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

Canada's Michael Smith
Genome Sciences Centre
www.bcgsc.ca

Libraries

RPCI-11

29,805 clones in set (92%)

189kb/46 fragments

RPCI-13

569 clones in set (2%)

138kb/29 fragments

Caltech-D

2,059 clones in set (6%)

146kb/35 fragments

Clones

32,433 BACs in set

30,561 have sequence coordinates

29,539 have sequence coordinates based on fingerprint anchoring

10,213 have unique double-ended coordinates based on end sequence

8,702 have sequence accessions

8,018 have Genbank records (Jan 2003)

- 4,967 finished
- 2,069 working draft
- 365 in-progress
- 569 low-pass

Coverage

3.042 Gb sequence assembly size

0.230 Gb (8%) assembly contig gaps

2.812 Gb (92%) assembled sequence

Sequence coverage determined using only BACs with sequence coordinates (94%)

2.789 Gb (99.2%) assembled sequence covered

0.024 Gb (0.8%) assembled sequence not represented by BACs with coordinates

Resolution

Depth of assembled sequence coverage by clones in the set:

1.077 Gb @ 1X

1.141 Gb @ 2X

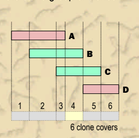
0.350 Gb @ 3X

0.151 Gb @ 4X+

57,876 unique clone covers in assembled sequence regions

47.0 kb average clone cover size (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.

1. Abstract

From the human fingerprint map constructed at Washington University Genome Sequencing Center, we have selected a set of 32,433 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 (94%) and Caltech-D (6%) libraries, selected to optimize size, coverage of the map and consistent overlap. The clones have been rearrayed into 384-well format. The identity of clones has been validated by fingerprinting. Following the first round selection of 29,035 clones, a combination of automated and visual fingerprint inspection identified 1,978 clones that did not match the fingerprints stored in the fingerprint map. 4,531 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 1,258 clones, with some chosen from outside the fingerprint map, to cover gaps larger than 10 kb. During the second round of fingerprint validation 413 clones were rejected.

The clone set covers 99% of the November 2001 version of the BAC fingerprint map. Using fingerprint-based localization, end sequence data and assembly coordinate data, 30,561 of the clones were localized within the genome and found to cover 2.788 Gb (99%+) of the assembled sequence. Approximately 35 Mb of this coverage was provided by clones not found in the fingerprint map. Approximately 82% of the assembly is covered at 1X and 2X in a 1:1 ratio. The sequence coverage of the set contains 729 sequence coverage gaps totaling 24Mb, with 46% of the gaps being smaller than 10kb. The average resolution of the clone set is 46kb.

This first version of the clone set is publicly distributed through the BACPAC Resources Centre (Children's Hospital, Oakland). We anticipate that the set will evolve as new versions of the sequence assembly and physical map are released. We are planning to create an analogous resource for the mouse and rat genomes.

2. Clone Selection, Validation and Localization

The primary source of BACs during the selection process was the fingerprint map. The BACs were chosen to provide complete coverage of the map and the selection enriched for BACs having end sequence and sequence records. Overlap in the map between neighbouring selections was determined by using the number conserved fragments (fragments found in the neighbouring and all intermedial clones). Clones were considered to overlap if they had >3 conserved fragments.

All the BACs in the set have been **validated by fingerprinting**. The experimental fingerprints were compared to those in the human physical map database to verify the identity of the clone. Coverage provided by any clones whose fingerprints did not match those in the map was duplicated with additional clones. Fingerprints of clones which were considered weak matches (5,272) were examined visually and 1,978 BACs were determined to require replacement (yellow clones in Figure 3). In the second round of validation, in which the first-round replacement clones were fingerprinted, 413 clones failed to match their FPC fingerprints. The coverage provided by these clones will be duplicated in subsequent versions of the set.

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. For 10,213 of clones (approximately 1/3 of our set) two-tailed BAC end sequence (BES) coordinates exist. To determine more precise location of the other clones in the set, an *in silico* approach was used. In this method, the clone's fingerprint was localized to a genomic assembly region using BES coordinates of neighbouring clones. More precise localization was performed by comparing *in silico* digests derived from windowed sequence regions within this neighbourhood. All BACs were subject to this method and fingerprint-based coordinates were found for 29,539 (91%) of the clones (20,076 of these clones had no previous coordinates).

