A Set of Rearrayed BAC Clones
spanning the human genome

Krzywinski M, Bosdet I, Smailus D, Mathewson C, Wye N, Barber S, Brown-John M
Chand S, Cloutier A, Masson A, Mayo M, Olson T, Lam W, MacAuley C
Osoegawa K†, Zhao S‡, de Jong PJ†, Schein J, Jones S, Marra M

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

Genome Sciences Centre

www.bcgsc.ca
info@bcgsc.ca

Libraries

RPCI-11

30,250 clones

189kb/46 fragments

RPCI-13

2,100 clones

138kb/29 fragments

Caltech-D

600 clones

146kb/35 fragments

Clones

32,850 BACs

30,500 had sequence coordinates based on fingerprint anchoring

10,500 have unique double-ended coordinates based on end sequence

5,000 are in the June 2002 UCSC Golden Path assembly

Coverage

> total assembly size

3.042 Gb

> assembly contig gaps

0.230 Gb (8%)

> assembled sequence

2.812 Gb (92%)

> clone set coverage

2.797 Gb (99.5%)

> coverage by clones with BES, fingerprint or Golden Path coordinates

2.764 Gb (98.8%)

> coverage infered by map topology

0.032 Gb (1.1%)

> coverage by clones not in the physical map

0.052 Gb (1.9%)

> gaps in clone set coverage

0.015 Gb (0.5%)

1. Abstract

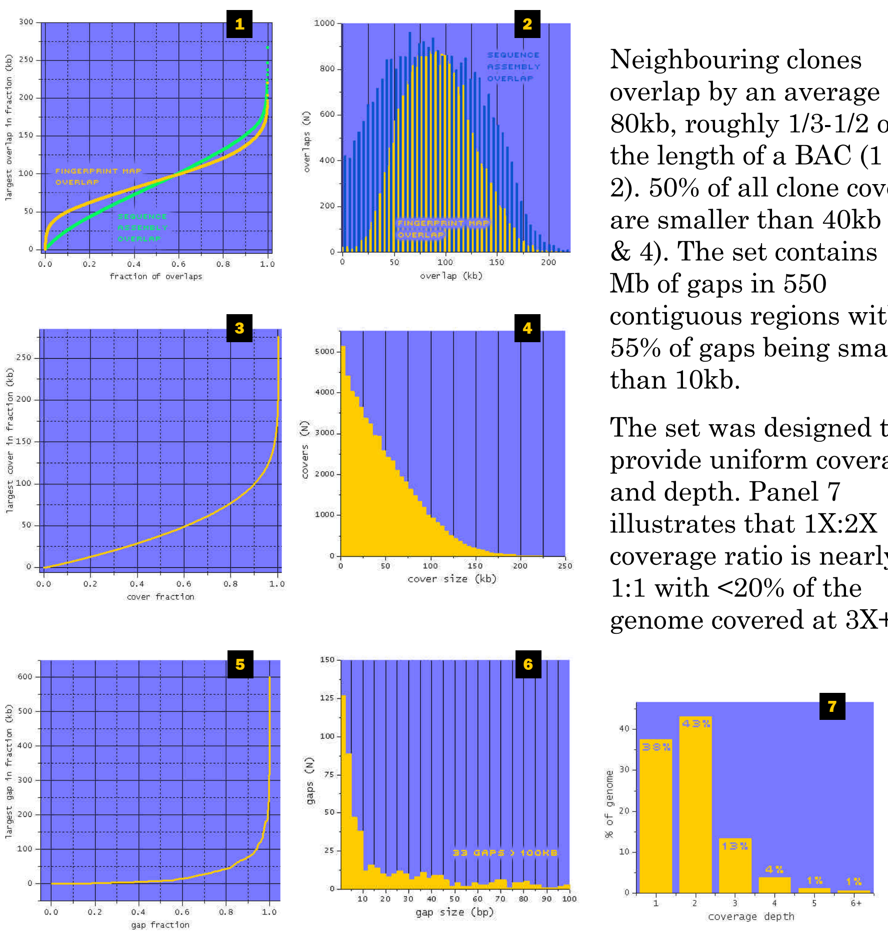
From the human fingerprint map constructed at Washington University Genome Sequencing Center, we have selected a set of 32,850 BACs that span the human genome. The purpose of the clone set is to serve as a genome-ordered set of probes for FISH and microarray-based BAC CGH experiments. The comprehensive coverage of this clone set makes it a valuable asset in both research and clinical contexts, in the search for understanding and detection of cancer-related chromosomal and expression alterations.

