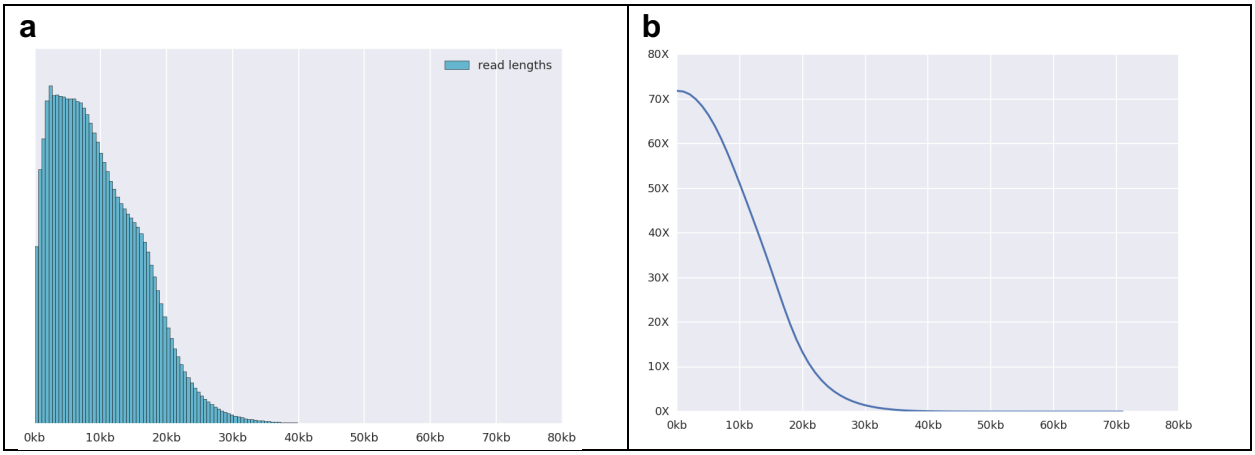
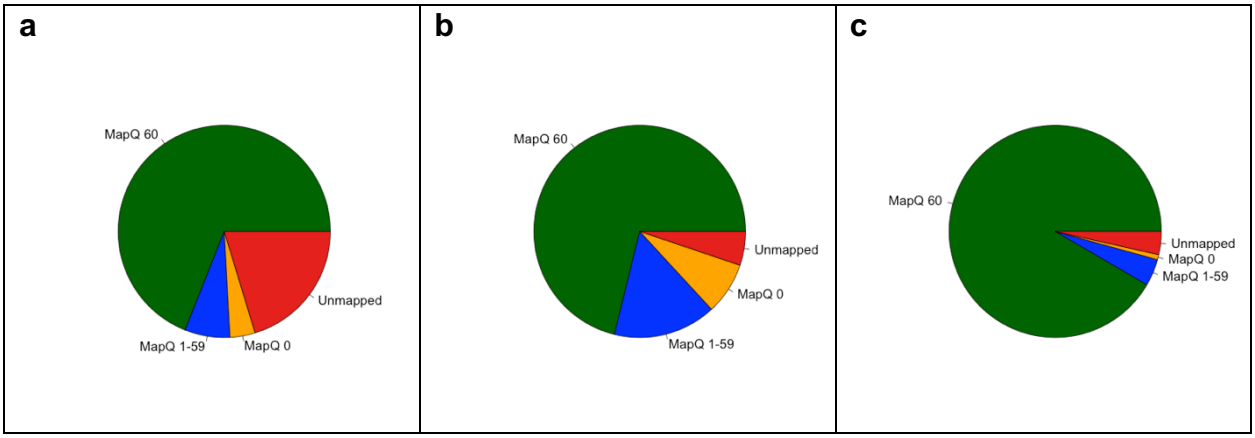
Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line

Maria Nattestad, Sara Goodwin, Karen Ng, Timour Baslan, Fritz J. Sedlazeck, Philipp Rescheneder, Tyler Garvin, Han Fang, James Gurtowski, Elizabeth Hutton, Elizabeth Tseng, Chen-Shan Chin, Timothy Beck, Yogi Sundaravadanam, Melissa Kramer, Eric Antoniou, John D. McPherson, James Hicks, W. Richard McCombie, Michael C. Schatz

Supplementary Tables and Figures



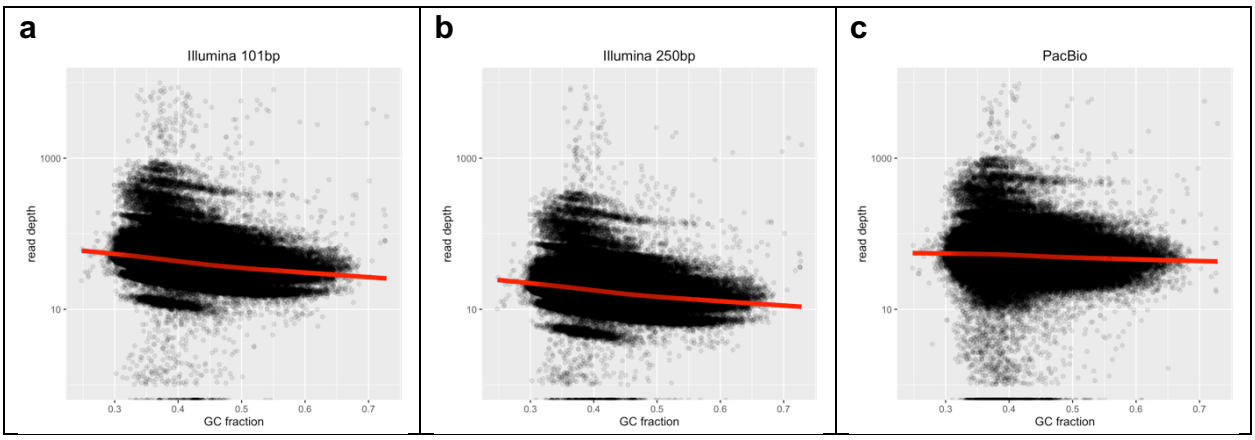
Supplementary Figure 1. Read length distribution of PacBio sequencing of the SK-BR-3 genome. Read lengths shown are of the longest subread from each circular PacBio read. (a) Histogram of read lengths. (b) Cumulative read depth as a function of minimum read length.



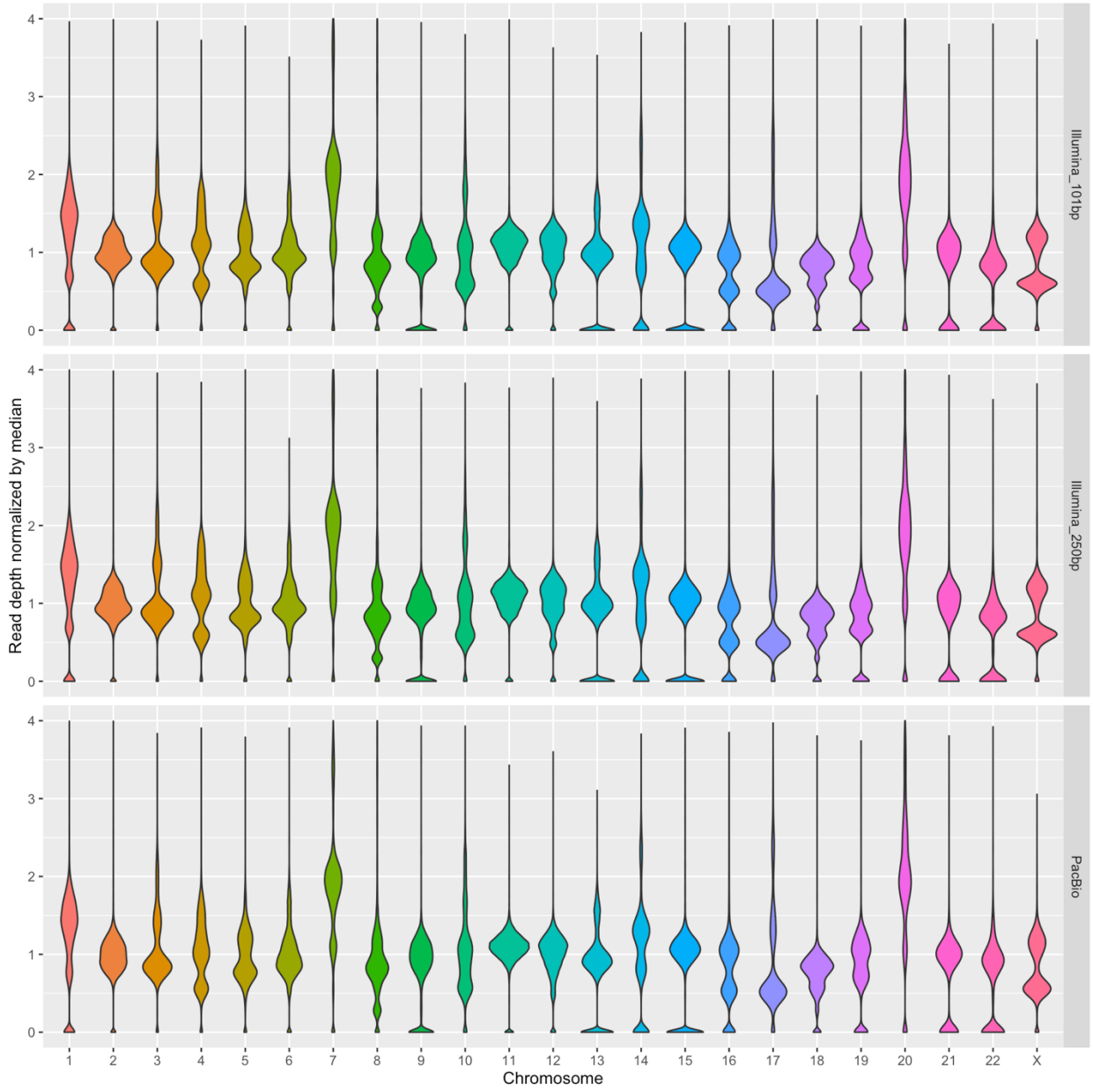
Supplementary Figure 2. Mappability of sequencing reads. Plots show the fraction of reads from each technology falling into each mapping quality category. (a) 101bp trimmed Illumina paired-end reads, where each read in the pair is evaluated separately. (b) 250bp untrimmed Illumina paired-end reads, where each read in the pair is evaluated separately. (c) PacBio long reads with an average read length of about 9.8kb.

	Illumina_101bp		Illumina_250bp		PacBio	
MapQ 60	549,180,512	69.00%	566,403,864	71.16%	20,681,145	91.61%
MapQ 1 to 59	54,639,912	6.86%	125,107,837	15.72%	924,192	4.09%
MapQ 0	30,522,037	3.83%	63,269,802	7.95%	170,158	0.75%
Unmapped	161,599,641	20.30%	41,160,599	5.17%	799,173	3.54%
Total	**795,942,102**	**100.00%**	**795,942,102**	**100.00%**	**22,574,668**	**100.00%**

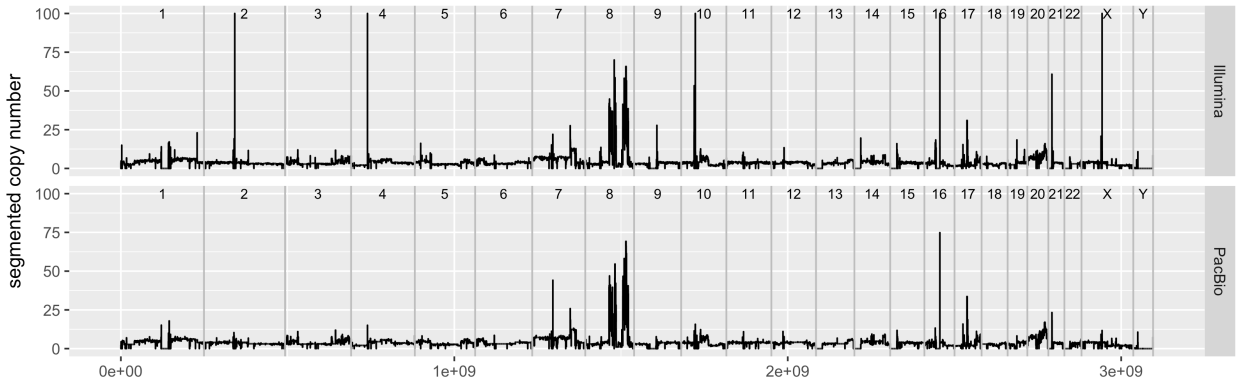
Supplementary Table 1. Read Mappability Analysis. Number of reads from each technology mapping with each category of mapping qualities.



Supplementary Figure 3. GC skew of the different sequencing technologies. GC content correlates with read depth for Illumina sequencing and to a lesser extent for PacBio sequencing. Read depth (y-axis) is shown on a log scale. A Lowess fit is shown in red. (a) Illumina trimmed 101bp dataset. (b) Illumina untrimmed 250bp dataset. (c) PacBio sequencing.



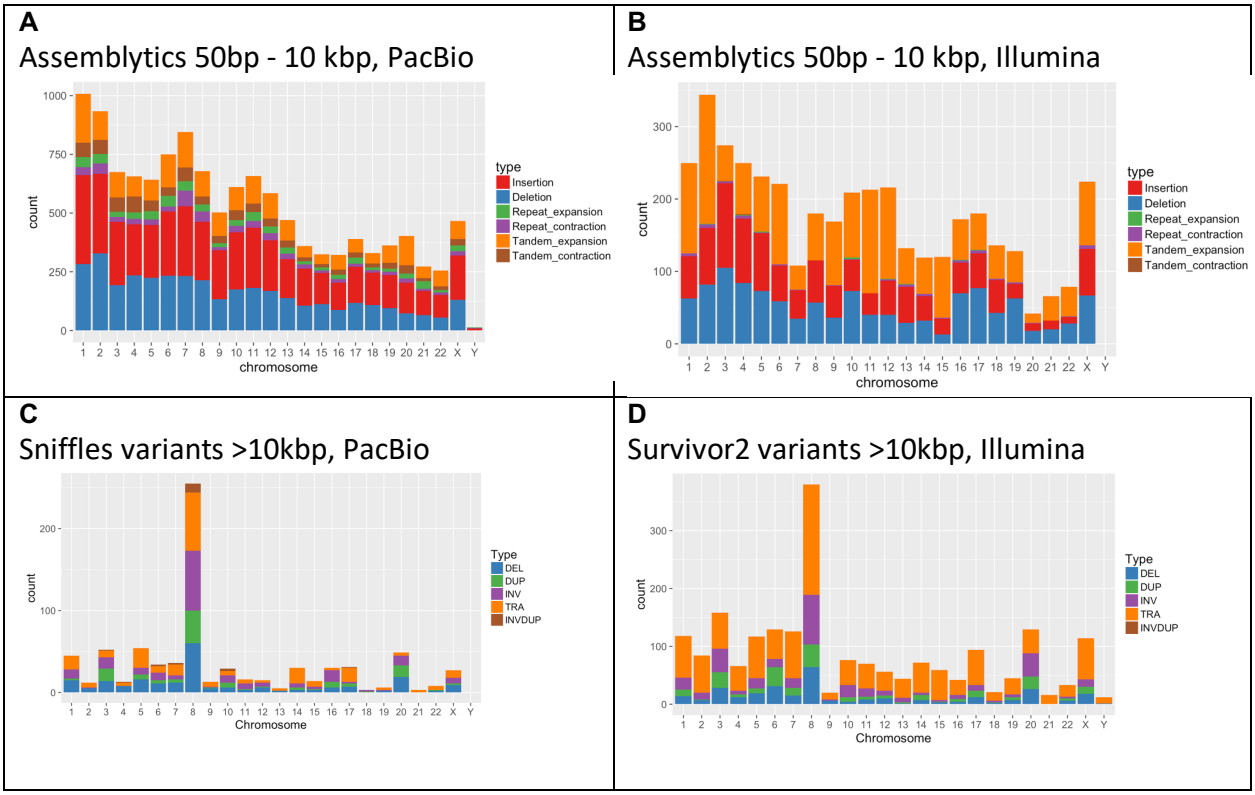
Supplementary Figure 4. Normalized read coverage by chromosome. Distribution of normalized read depth (dividing by the whole-genome median for each dataset), broken out by chromosome for each of the three technologies: (top) Illumina data trimmed to 101bp, (middle) Illumina data untrimmed at 250bp, and (bottom) PacBio long read sequencing.



Supplementary Figure 5. Segmented copy number profiles. Segmented copy number across the SK-BR-3 genome of the Illumina (top) and PacBio (bottom) sequencing datasets. Copy numbers are capped at 100 for visibility.

