Published online July 9, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 12 3651–3660
doi:10.1093/nar/gkh700

A set of BAC clones spanning the human genome

Martin Krzywinski, Ian Bosdet, Duane Smailus, Readman Chiu, Carrie Mathewson, Natasja Wye, Sarah Barber, Mabel Brown-John, Susanna Chan, Steve Chand, Alison Cloutier, Noreen Girn, Darlene Lee, Amara Masson, Michael Mayo, Teika Olson, Pawan Pandoh, Anna-Liisa Prabhu, Eric Schoenmakers[2], Miranda Tsai, Donna Albertson[3], Wan Lam[1], Chik-On Choy[4], Kazutoyo Osoegawa[4], Shaying Zhao[5], Pieter J. de Jong[4], Jacqueline Schein, Steven Jones and Marco A. Marra*

BC Cancer Agency Genome Sciences Center and [1]BC Cancer Agency, Vancouver, BC V5Z 4E6, Canada, [2]Department of Human Genetics, 417, University Medical Center Nijmegen, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands, [3]Cancer Research Institute, Box 0808, University of California at San Francisco, San Francisco, CA 94143-0808, USA, [4]BACPAC Resources, Children's Hospital Oakland, 747 52nd Street, Oakland, CA 94609, USA and [5]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Received May 21, 2004; Revised and Accepted June 22, 2004

ABSTRACT

Using the human bacterial artificial chromosome (BAC) fingerprint-based physical map, genome sequence assembly and BAC end sequences, we have generated a fingerprint-validated set of 32 855 BAC clones spanning the human genome. The clone set provides coverage for at least 98% of the human fingerprint map, 99% of the current assembled sequence and has an effective resolving power of 79 kb. We have made the clone set publicly available, anticipating that it will generally facilitate FISH or array-CGH-based identification and characterization of chromosomal alterations relevant to disease.

INTRODUCTION

Large insert bacterial artificial chromosome (BAC) (1) fingerprint maps have been developed for several organisms to create genome-ordered clone resources and often to provide resources for DNA sequencing efforts. For example, using BAC fingerprinting technology (2,3) we have generated for human (4), mouse (5), rat (6), cow (unpublished data) and other organisms, genome maps that each offer up to ~15-fold redundant coverage (on average, each region of the map is represented by 15 clones). Coverage redundancy is critical to achieving map contiguity and is used to provide evidence that individual clones are not cloning artefacts but high fidelity representations of the underlying genome. In genomic regions of single clone coverage this assurance is lacking, and indeed at least one report suggests that individual BACs may rearrange with a frequency of 10% (7). For mammalian-sized genomes, the number of clones fingerprinted to achieve a reasonable level of map contiguity is large. For example, in the case of the human BAC fingerprint map, more than 415 000 clones were fingerprinted (4). In the case of the mouse BAC fingerprint map, more than 300 000 clones were fingerprinted (5).

After maps are constructed, redundancy is unnecessary for a complete representation of the genome and most map driven sequencing efforts use the map to select a tiling set of clones that, as completely as possible, represent the map and therefore the underlying genome. For a typical mammalian genome, assuming 200 kb clone inserts, approximately 21 000 clones are sufficient to represent the genome. A clone set of this size is a substantial reduction from the size of the parent library or libraries used to construct the map. Such tiling sets have uses other than fuelling sequencing efforts. They lend themselves to a number of applications, ranging from focused studies on the structure and function of individual genes and gene families to fluorescence *in situ* hybridization and BAC array comparative genomic hybridization (8,9). These latter applications are particularly important for biomedical research.

We report here the results of our efforts to select a tiling set of clones representing the human genome. We used the available BAC fingerprint map, BAC end sequences and genome sequence in our clone selection process. We assessed the selected clones for coverage of the genome sequence and fingerprint map, and have made available online views of the clone set in the context of the map from which the clones were selected. The clone set has been rearrayed and is available from BACPAC Resources (bacpac.chori.org/pHumanMinSet.htm). The set has been used to make high-density microarrays for BAC array CGH (10). Detailed information about the clones in the set is available at mkweb.bcgsc.ca/bacarray.

MATERIALS AND METHODS

Map-based BAC set construction

For selection of clones, we used the human BAC fingerprint-based physical map (4) generated at Washington University Genome Sequencing Center. Clones were chosen from each of

*To whom correspondence should be addressed. Tel: +1 604 877 6084; Fax: +1 604 877 6085; Email: mmarra@bcgsc.bc.ca the 726 contigs to provide maximum coverage of the fingerprint map. Clones not assigned to contigs as well as buried (i.e. without a unique complement of fingerprint digest fragments) clones (11,12) were excluded from candidacy. We restricted our clone selection exercises to the readily available RPCI-11 (7), RPCI-13 (7) and Caltech D1/D2 (informa.bio.caltech.edu/Bac_info.html) libraries. The algorithm for clone selection was based on a clone-walking methodology, and each fingerprint map contig was treated independently. For each map contig, the starting set of clones eligible for selection consisted of the ordered set of clones assigned to the contig. The order of the clones was previously determined by the process of automated map creation (11,12) and subsequent manual curation at Washington University Genome Sequencing Center (www.genome.wustl.edu/projects/human). For our application, the availability of correctly ordered clones was key. Starting from the left end of each map contig, the first canonical (i.e. containing a unique complement of fingerprint digest fragments) clone from the ordered set was always selected. The next selection was chosen to have as close to, but no fewer than, four conserved bands with the previous selection. Conserved bands are defined as bands present in the fingerprints of two overlapping clones and in the fingerprints of all clones located between them. Conserved bands emanate from the same DNA and their use minimizes false positives in determining clone overlap, since bands found in multiple adjacent intermedial clones in the ordered map represent the same digest fragment. Clones <100 kb or >200 kb, or having fewer than 20 or more than 50 HindIII fragments were chosen only where map coverage could not be provided by other eligible clones. No clones <15 kb or with fewer than 5 HindIII fingerprint fragments were chosen. To assist with positioning the selected BACs on the genome sequence assembly, those clones with informative BAC end sequence (BES) (13,14) records (i.e. containing sufficient non-masked content with unambiguous sequence hits to the July 2003 genome assembly) were selected preferentially in regions where the extent and depth of coverage would not be negatively impacted. Clones with BES hit coordinates that were inconsistent with their position in the fingerprint map were avoided in cases where equivalent map coverage could be obtained by selecting another clone. During the selection process, we aimed to enrich, again where possible, the clone set with clones having either existing FISH information (15,16) or sequence data. Gaps in clone set coverage identified by sequence coordinate analysis were addressed by selecting additional clones that spanned these gaps.

