Summary of BAC sequence integration

By Zhangjun Fei

BAC sequence processing

Tomato BAC sequences version 548 (file name: bacs.v548.seq) were downloaded from SGN on July 6, 2010. This collection of BAC sequences contained 984 phase 3 BACs (total: 106,142,954 bp), 266 phase 2 BACs (total: 28,057,077 bp), and 208 phase 1 BACs (total: 23,226,440 bp). Phase 1 BACs were excluded from the integration.

Phase 2 and 3 BACs (1,250 with total length of 134,200,031 bp) were assembled into 658 contigs with a total length of 117,663,676 bp. The contigs were polished (corrected for base error) by Solexa sequences and further screened for E. coli genome sequences. Sequences matching E. coli genome and additional 100 bp from both sides of the matched regions were removed.

The contig sequences were split into smaller fragments based on the gaps they contain. There are three reasons why we need to split the contigs: 1) phase 2 sequences contain gaps (Ns) and we can't introduce these Ns into the genome assembly; 2) the order of the fragments in each phase 2 BAC could be wrong; 3) the strand information of some fragments in phase 2 BACs could be wrong. After splitting, a total of 1,118 sequences with a total length of 117,511,748 bp were obtained. These 1,118 sequences were used for final integration.

BAC sequence integration

Tomato BAC sequences were first aligned to scaffolds (v2.10) of genome assembly using megablast. The resulting alignments were manually curated to remove false alignments.

Based on the alignments, BAC sequences were integrated into the genome assembly following the guidelines described below:

1. If the whole BAC sequence can be aligned to a scaffold and in the alignments all the unmatched regions in the scaffold are gaps (Ns) or very short sequences (<500 bp), or there are deletions in the scaffold, the whole BAC is used to replace the corresponding regions in the scaffold.
2. If a BAC sequence can be aligned to a scaffold but in the alignments the unmatched regions in the scaffolds contain a long sequence (>500bp) and all others are gaps (Ns), only the matched regions of the BAC sequence are integrated.
3. If two parts of a BAC sequence are aligned to different locations of a scaffold (possible truncated or chimerical BAC, or genome assembly error) or aligned to two different scaffolds (possible chimerical BAC, or genome assembly error), the matched parts of the BAC sequences were integrated into the genome assembly separately.

4. Small scaffolds that are fully contained in an integrated BAC sequence were removed.
5. BAC sequences having no matches in the genome assembly or too many inconsistencies with the genome assembly were not integrated.

Conclusion

A total of 116,628,855 bp BAC sequences (99.25%) were integrated into the genome assembly and these BAC sequences removed 4,494,022 Ns from v2.1 scaffolds.