The fingerprint-based localization method was evaluated using 9,463 clones which had both BES coordinates and coordinates derived with the fingerprint-based method. The average difference in left, middle and right coordinate positions was 2±11kb, 8±14kb and -4±14kb. On average, 92±8% of the BES coordinate overlapped with the fingerprint-derived coordinate.

Overall, either BES, fingerprint-based or assembly **sequence coordinates exist for 30,561 clones (94%)**. Coordinates for the remaining 1,872 clones (all belonging to fingerprint map contigs) could not be determined. In these cases the map clones are impossible to anchor because of local variation in sequence assembly quality and state of completion. These clones are included in the set, however, to ensure full coverage of the fingerprint map. Clones with sequence coordinates were used to determine that the clone set covers 2.789 Gb of the assembled sequence (99.2%). This is likely to be an underestimate because we could not unambiguously place 6% of the clones in the set on the assembly.

Existing FISH data from CGAP provides hybridization location for 1,134 of the clones. There are 164 BACs associated with telomeres in the Human Telomere Sequencing and Mapping Project. Out of these, 45 are in the set with the remaining overlapping by an average 100 kb with the best clone set match.

The **resolution of the set** was determined using the concept of clone covers (see side panel). Among the clones with sequence coordinates, there are a total of 57,876 unique clone covers with an average size of 47.0 kb.

Genome Coverage

2.80 Gbp | 99%+ 24 Mb (729) gaps | N_{50} = 12.5 kb

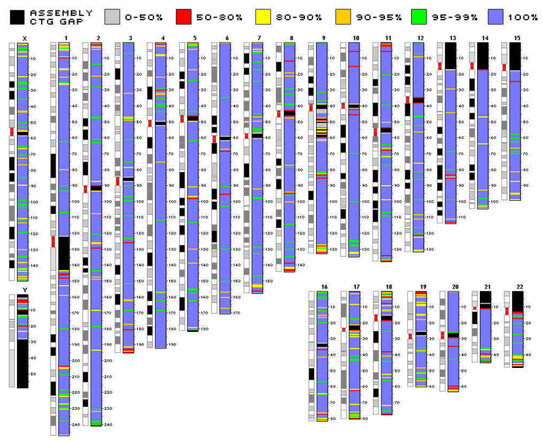


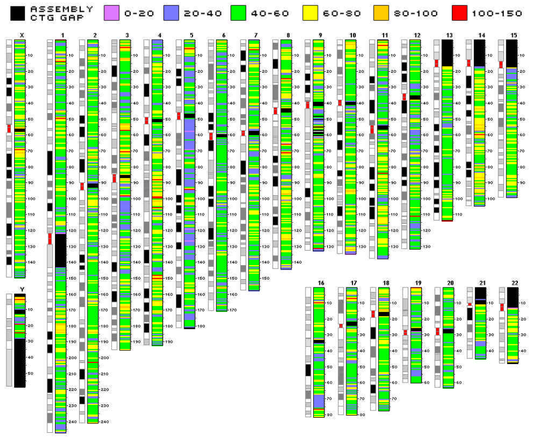
Figure 1

top | Representation of genomic coverage by the clone set calculated in 700kb sequence windows across the genome. Coverage was determined using clones with sequence coordinates derived from BES, fingerprint and assembly anchors. Regions in which no assembly information is available are coloured in black. Regions which are completely represented in the clone set are coloured blue (scale in Mb; cytogenetic bands are shown to the left of each chromosome with centromeres coloured in red)

bottom | Coverage resolution calculated by the average size of unique clone covers in the same 700kb sequence windows as displayed in the coverage diagram in the top panel.