We note that the extreme coverage spikes specific to the Illumina data are concentrated in poorly mappable regions of the genome, and thus likely represent alignment artifacts. Specifically, we extracted those regions with segmented coverage higher than 1000 and computed their “DangerTrack” scores. The DangerTrack scores ranges from 0 (completely trustworthy region) to 1000 (not accessible by short reads) and aggregates the mappability tracks from UCSC along with other highly suspicious regions found throughout the genome, such as those with abnormally high counts of structural variations identify by the 1000 Genomes project and Genome-In-A-Bottle. The DangerTrack score is discussed and published here: 10.12688/f1000research.11254.1.

The average score for the abnormally highly covered areas was 468.73, compared to the genome-wide average DangerTrack score of 107.16. To further assess the extreme high score of 468.73 we shuffled the regions using bedtools and intersected these random regions with the DangerTrack scores. This resulted in an average score of 75.91. Thus overall the high spikes in coverage correspond to regions that are highly disturbed and highly repetitive and thus not reliable for short read mapping.

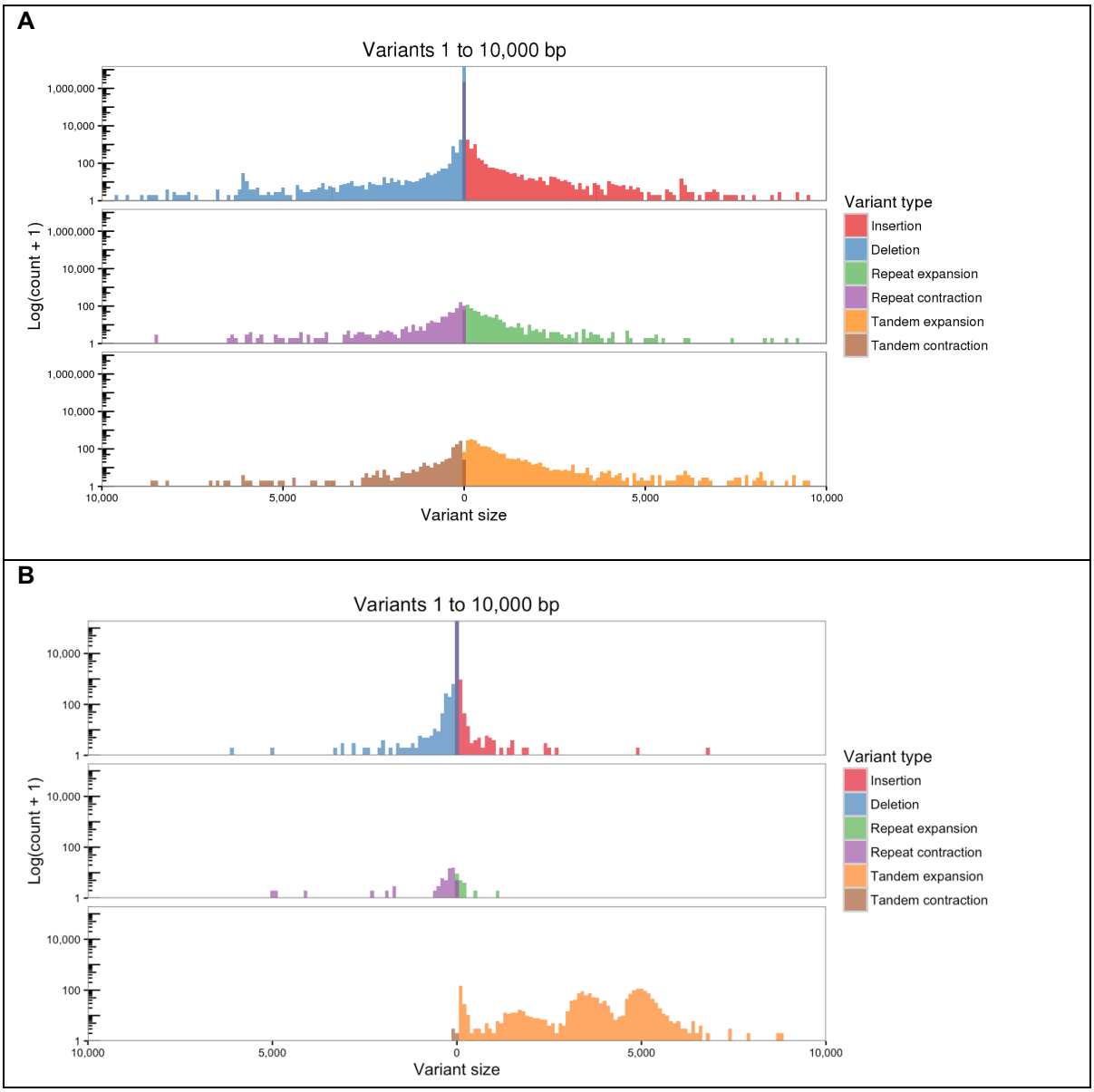


Supplementary Figure 6. Number of variants per chromosome. (A) Assemblytics variant calls from the PacBio Falcon assembly. (B) Assemblytics variant calls from the Illumina Allpaths-LG assembly. (C) Sniffles variant calls from PacBio read alignments. (D) Survivor consensus of Illumina paired-end read alignment-based variant calls where two of Lumpy, Manta, and Delly agree on a call.

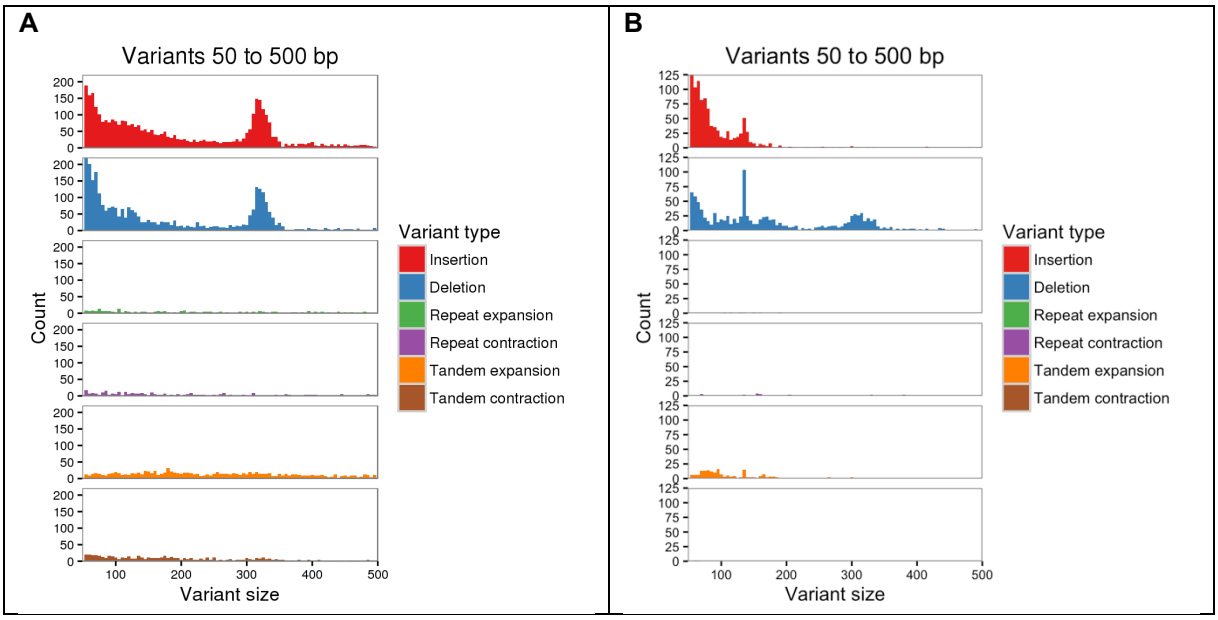
	PacBio Falcon Assembly		Illumina Allpaths-LG assembly	
Insertion				
Size range	**Count**	**Total bp**	**Count**	**Total bp**
1-10 bp	2262493	2942482	171013	359716
10-50 bp	24930	427225	11453	203301
50-500 bp	3839	727296	1058	93648
500-10000 bp	868	1619801	31	46511
Total	2292130	5716804	183555	703176
Deletion				
Size range	**Count**	**Total bp**	**Count**	**Total bp**
1-10 bp	15271816	17864905	177487	382440
10-50 bp	36863	612808	17425	307459
50-500 bp	3175	537178	1155	198094
500-10000 bp	533	1332189	52	74113
Total	15312387	20347080	196119	962106
Tandem expansion				
Size range	**Count**	**Total bp**	**Count**	**Total bp**
1-10 bp	5	22	0	0
10-50 bp	55	1968	0	0
50-500 bp	1175	292240	184	21989
500-10000 bp	1002	1642858	1525	6298306
Total	2237	1937088	1709	6320295
Tandem contraction				
Size range	**Count**	**Total bp**	**Count**	**Total bp**
1-10 bp	1	1	0	0
10-50 bp	25	1053	1	38
50-500 bp	607	114282	2	126
500-10000 bp	197	355360	0	0
Total	830	470696	3	164
Repeat expansion				
Size range	**Count**	**Total bp**	**Count**	**Total bp**
1-10 bp	14	75	2	9
10-50 bp	42	1315	6	185
50-500 bp	325	72375	7	982
500-10000 bp	278	455935	2	1683
Total	659	529700	17	2859
Repeat contraction				
Size range	**Count**	**Total bp**	**Count**	**Total bp**
1-10 bp	20	74	3	11
10-50 bp	83	2569	1	42
50-500 bp	340	67653	38	7252
500-10000 bp	203	390510	10	23123
Total	646	460806	52	30428
Total for all variants	17608889	29462174 bp	381455	8019028 bp
Total for all structural variants (size ≥ 50bp)	12542	7607677 bp	4064	6765827 bp

Supplementary Table 2. Comparison variants in the assemblies.

Table shows the counts and sizes of variant calls between the long-read and short-read assemblies.



Supplementary Figure 7. Histograms of variant sizes and types in the assemblies. (A) Long read FALCON assembly. (B) Illumina short reads Allpaths-LG assembly. Both plots show variants 1 bp to 10 kbp in size.



Supplementary Figure 8. Close up of variants found in the assemblies. Histograms of variant sizes by type detected from the PacBio Falcon (A) and Illumina Allpaths-LG (B) assemblies from 50 bp to 500 bp in size, showing clear Alu peaks around 320 bp for insertions and deletions in (A) but only a smaller peak for deletions and no insertion peak in (B).

	PacBio FALCON assembly	Illumina Allpaths-LG assembly
Deletions 300-350 bp	703	181
Deletions matching Alu elements	655 (93%)	165 (91%)
Insertions 300-350 bp	855	5
Insertions matching Alu elements	741 (87%)	2 (40%)

Supplementary Table 3. Analysis of variant ALU elements in the assemblies. Peaks around 320 bp in the distributions of insertions and deletions match Alu elements, but many more of these mobile element insertions and deletions are captured with the long-read FALCON assembly compared to a short-read, Allpaths-LG assembly. Matching Alu elements was determined by intersecting a RepeatMasker database on the reference for the deletions, and blasting extracted sequences from the assembly for insertions.

	DEL	INS	DUP	INV	TRA
50 bp - 1 kbp	136	166	17	0	N/A
long-range	23	9	11	14	11

Supplementary Table 4. Summary of COSMIC Cancer Gene census variants. There are 172 genes in the COSMIC Cancer Gene census whose transcribed sequences are hit by a total of 387 structural variants (minimum size 50 bp). Also see Supplementary Table 8 for details.

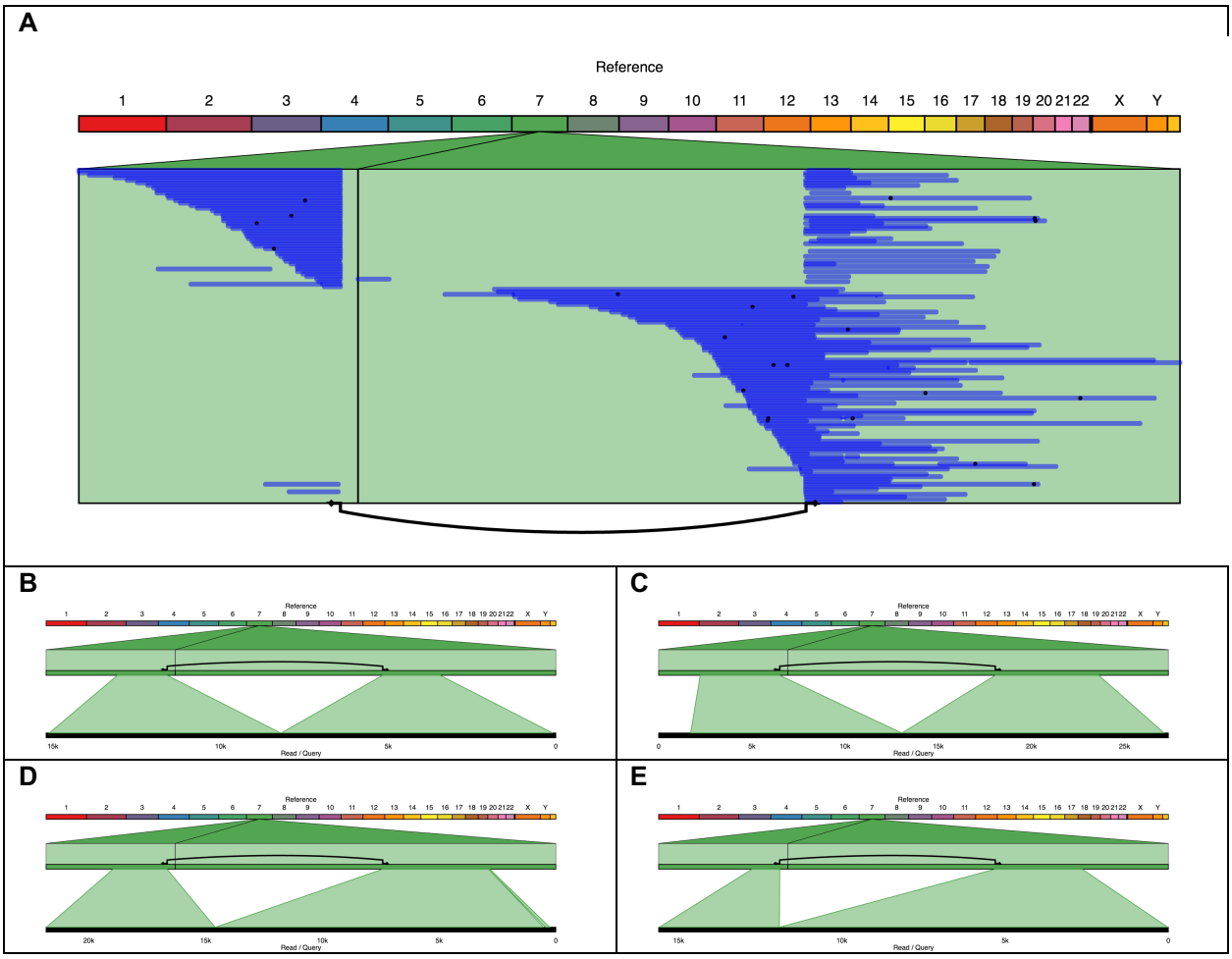
Supplementary Table 5. Summary of mapping based variant calls. See separate file.

Supplementary Table 6. Long Range Variants detected by Sniffles. See separate file.