The clones have been sampled from RPCI-11/13 (95%) and Caltech-D (5%) libraries, selected to optimize size, coverage of the map and consistent overlap, and have been rearrayed into 384-well format. The identity of clones has been validated by fingerprinting. Following the first round selection of 29,000 clones, a combination of automated and visual fingerprint inspection identified 2,000 clones that did not match the fingerprints stored in the physical map. 4,500 clones were added to the set to maximally conserve map coverage of the unmatched clones. Analysis of the set's sequence coverage (UCSC, 2002/06 assembly) resulted in the selection of an additional 1,300 clones to cover gaps larger than 10 kb.

The clone set covers 99.5% of the November 2001 version of the BAC fingerprint map. Using fingerprint-based localization, end sequence data and assembly coordinate data, the set was found to cover 2.80 Gb (99.5%) of the assembled sequence, with 50 Mb of assembly coverage provided by clones not found in the fingerprint map. Approximately 80% of the assembly is covered at 1X and 2X in a 1:1 ratio. The sequence coverage of the set contains 550 sequence coverage gaps totaling 15Mb, with 55% of the gaps being smaller than 10kb.

This first version of the clone set will be publicly distributed through Pieter de Jong. A web-based clone search and data mining portal will be available. We anticipate that the set will evolve as new versions of the sequence assembly and physical map are released. We are planning to create analogous resources for mouse and rat.

3. Coverage and Gap Characterization



Neighbouring clones overlap by an average of 80kb, roughly 1/3-1/2 of the length of a BAC (1 & 2). 50% of all clone covers are smaller than 40kb (3 & 4). The set contains 15 Mb of gaps in 550 contiguous regions with 55% of gaps being smaller than 10kb.

The set was designed to provide uniform coverage and depth. Panel 7 illustrates that 1X:2X coverage ratio is nearly 1:1 with <20% of the genome covered at 3X+.

Figure 2 Characterization of clone set neighbour overlap (1 & 2), distinct cover size and distribution (3 & 4) and coverage gap distribution (5 & 6). Proportion of coverage depth is shown in panel 7. All quantities are calculated based on sequence coordinates.

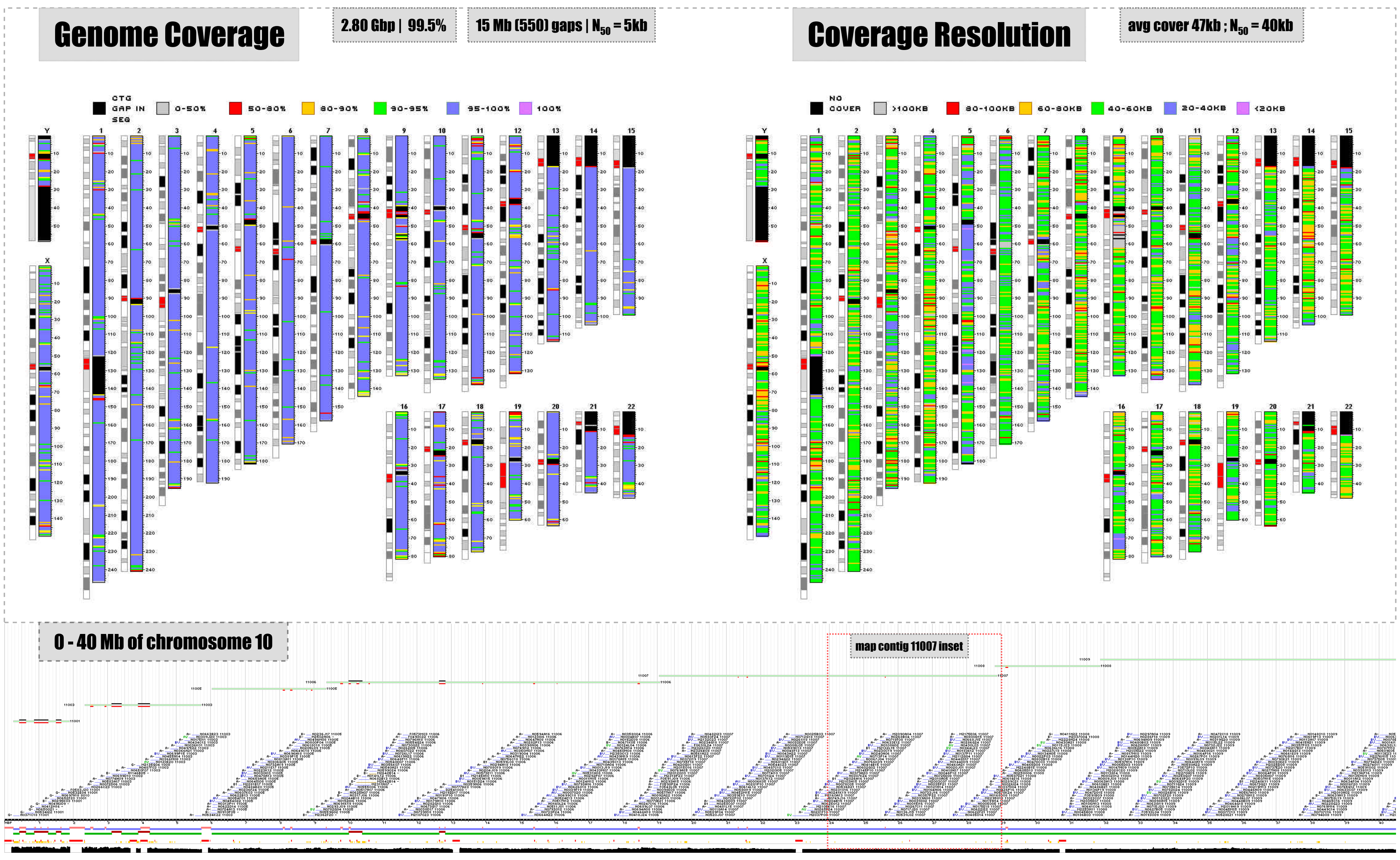
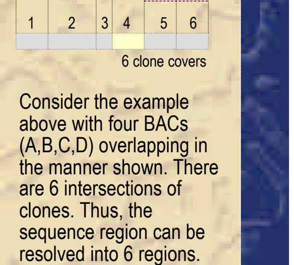