Clone validation and replacement

After the first round of map-based clone selection, which yielded 29 035 clones, all clones were digested using HindIII and fingerprints were generated as described elsewhere (2,3). Identification and sizing of bands in the clones' fingerprints was performed using BandLeader (17). All validation fingerprints were compared in an automated fashion to those stored in the physical map. In contrast to the validation fingerprints, the fingerprints in the map are sanitized—all fragments closer than 7 standard mobility units (this length unit corresponds to a size tolerance of 0.5% at 5 kb, 3% at 20 kb, 5% at 25 kb) have been replaced with a single fragment (4). This sanitization process, motivated by the historical difficulty in determining automatically restriction fragment copy number for multiple co-migrating fragments (multiplets), artificially lowers the apparent clone size by up to 30%. The validation fingerprints did not require sanitization, due to the availability of our new band calling software technology (17). The fingerprint comparison was made on the basis of the Sulston score (18), which corresponds to the probability that two fingerprints share similar fragments by chance. Each matching validation clone fingerprint was assigned a rank, from 1 to 10, indicating the strength of the match with the corresponding map clone fingerprint. Clones in the set with rank n had $n - 1$ map clones which were more similar than the expected map clone. Fingerprints of clones with a rank over 3 were visually examined (5272 clones). Fingerprints for 1978 clones in the set did not match their corresponding fingerprints in the physical map. The discrepancies could be categorized as resulting from clone tracking errors either during the generation of the fingerprint map or in the generation of our rearrayed clone set, from cross-well contamination, or from situations in which the validation fingerprinting process failed.

A second round of clone selection was performed to maintain the coverage represented by the 1978 failed clones. For each failed clone, neighboring clones were selected from the map to provide equivalent coverage. In total, 4531 clones were selected from the fingerprint map as replacements. These clones were sampled from RPCI-11, RPCI-13 and Caltech-D, in roughly the same proportion as for the final set (87%: 8%:5%). An additional 1258 clones were selected to close gaps >10 kb based on the June 2002 UCSC assembly (hg12). Approximately 755 of these clones were not in the physical map. A second round of fingerprint verification performed on the replacement clones identified 413 clones that did not match their map fingerprints. These clones were rejected from the set. The v1.0 set contains 32 432 validated clones and is available through BACPAC (www.chori.org/bacpac/pHumanMinSet.htm). Since the release of v1.0 we have removed 18 clones from the set, on account of poor fingerprint validation, and added an additional 441 clones to address gaps in coverage based on the July 2003 UCSC assembly (hg16). With these additions, the v1.1 set contains 32 855 clones.

Map coverage and representation validation

Map coverage was determined by analysing the total number and depth of cbmap units [a cbmap unit is equivalent to a detected and confirmed restriction fragment, regardless of fragment size (11,12)] that were covered by the clone set and the overlap distribution between map-adjacent clone selections. Coverage of cbmap units was determined by partitioning the cbmap scale into regions which were covered by (i) map clones, (ii) clones derived from the sampled libraries (RPCI-11, RPCI-13 and Caltech-D) and (iii) clones belonging in the clone set. The regions of coverage were treated as partitions on the cbmap scale and were compared using set operations such as union, intersection and difference to determine map representation and to identify gaps in coverage.

Overlap between map-adjacent clones in the set was calculated by computing the number of shared fragments, number of conserved fragments and the Sulston score between the clones. For cases where the number of shared fragments was fewer than 8, and the number of conserved fragments was fewer than 4 and the total size of shared fragments was <30 kb, overlap was considered weak and the clones were considered non-overlapping in the context of their map position.

To determine the overall representation of the fingerprint map in our clone set, all remaining canonical clones from the map that were not selected were compared to the selected set. For each map clone, the top 10 hits to the clone set, as ranked by the Sulston score, were extracted and the top hit in the same contig was identified as the closest match. Fingerprints were compared using a standard mobility tolerance of 7. The number of shared bands, overlap and Sulston score between the map clone and its closest match in the clone set was examined.

Sequence coverage

Sequence coverage was calculated by first determining precise sequence coordinates for as many clones as possible. The validation fingerprints of the clones were used to localize them to the genome in the following manner: for each clone, the region of the sequence from which the clone was derived was determined using the clone's map neighbours with BAC end sequence (13,14) hits. Five left map neighbours and five right map neighbours were identified and their BES hits were used to demarcate a region of the genome. Only neighbours whose BES hits landed on the same chromosome as the majority of the clones in the map contig were used. The clone's own BES hits were not used in determining the sequence region in case the clone's map position was incorrect or the BES hit coordinates did not reflect the actual position of the clone or were actually associated with another clone. The clone neighbourhood was enlarged by 1 Mb in both directions to minimize the effect of local inconsistencies. The neighbourhood assembly was digested *in silico* and the fingerprint of a sliding window of 120 fragments, created every 20 fragments, was matched to the clone's fingerprint. For the window position which corresponded to an *in silico* fingerprint that matched the most fragments, a sliding subwindow, having 1.4 times more fragments than the clone, was created every 5 fragments. The fingerprint matching was performed with a tolerance of 2% for fragments <10 kb, 3% for fragments 10–15 kb, 4% for fragments 15–25 kb and 5% for fragments >25 kb. This tolerance profile approximates a standard mobility tolerance of 7, the cutoff used to generate the fingerprint map. The subwindow which matched the most fragments was used to determine the clone coordinates. The clone was initially determined to start/end at the first fragment which was part of a 7 matched-fragment run which started with at least two matching fragments and had no more than 2 unmatched fragments. The end positions were subsequently refined by attempting to match fragments found in the clone's fingerprint which did not match the *in silico* digest from the demarcated sequence region to *in silico* fragments flanking the ends. We allowed for the detection of junction fragments, comprising both clone insert and a portion of the vector, by transforming each unmatched flanking *in silico* fragment by adding the edge fragments of the vector. These *in silico* anchors were accepted only in cases which satisfied all of the following: (i) over 70% of the fragments in the anchor matched the clone's fingerprint, (ii) the size of the anchor was between 0.2 and 2.0 times the clone's size, (iii) the anchor contained at least 15 matching fragments and (iv) the anchor fragment density, defined as the ratio of matching fragments to all fragments, excluding undetectable fragments (<600 bp) and poorly sized fragments (>30 kb), was at least 0.6. The final sequence coordinate for a clone was taken as the BES coordinates, assembly coordinates, or *in silico* coordinates, in this order of priority.

To assist in determining coverage by the clones in the set representing regions for which sequence coordinates could not be determined, the fingerprint map was used in the following manner: for every fingerprint map contig, an undirected graph of overlapping clones was created, separating the contig into strongly connected components. Two clones shared an edge in the same component if they had at least 4 conserved bands and were overlapped by more than 5 cbmap units. Clones which were not overlapped by 5 cbmap units had to share at least 6 conserved bands. For each strongly connected component, the leftmost and rightmost clone with sequence coordinates was used to locate the component within the assembly and this region was considered covered by the clone set.

