FISH data (for example JMF-8q21.1) is consistent, except for one clone (N005M18, 9q22, red underline), probably owing to a clone-tracking error (the placement of the associated accessioned sequence (AC022821) in this location is supported by sequence overlap with surrounding clone sequences (clone N0813B08, AC069139)). Chromosomal localization of clones using STSs derived from BAC end sequences (for example, COX_8) is also consistent, with one exception (N0028G16, chromosome 14 COX_14, red underline), probably owing to incorrect association of an end sequence with a clone name in the BAC end sequence database. GenBank accession numbers are indicated. Sequences were mapped to the associated clone using *in silico* restriction digests, BAC end sequences and sequence overlap. Around 11.5% of the accessioned sequences have an incorrect clone name in their GenBank record, so proper placement of the sequence relative to the physical map was achieved in this manner. N00792N11 and N0096I13 are associated with accessioned sequences AC026617 and AF181449, respectively. The incorrect clone name referenced in their sequence records is indicated. **e**, Markers associated with the GeneMap'99 radiation hybrid map. Several are associated with clones in this contig (**c**), further positioning this contig within the genome.

database. The combined ePCR and hybridized data sets contained 69,507 markers, including 1,659 polymorphic markers from the Généthon genetic map. We primarily used GeneMap'99 for further anchoring and ordering of contigs, as it has a substantial marker set (> 50,000), is well integrated with the Généthon genetic map and provides local ordering at < 1 Mb resolution. Once sequenced clones could be reliably associated with the fingerprinted clones[3], we could use the marker content of sequenced clones determined by ePCR to order and orient contigs more reliably. We used markers found on any of six maps (Généthon genetic map, Marshfield genetic map[11], WIBR YAC STS-content (http://carbon. wi.mit.edu:8000/cgi-bin/contig/phys_map), GeneMap'99, SHGC G3 radiation hybrid map (http://www-shgc.stanford.edu/Mapping/rh/index.html) and NCBI framework map (http://www.ncbi.nlm.nih.gov/genome/guide/)) to orient contigs with respect to the majority consensus of all maps examined.

Integration of specific mapping efforts

We integrated regional map data into the whole-genome BAC map from other genome centres (see list at http://www.nhgri.nih/), which enriched the map and helped in the selection of clones for sequencing as it minimized redundancy and improved coverage. The regional mapping data included those for chromosomes 12 (ref. 30), 14 (ref. 31) and Y (ref. 32), and 1, 6, 9, 10 13, 20, 22 and X (ref. 33). We also integrated mapping data for chromosome 19 (Lawrence Livermore National Laboratory, http://www-bio.llnl.gov/bbrp/genome/html/chrom_map.html) and a 20-Mb segment of chromosome 15 (University of Washington). Telomeric contigs were identified and positioned where possible, as described elsewhere in this issue[34].

Some mapping efforts employed clone resources other than the RPCI-11 BAC library. In these cases, clones were sent by these centres, fingerprinted at the Washington University Genome Sequencing Center (WUGSC) and incorporated into the whole-genome BAC map and FPC database. These clones included those from regions of chromosomes 5 (J. Cheng), 8 (A. Rosenthal and N. Shimizu[35]), 11 (Y. Sakaki) and 17 (J. Ramser). In addition, we used computer-generated restriction digests, or *in silico* digests, of sequences in GenBank to incorporate these clones into the whole-genome BAC map.

Table 3 Chromosomal assignment of contigs

Chromosome	Number of contigs	Estimated size (Mb)*
1	119	263
2	54	255
3	77	214
4	42	203
5	51	194
6	21	183
7	29	171
8	46	155
9	25	145
10	23	144
11	36	144
12	30	143
13	15	114
14	21	109
15	19	106
16	64	98
17	50	92
18	28	85
19	59	67
20	10	72
21	17	50
22	10	56
X	163	164
Y	8	59
UL†	229	–
Total	1,246	3,286

*Ref. 45
†UL consists of clone contigs that could not be reliably placed on a chromosome.

