Reconstructing complex regions of genomes using long-read sequencing technology

John Huddleston[1,6], Swati Ranade[2], Maika Malig[1], Francesca Antonacci[3], Mark Chaisson[1], Lawrence Hon[2], Peter H. Sudmant[1], Tina A. Graves[4], Can Alkan[5], Megan Y. Dennis[1], Richard K. Wilson[4], Stephen W. Turner[2], Jonas Korlach[2], and Evan E. Eichler[1,6]

1. Department of Genome Sciences, University of Washington, Seattle, WA, 98195, USA
2. Pacific Biosciences of California, Inc., Menlo Park, CA, 94025, USA
3. Department of Biology, University of Bari, Bari, 70126, Italy
4. The Genome Institute at Washington University, Washington University School of Medicine, St. Louis, MO, 63110, USA
5. Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey
6. Howard Hughes Medical Institute, University of Washington, Seattle, WA, 98195, USA

Correspondence to:

Evan E. Eichler, Ph.D.
Department of Genome Sciences
University of Washington School of Medicine
Foege S-413A, Box 355065
3720 15th Ave NE
Seattle, WA 98195-5065
E-mail: eee@gs.washington.edu

Running title: Assembling complex genomic regions with long reads

Keywords: segmental duplication, assembly, PacBio, Sanger, capillary

ABSTRACT

Obtaining high-quality sequence continuity of complex regions of recent segmental duplication remains one of the major challenges of finishing genome assemblies. In the human and mouse genomes, this was achieved by targeting large-insert clones using costly and laborious capillary-based sequencing approaches. Sanger shotgun sequencing of clone inserts, however, has now been largely abandoned leaving most of these regions unresolved in newer genome assemblies generated primarily by next-generation sequencing hybrid approaches. Here we show that it is possible to resolve regions that are complex in a genome-wide context but simple in isolation for a fraction of the time and cost of traditional methods using long-read single molecule, real-time (SMRT) sequencing and assembly technology from Pacific Biosciences (PacBio). We sequenced and assembled BAC clones corresponding to a 1.3 Mbp complex region of chromosome 17q21.31, demonstrating 99.994% identity to Sanger assemblies of the same clones. We targeted 44 differences using Illumina sequencing and find that PacBio and Sanger assemblies share a comparable number of validated variants, albeit with different sequence context biases. Finally, we targeted a poorly assembled 766 kbp duplicated region of the chimpanzee genome and resolved the structure and organization for a fraction of the cost and time of traditional finishing approaches. Our data suggest a straightforward path for upgrading genomes to a higher quality finished state.

INTRODUCTION

Complete high-quality sequence assembly remains a difficult problem for the *de novo* assembly of genomes (Alkan et al. 2011b; Church et al. 2011; Salzberg et al. 2012). Finishing of the human and mouse genome involved selecting large-insert BAC clones and subjecting them to capillary-based shotgun sequence and assembly (English et al. 2012). Sanger-based assembly of large-insert clones has been typically a time-consuming and expensive operation requiring the infrastructure of large genome sequencing centers and specialists focused on particular problematic or repetitive regions (Zody et al. 2008; Dennis et al. 2012; Hughes et al. 2012). Such activities can significantly improve the quality of genomes, including the discovery of missing genes and gene families. A recent effort to upgrade the mouse genome assembly, for example, resulted in the correction or addition of 2,185 genes, 61% of which corresponded to lineage-specific segmental duplications (Church et al. 2009). Within the human genome, there are over 900 annotated genes mapping to large segmental duplications. About half of these map to particularly problematic regions of the genome where annotation and genetic variation is poorly understood (Sudmant et al. 2010). Such genes are typically missing or misassembled in working draft assemblies of genomes. These include genes such as the *SRGAP2* family, which evolved specifically in the human lineage and is thought to be important in the development of the human brain (Charrier et al. 2012; Dennis et al. 2012). Other regions (e.g., 17q21.31 inversion) show incredible structural diversity, predispose specific populations to disease, and have been the target of remarkable selection in the human lineage (Stefansson et al. 2005; Zody et al. 2008; Steinberg et al. 2012). Such structurally complex regions were not resolved within the human reference sequence until large-insert clones were recovered and completely sequenced.

The widespread adoption of next-generation sequencing methods for *de novo* genome assemblies has complicated the assembly of repetitive sequences and their organization. Although we can generate much more sequence, the short sequence read data and inability to scaffold across repetitive structures translates into more gaps, missing data, and more incomplete references assemblies (Alkan et al. 2011a; Salzberg et al. 2012). Due to budgetary constraints, traditional capillary-based sequencing capacity as well as genome finishing efforts have dwindled in most sequencing centers leaving most of the complex regions of working draft genomes unresolved. Clone-based hierarchical approaches remain important for reducing the complexity of genomes, but even targeted sequencing of these clones using short-read data fails to completely resolve and assemble these regions due to the presence of highly-identical repeat sequences common in mammalian genomes. Here, we tested the efficacy of a method developed for finishing microbial genomes (Chin et al. 2013) to a 1.3 Mbp complex region of human chromosome 17q21.31 previously sequenced and assembled using traditional Sanger-based

approaches. We directly compared sequenced and assembled clones and validated differences to highlight advantages and limitations of the different technologies. We then applied the approach to a previously uncharacterized, highly duplicated region of the chimpanzee genome and show that we can rapidly resolve the structure and organization of the region using this approach.

RESULTS

For the purpose of this study, we initially selected eight BAC clones from a hydatidiform mole sample corresponding to a complex 1.3 Mbp region of 17q21.31 (Figure 1). The region was chosen because of its biomedical relevance and the difficulty it posed in the initial sequence and assembly of the human genome. Of the corresponding clone sequence, 55% consist of high-identity segmental duplications and the region is a site of large-scale structural polymorphisms that predisposes European and Mediterranean populations to recurrent microdeletions associated with the Koolen-DeVries syndrome (Zody et al. 2008; Steinberg et al. 2012). Although the targeted region is complex, we note that it does not contain any sequences that are recalcitrant to existing sequencing technologies. Its complexity lies in the presence of layers of common and low-copy repeat sequences, which complicates assembly at the whole-genome level and has typically required targeted clone-based approaches to resolve. We constructed 10 kbp insert sequence libraries and assembled PacBio sequence into consensus sequence contigs using the HGAP long-read assembler. Quiver was used to generate a final consensus with quality scores through the standard SMRT Analysis (v. 2.0.1) pipeline (Chin et al. 2013). In this study, the average subread length across all clones was 1.8 kbp (maximum length of 12.4 kbp) and sequence coverage ranged from 78- to 475-fold (average of 245-fold per clone). After vector trimming, we generated a single, linear contig sequence for each of the eight clones representing a total of 1,774,407 bp of “finished” sequence. We note that each of the eight clones assembled into a single contig (Tables 1 & S1), with six clones assembled from a single SMRT Cell of data each.

