Short and long-read genome sequencing methodologies for somatic variant detection; genomic analysis of a patient with diffuse large B-cell lymphoma

Supplementary figures and table captions

Hannah E Roberts[1]*, Maria Lopopolo[1]*, Alistair T Pagnamenta[1,2]*, Eshita Sharma[1], Duncan Parkes[1], Lorne Lonie[1], Colin Freeman[1], Samantha J L Knight[2], Gerton Lunter[3,4], Helene Dreau[5,6], Helen Lockstone[1], Jenny C Taylor[1,2¶], Anna Schuh[2,5,7¶], Rory Bowden[1], David Buck[1]

1. Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK

2. National Institute for Health Research Oxford Biomedical Research Centre, Oxford, UK

3. MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK

4. Department of Epidemiology, University Medical Centre Groningen, Groningen, NL

5. Oxford University Hospitals NHS Trust, Oxford, UK

6. Department of Haematology, University of Oxford, Oxford, UK

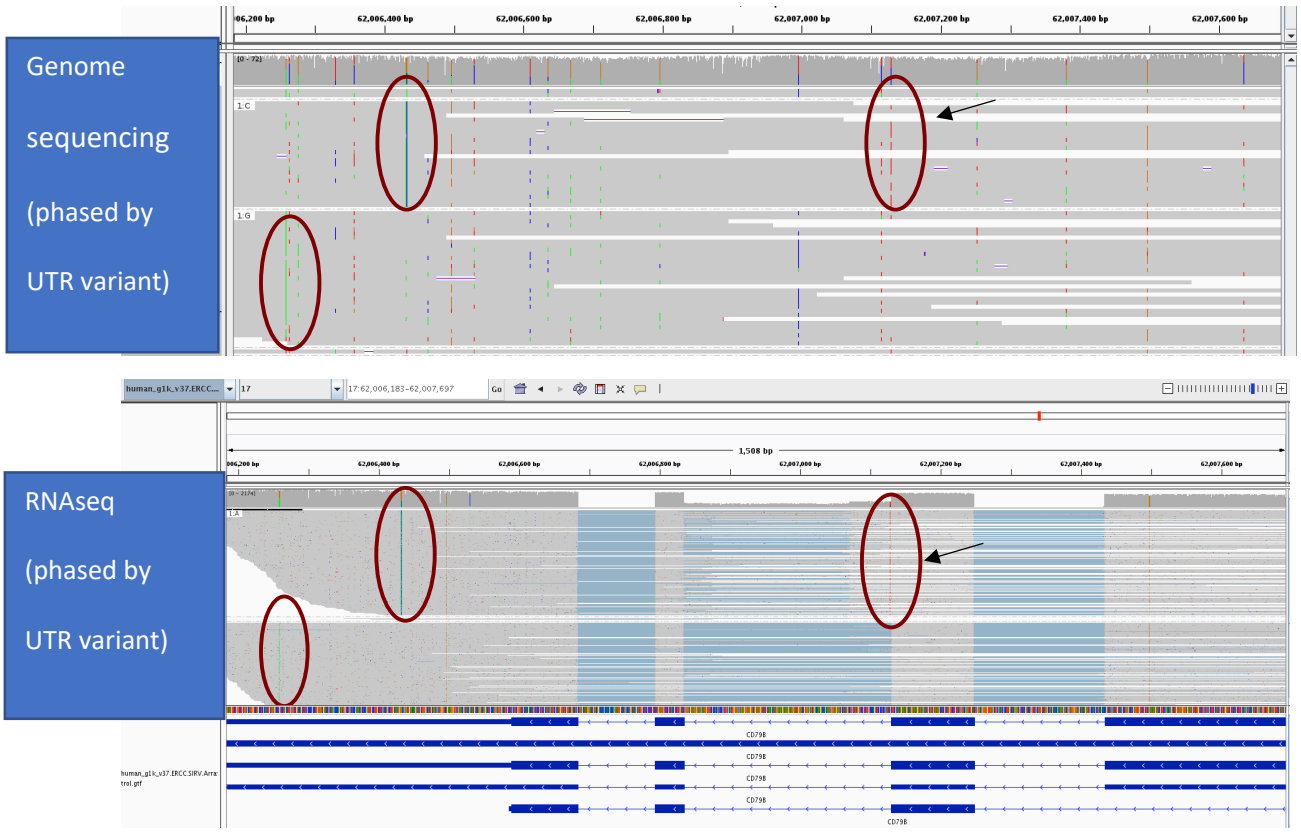
7. Department of Oncology, University of Oxford, Oxford, UK

