Supplementary Information for "*Improvement of structural variants detection using whole genome nanopore sequencing and comparative genome hybridization*"

Javier Cuenca-Guardiola, Belén de la Morena Barrio, Juan Luis García, Alba Sanchis-Juan, Javier Corral de la Calle and Jesualdo Tomás Fernández-Breis

Supplementary material

Supplementary Figures

Supplementary Figure 1

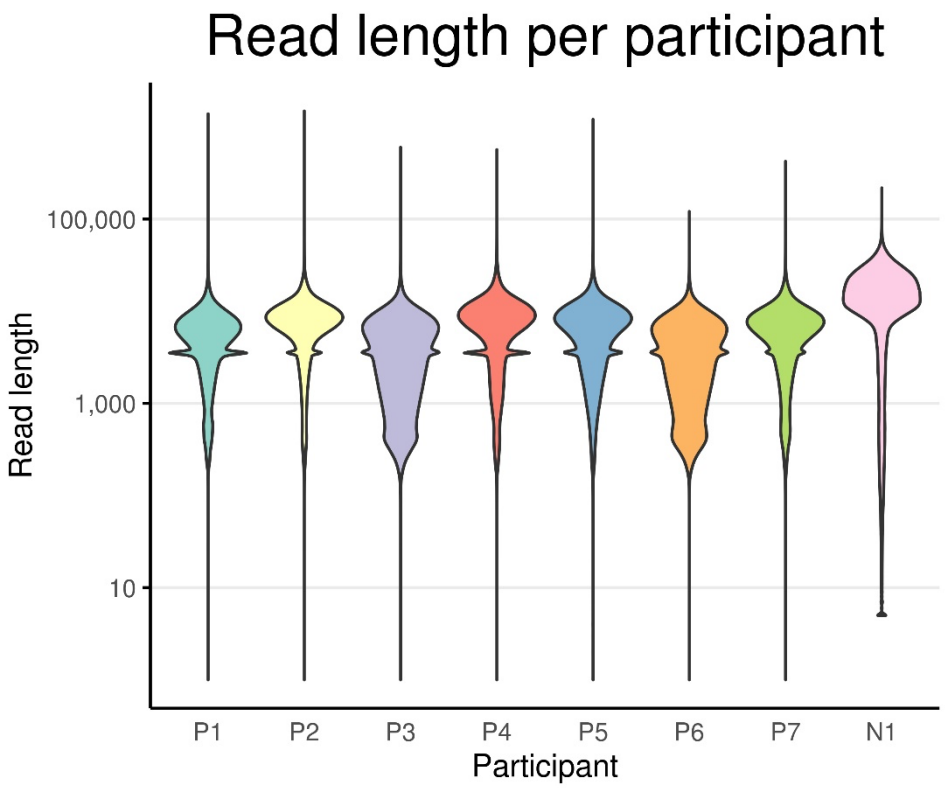

Fig. S1. Violin plots of read length distribution per patient. Each patient's shape is a different color. N1 refers to NA19240. The mean length is similar for all cases except for NA19240.

Supplementary Figure 2

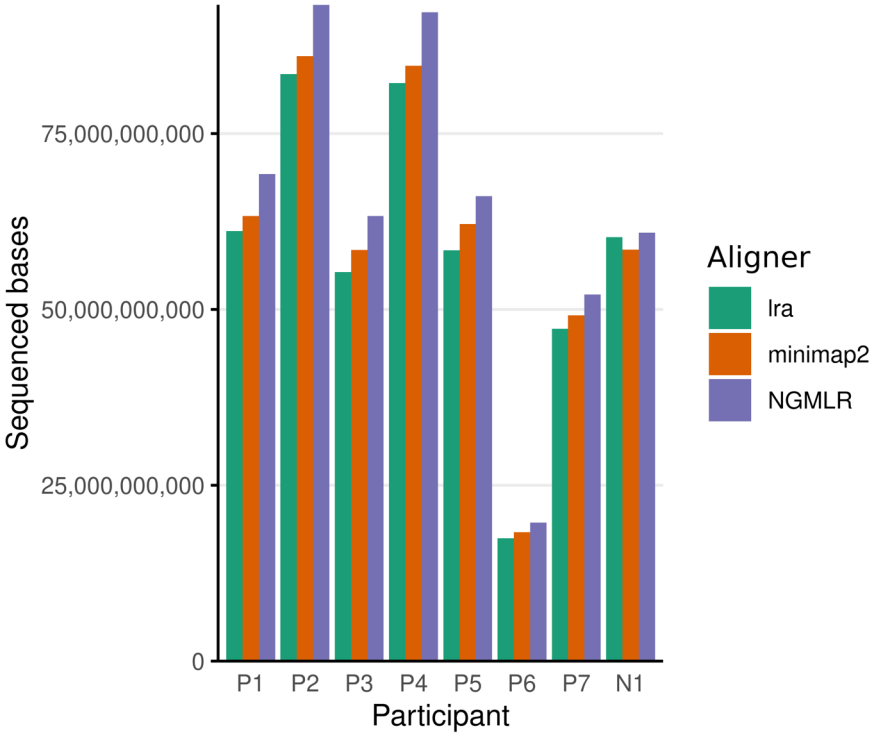

Fig. S2. Metrics of nanopore sequencing for each patient included in this study and NA19240 (N1): number of bases sequenced per aligner: dark green for minimap2, orange for NGMLR, purple for lra. P6's sequencing yielded fewer bases, while the data used from NA19240's whole genome sequencing was similar to the study cases.

