A Set of Rearrayed BAC Clones

Spanning the Human Genome

Martin Krzywinski, Jacquie Schein, Ian Bosdet, Duane Smailus, Calum MacAulay, Wan Lam, Steven Jones, Marco Marra

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

Genome Sciences Centre

www.bcgsc.ca
info@bcgsc.ca

Abstract

We have constructed a rearrayed set of BAC clones from the RPCI-11 and CalTechD libraries. The clones have been chosen from the human BAC physical map constructed at Washington University Genome Sequencing Centre. Our aim was to completely cover the entire genome with these BACs, as represented in the physical map, with a controlled degree of redundancy by using non-buried clones found in map contigs as the domain.

Our set of 29,052 size-filtered clones contains 99.6% of the fingerprint fragments found in the map and 97.5% of sequence found in the map. The rearray achieves 1X coverage for 45% of the map and 2X coverage for 46% of the map. Average coverage for any map clone by the best match in the rearray is 75% (108 kbp) with an average Sulston score of 10^{-14}.

We anticipate this resource will have several uses, including provision of a genome-ordered set of probes for fluorescent in-situ hybridization and provision of probes for microarray-based BAC-Comparative Genomic Hybridization (BAC-CGH) experiments. The rearray clone set will be fingerprinted to verify clone identity and to provide a high-quality map of the resource. Future plans include creating similar lists for mouse, rat, poplar and bovine genomes.

Clone Set Creation Algorithm

The human BAC physical map is a curated and sequence-validated resource, well suited for selection of a minimal set of clones spanning the genome. Using the Nov 2001 version of the map, our approach involves walking along each contig and selecting clones in the following fashion.

0. Repeat-masked BAC clone end sequences are searched by BLAST against repeat-masked human sequence (Aug 2001) to determine location of the clone, where possible. Hits with –log(Expect) > 20 or those with –log(Expect) > 50 and a match fraction of less than 80% are considered weak and discarded. For any clone, if BAC ends are closer than 2 kbp or further than 5 Mbp apart, they are discarded.

OUTER LOOP: foreach contig in map

1. All "non-buried" clones (canonical) in a contig are ordered by their left end positions (secondary sort by right end). Clones from RPCI-11 and CaltechD are marked as available and are eligible for the rearray. All other clones are considered unavailable. Clones that are buried or singletons (those not assigned to a contig) are not considered and are disregarded.

2. All admissible BAC end hits are used to determine the consensus chromosome for the map contig. Any clones with strong hits to other chromsomes are marked as inadmissable for the rearray (failed consensus). Out of the remaining clones, any clones farther than the length of the contig from the statistical mode BAC end position are also inadmissable. Hit polarity is used to determine left/right orientation of clones with only one admissable BAC end hit if a clone with both ends hits to the same sequence contig. Any clones whose single end cannot be categorized as left/right is marked as a middle hit and not used for overlap determination.

3. All admissible clones outside of the range of 100-200 kbp or 20-50 bands are marked and selected towards the rearray set only in places where a lack of coverage would otherwise occur.

4. The left-most clone is used as a starting point for a walk. To maximize coverage, the walk always starts at the first available clone, regardless of size, band or BAC end information. The following loop is iterated until the end of the contig is reached.

INNER LOOP: foreach clone in contig

5. A forward-looking local neighbourhood of clones is selected. This neighbourhood extends 20 cb map units past the current clone. For each clone in the neighbourhood, we determine the number of conserved bands (bands shared by the clones and all clones in between) and sequence overlap by using any available BAC end information.

6. From the neighbourhood clones we pick a single clone as the walk destination. Clones which pass the size/bands filter are always considered first. The evaluation proceeds in this order

i. right-most clone with sequence overlap is chosen

ii. right-most clone with admissible BAC ends is chosen. Clones with hits for both BAC ends must be at least 5 CB map units to the right. Clones with single BAC end hits must be at least 10 CB map units to the right.

iii. right-most clone with fewest conserved bands >4 is chosen

iv. if no such clones can be found, the next available clone is chosen, regardless of any size, band or BAC end information.

7. If the optimal clone, as determined solely by map position, fails size, band or BAC end position filters, then the next clone is chosen if overlap is not sacrificed. The algorithm is designed to minimize the number of sub-quality clones in the rearray, but choses them if they are the only alternative.

Maximized quantities

Coverage (by CB map units)
Clones which overlap by sequence
Clones with BAC end hits
Clones passing size/bands filter

Minimized quantities

Coverage by sub-standard clones
Clones found to lie on another chromosome (by BAC ends)
Clones found to lie far away from neighbours (by BAC ends)
Clones failing size/bands filter
High degree of overlap

Inset 1. The algorithm can be broadly described by this list of minimized and maximized characteristics. Where possible, the order in this list is used to rank the chracteristics.