For each sequenced clone, we aligned the Sanger and PacBio HGAP assembled sequence contigs using BLASR (Chaisson and Tesler 2012) and identified all sequence differences <50 bp in length. A total of 125 sequence differences were identified in 1.77 Mbp of aligned sequence resulting in 99.994% sequence identity between the assemblies (Table 2). In this estimate, we count the total number of base pairs encompassing a given insertion/deletion (indel) event as opposed to counting an indel as a single difference. Relatively few sequence differences (24 or 19%) were sequence substitutions. For example, five of the aligned clones showed no sequence substitution difference between the two assemblies. The bulk of sequence differences, instead, corresponded to insertions (81% or 101 aligned base-pair differences). We note marked asymmetry between assemblies with 76 insertions found in the PacBio assemblies and only 25 insertions in the Sanger assemblies (Figure S1). Simple repeats contributed to 48% of the differences with 47 differences occurring within homopolymer runs (Figure S2) and 13 within dinucleotide repeats.

Overall, the assembled contigs showed remarkable similarity in length (1,788 kbp Sanger vs. 1,774 kbp PacBio). The difference in length was due primarily to one clone where there was evidence of larger structural differences between the assemblies (Figure 2). The assembly of CH17-41F14 contained a 12 kbp complex higher-order repeat structure, which was expanded to 20 kbp in the Sanger assembly (Figure S3). Visualization of the corresponding read depth (Figure 2) confirmed a PacBio misassembly—i.e., a symmetric increase of reads in the collapsed region of CH17-41F14 (Figure 2). In addition to this PacBio misassembly, we discovered a 357 bp deletion in the Sanger assembly of CH17-41F14 (Table S2), which was subsequently confirmed as *bona fide* in the PacBio assembly based on alignment of Illumina reads from the same clone to both assemblies (Table S3). Interestingly, HGAP correctly assembled the clone CH17-227A2, which had been previously misassembled by an earlier long-read assembly algorithm, Allora, during our preliminary analysis (Figure S4).

To determine if the smaller sequence differences were errors in the PacBio or the Sanger assembly, we sequenced the same eight clones to high coverage (average 94-fold) using a Nextera Illumina sequencing pipeline (Adey et al. 2010). Short-read sequencing data were insufficient to assemble the complete insert, even in the case for clone CH17-170H8 where the longest exact repeat, 76 bases, is shorter than the read length (Table S4). Local alignment of the Illumina assemblies allowed us to unambiguously validate 44 differences between the assemblies (Table 3, Figure S5). Illumina sequencing supported 31 PacBio and 13 Sanger differences. The majority of variants supported in PacBio assemblies (97%) clustered within complex repetitive regions. For example, a 372 bp region in the Sanger assembly of CH17-169A24 accounted for 24 unambiguous differences, suggesting that this region had been misassembled in the Sanger assembly (Figure S6). The remaining validated PacBio difference corresponded to a homopolymer repeat. Similarly, the validated differences within the Sanger assemblies shared common features. Five (38%) of the validated Sanger variants occurred within simple repeat sequences. For the four validated Sanger variants within homopolymers, one of the alternate PacBio variants added an extra base, one added two bases, and one removed a single base. The remaining eight variants validated within the Sanger assemblies were evenly split between complex indels and mismatches where the PacBio assembly had potentially misassembled segments with no coverage between regions of normal coverage. We manually inspected the capillary traces for five of the eight clones at 23 total mismatch positions between Sanger and PacBio assemblies. Of these mismatches, 20 previously ambiguous mismatches were validated for the Sanger assemblies and three mismatches previously validated for PacBio were also supported by the capillary traces.

Since the average sequence coverage per clone was relatively high (245-fold), we performed two experiments to estimate the minimum coverage required to properly assemble clones into a single contig. In the first experiment, we randomly subsampled (100-fold coverage) and

assembled ~20 iterations per clone. We measured the success of these assemblies by their identity to the corresponding Sanger assembly, median number of assembled contigs, and total bases assembled per clone. The mean identity of assemblies ranged from 99.98% to 99.99%. Four of five clones had a median of one contig, while one clone, CH17-124M20, had a median contig count of two (Figure S11, Table S5). The identity between single-contig subsampled assemblies and their Sanger counterparts was 0.01% lower than the original HGAP assemblies with all reads (Tables 2 & S5). Interestingly, one assembly of the clone CH17-41F14 matched the length of the Sanger assembly by adding ~8 kbp in the complex repetitive region that had collapsed in the original assembly (Figure S12). However, this subsampled assembly had lower overall identity with the Sanger sequence at 99.95% compared to the original assembly's 99.99%. These results suggest that the previously recommended coverage of 100-fold for high-quality libraries is a minimum requirement for high-accuracy BAC assembly (Chin et al. 2013).

In the second experiment, we empirically assessed the efficacy of pooling BAC clones in individual SMRT Cells to determine if distinct assemblies of high quality could be produced. All pooled clones had been previously sequenced and assembled from single SMRT Cells with three out of four clones assembling into a single contig (Table S6). We tested a pool of two clones (Pool #1: CH251-75B17 and CH277-30K2) and a second pool of three clones (Pool #2: CH251-75B17, CH251-182P19, and CH277-80C4). In each case, clones were isolated independently and DNA normalized prior to library construction. Four of the five pooled clones assembled into single contigs. The clone CH277-30K2, which had previously assembled into one contig, assembled into two from the pooled data. Interestingly, the clone CH277-80C4 assembled into one contig from the pool with an additional 13 kbp of sequence compared to its single SMRT Cell assembly of five contigs. The single and pooled assemblies for the remaining clones were structurally concordant and ranged in alignment identity between 99.85% and 99.99%.

To demonstrate the utility of this approach for upgrading working draft assemblies, we identified five clones (CH251) corresponding to two complex segmental duplications within an orthologous region of the Smith-Magenis syndrome (SMS) in the chimpanzee (see Methods). We specifically selected this region because our previous analysis had shown it to be the site of complex lineage-specific duplications that had not been properly assembled in the chimpanzee genome (Sudmant et al. 2013). Moreover, the first chimpanzee analog of a genomic disorder had been identified within this region. The chimpanzee showed an SMS-like phenotype although the breakpoints of this rearrangement could not be reliably identified due to misassembly of the segmental duplications (Sudmant et al. 2013). Each chimpanzee BAC clone was sequenced and assembled as described above and each clone assembled into a single insert of the expected length. Two supercontigs were generated corresponding to 504 kbp of segmental duplication. This expanded to 766 kbp when including one orphan capillary clone sequence that had not yet been incorporated in the chimpanzee assembly (Table S7). A comparison to the current chimpanzee genome assembly (panTro4) showed that 241 kbp of sequence was completely

absent from the chimpanzee whole-genome assembly (Figure 3). The remaining 525 kbp which showed homology to sequence in panTro4 was distributed to six contigs, most of which were not localized (i.e., assigned to the random bin on unmapped chromosome). Only one ~44 kbp region was assigned correctly to a map location on chromosome 17. Alignment of supercontigs with the genome assembly revealed hundreds of small and large inconsistencies with an overall sequence identity of 94.69%.

