Ribbon: Intuitive visualization for complex genomic variation

Maria Nattestad, Robert Aboukhalil, Chen-Shan Chin, Michael C. Schatz

Supplementary Notes

Table of Contents

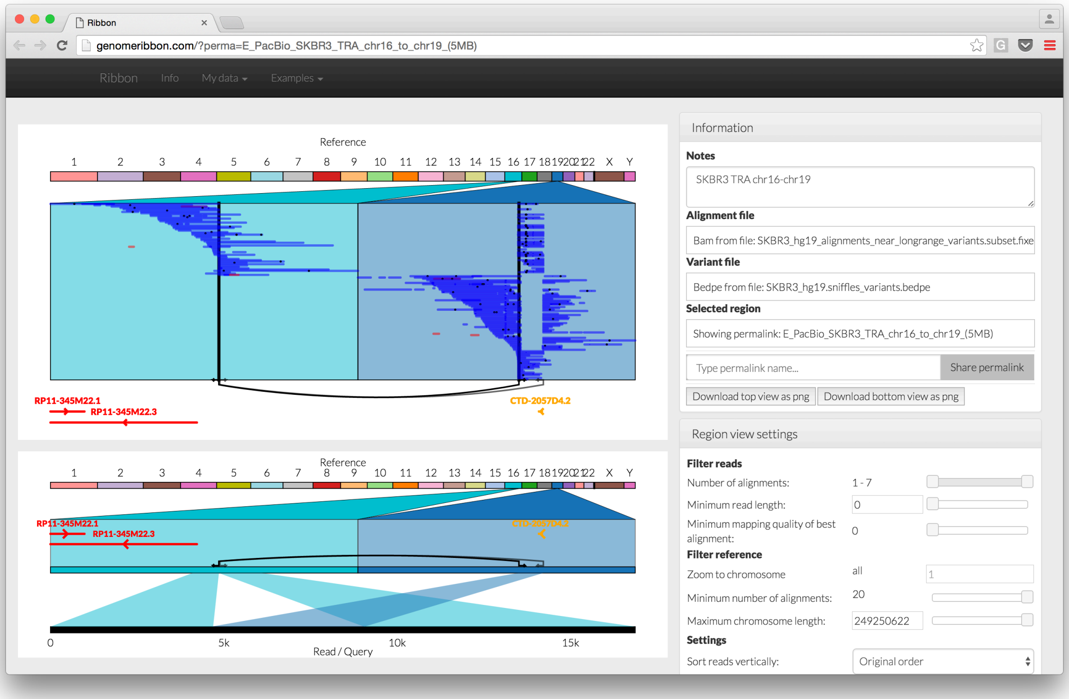
Supplementary Note 1: Ribbon features

Genomic visualization is now more important than ever for dissecting complex structural variations. However, if a variant connects two sequences that are far apart, most existing genome browsers such as IGV or the UCSC Genome Browser can only show one breakpoint at a time **(Supplementary Table 1)**. Even for smaller variants, split read alignments are not marked as being from the same read. Especially for long reads and assembled contigs, entire complex structural variants can be contained within a single query sequence, a perspective which is lost in the one-dimensional view of standard genome browsers. Ribbon is an alignment visualization tool that shows multiple regions of the reference in a single view and displays alignments both in the query sequence (read or assembled contig) and reference coordinates, giving the two-dimensional perspective of which part of the query maps where in the reference.

		IGV	Ribbon
Alignments	**Show alignments from a large BAM file quickly using an index**	yes	yes
	Can be used with long reads and short or paired-end reads	yes	yes
	Shows alignments to multiple reference chromosomes at once	no	yes
	Shows all alignments anywhere in the genome of all the reads in your region of interest	no	yes
	Indicates which part of a read maps where	no	yes
	Dot plot alignment view available	no	yes
	Multiple samples	yes	no
Variants	**Shows genes and any other features from a BED/VCF file**	yes	yes
	Shows long-range variants from a BEDPE file	no	yes
	Shows single nucleotide and small variants	yes	no
Navigation	**Navigation by scrolling**	yes	no
	Jump to a variant of interest	yes	yes
	Advanced sorting and filtering to find variants of interest	no	yes
	Jump to both breakpoints of a long-range variant	no	yes
Sharing	**Take a screenshot**	yes	yes
	Save a session and come back to it later	yes	yes
	Share an interactive snapshot with colleagues	no	yes

Supplementary Table 1. A comparison of features in IGV and Ribbon.

The main display of Ribbon contains two viewports, a *reference viewport* showing all alignments to the selected region(s) of the reference, and a *query viewport*, showing all alignments of a given sequence across the reference genome (**Supplemental Figure 1**). It also has a right side panel for inputs, settings, and filtering the data shown in those viewports.



genomeribbon.com/?perma=E_PacBio_SKBR3_TRA_chr16_to_chr19_(5MB)

Notes

SKBR3 TRA chr16-chr19

Alignment file

Bam from file: SKBR3_hg19_alignments_near_longrange_variants.subset.fixe

Variant file

Bedpe from file: SKBR3_hg19.sniffles_variants.bedpe

Selected region

Showing permalink: E_PacBio_SKBR3_TRA_chr16_to_chr19_(5MB)

Type permalink name...

Filter reads

Number of alignments: 1 - 7

Minimum read length: 0

Minimum mapping quality of best alignment: 0

Filter reference

Zoom to chromosome all 1

Minimum number of alignments: 20

Maximum chromosome length: 249250622

Settings

Sort reads vertically: Original order

Supplementary Figure 1 | Screenshot of Ribbon showing alignments of long reads from SK-BR-3. The reference viewport (top left) shows all reads/contigs aligned to the specified position(s) of the reference. The query viewport (bottom left) shows the details of an individual read or contig sequence aligned to the reference. The right side panel has various controls for filtering and querying the alignments.

1.1 Alignment inputs

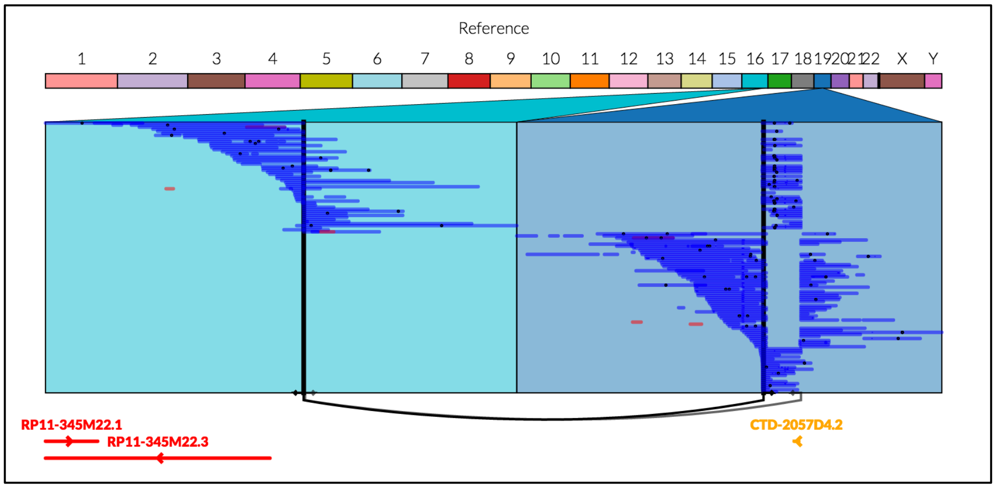
The inputs available include a BAM, SAM, and generic tab-delimited coordinate files. Coordinate files match those from MUMmer's show-coords utility[1] with the options '-lTH', which tells show-coords to add the lengths of the chromosomes and contigs (-l), make the output tab-delimited (-T) and exclude the header (-H). Since it is a simple tab-separated file, it can be generated from any aligner. This makes it extremely flexible so that any input type or custom filtering can be applied. There are no limitations on the size of the reference genome or number of sequences. For SAM/BAM files, the names and sizes of the reference chromosomes are taken from the header, and for coordinate files they are included in each row of the file.