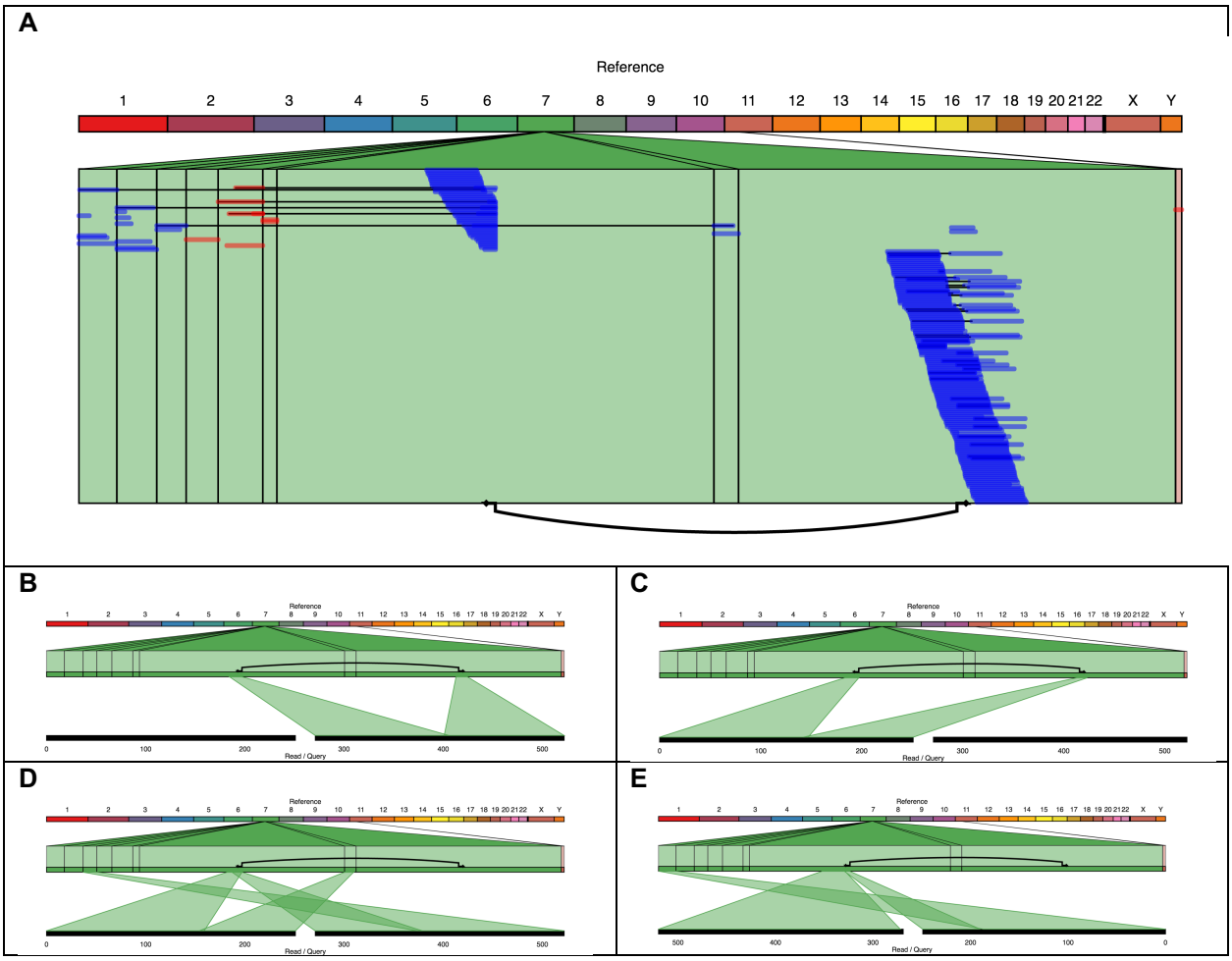
Supplementary Table 7. PCR Primers for Validation. See separate file.

Supplementary Table 8. Details of COSMIC Cancer Gene Variants. See separate file

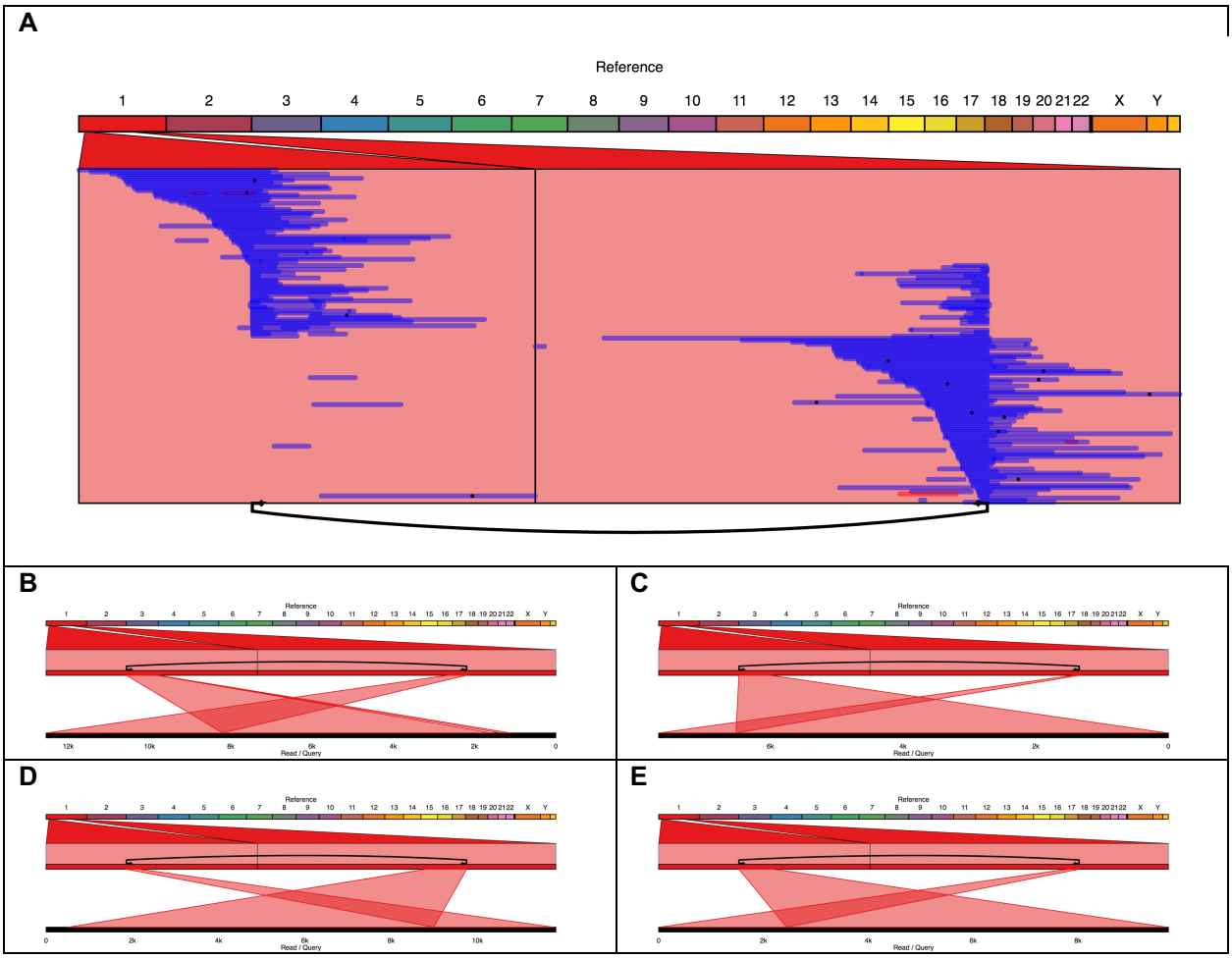
Supplementary Table 9. Gene Fusions detected by IsoSeq analysis. See separate file.



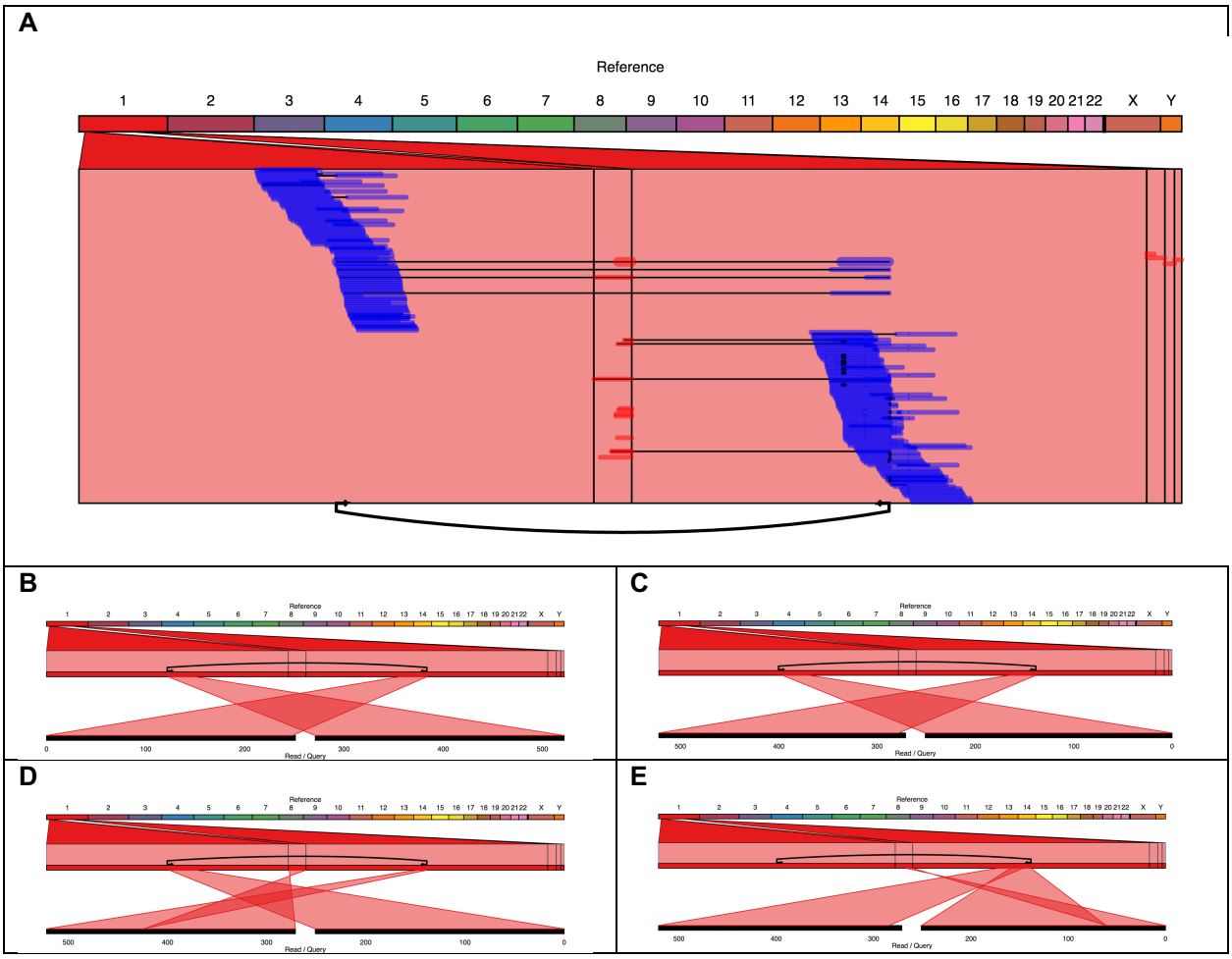
Supplementary Figure 9. Ribbon plot of a deletion detected by long reads. Ribbon plots of PacBio NGMLR alignments from a deletion (DEL145) called by Sniffles but by none of the short-read variant-callers tested (Delly, Manta, and Lumpy). **(A)** The multi-read view of the regions involved in this variant. The Sniffles call is shown as a connecting line underneath in black and NGMLR alignments as blue (forward) and red (reverse) where direction is with respect to the alignments at the breakpoints. **(B-E)** Single-read views for multiple reads supporting this variant, alignments from NGMLR.



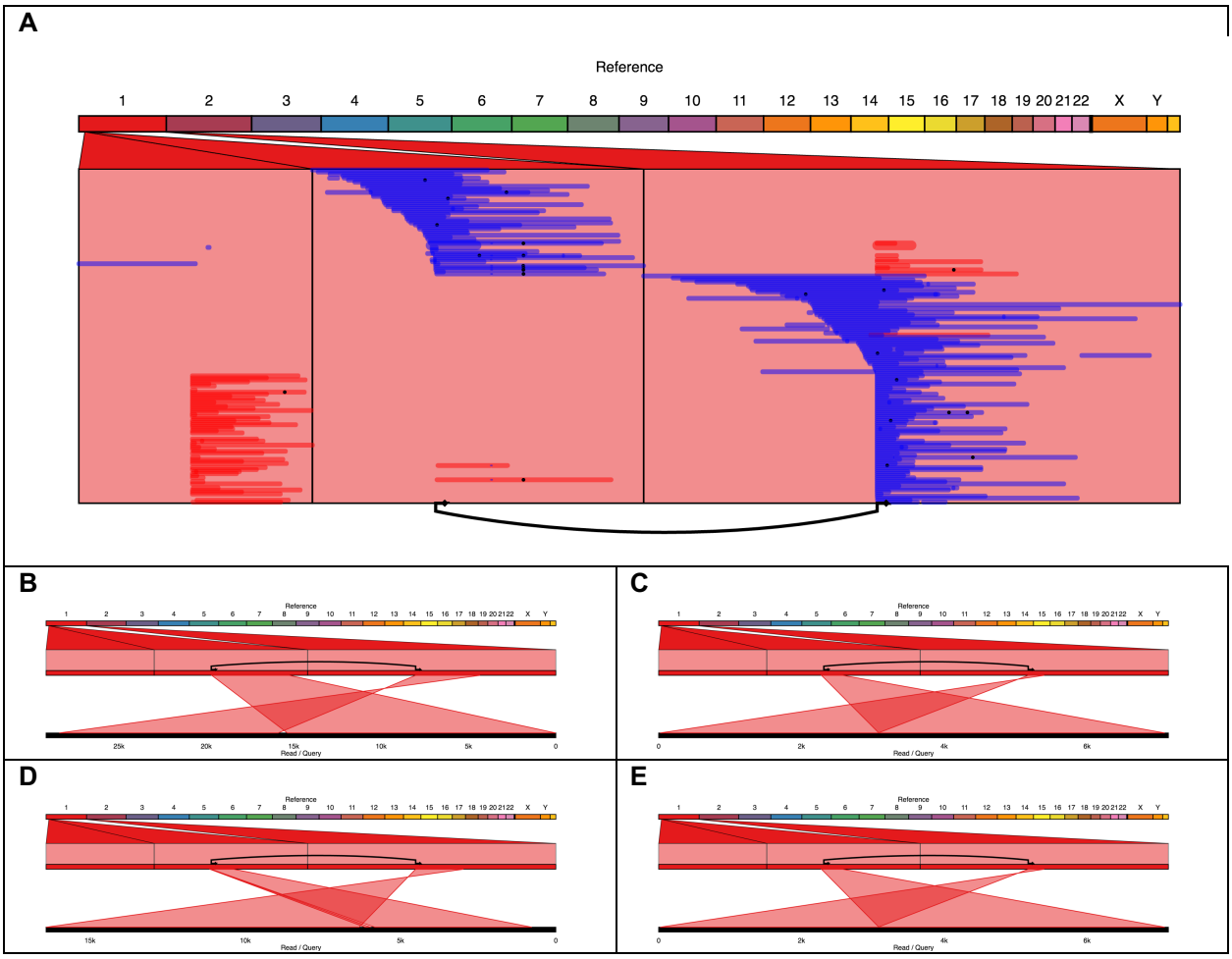
Supplementary Figure 10. Ribbon plot of short read alignments near candidate deletion. Ribbon plots of Illumina BWA-MEM alignments from the same deletion (DEL145) called by Sniffles but by none of the short-read variant-callers tested (Delly, Manta, and Lumpy). **(A)** The multi-read view of the regions involved in this variant. The Sniffles call is shown as a connecting line underneath in black and paired-end BWA-MEM alignments as blue (forward) and red (reverse) where direction is with respect to the alignments at the breakpoints. **(B-C)** Single-read views for the only two reads supporting the variant. **(D-E)** Single-read views for two other reads show alternate alignments in this region that could suggest repetitive sequence.