Accuracy of chromosomal positions

As an independent assessment of the accuracy of assigning a chromosomal position to contigs, we used aliquots of the BAC DNA from 96 fingerprinted clones (RPCI-11, clones 512M01–512O24) as FISH probes to metaphase chromosomes (see Methods). Of the 96 BACs examined, 87 were successfully assigned to a single chromosome band. The remaining clones either failed to label (six) or were associated with multiple chromosome bands (three). The chromosomal localization of 82 (94%) of the mapped BACs agreed unambiguously with the derived chromosomal assignment, based on STS content, of the contig into which the corresponding fingerprint had assembled. A single BAC mapped to one of the two positions that were equally well supported by the marker content of its associated contig. The remaining four BACs were associated with fingerprints in contigs that had no mapped marker content and thus were not previously localized. In summary, the FISH mapping data did not conflict with any of the chromosomal assignments of the contigs examined.

In addition, we selected a minimal tiling path of eight clones from a random contig. DNA remaining from the fingerprinting of these clones was used for FISH mapping. All eight clones co-localized to chromosomal segment 8q21.1. This was consistent with other FISH data (B. Trask; 8q21.1), radiation hybrid data (D. R. Cox; chromosome 8) and ePCR of 12 markers mapping to chromosome 8 (Fig. 2).

Accuracy of clone order

The integration of independent map data and the emerging sequence information enabled us to monitor the fidelity of the developing map. We regularly checked that the predicted clone order was reflected in the overlaps of the sequenced clones. The ongoing assignment of chromosomally positioned markers to clones and contigs provided a useful check for possible false joins between unrelated contigs. These checks for clone order and contig fidelity were carried out much more extensively once the draft genome sequence was assembled and additional marker data incorporated. Overall, local clone order agreed with the overlaps demonstrated by sequence.

Comparison of the chromosome 12 STS-content BAC map[30] with the fingerprint BAC map of the same chromosome provided an important test of the accuracy of clone ordering. The two maps were derived independently, but used the same RPCI-11 library. The maps are consistent in clone ordering and provide complimentary resources: the chromosome 12 STS-content BAC map provides more accurate contig anchoring and orientation, and our map provides more depth of clone coverage. Furthermore, these maps, while sharing some gaps, largely closed gaps for each other, underscoring the benefit of the complimentary mapping strategies. After integration, the resulting map consisted of four contigs on the short arm and 34 on the long arm, and this has been further reduced to 20 contigs by continued gap closure methods[30].

Duplications and repeats

Two problematic aspects of the genome still need to be resolved: large (> 150 kilobase (kb)) recently duplicated segments and smaller tandemly repeated sequences extending for > 100 kb. Analysis of the total clone population shows that about 1% of clones have unusually high numbers of closely related clones (>75% shared bands), indicative of large repeated sequences. In some cases, minor differences in band patterns have allowed some complex repeats to be tentatively teased apart, but many of these have yet to be investigated in detail at the sequence level (an exception is the Y chromosome[32]). In other cases, only more complete and finished sequence will clarify mapping data for these regions.

The presence of extensive smaller tandemly repeated sequences (which sometimes are not even successfully cloned) results in clones that resemble small insert and badly deleted clones, which we avoided including in the map. However, unlike the small and

deleted clones, tandem repeats are present in multiple independent clones that display a similar fingerprint pattern. Sequence analysis of some of these repeat sequences shows that they are related to centromeric and ribosomal-repeat-related repeats, among others.

Gaps in the map

The remaining gaps, currently fewer than 1,000, are likely to stem from a variety of causes. There may be overlaps between end clones that are too small to be detected by fingerprints and which will only be recognized once the end clones are sequenced. Gaps can also arise because of misassemblies in the map, particularly where a duplicated segment is inappropriately designated to represent just one region. Some gaps may be detected from analysis of other BAC DNA libraries constructed using different restriction enzymes. Other gaps may arise simply because clones are not recovered at sufficient frequencies in BAC or PAC large insert libraries—clones spanning these gaps could potentially be detected in YAC libraries or might need to be recovered using special approaches.

