bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# HQAlign: Aligning nanopore reads for SV detection using current-level modeling

Dhaivat Joshi, Suhas Diggavi*, Mark J.P. Chaisson, and Sreeram Kannan †

**Abstract**

**Motivation:** Detection of structural variants (SV) from the alignment of sample DNA reads to the reference genome is an important problem in understanding human diseases. Long reads that can span repeat regions, along with an accurate alignment of these long reads play an important role in identifying novel SVs. Long read sequencers such as nanopore sequencing can address this problem by providing very long reads but with high error rates, making accurate alignment challenging. Many errors induced by nanopore sequencing have a bias because of the physics of the sequencing process and proper utilization of these error characteristics can play an important role in designing a robust aligner for SV detection problems. In this paper, we design and evaluate HQAlign, an aligner for SV detection using nanopore sequenced reads. The key ideas of HQAlign include (i) using basecalled nanopore reads along with the nanopore physics to improve alignments for SVs (ii) incorporating SV specific changes to the alignment pipeline (iii) adapting these into existing state-of-the-art long read aligner pipeline, minimap2 (v2.24), for efficient alignments.

**Results:** We show that HQAlign captures about $4-6\%$ complementary SVs across different datasets which are missed by minimap2 alignments while having a standalone performance at par with minimap2 for real nanopore reads data. For the common SV calls

*SD is corresponding author: suhas@ee.ucla.edu

†DJ and SD are at the University of California, Los Angeles. MC is at the Department of Quantitative and Computational Biology, University of Southern California, Los Angeles. SK is at the University of Washington, Seattle.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

between HQAlign and minimap2, HQAlign improves the start and the end breakpoint accuracy for about $10-50\%$ of SVs across different datasets. Moreover, HQAlign improves the alignment rate to 89.35% from minimap2 85.64% for nanopore reads alignment to recent telomere-to-telomere CHM13 assembly, and it improves to 86.65% from 83.48% for nanopore reads alignment to GRCh37 human genome.

**Availability:** `https://github.com/joshidhaivat/HQAlign.git`

# 1 Introduction

Structural variations (SVs) are genomic alterations of size at least 50 bp long, including insertions, deletions, inversions, duplications, translocations or a combination of these types [1]. The study of these genetic variations has an important role in understanding human diseases, including cancer [2], and begins with sequence alignment from the sample back to the reference genome. Accurate alignment of short reads from high throughput sequencing poses a challenge, especially, in the repetitive regions of the genome which are also the hotspots of nearly 70% of the observed structural variations [3].

Long read sequencing technologies have addressed this problem by producing reads that are longer than the repeat regions, therefore, enabling the detection of variants in the repeat regions at the cost of higher error rates than short read sequencing technologies. This high error rates in the long reads lead to non-contiguous alignment which poses a challenge in variant detection problem, especially, in the repeat regions.

Nanopore sequencing [4, 5] is a long read sequencing technology that provides reads (with average read length  10-kb and the longest read sequenced more than 2-Mb long) that can span these repetitive regions but it has a high error rate of (average) 10%. This high error rates result in low accuracy alignments [6] using state-of-the-art methods including minimap2 (v2.24) [7] which is a fast method designed for the computationally challenging task of long sequence alignment. This problem is further amplified in the repetitive regions such as variable-number tandem repeats (VNTR) region that accounts for a significant fraction of SVs [8, 9].