* contributed equally

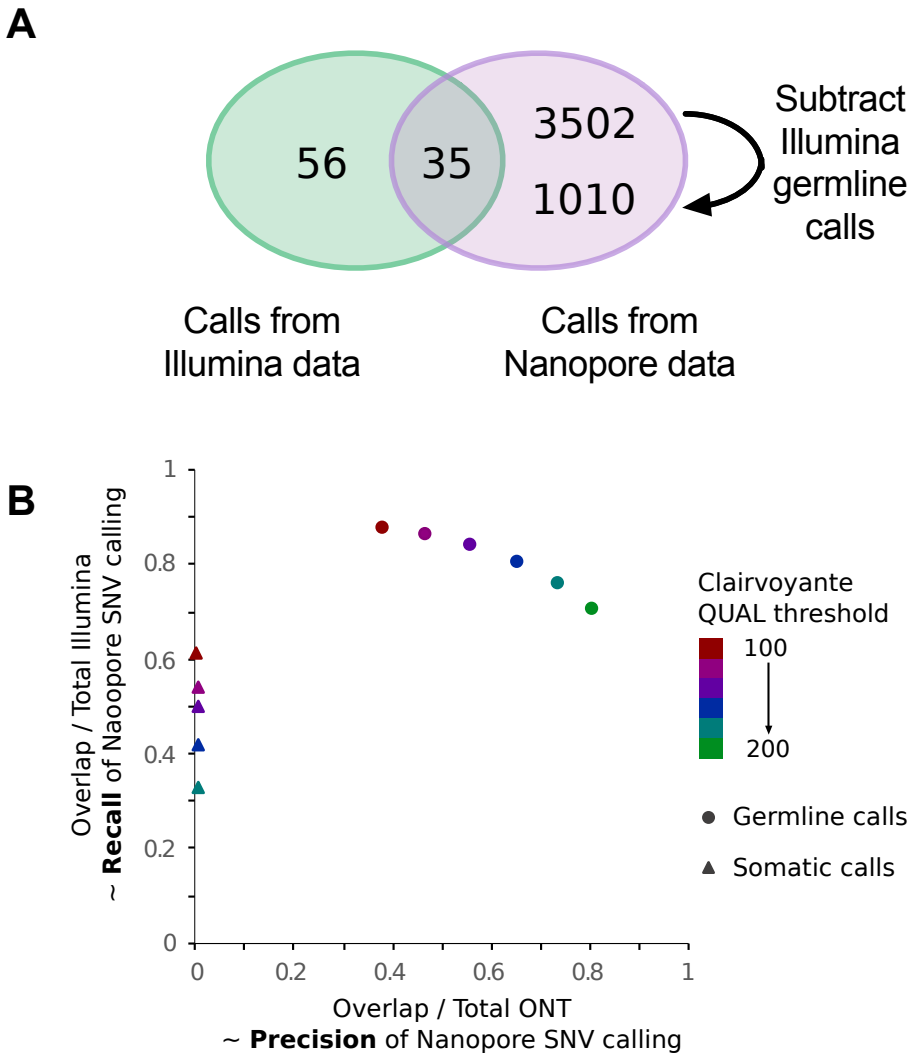
¶ correspondence to jenny.taylor@well.ox.ac.uk and anna.schuh@oncology.ox.ac.uk

Supplementary Information: Short vs long-read genome sequencing for somatic variant detection

Supplementary Figures



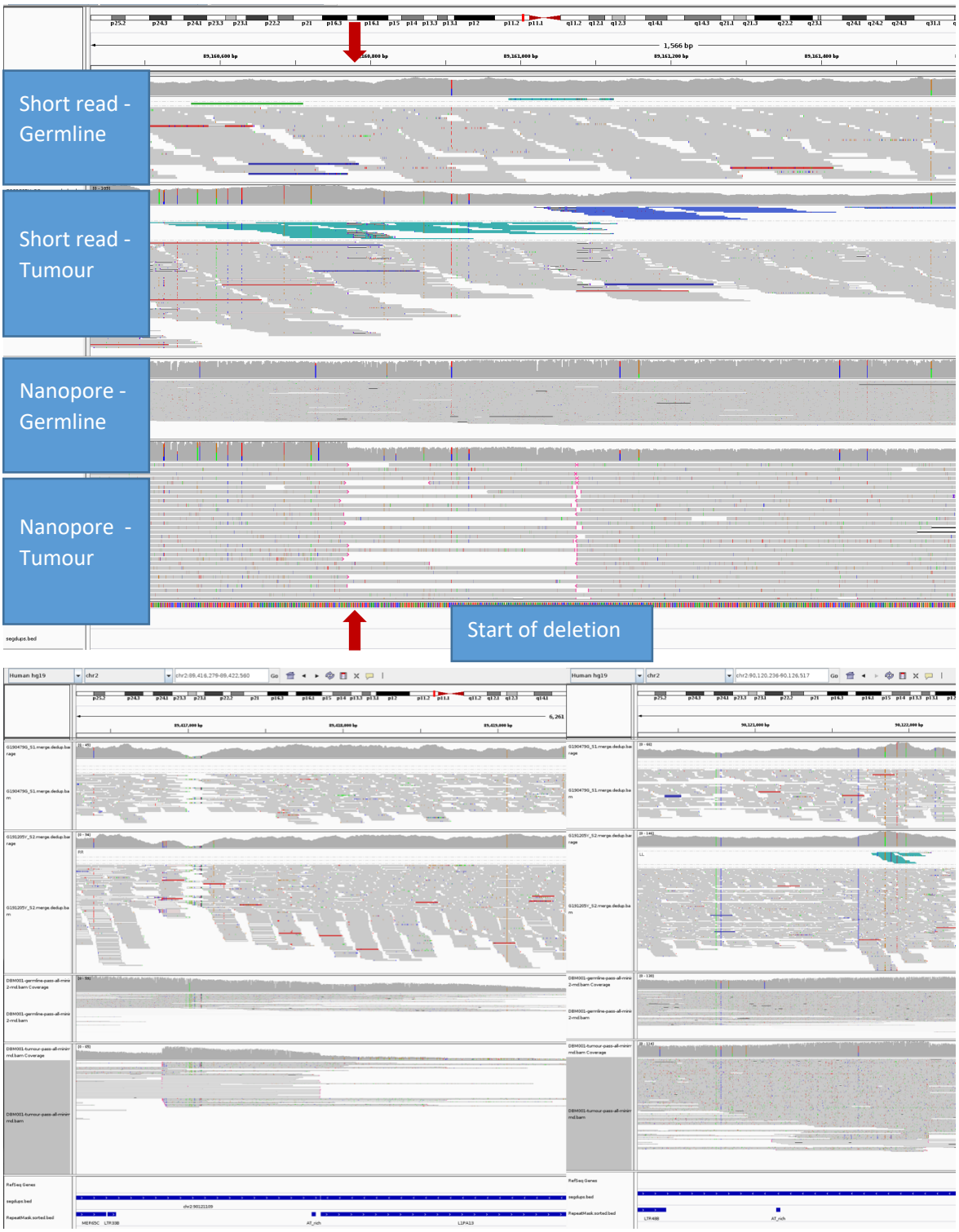
Supplementary Figure S1: The splice donor site mutation (c.552+1G>A, NM 001039933.1) can be accurately phased across multiple exons with the use of long Nanopore cDNA reads. Top: IGV screenshot showing reads from tumour DNA sample, which enable phasing of the somatic variant with one of the 3'-UTR variants. Bottom: Long Nanopore cDNA reads from the tumour RNA sample show that the splice donor variant results in intron retention and partial intron retention. In both screenshots, reads are grouped according to the base at one of the heterozygous SNPs in the 3'-UTR (circled in red). The somatic splice donor site mutation is circled in red and highlighted with an arrow.



