Supplementary material

Jain et al. *Long read mapping to repetitive reference sequences using Winnowmap2*

Content summary

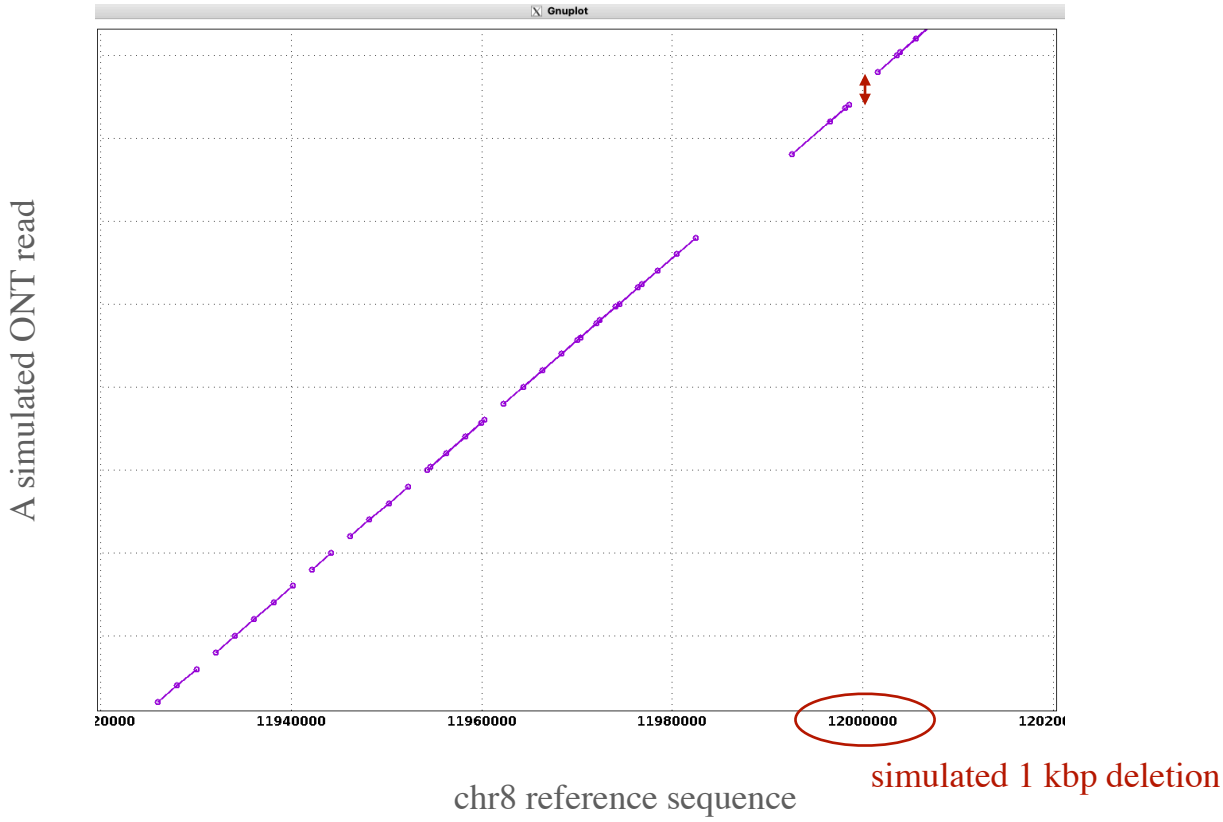

Fig. S1. Several MCAS alignments computed by Winnowmap2 are shown for a simulated ONT read using a dot-plot. MCASs surrounding the non-reference SV allele are correctly aligned to the mutated chr8 reference sequence. A purple dot indicates either the start or the end of an MCAS alignment. MCASs can have variable length and can also overlap with each other. These MCAS alignments are joined together in a final step by Winnowmap2.

