SUPPLEMENTARY FIGURES

SUPPLEMENTARY TABLES

SUPPLEMENTARY FILES

Supplemental Methods

File S1: Sample Celera 8.3 MHAP specification file, with default small/haploid genome parameters. Parameters altered in this study were merSize, -k, --num-hashes, and assembleCoverage.

File S2: Sample Celera 8.3 MHAP specification file, with large/diploid genome parameters. Parameters altered in this study were merSize, -k, --num-hashes, and assembleCoverage.

File S3: Sample specification file used to run Canu assembler with 4% error rate

File S4: Sample specification file used to run FALCON

File S5: Spec file used to construct BLASR-corr Cel8.3 assembly

File S6: SLURM script used to construct BLASR-corr Cel8.3 assembly. Note: this script did not appear to properly allocate resources on our cluster, resulting in a long (~17 days) assembly time. Configured properly, assembly should be much faster.

File S7: SLURM file used to run Canu 1.2 using BLASR corrected reads (Canu-corr assembly). This uses the default Canu settings but skips read correction.

File S8: SLURM job handler file used to to run Celera 8.3 assembler

File S9: Custom Repbase repeat library used to annotate assemblies.

File S10: Perl script used to annotate assembly from BLAST output

File S11: GFF annotation file for the major Rsp locus in the PBcR-BLASR assembly, constructed using custom scripts.

File S12: GFF annotation file for the 1.688 loci in the PBcR-BLASR assembly, constructed using custom scripts.

File S13: GFF annotation file for the minor Rsp locus in the PBcR-BLASR assembly, constructed using custom scripts.

Figure S1. *Rsp*, *Bari1* and *AAGAG* clusters on chromosome *2R*

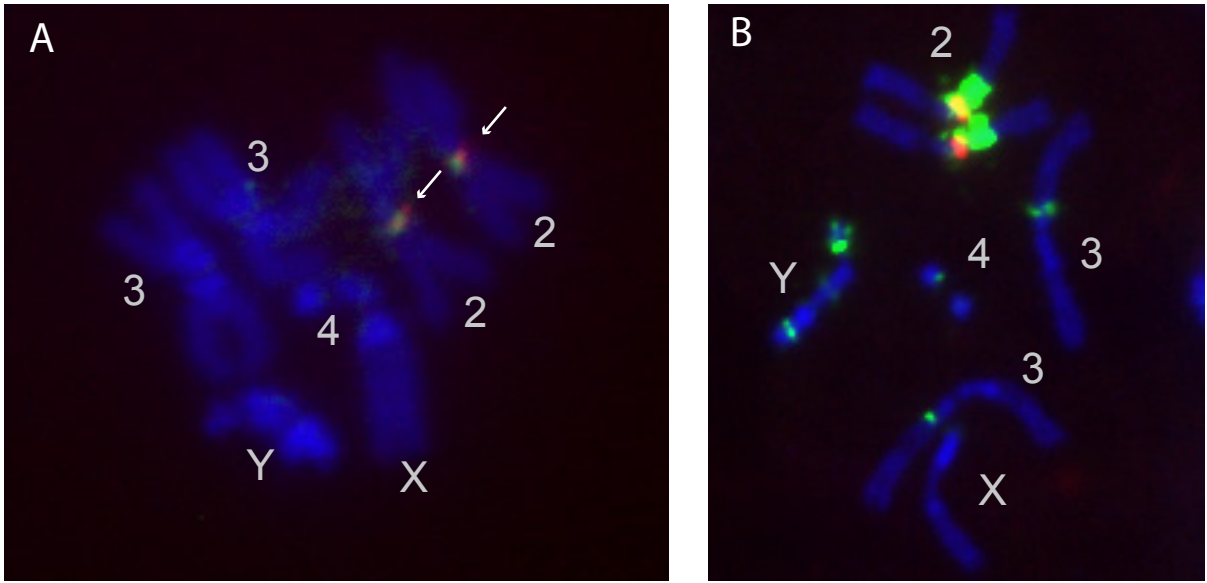


Fig S1. FISH image of *D. melanogaster* mitotic chromosomes from 3rd instar larvae. DNA is stained with DAPI (blue). A. *Rsp* (red) and *Bari1* (green) are close together on chromosome *2R*. B. *AAGAG* (green) is widespread throughout the genome. On chromosome 2, it spans the centromere and is distal to *Rsp* on chromosome *2R*.