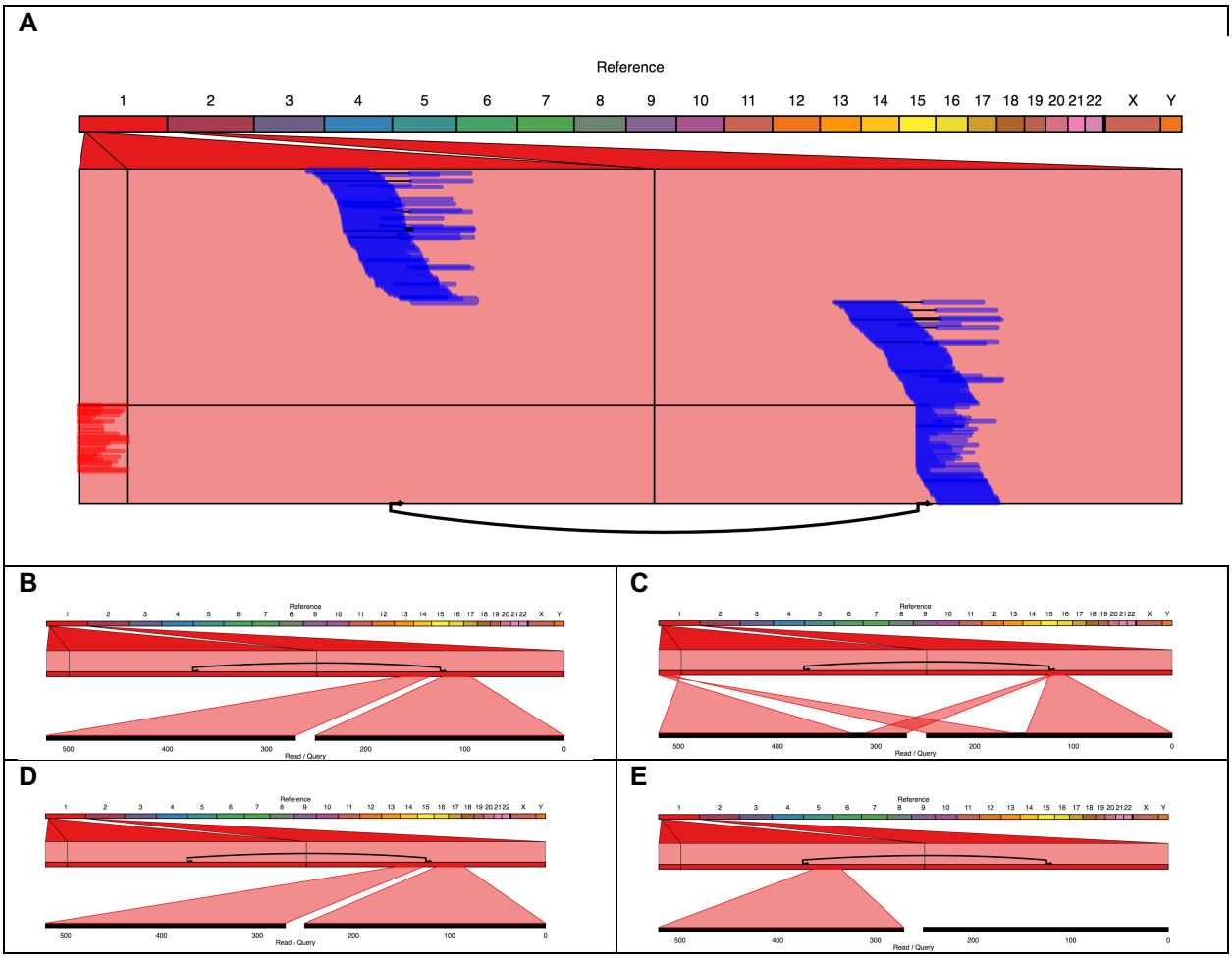
Supplementary Figure 11. Ribbon plot of duplication identified by long reads. Ribbon plots of PacBio NGMLR alignments from a duplication (DUP3) called by Sniffles but by none of the short-read variant-callers tested (Delly, Manta, and Lumpy). **(A)** The multi-read view of the regions involved in this variant. The Sniffles call is shown as a connecting line underneath in black and NGMLR alignments as blue (forward) and red (reverse) where direction is with respect to the alignments at the breakpoints. **(B-E)** Single-read views for multiple reads supporting this variant, alignments from NGMLR.



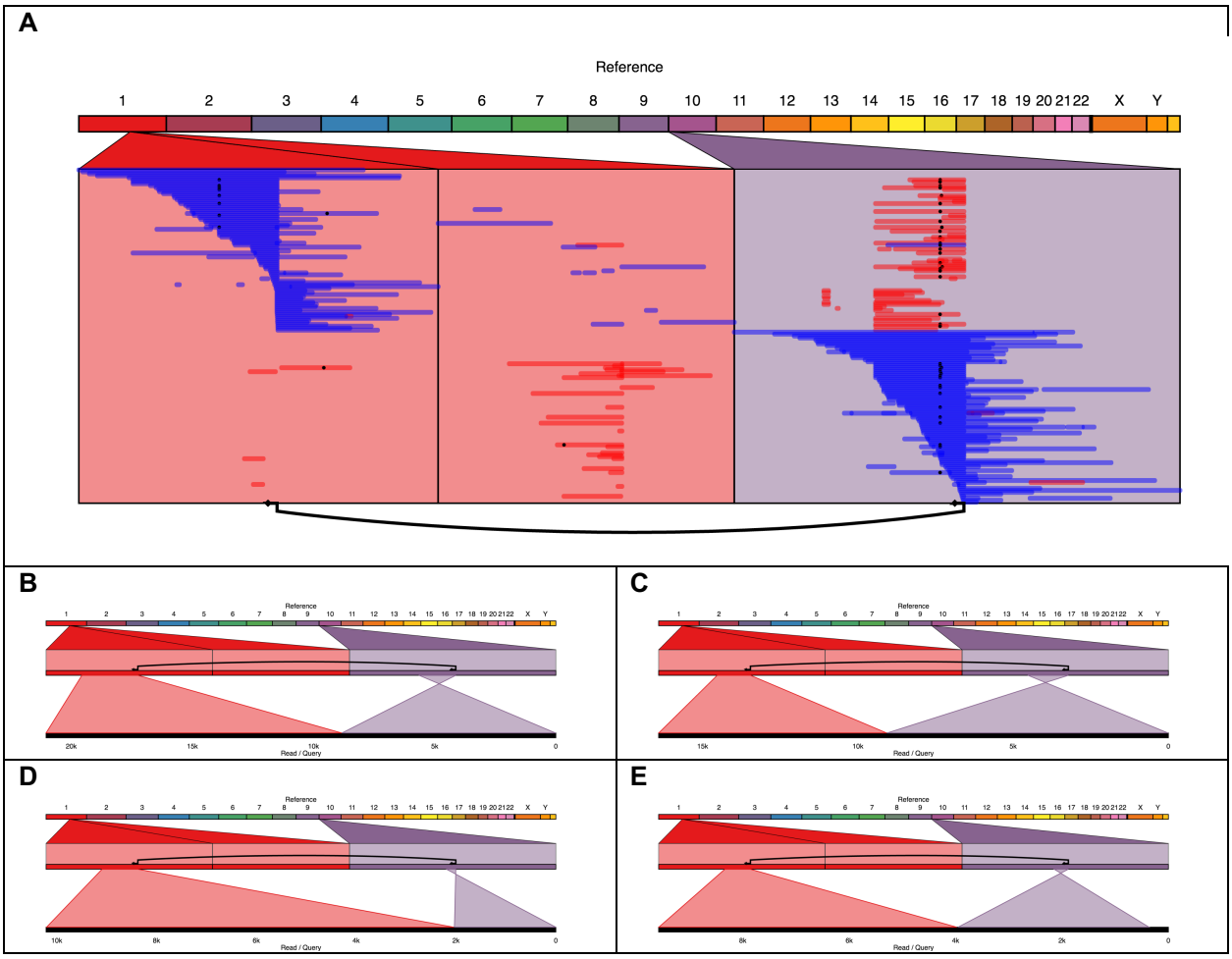
Supplementary Figure 12. Ribbon plot of short read alignments near duplication detected by long reads. Ribbon plots of Illumina BWA-MEM alignments from the same duplication (DUP3) called by Sniffles but by none of the short-read variant-callers tested (Delly, Manta, and Lumpy). **(A)** The multi-read view of the regions involved in this variant. The Sniffles call is shown as a connecting line underneath in black and paired-end BWA-MEM alignments as blue (forward) and red (reverse) where direction is with respect to the alignments at the breakpoints. **(B-C)** Single-read views for two reads with discordant pairs supporting the variant, but no reads are split supporting this variant. **(D-E)** Single-read views for two other reads show alternate alignments in this region that could suggest repetitive sequence.



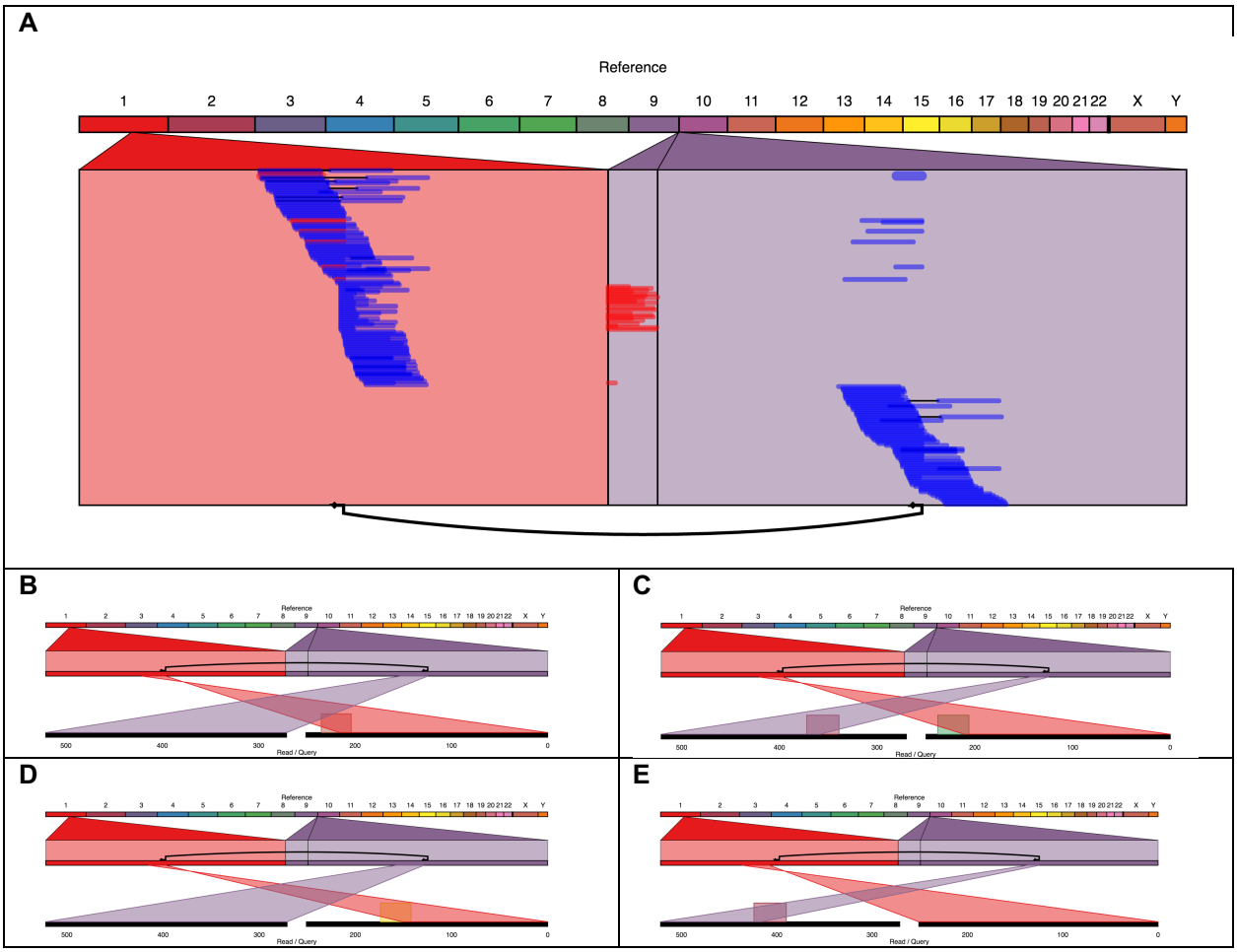
Supplementary Figure 13. Ribbon plot of inversion detected by long reads. Ribbon plots of PacBio NGMLR alignments from an inversion called by Sniffles but by none of the short-read variant-callers tested (Delly, Manta, and Lumpy). **(A)** The multi-read view of the regions involved in this variant. The Sniffles call is shown as a connecting line underneath in black and NGMLR alignments as blue (forward) and red (reverse) where direction is with respect to the alignments at the breakpoints. **(B-E)** Single-read views for multiple reads supporting this variant, alignments from NGMLR.



Supplementary Figure 14. Ribbon plot of short read alignments near inversion detected by long reads. Ribbon plots of Illumina BWA-MEM alignments from the same inversion (INV5) called by Sniffles but by none of the short-read variant-callers tested (Delly, Manta, and Lumpy). **(A)** The multi-read view of the regions involved in this variant. The Sniffles call is shown as a connecting line underneath in black and paired-end BWA-MEM alignments as blue (forward) and red (reverse) where direction is with respect to the alignments at the breakpoints. **(B-E)** Single-read views for four reads showing different alignment patterns, none of which are split between the two regions. No reads are split or show discordant read pair evidence for this variant.



Supplementary Figure 15. Ribbon plot of inversion detected by long reads. Ribbon plots of PacBio NGMLR alignments from a translocation (TRA1327) called by Sniffles but by none of the short-read variant-callers tested (Delly, Manta, and Lumpy). **(A)** The multi-read view of the regions involved in this variant. The Sniffles call is shown as a connecting line underneath in black and NGMLR alignments as blue (forward) and red (reverse) where direction is with respect to the alignments at the breakpoints. **(B-E)** Single-read views for multiple reads supporting this variant, alignments from NGMLR.



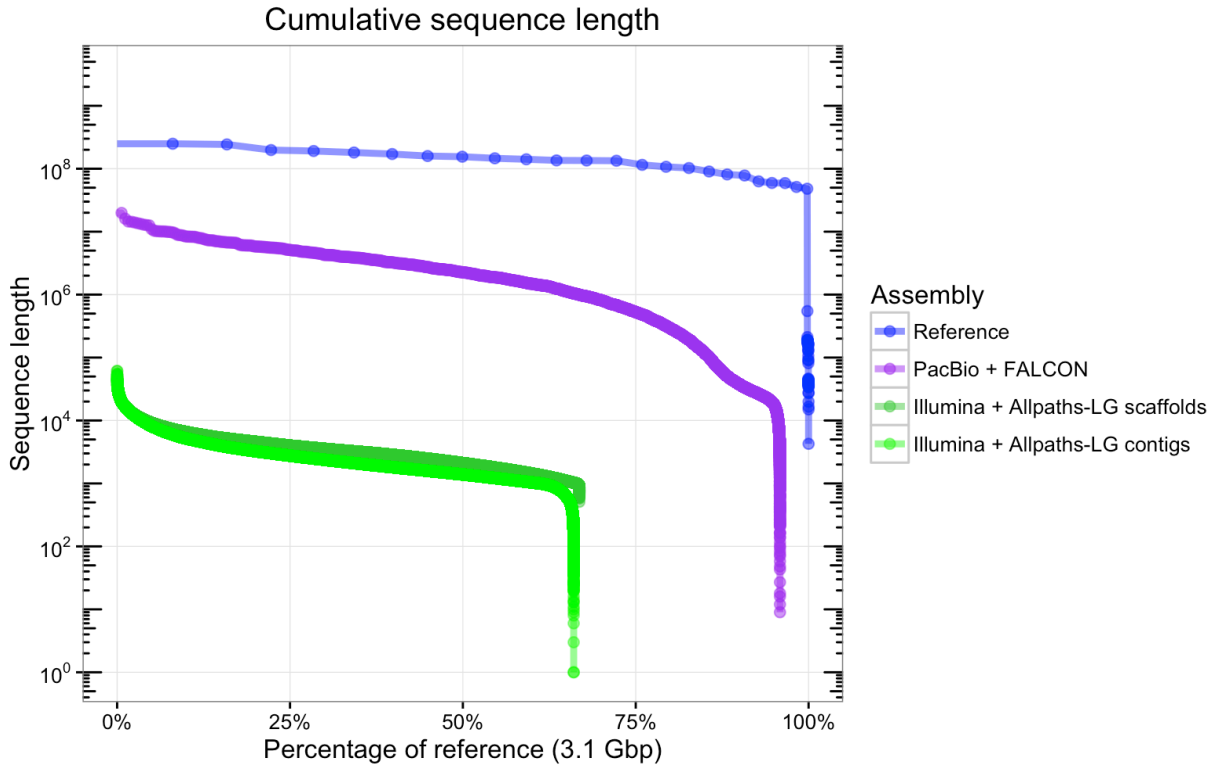
Supplementary Figure 16. Ribbon plot of short read alignments near inversion detected by long reads. Ribbon plots of Illumina BWA-MEM alignments from the same translocation (TRA1327) called by Sniffles but by none of the short-read variant-callers tested (Delly, Manta, and Lumpy). **(A)** The multi-read view of the regions involved in this variant. The Sniffles call is shown as a connecting line underneath in black and paired-end BWA-MEM alignments as blue (forward) and red (reverse) where direction is with respect to the alignments at the breakpoints. **(B-E)** Single-read views for four representative reads that have alignments to both regions with seemingly spurious alignments in between. All of the split reads between these two regions show alignments to various other chromosomes that suggest the presence of a repetitive sequence, which may contribute to the inability of all three short-read variant-callers tested to capture this interchromosomal fusion.