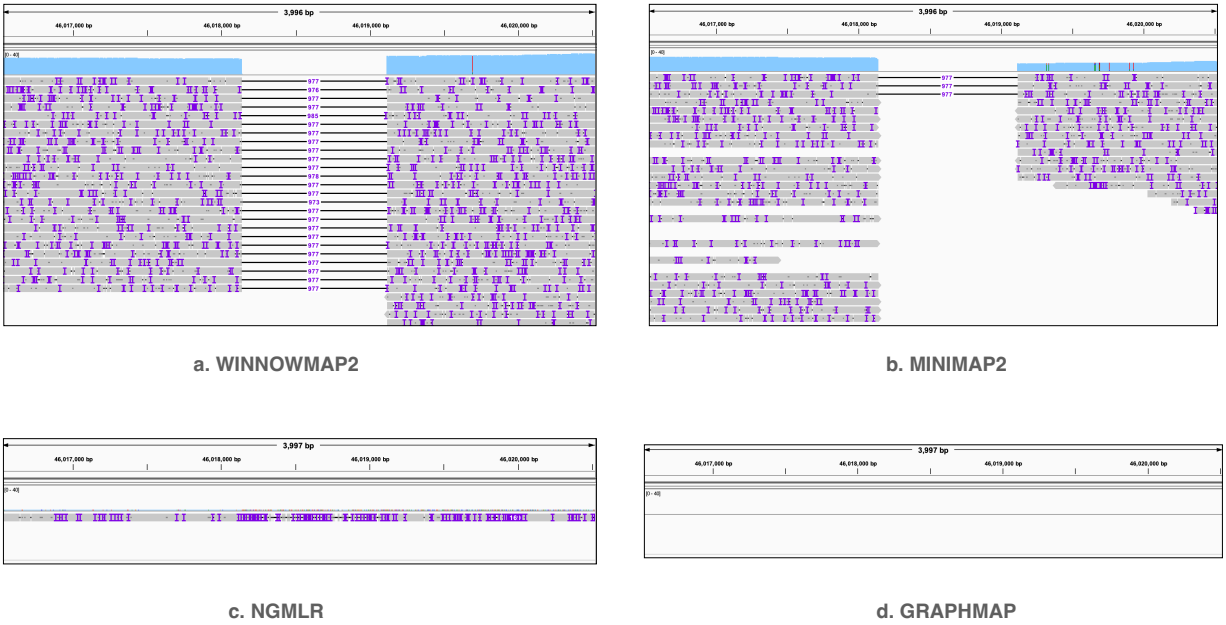

Fig. S2. IGV visualization of alignment pileup near a simulated deletion SV at locus 46,018,137 bp of chromosome 8. The locus is within the centromeric satellite DNA array of chromosome 8. The sky-blue-colored track on top of each plot shows mapping-coverage using a uniform y-axis scale (0-40). The grey-colored line segments show individual primary read alignments. IGV uses purple markers to indicate presence of indels within read alignments. NGMLR, minimap2, graphmap show reduced coverage due to allelic bias whereas Winnowmap2 shows good coverage in this region. Consistent large deletions in the middle of winnowmap2 read alignments are distinctly visible due to simulated SV of size 977 bp.