Supplementary Figure S2: A) Overlap between somatic SNV calls generated from Nanopore vs Illumina reads covering chromosome 22. Freebayes was used to call SNVs in the Nanopore data. As with the results for chromosome 17 shown in Figure 1, there is more than an order of magnitude difference between the number of calls in the overlap between the two data sets and the number of calls from the Nanopore data, even after stringent filtering and subtracting the Illumina germline calls. B) The performance of Clairvoyante on the Nanopore data for chromosome 17 is shown. The correspondence between Clairvoyante and Illumina callsets is indicated for germline calls (circles) and somatic calls (triangles). The values displayed on the x and y axes approximate the precision and recall of Nanopore SNV calling

respectively, under the assumption that the Illumina calls approximate the truth set. For germline calls the colours represent different QUAL score thresholds used to define the 'PASS' filter (applied uniformly to homozygous and heterozygous calls), from red=100 to green=200. The same colouring is used for somatic calls, but here the variable QUAL score referred to is that used for filtering tumour calls prior to subtraction of Clairvoyante germline calls with minimal filtering applied (QUAL >0). Even following subtraction of this widest set of germline calls a large number of putative somatic calls remained, of which only 0.1-0.4% (QUAL 100 – 180) were found in the Illumina somatic call set.

Supplementary Information: Short vs long-read genome sequencing for somatic variant detection



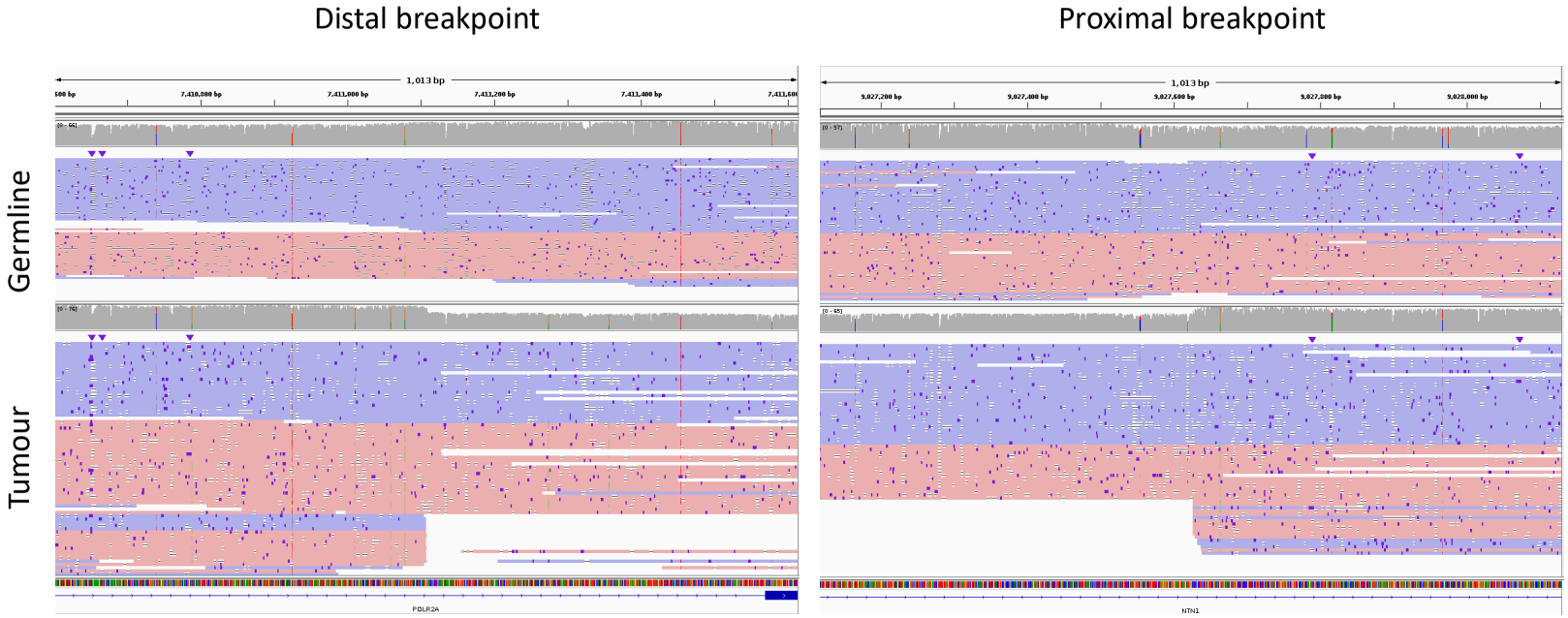
Supplementary Figure S3. Example of a somatic deletion near the centromere on the p arm of chromosome 2, which is incorrectly called as an inversion in the short-read data. Top:

Illumina and Nanopore read alignments at the distal breakpoint (breakpoint of interest indicated by red arrows – note that a second breakpoint corresponding to a smaller inversion can also be seen, and is called as an inversion in both data sets). The four loaded tracks in all screen shots correspond to the Illumina germline bam, Illumina tumour bam, Nanopore germline bam and Nanopore tumour bam, from top to bottom. The long-read alignments clearly show a drop in coverage associated with this breakpoint. Paired-end short reads are coloured according to pair orientation and insert size, hence the many turquoise reads in the Illumina tumour bam suggest an inverted breakpoint, although a few reads (red) support the deletion called in the Nanopore data. Bottom, LHS: IGV screen shot of proximal breakpoint suggested by long-read data; RHS: IGV screen shot of proximal breakpoint suggested by short-read data. The segmental duplications track second from bottom in these screen shots shows that there is a segdup linking these two suggested proximal breakpoints. Several of the long Nanopore reads mapping to the left-hand breakpoint extend beyond the end of the segdup (whereas this is not the case with any of the long reads supporting the right-hand breakpoint), hence suggesting the deletion as the true underlying SV.