To assess the accuracy of this new assembly, we mapped both chimpanzee BAC-end and fosmid-end sequences (BES and FES) to the assembled contigs (Figures 4 & S7) restricting alignments to high-quality base pairs from the capillary traces (Phred quality score >30). The mean identity of all BES alignments was 99.72% (16,174/16,220 high-quality bases). Twelve clones mapped concordantly (mean identity of 99.99%), five mapped discordantly with both ends (mean identity of 99.32%), and six mapped with one end only (mean identity of 99.03%). Alignment of concordant and discordant chimpanzee FES to CH251 supercontigs showed a mean alignment of 99.98% (156,955/156,991 high-quality bases). A total 138 clones mapped concordantly (mean identity of 99.98%) while 17 mapped discordantly in pairs (mean identity of 99.98%) and 20 mapped with only one end (mean identity of 99.99%). We note that the assembled contigs are largely composed of high-identity duplications and many of the lower-identity discordant read-pairs likely originate from paralogs or alternate structural haplotypes within the chimpanzee. Importantly, analysis of the fosmid insert size distribution based on mapped FES to the chimpanzee supercontig shows a tight insert size distribution (37 +/- 3 kbp) revealing that 99% of the assembly was spanned correctly by fosmid end sequence pairs of high identity (Figure S8). These data confirm the order, orientation, and sequence accuracy of the clone-based assembly of this complex region of the chimpanzee genome (Figure S7).

DISCUSSION

Our data suggest that SMRT sequencing of large-insert clones can significantly improve sequence assembly within complex repetitive regions of genomes, including segmental duplications. Clones assembled both with capillary-based and SMRT sequencing compared favorably in length and sequence accuracy (99.994%). The most common error within the assembled clones was the addition of a single base pair particularly in homopolymer runs, which is consistent with previous reports of potential artifacts of SMRT technology (Au et al. 2012; Okoniewski et al. 2013). High sequence coverage (>90-fold) and the single-base-pair error correction model afforded by Quiver were key to accurate assembly. In addition, the long reads were critical to traversing common repeats. It is instructive, for example, that sequence collapses were restricted to the largest and most identical tandem repeats within each clone. The HGAP assembler (Chin et al. 2013), which sub-selects the longest reads upon which to scaffold an assembly, readily resolved a 2 kbp artifactual duplication from our preliminary Allora assemblies but was unable to fully resolve a 20 kbp higher-order tandem repeat. In the case of the latter, it is

interesting to note that the longest read generated for the clone CH17-41F14 (~12.3 kbp) was shorter than the tandem duplication and that the largest repeat sequence generated in this assembly was ~12 kbp. We predict that larger clone libraries and longer subread lengths will be required to resolve these most problematic regions. Automated gel electrophoresis systems such as BluePippin (Sage Science, Beverley, MA) may be particularly useful in this regard because they facilitate the preparation of larger insert libraries that can traverse larger repeats.

Despite these limitations, application of the PacBio sequencing approach confers significant advantages in terms of cost, labor, and throughput. Sequencing centers recently estimated that finishing a single BAC clone to high quality (QV>45) using capillary-based approaches now costs between $4000 to $5000 per clone. Approximately 30-50 clones per month could be completely sequenced and assembled given a staff of three to four dedicated persons within The Genome Institute at Washington University. We estimate that with one PacBio RS sequencing machine, a single technician with part-time bioinformatics support can produce ~100-120 clone assemblies per month with ~85% being completely finished with an estimated error of 1 mismatch/10,000 using the HGAP/Quiver assembly approach. The cost per finished clone is estimated at approximately $625 (per SMRT Cell)—based on a survey of cost-center rates of five centers currently operating PacBio RS machines. Of course, the cost decreases and throughput increases multifactorially if BACs are pooled. We note that our benchmark pooling experiments were performed with a PacBio RS machine with 75,000 ZMWs (zero-mode waveguides). Current upgrades (PacBio RS II) double the number of productive ZMWs (n = 150,000) and increase movie times making larger BAC pooling schemes feasible. We caution that target regions frequently harbor internal large repeats and automated assembly benefits from both high coverage and the reduced complexity of the large-insert target. Downsampling and pooling experiments highlight the need for sufficient coverage (>=100X) and high-quality DNA libraries for each clone. Barcoding, which is now possible, may further improve pooling of multiple clones within a single SMRT Cell (http://www.pacificbiosciences.com/pdf/TN_Multiplexing_Targeted_Sequencing_Using_Barcodes.pdf).

One approach to improve existing working draft genome assemblies would be to leverage the extensive BES data for many mammalian genomes to select large-insert clones spanning gaps and repeats and mapping to collapsed regions of segmental duplication. All BAC clones mapping to a problematic region (as well as extending 50-100 kbp outside of it for anchoring purposes) could be selected and sequenced to high coverage in 96-well pools using a Nextera Illumina-based sequencing protocol (Adey et al. 2010). Although *de novo* assemblies of 150 bp Nextera reads tend to fragment within homopolymer and SINE/Alu repeats, the mapping positions of short reads from clones could be used to define an optimal tiling path of clones (~10-20 clones per region). Once a tiling path of clones has been established for each region, clones could be sequenced in pools of 2-3 clones, assembled using HGAP/Quiver, and validated by mapping

fosmid paired-end sequences to the final assemblies. Clones that failed to assemble into a single contig could be subjected to higher coverage using one clone prep per SMRT Cell. For genomes without fosmid end sequence data, gel-extraction of DNA and sequencing using orthogonal chemistries would be an important development to enable validation of *de novo* assemblies.

It should be emphasized that this procedure is a targeted one rather than a genome-wide approach. Other strategies have been described to upgrade draft genome assemblies by leveraging long-read sequence data or long-range information provided from Hi-C sequence data (English et al. 2012; Burton et al. 2013). While these methods systematically improve chromosomal contiguity across the genome (as measured by N50 contig length), they fail to accurately assemble the most complex regions of segmental duplications (Burton et al. 2013). Regions targeted by our approach are frequently missing or grossly misassembled by whole-genome shotgun sequence assembly using either capillary or next-generation sequencing platforms (Alkan et al. 2011b), still requiring high-quality sequencing of large-insert clones to correctly resolve. Analysis of the mouse and human genomes suggests that these typically correspond to 300-500 regions (~140-150 Mbp) per genome, including in some cases almost entire chromosomes, such as the Y chromosome (Hughes et al. 2012). The approach we have described provides a strategy to resolve these more structurally complex regions during the final stages of assembly, ensuring that the 1000-2000 genes mapping therein become incorporated within future mammalian genome assemblies (Alkan et al. 2011b; Church et al. 2011).

