

Dysgu variant call case examples

To better understand the different mechanisms that distinguish conventional linear reference alignment from graph-pangenome alignment methods, we investigated a few cases using Integrative Genomics Viewer, IGV¹. We selected different SV types from the truth set which were called by Dysgu exclusively in either linear or graph-based alignment. The BAM alignment files from both alignment approaches were loaded together with the PacBio Long-read alignment which was the main technology used for truth set generation. Variant call format files were also loaded including the truth set and the true positive VCF files produced by Truvari for each call set, which describes variant calls in the caller representation.

Example 1: Chr7 Insertion

In the linear alignment, displayed region has a mean coverage of 34X and a structural variant call was detected near the approximate breakpoint location. However, the variant was incorrectly classified as a translocation (TRA), no insertion length was determined, and the call was ultimately filtered due to low confidence. The alignment also showed many reads with mates mapped to different chromosomes, indicating substantial mapping ambiguity in the region.

Graph-based alignment using Giraffe produced a more accurate structural variant call. The mean coverage in the region was 37X, and the called breakpoint was located closer to the long-read truth position. Dysgu correctly identified the event as an insertion and estimated the insertion length to be 286 bp, compared with the true insertion length of 320 bp. Unlike the linear alignment result, the variant passed filtering criteria. Additionally, instead of extensive discordant interchromosomal mate mappings, many reads in the graph alignment had unmapped mates, suggesting reduced ambiguity in read placement relative to the linear alignment approach.

Further inspection of the breakpoint region revealed an 18 bp segment with elevated coverage at the insertion site: AGAAAGTGGGAGAGAAGG. The long-read alignment which is the basis for the truth set showed that this sequence forms part of a duplicated flanking region contained within the inserted allele, with the SV alt sequence beginning with: TAGAAAGTGGGAGAGAAGGCCAGGTG.... This flanking duplication creates ambiguity in breakpoint placement. Both the graph-based and linear short-read alignments failed to fully represent the duplicated 18 bp sequence as part of the insertion, causing the Dysgu graph-based call to report a breakpoint shifted by 18 bp relative to the long-read truth set. Despite this remaining ambiguity, the graph-based alignment substantially improves resolution and breakpoint localization in this case.

Example 2: Chr1 Deletion

A second example involved a 56 bp deletion at chr1:57,395,726–57,395,782. In the conventional linear alignment, the region had a mean coverage of 42X, but no structural variant call was produced. Inspection of the alignments revealed a high number of clipped reads at the breakpoint region, suggesting that the linear aligner was unable to perform gapped alignment at the deletion site.

In contrast, graph-based alignment using Giraffe enabled successful detection of the variant. The region showed a mean coverage of 47X, and Dysgu correctly identified both the structural variant type and the deletion length. Although the reported breakpoint position was shifted relative to the truth-set coordinates, the read alignments more cleanly spanned the deleted region rather than terminating as clipped alignments, indicating improved representation of deletions events shorter than read length.

Comparison with the long-read alignment and truth set showed that the deleted 56 bp sequence occurs twice in the reference genome. Both the graph-based Dysgu call and the long-read analysis indicated that one copy of the duplicated sequence was absent in the sample. Consequently, the breakpoint ambiguity likely reflects the repetitive nature of the locus, where both reported structural variant positions are biologically plausible representations of the same deletion event.

¹ James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. [Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 \(2011\)](#). A public access version is also available: [PMC3346182](#).

Example 3: Chr14 Insertion

A third example involved a 329 bp insertion at chr14:49,725,357 that was detected by Dysgu using conventional linear alignment but not identified in graph-based alignment. In the linear alignment, the region had a mean coverage of 39X, and Dysgu produced a structural variant call close to the expected breakpoint position and insertion length. However, the event was classified as a duplication rather than an insertion. The alignment coverage showed a clearly delineated and evenly increased depth across the affected region, supporting the interpretation that the event may represent a tandem duplication rather than a simple insertion. Numerous clipped reads, together with several supplementary alignments, provided informative evidence for the breakpoint and insertion boundaries. In addition, several mate pairs with unusually low insert sizes were observed, which may also be consistent with a duplication event.