Figure 1 *top left* | representation of genomic coverage in 700kb sequence windows in the clone set, as determined by sequence coordinates derived from BES, fingerprint and assembly membership anchors (scale in Mb; cytogenetic coordinates slightly differerent than assembly coordinates; contig gaps in assembly at end of chromosomes are not shown)| *top right* | coverage resolution calculated by the average size of unique clone covers in 700kb sequence windows | *bottom* | enlarged portion of chromosome 10 (0-40Mb) showing a sequence-coordinate view of physical map contigs (top), individual BAC clones (middle) and coverage indicators (bottom). Portion of contig 11007 is shown in Figure 3.

Resolution

The resolution of the set was determined by using the concept of **clone covers**. The set of covers is found by intersecting the cover of every clone with those of all its neighbours. Any base pair location will be covered by a group of clones. The **cover** is the largest contiguous sequence region covered by the same group of clones.



Consider the example above with four BACs (A,B,C,D) overlapping in the manner shown. There are 6 intersections of clones. Thus, the sequence region can be resolved into 6 regions. For example, if BACs B and C show positive hybridization in an experiment, the probe can be localized to the fourth cover. The smaller the average size of the cover, the higher the effective resolution of a clone set. Figure 1 (top left) shows the distribution of the average cover size across the genome. Figure 2 (3 & 4) shows the distribution of cover sizes.

2. Validation and Localization

All clones in the set have been fingerprinted using a high-resolution agarose electrophoresis method. The experimental fingerprints were compared to those in the human physical map database to verify the identity of the clone. Fingerprints of clones which were considered weak matches (5,200) were examined visually and 2,000 BACs were determined to require replacement (yellow in Figure 1 left panel).

End Sequence and Fingerprint Anchors

Localization of clones within the genomic assembly was performed using end sequence coordinates (where available), fingerprint-based localization (done for all clones) and assembly coordinates for those clones which contributed sequence to the Golden Path. End sequence coordinates were used in preference to fingerprint anchors, which were prefered over assembly coordinates.

Approximately 30,500 (92%) of clones were localized in the genome using their fingerprints. In this method, the neighbourhood in which the clone was expected to be found was determined on the basis of the clone's map neighbours which already were anchored to the genome with end sequence or assembly coordinates. The clone was localized within the neighbourhood using a series of progressively smaller sample windows, in which an in-silico digest of the sampled sequence was compared to the clone's fingerprint to determine location.