Coverage of the genome

To estimate the fraction of the genome that was represented in the whole-genome map, we analysed chromosomes 21 and 22. Using *in silico* digest methods, we estimated the coverage of the fingerprint map encompassing finished chromosomes 21 (ref. 36) and 22 (ref. 37). Simulated 175-kb clones were created and digested *in silico* from the contiguous sequences for these chromosomes; each clone overlapped by 40%. We compared these digested simulated clones against the FPC database at high fingerprint overlap stringency. For chromosome 21, 316 simulated clones were created, of which 315 had at least 15 *Hin*dIII restriction fragments; clones containing fewer bands are difficult to compare. Of the 315 simulated fingerprints, 309 (98%) matched a related clone in the whole-genome BAC fingerprint FPC database. Similarly, for chromosome 22, 308 simulated clones were created. Of those, 303 had more than 15 *Hin*dIII restriction fragments and, when compared to the entire FPC fingerprint database, 297 (98%) found a related clone. This analysis was repeated using a 210-kb *in silico* clone size with 90% overlap, with similar results. Each of these chromosomes has four sequence gaps that are estimated to encompass 1.6% of the chromosome; therefore, the confirmed clone coverage of the euchromatic region of these chromosomes is around 96%. Collectively, these chromosomes represent approximately 3% of the genome. It is probably reasonable to extrapolate that this level of clone coverage will be found throughout the genome.

Clone selection for sequencing

The whole-genome BAC map was, and continues to be, used to select clones for sequencing. We devised algorithms for automatic high-throughput selection of BACs, which specifically chooses clones from contigs lacking sequenced clones ('seed clones') and clones that extend from already selected clones (see Supplementary Information). We took several issues into account when developing these programs. First, we had to devise methods compatible with a constantly and rapidly evolving map, which had considerable new information added to it each week. Second, we had to avoid clones representing genomic regions already sequenced from libraries other than RPCI-11. Third, we wished to select only clones not deleted or otherwise rearranged and thus faithfully represent the underlying genome.

The fingerprint map was used initially to identify nonredundant seed clones for sequencing when only a small portion of the RPCI-11 clones had been fingerprinted. As described above, we used all available forms of mapping data to localize the clones, and thus the contig. Once an appropriate contig was identified, the program looked for the largest clone (smaller than 225 kb to avoid artefacts) in the contig. The program also checked the fidelity of the clone by comparing its bands against other clones. We avoided end clones, as they inevitably had bands that could not be confirmed. In addition, a clone registry was developed (NCBI) to track clones selected for sequencing by any centre, and contigs with these clones were also avoided, as were contigs containing clones with similarities to other clones in GenBank as detected by *in silico* digest.

The next step in automated clone selection was to extend progressively from the seed clones using tools to search for appropriately overlapping clones. Neighbouring clones were evaluated using the overlap statistic to provide a tentative clone order. Clones within a specified range of overlap statistic were evaluated for the total size of shared bands. The amount of acceptable overlap was also specified (typically 25%). Any candidate in turn was evaluated against an intermediately positioned clone to ensure that the overlap was genuine and was compared to existing data using the clone registry and *in silico* digests to avoid redundancy.

These automated tools were used until late January 2000, when the manually evaluated contigs became available, allowing the selection of minimal tiling paths based on these clone orders. In the course of generating the working draft, more than 10,000 BAC clones were selected for the sequencing pipelines at the WUGSC, Whitehead Institute for Biomedical Research and the Stanford Genome and Technology Center using these tools and the evolving whole-genome BAC map. A check of 518 overlaps between finished clones selected both manually and through the automated methods at WUGSC shows that they have an average overlap of 47.5 kb with their neighbours, or about 28%: this is an acceptable degree of overlap, given the relatively dense seeding that occurred, and the importance placed on achieving coverage.