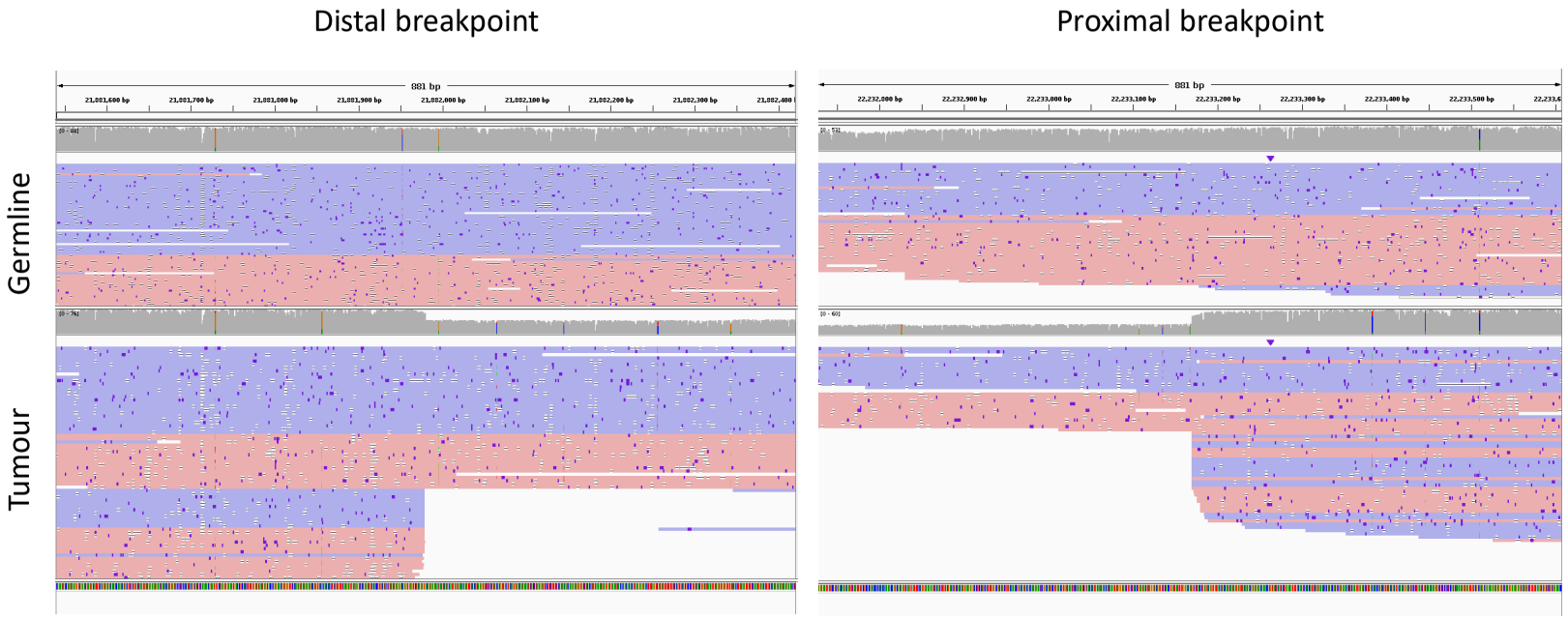
Supplementary Information: Short vs long-read genome sequencing for somatic variant detection

Supplementary Figure S4 (A – G): Screenshots of read alignments in the Integrative Genomics Viewer (IGV) supporting the large variants of clinical interest described in the main text. Reads correspond to the Nanopore data unless labelled otherwise. Nanopore reads are coloured red/blue according to plus/minus strand