METHODS

PacBio DNA Preparation. BAC DNA from CHORI-17 (CH17) and CHORI-251(CH251) clone libraries (http://bacpac.chori.org) was isolated using a High Pure Plasmid Isolation Kit from Roche Applied Science per manufacturer instructions using 10 mL LB media with Chloramphenicol selective marker. We isolated ten preps per BAC yielding ~10 µg of starting material.

PacBio Sequencing. Approximately 5 µg of BAC DNA was mechanically sheared to a size of ~8 kbp, using the Hydroshear® system and large assembly at a shearing speed of 9 for 20 cycles per manufacturer instructions. SMRTbell® libraries were prepared for each sample by ligation of hairpin adaptors at both the ends (Travers et al. 2010), using PacBio DNA Template Prep Kit 2.0 (3–10 kbp) for SMRT Sequencing with C2 chemistry on the PacBio® RS according to manufacturer instructions. Libraries were purified using (0.45X) Agencourt® AMPure® beads to remove short sheared inserts below 1.5 kbp. The sheared DNA template was characterized for size distribution using an Agilent Bioanalyzer 2100 along with a 12k chip and the means from the fragment distribution were between 7 to 9 kbp, while the overall fragment inserts distribution ranged from ~2 kbp to 13 kbp (Figure S9). Sequencing primers were annealed to the templates at a final concentration of 5 nM template DNA and DNA polymerase enzyme C2 was complexed per manufacturer's recommendation for small-scale libraries. DNA/Polymerase Binding Kit 2.0

(PacBio) was used for setting up enzyme template-complexes and libraries were loaded on to the 75,000 zero-mode waveguides (ZMWs) following instructions in the complex setup and loading calculator provided by the manufacturer. Sequencing Kit 2.0 (PacBio) was used for sequencing using 45 min sequence capture protocol along with stage start to maximize subread length, on the PacBio–RS. With the exception of accidental and intentional pooling, each SMRT Cell contained a single BAC. For pooling experiments, libraries were made individually following "Reduced Input 10 kbp Template Preparations" per manufacturer instructions. BACs were pooled with finished library using roughly equimolar concentrations and sequenced in a standard diffusion run.

Clone Sequence Assembly. *De novo* assembly of BAC inserts was performed using the standard SMRT Analysis (v. 2.0.1) pipeline. Reads were masked for vector sequence (pBACGK1.1) and assembled with HGAP followed by consensus sequence calling with Quiver (Chin et al. 2013) (Figure S10). HGAP creates a scaffold assembly using the longest reads (e.g., >7 kbp) as seeds to recruit additional subreads as a scaffold while Quiver is a multi-read consensus algorithm that takes advantage of the full information from the raw pulse and base call information generated during SMRT sequencing. Final assembly was performed using a minimum read length of 500 bp and minimum read quality of 0.80 on a PC cluster (eight cores/10 GB of RAM) running RedHat 6 SE. We screened unsplit PacBio reads in FASTA format with cross_match using the recommended settings for contamination screening (-minmatch 10 -minscore 20 -screen). PacBio assemblies were reviewed for misassembly by visualizing read depth of PacBio reads in Parasight (http://eichlerlab.gs.washington.edu/jeff/parasight/index.html) using coverage summaries generated during the resequencing protocol. Sanger assemblies were obtained from NCBI by accession ID (Table S8). *De novo* assembly of short-read data was performed with iCAS (ftp://ftp.sanger.ac.uk/pub/badger/aw7/icas_README).

Illumina Sequencing of BAC Clones. BAC DNA isolation and library preparation was performed as described by Steinberg et al. (2012).

Sequence Alignment. We compared Sanger and PacBio assemblies for each clone using BLASR (Chaisson and Tesler 2012) (-maxLCPLength 16 -bestn 1 -m 0) and visualized these for larger structural rearrangements using Miropeats (Parsons 1995). Alignment identity was calculated from the total number of single-base-pair matches between assemblies divided by the total number of contiguous mismatch events, including substitutions, insertions, and deletions. From the BLASR alignments, we determined the coordinates for each mismatch in both assemblies to create a set of PacBio and Sanger variant pairs. We annotated a subset of these variants that qualified as components of homopolymers, dipolymers, or GC-rich regions based on the context of their adjacent bases. We identified the corresponding regions for the two chimpanzee supercontigs in panTro4 using NCBI's default MEGABLAST settings and aligned the resulting sequences to the supercontigs with BLASR and Miropeats (s = 1000).

To validate the differences we observed between PacBio and Sanger assemblies, we mapped 76 bp Illumina reads from each BAC to both assemblies and chose the variant in each difference that was unambiguously supported by the short reads. Clone pools were sequenced to high coverage using the Nextera protocol described above and mapped with mrsFAST 2.4.0.4 in single-end mode with an edit distance of zero to ensure that only reads with perfect matches

counted as support for variants. For a variant to be supported by the short reads, we required at least one read to span the variant and anchor in sequence that was neither homopolymer nor dipolymer. If one variant in a difference had short-read support and the alternate variant did not, the variant with support was considered validated. In the case where neither or both variants in a difference had support, the difference was considered ambiguous. We performed the same experiment with whole-genome sequence (WGS) from three high-coverage individuals (NA12891, NA18507, and NA18508). Differences between calls from BAC and WGS reads were attributed to potential cell-line variants.

DATA ACCESS

All sequence assemblies are publicly available in GenBank through accessions AC254814-AC254826. Accessions are linked to clone name in Tables S7 & S8.

FIGURE LEGENDS

Figure 1: 17q21.31 genomic target region. a) Tiling path of eight large-insert BAC clones sequenced and assembled using both PacBio- and Sanger-based approaches. Clones were selected from a haploid complete hydatidiform mole source (CH17). b) Gene annotation (RefSeq) and segmental duplication organization was obtained from GRCh37 using a custom liftover coordinate conversion tool that accounted for the difference in copy number between the mole haplotype and the reference. c) Alignment of supercontigs built from the same eight clones using PacBio and Sanger assemblies. Sequence differences (vertical blue lines) and internal duplications (gray) are shown. The two supercontigs are 99.99% identical, excluding a collapsed higher-order repeat at the end of the PacBio assembly of CH17-41F14.

Figure 2: Concordant and discordant PacBio assemblies. a) Alignment between PacBio (top) and Sanger (bottom) assemblies for CH17-227A2 using Miropeats (Parsons 1995) shows virtually no differences. Note the uniform sequence coverage between 200-300 fold. Mismatches/indels are indicated by vertical blue lines. b) Alignment between PacBio and Sanger assemblies for clone CH17-41F14. A spike of increased sequence coverage across the internal repeat and the reduced complexity of the repeat compared to the Sanger assembly clearly define a collapse of a higher-order repeat from 20 kbp to 12 kbp within the PacBio assembly. The uniformity of sequence coverage may be used as one indicator of potential misassembly.