BAM files are treated specially in Ribbon because the file is not actually uploaded, but rather Ribbon reads the header for reference information and then uses the BAI index file to fetch reads from particular regions of interest in the BAM file using the bam.iobio library [2,3] for remote bam files by URL and using samtools compiled to WebAssembly for local bam files. This greatly reduces the processing requirements for Ribbon, and makes it practical to use in low memory environments, including on mobile devices.

The location to display in the reference viewport can be selected manually, or by selecting variants to fetch all reads nearby (see "Optional variant and feature inputs" section). Other alignments for the same read are found using the SA tag in SAM/BAM data and shared read-names in the coordinates data. For example, when querying a BAM file for a small region on chromosome 1, all alignments will be shown for the reads with alignments within 100bp of that location. In addition, if any of those reads has an SA tag showing additional alignments to another location, either on that chromosome or another chromosome, then those additional genomic locations will also be shown in the viewport. This makes it possible to easily see whether other reads are sharing this secondary/supplementary alignment, especially for structural variation analysis (**Supplemental Notes 2-4**). Thus, for each BAM/SAM record, the main alignment listed on the line is shown along with all its other alignments (from the SA tag) anywhere in the genome. Ribbon has support for single reads as well as paired-end reads, and it automatically detects the type of data from the BAM/SAM flags, treating paired-end reads as shown in **Supplementary Figure 9**.

1.2 Reference viewport: Multi-read alignment view

The top viewport displays multiple query sequences (reads or contigs) from one of the inputs listed above, showing all alignments for each query across the whole reference genome **(Supplemental Figure 2)**. In IGV and most other visualization packages, multiple alignments from the same read are not associated in any way. In Ribbon, all alignments from the same read are shown in the same row of the display, making it straightforward to match up which reads are supporting a translocation or another multi-breakpoint variant. Alignments are shown in blue when they are mapped in the forward direction and red when mapped in the reverse complement direction. These orientations are optionally recalculated on the fly to make the main alignment of each read appear in the forward direction. For a coordinate file, the main alignment is by default the longest, while for a BAM/SAM file it is the one not appearing in the SA tag. By clicking on a read/query, the user can select a read to show in the single-read detail view below the reference viewport.

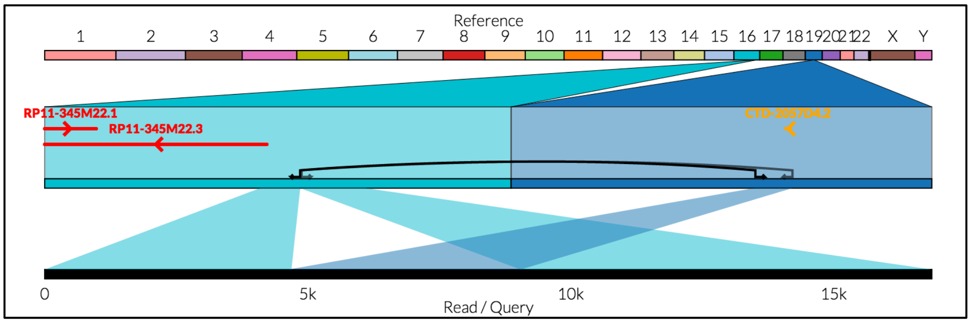


Supplementary Figure 2 | Reference viewport in Ribbon showing alignments in the forward (blue) and reverse (red) directions, zooming in on the parts of the reference that have alignments for these reads.

1.3 Query viewport: Single-read or single-contig detail view

There are many cases where it is important to know which part of a read aligns where, including for manually curating variant calls, inspecting how aligners and variant-callers are performing in or near complex variations, and visualizing the genes and other genomic features near identified variants. In this single-read query viewport, all of the alignments of the selected query sequence (read or contig) are shown along that query sequence (**Supplemental Figure 3**). This highlights the pieces of the reference that this sequence maps to, as opposed to the multi-read view that shows all pieces of the reference where any of the reads align.

When selecting between different query sequences for the viewport, Ribbon can optionally display all of the aligned reference sequences for that read or maintain the previous set of reference sequences. This option helps keep the context intact when switching between different reads, especially when inspecting alignments spanning complex structural variations. This view is meant to make it clear which part of the read aligns where in the reference, including if parts of the read map nowhere or if alignments are overlapping each other. This is particularly useful, for example, to inspect split read mappings flanking a structural variation, as the aligner may shift the specific position of the breakpoint depending on other errors or noise in the sequences.



Supplementary Figure 3 | Single-read query viewport in Ribbon showing a single read (black bar) with all of its alignments above.

1.4 Optional variant and feature inputs

After choosing a set of alignments to view, Ribbon has the option to upload a file containing structural variants in VCF or BED format. This will generate a table of the variants with advanced sorting and filtering options to help narrow down the file to some variants of interest. In both the multi-read and single-read views, variants show up as rectangles in the reference coordinate space, colored by variant type when available.

A BEDPE file can also be uploaded to show long-range variants, namely rearrangements that exist at two genomic locations, such as a fusion of two chromosomes or a translocation between sequences far away on a single chromosome. A BEDPE file can be obtained by running the variant-callers LUMPY[4] on short-read next-generation sequencing data or Sniffles[5] on long read sequencing data. VCF files containing long-range (2-breakpoint) variants can be converted to BEDPE using a Python script available at http://genomeribbon.com/longrange_vcf_to_bedpe.py, which can be run from the command line and parses the VCF output of various variant-callers (which differ greatly in how they report the second breakpoint) to reveal both breakpoints of the variants and enable querying from a BAM file at both positions. BEDPE input has a similar table for filtering and sorting rearrangements, which can also be used to query a BAM file at those positions. BEDPE rearrangements show up as connections between two genomic locations in the reference space, with arrows indicating the direction of read alignment gleaned from the strands in the BEDPE file.

Finally, genomic features can also be uploaded as a BED file, which could be used for instance to annotate genes or repetitive elements. These can be shown either as rectangles or with arrows indicating directionality, which is especially useful for genes **(Supplementary Figure 1-3).**

1.5 Filters, sorting, information, and sharing data

The panel on the right side contains a variety of controls to zoom, filter, get details, and otherwise interact with the data. Using the panel, users can narrow down the reference positions shown based on chromosome length or number of alignments to each position. User can also filter query sequences (reads or contigs) shown according to length, number of alignments, or mapping quality. They can also select specific chromosomes or reads using a live search feature or clicking on them in the main view. Reads can be sorted and oriented based on various criteria. In the single-read view, alignments can be filtered by mapping quality or length, and the view can be transformed from a ribbon plot to a dot plot. In the multi-read view insertions and deletions contained within the CIGAR string (SAM/BAM only) can be shown with different marks, including an option to show their sizes. The single-read query view can also show insertions and deletions, which here are indicated by splitting the alignments. This makes it possible to show the relative sizes of the events, even of insertions.

1.6 Availability and Requirements

Ribbon is available on the web at http://genomeribbon.com, and the code is open-source at https://github.com/marianattestad/ribbon. Ribbon can be deployed locally by following the instructions available on the GitHub page.