Supplementary Figure 4A. Deletion of *TP53*

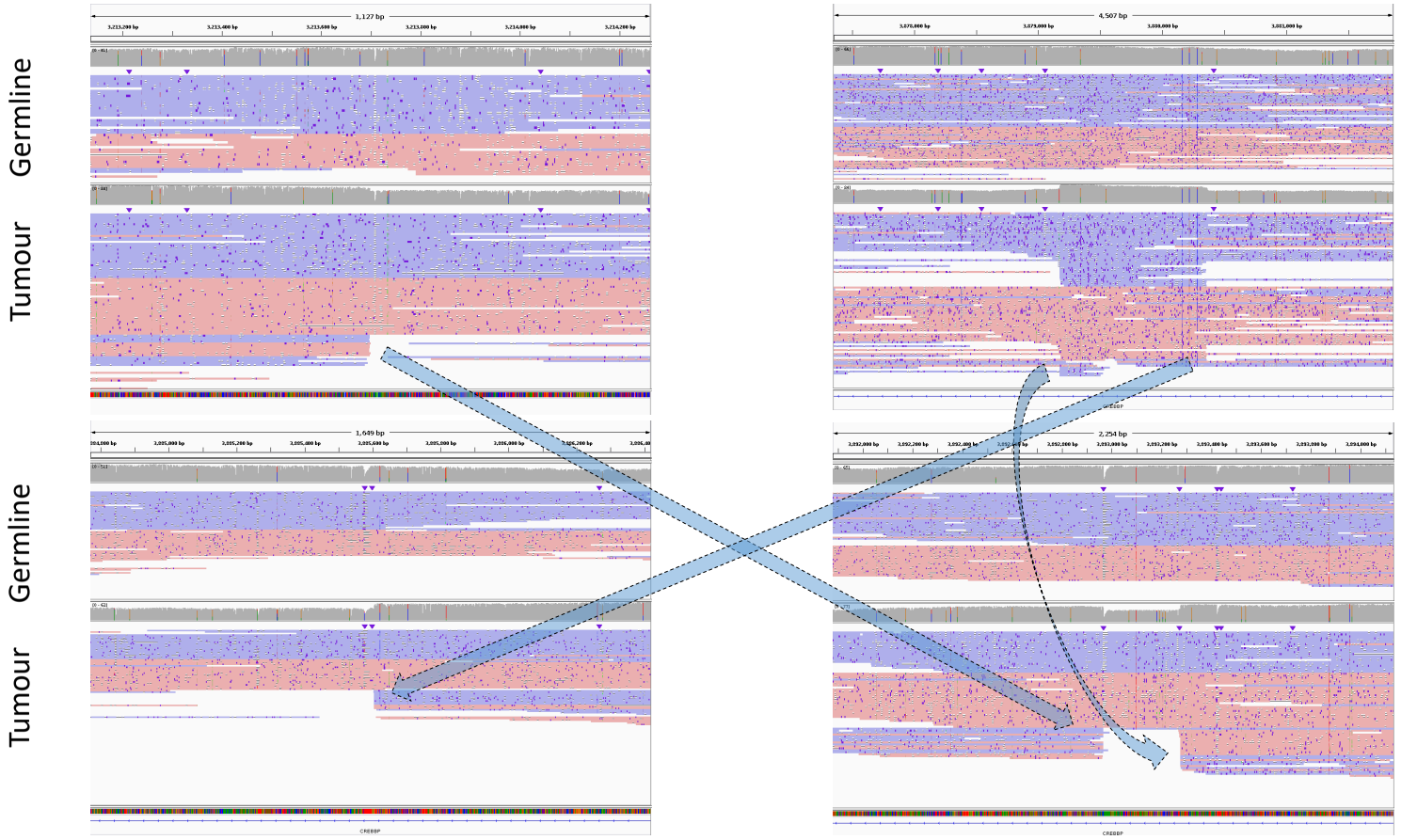


Supplementary Figure 4B. Deletion of *CDKN2A/2B*



Supplementary Figure 4C. Complex deletion/inversion involving *CREBBP*

Same complex SV as shown in Figure 4.

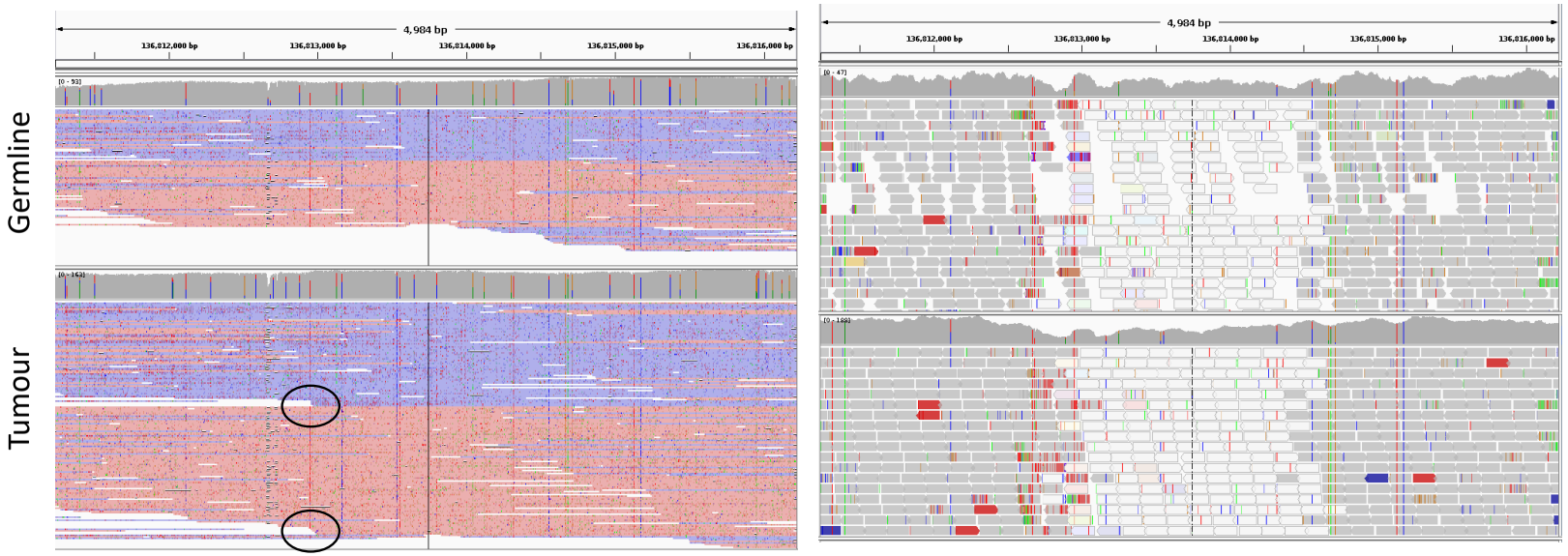


Supplementary Figure 4D. High copy gain involving *BCL6* and *PIK3CA*

High copy gain of whole of chromosome 3 (see circos plot shown in Figure 3), with the CNV annotation segmented by a ~4Mb deletion. Deletion means high copy gain -> copy gain