Clone Set Characterization

Rearray Statistics

29,052 clones

95% RPCI-11
5% Caltech D1/D2

20% sequence clones (full X)

5% failed size/band filter
52% have BAC end hits
  42% both ends
  10% left only
  12% right only
  36% unclassified
19% overlap by sequence with next pick

4.3 Gbp total size
2.9 Gbp unique band size

Table 1. Global rearray statistics.

Global rearray statistics are shown in Table 1. The average map size of the rearray clones is 147 kbp and 34 bands. This very closely approximates the distribution of clones in the human physical map.

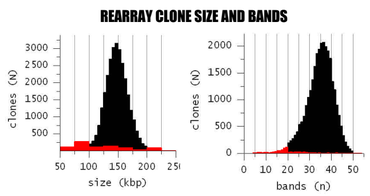


Figure 1 & 2. Size and band distributions showing all clones (black) and clones which failed size/bands filter (red).

Below are neighbour overlap statistics based on fingerprints in the human physical map. All fingerprint comparisons were carried out using a tolerance of 7 std mobility units.

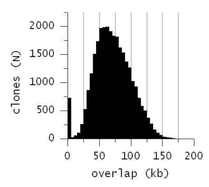


Figure 3. Distribution of overlap between rearray neighbours, as determined by the size of shared bands (average 71 kbp).

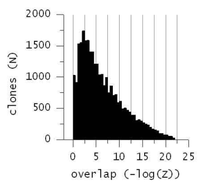


Figure 4. Distribution Sulston scores between rearray neighbours (average 6.4).

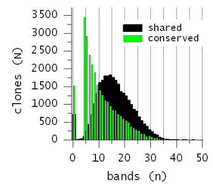


Figure 5. Distribution of the number of shared conserved bands between neighbours (average shared 16, conserved 10).

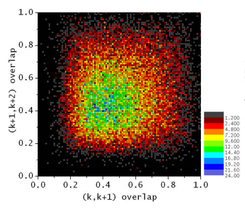


Figure 6. Frequency colour map of fractional map overlap between (k,k+1) clones and (k+1,k+2) clones. This lag plot is useful in exposing contiguous areas of deep or shallow coverage.

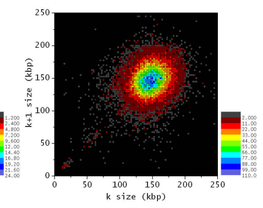


Figure 7. Frequency colour map of the size of clone k+1 vs clone k. This lag plot is useful in investigating areas covered by contiguous runs of small or large clones.

Figures 6 & 7 show how neighbouring rearray clones relate in terms of overlap and size. Areas of the map predominantly composed of small clones (e.g. contigs 25000-25097) show up as a cluster on the size lag plot near 25 kbp.

QUALITY CONTROL

Rearray clones will be fingerprinted to provide a set of high-resolution digest data to verify clone identity and placement.

Determination of Map Coverage

To illustrate map coverage, Table 2 shows the distribution of coverage for CB map units. Only 0.6% of units which contain clones are not covered by the rearray. These are areas with only unavailable of virtual clones. The CB map units within intracontig gaps are not counted towards the coverage determination because there are no actual clones within these areas.

CB Map Unit Coverage

6,828 units in intracontig gaps
607,890 units with clones

99.4% covered by rearray
  98.7% optimal clones
    45/46/8% 1/2/3 overage
  0.7% failed size/band filter
    79/18/3% 1/2/3+x coverage
  0.6% sequence overlap
0.6% not covered by rearray

Table 2. Map coverage by CB unit.

To assess how well the rearray represents map clones, every canonical, non-electronic map clone that is not in the rearray (113,000) was paired with its best match in the rearray. A high degree of map coverage would be indicated by these hits being on the same map contig, having low Sulston scores, high number of matching bands, and small left-end distance. The Sulston score was used to select the top 10 hits to the rearray. Figures 8-12 show the distribution of various statistics illustrating the extent of coverage.

For each map clone's top 10 hits, the highest ranking hit on the same contig was used for Figures 9-12. Only 0.6% of map clones did not show same contig hits in the top 10 list. These are (i) typically shorter clones (average 126 kbp) whose coverage is split between two rearray clones, or (ii) unavailable clones which do not overlap available clones.

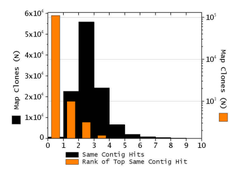


Figure 8. Distribution of the number of same contig hits among top 10 between map and rearray clones (black). The rank of the highest scoring same contig hit is shown in orange.

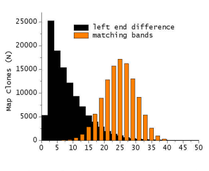


Figure 9. Distribution of left end difference (cb map units) between map clones and their best match in the rearray (black, average 10.6) and the number of matching bands in this pair (orange, average 24.4).