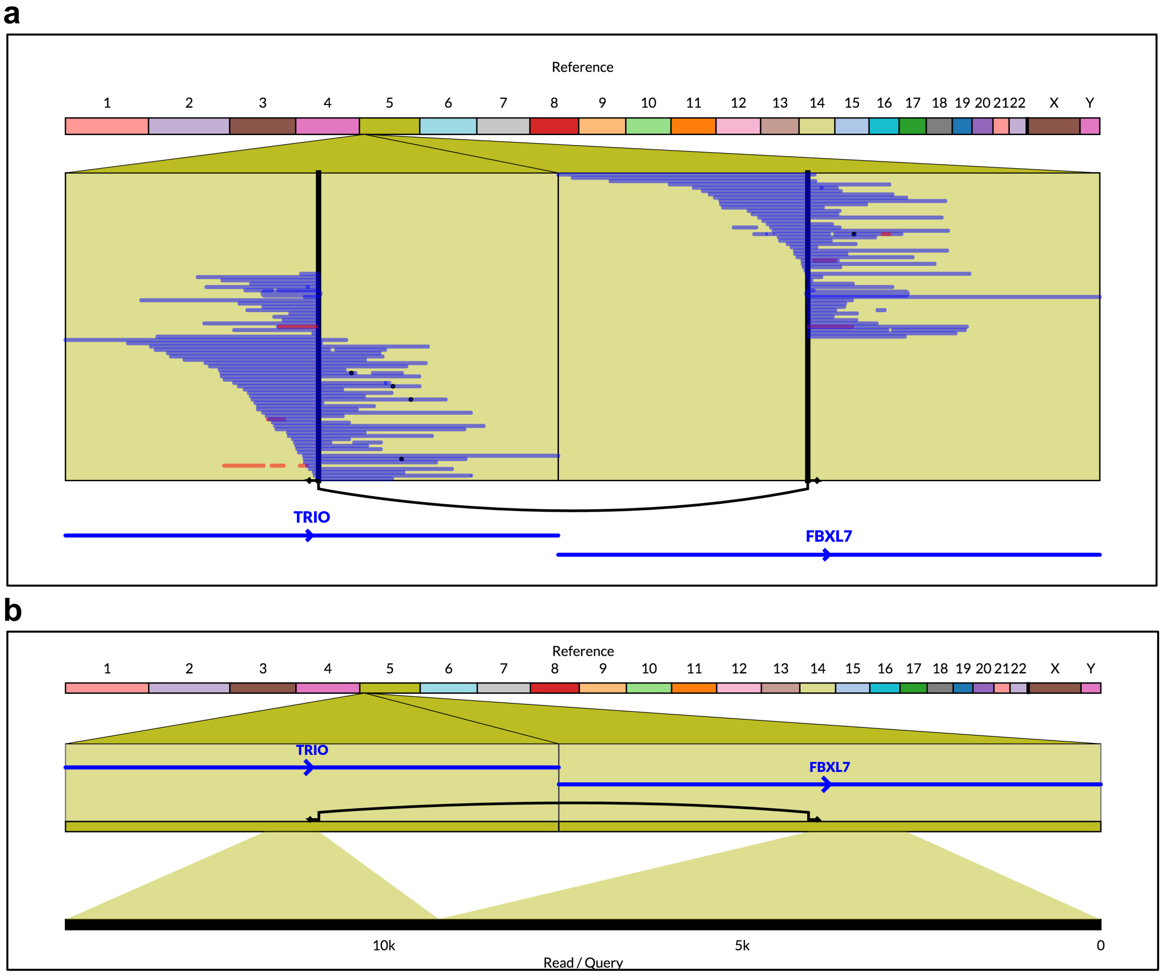
Supplementary Note 2: Using Ribbon to investigate complex gene fusions in a cancer genome

The SK-BR-3 breast cancer cell line has a highly rearranged and amplified genome, and it is one of the most commonly studied breast cancer cell lines, including basic and pre-clinical research into Her2-amplification and in several studies of gene fusions. We recently sequenced the whole SK-BR-3 genome using PacBio SMRT sequencing to an average coverage of 72X with an average read length of 9 kb[6]. Read alignments were computed using BWA-MEM with the PacBio option, and variant-calling was done using Sniffles[5]. We also performed IsoSeq long-read sequencing of the transcriptome to identify gene fusion candidates. Gene fusion candidates were evaluated by SplitThreader[7] according to whether genomic variants exist that link the putative fusion genes.

Here we investigate the genomic evidence behind a number of these putative gene fusions supported by both IsoSeq and SplitThreader, some of which were also found in the transcriptome using RNA-sequencing in previous studies[8-10]. Ultimately, it is crucial to see the specific read alignments supporting each variant in a call set for purposes of manual curation, which was the motivation behind creating Ribbon. Using Ribbon, we analyzed several of the gene fusions found in the extremely rearranged genome of SK-BR-3 in order to see the alignment evidence behind the variant calls and evaluate whether any of these fusions are involved in more complex events.

2.1 TRIO-FBXL7 gene fusion

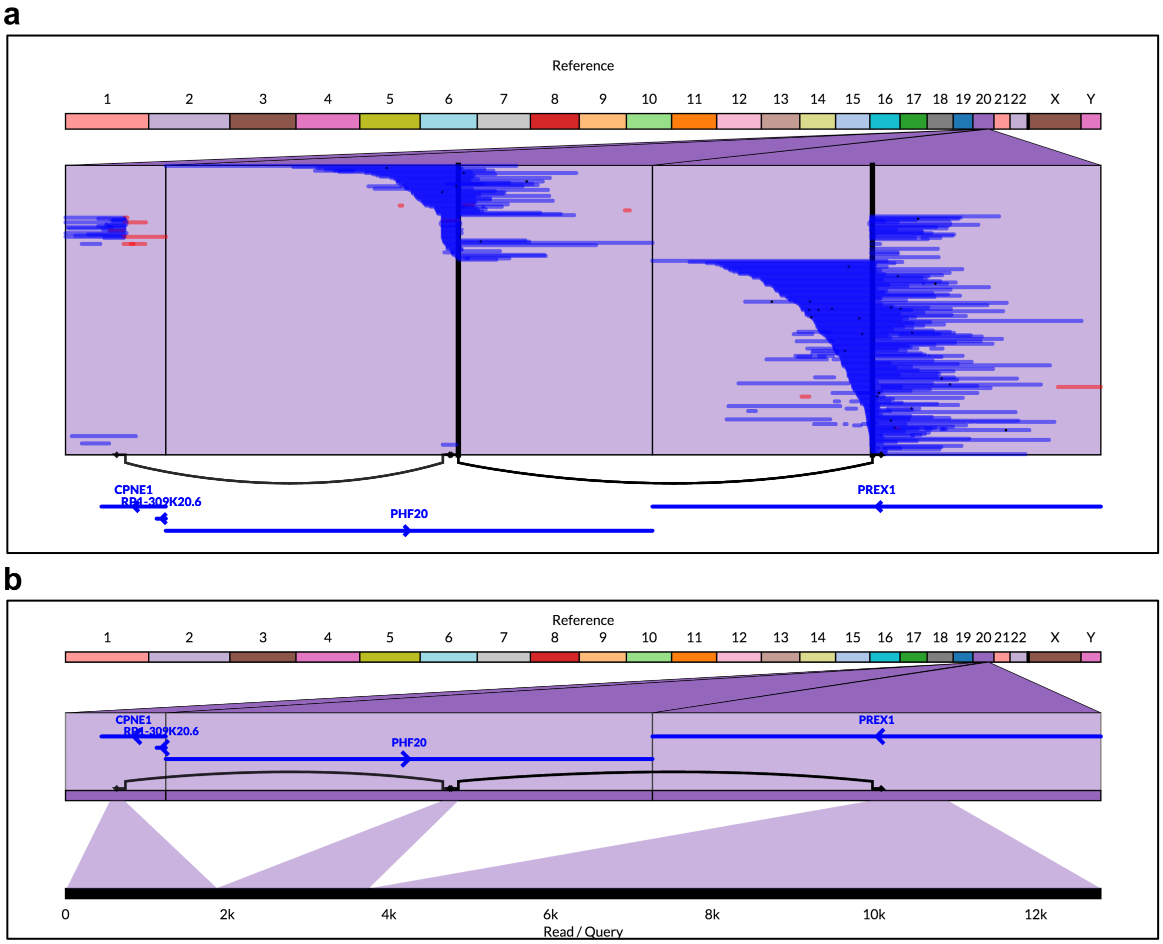
One of several simple gene fusions in SK-BR-3 is TRIO-FBXL7, which takes place within chr5 as a large deletion **(Supplementary Figure 4).** This is clearly shown in the reference and query viewports in Ribbon by a number of "split-read" alignments spanning across the deletion boundary.