Sequence map of the human genome

Although the whole-genome BAC map was constructed primarily to exploit the coverage of high-redundancy BAC libraries for use in sequencing the human genome, it has served to integrate the sequences in GenBank[38] with the physical map. This was needed to guide the long-range assembly of the working draft sequence and to identify all remaining gaps in this sequence map so that spanning clones could be selected. By using *in silico* digests to generate fragment size information that could be compared to the fingerprints in the FPC database, virtually all except for short sequences (such as individual cosmids) in GenBank were positioned onto the whole-genome BAC map (see Methods). Additional information, such as BAC end sequence alignment and clone sequence overlap, was used to augment the *in silico* digest placement (if needed) and, in some cases, multiple sequences were positioned as part of larger assemblies of overlapping sequences (NT segments, NCBI). From these analyses, we determined that as many as 11.5% of the sequences in GenBank had incorrect clone names referenced in their GenBank records, probably owing mostly to data tracking and clone retrieval errors at the genome centres. A consequence of these naming errors was that many contigs contained clones associated with multiple markers determined by ePCR that mapped collectively to a single region of the genome, but were inconsistent with the remaining clone-to-marker associations in the contig. This was a direct result of incorrect clone names being associated with sequences and hence, incorrect assignment of the markers to those clones in the FPC database. Mapping of sequences to the clone map corrected the naming errors and resolved seemingly out of place ePCR markers once the sequence in which they were detected was properly situated within the ordered contigs. We have found that the correct clone can be retrieved 95% of the time using the whole-genome BAC map, as judged by comparing the fingerprint obtained for the retrieved clone with that in the database. Some clones could not be retrieved owing to growth failures, and others represent data-tracking errors within the fingerprint set. The high level of redundancy of the whole-genome BAC map allows a substitute clone to be readily selected to replace the 5% of clones that are not recovered on the first attempt.

Once the sequences were aligned to the whole-genome BAC fingerprint map, we used these data as a foundation for determining a nonredundant sequence path across the genome. The map order and placement of the sequences with respect to the whole-genome BAC map were considered in the sequence assembly to minimize errors due to potential false assignment of overlaps between related but not identical sequences. The BAC map placements were used as a localization guide only and did not completely constrain the sequence assembly, to avoid any propagation of errors and imprecision of clone placement. The analysis of markers identified within the genome sequence enables a detailed comparison of the whole-genome BAC map with other established landmark-content, radiation hybrid and genetic maps[3]. There was overall agreement between the sequence assembly that overlays the whole-genome BAC map and other existing maps, with local exceptions. In most instances, these local disagreements indicated the need simply to reverse the current orientation of the underlying BAC contig, and this has been done in the present version of the map.

Conclusions

The whole-genome BAC map allowed the integration of a range of data, including FISH cytogenetic clone localizations, landmark data obtained by PCR and hybridization screening, clones from other libraries with associated map data, and working draft and finished clone sequence and associated ePCR landmarks. New data will continue to be incorporated into this growing database. The entire FPC database of the human genome BAC fingerprint map can be obtained from http://genome.wustl.edu/gsc/human/Mapping/index.shtml. A searchable AceDb[39] version of the whole-genome BAC map is also accessible at http://genome.wustl.edu/ gsc/Search/db.shtml, and an overview of the map is available as Supplementary Information.

This clone-based map has been vital for the accurate assembly of the human genome sequence[3]. The BAC clones comprising the clone-based map also provide an integrated resource for analysis of chromosome structure, comparative genome hybridization[40] and functional genetics, including gene inactivation[41]. Together, the human genome clone map and the anchored sequence map provide synergistic resources for future analysis of the human genome. □