Nanopore data

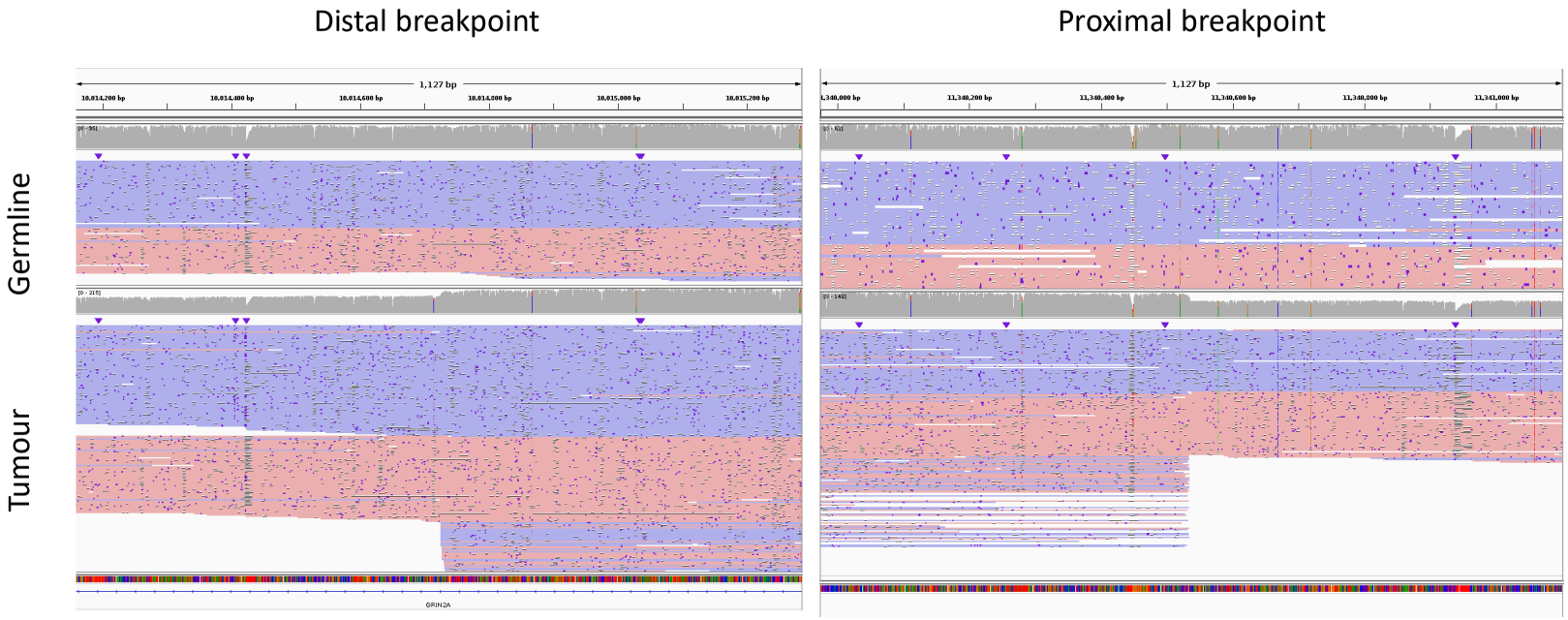
Illumina data



Circles highlight clipped reads

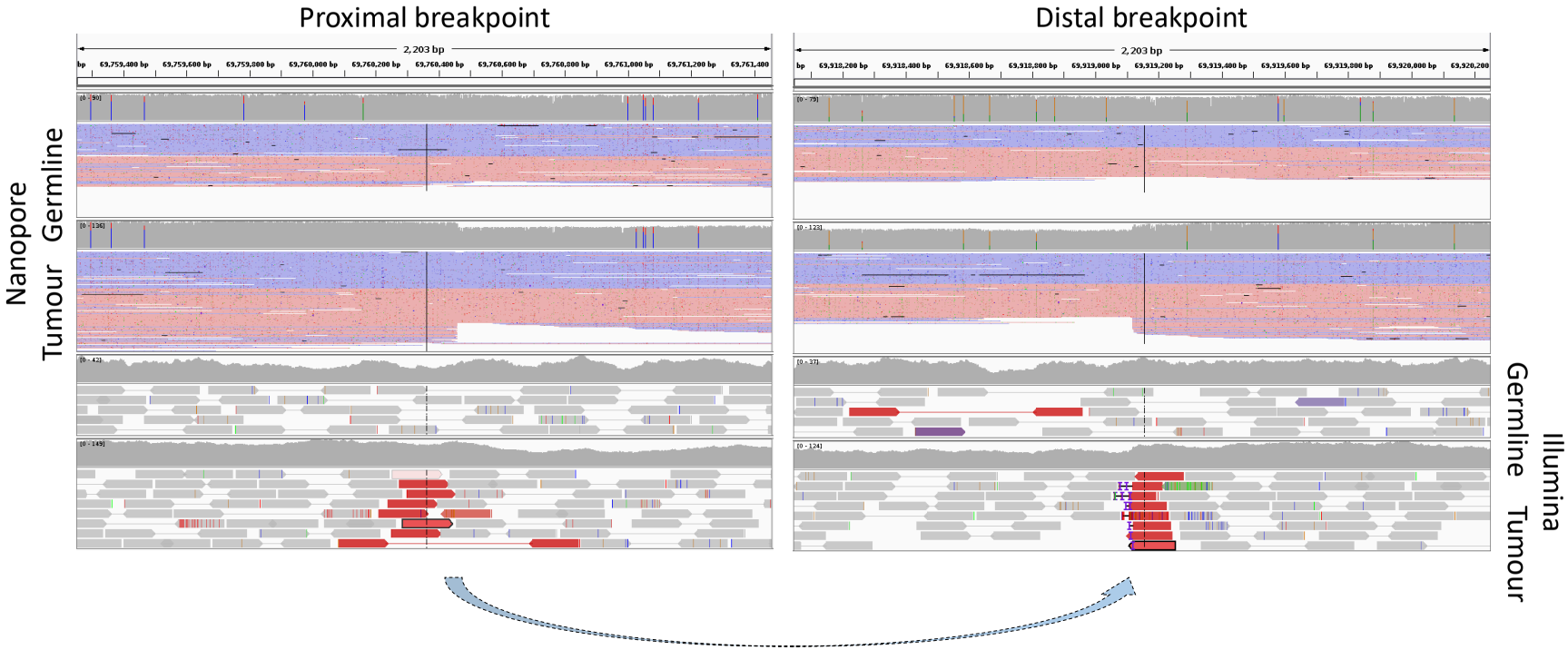
Suspected BP in 2kb LINE element where Illumina reads suffer from low mapping quality