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

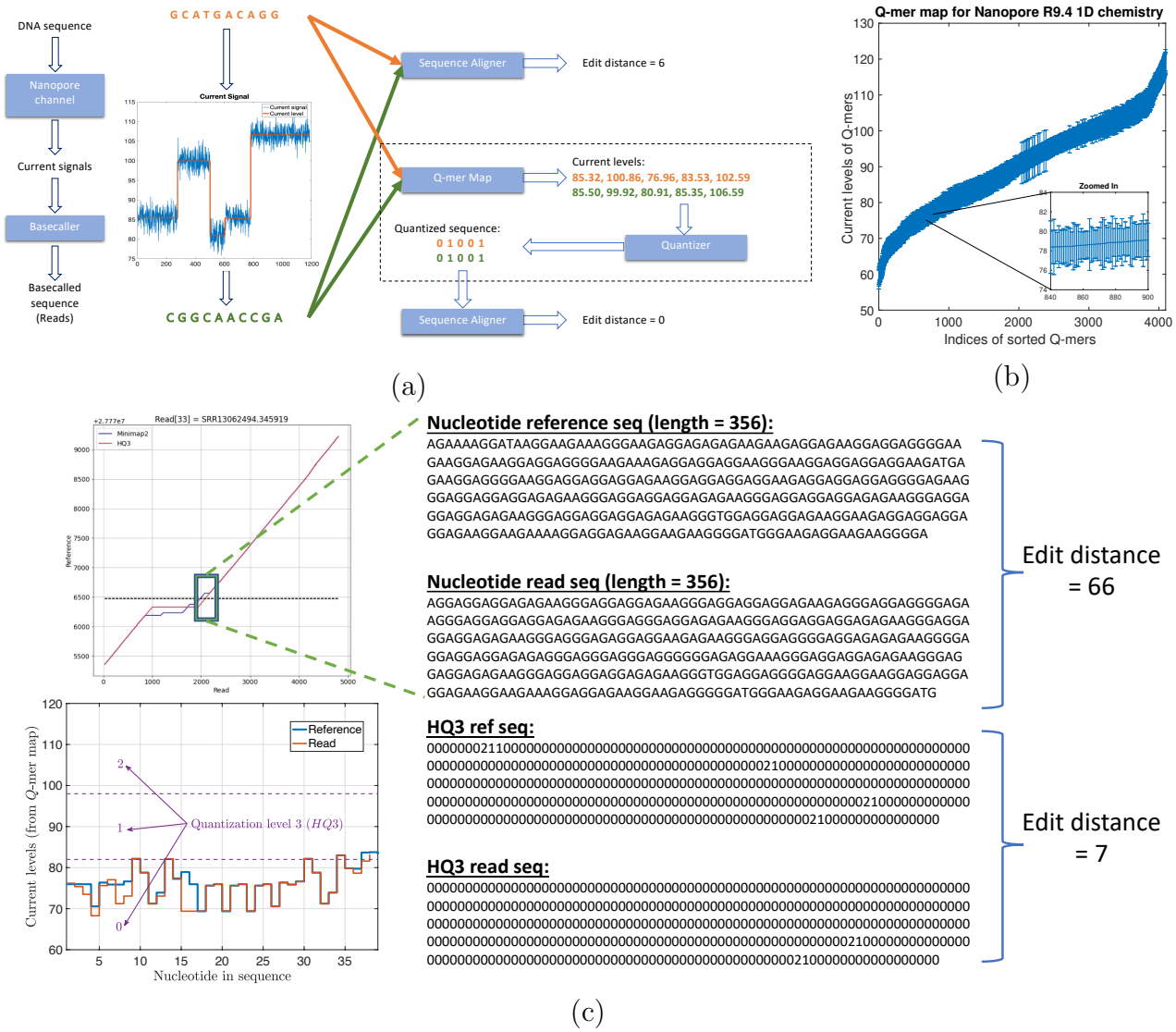



Figure 1: (a) An example to illustrate the error biases in nanopore basecalled reads which can be resolved through the $Q$-mer map ability of HQAlign to perform accurate alignment despite of the errors (the edit distance used here is domain specific and is used to demonstrate accuracy of the alignment). (b) $Q$-mer map for Nanopore R9.4 1D flow cell (for $Q = 6$). It represents the physics of nanopore. The median current value along with the standard deviation (as error bars) are plotted for all 6-mers in the $Q$-mer map for R9.4 1D nanopore flow cell (the $Q$-mers are sorted in increasing median current levels). Note that the difference between the median current levels of any two consecutive $Q$-mers is very small, therefore, resulting in large overlaps. (c) An example from PromethION R9.4.1 ONT data in the neighborhood of a SV in repeat region showing the two different nucleotide sequences have similar current levels and therefore, the edit distance as observed through the lens of quantized sequence is significantly lower in $HQ3$.