Supplementary Figure 4 | TRIO-FBXL7 gene fusion. This is an example of a simple 1-step gene fusion taking place through a single variant, shown in the multi-read (a) and single-read (b) views it is clear to see the split reads supporting this gene fusion. Most gene fusions observed in SK-BR-3 are direct gene fusions through only one variant.

2.2 PREX1-PHF20-CPNE1 triple gene fusion

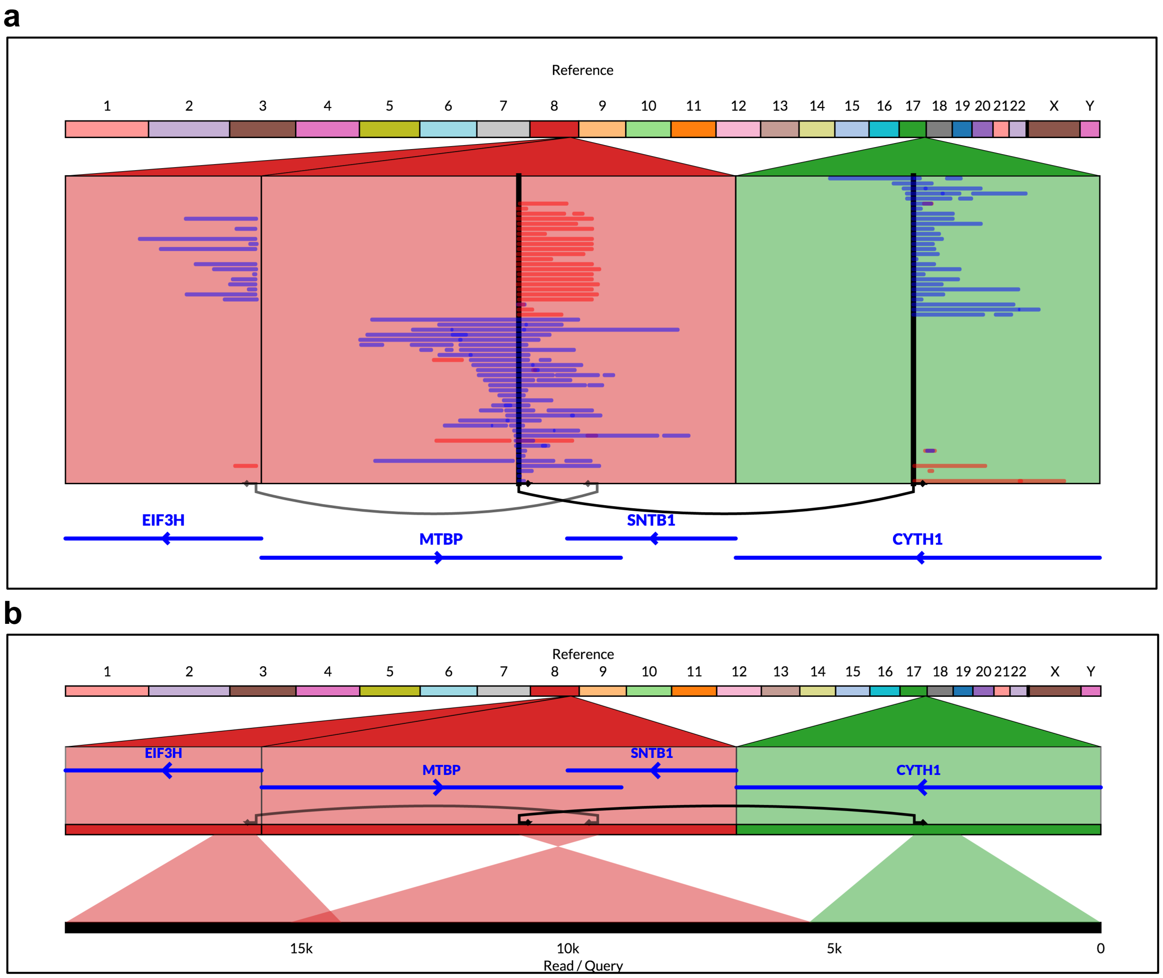
In addition to the simple gene fusions, some of the fusions identified by SplitThreader are two-step fusions, meaning the genes are linked through a series of two variants. PREX1-PHF20-CPNE1 has been previously shown as a triple gene fusion found through RNA-seq with the genomic link validated using PCR[9]. Thanks to a combination of long-read sequencing and Ribbon's visualization that shows all alignments for each read, Ribbon identifies, for the first time, the individual reads capturing the entire triple gene fusion **(Supplementary Figure 5)**.



Supplementary Figure 5 | PREX1-PHF20-CPNE1 triple gene fusion in SK-BR-3. The multi-read view (a) shows all the reads with alignments near the variants in PHF20 and PREX1, including several that have alignments to both of them in addition to CPNE1. The single-read view (b) shows one of the reads that has alignments to all three genes and contains both of the variants in this 2-step gene fusion.

2.3 CYTH1-MTBP-EIF3H triple gene fusion

We discovered a second triple gene fusion in SK-BR-3: CYTH1-MTBP-EIF3H, which has only been observed previously as a link between CYTH1 and EIF3H in RNA-sequencing. In one study, researchers looked for a genomic link between CYTH1 and EIF3H and found nothing, citing low coverage[9]. Until now, there were no previous reports of a genomic link between CYTH1 and EIF3H in the SK-BR-3 cell line. For this fusion, we first found IsoSeq evidence matching the observation that CYTH1 and EIF3H are fused in the RNA, but also that CYTH1 and MTBP have a fusion transcript with lower expression than the CYTH1-EIF3H fusion. SplitThreader found two variants that link CYTH1 and EIF3H through MTBP (**Supplementary Figure 6)**. In addition to seeing the evidence behind each of the two variants, we identify reads in Ribbon that contain both variants and therefore show direct evidence of a genomic link between all three genes.