Methods

Regional approach to large-scale physical map construction

The general approach involved screening genomic BAC and PAC libraries by PCR or by probe hybridization using overgo probes[42] to identify clones corresponding to specific STS markers. Overgo probes are made by filling in the single-stranded overhangs of two overlapping oligonucleotides using radiolabelled nucleotides and Klenow polymerase. Typically, we used two 24-mers overlapping by 8 bp to generate a radiolabelled double-stranded 40-mer. Overgo probes were arranged in three-dimensional arrays with six probes on each axis (giving 216 probes each). A five-directional pooling strategy allowed resolution of 80–90% of all markers with only 30 hybridizations. More than 25,000 human and mouse markers have been associated with BACs using this probe type at the WUGSC (J. McPherson). Once identified, fingerprints were generated from marker-positive clones using *Hin*dIII restriction enzyme digests with fragment separation on 1% agarose gels[43], analysed using Image (http://www.sanger.ac.uk/Software/Image/)[20,21] and the fingerprints examined manually within FPC to build contigs and to select clones for sequencing that span contigs. Manual editing of the automated band calls was required because of inconsistencies in band identification.

Fluorescence *in situ* hybridization

Probes were generated from aliquots of the BAC DNA used to generate the *Hin*dIII fingerprints using the Prime-it Fluor labelling kit (Stratagene), which incorporates fluor-12-dUTP into the probe fragments by random priming. Probes were hybridized to chromosome spreads on slides with competitor DNA present. Slides were processed essentially according to standard methods[44]. Data were collected and analysed using a Zeiss Axiophot microscope equipped with the Genus camera setup and software (Applied Imaging Corporation).

Integration of sequenced clones using synthetic fingerprints from *in silico* digests

BAC-sized clones were simulated from finished contiguous sequenced regions of DNA. The sequences were cut into 175-kb fragments each with 40% overlap with the previous segments. Bands less than 600 bp were then removed from consideration to be consistent with the fingerprint data. Fingerprint data were converted from mobilities to sizes and clones from the fingerprinting effort could then be directly compared to sequenced clones from any library or group (when comparing size data, the FPC tolerance variable was set to 10). For clones that were not finished, each contig of the sequence was digested and all end fragments were removed. The remaining fragments were summed to create an *in silico* digest for unfinished clones.

Received 28 November; accepted 27 December 2000.