However, these errors in nanopore sequencing have a bias induced from nanopore physics which is missed by many long read aligners since they consider the errors as independent insertions,

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

deletions, and substitutions. In nanopore sequencing, a DNA strand migrates through the nanopore, and an ionic current according to the nucleotide sequence in or near the nanopore is established. However, because of the physics and non-idealities of the nanopore sequencing, each current level recorded depends on a $Q$-mer (a set of $Q$ consecutive nucleotide bases which influence the measurement in the nanopore) [10, 11, 12]. These current readings are translated back to nucleotide sequences by basecalling algorithms. Therefore, the error biases could be introduced in basecalling, especially, between different $Q$-mers that have similar current levels. This similarity in the median current levels for different $Q$-mers is captured by the $Q$-mer map as shown in Figure 1b. A $Q$-mer map represents the median current level for different $Q$-mers ($Q = 6$) for nanopore flow cell. It is evident from this figure that there is a significant overlap between the current levels observed for different $Q$-mers migrating the nanopore. We propose a new alignment method, HQAlign (which is based on QAlign [13]), which is designed specifically for detecting SVs while incorporating the error biases inherent in the nanopore sequencing process. HQAlign pipeline is modified to enable detection of inversion variants which was not feasible with the earlier QAlign pipeline (refer to methods section 2.2 for details).

HQAlign takes the dependence of $Q$-mer map into account to perform accurate alignment with modifications specifically for discovery of SVs. Figure 1a gives an example where a DNA sequence (GCATGACAGG) is sequenced incorrectly as (CGGCAACCGA) due to the error bias in nanopore sequencer. Therefore, the sequences are different in the nucleotide space but they are identical in the $Q$-mer map space. It is important to note that *no additional soft information is used* to establish this identity such as raw nanopore current values for the nanopore reads. Instead, the nucleotide sequences that have indistinguishable current levels from the lens of the $Q$-mer map are mapped to a common quantized sequence. A nucleotide sequence is converted to a quantized sequence by first converting the nucleotide sequence to a sequence of current levels using the $Q$-mer map and then converting the sequence of (continuous) current levels to a (finite level) quantized sequence by hard thresholding the current levels (refer to Supplementary material section 1.1 for more details). Therefore, the additional information about the raw current signals is not used in the quantization process but only the $Q$-mer map is utilized.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

This process is explained in detail in Supplementary Figure 1. Further, the quantization of continuous current levels to finite discrete levels enables the use of existing software pipelines of state-of-the-art long read aligners such as minimap2 as the core seed and extend algorithm for the alignment of quantized sequences.

In HQAlign, we first perform the alignment of reads onto genome using minimap2 to determine the region of interest where a read can possibly align to, and then re-align the quantized read to the quantized genome region from the first step. This helps in performing an accurate alignment of the read to the region of genome without dropping the frequently occurring seed matches from the chain in minimap2 algorithm while taking the error biases of nanopore sequencing into account through quantized sequences. Moreover, HQAlign pipeline enables detection of inversion variants unlike QAlign pipeline. In QAlign, the quantized reverse complement of read is aligned separately to the quantized genome, therefore, the alignment of inverted sequence is not observed in QAlign (as shown in Figure 2a). However, in HQAlign, we have modified the minimap2 pipeline to align the reverse complement of quantized read along with the aligning the forward quantized read sequence to the quantized genome, simultaneously. This is necessary for detection of inverted alignment using quantized sequences because unlike nucleotide sequences where minimap2 can inherently produce the reverse complement of the input nucleotide query, the reverse complement of the quantized sequence is to be computed separately. Further, HQAlign is about 2x faster than QAlign as the seed search domain is reduced to a region of interest determined in the first step of the pipeline for the quantized sequences (as shown in Table 1).