Supplementary Figure 4E. Duplication of *CIITA*



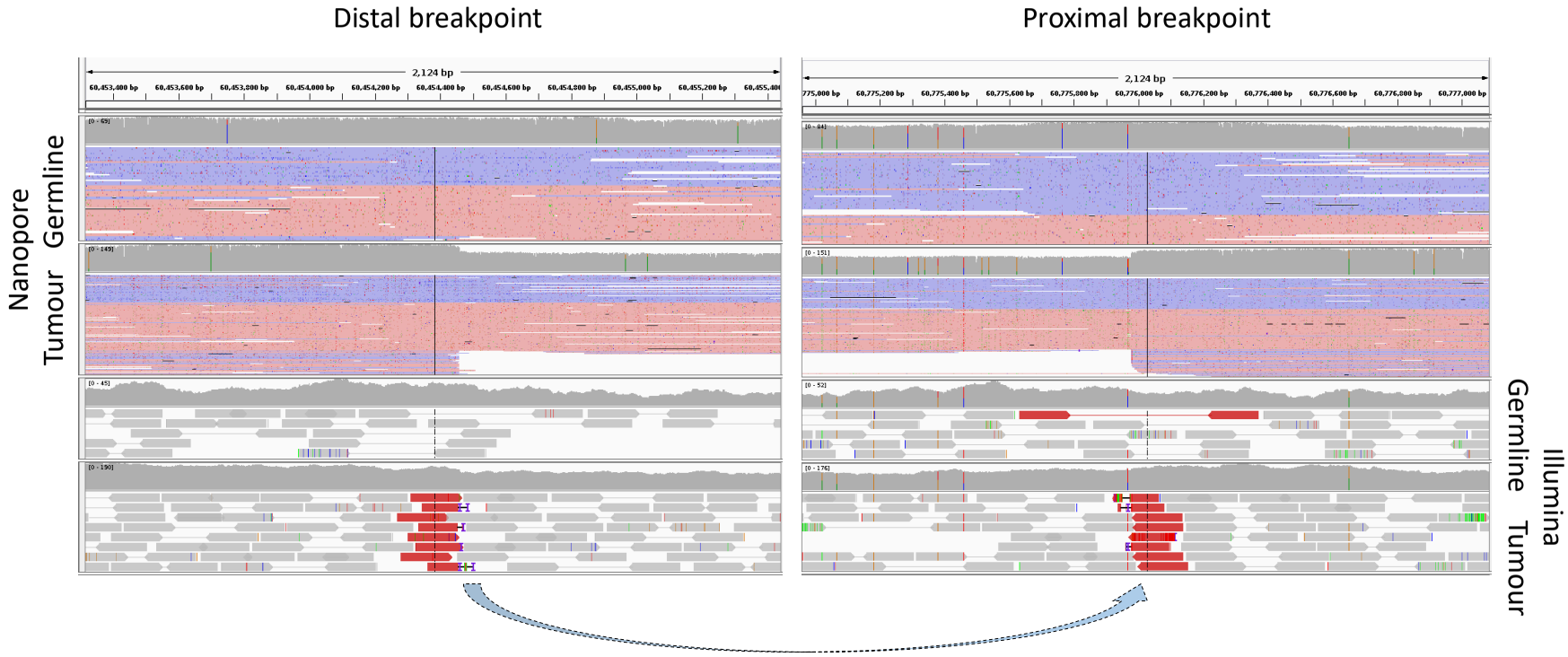
Supplementary Figure 4F. Duplication of *CARD11*

Copy gain of whole chromosome 7 (see circos plot shown in Figure 3) with CNV call segmented by a 159kb deletion. Deletion means copy gain -> diploid