Coverage Resolution

avg cover 46kb ; N_{50} = 40kb



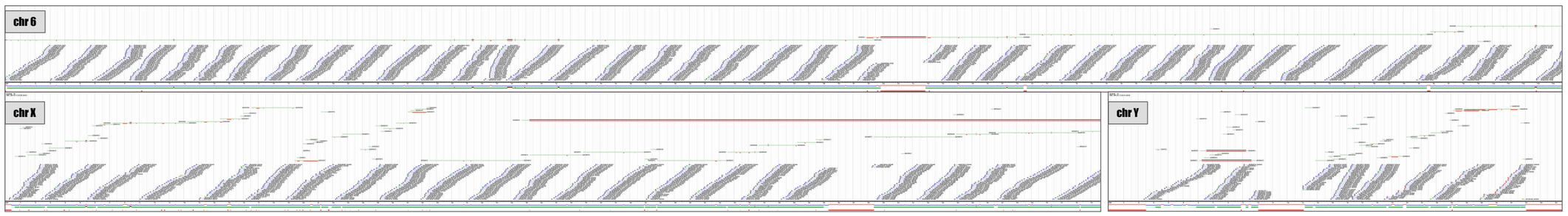


Figure 2

Detailed view of regions of chromosomes 6, X and Y. By using the fingerprint map we have been able to construct a set which provides coverage in regions where the assembly is incomplete. Sequence coordinate type flags shown to the left of the clone are: B are BES coordinates and A are fingerprint-based coordinates. Extent of fingerprint map contigs is shown above the clone stacks. Regions of the fingerprint contig coloured in green indicate coverage by clones with sequence coordinates. Regions shown in yellow represent coverage determined by position in the fingerprint map by clones without coordinates. Regions shown in black indicate an apparent break in the fingerprint map contig.

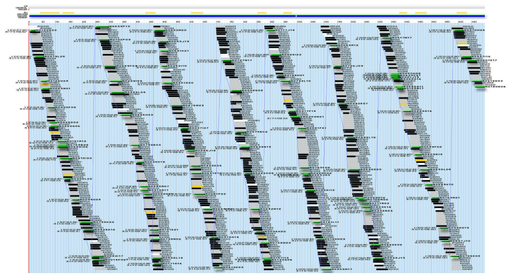
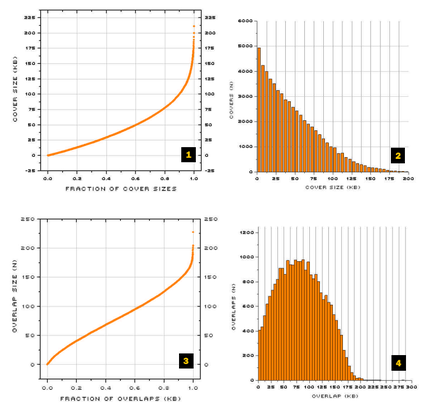


Figure 3

Fingerprint map view of clones in fingerprint map contig 6006. Clones in the set are coloured as follows: set BACs (green), failed during validation (yellow). Other clones in the map are canonical (black) and buried (gray), virtual (white). Sequence overlap between adjacent clones is indicated by a joining blue line with detailed overlap information shown to the left of the clone.

3. Coverage and Gap Characterization



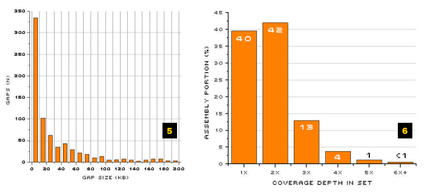
Neighbouring clones overlap by an average of 83kb, roughly 1/2 of the length of a BAC. The aggregate size of all clones in the clone set is approximately 4.6Gb.

On average, a canonical clone selected randomly from the fingerprint map overlaps its nearest neighbour in the set by 110 kb and in 90% of cases clones overlap by more than 80kb. Thus, any map clone can be associated with one or more clones from the clone set which provide equivalent coverage.

Figure 4

panels 1 & 2 | Characterization of clone set covers which are used to estimate the resolution of the set. Half of the covers are smaller than 40kb and 80% of the covers are smaller than 75kb.

panels 3 & 4 | Clone overlap for overlapping neighbouring clones with sequence coordinates. The average overlap is 83kb, approximately 1/2 of the length of a BAC



panel 5 | Characterization of coverage gaps in the clone set. The 729 gaps in coverage represented here are likely to be an overestimate because 6% of the clones in the set have not been localized within the assembly and a fraction of these may span the gaps.

panel 6 | Coverage depth by clone set BACs. Approximately 82% of the genome is covered in a 1:1 ratio at 1X:2X. This depth profile, calculated based on sequence coordinates, matches closely the profile determined by using the fingerprint map. Less than 6% of the genome is covered at 4X+.

4. Resource Availability

Additional information and news relating to the availability, distribution, and annotation of the set can be found at the Genome Sciences Centre web page, http://www.bcgsc.ca.

The BAC set clones are distributed through the **BACPAC Resource Centre** (www.chori.org/bacpac/pHumanMinSet.htm, Pieter de Jong, Children's Hospital, Oakland). The clone set has been rearrayed into chromosome-specific sub-sets, also distributed through BACPAC.

Acknowledgements

Resources

Clone libraries | RPCI-11/13: www.chori.org/bacpac; CalTechD 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

NHGRI

Genome Canada | Cancer Genomics: Victor Ling, Connie Eaves, Marco Marra