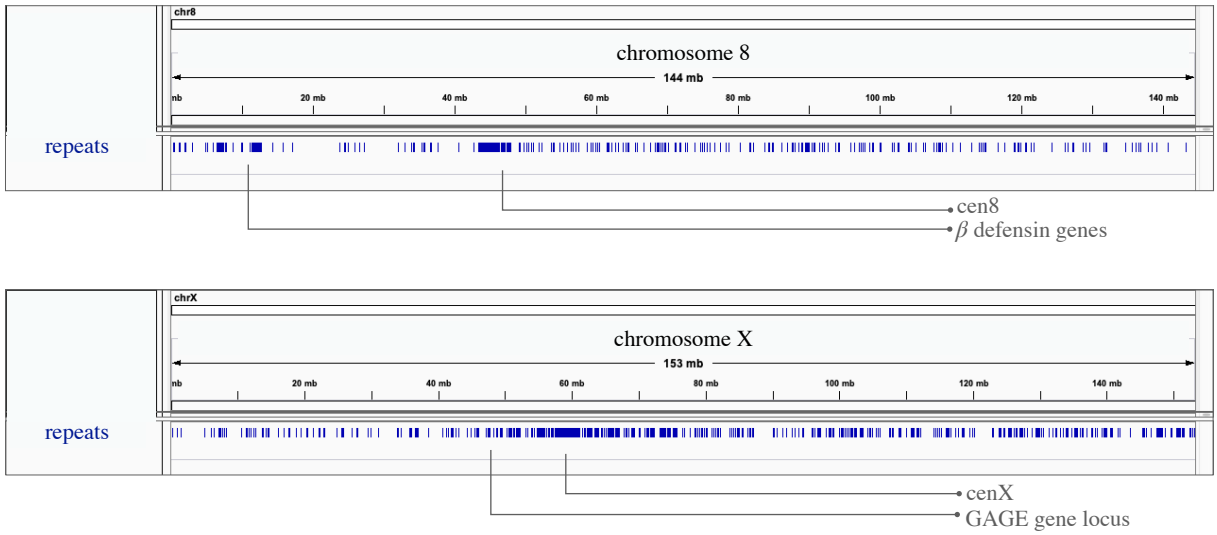

Fig. S3. De novo annotation of long near-identical duplications in CHM13 chromosome 8 and X respectively. In each case, these duplications (length≥ 10 kbp, identity≥ 95%) were identified by executing self-alignment of these chromosomes using Mashmap. Mashmap (https://github.com/marbl/MashMap) includes a script that generates these intervals in bed format. A few repeat units are also labeled in the above figure.

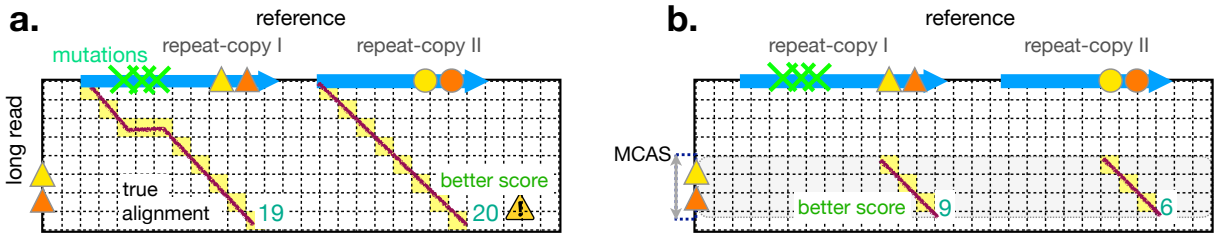

Fig. S4. Illustration of MCASs using a DP alignment scoring matrix. Similar to Figure 1 in the main text, PSVs are shown using colored triangle and dot markers. Alignments are annotated with their hypothetical scores. The left figure illustrates the effect of allelic-bias. The score of a true alignment spanning non-reference alleles can be lower compared to the score of an incorrect alignment. On the right side, an MCAS of a read which does not span non-reference alleles can achieve correct and unique placement to a reference. MCASs are subsequently processed during a final consolidation step to obtain correct alignments.

chromosome 8, total calls simulated 510/100/490

Dataset		**Winnowmap2**	**Winnowmap**	**minimap2**	**NGMLR**
hifi 20x	Total calls	500/98/479	494/85/474	492/85/469	493/95/454
	FN calls	10/2/11	16/15/16	18/15/21	17/5/36
	FP calls	0/0/0	1/0/1	3/0/3	6/2/0
hifi 40x	Total calls	510/100/490	503/87/480	501/87/475	497/97/466
	FN calls	0/0/0	7/13/10	9/13/15	13/3/24
	FP calls	2/0/1	5/0/2	6/0/4	12/4/0
ONT 20x	Total calls	502/97/484	494/86/478	493/85/476	479/93/452
	FN calls	8/3/6	16/14/12	17/15/14	31/7/38
	FP calls	0/0/0	3/0/2	4/0/4	2/0/0
ONT 40x	Total calls	510/100/489	505/88/484	502/88/479	497/97/468
	FN calls	0/0/1	5/12/6	8/12/11	13/3/22
	FP calls	0/0/1	7/1/6	8/2/8	6/3/0