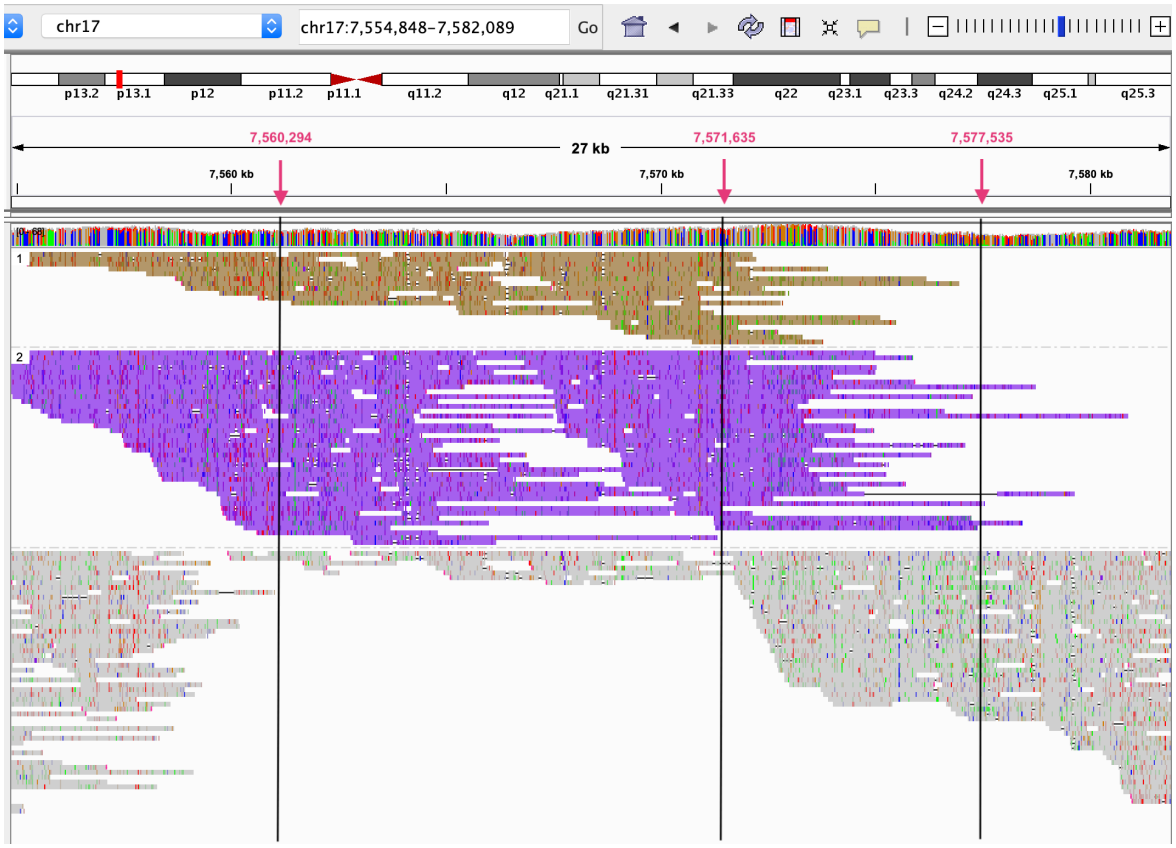


Supplementary Figure 4G. High copy gain involving *MYD88*

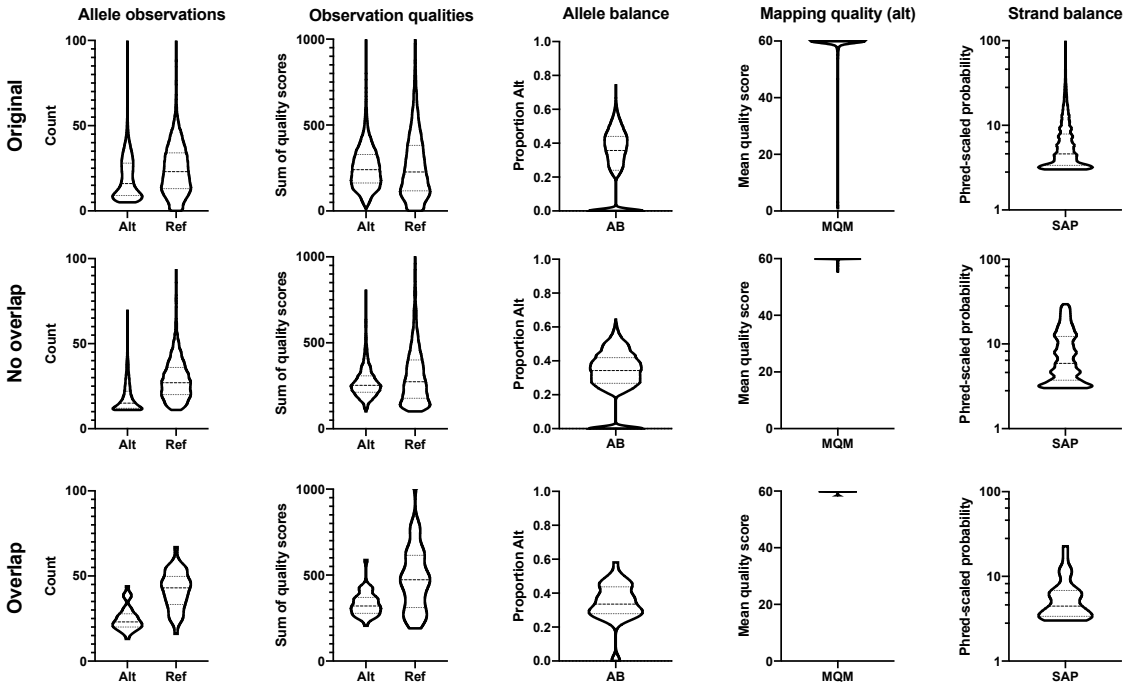
High copy gain of whole of chromosome 3 (see circos plot shown in Figure 3), with the CNV annotation segmented by a 322kb deletion. Deletion means high copy gain -> copy gain



Supplementary Information: Short vs long-read genome sequencing for somatic variant detection



Supplementary Figure S5: Long reads from the tumour sample mapped with minimap2 are shown in IGV, coloured by phase set and grouped according to haplotype. Phase sets and haplotypes were determined by WhatsHap. The region shown includes the p.Arg249Met mutation (located at 17:7,577,535 and indicated by the rightmost pink arrow). The germline heterozygous SNPs at the start and end of the phase set detailed in Figure 5B are also indicated by the left and centre pink arrows.



Supplementary Figure S6: Properties of Nanopore SNV calls. Top – all calls. Middle – filtered calls that don't overlap with short-read calls. Bottom – filtered calls that do overlap with short-read calls

Supplementary Information: Short vs long-read genome sequencing for somatic variant detection

Supplementary Table captions

Supplementary Table S1: Somatic SNVs of potential clinical relevance detected by short-read genome sequencing. Three of the 13 SNVs were located on chr17. Variant annotation was using the Ensembl Variant Effect Predictor web interface (http://grch37.ensembl.org/Homo_sapiens/Tools/VEP), performed on 10th December 2019. Although listed here separately, the three variants in IGLL5 and two variants in IGKV1D-17 lie in close proximity and may represent single mutational events

Supplementary Table S2: Coverage, mapping quality and base quality in the long read data at breakpoints of the 3 deletions, 1 inversion and 1 translocation that were called only in the short-read data but assessed to be true somatic variants

Supplementary Table S3: Somatic SVs detected by one or more pipeline. Breakpoint (BP) positions are given with reference to the GRCh37 genome. SV types include TRA (translocations), DUP (duplications), DEL (deletions) and INV (inversions). Method(s) used to detect SV are indicated by I (Illumina), M (nanopore pipeline with minimap2 mapping) and N (nanopore pipeline with ngmlr read mapping). N/A, SV length not available for translocation events. Comments reflect either the reason a call was missed by some methods (for TRUE calls), or the reason a call could have been falsely made (for FALSE calls), where such reasons were easily classifiable. Background ploidy of the chromosome (of BP1) is included for true calls to aid assessment of whether an SV results in a CN gain or loss (e.g. the deletions in chr 3 are against a background ploidy of 4 and hence result in a CN gain while regions outside the deletions have a high CN gain)