In contrast, graph-based alignment failed to produce any structural variant call in the region. Mean coverage across the locus was 36X, although local peak coverage increased substantially when low mapping quality reads clustered within the putative insertion region were included. Compared with the linear alignment, relatively few reads help define the variant boundaries. Instead, the region was dominated by reads with low mapping quality, suggesting substantial mapping ambiguity, which may itself reflect the presence of a large duplication event, where repeated sequence content reduces the ability of graph-based alignment to uniquely position reads. Consequently, despite evidence of structural variation in the region, the graph-based approach did not provide sufficient confidently aligned breakpoint-supporting reads for Dysgu to generate a call.

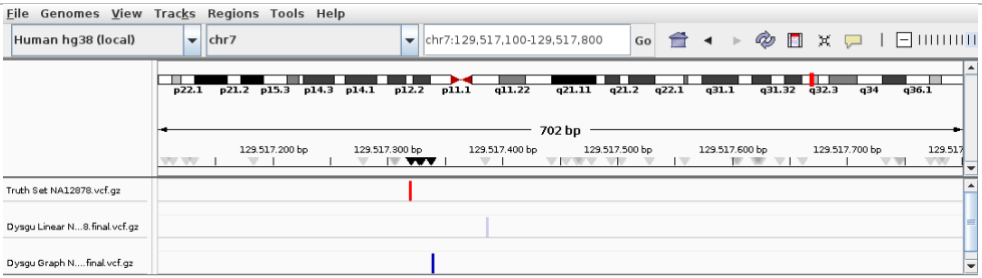
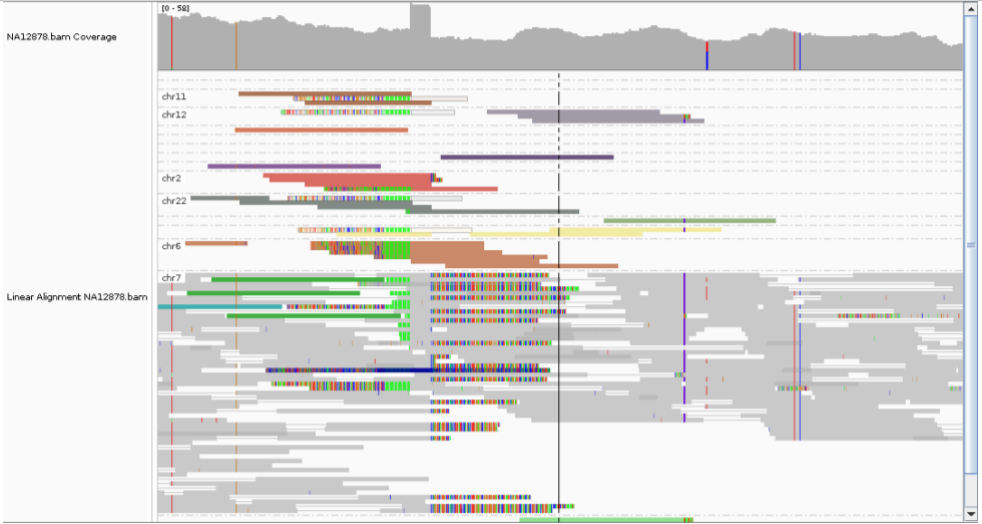
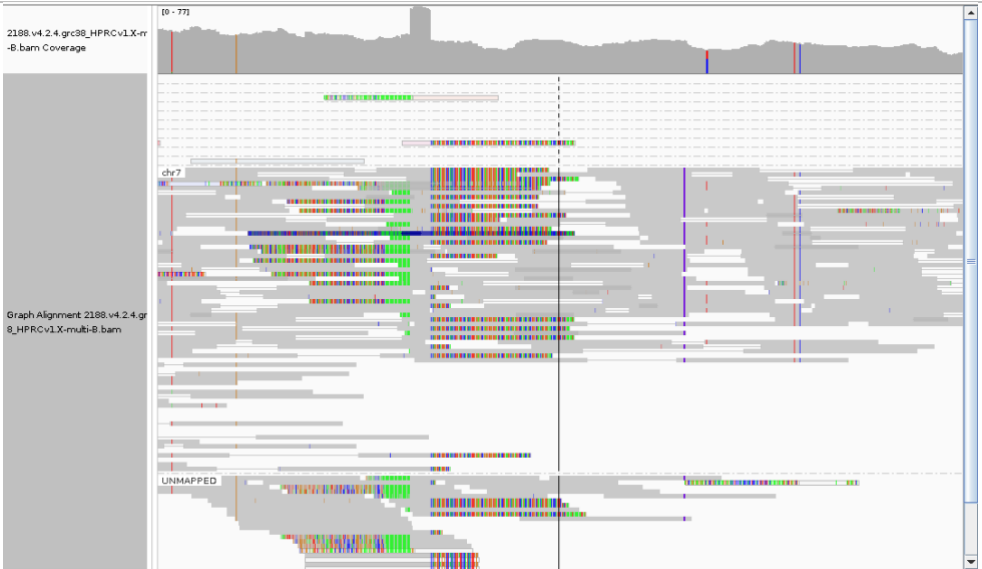
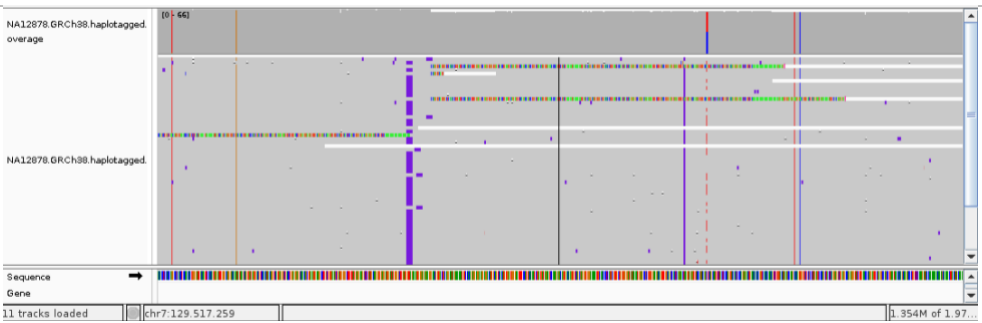
Interestingly in this case the Dragen SV-caller is similarly also able to call the variant using the Dragen alignment to a linear reference but not to the Dragen graph-augmented reference. This case also illustrates a limitation of strict structural variant type matching, which may cause otherwise informative calls to be missed.