chromosome X, total calls simulated 488/100/512

Dataset		**Winnowmap2**	**Winnowmap**	**minimap2**	**NGMLR**
hifi 20x	Total calls	476/98/496	470/81/481	469/79/482	468/96/470
	FN calls	12/2/16	18/19/31	19/21/30	20/4/42
	FP calls	1/1/0	4/1/0	4/1/0	6/6/0
hifi 40x	Total calls	485/99/509	473/81/493	472/80/496	471/97/486
	FN calls	3/1/3	15/19/19	16/20/16	17/3/26
	FP calls	3/0/0	8/2/0	9/3/3	14//12/1
ONT 20x	Total calls	476/99/507	467/83/496	467/84/500	457/90/475
	FN calls	12/1/5	21/17/16	21/16/12	31/10/37
	FP calls	2/0/1	6/1/1	7/1/1	5/3/1
ONT 40x	Total calls	485/99/510	480/84/503	481/84/507	472/97/494
	FN calls	3/1/2	8/16/9	7/16/5	16/3/18
	FP calls	2/0/1	8/2/2	12/3/2	8/9/1

Table S1. Structural variant accuracy evaluation using chromosome 8 and chromosome X as reference sequences respectively. In this experiment, the following three types of SVs were simulated using SURVIVOR: deletions, inversions and insertions. Accordingly all figures of the form $x/y/z$ indicate x deletions, y inversions and z insertions respectively. This table provides a detailed breakdown of the plot shown in main text (Figure 3).

Data set	Reference	Method	Unmapped reads	Incorrectly mapped reads
PacBio-CLR (5× coverage)	Human chrX (CHM13)	Winnowmap2	0%	0.03%
PacBio-CLR (5× coverage)	Human chrX (CHM13)	minimap2	0%	0.15%
PacBio-CLR (5× coverage)	Human WG (GRCh38)	Winnowmap2	0%	2.0%
PacBio-CLR (5× coverage)	Human WG (GRCh38)	minimap2	0%	1.9%

Table S2. Evaluation of read mapping accuracy by checking the fraction of reads that were incorrectly placed compared to their simulated origin. This experiment repeats the simulated benchmarking method adopted in [11] to test Winnowmap2 and minimap2. Here we simulated PacBio CLR reads with mean error rate of 10% and mean read length of 15 kbp using PBSIM from T2T chromosome assemblies of chromosome X (v0.7) as well as GRCh38 human reference. We used Minimap2's `paftools` utility to evaluate mapping accuracy. Unlike the experiment conducted using SURVIVOR method, here reads were directly simulated from the sequence on which they are mapped.

Tool	Purpose	Command line parameters
Winnowmap2 (v2.03)	HiFi read mapping	`-W repetitive_k15.txt -ax map-pb --MD ref.fasta hifi.fq.gz`
	ONT read mapping	`-W repetitive_k15.txt -ax map-ont --MD ref.fasta ont.fq.gz`
minimap2 (v2.18)	HiFi read mapping	`-t 24 -ax asm20 --MD ref.fasta hifi.fq.gz`
	ONT read mapping	`-t 24 -ax map-ont --MD ref.fasta ont.fq.gz`
NGMLR (v0.2.7)	HiFi read mapping	`-t 24 -r ref.fasta -q hifi.fq.gz -o output.sam`
	ONT read mapping	`-t 24 -x ont -r ref.fasta -q ont.fq.gz -o output.sam`
Winnowmap (v1.01)	HiFi read mapping	`-W repetitive_k19.txt -t 24 -ax asm20 --MD ref.fasta hifi.fq.gz`
	ONT read mapping	`-W repetitive_k15.txt -t 24 -ax map-ont --MD ref.fasta ont.fq.gz`
graphmap (v0.5.2)	ONT read mapping	`align -r ref.fasta -d ont.fq.gz -o output.sam -t 24`
Sniffles (v1.0.11)	SV calling	`-n -1 -t 8 -m alignments.bam -v output.vcf`
SURVIVOR (v1.0.6)	SV simulation	`simSV ref.fasta parameter_file 0 1 alternate`
	SV evaluation	`eval SV.vcf truth.bed 50 results`
SVanalyzer(v0.36)	SV evaluation	`SVbenchmark --ref ref.fasta --test test.vcf --truth truth.vcf --includebed giab.bed --testfilter PASS --truthfilter PASS --normdist 1.00 --normsizediff 1.00 --normshift 1.00`
PBSIM (commit:e014b1)	HiFi read simulation	`--depth 20 --data-type CLR --accuracy-mean 0.999 --accuracy-min 0.99 --length-min 18000 --length-mean 20000 --length-max 22000 --model_qc model_qc_clr ref.fasta`
NanoSim(v2.6.0)	ONT read simulation	`read_analysis.py genome -i ont.fq.gz -rg ref1.fasta -ga output.sam -t 28 -o train, simulator.py genome -rg ref2.fasta -c train -med 50000 -sd 0.5 -t 28 -n $NUM`
Mashmap (commit:ffeef4)	Repeat annotation	`mashmap -r ref.fasta -q ref.fasta -f none -s 10000 --pi 95, python denovo_repeat_annotation.py mashmap.out 10000 95 > tmp.bed, bedtools merge -i tmp.bed`
bedtools(v2.29.2)	Filter SVs within repeats	`bedtools intersect -a output.vcf -b repeats.bed -u -wa > repeats.vcf`