Figure 3: Upgrading a chimpanzee genomic region. Sequence and assembly of six large-insert clones (CH251) from two segmental duplication blocks (red and green) are aligned to their corresponding sequences from the 17p11.2 Smith-Magenis region of the chimpanzee reference assembly (panTro4). Clones were sequenced and assembled from the a) distal and b) proximal segmental duplication blocks. The PacBio assembly was compared to the corresponding working draft sequences from panTro4. The alignment identity of panTro4 contigs without gap sequence and the PacBio supercontigs is 94.69% over 525 kbp of aligned sequence. 31% (241/766 kbp) of

the chimpanzee sequence is missing within the working draft assembly. The average sequence identity for Phred >30 base pairs from BES mappings was 99.72% (16,174/16,220 high-quality bases) and 99.98% (156,955/156,991 high-quality bases) from FES mappings. Gaps in the panTro4 contigs are indicated in red. Gene annotations are shown based on a custom liftover from RefSeq annotations of GRCh37 in the corresponding regions of 17p11.2. The missing sequence corresponds to high-identity segmental duplications (orange bars represent segmental duplications predicted by whole-genome shotgun sequence detection or WSSD). The clone CH251-545A24 was previously sequenced with capillary sequencing (Accession: AC183294).

F**igure 4: Support for chimpanzee supercontig architecture from clone end mappings.** Concordant BES and FES alignments confirm order and orientation of a) distal and b) proximal chimpanzee supercontig assemblies. 125 paired-end sequences that map with >99.8% sequence identity are depicted. Both analyses support high-quality assembly of these complex regions of the chimpanzee genome.

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.

ACKNOWLEDGMENTS

We thank C. Campbell, D. Alexander, and A. Klammer for helpful conversations, K. Mohajeri and L. Harshman for assistance in sample preparation, and T. Brown for assistance in manuscript preparation.

AUTHOR CONTRIBUTIONS

E.E.E. and S.W.T. designed experiments; F.A. prepared DNA; M.M. and S.R. prepared libraries and generated sequence data; P.H.S. and M.Y.D. identified clones for sequencing; J.H., L.H., M.C., T.A.G. and C.A. performed bioinformatics analyses; T.A.G. and R.K.W. performed targeted capillary sequencing of clones; J.H. and E.E.E. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests. S.R., L.H., S.W.T. and J.K. are employees of Pacific Biosciences, Inc., a company commercializing DNA sequencing technologies, and E.E.E. is on the scientific advisory boards for Pacific Biosciences, Inc., SynapDx Corp., and DNAnexus, Inc.

REFERENCES

Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X et al. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* **11**(12): R119.

Alkan C, Cardone MF, Catacchio CR, Antonacci F, O'Brien SJ, Ryder OA, Purgato S, Zoli M, Della Valle G, Eichler EE et al. 2011a. Genome-wide characterization of centromeric satellites from multiple mammalian genomes. *Genome Res* **21**(1): 137-145.

Alkan C, Sajjadian S, Eichler EE. 2011b. Limitations of next-generation genome sequence assembly. *Nat Methods* **8**(1): 61-65.

Au KF, Underwood JG, Lee L, Wong WH. 2012. Improving PacBio long read accuracy by short read alignment. *PLoS One* **7**(10): e46679.

Burton J, Adey A, Patwardhan RP, Qiu R, Kitzman J, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Methods* **in press**.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**: 238.

Charrier C, Joshi K, Coutinho-Budd J, Kim J-E, Lambert N, de Marchena J, Jin WL, Vanderhaeghen P, Ghosh A, Sassa T et al. 2012. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**(4): 923-935.

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods* **10**(6): 563-569.

Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M et al. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLOS Biol* **7**(5): e1000112.

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen HC, Agarwala R, McLaren WM, Ritchie GR et al. 2011. Modernizing reference genome assemblies. *PLoS Biol* **9**(7): e1001091.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H et al. 2012. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**(4): 912-922.

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**(11): e47768.

Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding Y, Buhay CJ, Kremitzki C et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**(7387): 82-86.

Minoche AE, Dohm JC, Himmelbauer H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol* **12**(11): R112.

Okoniewski MJ, Meienberg J, Patrignani A, Szabelska A, Matyas G, Schlapbach R. 2013. Precise breakpoint localization of large genomic deletions using PacBio and Illumina next-generation sequencers. *Biotechniques* **54**(2): 98-100.

Parsons J. 1995. Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci* **11**: 615-619.

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M et al. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res* **22**(3): 557-567.

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG et al. 2005. A common inversion under selection in Europeans. *Nature genetics* **37**(2): 129-137.

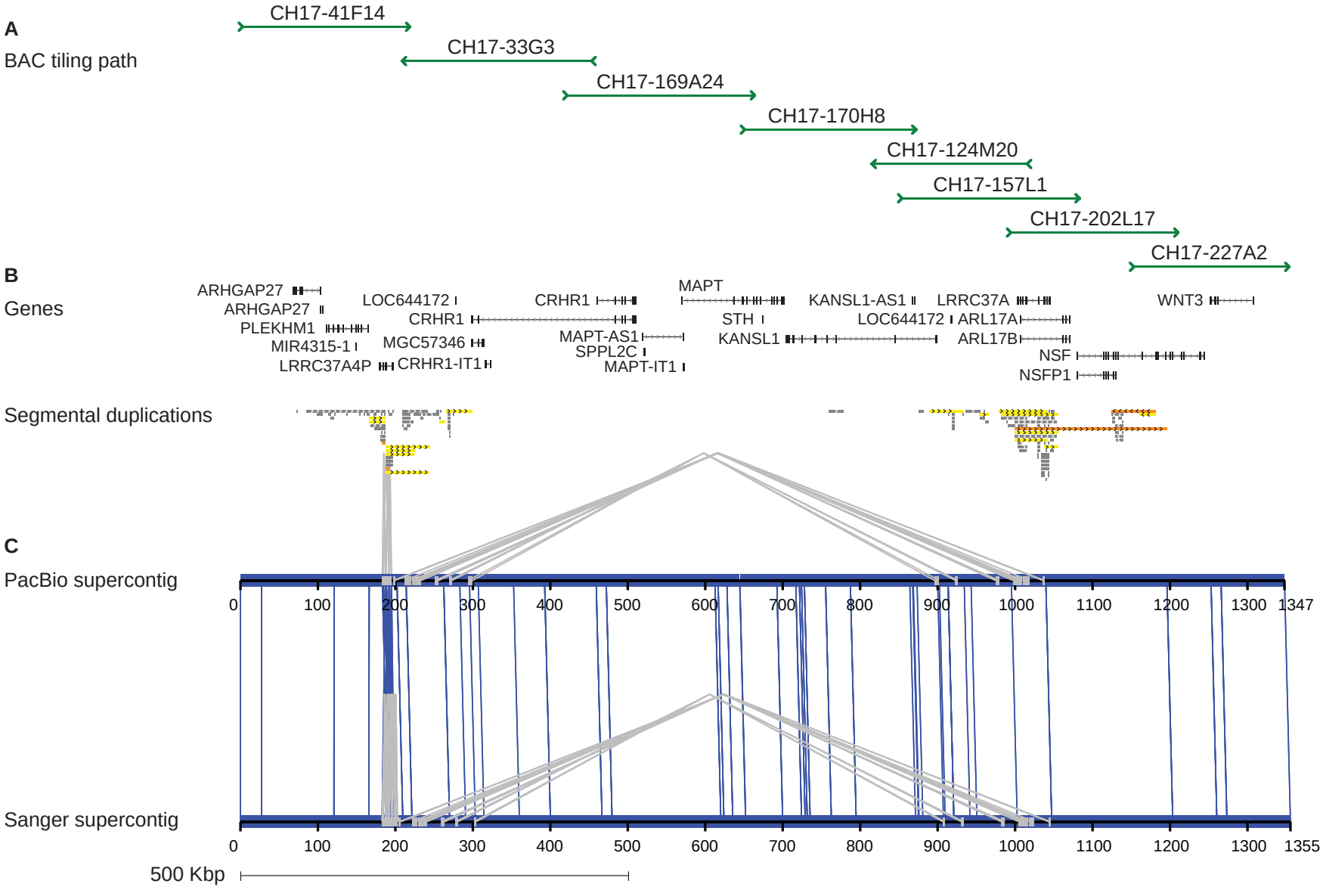
Steinberg KM, Antonacci F, Sudmant PH, Kidd JM, Campbell CD, Vives L, Malig M, Scheinfeldt L, Beggs W, Ibrahim M et al. 2012. Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nature genetics* **44**(8): 872-880.