The fingerprint anchoring was evaluated using 9,500 clones for which both anchors and end sequence coordinates existed. 90% of anchors were within 2 HindIII fragments of the end sequence coordinates. 80% of anchors were within 5 HindIII fragments. The average overlap between the anchor and end sequence coordinates was 93%.

FISH & Telomere and Centromere Representation

Existing FISH data from CGAP provides hybridization location for 1,175 of the clones in the set with 103 hybridizations to centromeric bands (p11(.1)/q11(.1)) and 240 hybridizations to telomeric bands (first/last). A portion of the FISHed clones hybridize to multiple regions, some spanning multiple chromosomes.

There are 164 BACs associated with telomeres in the CGAP repository. Out of these, 53 are in the set with the remaining 111 overlapping by an average 100 kb with the best clone set match.

When additional clone markers for important regions which are found to be under-represented in our clone set are identified, they will be added to improve coverage.

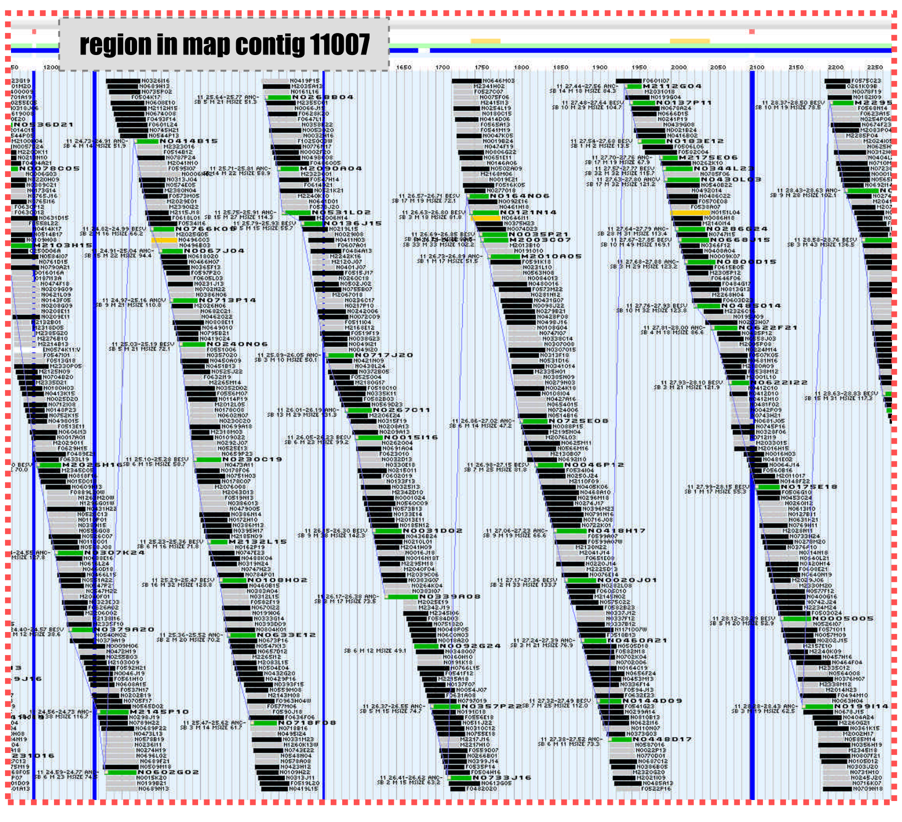


Figure 3 Map-based view of clones in a region of contig 11007 highlighted in Figure 1. This figure shows clone map-based clone positions with clones coloured according to the following scheme: set BACs (green), replaced BACs (yellow) and canonical (black) and buried (grey) map clones. Sequence overlap with the next clone in the set is shown to the left of the clone.

Acknowledgments

Clone libraries | RPCI-11/13: www.chori.org/bacpac; CalTechD http://www.tree.caltech.edu/ | *BAC physical map* | Washington University Genome Sequencing Centre www.genome.wustl.edu | *Golden path* | International Human Genome Sequencing Consortium www.ncbi.nlm.nih.gov/genome/seq; GigAssembler genome.ucsc.edu | *BAC end database* | TIGR www.tigr.org | *Human Telomere Mapping and Sequencing Project* | www.wistar.upenn.edu/Riethman

Affiliations

†Children's Hospital Oakland Research Institute, Oakland, CA, USA

‡The Institute for Genomic Research, Rockville, MD, USA

Funding

Genome Canada | Cancer Genomics: V Ling, C Eaves, M Marra