Table S3. Command line parameters that were used to execute various tools for this study.

Simulated long reads from T2T chromosomes 8 and X that were used in this study

Dataset	**Read count**	**N50**	**Min. length**	**Max. Length**	**Notes**
HiFi (chr8, 20x)	146,118	19,979	18,000	22,000	simulated by PBSIM
HiFi (chr8, 40x)	292,286	19,972	18,000	22,000	simulated by PBSIM
HiFi (chrX, 20x)	154,869	19,969	18,000	22,000	simulated by PBSIM
HiFi (chrX, 40x)	309,745	19,971	18,000	22,000	simulated by PBSIM
ONT (chr8, 20x)	51,993	63,283	8,915	369,183	simulated by NanoSim
ONT (chr8, 40x)	103,723	63,503	4,498	497,739	simulated by NanoSim
ONT (chrX, 20x)	54,073	64,260	11,076	331,049	simulated by NanoSim
ONT (chrX, 40x)	110,227	63,763	1,272	342,759	simulated by NanoSim

Real human HG002 sequencing data used in this study

Dataset	**Read count**	**N50**	**Min. length**	**Max. Length**	**Notes**
HiFi (HG002, 35x)	8,449,287	12,885	47	30,581	15 kbp library
ONT (HG002, 35x)	12,563,983	50,819	1	543,308	Guppy 3.6.0 PromethION
ONT (HG002, 50x)	19,328,993	50,380	1	543,308	Guppy 3.6.0 PromethION
ONT (HG004, 90x)	29,319,334	48,060	1	1,197,859	Guppy 3.6.0 PromethION
ONT (HG007, 45x)	4,986,802	50,117	1	647,447	Guppy 3.5.1 PromethION

Table S4. Length statistics of simulated and real long read sequencing data sets used in this study.

Supplementary Note S1

Lemma 1. *Computing MCAS(i)* $\forall\ 0 \leq i < |Q|$ *requires* $O(|Q||R|)$ *time and* $O(|R|)$ *space.*

Proof. Assume any appropriate linear or affine gap scoring function is being used. Denote p^{th} character in read Q as $Q[p]$ and a substring ranging from positions p to q as $Q[p, q]$ with both ends inclusive. Consider the following algorithm to compute MCAS(i). Compute semi-global DP alignment of $Q[i, i + j]$ to reference R while iterating the variable j from 0 to c. Here c is the maximum length allowed for a valid MCAS. Computing a row of the alignment score table requires $O(|R|)$ time. As a new row is computed in an iteration, we need to check whether the best-scoring alignment satisfies the confidence criteria, i.e., whether its score compared to the second-best non-overlapping alignment exceeds by a user-specified threshold. Using a method from Waterman and Eggert (PubMed: 2448477), the second-best non-overlapping alignment can be computed in $O(j^2)$ additional time. Therefore, asymptotic time spent per row is $O(|R| + j^2)$. As j is bounded by the constant c, asymptotic time spent to compute MCAS(i) remains $O(|R|)$. Therefore, computing all MCASs requires $O(|Q||R|)$ time. Asymptotic space complexity of the above algorithm is $O(|R|)$. □