Rearraying

The initial set of 29 035 clones was rearrayed at the Genome Sciences Center, Vancouver, BC, using a QPix colony picking platform (Genetix). The additional 5789 replacement clones were rearrayed manually at BACPAC Resources, Children's Hospital Oakland Research Institute (5174 RPCI-11 and RPCI-13 clones) and at the Genome Sciences Center (615 Caltech-D clones).

The rearraying of replacement clones from RPCI-11 and RPCI-13 libraries was performed in two steps to minimize manual operational error, well to well cross-contamination and non-growth wells. Clones for each library were first coordinated in 96-well format in the order of plate, row and column. BACs were inoculated, grown in 96 deep well blocks and kept at −80°C until all clones were rearrayed. The BACs were then condensed into 384-well plates using 96-pin tools.

RESULTS

Clone selection

An important resource for the initial selection of the clones was the human BAC fingerprint map (4). The fingerprint map is a manually curated and mature dataset which covers >99% of the genome at an average depth of 15X. We used the redundancy of clone coverage to specify desired clone overlap and size characteristics with the goal of achieving the best representation possible of both the fingerprint map and the genome sequence. We did not place special emphasis on selecting a 'minimal' tiling set, as was attempted for sequencing the genome, but instead restricted the number of clones selected to the number (approximately 30 000) that could readily be placed on glass slide 'microarrays' with existing slide printing technology. As a result, there are some clones in the set that overlap extensively with their neighbours. The extent of all clone overlap relationships is known.

We created the clone set from 3 of the 18 libraries used to construct the fingerprint map: RPCI-11 (7), RPCI-13 (7) and Caltech D1/D2 (informa.bio.caltech.edu/Bac_info.html). Clones from these libraries make up 98% of the 415 583 clones in the fingerprint map, 96% of the 158 082 canonical map clones assigned to contigs and 74% of the clones in the 'tiling path' set of clones that contribute 79% to the finished human genome sequence (19). Central to our purpose of creating a public clone resource, clones from these libraries are readily accessible to the scientific community.

Clone validation

Following rearraying, each clone in the set was assayed to confirm its identity and to provide data for localizing the clone in the genome. The first round of clone selection yielded 29 035 clones, representing 99% of the fingerprint map (Materials and Methods). Matches for all clones whose validation fingerprints matched the corresponding map fingerprints poorly (5272 clones) were examined manually (Materials and Methods). This included 2784 clone fingerprints identified by automated analysis as potential mismatches. The manual review revealed 1978 clones with verification fingerprints that did not match their map counterparts. This group of clones was considered ineligible for inclusion within the clone set, and consisted of 1143 clones that did not match the fingerprints in the human fingerprint map and 835 clones with fingerprints indicative of various technical failures, including cross-well contamination, or gross insert deletion. To maintain coverage in areas represented by these failed clones and fill in subsequently identified sequence coverage gaps, an additional 5789 BACs were selected and fingerprint-verified. Following the rearraying of replacement clones and fingerprint match verification, 413 clones failed the validation procedure and were removed from the clone set. Through analysis of the clone set coverage of the most recent sequence assembly (UCSC, hg16, July 2003) and fingerprint map data (WashU Genome Sequencing Center, September 17, 2003), we have selected an additional 441 clones to replace previously failed clones and to fill in newly detected coverage gaps. This second round of replacements clones is currently undergoing fingerprint verification. The majority of the clones in the current 32 855 clone version of the set are from RPCI-11 (92%) with the remaining 2% from RPCI-13 and 6% from Caltech D libraries.

Properties of the clone set

Out of a total of 32 855 clones in the validated set, 32 209 (98%) are represented in the fingerprint map, with 31 990 of these localized to contigs. The 646 clones not found in the fingerprint map were selected, based on the sequence assembly coordinates of their BES matches, to provide coverage of the sequence assembly. Based on analysis of our validation fingerprints, the average clone size and HindIII fragment counts for each library are 189 kb/46 (RPCI-11), 160 kb/37 (RPCI-13) and 146 kb/35 (Caltech-D). The average sizes of the clone set members based on BES data are 176 kb for RPCI-11 clones and 140 kb for Caltech-D, indicating that the sizes of the validation fingerprints overestimate the sequence-predicted sizes by 4–7%. This difference is in part due to the fact that HindIII does not cleave the BAC DNA at the precise cloning site employed during library construction, but instead results in fingerprints containing vector–insert junction fragments of unpredictable size.

During our selection procedure, we preferentially selected clones which had sequence accessions (BES records or insert sequence) or prior FISH data. Genbank (20) sequence accessions, excluding BES records, were available for 7345 clones in the set. Of these, the records indicated that 5196 clones were finished, 2014 clones were in the working draft form, 121 clones were in progress and 533 clones had been sequenced to low coverage. Paired-end BES coordinates or assembly coordinates were available for 15 486 of the clones (47% of clones in the set), providing a scaffold for localization of clones, including those without sequence information. In the set, 1179 clones (3.5%) had previously generated FISH data (15,16) (genome.ucsc.edu/goldenPath/hg16/database/fishClones.txt.gz).

Clone localization

We mapped the location of 96% of the clones in the set onto the genome sequence (UCSC, hg16, July 2003). 12 598 of the clones in the set were unambiguously positioned in the genome based on comparison of their BAC end sequences to the genome sequence and 2888 based on their coordinates in the sequence assembly. The remaining 16 200 clones were mapped onto the genome assembly by comparing their fingerprint patterns to computer-generated fingerprints derived from the genome sequence (*in silico* fingerprint mapping; Materials and Methods). This process was facilitated by the large fraction of clones (47%) that were mapped onto the genome by virtue of BES alignments or other sequence data, and by the generally accurate and lengthy BAC fingerprint contigs contained in the human map (4). The key to accurate and reliable *in silico* mapping involved identifying, for each clone to be mapped, neighbouring flanking clones within map contigs that were linked to the sequence through BES alignments. These flanking clones were used to demarcate the neighbourhood of the sequence assembly to which the mapping was applied (sequence coverage; Materials and Methods). Using this approach, mapping against the entire genome was generally not required, thus substantially reducing the complexity of the clone localization problem by focusing the exercise to the local neighbourhood defined by the flanking clones.