As a subset of map clones, the **rearray represents 99.4% of the fingerprint fragments in the map**. On average, a canonical map clone will overlap by 75% of its size with a rearray clone at a Sulston score of log Z = -14.4. 1X and 2X coverage by rearray clones accounts for 91% of the map and is proportioned equally - this feature increases the effective resolution of any hybridization or microarray experiments.

BEST MAP MATCH SCORE AND OVERLAP

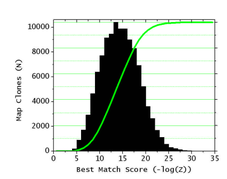


Figure 10. Distribution of the Sulston score, Z, for the same best contig hit (average 14.4). The cumulative distribution is shown in green (95% clones >7.5, 90% clones >8.8).

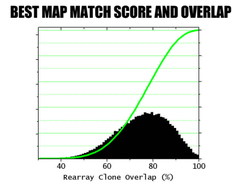


Figure 11. Distribution of the fractional overlap (sequence) between rearray neighbours (average 75.5, 95% clones >55% overlap, 90% clones >60% overlap).

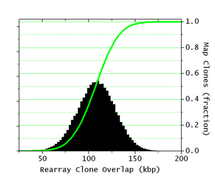
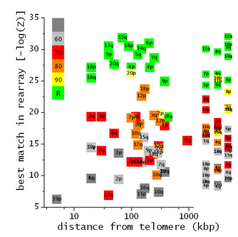
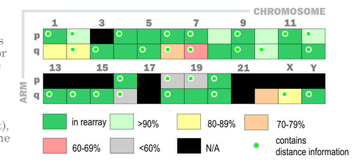


Figure 12. Distribution of the absolute overlap (sequence) between rearray neighbours (average 108 kbp, 95% clones > 74 kbp, 90% clones > 81 kbp).



About 110 RPCI-11 canonical map BACs associated with subtelomeric regions were matched to the rearray. For about half of these clones, some telomere distance information was available. Figure 13 shows the representation of these clones in the rearray.

Figure 13. Representation of subtelomeric map clones in the rearray. Clones for which distance from the telomere is not known are shown offset (left). The set of best rearray candidate for each telomere is shown (right), coloured by overlap to the BAC probes.



Determination of Sequence Coverage

Sequence coverage was evaluated using in-silico restriction digests from human sequence, counting the number of restriction fragments represented in the rearray. All possible non-overlapping 100 kb fragments without sequence gaps were created from sequence (17,000, approximately 56% of the genome) and digested in-silico with HindIII. Junction fragments and fragments smaller than 600 bp were removed and multiplets within 7 std mob units were collapsed to the largest fragment. This sanitization was done to mimic as closely as possible the sanitization state of clones in the fingerprint map. As a control, the same comparison was done for all canonical map clones, as well as for all available map clones.

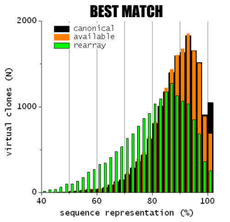


Figure 14. Distribution of sequence overlap between the virtual clone and its best match in the canonical, available and rearray clone sets (averages: rearray 80.5, available 87.9, canonical 88.4).

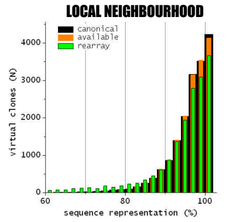


Figure 15. Distribution of sequence overlap between the virtual clone and the local neighbour of its best match in the canonical, available and rearray clone sets. This represents the extent of sequence representation in these clone sets (averages: rearray 93.3, available 95.6, canonical 95.7).

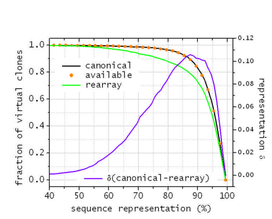


Figure 16. Right end cumulative distributions for histograms from Figure 14. The difference in the cumulative distribution between the rearray and the canonical clone set is shown. The canonical and available map sets perform nearly identically, since the map provides deep coverage and about 85% of the clones are available.

Sequence Coverage

	coverage	unmatched(matched) bands
Rearray	93.3(0.975)	2.0(90%)
Available Map	95.6(0.999)	1.6(93%)
Canonical Map	95.7(1.000)	1.5(93%)

Table 3. Sequence coverage.

Using our algorithm to determine sequence coverage, the rearray is found to represent 93.3% (2.0 digest fragments unmatched) of the tested sequence (56% of entire genome). As a control, the canonical map set was found to represent 95.7% (1.5 digest fragments unmatched) of the tested sequence. Since the rearray cannot represent more sequence than the map, being a subset of the map, normalizing by the control result is justifiable and removes dependance on the algorithm parameters. With this method, we find that **the rearray achieves 97.5% sequence coverage**, when compared to the map.