Supplementary Figure 3

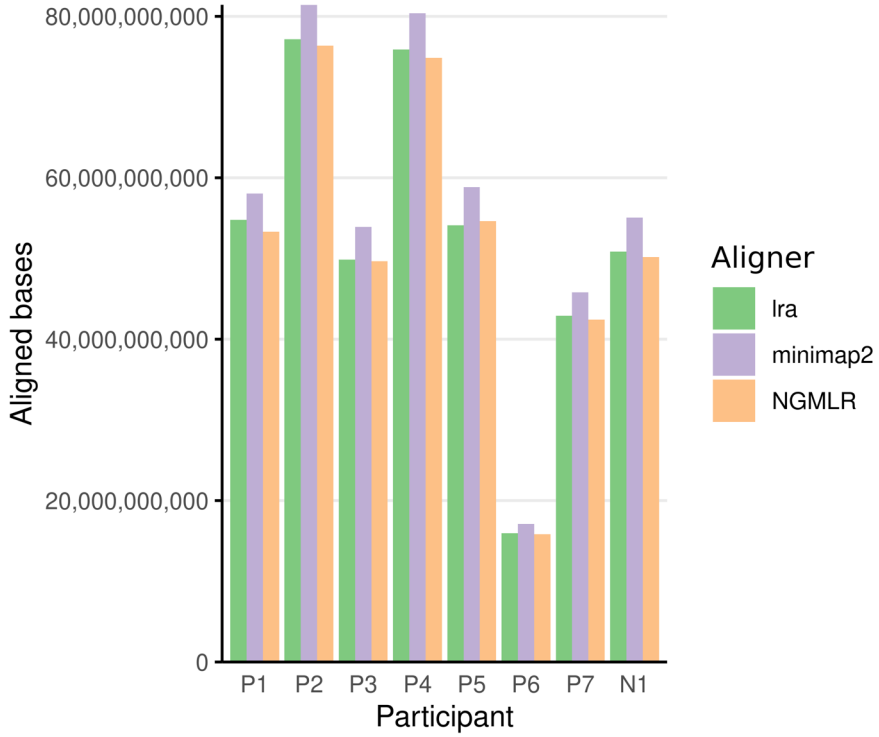

Fig. S3. Metrics of nanopore sequencing for each patient included in this study and NA19240 (N1): number of bases aligned: green for minimap2, violet for NGMLR, orange for lra. As for the number of bases sequenced, P6's sequencing produced fewer, while NA19240 data was similar to our study cases.