Figure S2. Molecular validation of *Rsp* locus

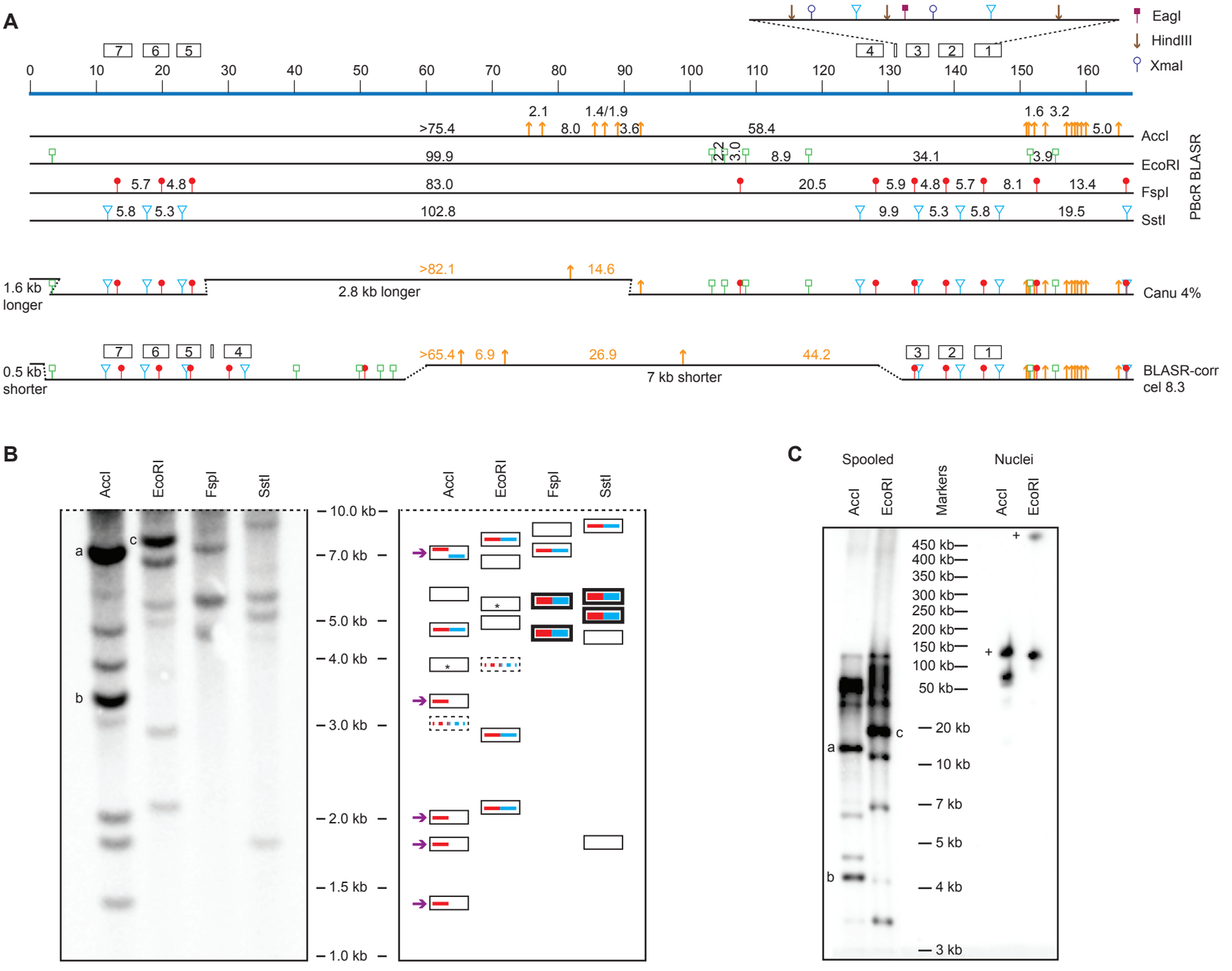


Figure S2: A. Schematic showing the extent of the assembled *Rsp* array (blue line as in Figure 2A) as well as the approximate locations of restriction endonuclease sites (*AccI, EcoRI, FspI, SstI*) within the PBcR-BLASR, Canu 4% and BLASR-corr cel 8.3 assemblies. Predicted fragment sizes based on *in silico* digestion of the *Rsp* locus in the PBcR-BLASR assembly in Geneious are also indicated for each of these enzymes. Because there are additional *Rsp* repeats proximal (left on the diagram, see text) to the assembled array, it is not possible to determine the size of the proximal fragment(s) for any of the restriction enzymes (e.g. a proximal *AccI* fragment is greater than 75 kb). The position of *G5 Jockey* repeats are indicated by the numbered boxes. The location of the15-kb PCR product and the positions of the restriction enzyme sites used to produce Figure 2A are shown above the blue line. B. Genomic blot and schematic interpretation as described in Figure 2B except the *AccI* bands inconsistent with the Canu 4% assembly are also indicated (arrows). Sizes of the hybridizing bands can be compared to the *in silico* predictions in (A). Select bands are indicated with a letter (a,b,c) and their sizes compared to the same fragments in (C). C. Pulse-field gel of spooled DNA (left) and high molecular weight DNA (nuclei, right) digested with the indicated restriction enzymes and probed with *Rsp*. The intensity of the bands labeled a,b, c (left) anchor the pattern seen in (C), but the *Rsp* fragments appear to run slower under the pulse field conditions. For the lanes on the right, the size of the smaller fragment in each digest can be compared to the *in silico* prediction. + unpredicted band which may represent hybridization to repeats proximal to those in the assembly.

Fig. S3: Major *Rsp* locus in Canu 4% assembly

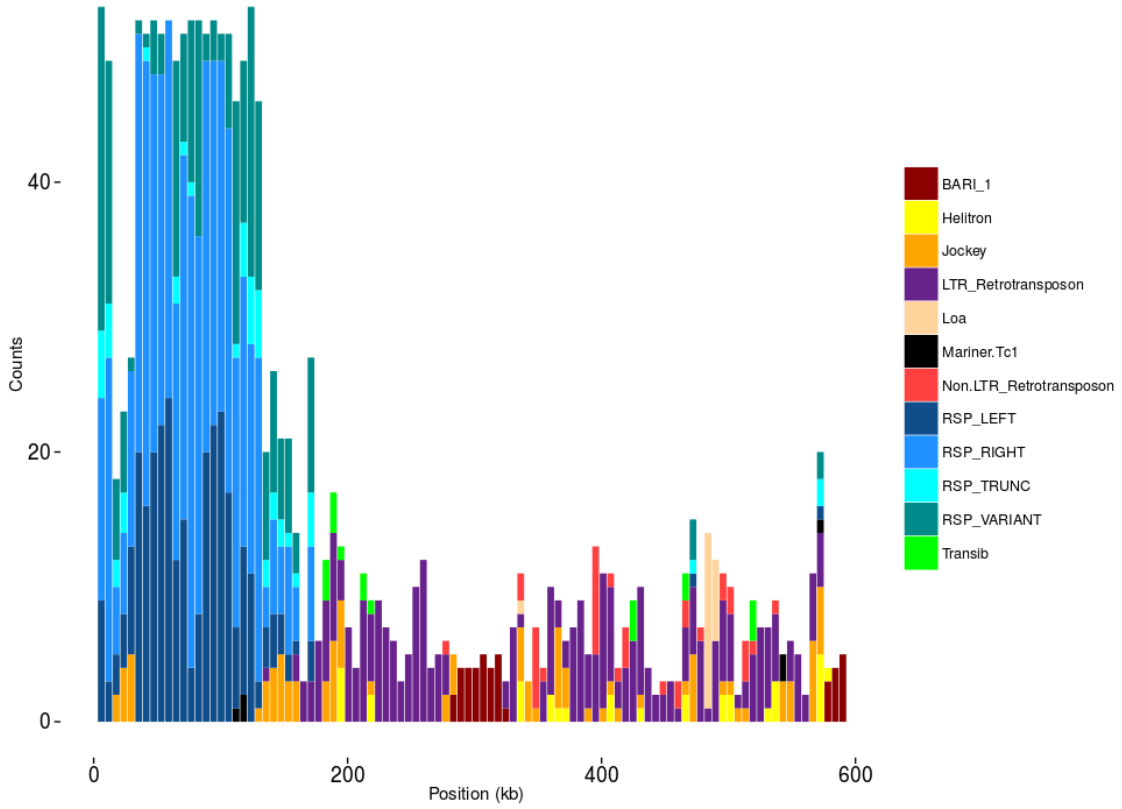


Figure S3: Major *Rsp* locus in Canu 4% assembly. Counts for each element in our custom Repbase library in 5-kb windows across the *Rsp* locus in our Canu 4% assembly. The main block of *Rsp* repeats is approximately equal in repeat number to the PBcR-BLASR assembly, and orientation of the repeats is supported by restriction digest and Southern blot analysis. The contig extends the PBcR-BLASR contig ~300 kb distally, which includes an additional 10 variant *Rsp* repeats (minor *Rsp* locus) as well as the *Bari1* clusters. These additional repeats are also present in the PBcR-BLASR assembly, but on a separate contig.

Figure S4. Coverage of deep Illumina reads over *Rsp* locus in the MHAP 20_1500_25X assembly

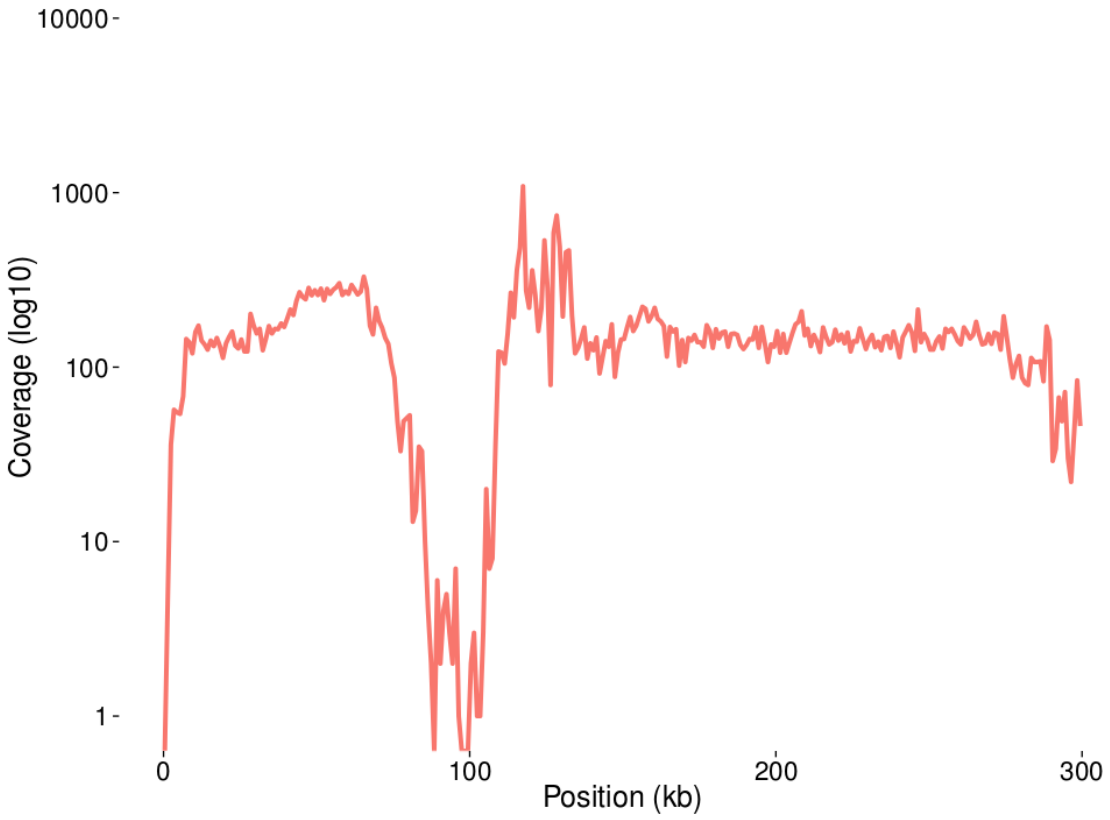


Figure S4: Coverage of deep Illumina reads (SRA accession ERR701706) over *Rsp* locus in MHAP 20_1500_25X assembly mapped using the –very-sensitive settings in Bowtie2. Coverage was plotted in 1-kb windows using Bedtools and is shown on a log scale.