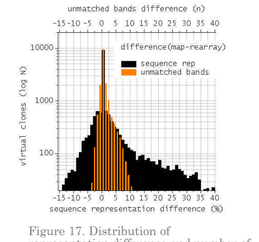
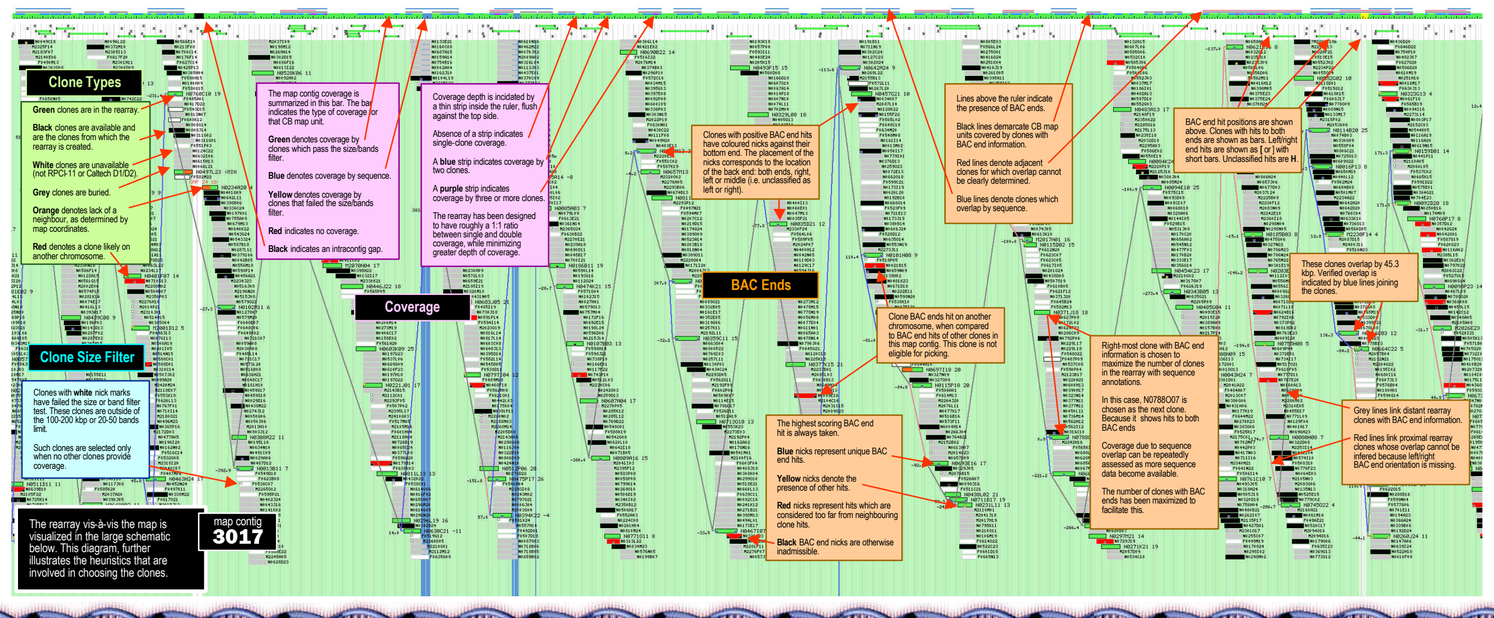


Figure 17. Distribution of representation difference and number of unmatched bands between the map and rearray clone set. Negative differences can be explained by the fact that the local neighbourhood for the best match in the map and in the rearray may be different.



Dr. Michael Smith
1932-2000
Founding Director

Acknowledgements

External Resources

Clone libraries RPCI-11: Osoegawa et al. (2001) Genome Res 11(3):483-496), CalTechD: Knight and Lese et al. (2000) Am J Hum Genet, 67(2):320-332

Fingerprinting Marra M etal. (1997) Genome Res 7:1072-84

BAC physical map Washington University Genome Sequencing Centre [www.genome.wustl.edu]

Golden path generation: International Human Genome Sequencing Consortium [www.nih.gov/news/pr/jun2000/nhgri-26.html], assembly: GigAssembler [genome.ucsc.edu/goldenPath/algo.html] by Jim Kent

BAC End database: TIGR [www.tigr.org]

Human Telomere Mapping and Sequencing Project Riethman H et al. [www.wistar.upenn.edu/Riethman]

Genome Sciences Centre Team

Bioinformatics Jacquie Schein, Chris Fjell, Steven Jones, Marco Marra

Laboratory Ian Bosdet, Duane Smailus, Carrie Mathewson, Natasja Wye

Associates Wan Lam, Calum MacAuley, Adrian Ishkanian

Funding

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