To validate our *in silico* fingerprint mapping clone localization approach, we examined the set of 11 763 clones that were localized in the sequence by both BES hits and by *in silico* mapping. Genomic intervals defined strictly by alignments of the clone validation fingerprints to the *in silico* fingerprints derived from the genome sequence were generally accurate, with 11 620/11 763 (~99%) of the *in silico* mappings overlapping with BES intervals. *In silico*-calculated clone spans averaged 7 kb less than the lengths of the corresponding intervals for the same clones as defined by BES alignments to the genome sequence. This difference is due to the conservative nature of the *in silico* localization algorithm (Materials and Methods), which tends to yield coordinates which are subsumed by the interval formed by the BES hits. The median differences in left end, middle and right end alignments to the genome and the 90 percentile ranges, comparing the *in silico* and BES coordinates, were 1.3 (−11 to 9) kb, 0.1 (−11 to 9) and –1.8 (−20 to 8) kb. Negative values indicate that the BES coordinate was to the left of the *in silico* coordinate. Distributions of these differences across the clone set are shown in

Figure 1. On average, 98% of a clone's *in silico* coordinate interval overlapped with the BES coordinates, and on average 95% of the BES coordinate overlapped with the *in silico* interval. In only 35/11 620 cases did the *in silico* coordinate interval overlap by less than half of its extent with the BES coordinate. The average proportion of sequence-confirmed restriction fragments for these cases was 0.85, compared to 0.96 for the 11 620 clones in the test set.

In only 143 cases was the BES coordinate located outside of the coordinate specified by *in silico* mapping. In 35 of these cases, the coordinates were mapped to the same chromosome, with an average distance between the closest ends of the two coordinates of 85 kb. In the remaining 108 cases, the BES and *in silico* mappings were assigned to different chromosomes. We speculate that these inconsistencies are due to ambiguous coordinate assignment in regions containing repeats or sequence assembly gaps, and possibly also to laboratory tracking errors causing misidentification of the BES records.

In total, either BES, *in silico* or assembly coordinates exist for 31 686 clones in the set. For the 1169 remaining clones precise sequence coordinates could not be unambiguously determined. These clones had no BES data, did not contribute to the sequence assembly and had fingerprints which failed to significantly match the assembly region expected to contain the clone. We speculate that localization for these clones failed due to un-sequenced or unfinished genomic regions.

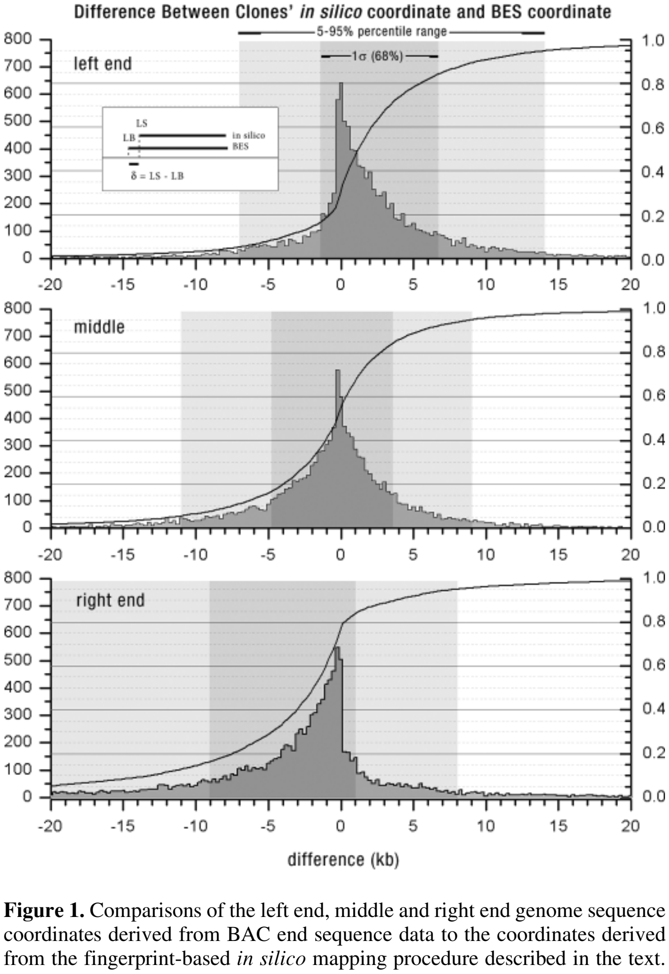

Figure 1. Comparisons of the left end, middle and right end genome sequence coordinates derived from BAC end sequence data to the coordinates derived from the fingerprint-based *in silico* mapping procedure described in the text. Shown are distributions depicting the frequency of occurrence of differences in the coordinates. In general, a narrower distribution corresponds to better correlation between coordinates derived from the BES data and the *in silico* approach. Shaded regions indicate the proportion of the comparisons falling into indicated intervals. The X-axis scale is in kilobases; the Y-axis depicts number of comparisons. The line indicates the cumulative distribution of the differences in localization.

Clone set representation of the sequence assembly

Coverage of the human genome sequence by the clone set was assessed using the clone sequence coordinates derived as described above. BES coordinates were used where available, and in other cases clone assembly coordinates (21) were preferred over fingerprint-based placement.

Gaps and undetermined portions in the sequence assembly were not used in the assessment of sequence coverage of the clone set. Such regions currently represent an estimated 0.227 Gb (7%) out of a total assembly size of 3.070 Gb. Using the 31 686 clones with sequence coordinates, we determined that 2.827 Gb, corresponding to >99.4% of the assembled sequence (July 2003), is represented by the clone set (Figure 2). Clones not found in the BAC fingerprint map, selected to fill in gaps >10 kb, contribute 18 Mb towards coverage of the genome sequence. We estimate therefore that the fingerprint map itself covers ~2.809 Gb, or at least 99% of the current sequence assembly. This is more than the previously predicted coverage of 96%, computed using analysis of chromosomes 21 and 22 (4), and indicates that, at least in the case of human, the fingerprint map provides near-complete coverage of the sequenced portion of the genome. For other mammalian genomes (i.e. rat and mouse) it will be interesting to repeat this coverage assessment as the genome sequence assemblies mature.

Excluding regions of the assembly for which sequence information is known not to be available (e.g. telomeres, centromeres, internal gaps), our analysis reveals the existence in the clone set of 526 detectable gaps in sequence coverage, totalling 13.7 Mb. The gaps arise largely from regions for which coverage cannot be achieved by using RPCI-11, RPCI-13 and Caltech-D clones. Of the gaps, 210 are smaller than 10 kb, and these total 0.7 Mb (Figure 3). Given that clones from a number of libraries (www.ncbi.nlm.nih.gov/genome/clone) have been sequenced to generate the human genome sequence assembly, it is not surprising that the largest gaps are found in genome regions not represented by clones in the RPCI-11, RPCI-13 or Caltech-D libraries. It is possible to address these gaps by including individual clones from other libraries in future versions of the clone set. Our estimation of gaps likely represents an overestimate of the actual gaps in the clone set because sequence assembly coordinates based on *in silico* mappings are typically smaller than the actual clone insert size.