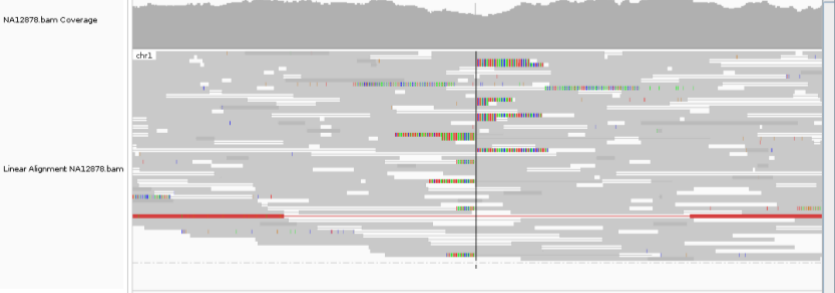
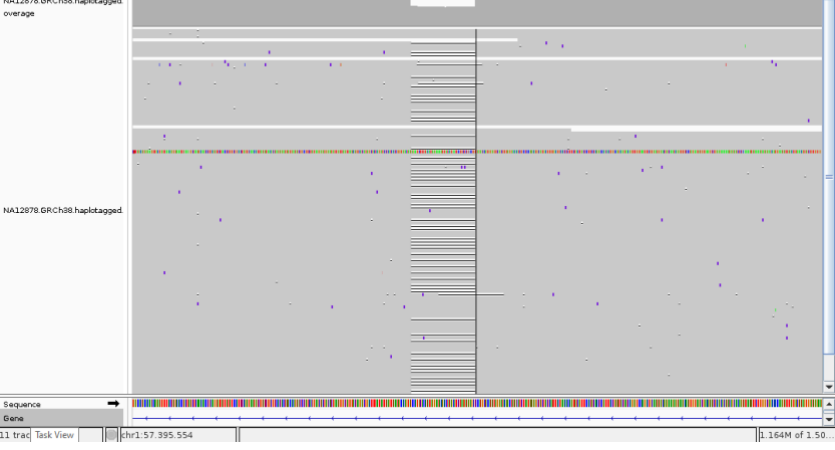
Example 4: Chr16 Complex Inversion Site

A complex structural variant region on chr16 containing multiple events in the truth set, which defines a 944 bp inversion at chr16:48,871,380, followed by a 164 bp deletion at chr16:48,871,896, and a 407 bp insertion at chr16:48,872,324. This region illustrates the difficulty of accurately representing clustered or overlapping structural variants using short-read data, where different callers and alignment strategies may capture only partial aspects of the underlying haplotype structure.