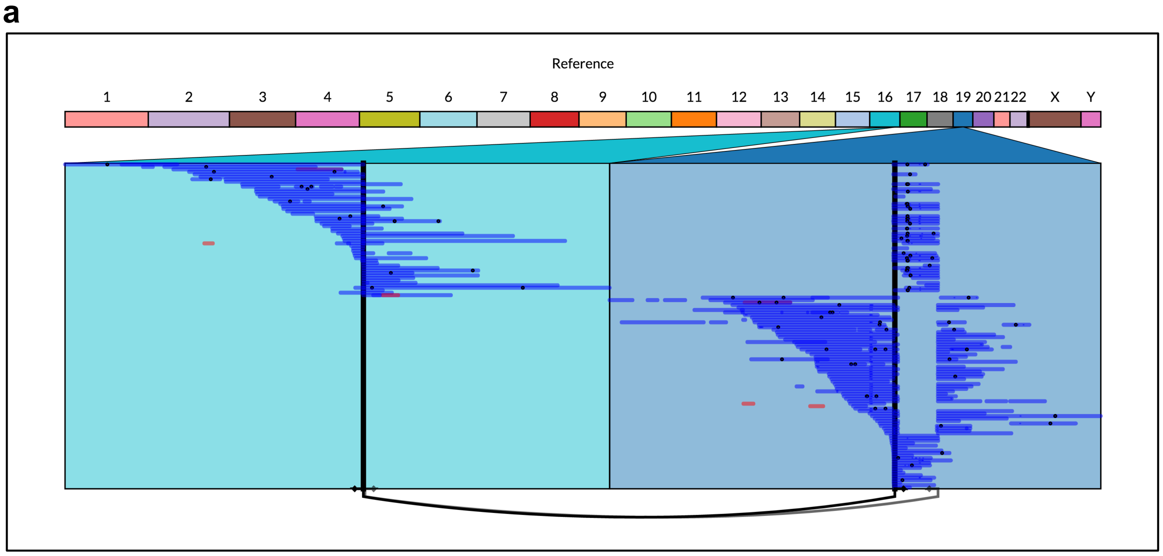
Supplementary Figure 6 | CYTH1-MTBP-EIF3H gene fusion in SK-BR-3. The multi-read view (a) is filtered to show only reads with 3 alignments in order to highlight the reads supporting this fusion. This gene fusion takes place through two variants, both of which are captured several individual reads, one of which is shown here (b).

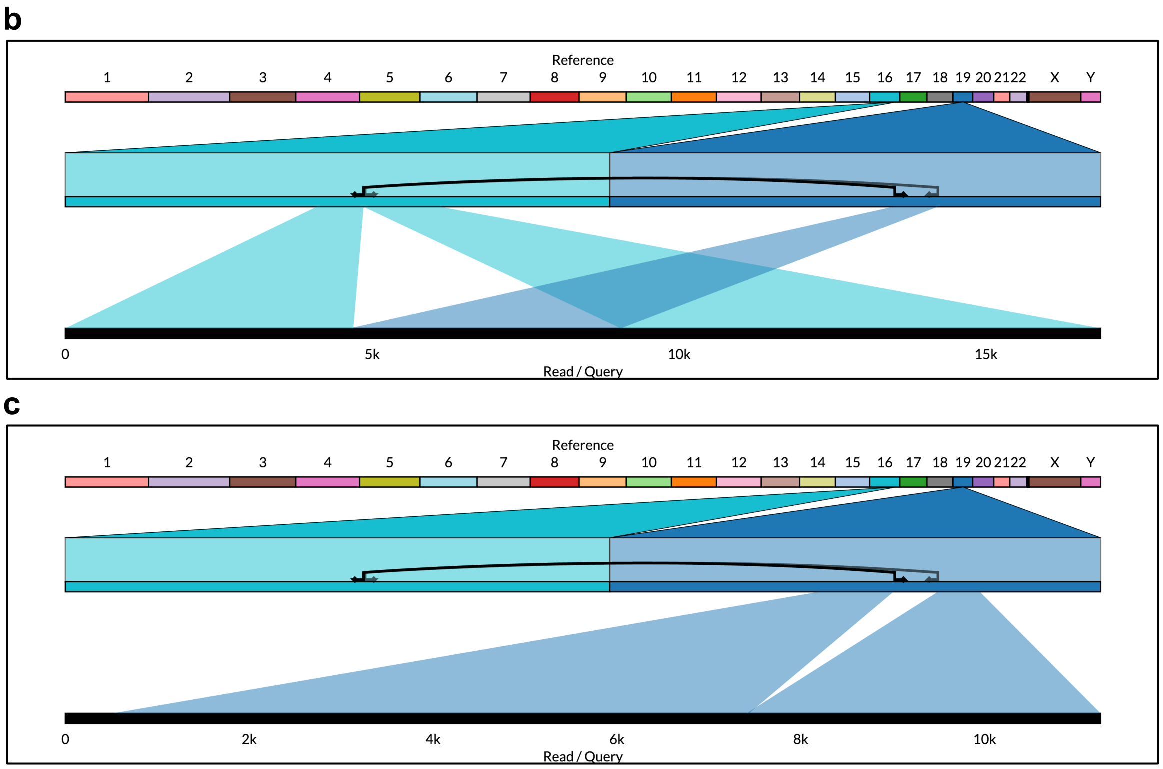
Supplementary Note 3: Complex variant signatures in Ribbon

Here we highlight two specific genomic variants in the SK-BR-3 breast cancer cell line found using SMRT long-read sequencing, BWA-MEM alignment, and Sniffles variant-calling as outlined in **Supplementary Note 2**. We also show a deletion found in the A549[11] non-small cell lung carcinoma (NSCLC) cell line using Illumina paired-end sequencing. The purpose of this section is to showcase the signatures of different variant types that can be inspected with Ribbon, viewing alignments from both long-read and short-read paired-end sequencing.

3.1 A canonical translocation

Supplementary Figure 7 shows a canonical translocation where a sequence has been excised from chr19 (between positions 30,388,930 and 30,393,102 bp) and inserted into chr16 (at position 79,864,900 bp). This event is made up of two variants, one supported by 16 split reads the other by 17 split reads. Looking at these breakpoints separately in a one-dimensional genome browser such as IGV, it is not apparent that the insertion of this sequence into chr16 is captured entirely within some of the reads, such as the one shown in **Supplementary Figure 7b**. The same sequence inserted into chr16 is also the one that is deleted in chr19, which **Supplementary Figure 7a** clearly shows in the multi-read view. It is also apparent that both the insertion and the deletion are homozygous, since there are no reads in the chr19 locus without the deletion and none in the chr16 locus without the insertion. Ribbon simultaneously enables a high-level overview and a detailed view of individual reads, showing that this variant is a real translocation where a sequence was actually excised from one location and inserted into another.