Clone set representation of the BAC fingerprint map

Although the calculation of map coverage appears subordinate to that of assembled sequence coverage, assessment of map coverage is relevant because some regions of the fingerprint map may not be represented in the sequence assembly. However, a precise assessment of fingerprint map coverage by the clone set is more difficult to determine automatically than such an assessment of sequence coverage. This is because the

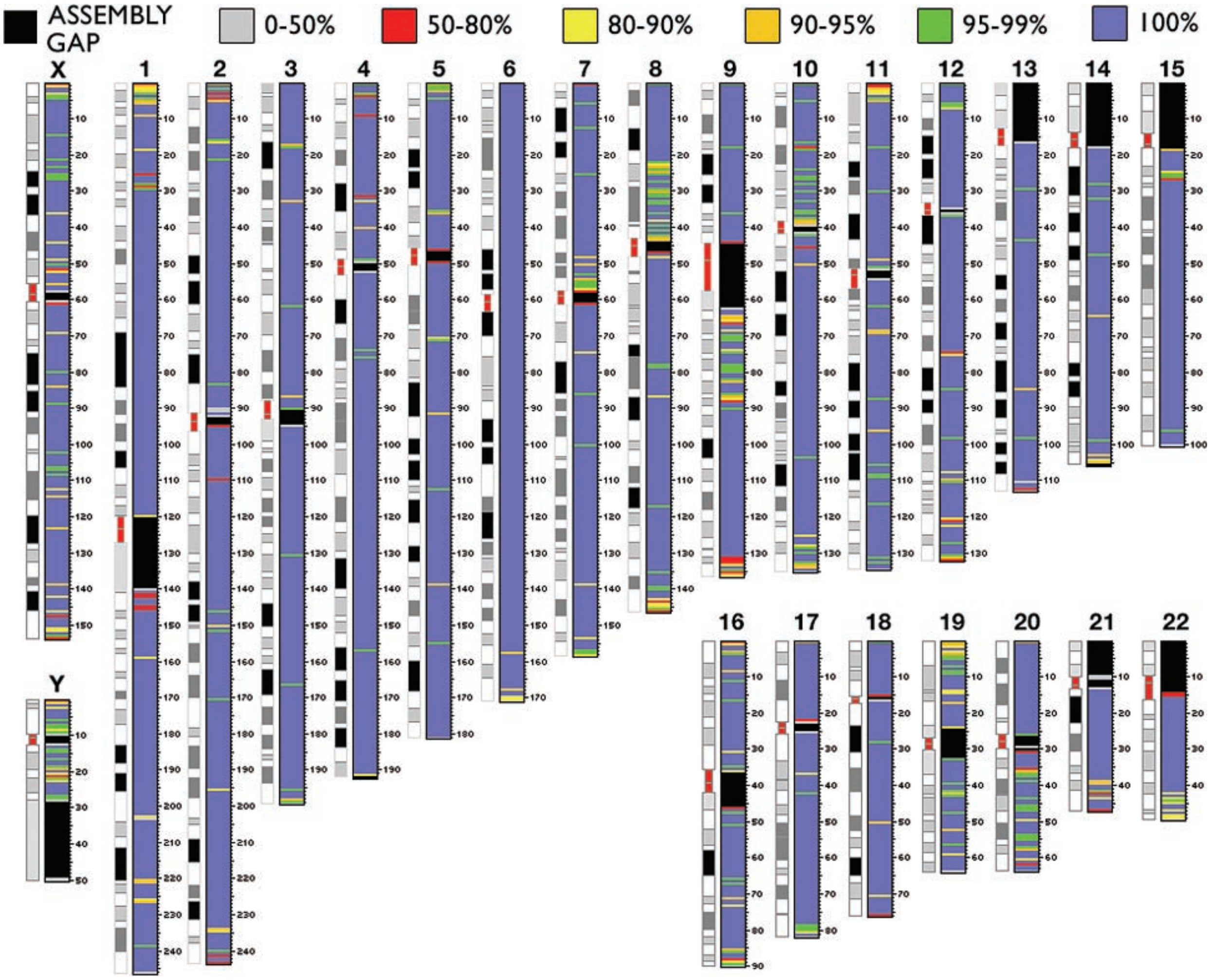

Figure 2. Coverage of the sequence assembly provided by the clones in the set. For each chromosome, the coverage of adjacent 700 kb regions is plotted as a colour map. Regions in the assembly without sequence information appear as black areas. Bright blue regions correspond to 100% coverage. Distance scale is in megabases.

fingerprint map scale metric, measured in cbmap units, is not locally linear with the sequence scale and because not all restriction fragments are detected by the fingerprinting method (2,3). An additional confounding factor is that fingerprints in the map were processed to remove multiple co-migrating fragments (4). We evaluated coverage of the fingerprint map in two ways. First, we examined the coverage of the map provided by the clones in the set in terms of the representation of contigs as well as of the cbmap unit scale. Second, we evaluated the overlap in the fingerprint map between all adjacent clones in the set using the validation fingerprint data. Briefly, all fingerprint map contigs larger than 2 clones and containing clones from RPCI-11, RPCI-13 or Caltech-D are represented in the clone set. The set provides coverage of 98% of the cbmap units, with a median cbmap gap of 6. For 1.5% of adjacent set clones, overlap by fingerprint could not be unambiguously inferred. Refer to the Supplementary Material for additional details.

Resolving power and other properties of the clone set

To determine the spatial resolving power of the clone set, we computed all unique intersections between clones in the set using their sequence coordinates. This process can be visualized as locating both ends of each clone on the sequence and evaluating the distances between closest end positions. This distance between adjacent ends, which we call a 'clone cover', defines the smallest resolvable region. The average size of clone covers is 47 kb. The effective resolving power of the set is 79 kb (Figures 4 and 5), and was calculated using a weighted average of the clone cover size, where the weights are given by the fraction of the sequence represented in covers of a given size. This figure is a reasonable approximation of the theoretical resolution achievable with this set in array-CGH experiments [see e.g. (9,22,23)].

Using our analysis of clone covers, we were able to derive other useful statistics for the clone set. The median sequence overlap between neighbouring clones is 83 kb, which corresponds to 50% of the length of the clones, with 90% of neighbouring overlaps in the range of 14–160 kb. The average coverage depth of the clone set, based on sequence coordinates is 1.9X (Figure 6), which substantiates the calculation we performed using cbmap units as described above. The ratio of 1- to 2-fold coverage of the genome is approximately 1:1, with 1.02 Gb of the sequence assembly spanned by a single clone and 1.16 Gb spanned by two overlapping clones.