Fig. S5: Coverage of deep Illumina reads over *Rsp* contig in Canu 4% assembly

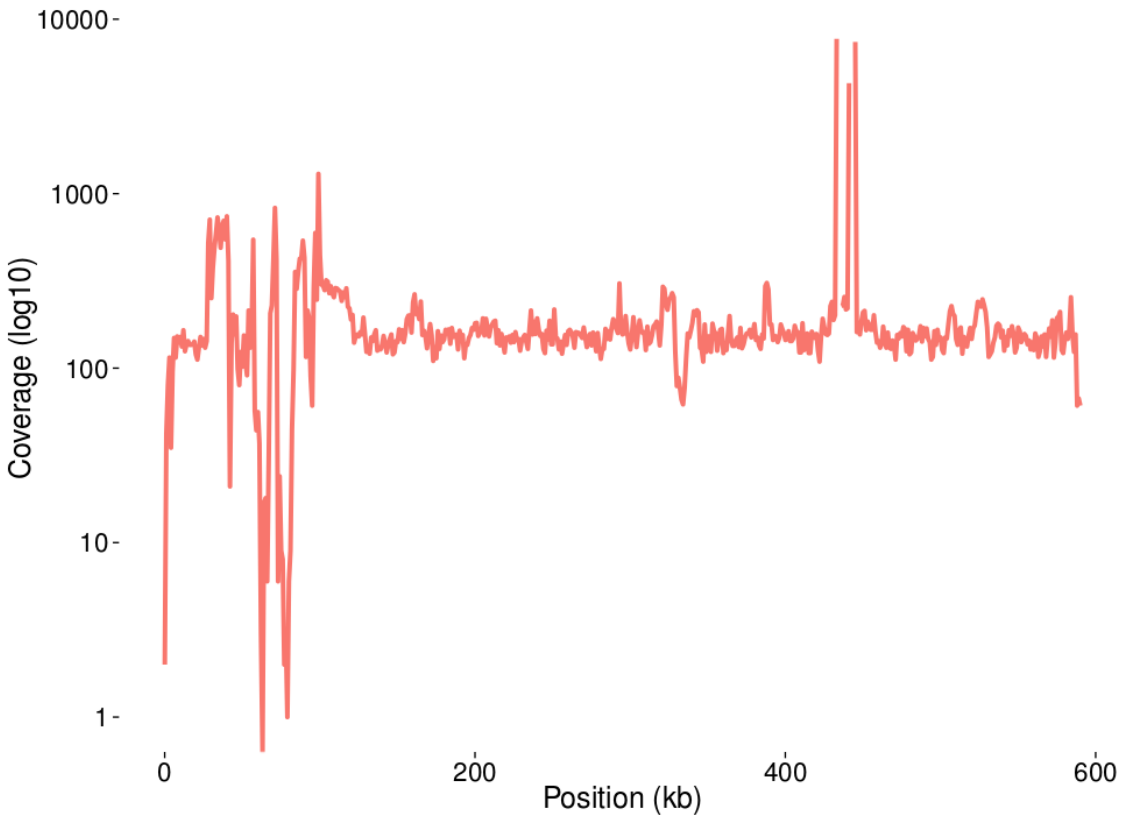


Figure S5: Coverage of deep Illumina reads (SRA accession ERR701706) over *Rsp* locus in Canu 4% error-rate assembly mapped using –very-sensitive settings in Bowtie2. Coverage was plotted in 1-kb windows using Bedtools and is shown on a log scale.

Figure S6. Coverage of deep Illumina reads over *Rsp* locus in PBcR-BLASR assembly

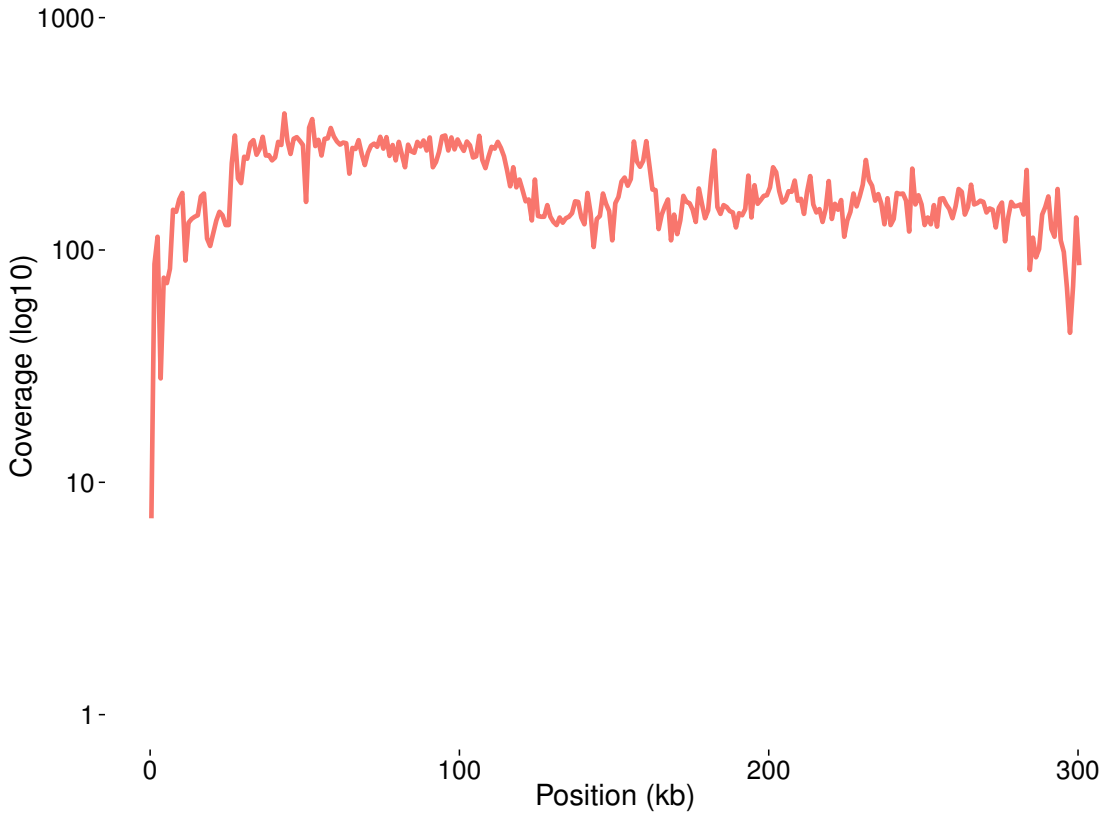


Figure S6: Coverage of deep Illumina reads (SRA accession ERR701706) over *Rsp* locus in PBcR-BLASR assembly mapped using the –very-sensitive settings in Bowtie2. Coverage was plotted in 1-kb windows using Bedtools and is shown on a log scale.