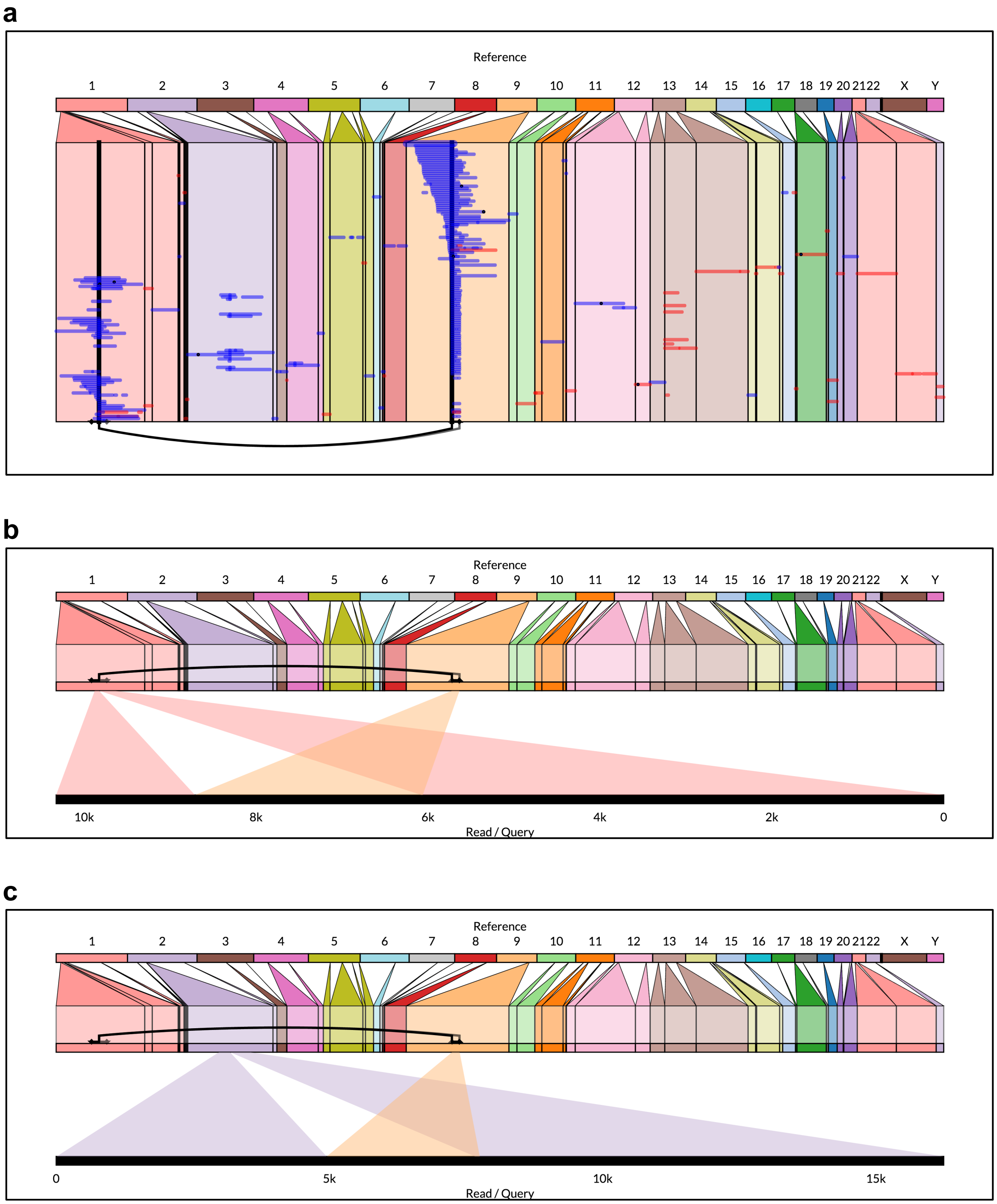


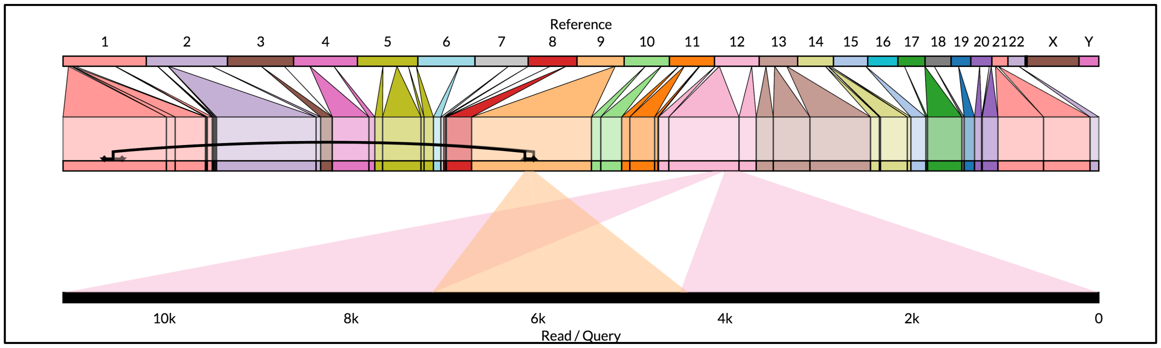


Supplementary Figure 7 | A classic translocation showing both deletion of a sequence in chr19 and insertion of the same sequence into chr16. (a) Multi-read view. (b) Single-read view of the insertion locus. (c) Single-read view of the deletion locus.

3.2 Interspersed duplications

Supplementary Figure 8 shows an interspersed duplication variant indicating that the sequence from chr9 (between positions 114,404,939 and 114,402,124 bp) is duplicated and inserted into chr1 (at position 16,329,400 bp), and we find good support for this variant in the alignments. Similar to the translocation above, this event is made up of two variant calls, one supported by 10 and the other supported by 11 split reads. In contrast to the translocation, the original sequence has not moved. Initially only the insertion into chr1 was detected, but upon further inspection of the alignments using Ribbon, the same sequence was found to be inserted into chr2 (at position 65.7 Mbp) and chr12 (at position 14.3 Mbp) as well, with very clean breaks between alignments found in all reads showing this insertion into the three different regions.

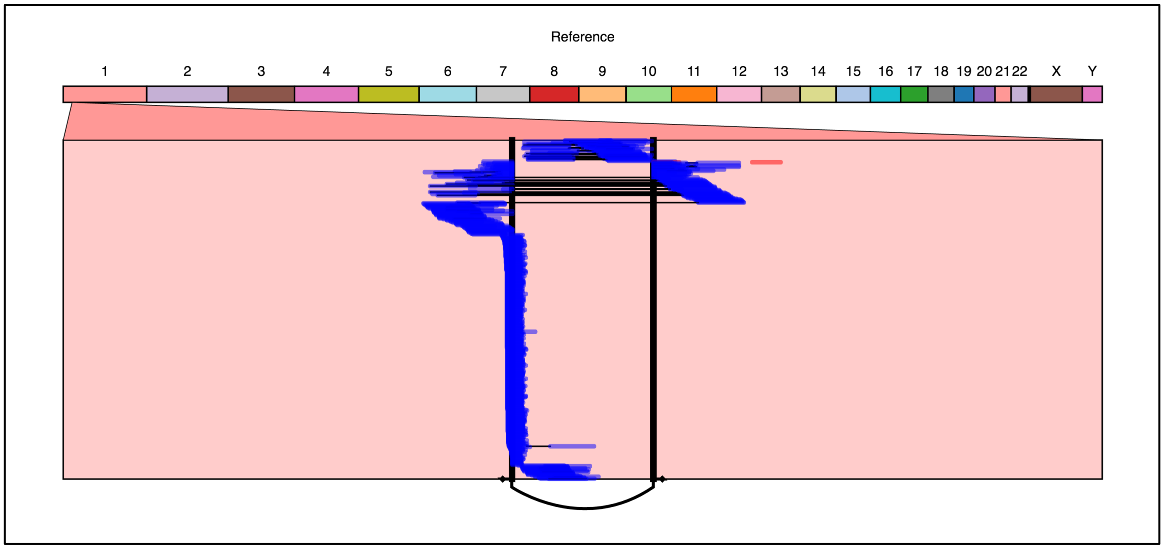

a
Reference
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
b
Reference
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
10k 8k 6k 4k 2k 0
Read / Query
c
Reference
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
0 5k 10k 15k
Read / Query



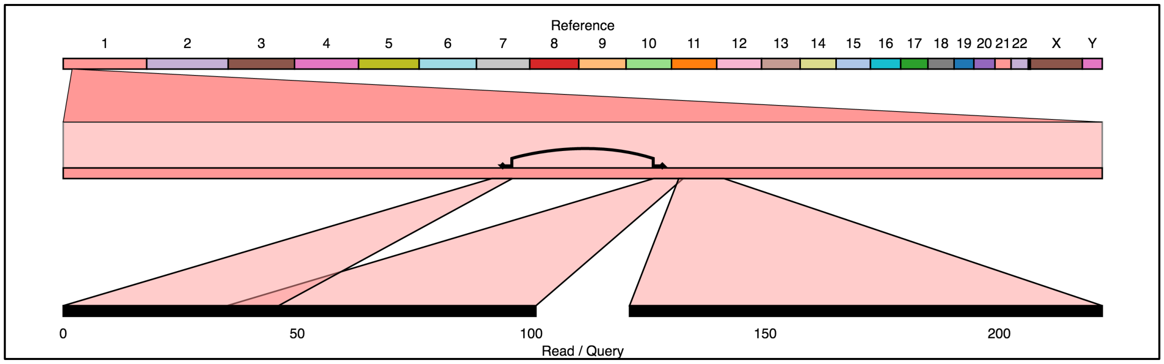
Supplementary Figure 8 | Insertions of the same sequence from chr9 into three different places on chr1 (b), chr2 (c), and chr12 (d). The variant used to navigate to this region was the insertion into chr1 (a).