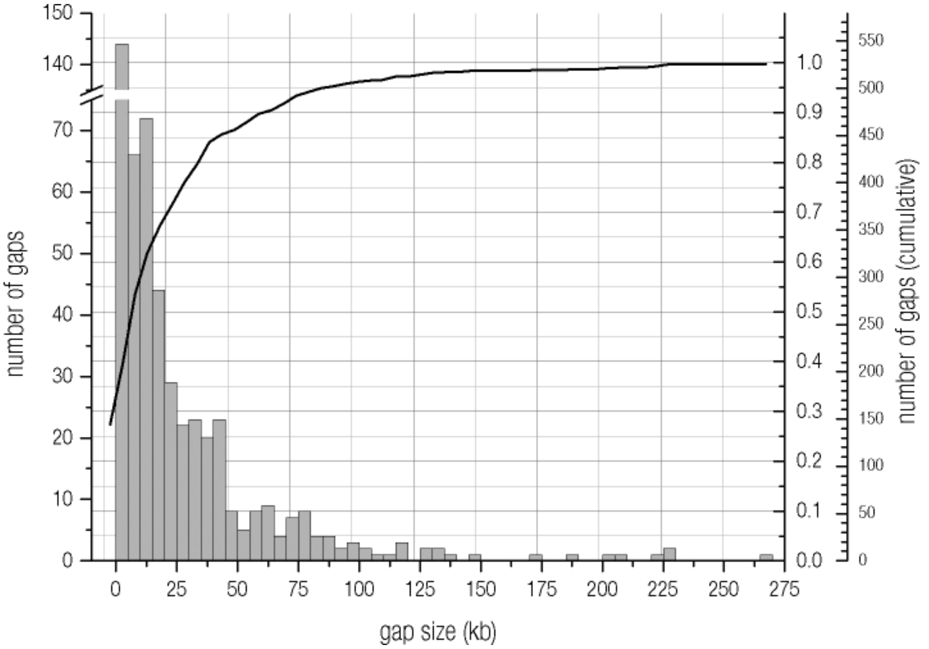

Figure 3. Distribution of gaps in sequence coverage. Location and sizes of gaps is determined by using all available sequence coordinates for clones in the set. Of the gaps, 40% are <10 kb. Many gaps in this distribution may not be real, but instead result from our conservative algorithm for *in silico* sequence coordinate determination. There are 1169 BACs without sequence coordinates and therefore without explicit localization in the genome. When the genome sequence is complete it is possible that many of these BACs will be localized on the sequence.

Regions spanned at 3-fold coverage total 0.42 Gb, and deep coverage of at least 4-fold spans 0.19 Gb of the genome (Figure 6). Coverage at high depth in our clone set typically occurs in regions where additional clones were added to the set to replace clones that failed validation, as it was not always possible to find a single clone providing equivalent coverage during the replacement process.

The precise determination of telomere representation in the clone set is complicated by the difficulties in isolating and sequencing clones derived from them. We attempted to measure the representation of telomeric regions in the clone set using 210 unique BAC telomeric markers derived from the RPCI-11/13 and Caltech-D libraries by the Human Telomere Sequencing and Mapping Project (24). Out of the 46 different telomeres (X and Y telomeres are considered identical), these 210 BACs mark unique sequences near 40 telomeres. Telomeres 13p, 14p, 15p, 21p and 21q are not represented in this set. Our clone set contains 64 of these 210 telomeric markers, 9 of which do not have explicit sequence coordinates. Of the remaining 146 markers, 116 are in the fingerprint map and these BACs overlap (by analysis of fingerprint data) with the best match in the clone set by an average of 110 kb (e.g. 102 of the 116 overlap by >83 kb). The 64 markers included directly in our set and the clones sharing substantial overlap with the 102 telomere BAC markers provide good representation of the content of the BAC telomeric markers. The markers of the 6q telomere are not represented. We anticipate that by creating a clone set which represents full coverage of the fingerprint map (as opposed to only those regions of the map represented by sequence data), the issue of telomere and centromere representation is mitigated somewhat, although genomic regions un-clonable in BACs, or not present in the map, or represented only in difficult to obtain and distribute libraries will not be represented in this version of the clone set.

DISCUSSION

We have created and made available a clone set which represents at least 99% of the current human genome sequence assembly. The clone set contains 32 855 clones, with 32 432 already rearrayed from the RPCI-11, RPCI-13 and Caltech-D libraries and the additional 441, comprising the Version 1.1 clone set update, soon to be made available. The clone set is portable and affordable, and offers interested researchers the opportunity to acquire a reasonably complete representation of the human genome in a modest number (93) of 384-well plates. We envision various uses of the clone set, including the development of FISH probes and for BAC array CGH. Towards this latter goal, we have undertaken a preliminary validation of the clones in the set, measuring their performance in array CGH (10).

A potential issue for some users of the clone set is the extent to which clones in the set could be related to the genome sequence. Early on during the construction of the clone set, when the sequence assembly was still evolving, we considered using only clones sequenced by the Human Genome Sequencing Project. However, we found that clones populating sequencing queues were frequently not optimal for various reasons. These included clones of sub-optimal length and clones represented in libraries not readily available to us and others. The frequently changing nature of the clones listed in sequencing queues was likewise confounding. And obviously, if restricted to only sequenced clones, we would be unable to sample regions of the genome un-represented in sequenced clones,