Figure S7. Coverage of raw PacBio reads mapped against PBcR-BLASR assembly

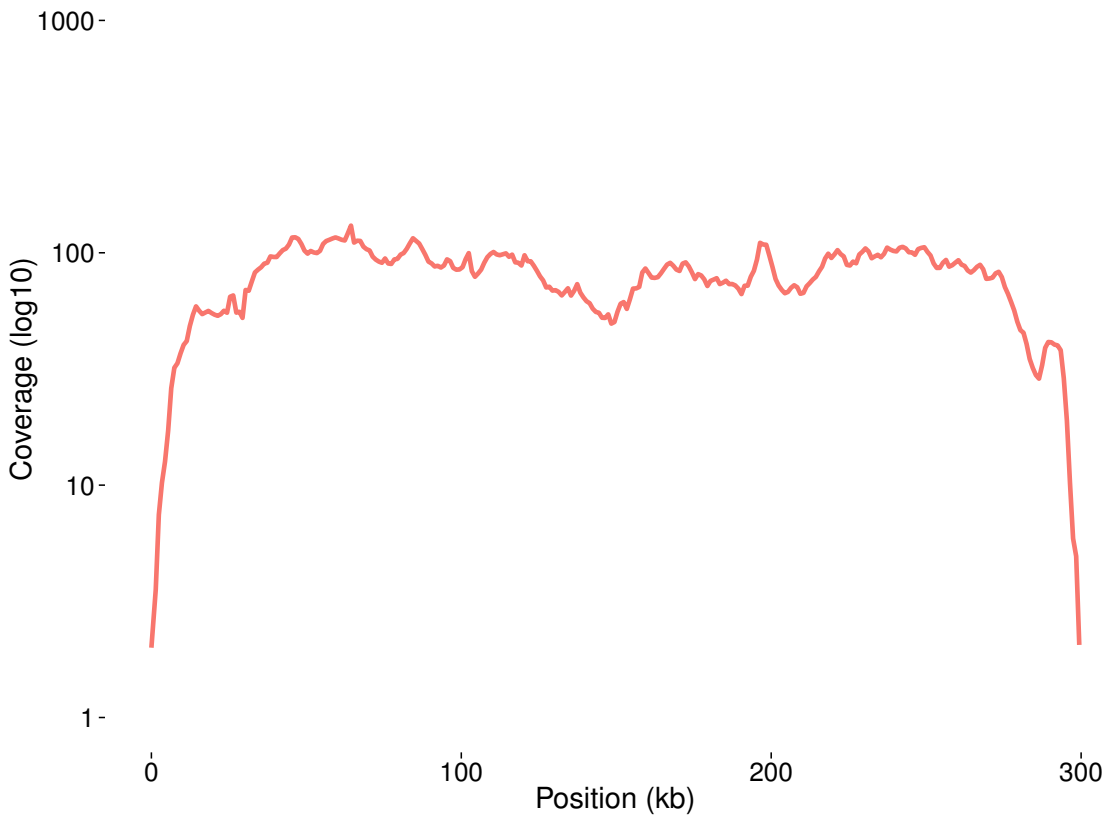


Figure S7: Coverage of raw PacBio reads mapped against PBcR-BLASR assembly. Reads were mapped using BLASR included in SMRT Analysis v. 2.3. Coverage shown on a log(10) scale.

Figure S8: Coverage of raw PacBio reads mapped against BLASR-corr Cel8.3 assembly

Figure S8: Coverage of raw PacBio reads mapped against BLASR-corr Cel8.3 assembly. Reads were mapped using BLASR included in SMRT Analysis v. 2.3. Coverage shown on a log(10) scale.

Figure S9. Mummer dotplot alignment of *Rsp* loci

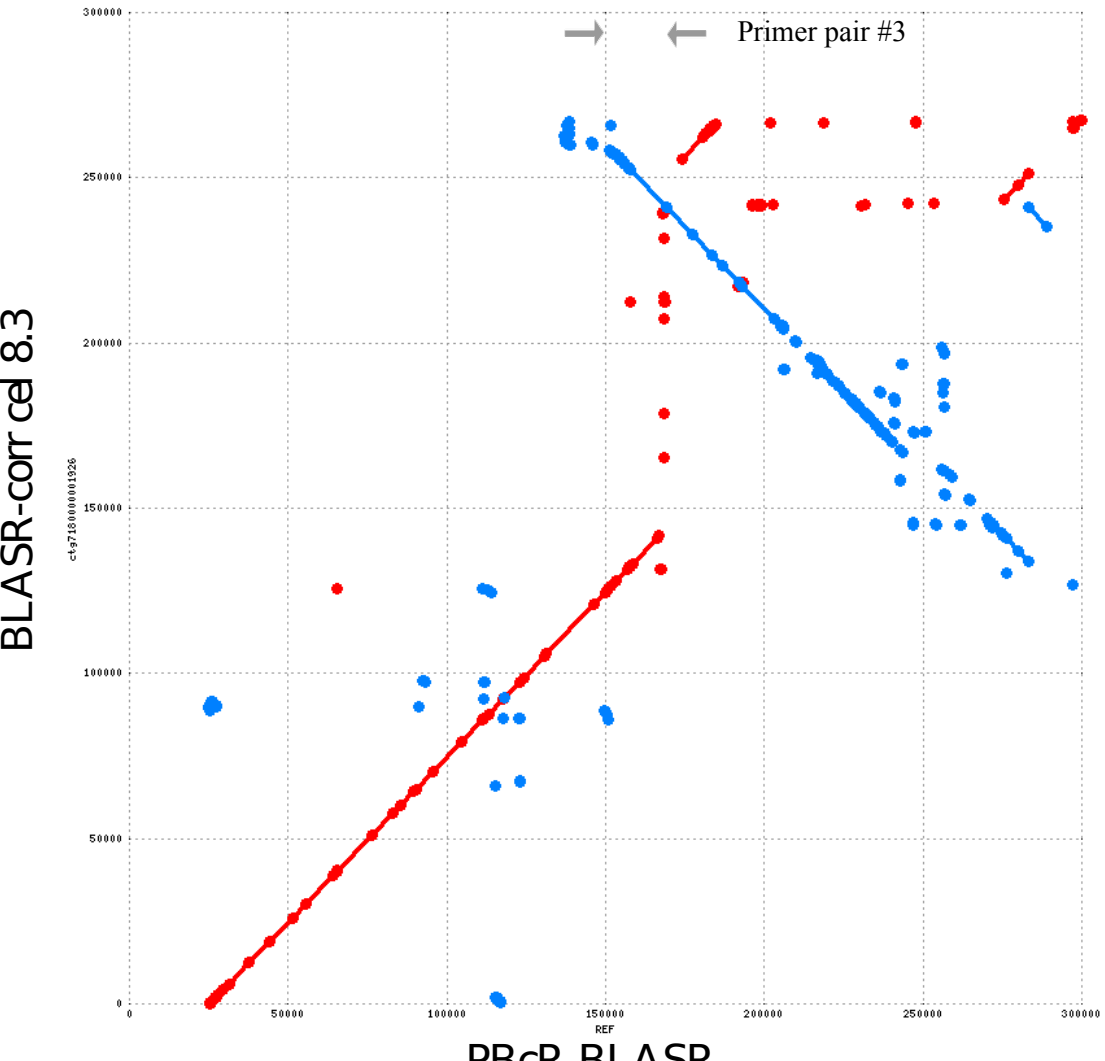


Figure S9: Mummer dotplot alignment of *Rsp* loci contigs from the two best-supported assemblies (PBcR-BLASR and BLASR-corr Cel8.3). Red points indicate identical strand, blue indicates opposite strand. Approximate locations of long PCR primers used to verify the PBcR-BLASR orientation are indicated with grey arrows.