Supplementary Note 1: De novo assembly-based variant detection

We generated a de novo genome assembly of SK-BR-3 from our long-read PacBio dataset using the Falcon assembler[1]. For comparison, we also assembled the genome with the widely used Allpaths-LG assembler[2] to create a short-read assembly using a combination of an overlapping paired-end library and two mate-pair libraries sequenced to similar amounts of coverage.

The assembly was generated from the SMRT-sequencing reads using FALCON[1] on the DNAnexus platform. To produce a short-read assembly for comparison, the overlapping fragment library and the mate libraries Illumina reads were assembled using Allpaths-LG[2]. For assembly-based variant-calling, alignment of the PacBio assembly contigs and Illumina assembly contigs (not scaffolds) to hg19 was computed using MUMmer[3], and Assemblytics[4] was used to call structural variants.

For the short-read assembly, Illumina sequencing was performed on 180 bp paired end overlapping library (2x100 bp reads), as well as 2-3 kbp and 5-10 kbp mate-pair libraries. The contiguity of the long-read assembly is over one thousand-fold better than the short-read assembly, with a contig N50 of 2.4 Mbp compared to 2.1 kbp from short reads, also far surpassing the scaffold N50 of 3.2 kbp. The high quality long-read assembly also allowed for a much more comprehensive view into structural variations compared to the short-read assembly.



Supplementary Figure 17. Comparison of short and long read assemblies. Plot shows sequence lengths from de novo assemblies created using long reads (PacBio+FALCON, purple) and short reads (Illumina+Allpaths-LG, green), where the short-read Allpaths-LG assembly is shown as both the full scaffolded assembly (dark green) and the unscaffolded contigs (light green). The hg19 reference genome sequence lengths are shown in blue for reference.

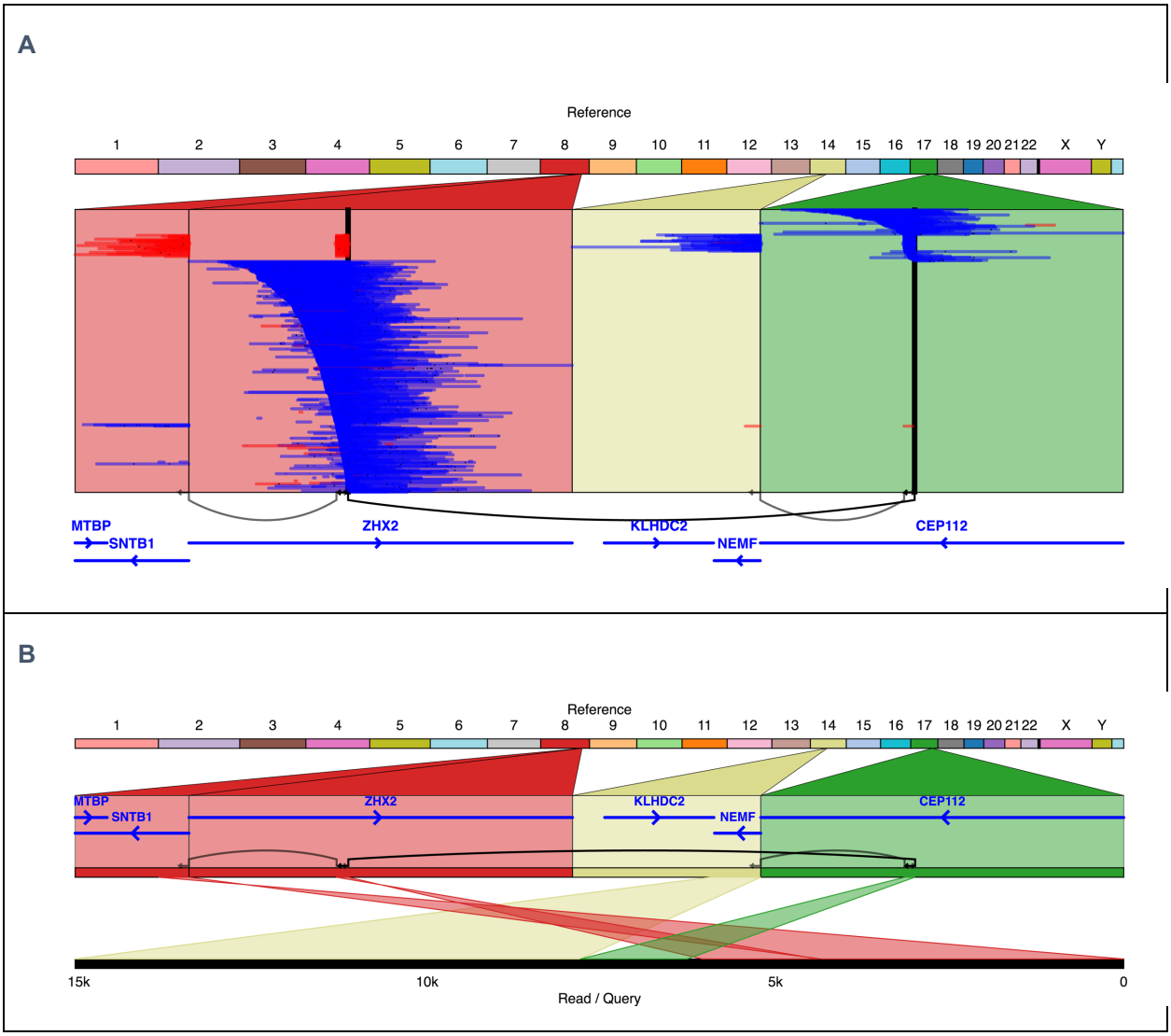
The long-read assembly has two clear advantages over the short-read assembly: a thousand-fold higher contiguity and a greater fraction of the genome is represented. To see if this would enable the long-read assembly to better capture the same variation in the genome that we saw using the alignment-based Sniffles variant-calling, we used Assemblytics, our assembly-based variant-caller on the PacBio Falcon assembly and on the Illumina Allpaths-LG assembly contigs.

We compared the genome assembly from Falcon with the reference genome using MUMmer[3] for alignment and Assemblytics[4] to characterize small indels and structural variants both between discordant alignments and within individual alignments. These variants were found in the range 50 bp and 10 kbp in size and characterized as insertions, deletions, tandem expansions, tandem contractions, repeat expansions, and repeat contractions by Assemblytics. Using the PacBio assembly, we detected a total of 17,608,889 variants, spanning 29.46 Mbp of the genome, of which 12,542 are structural variants of at least 50 bp in size, affecting 7.61 Mbp (**Supplementary Table 2**). The similar size distributions of insertions and deletions from the assembly reflects the same trend seen from Sniffles alignment-based variant-calling.

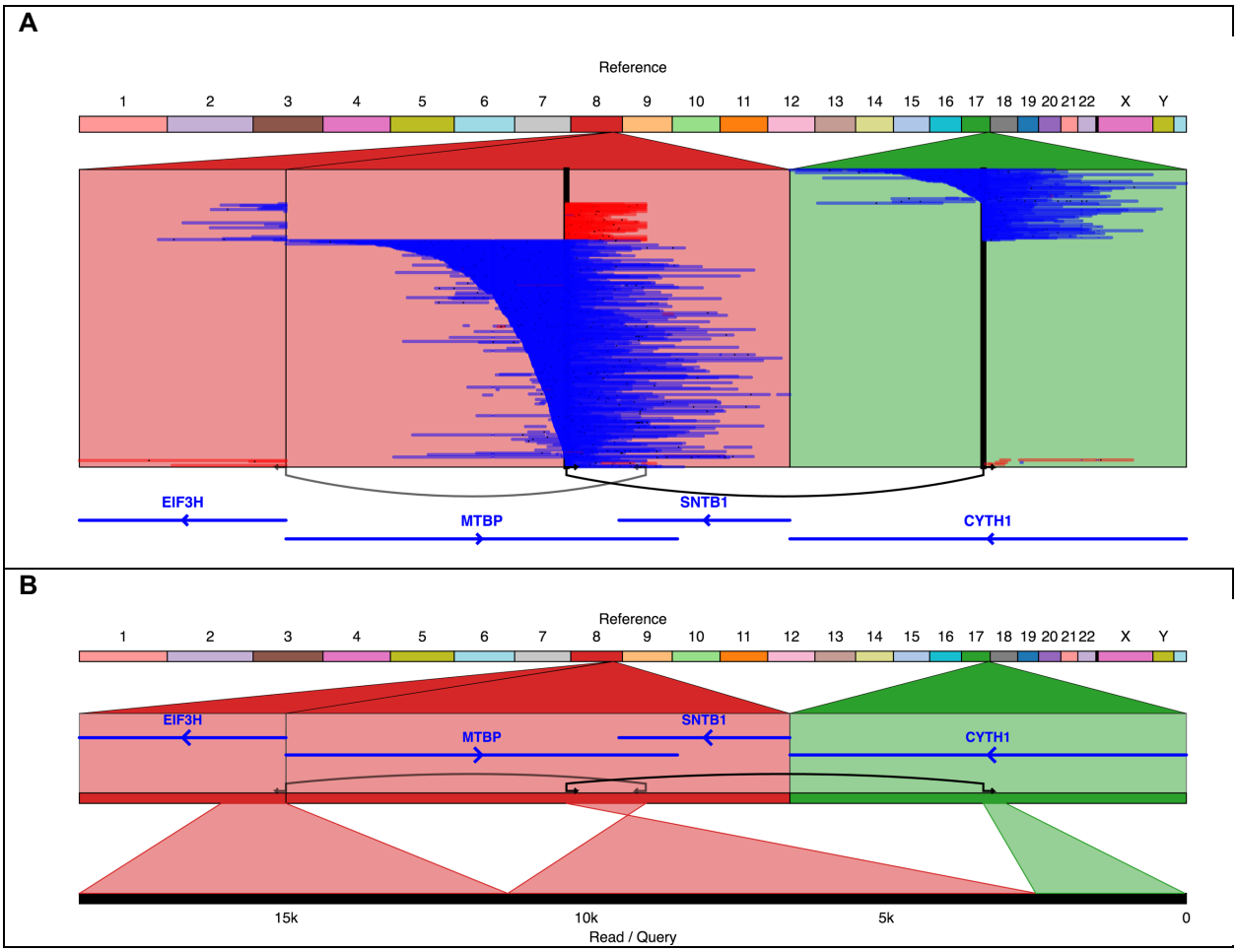
The long-read assembly captures 12,542 structural variants, while the short-read assembly only captures 4,064, less than a third as many (**Supplementary Table 2**). Structural insertions and deletions from the long-read assembly show larger counts for small variants with a decreasing frequency as size increases, except for a clear enrichment in the size range 300-350 bp (**Supplementary Figure 7A**). Suspecting that this enrichment is due to movement of Alu elements within human populations, we further investigated insertions and deletions within this size range. We noted that 655 of these deletions intersect Alu elements in the RepeatMasker hg19 database, and 741 insertions when extracted from the assembly match sequences in the human Alu BLAST database[5]. These mobile element indels are captured far better using the long-read FALCON assembly, whereas only 165 Alu deletions and a mere 2 Alu insertions are found from the short-read Allpaths-LG assembly (**Supplementary Figure 7B**, **Supplementary Table 3**).

Supplementary Note 2: Gene fusions with genome and transcriptome evidence

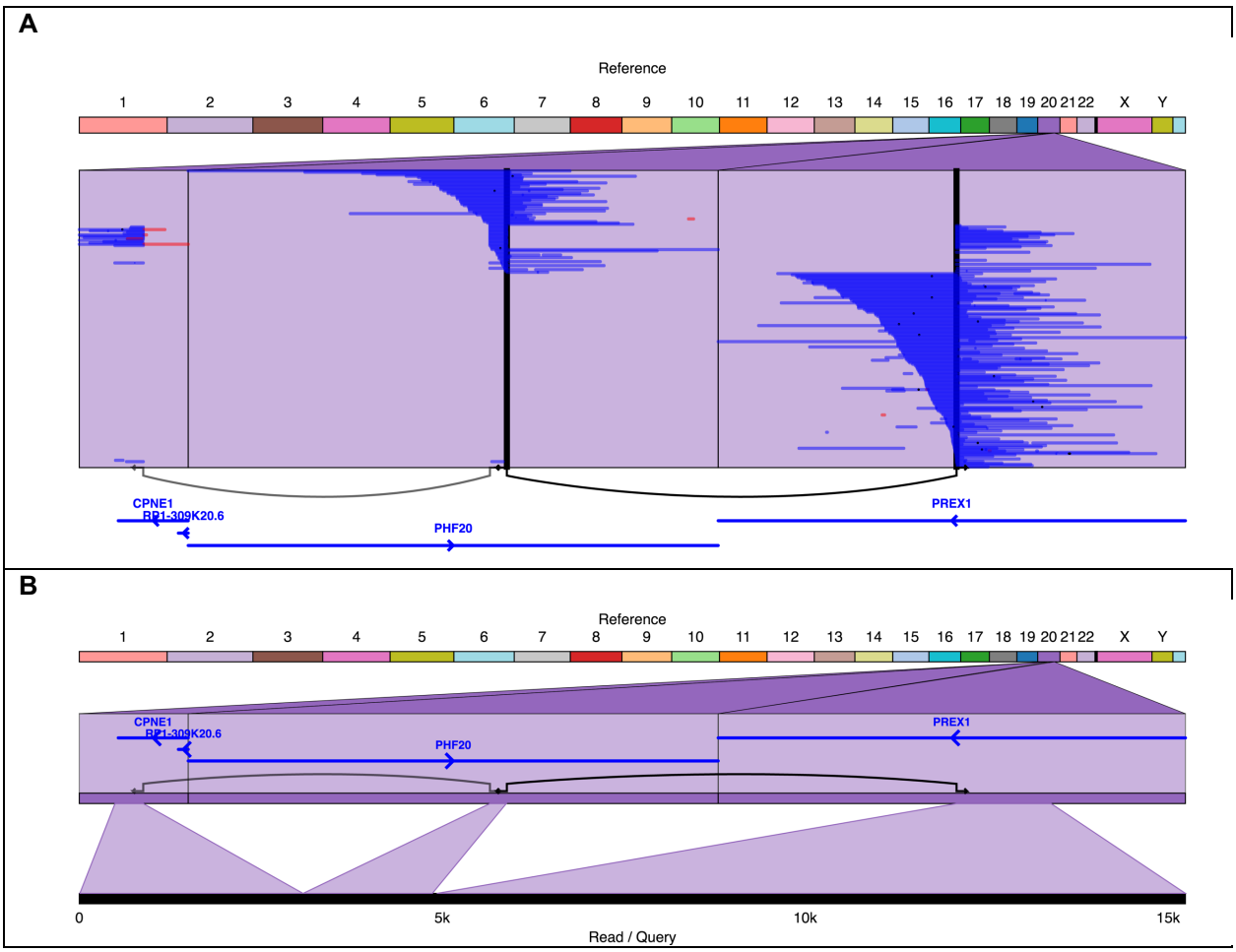
All figures in this section are generated using Ribbon to visualize the variants involved in gene fusions. The top multi-read view (A) in each figure is filtered to show only regions with at least 10 alignments; otherwise, all reads that have alignments at the relevant breakpoints are shown including all of their other alignments everywhere else in the genome. In each figure, one of the supporting reads is selected for (B) that clearly showcases all of the variants involved in the gene fusion.