In the graph-based alignment, Dysgu correctly identified the position and approximate size of the first inversion event, although the variant was misclassified as a deletion. The insertion at the third truth-set variant is placed correctly, with a size slightly smaller than reported in the truth set. In contrast, the linear-alignment-based Dysgu call correctly localized the second variant site corresponding to the 164 bp deletion but classified the event as a 407 bp inversion instead. Thus, while neither approach fully reconstructed the true combination of structural events, both identified evidence of substantial structural.

In regions containing interfering events, different alignment frameworks may produce different but partially informative representations of the same underlying structural rearrangement. Consequently, manual inspection remains essential for interpretation of complex structural variation, particularly in clinical contexts. In such settings, the ability of a caller to reliably flag a region as structurally abnormal may be more valuable than perfectly matching what may be determined more precisely from long read data.

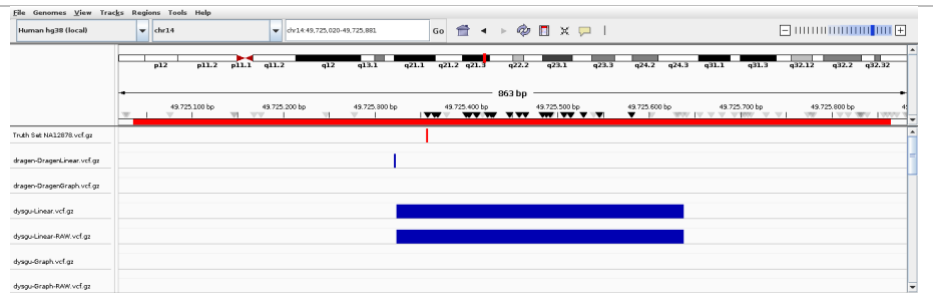
Example 1 Truth set defines a 320 bp insertion at position: chr7: 129,517,320	IGV overview 
Linear alignment with reads displayed as pairs and grouped to show interchromosomal mate mapping.	
Graph alignment representation has slightly increased read depth and less interchromosomal activity.	
Long read alignment with insertion site as purple band.	

Example 2 56 bp deletion at position: chr1:57,395,726–57,395,782	IGV overview 
Linear alignment with reads clipping at breakpoint.	
Graph alignment with gapped read alignment.	
Long read alignment visualizes similarly with gap shifted upstream by the length of deletion.	