Fig. S10: Nucleotide substitution error distribution

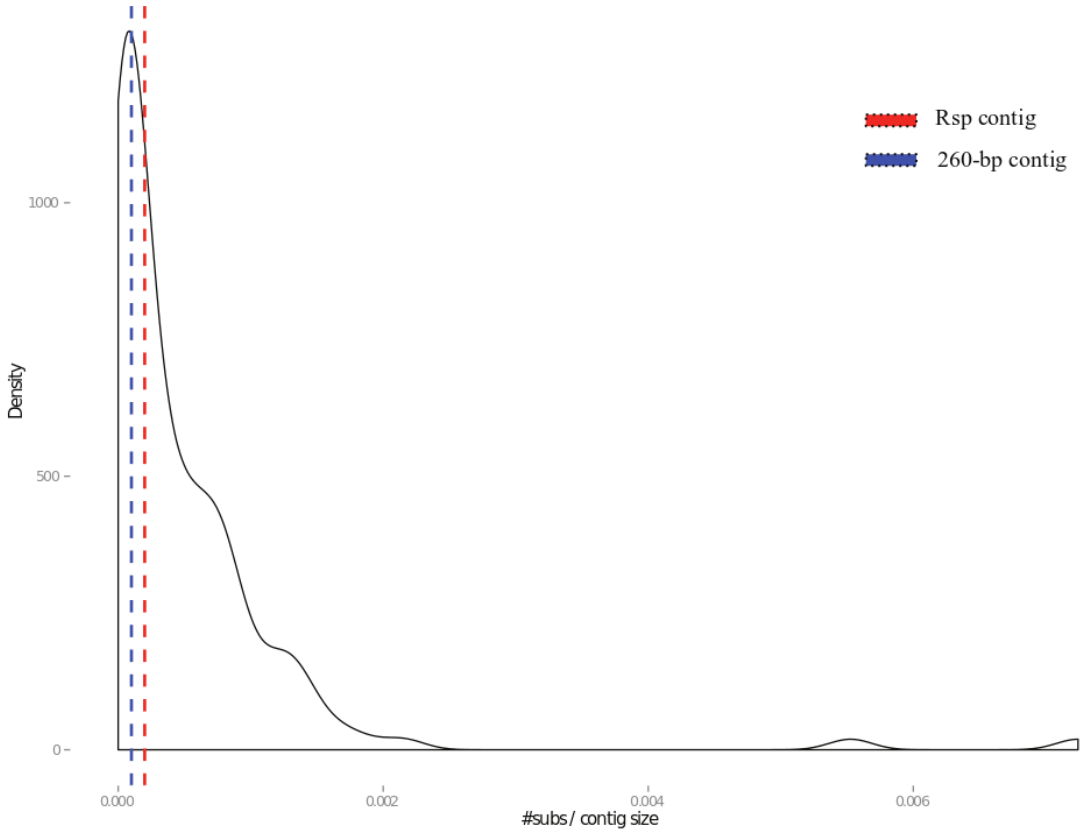


Figure S10: Density plot showing distribution of nucleotide substitutions fraction for each contig in the PBcR-BLASR assembly. Illumina reads were mapped to the assembly using Bowtie2, and Pilon was used to generate variant call format (vcf) files. SNPs for each contig were summed and divided by the total contig length.

Fig S11. Indel distribution

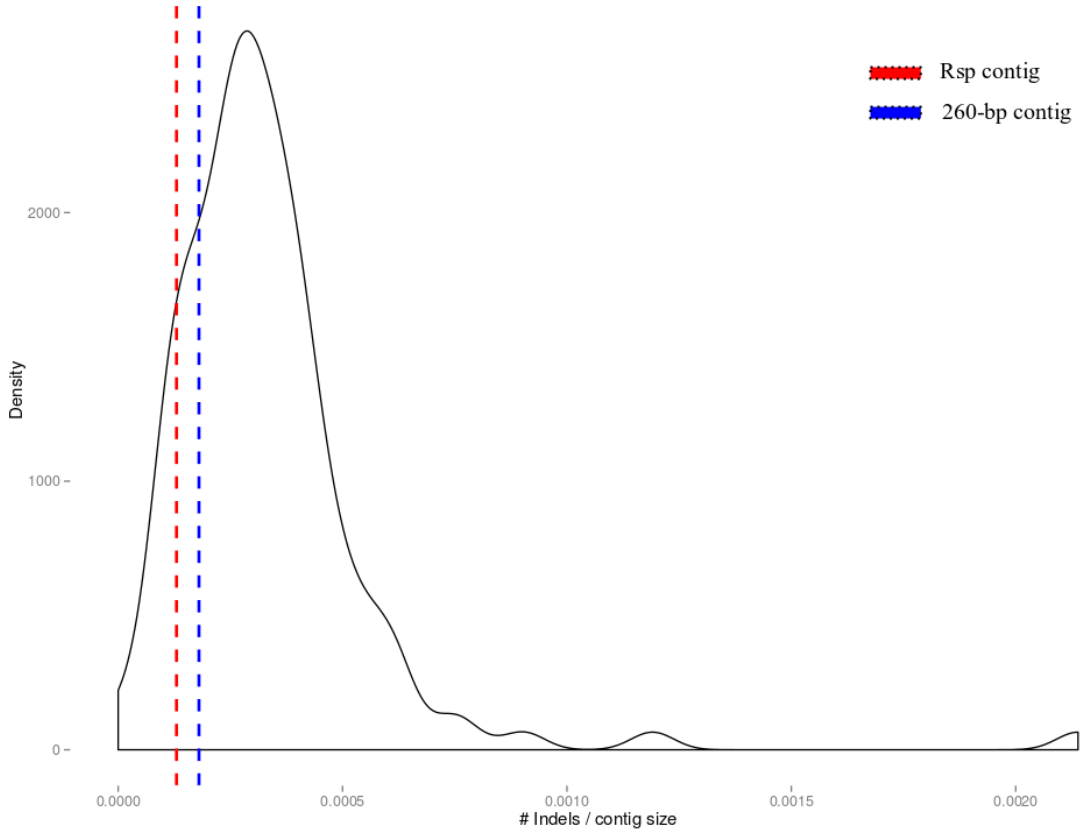


Figure S11: Density plot showing distribution of indel fraction for each contig in the PBcR-BLASR assembly. Illumina reads were mapped to the assembly using Bowtie2, and Pilon was used to generate variant call format (vcf) files. Indels for each contig were summed and divided by the total contig length.