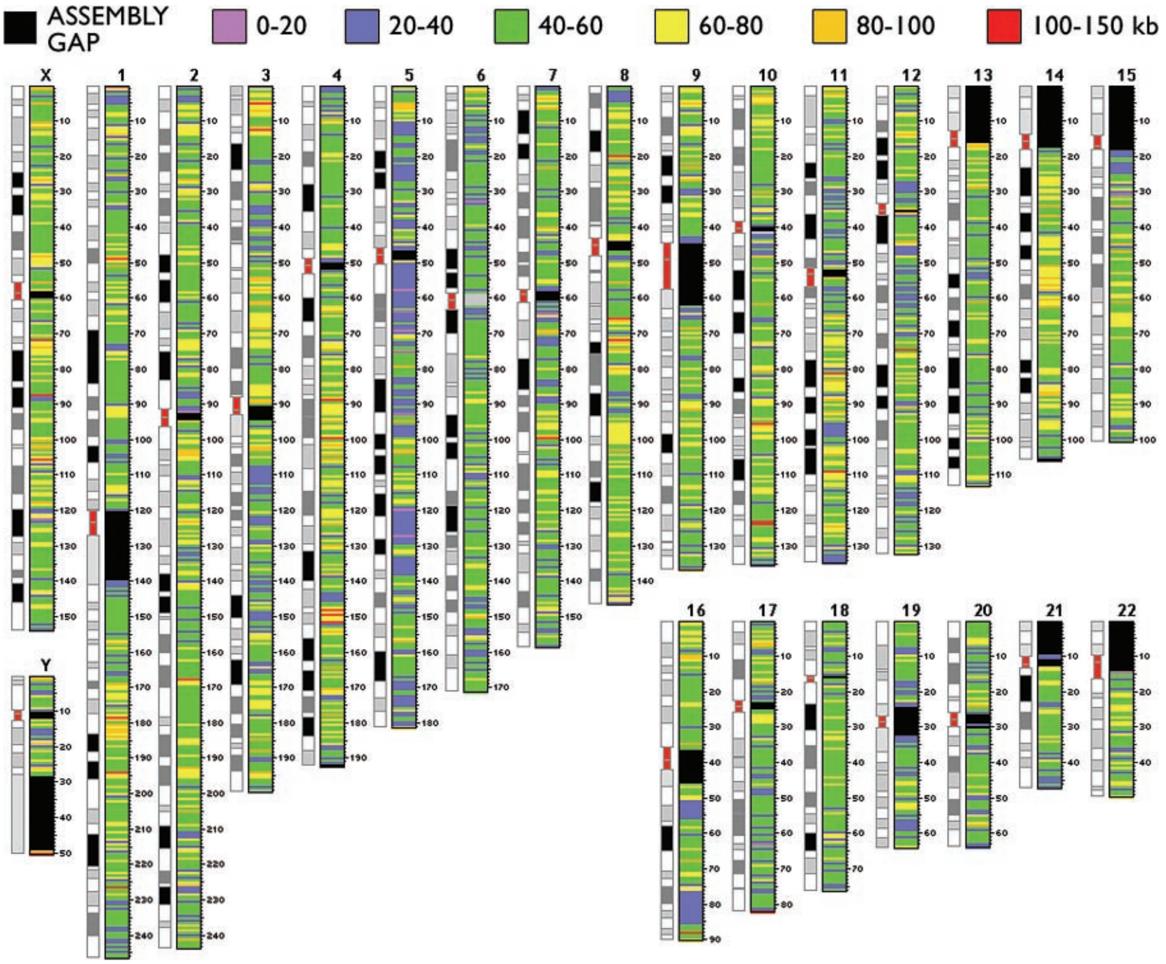

Figure 4. Resolution provided by the clones in the set. For each chromosome, the average clone cover size is coded by colour. There are a total of 61 656 clone covers with an average cover size of 47 kb and an effective resolution of 79 kb. Regions in the assembly without sequence information appear as black areas. Distance scale is in megabases.

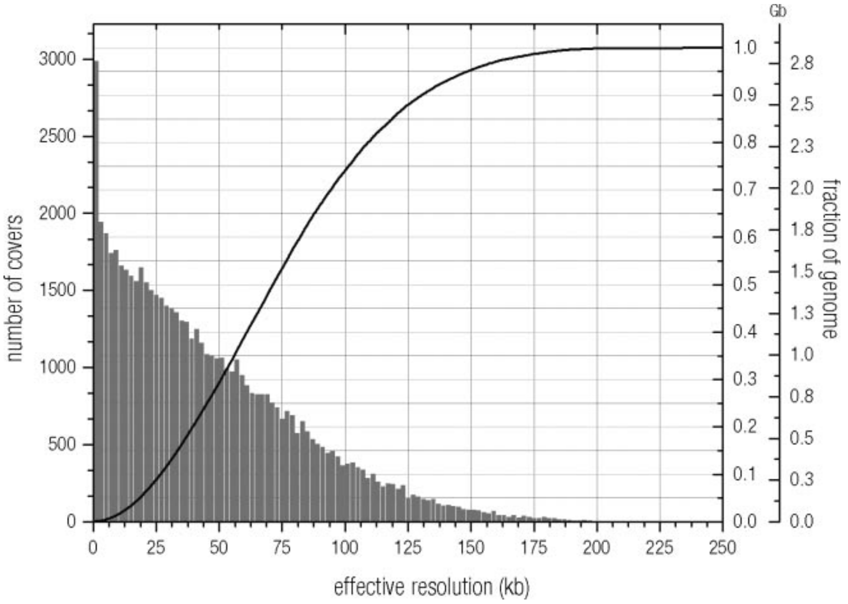

Figure 5. Distribution of clone cover sizes and cumulative distribution of the effective resolution of our clone set. The average effective resolution is 79 kb. This value is obtained by averaging the cover sizes at randomly sampled points of the sequence assembly. 95% of the genome can be resolved by the clone set at a level of 150 kb, or better.

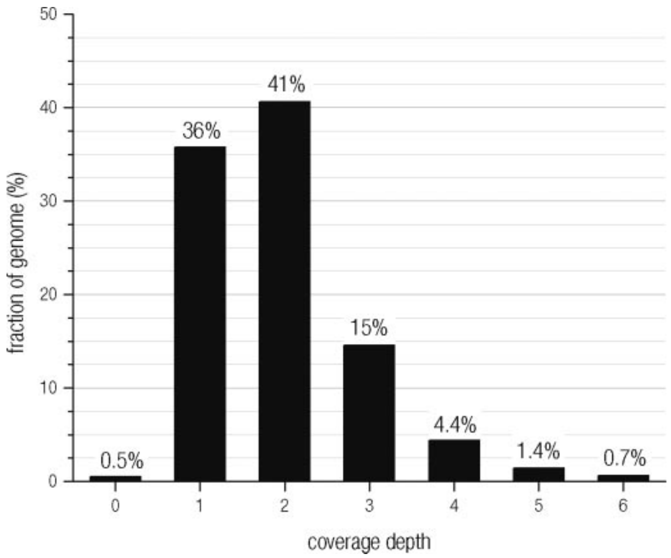

Figure 6. Distribution of coverage depth of the clone set. The depth is represented by the number of clone set BACs with sequence coordinates that span a given region of sequence.

yet represented in the BAC fingerprint map. Finally, the localization of the ends of even sequenced clones is non-trivial in cases where the full clone insert has not been finished and where useful BAC end sequences are not available. For all of these reasons, we felt that a general method to associate clones in the set to the genome sequence, even in the absence of sequence data from the clones, was necessary. The main challenge was in demonstrating that our fingerprinting procedure yielded restriction fragment size information that could be compared directly to *in silico* fingerprints derived from the sequence, yielding reliable estimates of clone coordinates and span on the sequence. This in turn was aided by the fact that half of the clones in the set could be directly related to the sequence through either informative BAC end sequences or the availability of other sequence information. Once we were satisfied that we could localize clones accurately on the sequence assembly using a combination of sequence and fingerprint data, we enjoyed substantial flexibility in the selection of clones.

