SUPPLEMENTARY MATERIAL

Clone set representation of the BAC fingerprint map

Three small contigs are not represented in the clone set: ctg20907 (2 clone contig), ctg25064 (4 clone contig which does not contain clones from libraries that we sampled) and ctg25101 (2 clone contig). In order to evaluate representation of each contig, we calculated the coverage of the clones in the clone set in terms of cbmap units. The cbmap unit is the distance metric used in the BAC fingerprint map. Each unit on this scale corresponds to a single detected fingerprint restriction fragment that is confirmed by at least two clones, but is not associated with a restriction fragment size. Hence, local genome distances measured in cbmap units cannot be precisely related to a distance in sequence coordinate space. Each clone in a fingerprint map contig is positioned relative to all other clones using these cbmap coordinates. The fingerprint map contains clones spanning 600,376 cbmap units, out of which 599,622 are covered by clones in RPCI-11, RPCI-13 and Caltech-D libraries. The clone set provides coverage for 586,343 cbmap units, corresponding to map scale coverage of 98%. We found that the cbmap unit gaps were on average 8 units in size (median 6 units) with the largest gap of 75 units. It is important to emphasize that gaps in cbmap unit coverage do not necessarily indicate gaps in sequence coverage. Specifically, during manual editing of clone order in the fingerprint map, precise cbmap positions of clones are frequently adjusted making the precise use of the unit impossible.

Our second method of evaluating map coverage was to examine extent of overlap of all map-adjacent clones in the set. Out of 29,949 such clone pairs, the median number of conserved bands between clone pairs was 12 (56 kb). We found 382 pairs (1.3%) in

which fingerprint overlap was weak and actual overlap could not be accurately inferred. These 382 clone pairs on average shared 5 digest fragments, in contrast to an average of 23 fragments shared by the remaining 29,567 pairs. For 94 of the 382 pairs, we found verifiable sequence overlap between the clones based on their BAC-end, assembly or in silico positions. In these circumstances, the overlap was too small to reliably detect using fingerprint data. For 190 of the 382 pairs, one or both of the clones could not be localized on the sequence. For the remaining 98 cases we found that the clones failed to overlap based on sequence information. The median size of the gap between clones in these pairs was 20kb. It is possible that at least some of these gaps have arisen as a result of inconsistencies between different versions of the fingerprint map. For example, clones were selected using one version of the map, and the selections assessed using the latest updated version of the map.

The depth at which the clone set covers the genome can be approximated using the cbmap unit coverage of the fingerprint map, with the assumption that the ratio of cbmap units to in silico-predicted restriction fragments is relatively constant over large map distances. In the fingerprint map, 37% of cbmap units are covered by only one clone in our clone set and another 43% are covered by the overlapping region of two clones. Finally 18% are covered by 3 or more clones. 2.3% of the cbmap units are not represented in the clone set. The average coverage depth was calculated to be 1.8X. A similar analysis of coverage depth carried out with sequence coordinates corroborates this result (Fig. 4).