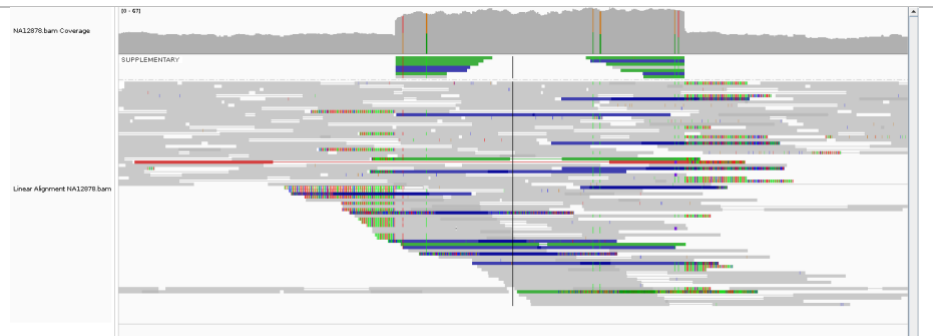
Example 3

329 bp insertion at position:
chr14:49,725,357

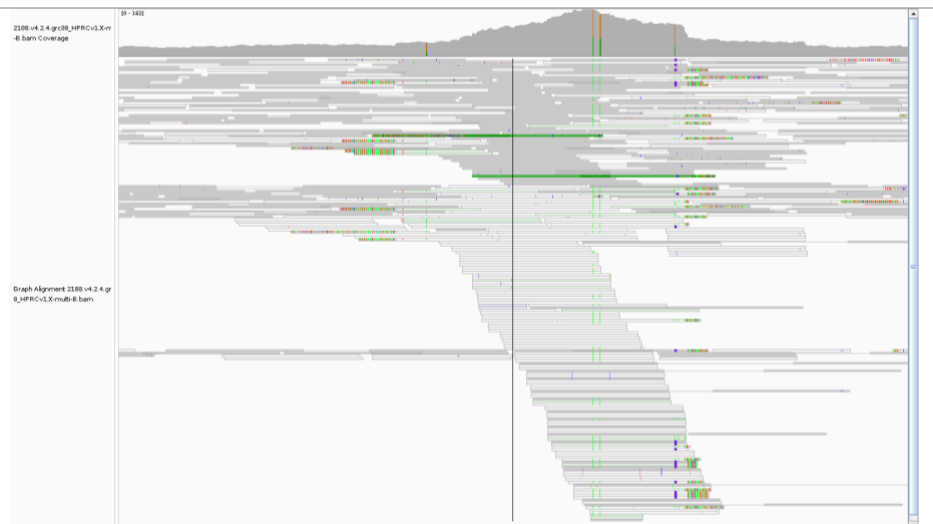
IGV overview



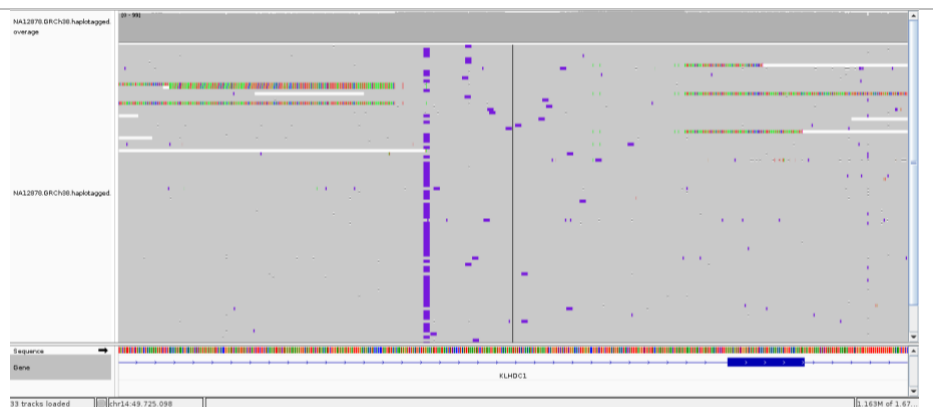
Linear alignment with
supplementary reads at the to. Low
insert size read pairs in dark blue.

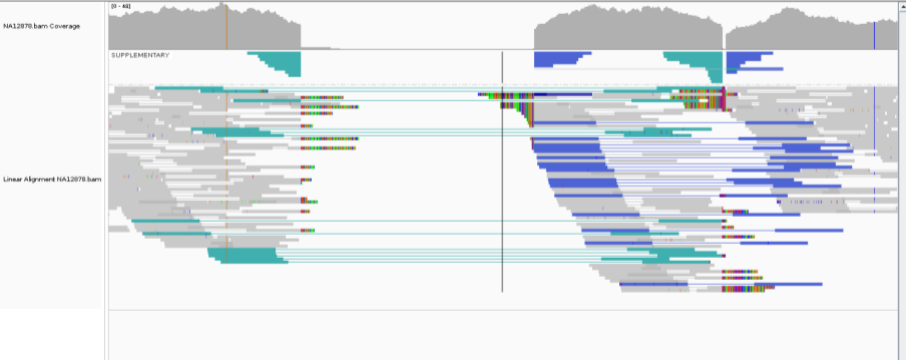


Graph alignment also show
increased coverage at site but with
poorly defined boundaries.



The insert sequence defined by the
long read alignment matches the
reference sequence at the insert
site which is indicative of a tandem
duplication.



Example 4 944 bp inversion at position: chr16:48.871.380, 164 bp DEL at position: 48.871.896, and a 407 bp INS at position: 48.872.324.	IGV overview 
Linear alignment With many read pairs in Forward-Forward orientation (turquoise) or Reverse-Reverse orientation (blue).	
Graph alignment loses inversion evidence due to low MAPQ of most reads with atypical read pair orientation.	
Long read alignments become difficult to interpret visually due to extensive soft clipping.	