Figure S12: R6 vs PBcR-BLASR assembly comparison

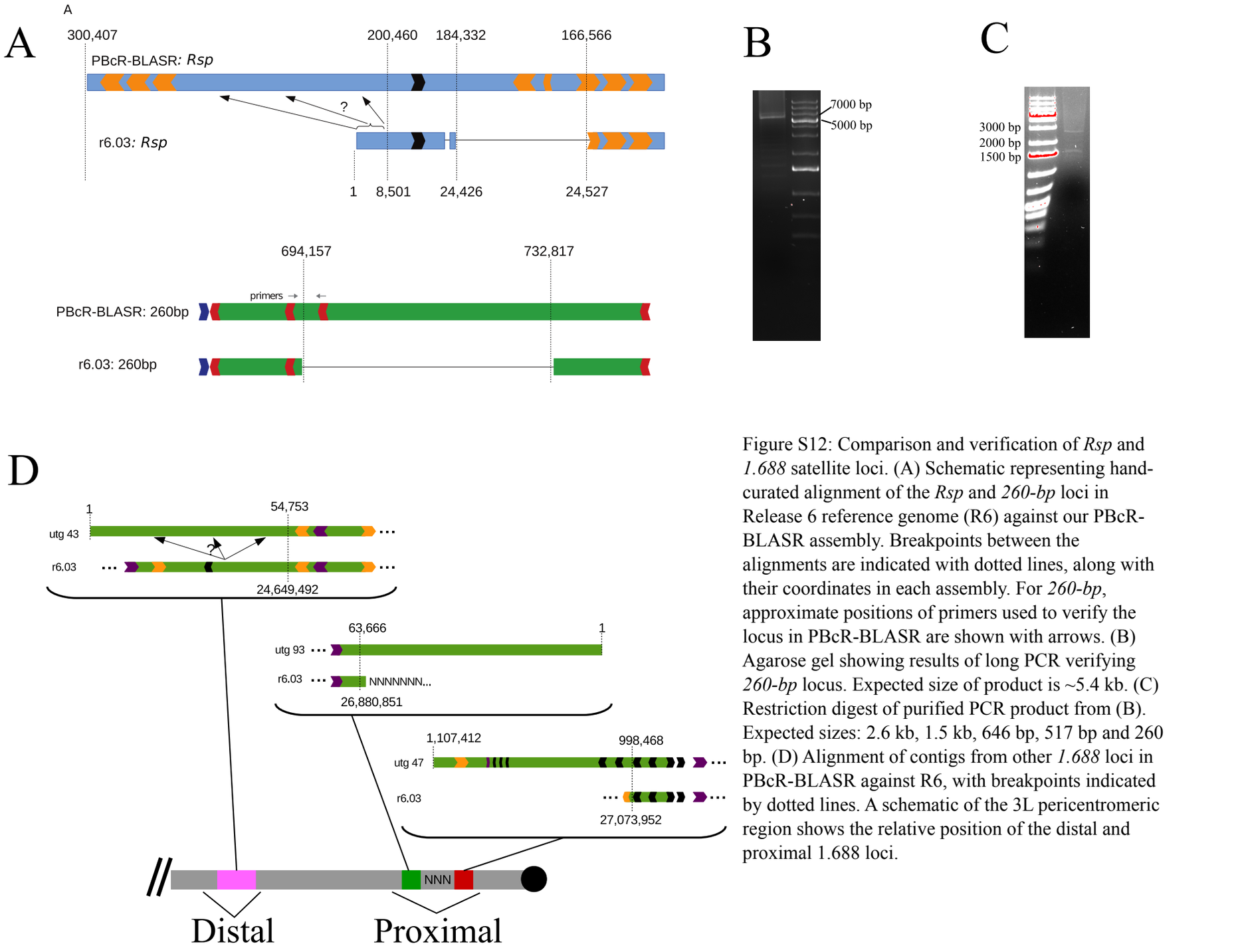


Figure S12: Comparison and verification of *Rsp* and *1.688* satellite loci. (A) Schematic representing hand-curated alignment of the *Rsp* and *260-bp* loci in Release 6 reference genome (R6) against our PBcR-BLASR assembly. Breakpoints between the alignments are indicated with dotted lines, along with their coordinates in each assembly. For *260-bp*, approximate positions of primers used to verify the locus in PBcR-BLASR are shown with arrows. (B) Agarose gel showing results of long PCR verifying *260-bp* locus. Expected size of product is ~5.4 kb. (C) Restriction digest of purified PCR product from (B). Expected sizes: 2.6 kb, 1.5 kb, 646 bp, 517 bp and 260 bp. (D) Alignment of contigs from other *1.688* loci in PBcR-BLASR against R6, with breakpoints indicated by dotted lines. A schematic of the 3L pericentromeric region shows the relative position of the distal and proximal 1.688 loci.

Figure S13. Map of centromere-proximal *G2* contig

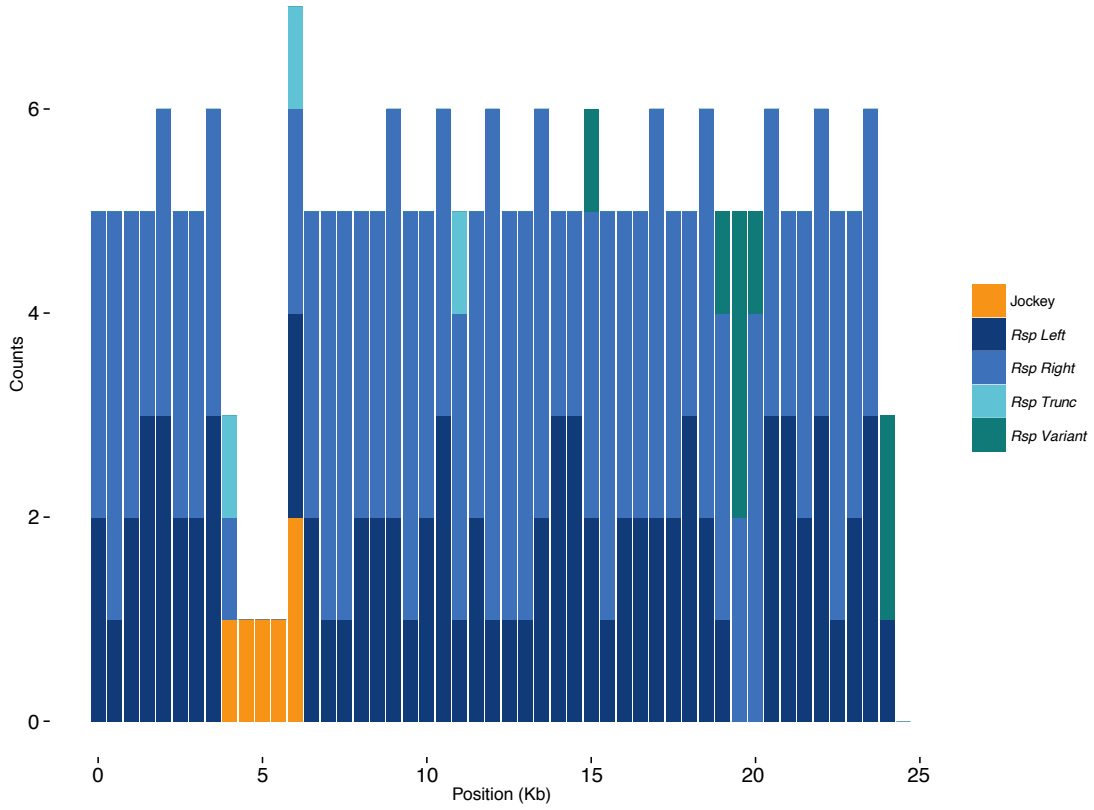


Figure S13: Map of the centromere-proximal *G2-Rsp* contig assembled using corrected PacBio reads. Counts for each repetitive elements in our custom Repbase library are plotted in 500 bp windows across the contig.