Version 1 of the clone set has been available through the BACPAC Resources Center (bacpac.chori.org) since February 2003. Since that time, 18 requests for the entire clone set have been received. Derivatives of the clone set have also been produced (www.chori.org/bacpac/pHumanMinSet.htm), often to represent specific chromosomes or chromosome arms, and four requests for these have been received. Hence, the resource is of interest, and we are constructing similar resources for other organisms of major importance to biomedical research. As we gain experience with the performance of clones in the human set, either directly or through the feedback of others, it is our intention to upgrade the clone set, producing an enhanced Version 2. Clone annotations (including clone names, chromosomal and sequence position assignments, clone covers, plate locations and fingerprint data), views of the fingerprint map complete with the identities and positions of selected clones and a UCSC Genome Browser Track are all available on our web site at mkweb.bcgsc.ca/bacarray.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We acknowledge useful discussions involving Dr Calum MacAulay and Dr Dan Pinkel, who encouraged us to move forward with a clone set spanning the human genome. M.A.M. and S.J. are scholars of the Michael Smith Foundation for Health Research. The authors wish to acknowledge support from the BC Cancer Foundation, Genome British Columbia, Genome Canada and the National Human Genome Research Institute (USA). The work presented here was funded by a Genome British Columbia Grant (Cancer Genomics) to Victor Ling, Connie Eaves and M.A.M.

REFERENCES

1. Shizuya,H., Birren,B., Kim,U.J., Mancino,V., Slepak,T., Tachiiri,Y. and Simon,M. (1992) 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.
2. Marra,M.A., Kucaba,T.A., Dietrich,N.L., Green,E.D., Brownstein,B., Wilson,R.K., McDonald,K.M., Hillier,L.W., McPherson,J.D. and Waterston,R.H. (1997) High throughput fingerprint analysis of large-insert clones. *Genome Res.*, **7**, 1072–1084.
3. Schein,J., Kucaba,T.A., Sekhon,M., Smailus,D., Waterston,R.H. and Marra,M.A. (2004) High-throughput BAC fingerprinting. In Zhao,S. and Stodolsky,M. (eds), *Methods in Molecular Biology. Vol. 255: Bacterial Artificial Chromosomes: Library Construction, Physical Mapping and Sequencing*. Humana Press Inc., Totowa, NJ, Vol. 1, pp. 143–156.
4. McPherson,J.D., Marra,M., Hillier,L., Waterston,R.H., Chinwalla,A., Wallis,J., Sekhon,M., Wylie,K., Mardis,E.R., Wilson,R.K. *et al*. (2001) A physical map of the human genome. *Nature*, **409**, 934–941.
5. Gregory,S.G., Sekhon,M., Schein,J., Zhao,S., Osoegawa,K., Scott,C.E., Evans,R.S., Burridge,P.W., Cox,T.V., Fox,C.A. *et al*. (2002) A physical map of the mouse genome. *Nature*, **418**, 743–750.
6. Krzywinski,M., Wallis,J., Gosele,C., Bosdet,I., Chiu,R., Graves,T., Hummel,O., Layman,D., Mathewson,C., Wye,N. *et al*. (2004) Integrated and sequence-ordered BAC- and YAC-based physical maps for the rat genome. *Genome Res.*, **14**, 766–779.
7. Osoegawa,K., Mammoser,A.G., Wu,C., Frengen,E., Zeng,C., Catanese,J.J. and de Jong,P.J. (2001) A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.*, **11**, 483–496.
8. Solinas-Toldo,S., Lampel,S., Stilgenbauer,S., Nickolenko,J., Benner,A., Dohner,H., Cremer,T. and Lichter,P. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.
9. Pinkel,D., Segraves,R., Sudar,D., Clark,S., Poole,I., Kowbel,D., Collins,C., Kuo,W.L., Chen,C., Zhai,Y. *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet.*, **20**, 207–211.
10. Ishkanian,A.S., Malloff,C.A., Watson,S.K., DeLeeuw,R.J., Chi,B., Coe,B.P., Snijders,A., Albertson,D.G., Pinkel,D., Marra,M.A. *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genet.*, **36**, 299–303.
11. Soderlund,C., Humphray,S., Dunham,A. and French,L. (2000) Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res.*, **10**, 1772–1787.
12. Soderlund,C., Longden,I. and Mott,R. (1997) FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.*, **13**, 523–535.
13. Zhao,S. (2000) Human BAC ends. *Nucleic Acids Res.*, **28**, 129–132.
14. Zhao,S., Malek,J., Mahairas,G., Fu,L., Nierman,W., Venter,J.C. and Adams,M.D. (2000) Human BAC ends quality assessment and sequence analyses. *Genomics*, **63**, 321–332.

15. Cheung,V.G., Nowak,N., Jang,W., Kirsch,I.R., Zhao,S., Chen,X.N., Furey,T.S., Kim,U.J., Kuo,W.L., Olivier,M. *et al*. (2001) Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature*, **409**, 953–958.
16. Strausberg,R.L., Buetow,K.H., Greenhut,S.F., Grouse,L.H. and Schaefer,C.F. (2002) The cancer genome anatomy project: online resources to reveal the molecular signatures of cancer. *Cancer Invest*., **20**, 1038–1050.
17. Fuhrmann,D.R., Krzywinski,M.I., Chiu,R., Saeedi,P., Schein,J.E., Bosdet,I.E., Chinwalla,A., Hillier,L.W., Waterston,R.H., McPherson,J.D. *et al*. (2003) Software for automated analysis of DNA fingerprinting gels. *Genome Res*., **13**, 940–953.
18. Sulston,J., Mallett,F., Staden,R., Durbin,R., Horsnell,T. and Coulson,A. (1988) Software for genome mapping by fingerprinting techniques. *Comput. Appl. Biosci*., **4**, 125–132.
19. The International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
20. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res*., **30**, 17–20.
21. Kent,W.J. and Haussler,D. (2001) Assembly of the working draft of the human genome with GigAssembler. *Genome Res*., **11**, 1541–1548.
22. Snijders,A.M., Nowak,N., Segraves,R., Blackwood,S., Brown,N., Conroy,J., Hamilton,G., Hindle,A.K., Huey,B., Kimura,K. *et al*. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genet*., **29**, 263–264.
23. Fiegler,H., Carr,P., Douglas,E.J., Burford,D.C., Hunt,S., Smith,J., Vetrie,D., Gorman,P., Tomlinson,I.P. and Carter,N.P. (2003) DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer*, **36**, 361–374.
24. Knight,S.J., Lese,C.M., Precht,K.S., Kuc,J., Ning,Y., Lucas,S., Regan,R., Brenan,M., Nicod,A., Lawrie,N.M. *et al*. (2000) An optimized set of human telomere clones for studying telomere integrity and architecture. *Am. J. Hum. Genet*., **67**, 320–332.