Text Box 1. Completing the *C. elegans* genome sequence

At publication in 1998, there were tens of unfinished YACs and three unfinished cosmids and fosmids. These clones were all completed over the next year or two using the array of methods available for clone finishing (International Human Genome Sequencing Consortium 2004). In addition, we corrected ~20 misassembled, ambiguous, or deleted regions along with ~200 single base corrections (mostly in early projects) stemming from detailed analysis of Expressed Sequence Tags (ESTs) (McCombie et al. 1992; Waterston et al. 1992; Kohara 1996; The *C. elegans* Genome Sequencing Consortium 1998) and other data including community feedback.

More significantly, there remained two internal map gaps on Chromosomes III and IV, respectively, where no spanning clones were available, and three telomeric (Chromosome II right, where left and right are with reference to the genetic map) or subtelomeric (Chromosome I left and Chromosome X left) gaps. The telomere clone cTel33B (one from a set of eleven isolated by Wicky et al. 1996) eventually overlapped Y74C9 as its sequence was completed, capping the left end of Chromosome I. Plasmid cTel7X was linked to Y35H6 on the left end of Chromosome X through three PCR fragments, capping that chromosome end.

The internal gaps persisted despite the high redundancy of the initially mapped clones (some 30-fold from YAC, cosmid, and fosmid clones) and after screening a new BAC library (Exelixis, http://www.exelixis.com, pers. comm.). Given the rarity of these regions in large insert clone libraries, we turned to a strategy of directly subcloning and shotgun-sequencing a restriction fragment from whole genomic DNA for these internal gaps and the uncloned telomere from Chromosome II right.

The regions containing the internal gaps and the remaining telomere were mapped by macrorestriction Southern-blot analysis, using probes derived from the known flanking sequence. To obtain useful purity of the fragments, we adopted a successive digest scheme, using pulsed field gel electrophoresis (PFGE) to isolate the product of the first digest, digesting this in situ with a second enzyme, and subcloning the isolated DNA from a second PFGE purification. Inevitably these libraries were contaminated with copurifying DNA (50%–95% contaminated), but the dominant contig was easily identified in each case and the rest accounted for with known sequence.

The spanning sequence for the internal gaps was in each case a small fraction of the size predicted by Southern blots (6 kb vs. the predicted 250 kb and 20 kb vs. 70 kb for Chromosomes III and IV, respectively). Perhaps the fragment mobility in PFGE can be anomalous at high concentrations (Doggett et al. 1992) (we used 50–100 µ/mL) or result from unusual sequence features, which might also account for the poor representation of the regions in libraries. The telomere segment was in better agreement (82 kb vs. 90 kb predicted), with the difference accounted for at least in part by exclusion of the telomere repeat from the assembled sequence.