Supplementary Figure 4

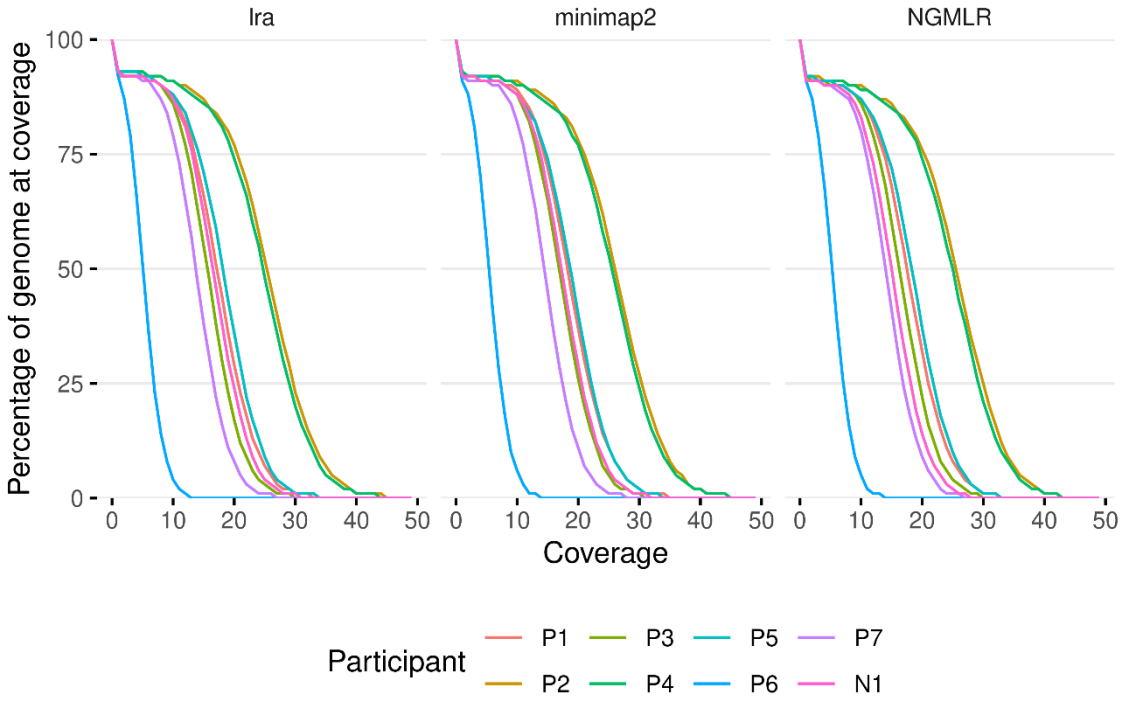

Fig. S4. Percentage of the genome covered by nanopore sequencing for each patient according to lra, minimap2 or NGMLR analysis. Participants are color-coded. N1 refers to NA19240. As expected by the low base count, P6's coverage was lower, not reaching 20x depth in any position. Again, the data from NA19240 was similar to the study cases.

Supplementary Tables

Supplementary Table 1

This table is provided as a standalone file. The first row contains the description of the column, for example, “Sniffles RE (minimap2)” contains the RE value (read support) for Sniffles after having used minimap2.

The table contains aCGH coordinates and CNV type for the variants in the study, and for each caller, coordinates, SV type, read support (RE), and for SVIM, the QUAL value that SVIM computes. Additionally, it includes the disCoverage results for these SVs. The p-value thresholds are the same than in the main text: “***”: p-value <= 1e-13482, “**”: 1e-3117. For this table, an arbitrary value was chosen for “*”: p-value <= 1e-1000, it is worth mentioning that variants have only been considered supported when below the first threshold. After this, information is presented for the hg19 pipeline. Coordinates prefixed by an exclamation mark (!) indicate that the SV was not considered found, either for a difference in size larger than 30%, or a read support below the threshold, but are included for completion’s sake.

Supplementary Table 2

Tool	Type of software	Input	Output	Type of SV it can process	Genotypes SVs?
lra	Aligner	Reads	Alignment	-	-
minimap2	Aligner	Reads	Alignment	-	-
NGMLR	Aligner	Reads	Alignment	-	-
cuteSV	Variant caller	Alignment	Variant calls	DEL, DUP, INV, INS, TRA	Yes
NanoVar	Variant caller	Alignment	Variant calls	DEL, DUP, INV, INS, TRA	Yes
Sniffles	Variant caller	Alignment	Variant calls	DEL, DUP, INV, INS, TRA	Yes
SVIM	Variant caller	Alignment	Variant calls	DEL, DUP, INV, INS, TRA	Yes*
SURVIVOR	Variant combiner	Variant calls	Variant calls	DEL, DUP, INV, INS, TRA	No (accepts genotyped input)
mosdepth	Coverage analysis tool	Alignments	Coverage	-	-
disCoverage	Other	Variant calls, coordinates	Coverage support for variants	DEL, DUP	No

Table S2 This table summarizes the aligners and variant callers employed in this work, alongside other well-established tools. The software disCoverage has been added to clarify its function. Aligners generate mappings of reads against a reference genome. Variant callers take these mappings, detect patterns of variation, and generate calls for SVs. Additionally, callers may genotype variants. Some of them, such as SVIM, can run minimap2 or other aligners, but we have not considered them aligners. This table is a summary, and these tools have many uses and applications, only those relevant for this work are listed. *: Genotyping is described as stable.

Supplementary Table 3

Aligner	SV type	Total	<1 kb	1-10 kb	10-50 kb	>50 kb
lra	INS	6 763 (36.2%)	6 187 (40.6%)	575 (18.2%)	1 (0.59%)	0 (0.0%)
	DEL	6 942 (37.9%)	6 177 (41.2%)	738 (25.0%)	27 (9.78%)	0 (0.0%)
	DUP	41 (3.7%)	41 (6.8%)	0 (0.0%)	0 (0.00%)	0 (0.0%)
minimap2	INS	7 009 (37.6%)	6 355 (41.7%)	652 (20.6%)	2 (1.18%)	0 (0.0%)
	DEL	7 817 (42.7%)	6 989 (46.7%)	795 (26.9%)	32 (11.59%)	1 (1.1%)
	DUP	153 (13.8%)	130 (21.5%)	19 (4.9%)	4 (4.49%)	0 (0.0%)
NGMLR	INS	6 503 (34.8%)	5 902 (38.7%)	600 (19.0%)	1 (0.59%)	0 (0.0%)
	DEL	7 331 (40.1%)	6 532 (43.6%)	770 (26.1%)	29 (10.51%)	0 (0.0%)
	DUP	182 (16.4%)	163 (26.9%)	15 (3.9%)	4 (4.49%)	0 (0.0%)

Table S3. Performance of variant calling after each aligner, split by SV size, for NA19240 data. As SV size increases, recall diminishes, although this is particularly steep over 50 kb. The truth set included 231 inversions with length lower than 1000 bp, none of them were found, and 75 SVs labeled as CNV, which did not allow SURVIVOR to match them against the calls generated by our pipeline. When comparing sets ignoring SV types, they were not found, either. INS: insertions, DUP: duplications, DEL: deletions.