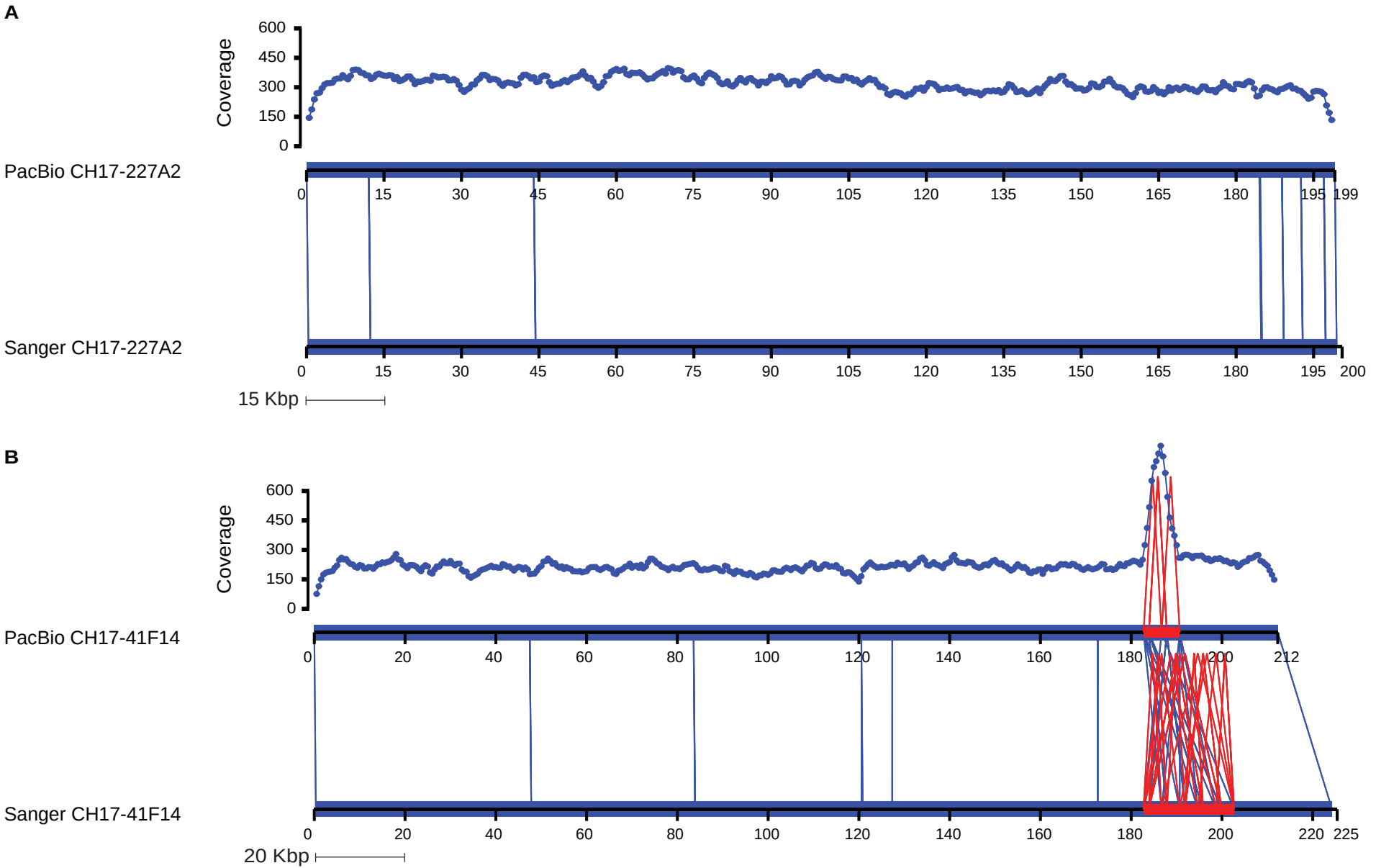
Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, Antonacci F et al. 2013. Evolution and diversity of copy number variation in the great ape lineage. *Genome Res* **23**(9): 1373-1382.

