Figure S1: Comparison of PacBio and Sanger indel lengths. Total number of indel events by length and sequence type.

Figure S2: Distribution of homopolymer sequence lengths associated with PacBio/Sanger assembly mismatches. Colors indicate validation status by Illumina reads mapped to PacBio- and Sanger-based assemblies of BAC clones. The dashed vertical line indicates the potential maximum length of homopolymer sequences Illumina HiSeq machines can accurately sequence (Minoche et al. 2011).

Figure S3: Dotplot alignment of Sanger assembly for CH17-41F14. The dotplot alignment (word size = 20) of the Sanger assembly for the clone CH17-41F14 indicates the complex repetitive sequence near the end of the clone.

Figure S4: Allora vs. HGAP assembly. a) Alignment of the Allora assembly for CH17-227A2 against the Sanger assembly with a decrease in PacBio coverage over a repeat structure indicating a misassembly. b) Alignment of the HGAP assembly for CH17-227A2 against the Sanger assembly. The incorrectly expanded repeat structure in the Allora assembly is resolved by HGAP with a seed cutoff of 5,800 bp.

Figure S5: Composition of mismatches between PacBio and Sanger assemblies. Mismatches between assemblies of BAC clones are shown by validation status. Colors indicate the type of difference between sequences.

Figure S6: Pairwise alignment of a 372 bp misassembled region from Sanger assembly with PacBio assembly. Sequences shown in red indicate mismatches within the assembly that have Illumina support for PacBio sequence.

Figure S7: Alignment of CH251 clone end sequences to supercontigs built from the complete BAC inserts. a) Orientation of BES is indicated by the direction of the sequence arrows. Alignments shown are all at >99.8% identity. The mean identity of BES alignments was 99.72% (16,174/16,220 high-quality bases). Twelve clones mapped concordantly (mean identity of 100%), five mapped discordantly with both ends (mean identity of 99.32%), and six mapped with one end only (mean identity of 99.03%). b) Alignment of concordant and discordant chimpanzee fosmid end mappings to CH251 supercontigs. The mean identity of fosmid end alignments was 99.69% (245,005/245,758 high-quality bases). A total 181 clones mapped concordantly (mean identity of 99.82%) while 39 mapped discordantly in pairs (mean identity of 99.56%) and 76 mapped with only one end (mean identity of 99.18%).

Figure S8: Insert size distribution of concordant fosmid end mappings to chimpanzee supercontigs. The mean (stddev) for Contig A is 36,773 bp (2643) and for Contig B it is 37,306 bp (3016).

Figure S9: Post-filter distribution of PacBio read length and quality for all eight clones. a) CH17-124M20; b) CH17-157L1; c) CH17-169A24; d) CH17-170H8; e) CH17-202L17; f) CH17-227A2; g) CH17-33G3; and h) CH17-41F14.

Figure S10: BAC assembly pipeline. Flowchart of assembly process including management of raw reads in HDF5 (.bas.h5) files through vector screening, assembly, and refinement.

Figure S11: Assembly results for clones subsampled at 100X coverage. The number of assembled contigs for ~20 assemblies per clone based on subsampling reads to 100X coverage. Clone CH17-169A24 was omitted due to the presence of multiple BACs in one SMRT Cell and clones CH17-170H8 and CH17-33G3 were omitted due to contamination in SMRT Cells.

Figure S12: Assembly of complex sequence at 100X coverage. One assembly of CH17-41F14 at 100X coverage of PacBio reads from 19 subsampling iterations, shown here aligned to the Sanger assembly of the clone, nearly recreates the most complex region of the clone, which is collapsed when assembled with higher coverage. The alignment identity of this assembly with the Sanger sequence is 99.95% compared with the alignment identity of 99.99% between the original assembly of the clone with all reads.