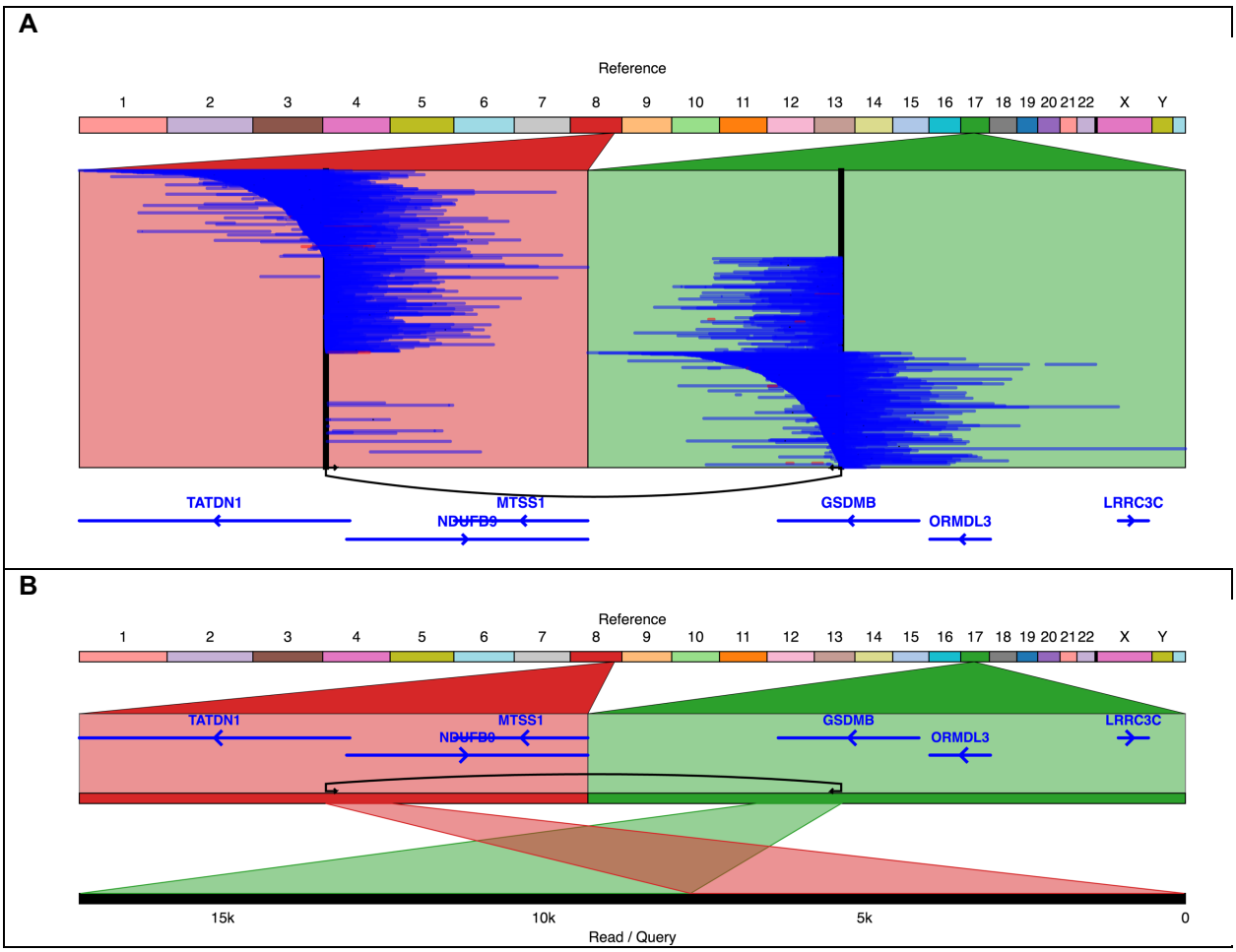
Supplementary Figure 18. Ribbon plot of "3-hop" KLHDC2-SNTB1 gene fusion captured by long reads. This is a "3-hop" gene fusion in SK-BR-3 created by a series of three variants (A). These variants are captured together in several individual SMRT sequencing reads, one of which is shown in (B).



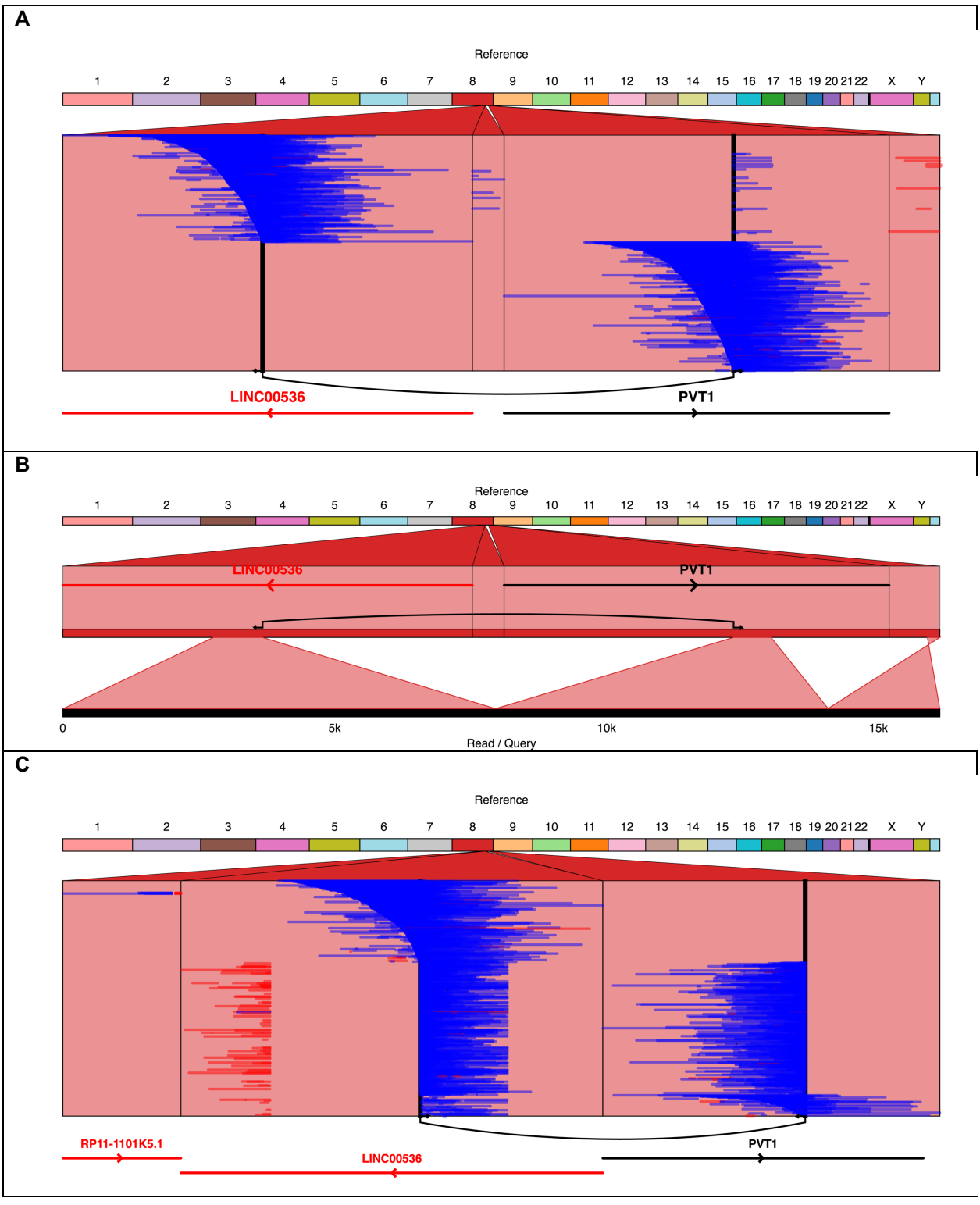
Supplementary Figure 19. Ribbon plot of "2-hop" CYTH1-EIF3H gene fusion captured by long reads. This is a "2-hop" gene fusion in SK-BR-3 created by a series of two variants (A). These variants are captured together in several individual SMRT sequencing reads, one of which is shown in (B).

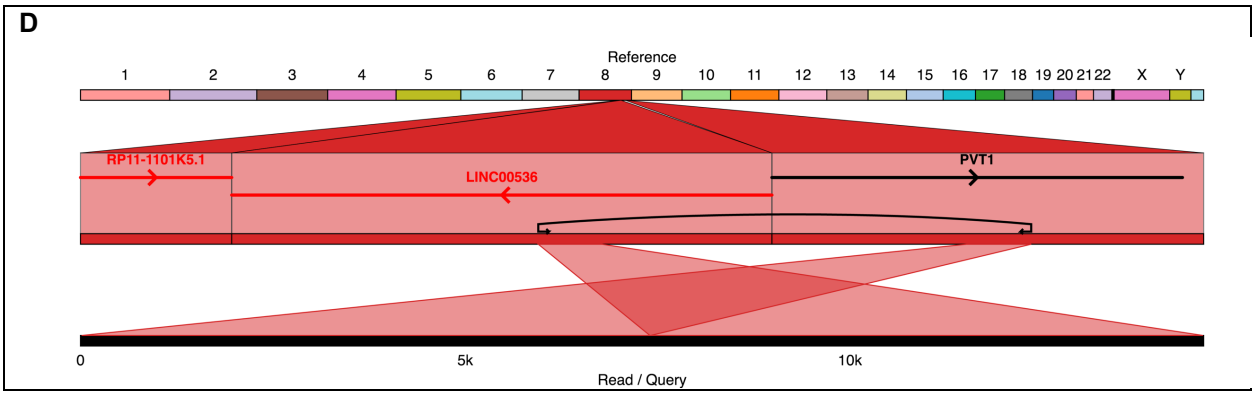


Supplementary Figure 20. Ribbon plot of "2-hop" CPNE1-PREX1 gene fusion captured by long reads. This is a "2-hop" gene fusion in SK-BR-3 created by a series of two variants (A). These variants are captured together in several individual SMRT sequencing reads, one of which is shown in (B).

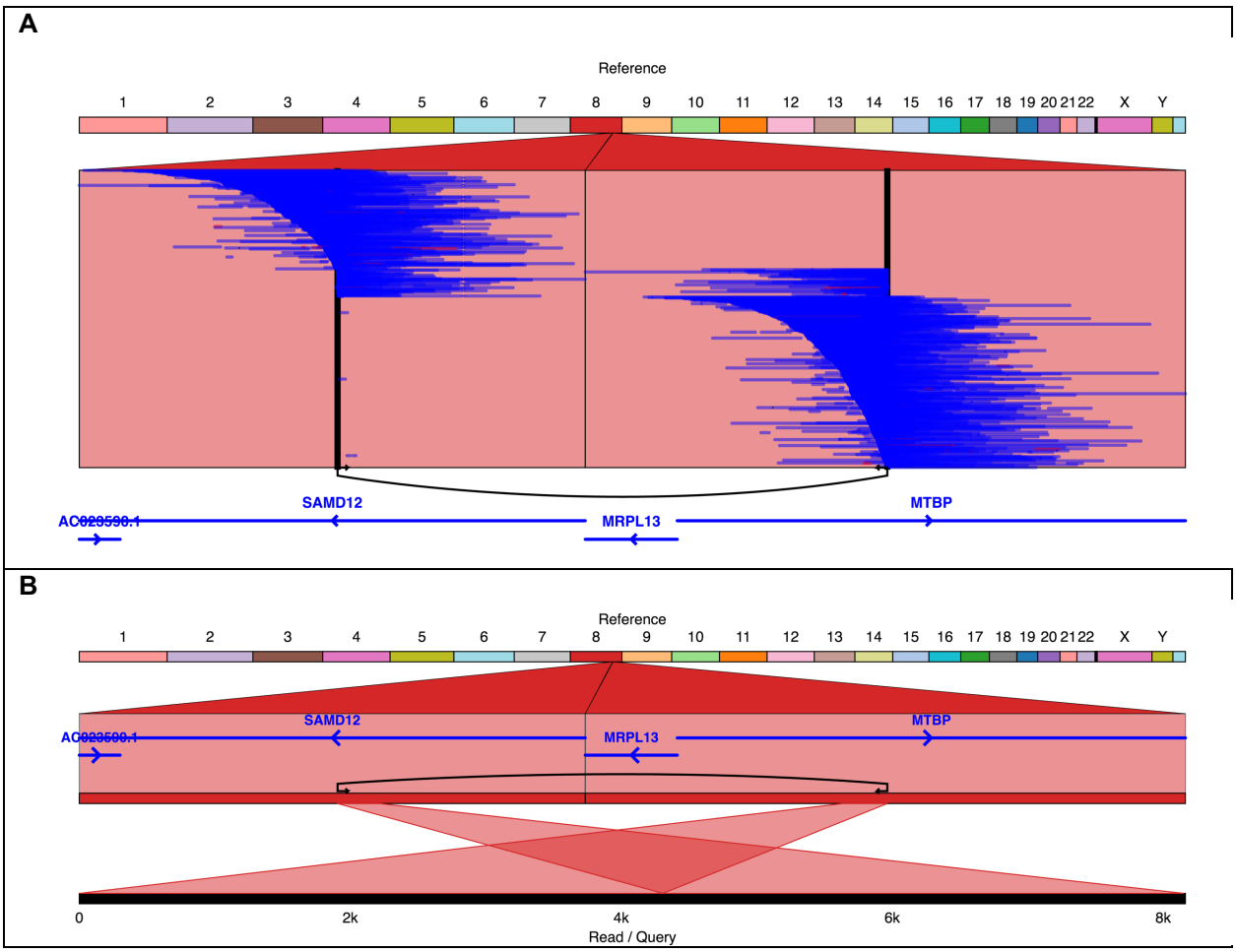


Supplementary Figure 21. Ribbon plot of GSDMB-TATDN1 gene fusion captured by long reads. This is a gene fusion in SK-BR-3 created by a single variant (A). This variant is captured in several individual SMRT sequencing reads, one of which is shown in (B).

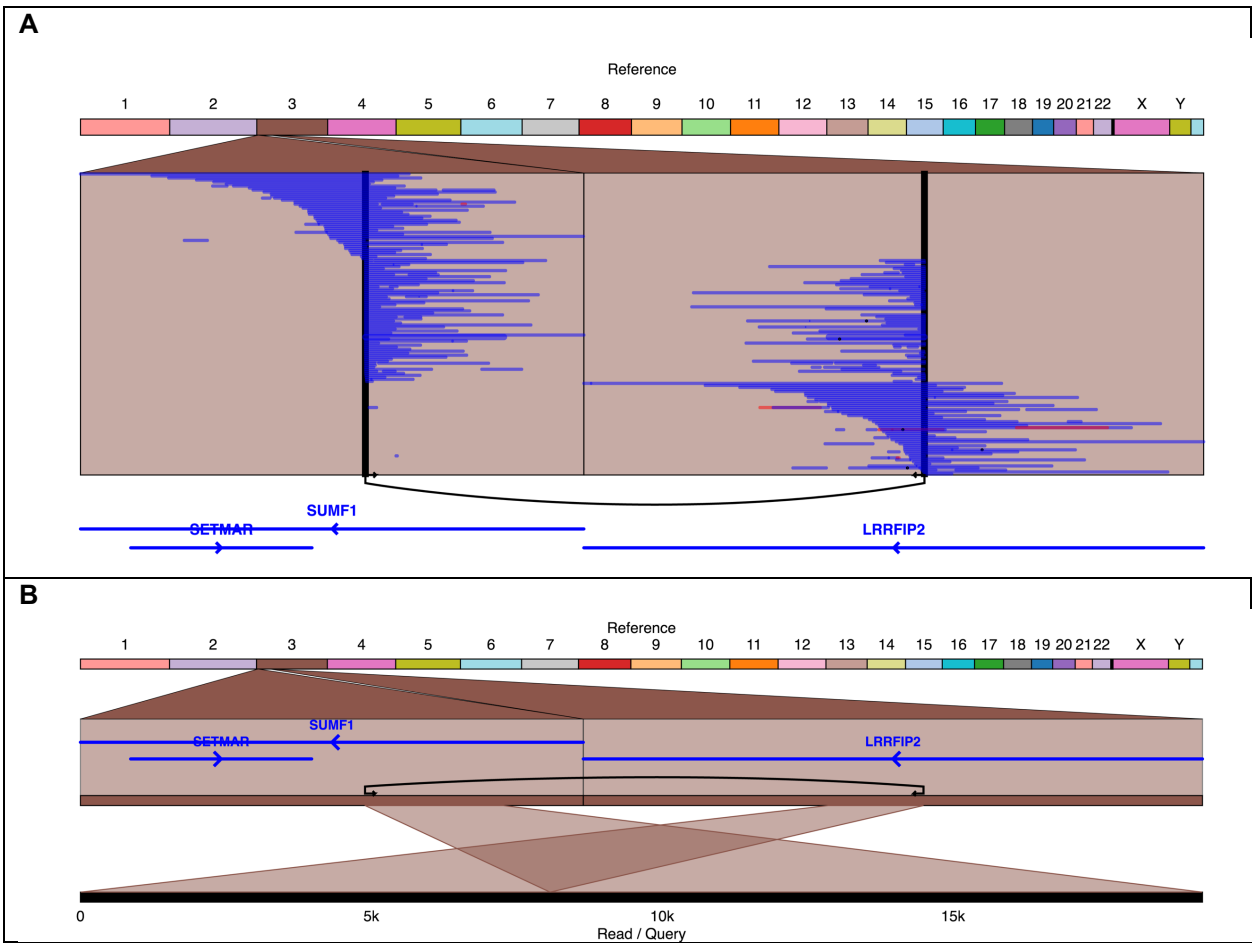

A
Reference
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
LINC00536
PVT1
B
Reference
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
LINC00536
PVT1
0
5k
10k
15k
Read / Query
C
Reference
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
RP11-1101K5.1
LINC00536
PVT1