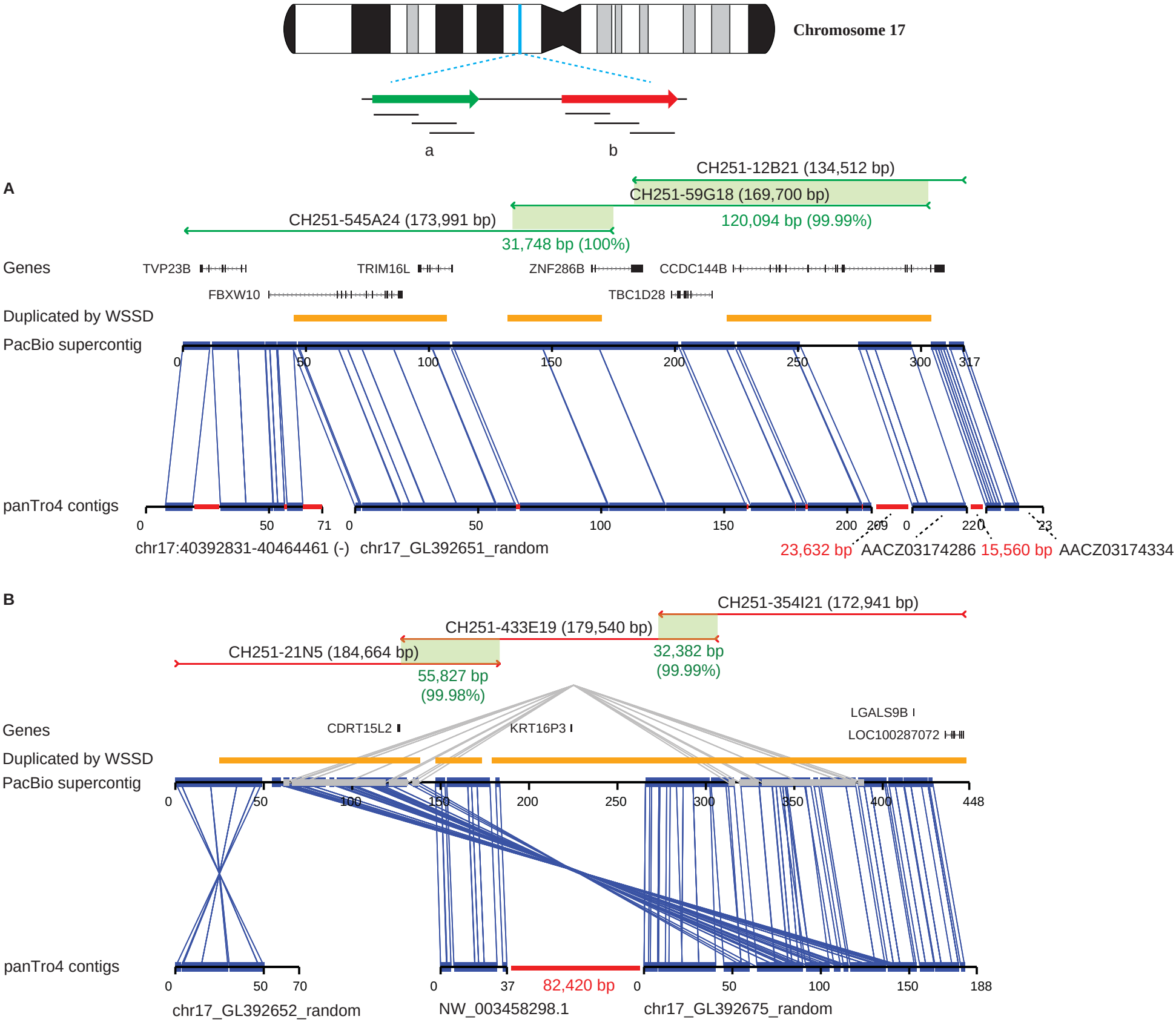
Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Eichler EE. 2010. Diversity of human copy number variation and multicopy genes. *Science* **330**(6004): 641-646.

Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res* **38**(15): e159.

Zody MC, Jiang Z, Fung HC, Antonacci F, Hillier LW, Cardone MF, Graves TA, Kidd JM, Cheng Z, Abouelleil A et al. 2008. Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nature genetics* **40**(9): 1076-1083.


A
BAC tiling path
CH17-41F14
CH17-33G3
CH17-169A24
CH17-170H8
CH17-124M20
CH17-157L1
CH17-202L17
CH17-227A2
B
Genes
ARHGAP27
ARHGAP27
PLEKHM1
MIR4315-1
LRRC37A4P
LOC644172
CRHR1
MGC57346
CRHR1-IT1
CRHR1
MAPT-AS1
SPPL2C
MAPT-IT1
MAPT
STH
KANSL1
KANSL1-AS1
LOC644172
LRRC37A
ARL17A
ARL17B
WNT3
NSF
NSFP1
Segmental duplications
C
PacBio supercontig
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1347
Sanger supercontig
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1355
500 Kbp


A
Coverage
600
450
300
150
0
PacBio CH17-227A2
0 15 30 45 60 75 90 105 120 135 150 165 180 195 199
Sanger CH17-227A2
0 15 30 45 60 75 90 105 120 135 150 165 180 195 200
15 Kbp
B
Coverage
600
450
300
150
0
PacBio CH17-41F14
0 20 40 60 80 100 120 140 160 180 200 212
Sanger CH17-41F14
0 20 40 60 80 100 120 140 160 180 200 220 225
20 Kbp


Chromosome 17
a
b
A
CH251-12B21 (134,512 bp)
CH251-59G18 (169,700 bp)
120,094 bp (99.99%)
CH251-545A24 (173,991 bp)
31,748 bp (100%)
Genes
TVP23B
TRIM16L
ZNF286B
CCDC144B
FBXW10
TBC1D28
Duplicated by WSSD
PacBio supercontig
0
50
100
150
200
250
300
317
panTro4 contigs
0
50
71
0
50
100
150
200
209
0
220
23
chr17:40392831-40464461 (-)
chr17_GL392651_random
23,632 bp
AACZ03174286
15,560 bp
AACZ03174334
B
CH251-354I21 (172,941 bp)
CH251-433E19 (179,540 bp)
32,382 bp
(99.99%)
CH251-21N5 (184,664 bp)
55,827 bp
(99.98%)
Genes
CDRT15L2
KRT16P3
LGALS9B
LOC100287072
Duplicated by WSSD
PacBio supercontig
0
50
100
150
200
250
300
350
400
448
panTro4 contigs
0
50
70
0
37
82,420 bp
0
50
100
150
188
chr17_GL392652_random
NW_003458298.1
chr17_GL392675_random