Figure 1c demonstrate an example from real ONT reads data in a repeat region (note that a pattern of a few consecutive nucleotide bases is repeated in the example) that is flanking around an insertion structural variant. Minimap2 alignment of nucleotide reference and read (both of length 356 from the region highlighted with a box) have an edit distance of 66 whereas the $HQ3$ alignment ($HQ3$ is an alignment from HQAlign pipeline where the nucleotide sequences are translated to three level quantized sequences, refer to section 2.2 for details) of quantized reference and read sequences from the same region have a significantly smaller edit distance

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of 7. This is because the current level sequence (by converting the nucleotide sequences using the $Q$-mer map in Figure 1b) for the reference and the read are very similar. Therefore, the sequences that are far apart in nucleotide space are inherently very similar in the $HQ3$ space in terms of the edit distance in the transformed space.

We show that HQAlign gives significant performance improvements in quality of read alignment across real and simulated data. The well-aligned reads (a read is defined as well aligned if at least 90% of the read is aligned on the genome with a mapping quality more than 20) improves to 86.65% with $HQ3$ from 83.48% with minimap2 (v2.24) for the alignment of ONT reads from HG002 sample to GRCh37 human genome. The metric improves to 89.35% from 85.64% for HG002 reads alignment to T2T CHM13 assembly [14], and improves to 81.57% from 81.01% for the simulated reads data. These results are presented in the results section 3.1.2 Table 2.

In terms of SV detection, HQAlign has F1 score at par with minimap2 (v2.24) with Sniffles2 [15] as the variant calling algorithm across both real and simulated dataset (Table 3). However, both HQAlign and minimap2 captures many complementary calls ( $4-6\%$) which are missed by the other method (as shown in Figure 6, 7, 8, 9, 10). For instance, the complementary HQAlign calls are SVs that are uniquely called by HQAlign or labeled missed in minimap2 due to breaking in the SV and vice-versa for the complementary calls in minimap2. Further, the analysis of common true positive SV calls in HQAlign and minimap2 against the truth set shows that HQAlign has on average a significant improvement ($10-50\%$, from the slope of the regression line in Figures 11, 12, 13, 14, and weighted average across all datasets for 39% SVs) in the breakpoint accuracy than minimap2 for the calls with difference in breakpoint greater than 50 bp (breakpoint accuracy is determined from the difference in the start and end breakpoints of a SV with respect to the match SV in truth set, therefore, lower the difference higher is the breakpoint accuracy, refer to section 2.3 for precise definition). Moreover, for the common true positive calls, HQAlign has (on average) better SV length similarity than minimap2 (when SV length similarity is less than 0.95, SV length similarity is a measure of how similar is the length of SV from an alignment method relative to the match SV in truth set; refer to section 2.3 for

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

a precise definition) as shown in Figures 11, 12, 13, and 14.