1. Olson, M. V. *et al.* Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl Acad. Sci. USA* **83,** 7826–7830 (1986).
2. Coulson, A., Sulston, J., Brenner, S. & Karn, J. Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **83,** 7821–7825 (1986).
3. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
4. Wilson, R. K. & Mardis, E. R. in *Analyzing DNA* (eds Birren, B., Green, E. D., Klapholz, S., Myers, R. M. & Roskams, J.) 398–454 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1997).
5. Doggett, N. A. *et al.* An integrated physical map of human chromosome 16. *Nature* **377,** 335–365 (1995).
6. Ashworth, L. K. *et al.* An integrated metric physical map of human chromosome 19. *Nature Genet.* **11,** 422–427 (1995).
7. Crollius, H. R. *et al.* An integrated YAC map of the human X chromosome. *Genome Res.* **6,** 943–955 (1996).
8. Bouffard, G. G. *et al.* A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. *Genome Res.* **7,** 673–692 (1997).
9. Whitehead Institute for Biomedical Research. YAC STS-content map of the human genome (cited October 2000) ⟨*http://carbon.wi.mit.edu:8000/cgi-bin/contig/phys_map*⟩ (1997).
10. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380,** 152–154 (1996).
11. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63,** 861–869 (1998).
12. Ioannou, P. A. *et al.* A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nature Genet.* **6,** 84–89 (1994).
13. Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA* **89,** 8794–8797 (1992).
14. Marra, M. *et al.* A map for sequence analysis of the *Arabidopsis thaliana* genome. *Nature Genet.* **22,** 265–270 (1999).
15. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
16. Hoskins, R. A. *et al.* A BAC-based physical map of the major autosomes of *Drosophila melanogaster* [published erratum appears in *Science* **288,** 1751 (2000)]. *Science* **287,** 2271–2274 (2000).
17. Pennisi, E. Genomics. Mouse sequencers take up the shotgun. *Science* **287,** 1179–1181 (2000).
18. Bouck, J. B., Metzker, M. L. & Gibbs, R. A. Shotgun sample sequence comparisons between mouse and human genomes. *Nature Genet.* **25,** 31–33 (2000).
19. Osoegawa, K. *et al.* A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.* (in the press).
20. Sulston, J. *et al.* Software for genome mapping by fingerprinting techniques. *Comput. Appl. Biosci.* **4,** 125–132 (1988).
21. Sulston, J., Mallett, F., Durbin, R. & Horsnell, T. Image analysis of restriction enzyme fingerprint autoradiograms. *Comput. Appl. Biosci.* **5,** 101–106 (1989).
22. Soderlund, C., Humphray, S., Dunham, A. & French, L. Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res.* **10,** 1772–1787 (2000).
23. Soderlund, C., Longden, I. & Mott, R. FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* **13,** 523–535 (1997).
24. Schuler, G. D. *et al.* A gene map of the human genome. *Science* **274,** 540–546 (1996).
25. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282,** 744–746 (1998).
26. The International Radiation Hybrid Mapping Consortium. A new gene map of the human genome: GeneMap'99. (cited October 2000) ⟨*http://www.ncbi.nlm.nih.gov/genemap*⟩ (1999).
27. Zhao, S. *et al.* Human BAC ends quality assessment and sequence analyses. *Genomics* **63,** 321–332 (2000).
28. The BAC Resource Consortium. Integration of autogenetic landmarks into the draft sequence of the human genome. *Nature* **409,** 953–958 (2001).
29. Schuler, G. D. Electronic PCR: bridging the gap between genome mapping and genome sequencing. *Trends Biotechnol.* **16,** 456–459 (1998).
30. Montgomery, K. *et al.* A high-resolution map of human chromosome 12. *Nature* **409,** 945–946 (2001).
31. Bruls, T. *et al.* A physical map of human chromosome 14. *Nature* **409,** 947–948 (2001).
32. Tilford, C. A. *et al.* A physical map of the human Y chromosome. *Nature* **409,** 943–945 (2001).
33. Bentley, D. R. *et al.* The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409,** 942–943 (2001).
34. Reithman, H. C. *et al.* Integration of telomere sequences with the draft human genome sequences. *Nature* **409,** 948–951 (2001).
35. Asakawa, S. *et al.* Human BAC library: construction and rapid screening. *Gene* **191,** 69–79 (1997).
36. The chromosome 21 mapping and sequencing consortium. The DNA sequence of human chromosome 21. *Nature* **405,** 311–319 (2000).
37. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).
38. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **28,** 15–18 (2000).
39. Eeckman, F. H. & Durbin, R. ACeDB and macace. *Methods Cell Biol.* **48,** 583–605 (1995).
40. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet.* **20,** 207–211 (1998).
41. Capecchi, M. R. Choose your target. *Nature Genet.* **26,** 159–161 (2000).
42. Ross, M. T., LaBrie, S., McPherson, J. & Stanton, V. P. in *Current Protocols in Human Genetics* (eds Dracopoli, N. C. *et al.*) 5.6.1–5.6.5 (Wiley, New York, 1999).

43. Marra, M. A. *et al.* High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7,** 1072–1084 (1997).
44. Lichter, P., Boyle, A. L., Cremer, T. & Ward, D. C. Analysis of genes and chromosomes by nonisotopic in situ hybridization. *Genet. Anal. Tech. Appl.* **8,** 24–35 (1991).
45. Morton, N. E. Parameters of the human genome. *Proc. Natl Acad. Sci. USA* **88,** 7474–7476 (1991).