Fig S14: *G5* phylogeny

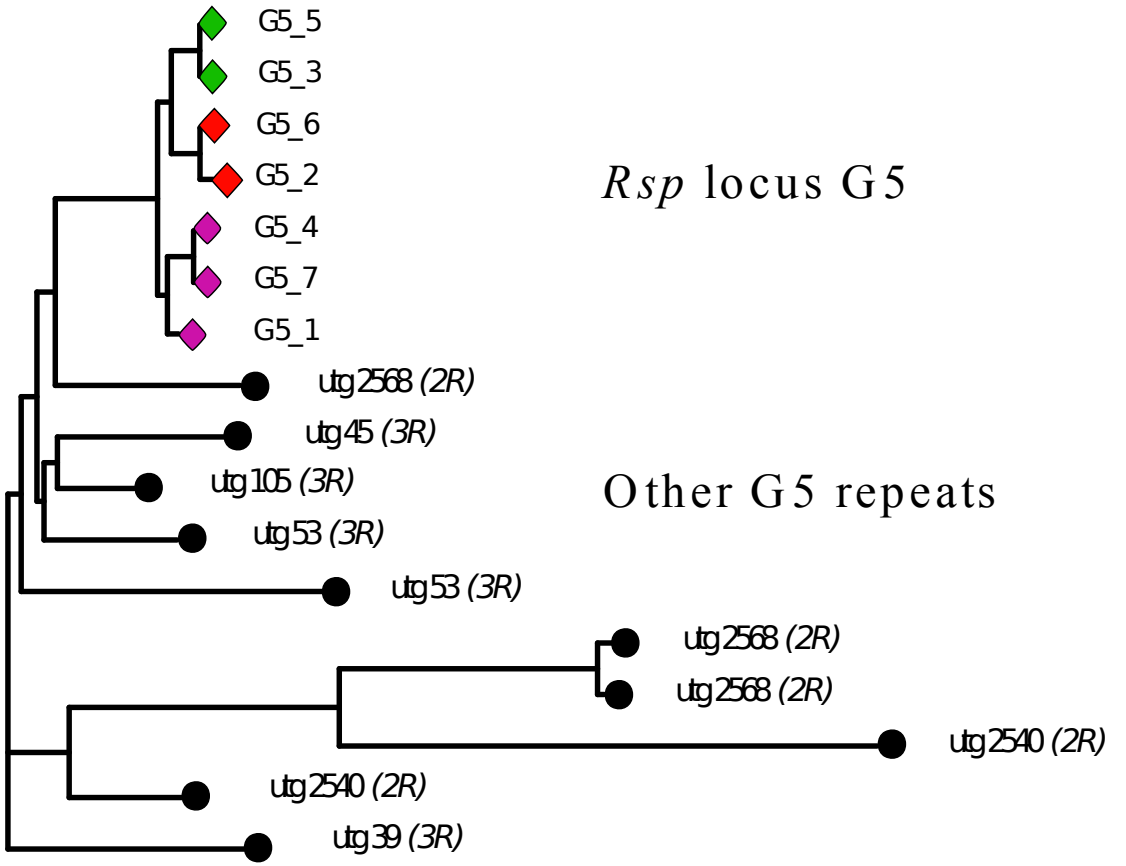
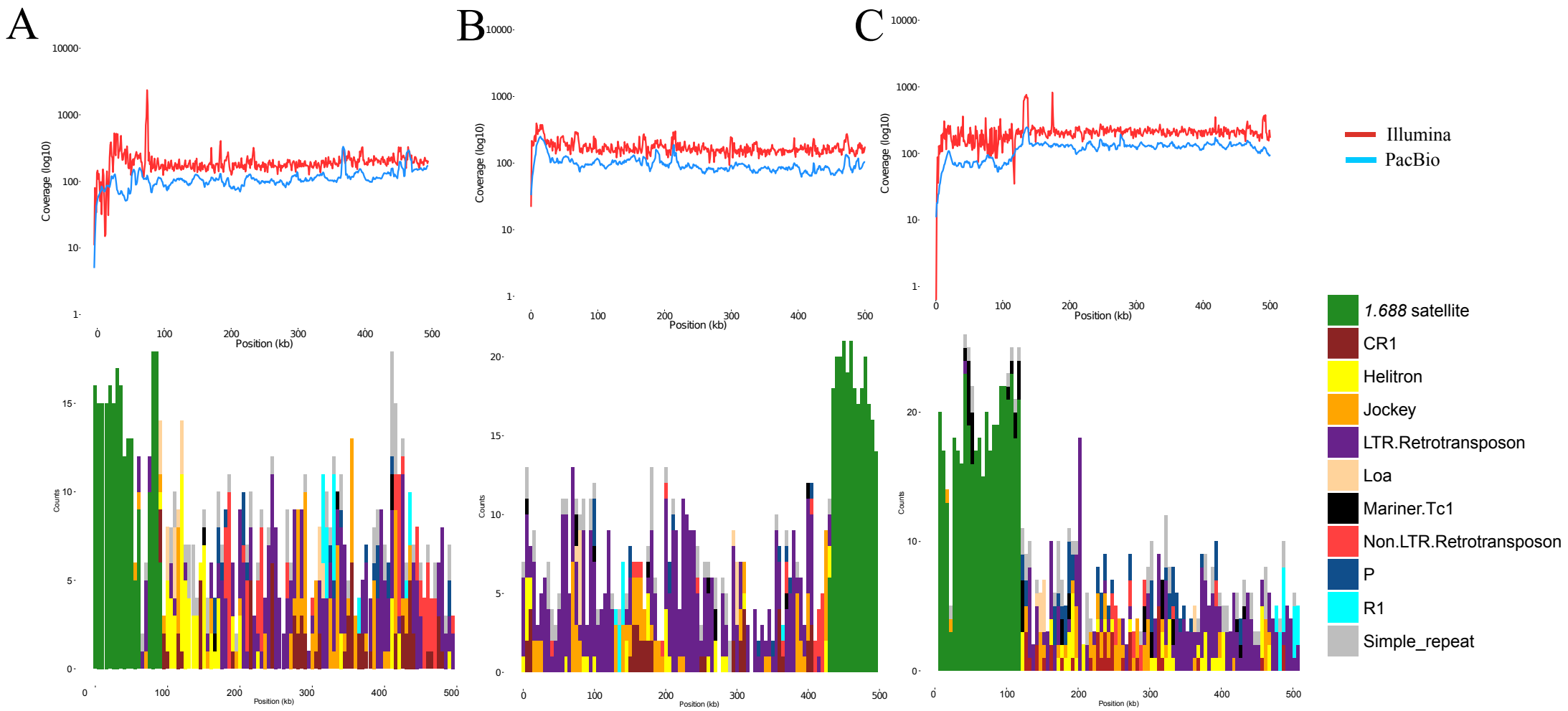


Figure S14: RaxML tree of *G5* repeats at *Rsp* locus and all other *G5* repeats > 1000bp in the PBcR-BLASR assembly. Tree was built in Geneious with 100 bootstraps. The other *G5* repeats on *2R* are all at least 2 Mb distal to the *Rsp* locus *G5* cluster.