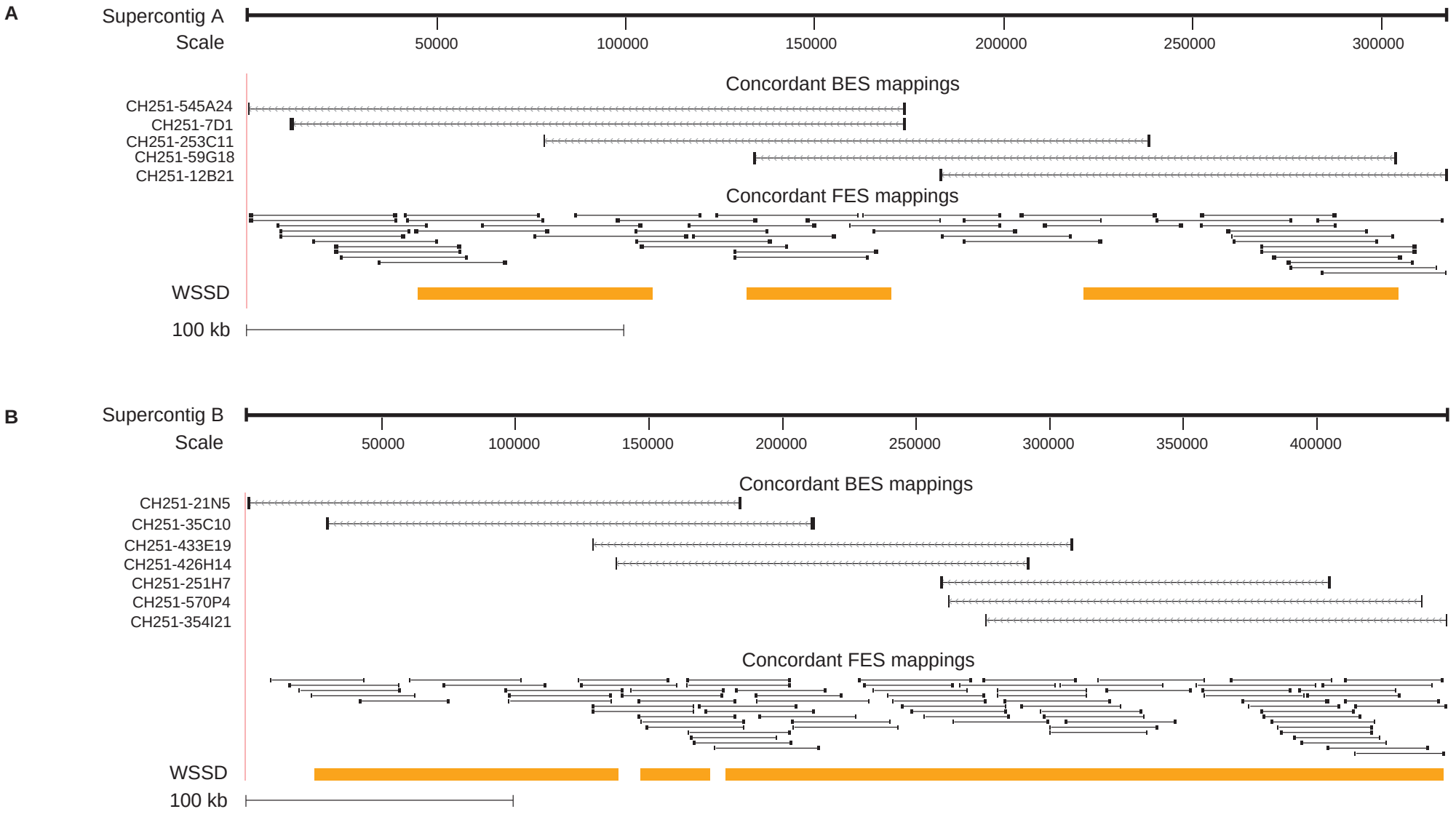

A
Supercontig A
Scale
50000
100000
150000
200000
250000
300000
Concordant BES mappings
CH251-545A24
CH251-7D1
CH251-253C11
CH251-59G18
CH251-12B21
Concordant FES mappings
WSSD
100 kb
B
Supercontig B
Scale
50000
100000
150000
200000
250000
300000
350000
400000
Concordant BES mappings
CH251-21N5
CH251-35C10
CH251-433E19
CH251-426H14
CH251-251H7
CH251-570P4
CH251-354I21
Concordant FES mappings
WSSD
100 kb

Table 1. CH17 clone summary

Clone	Duplications[a] (Kbp)	PacBio coverage	Illumina coverage	Sanger size (bp)	PacBio size (bp)	SMRT cells	Contigs
CH17-124M20	184	475	117	202,892	202,859	1	1
CH17-157L1	210	186	95	230,865	230,921	1	1
CH17-169A24	0	78	68	243,129	242,237	1	1
CH17-170H8	74	240	91	223,520	222,143	5[b]	1
CH17-202L17	204	263	95	217,579	217,211	1	1
CH17-227A2	110	312	109	200,520	201,802	1	1
CH17-33G3	87	177	82	244,867	244,942	2[b]	1
CH17-41F14	123	226	92	225,391	212,292	1	1

[a] Duplications annotated by DupMasker

[b] Sequenced to higher coverage due to contamination in DNA library

Table 2. Summary of alignments between PacBio and Sanger assemblies

Clone	PacBio coverage	Matches[a]	Substitutions[b]	PacBio Insertions[c]	Sanger Insertions[d]	Mismatches[e]	Per-base Identity[f]	Per-event Identity[g]
CH17-124M20	475	202,813	0	15 (10)	4 (4)	19 (14)	0.999906	0.999931
CH17-157L1	186	230,782	0	12 (10)	3 (3)	15 (13)	0.999935	0.999944
CH17-169A24	78	243,011	18 (16)	27 (13)	6 (6)	51 (35)	0.999790	0.999856
CH17-170H8	240	223,424	0	13 (12)	0	13 (12)	0.999942	0.999946
CH17-202L17	263	217,482	0	2 (2)	1	3 (3)	0.999986	0.999986
CH17-227A2	312	200,447	0	0	3 (3)	3 (3)	0.999985	0.999985
CH17-33G3	177	244,778	2 (2)	5 (4)	4 (4)	11 (10)	0.999955	0.999959
CH17-41F14	226	217,376	4 (4)	349 (2)	7,991 (4)	8,344 (10)	0.963034	0.999954

[a] Matching bases determined by BLASR alignment of PacBio and Sanger assemblies

[b, c, d, e] Total differences between assemblies by base and by unique event in parentheses

[f] % identity between assemblies based on total matches divided by matches plus mismatch bases

[g] % identity between assemblies based on total matches divided by matches plus mismatch events

Table 3. Total mismatches between assemblies validated by Illumina reads

Clone	Total mismatches[a]	PacBio supported[b]	Sanger supported[c]	Ambiguous[d]	Homopolymer	Dipolymer	GC rich
CH17-124M20	19	2	1	16	11	7	3
CH17-157L1	15	0	2	13	11	1	2
CH17-169A24	51	24	6	21	2	3	2
CH17-170H8	13	0	0	13	11	0	0
CH17-202L17	3	0	0	3	1	1	0
CH17-227A2	3	0	0	3	2	0	0
CH17-33G3	11	3	0	8	6	0	2
CH17-41F14	10	2	4	4	3	1	0
Total	125	31	13	81	47	13	9

[a] Total base pair mismatches between assemblies in events < 50 bp

[b] PacBio bases with more support by Illumina reads than the corresponding Sanger bases

[c] Sanger bases with more support by Illumina reads than the corresponding PacBio bases

[d] Mismatch bases that had no Illumina support for either assembly or support for both assemblies

Reconstructing complex regions of genomes using long-read sequencing technology

John Huddleston, Swati Ranade, Maika Malig, et al.

Genome Res. published online January 13, 2014
Access the most recent version at doi:10.1101/gr.168450.113