Supplementary Information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank everyone who has contributed to mapping the human genome by providing data to GenBank and other publicly accessible web sites. As much as possible, these data have been incorporated into the map presented here. The fingerprinting project was funded as part of the Human Genome Project sequencing initiative of the NHGRI. The Keio group was supported in part by the Fund for Human Genome Sequencing Project from the JST and the Fund for "Research for the Future" Program from the Japan Society for the Promotion of Science. The RIKEN GSC group is supported by the Special Fund for Human Genome Sequencing from Science and Technology Agency, Japan and a Grant-in-Aid Scientific Research on Priority Area, "Genome Science" from Monbusho, Japan. Multiple US groups were funded by NIH/NCI and DOE. The MPIMG acknowledge grants from the Max-Planck-Society and the Federal German Ministry of Education, Research and Technology (BMBF) through Projekttraeger DLR, in the framework of the German Human Genome Project. Data management throughout the project was facilitated by using a suitably modified AceDb database (R. Durbin and J. Thierry-Mieg). For a complete list of all authors, see Supplementary Information.

Correspondence and requests for materials should be addressed to J.D.M. (e-mail: jmcphers@watson.wustl.edu).

The International Human Genome Mapping Consortium*

Washington University School of Medicine, Genome Sequencing Center: John D. McPherson[1], Marco Marra[1]*, LaDeana Hillier[1], Robert H. Waterston[1], Asif Chinwalla[1], John Wallis[1], Mandeep Sekhon[1], Kristine Wylie[1], Elaine R. Mardis[1], Richard K. Wilson[1], Robert Fulton[1], Tamara A. Kucaba[1], Caryn Wagner-McPherson[1] & William B. Barbazuk[1]

Wellcome Trust Genome Campus: Simon G. Gregory[2], Sean J. Humphray[2], Lisa French[2], Richard S. Evans[2], Graeme Bethel[2], Adam Whittaker[2], Jane L. Holden[2], Owen T. McCann[2], Andrew Dunham[2], Carol Soderlund[2]*, Carol E. Scott[2] & David R. Bentley[2]

National Center for Biotechnology Information: Gregory Schuler[3], Hsiu-Chuan Chen[3] & Wonhee Jang[3]

National Human Genome Research Insititute: Eric D. Green[4], Jacquelyn R. Idol[4] & Valerie V. Braden Maduro[4]

Albert Einstein College of Medicine: Kate T. Montgomery[5], Eunice Lee[5], Ashley Miller[5], Suzanne Emerling[5] & Raju Kucherlapati[5]

Baylor College of Medicine, Human Genome Sequencing Center: Richard Gibbs[6], Steve Scherer[6], J. Harley Gorrell[6], Erica Sodergren[6], Kerstin Clerc-Blankenburg[6], Paul Tabor[6], Susan Naylor[7] & Dawn Garcia[7]

Roswell Park Cancer Institute: Pieter J. de Jong[8]*, Joseph J. Catanese[8]*, Norma Nowak[8] & Kazutoyo Osoegawa[8]*

Multimegabase Sequencing Center: Shizhen Qin[9], Lee Rowen[9], Anuradha Madan[9], Monica Dors[9] & Leroy Hood[9]

Fred Hutchinson Cancer Research Institute: Barbara Trask[10], Cynthia Friedman[10] & Hillary Massa[10]

The Children's Hospital of Philadelphia: Vivian G. Cheung[11], Ilan R. Kirsch[12], Thomas Reid[12] & Raluca Yonescu[12]

Genoscope: Jean Weissenbach[13], Thomas Bruls[13] & Roland Heilig[13]

US DOE Joint Genome Institute: Elbert Branscomb[14], Anne Olsen[14], Norman Doggett[14], Jan-Fang Cheng[14] & Trevor Hawkins[14]

Stanford Human Genome Center and Department of Genetics: Richard M. Myers[15], Jin Shang[15], Lucia Ramirez[15], Jeremy Schmutz[15], Olivia Velasquez[15], Kami Dixon[15], Nancy E. Stone[15] & David R. Cox[15]

University of California, Santa Cruz: David Haussler[16,17], W. James Kent[18], Terrence Furey[17], Sanja Rogic[17] & Scot Kennedy[19]

British Columbia Cancer Research Centre: Steven Jones[20]

Department of Genome Analysis, Institute of Molecular Biotechnology: André Rosenthal[21], Gaiping Wen[21], Markus Schilhabel[21], Gernot Gloeckner[21], Gerald Nyakatura[21]*, Reiner Siebert[22] & Brigitte Schlegelberger[22]