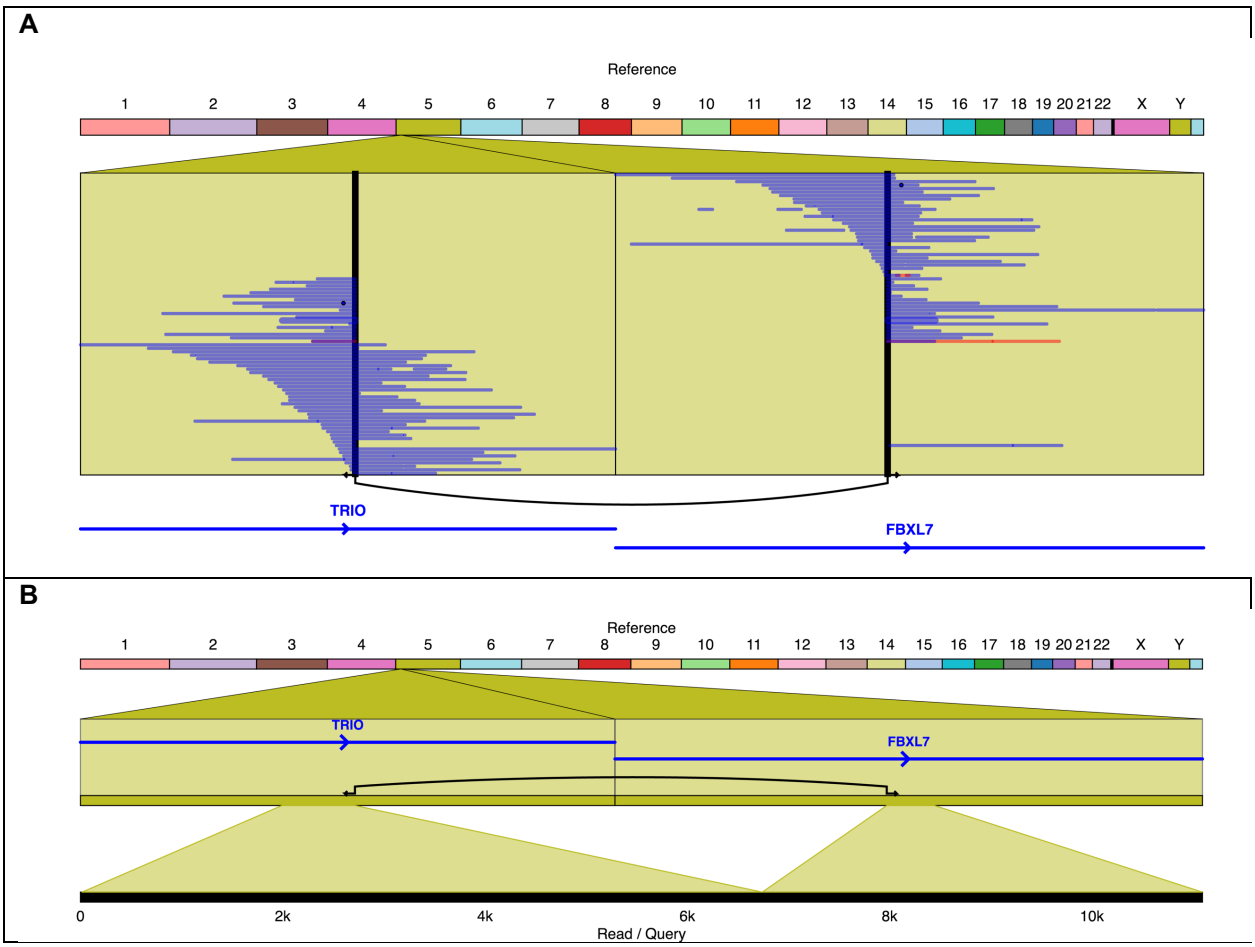
Supplementary Figure 22. Ribbon plot of LINC00536-PVT1 gene fusion captured by long reads. This is a gene fusion in SK-BR-3 created by two separate variants (A and C). One of the individual SMRT sequencing reads supporting the variant in (A) is shown in (B), while one of the reads supporting the variant in (C) is shown in (D).



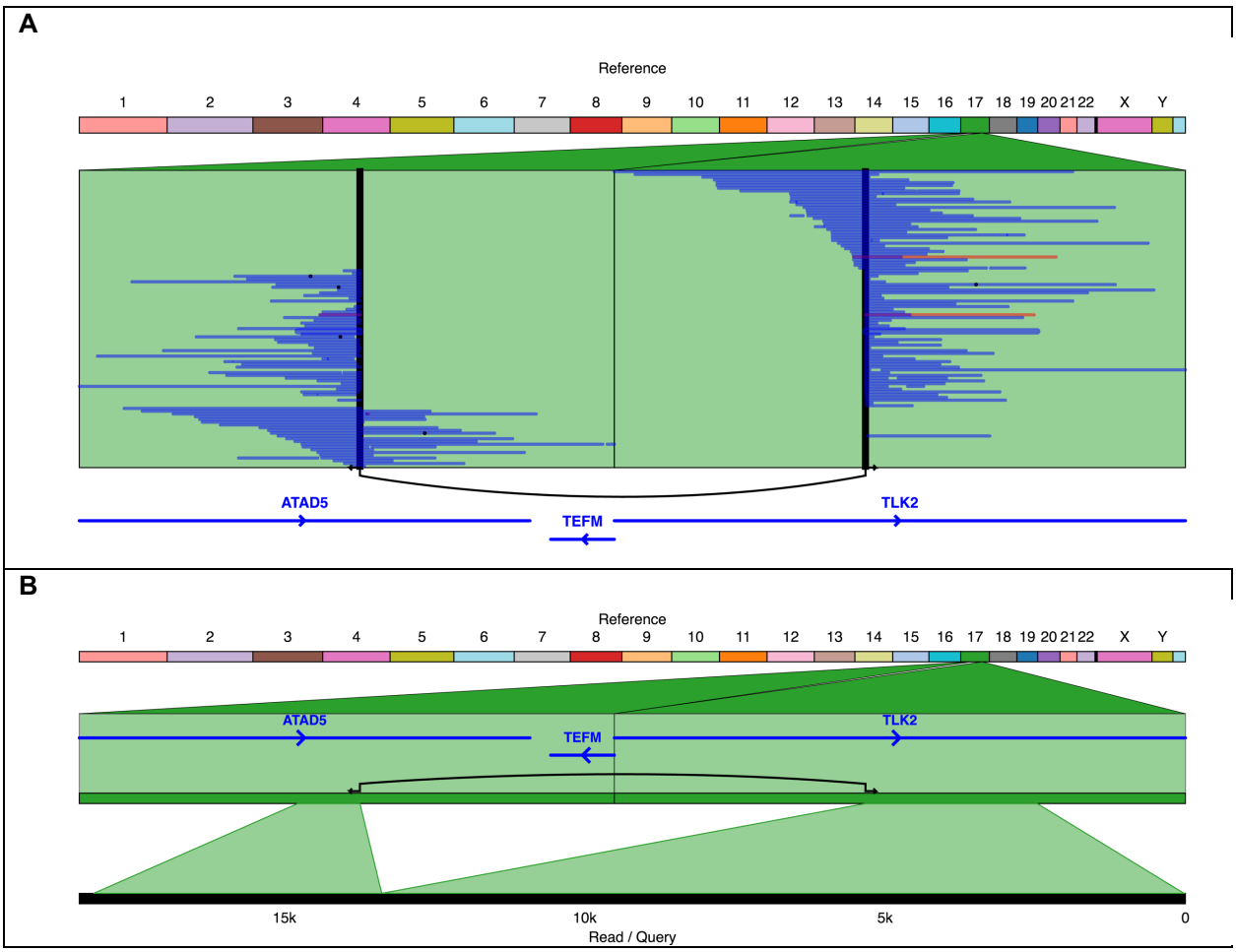
Supplementary Figure 23. Ribbon plot of MTBP-SAMD12 gene fusion captured by long reads. This is a gene fusion in SK-BR-3 created by a single variant (A). This variant is captured in several individual SMRT sequencing reads, one of which is shown in (B).



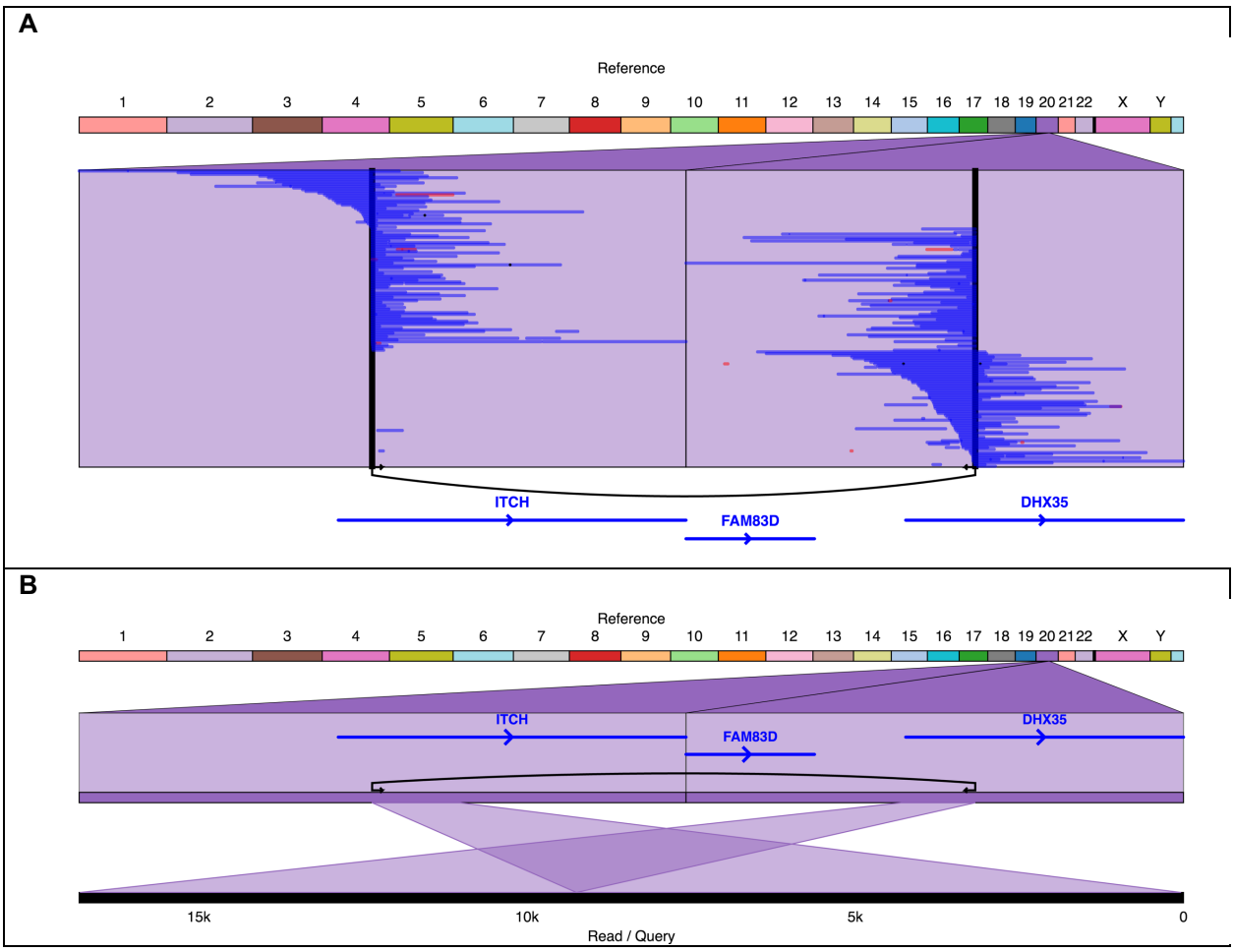
Supplementary Figure 24. Ribbon plot of LRRFIP2-SUMF1 gene fusion captured by long reads. This is a gene fusion in SK-BR-3 created by a single variant (A). This variant is captured in several individual SMRT sequencing reads, one of which is shown in (B).



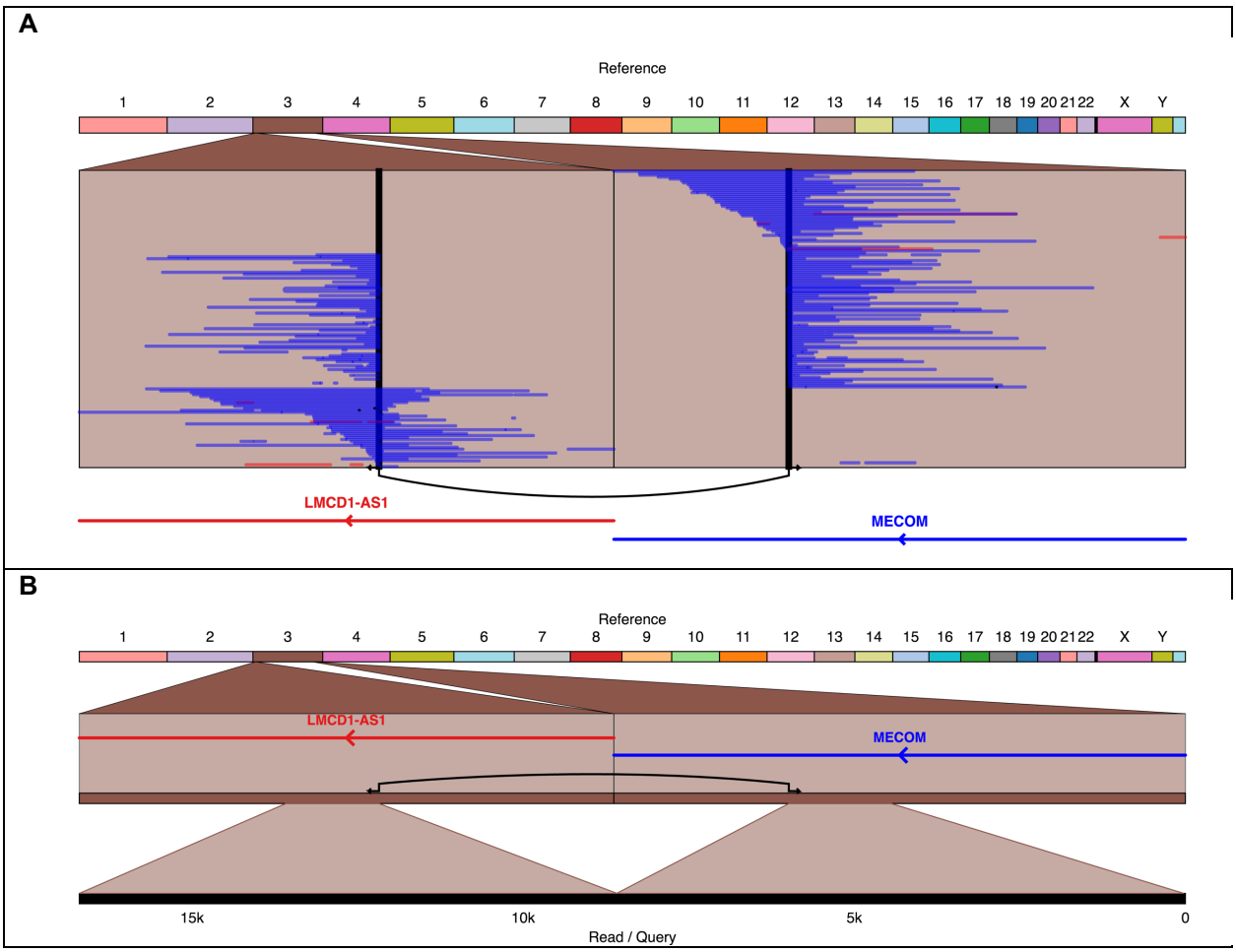
Supplementary Figure 25. Ribbon plot of FBXL7-TRIO gene fusion captured by long reads. This is a gene fusion in SK-BR-3 created by a single variant (A). This variant is captured in several individual SMRT sequencing reads, one of which is shown in (B).