# 2 Methods

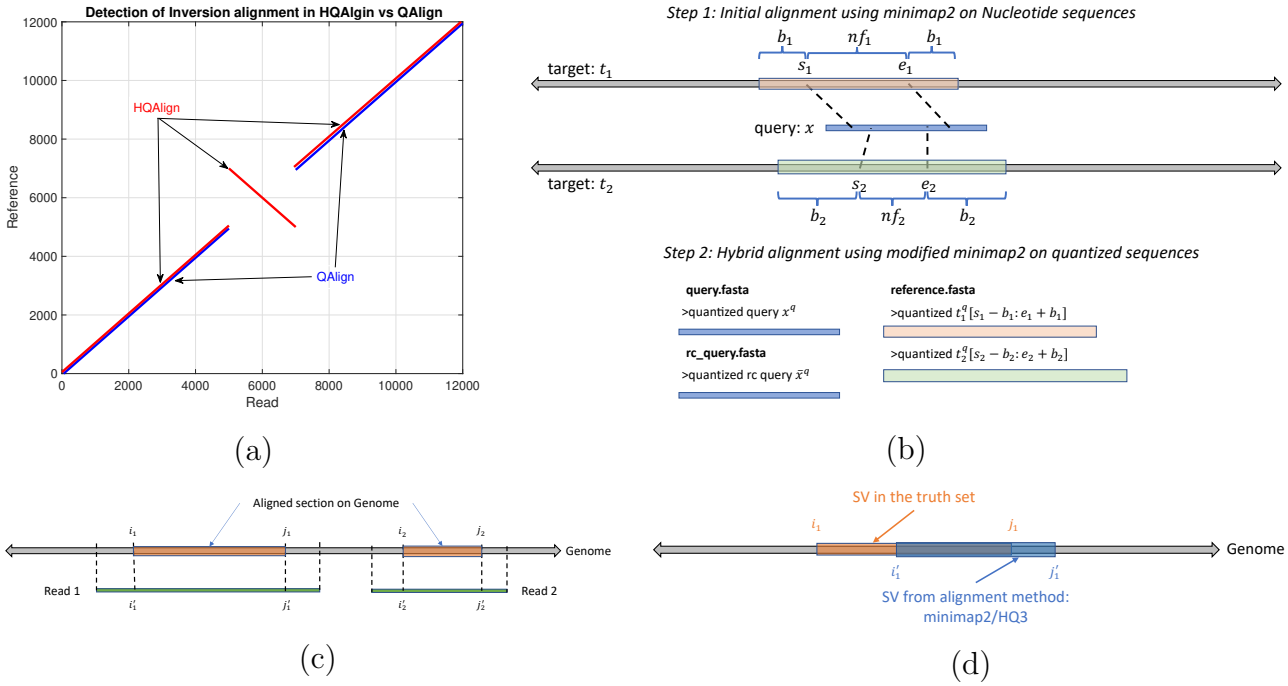



Figure 2: (a) An example to demonstrate the ability of HQAlign pipeline to align inverted sequences where QAlign fails (b) An example of HQAlign pipeline. (c) An example of read-to-genome alignment. (d) Comparison of SV in truth set to SV determined by method: minimap2/$HQ3$.

The HQAlign strategy consists of two steps: (1) the initial alignment of the standard base-called query sequence $x$ to a target sequence $t$ using Minimap2. This primitive step identifies the region of interests on the target sequence where $x$ aligns. (2) the hybrid step is re-aligning the query $x$ only to the region of interest determined in the first step using a modified pipeline for QAlign method. QAlign method converts the nucleotide sequences in both query $x$ and target region $t$ to quantized (three levels) sequence, and then aligns these quantized sequences using any state-of-the-art aligner. We have modified the minimap2 (v2.24) pipeline for the alignment of the quantized sequences in the second step so that it enables detection of the inversion variants in alignment using the quantized sequences. This is explained in detail in section 2.2. Further, this strategy is about 2x faster than QAlign standalone as it narrows down the seed

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

search domain for lower alphabet size (*e.g.* three levels) in QAlign. This strategy is explained in Figure 2b, and mathematically in the following sections.

Table 1: Comparison of computation time for alignment of 500k randomly sampled ONT reads to CHM13 assembly using 20 threads for each method.

| **Method** | **minimizer (seed) length** | **Real time (in seconds)** | **CPU time (in seconds)** |
|---|---|---|---|
| QAlign ($Q3$) | 18 | $10,312$ | $177,516$ |
| HQAlign ($HQ3$) | 18 | $5,033$ | $76,375$ |

## 2.1 Initial alignment

The nucleotide query $x$ is aligned to a nucleotide target sequence $t$ using minimap2. This is similar to aligning a read to a genome with one chromosome. Here we consider only one chromosome in target $t$ for simplicity but the method generalizes to multiple chromosomes in $t$ such as $t = (t_1, t_2, \dots, t_m)$ (this generalization is explained in detail in Supplementary section 1.2). This step identifies the region of interests on the target $t$, say, $t[s_i : e_i]$, where $i \in \{1, 2, 3, \dots\}$ represent one or more alignments on $t$ and $s_i$ and $e_i$ are the corresponding start and end location of each alignment $i$ on target $t$, respectively.