Departments of Human Genetics and Pediatrics, University of California: Julie Korenberg[23] & Xiao-Ning Chen[23]

RIKEN Genomic Sciences Center: Asao Fujiyama[24], Masahira Hattori[24], Atsushi Toyoda[24], Tetsushi Yada[24], Hong-Seok Park[24] & Yoshiyuki Sakaki[24]

Department of Molecular Biology, Keio University School of Medicine: Nobuyoshi Shimizu[25], Shuichi Asakawa[25], Kazuhiko Kawasaki[25], Takashi Sasaki[25], Ai Shintani[25], Atsushi Shimizu[25], Kazunori Shibuya[25], Jun Kudoh[25] & Shinsei Minoshima[25]

Max-Planck-Institute for Molecular Genetics: Juliane Ramser[26], Peter Seranski[26,27], Celine Hoff[26,27], Annemarie Poustka[26,27], Richard Reinhardt[26] & Hans Lehrach[26]

1, Washington University School of Medicine, Genome Sequencing Center, Department of Genetics, 4444 Forest Park Boulevard, St. Louis, Missouri 63108, USA; 2, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK; 3, National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA; 4, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA; 5, Department of Molecular Genetics, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA; 6, Baylor College of Medicine, Human Genome Sequencing Center, Houston, Texas, USA; 7, University of Texas, San Antonio, Texas, USA; 8, Roswell Park Cancer Institute, Buffalo, New York 14263, USA; 9, Multimegabase Sequencing Center, Institute for Systems Biology, Seattle, Washington 98105, USA; 10, Division of Human Biology, Fred Hutchinson Cancer Research Institute, Seattle, Washington 98109, USA; 11, Department of Pediatrics, The Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; 12, Genetics Department, Medicine Branch, National Cancer Institute, Washington DC, USA; 13, Genoscope, Centre National de Séquencage, 2 Rue Gaston Crémieux, CP 5706, 91057 Evry, France; 14, US DOE Joint Genome Institute, Walnut Creek, California, USA; 15, Stanford Human Genome Center and Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA; 16, Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, California 95064, USA; 17, Department of Computer Science, University of California, Santa Cruz, Santa Cruz, California 95064, USA; 18, Department of Biology, University of California, Santa Cruz, Santa Cruz, California 95064, USA; 19, Department of Mathematics, University of California, Santa Cruz, Santa Cruz, California 95064, USA; 20, British Columbia Cancer Research Centre, 600 West 10th Avenue, Room 3427, Vancouver, British Columbia V5Z 4E6, Canada; 21, Dept. of Genome Analysis, Institute of Molecular Biotechnology, Beutenbergstrasse 11, D-07745 JENA, Germany; 22, Institute of Human Genetics, University of Kiel, Germany; 23, Departments of Human Genetics and Pediatrics, University of California, Los Angeles, California, USA;

24, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan; 25, Department of Molecular Biology, Keio University School of Medicine, 35 Shinanomachi Shinjuku-ku, Tokyo 160-8582, Japan; 26, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, D-14195, Berlin, Germany; 27, Abteilung Molekulare Genomanalyse, Deutsches Krebstorschungszentrum, Im Neuenheimer Feld 280, 69120, Heidelberg, Germany.

** Present addresses: British Columbia Cancer Research Centre, 600 West 10th Avenue, Room 3427, Vancouver, British Columbia V5Z 4E6, Canada (M.M.); Clemson University Genome Institute, 100 Jordan Hall, Clemson University, Clemson, South Carolina 29634-5727, USA (C.S.); Children's Hospital Oakland Research Institute, BACPAC Resources, Oakland, California 94609, USA and Pfizer Global Research & Development, Alameda Laboratories, Alameda, California 94502 USA (P.J.d.J., J.J.C., K.O.); MWG-Biotech AG, Ebersberg, Germany (G.N.).*