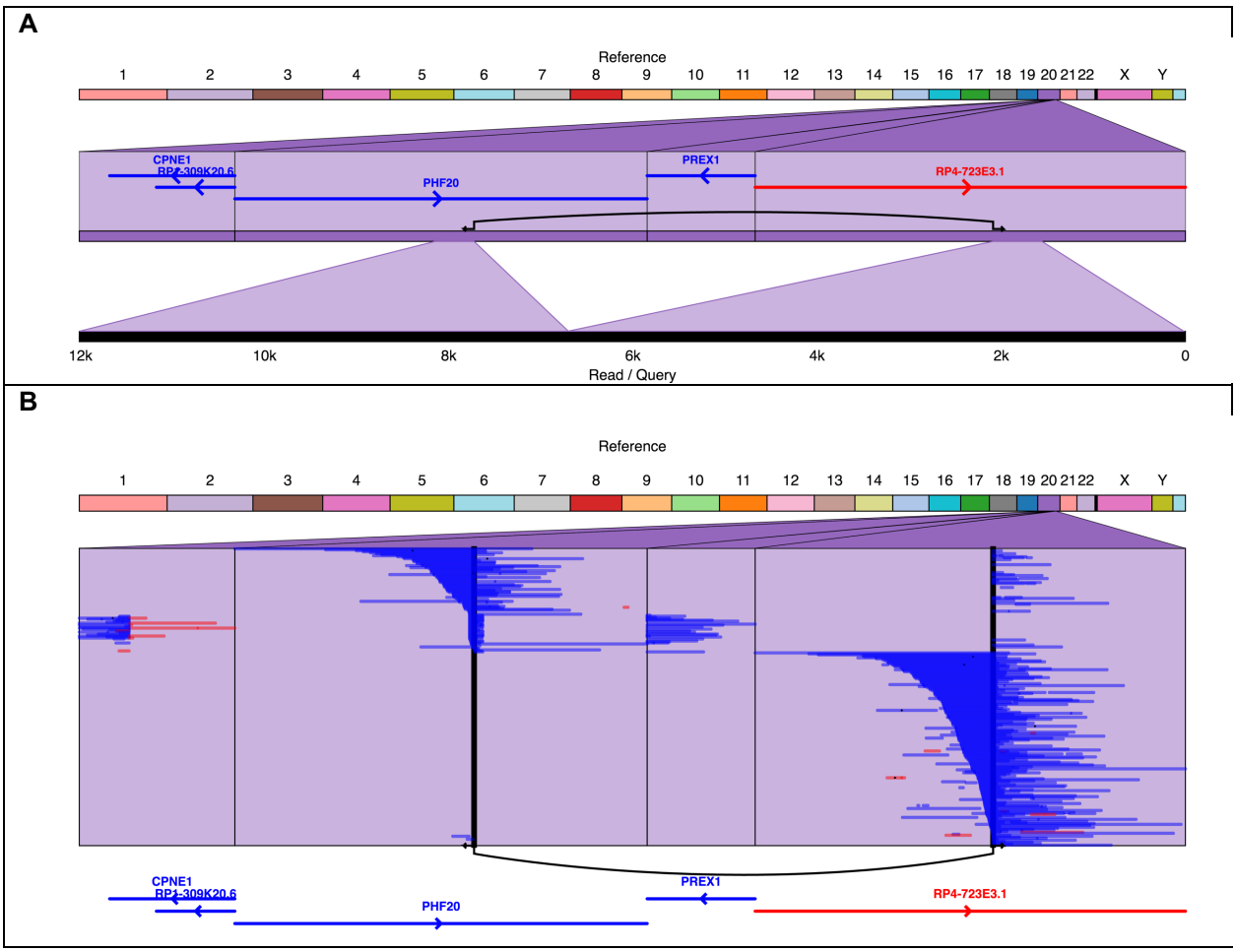
Supplementary Figure 26. Ribbon plot of ATAD5-TLK2 gene fusion captured by long reads. This is a gene fusion in SK-BR-3 created by a single variant (A). This variant is captured in several individual SMRT sequencing reads, one of which is shown in (B).



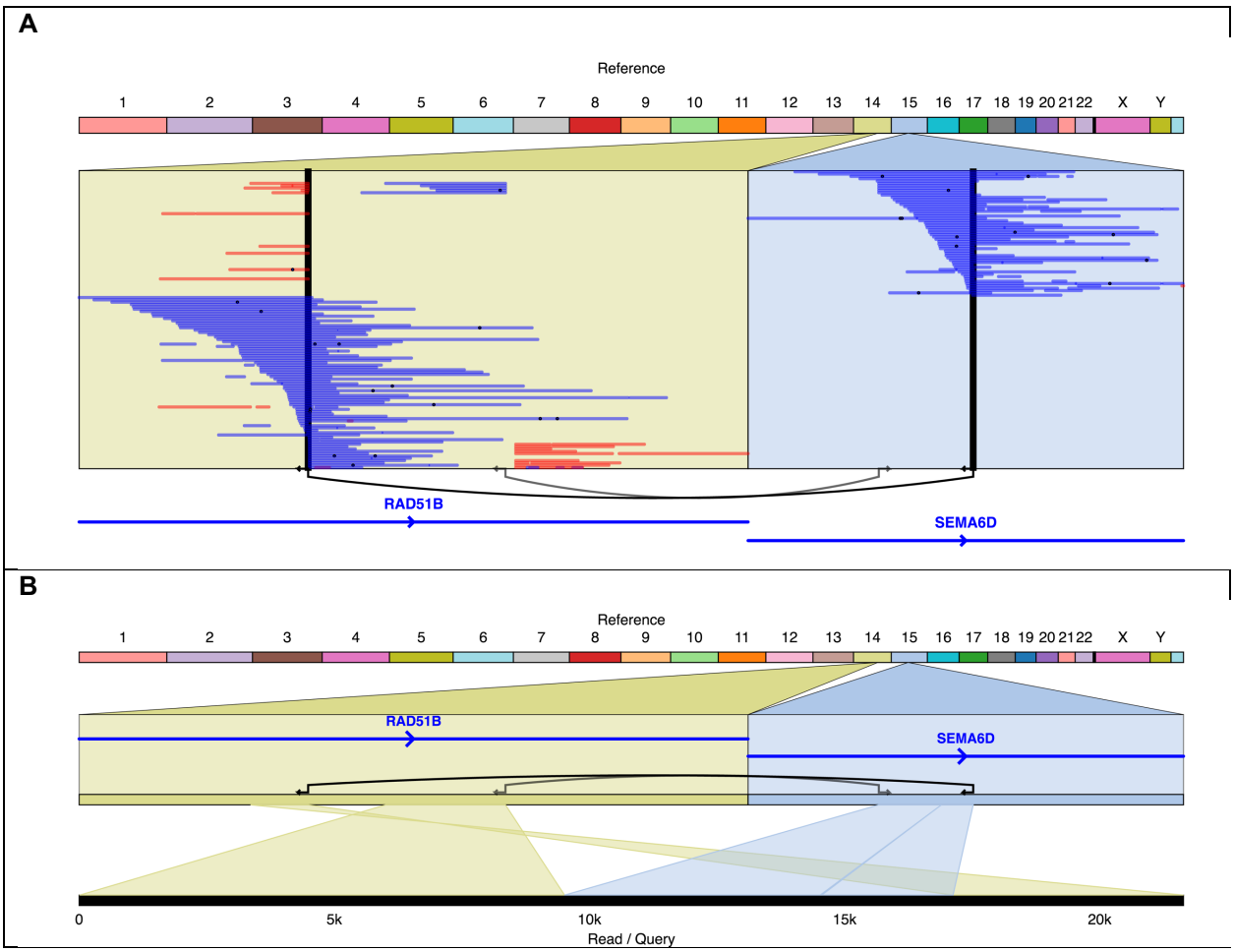
Supplementary Figure 27. Ribbon plot of DHX35-ITCH gene fusion captured by long reads. This is a gene fusion in SK-BR-3 created by a single variant (A). This variant is captured in several individual SMRT sequencing reads, one of which is shown in (B).



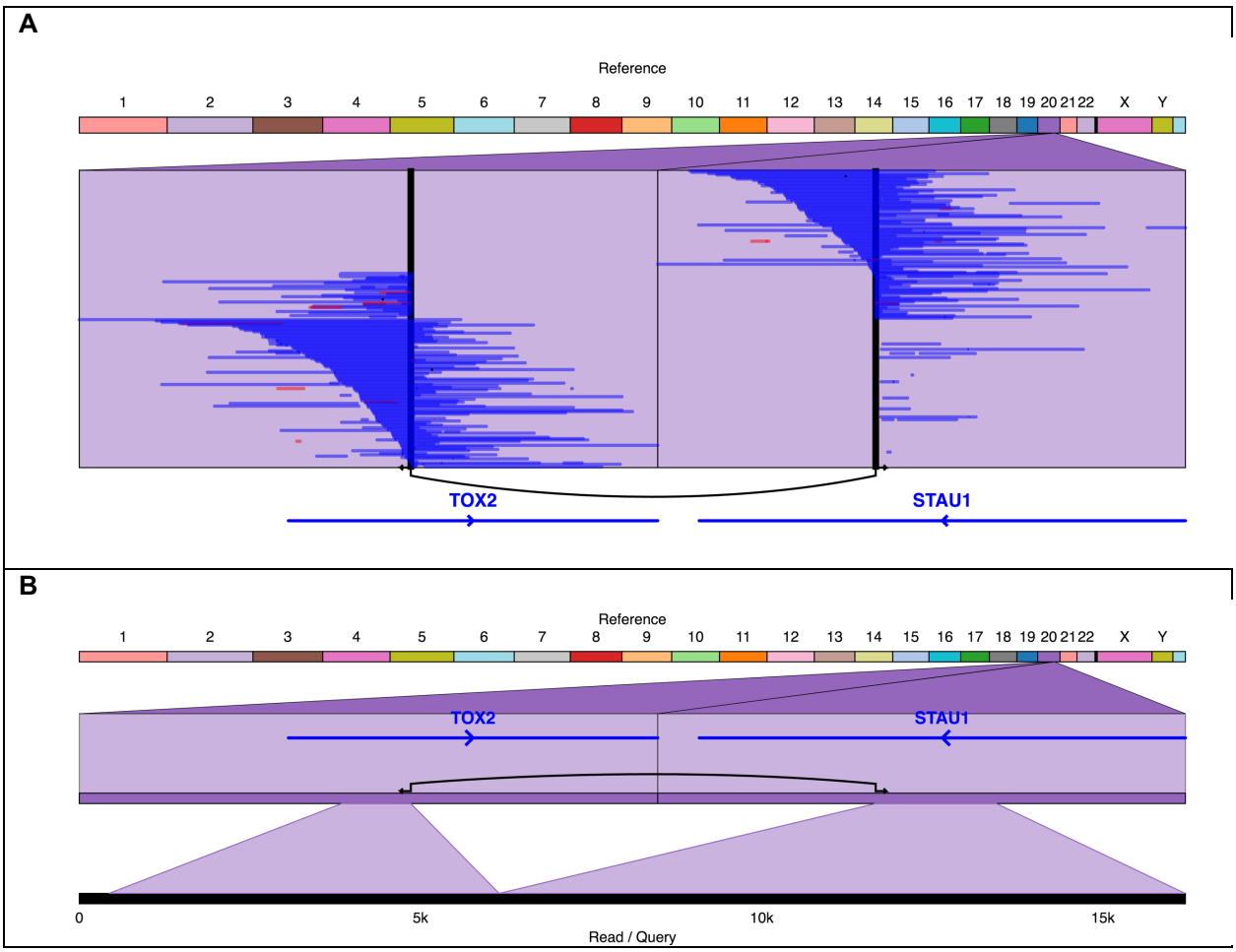
Supplementary Figure 28. Ribbon plot of LMCD1-AS1 – MECOM gene fusion captured by long reads. This is a gene fusion in SK-BR-3 created by a single variant (A). This variant is captured in several individual SMRT sequencing reads, one of which is shown in (B). MECOM is protein-coding while LMCD1-AS1 is antisense.



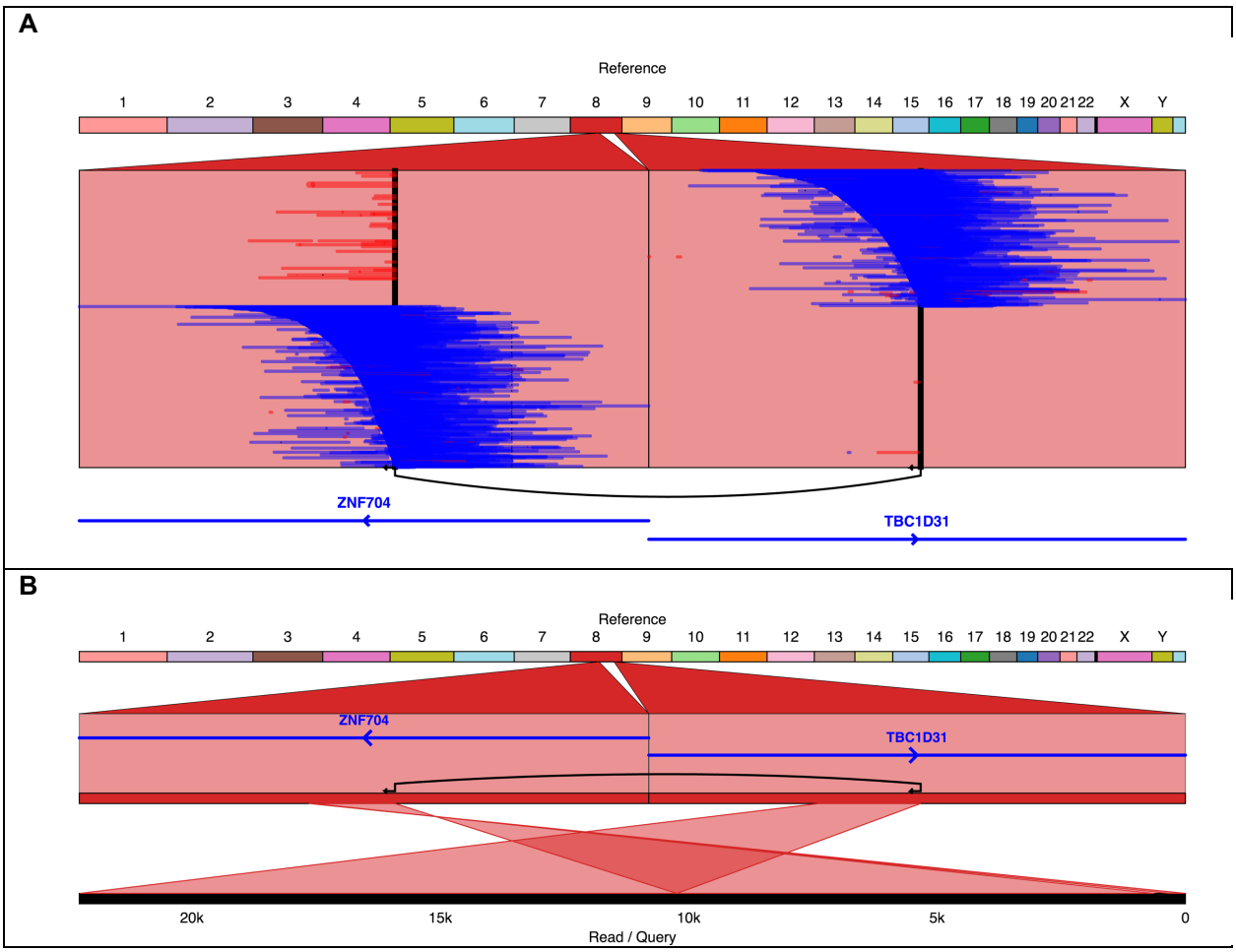
Supplementary Figure 29. Ribbon plot of PHF20 – RP4-723E3.1 gene fusion captured by long reads. This is a gene fusion in SK-BR-3 created by a single variant (A). This variant is captured in several individual SMRT sequencing reads, one of which is shown in (B). PHF20 is protein-coding while RP4-723E3.1 is a lincRNA.



Supplementary Figure 30. Ribbon plot of RAD51B-SEMA6D gene fusion captured by long reads. The gene fusion in SK-BR-3 created by two separate variants (A). Both of these variants are captured in several individual SMRT sequencing reads, one of which is shown in (B).



Supplementary Figure 31. Ribbon plot of STAU1-TOX2 gene fusion captured by long reads. The gene fusion in SK-BR-3 created by a single variant (A). This variant is captured in several individual SMRT sequencing reads, one of which is shown in (B).



Supplementary Figure 32. Ribbon plot of TBC1D31-ZNF704 gene fusion captured by long reads. The gene fusion in SK-BR-3 created by a single variant (A). This variant is captured in several individual SMRT sequencing reads, one of which is shown in (B).

Supplemental References

1. Chin, C.-S. *et al. Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing*. *bioRxiv* 056887 (Cold Spring Harbor Labs Journals, 2016). doi:10.1101/056887
2. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences* **108,** 1513–1518 (2011).
3. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5,** R12 (2004).
4. Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw369
5. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A Greedy Algorithm for Aligning DNA Sequences. *http://www.liebertpub.com/cmb* **7,** 203–214 (2004).