3.3 Microhomology-mediated deletion

A549 is a widely studied non-small cell lung carcinoma cell line. Here we show a 315bp deletion found on chr1 from previously published Illumina paired-end sequencing data, using SpeedSeq[9] to align the reads and call variants. The deletion shows signs of possible microhomology at the breakpoint: the 11 bp overlap in alignments on the read suggest sequences on both sides of the breakpoint in the reference match each other, since they both align to the same sequence on the read. Microhomology has been previously shown to mediate deletions, making the overlap in alignments a critical feature for understanding this variant.



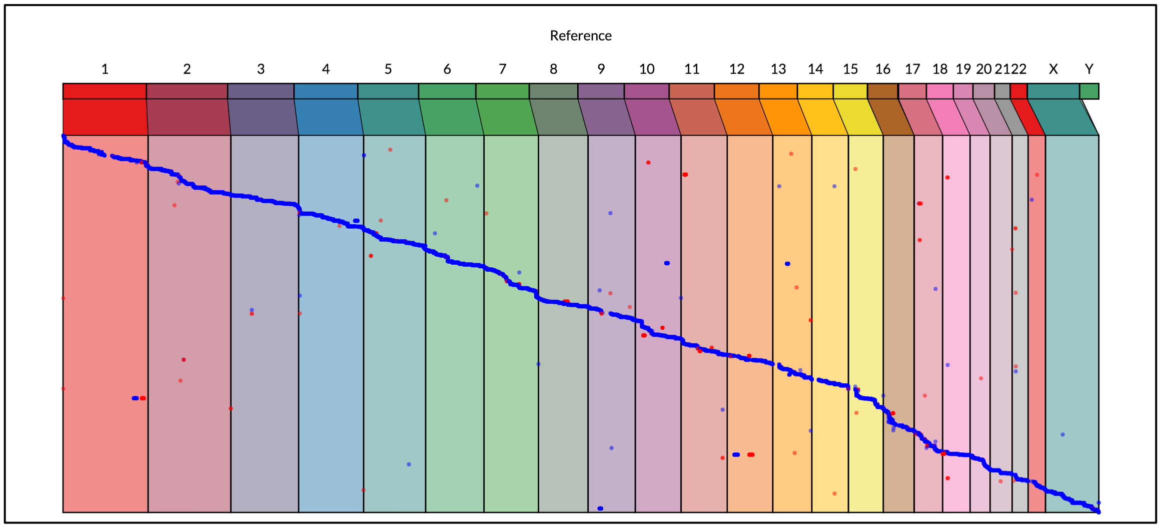
b



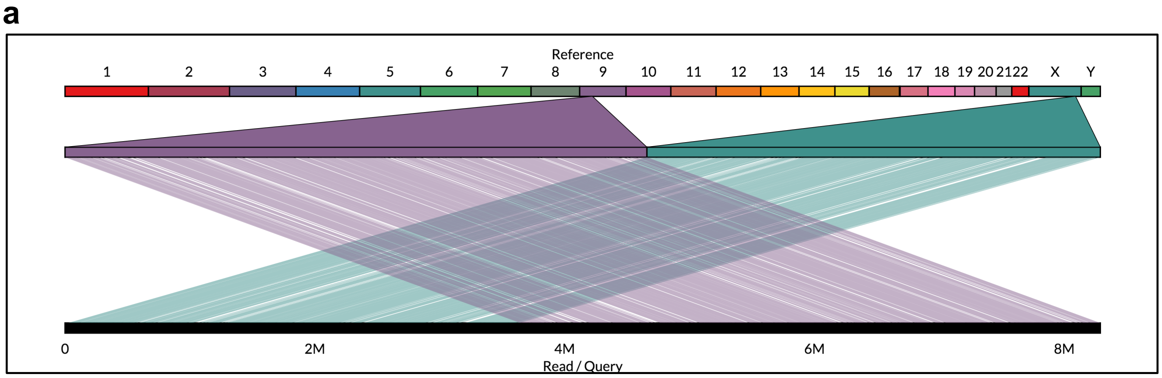
Supplementary Figure 9 | A deletion with microhomology at the breakpoint captured with Illumina paired-end sequencing. Paired-end reads are shown with a black connecting line in the reference view (a) and as two black bars in the query view (b).

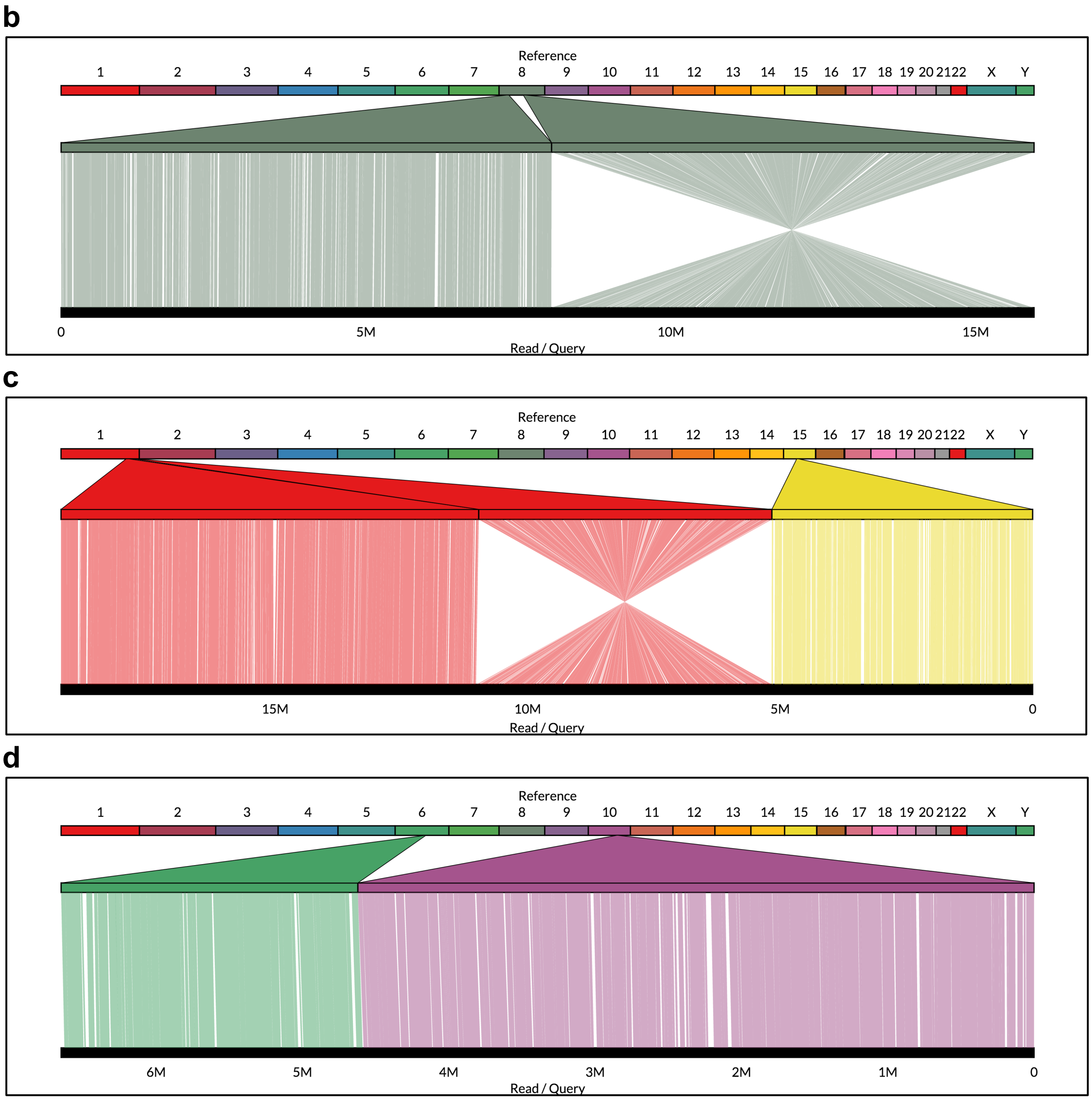
Supplementary Note 4: Using Ribbon to compare the gorilla and human genomes