Figure S15: structure of *1.688* loci



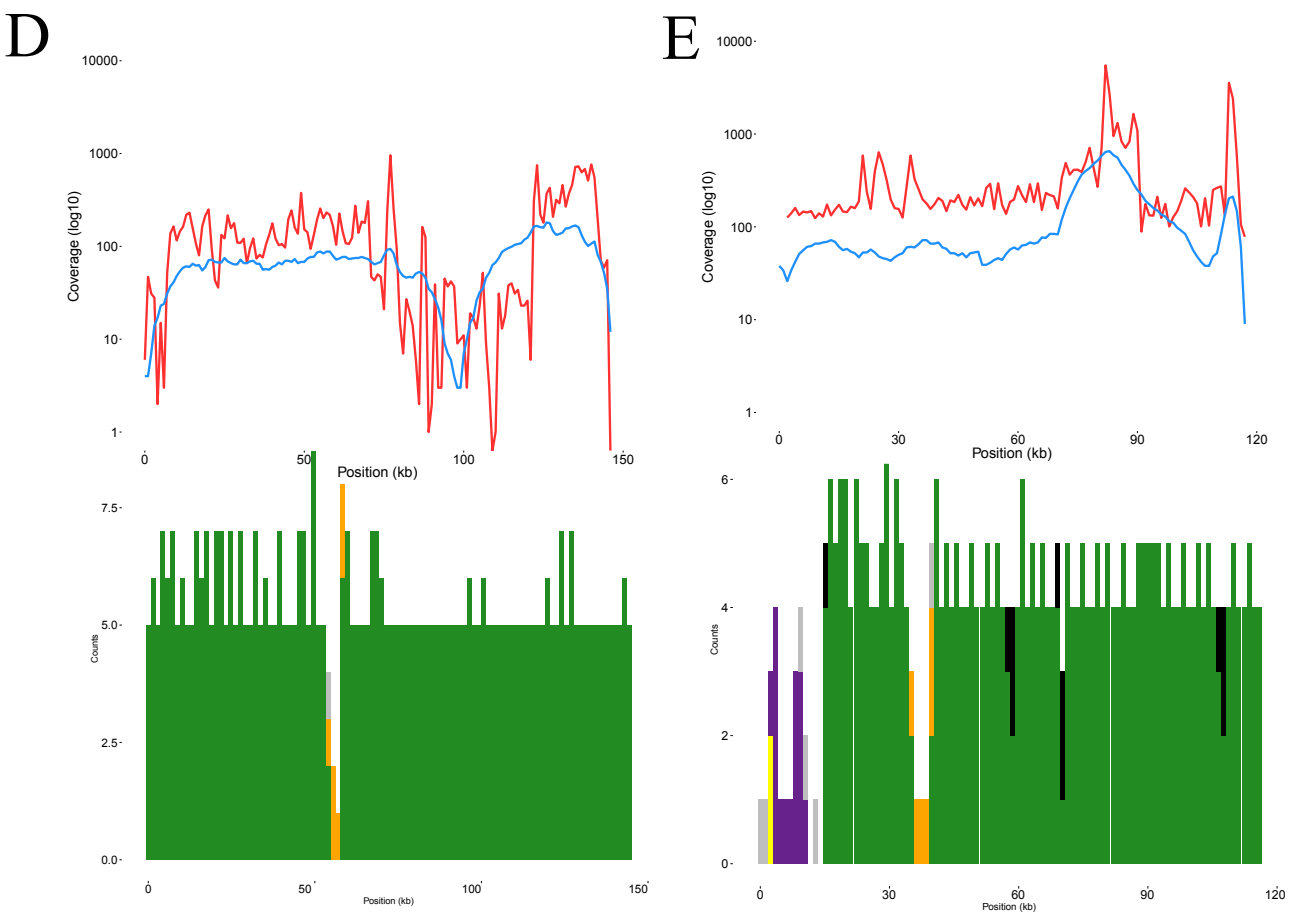


Figure S15: *1.688* loci in PBcR-BLASR assembly. Counts for each element in our custom Repbase library are shown in windows across five contigs. Green bars represent *1.688* family satellites (*353-bp* or *356-bp*). Contigs in A-C correspond to two loci in the *3L* pericentric heterochromatin ("distal" and "proximal") that are about 2Mb from each other. Coverage of Illumina reads (red lines) and PacBio reads (blue lines) in 1kb windows are shown above for each contig. (A) utg 43 (distal *3L* locus) (B) utg 93 (proximal *3L* locus) (C) utg 47 (proximal *3L* locus) (D) utg 564 (unknown) (E) utg 565 (unknown)

Table S1: summary of PBcR-MHAP assemblies

	Diploid/large genome?	kmer	sketch size	coverage	# *Rsp*	# Rsp conti	*Rsp score**	# *260-bp*	# *260-bp* cc	*260 score***
r6.03	n/a	n/a	n/a	n/a	343	9	38.1	206	57	3.6
PBcR BLASR*	n/a	14	n/a	25	1088	3	362.7	284	13	21.8
blasr corr cel8.3	n/a	14	1500	25	923	3	307.7	505	46	11.0
PBcR-MHAP 8.2**	n	16	1024	25	251	4	62.8	172	16	10.8
PBcR-MHAP 8.2 EK	n	16	1024	25	256	5	51.2	163	13	12.5
PBcR-MHAP 8.3 EK	n	16	1024	25	616	8	77.0	283	39	7.3
PBcR-MHAP 20_500_25X	y	20	500	25	1964	6	327.3	384	60	6.4
PBcR-MHAP 20_2000_25X	y	20	2000	25	1512	5	302.4	439	42	10.5
PBcR-MHAP 25_1500_25X	y	25	1500	25	2318	9	257.6	385	61	6.3
PBcR-MHAP 25_2000_25X	y	25	2000	25	2441	10	244.1	398	46	8.7
PBcR-MHAP 14 20X	y	14	1500	20	1291	5	258.2	341	41	8.3
PBcR-MHAP 14_25X	y	14	1500	25	2362	11	214.7	391	45	8.7
PBcR-MHAP 16 20X	y	16	1500	20	1260	4	315.0	374	37	10.1
PBcR-MHAP 18 25X	y	18	1500	25	1858	6	309.7	513	48	10.7
PBcR-MHAP 20 25X	y	20	1500	25	1486	5	297.2	391	41	9.5
PBcR-MHAP 20 20X	y	20	1500	20	1986	8	248.3	445	39	11.4

*Sergey Koren and Adam Phillipy
**Berlin et al 2015
***Score refers to # repeats / # of contigs

Table S2: summary of FALCON assemblies

	-min_cov	-min_size	# *Rsp*	# *Rsp* contigs	*Rsp score**	# 260-bp	# *260-bp* contigs	*260 score**
falcon_20_2500	20	2500	438	5	87.6	289	11	26.2727273
falcon_20_5000	20	5000	295	4	73.75	282	10	28.2
falcon_20_7500	20	7500	391	4	97.75	289	12	24.0833333
falcon_5_10000	5	10000	370	4	92.5	283	12	23.5833333
falcon_5_2500	5	2500	423	6	70.5	343	20	17.15
falcon_5_5000	5	5000	622	7	88.8571429	320	16	20
falcon_5_7500	5	7500	546	6	91	290	17	17.0588235
falcon_10_10000	10	10000	365	4	91.25	342	18	19
falcon_10_2500	10	2500	385	6	64.1666667	326	18	18.1111111
falcon_10_5000	10	5000	565	5	113	317	17	18.6470588
falcon_10_7500	10	7500	544	5	108.8	293	15	19.5333333
falcon_15_10000	15	10000	386	4	96.5	296	17	17.4117647
falcon_15_2500	15	2500	438	5	87.6	292	12	24.3333333
falcon_15_5000	15	5000	602	5	120.4	282	11	25.6363636
falcon_15_7500	15	7500	290	4	72.5	297	12	24.75
falcon_20_10000	20	10000	391	4	97.75	136	14	9.71428571
falcon_min12000		12000	665	6	110.833333	372	20	18.6
falcon_min5000		5000	1169	11	106.272727	403	33	12.2121212
falcon_ek			703	6	117.166667	455	32	14.21875

*Score refers to # repeats / # of contigs

Table S3: summary of Canu assemblies

	Error rate	# *Rsp*	# *Rsp* contigs	*Rsp* score*	# *260*	# *260* contigs	*260* score*
canu 0.02	0.02	689	4	172.25	121	18	6.7222222
canu default (0.025)	0.025	954	6	159	255	18	14.166667
canu 0.035	0.035	1295	4	323.75	277	14	19.785714
canu 0.04	0.04	1114	3	371.33333	265	15	17.666667
canu-corr	N/A	1065	3	355	466	29	16.068966

*Score refers to # repeats / # of contigs

Table S4: Rsp and 260-bp loci breakpoints in R6 vs PbcR-BLASR

Assembly: contig			
PbcR-BLASR: utg_2021	300407-166566		
R6.03: 2R		24527-end	
PbcR-BLASR: utg_2031		694157-732817	
R6.03: 2L	1-23,112,003		23,112,004-end

Table S5: simple satellites in reads and assembly

Sample	AAGAG count (# bases)	AATAT count (# bases)	Total # bases	Percent AAGAG total	Percent AATAT tot
Raw reads	108057261	39477773	15744118797	0.686334131	0.250746158
Error corrected reads	2761071	954412	8884162838	0.031078573	0.010742847
PBcR-BLASR	7620	68819	138490501	0.005502182	0.049692217

Supplemental methods

Slot Blots

Genomic DNA (100 ng to 600 ng) was denatured (final concentration 0.25 N NaOH, 0.5 M NaCl) for 10 mins at room temperature and then quick cooled. Slot blots were performed as recommended using a 48-well BioDot SF microfiltration apparatus (Bio-Rad). Each blot was first hybridized with a biotinylated rp49 RNA probe (generated as per above with primers: T7_rp49REV 5'-GTAATACGACTCACTATAGGGCAGTAAACGCGGTTCTGCATG-3 and rp49FOR 5'-CAGCATACAGGCCCAAGATC-3'). The membrane was then stripped with a 100° C solution of 0.1X SSC/ 0.5% SDS (3 times for ~20 mins) and re-hybridized with the *Rsp* probe as per the above Southern analysis. Signals were quantitated using the ImageLab software (BioRad).