The latest gorilla genome assembly[12] Susie3 was aligned against the GRCh38 human reference genome using MUMmer[1]. An overview of the alignments is given in **Supplementary Figure 10**, showing how much of the sequence of the human genome has some alignment to the gorilla genome. From this view, it also appears that Susie3 has no alignments to the human chrY, as expected since the gorilla is female. We also showcase some of the major structural differences between the genomes of these related species using Ribbon in **Supplementary Figure 11**.



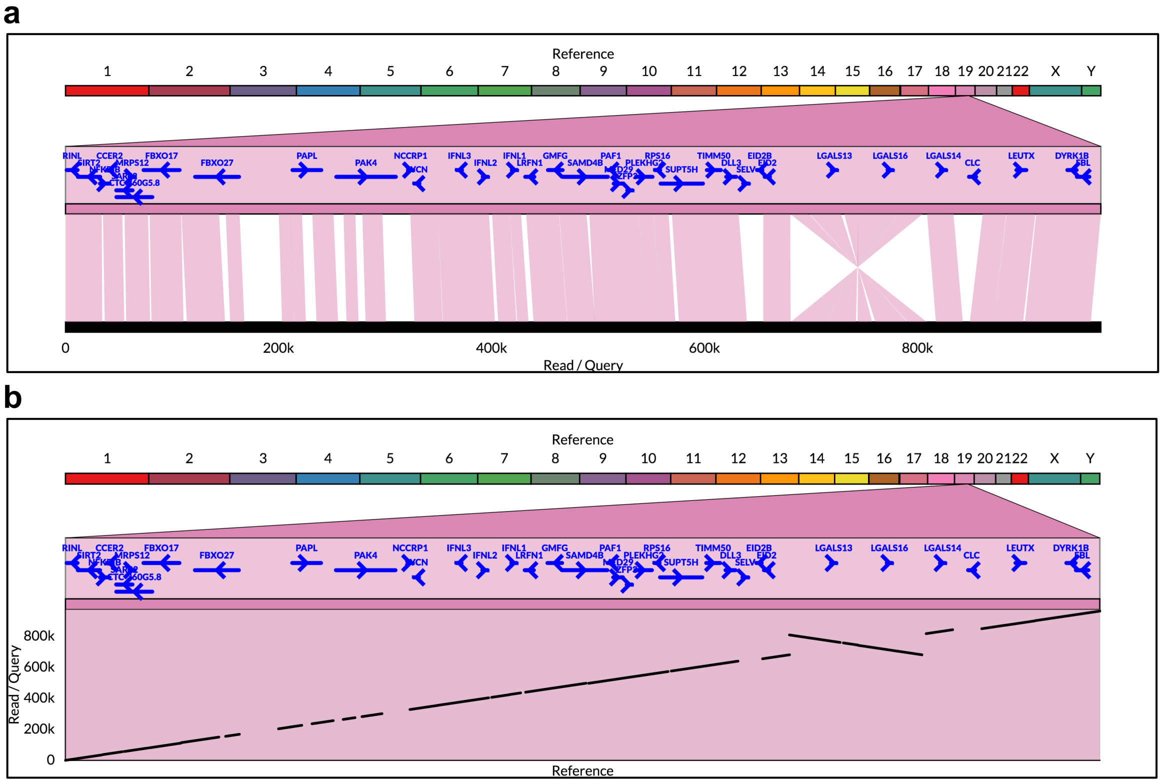
Supplementary Figure 10 | An overview of the alignments of the gorilla assembly against GRCh38. The colorful chromosomes in the background and across the top of the view represent the human chromosomes, and the alignments are contigs from the gorilla assembly shown in blue for forward direction and red for alignments in the reverse complement direction. The directions (colors) of the alignments are standardized relative to the longest alignment for each contig.





Supplementary Figure 11 | Large structural differences between gorilla (query) and human (reference). (a) Sequences on chr9 and chrX in human are fused in gorilla. (b) Sequences far apart on chr8 in human are fused in gorilla and one is inverted relative to the other. (c) Sequences on chr 1 and chr15 in human are fused in gorilla with some of the chr1 sequence inverted. (d) Sequences on chr6 and chr10 in human are fused in gorilla.

The original publication of this gorilla assembly[12] explored an inversion mapping to human chr19 near the 39.6 Mb position on GRCh38, which was flanked by nearby deletions on both sides. **Supplementary Figure 10** shows alignments of this same position in Ribbon as both a dot plot and a ribbon plot. The inversion is clearly visible, as are the two deletions on either side near the inversion. The first deletion to the left of the inversion in **Supplementary Figure 12** contains the gene SELV which is present in human but not in the gorilla genome. The second deletion is located to the right of the inversion and contains the gene CLC. These observations can be made almost instantaneously within Ribbon, and are consistent with what was reported in the original publication. It is also clear to see that the genes SELV and CLC are not present in the Susie 3 genome and that a large segment of that contig is inverted, and Ribbon produces attractive figures that can be used to explore and interpret patterns like this in alignments of any assembly or genome to another.



Supplementary Figure 12 | The gorilla genome has one region inverted compared to human flanked by nearby deletions on both sides. The inversion includes the genes LGAL513 and LGAGL516, and the deletions on either side include the genes SELV and CLC. Both the ribbon plot (a) and the dot plot (b) are generated by Ribbon.

Supplementary References

1. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5,** R12 (2004).
2. Miller, C. A., Qiao, Y., DiSera, T., D'Astous, B. & Marth, G. T. bam.iobio: a web-based, real-time, sequence alignment file inspector. *Nat Meth* **11,** 1189 (2014).
3. Down, T. A., Piipari, M. & Hubbard, T. J. P. Dalliance: interactive genome viewing on the web. *Bioinformatics* **27,** 889–890 (2011).
4. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15,** R84 (2014).
5. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single molecule sequencing. *bioRxiv* 169557 (2017). doi:10.1101/169557
6. Nattestad, M. *et al.* Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. (2017). doi:10.1101/174938
7. Nattestad, M., Alford, M. C., Sedlazeck, F. J. & Schatz, M. C. SplitThreader: Exploration and analysis of rearrangements in cancer genomes. *bioRxiv* 087981 (2016). doi:10.1101/087981
8. Edgren, H. *et al.* Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol.* **12,** R6 (2011).
9. Chen, K. *et al.* BreakTrans: uncovering the genomic architecture of gene fusions. *Genome Biol.* **14,** R87 (2013).
10. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* **12,** R72 (2011).
11. Suzuki, A. *et al.* Aberrant transcriptional regulations in cancers: genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. *Nucl Acids Res* **42,** 13557–13572 (2014).
12. Gordon, D. *et al.* Long-read sequence assembly of the gorilla genome. *Science* **352,** aae0344–aae0344 (2016).