## 2.2 Hybrid alignment

In this step, the query $x$ is re-aligned to an extended region of interest on the target $t[s_i^q : e_i^q]$ using the modified pipeline for QAlign method, where $s_i^q = s_i - b_i$ and $e_i^q = e_i + b_i$, $b_i = (1 - f_i + 0.25)n$ is an appended extension of the region of interest on target, $f_i = (e_i - s_i)/n$ is the fraction of read aligned in initial step, and $n$ is the length of the query $x$. The nucleotide query $x$ and the nucleotide extended target $t[s_i^q : e_i^q]$ are converted to the quantized query $x^q$ and quantized extended target $t^q[s_i^q : e_i^q]$, respectively, using the quantization method demonstrated in QAlign (refer to Supplemental section 1.1 for more details on quantization process). It is important to note that we do not use any additional soft information such as raw current signals from nanopore sequencing in the quantization process, instead, we translate the basecalled nucleotide

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

reads to current levels using the $Q$-mer map (in Figure 1b) and then hard threshold the current levels to finite (*e.g.* three) levels to get the quantized ($HQ3$) reads (refer Supplementary Figure 1 for more details on quantization process). These quantized sequences are then aligned using a modified pipeline of minimap2 (v2.24). We have modified minimap2 pipeline for this hybrid step to accept the quantized reverse complement query $\bar{x}^q$ as an input which helps in indentifing the inversion SVs in contiguous alignment with quantized sequences which was not possible with the earlier QAlign method as shown in Figure 2a. QAlign uses the default minimap2 pipeline for the alignment of quantized sequences which inherently aligns both the forward and the reverse complement strand of the sequences in nucleotide domain. However, the quantized reverse complement sequence cannot be computed given only the forward quantized sequence, therefore, QAlign separately aligns both quantized forward and quantized reverse complement sequence. This method, however, fails to identify an inverted alignment as shown in Figure 2a. Therefore, in HQAlign, we have modified the minimap2 pipeline to enable alignment using both quantized forward and quantized reverse complement sequence, simultaneously. Note that the quantized alignment employs a different minimizer length `k` = 18 in minimap2 for ternary ($HQ3$) quantization.

We define several metrics that are used for the performance evaluation of HQAlign against minimap2 (these metrics are used from the earlier QAlign method [13]).

(i) **well-aligned:** Consider in Figure 2c, *Read 1* aligns at location $i_1$ through $j_1$ on the genome determined using nucleotide alignment. We say that the read is well-aligned, if at-least 90% of the read is aligned onto the genome (i.e., $j_1 - i_1 \geq 0.9(\text{length}(\textit{Read 1}))$), and has high mapping quality (greater than 20). This metric quantifies the reads that are mapped almost entirely to the reference.

(ii) **normalized edit distance:** In order to compare the quality of the alignments at fine-grained level, we further define normalized edit distance. The normalized edit distance

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

for nucleotide alignment is defined as

$$\frac{\text{edit_distance}\{r; G[i_1 : j_1]\}}{\text{length}(r)} \tag{1}$$

and for quantized alignment is

$$\frac{\text{edit_distance}\{r; G[i_1^q : j_1^q]\}}{\text{length}(r)} \tag{2}$$

where $i_1, j_1$ are the start and end location of alignment on genome in nucleotide space and $i_1^q, j_1^q$ are the start and end location of alignment on genome in the quantized space, $r$ is the entire read and $G$ is the genome as shown in Figure 2c. It is important to note that for computing the normalized edit distance for alignments in the quantized space, we only leverage the information of location of the alignment on genome from quantized space, *i.e.* $i_1^q$ and $j_1^q$, but the edit distance between read and the aligned section on genome is computed on the nucleotide sequences. This metric gives a measure of the distance similarity between two sequences, especially, used for the real data where the truth of sequence sampling location is not known.

(iii) **normalized alignment length:** Another metric at the fine-grained level is normalized alignment length, which is the ratio of the length of the section on genome where a read aligns to the length of the read. It is

$$\frac{j_1 - i_1}{\text{len}(r)} \tag{3}$$

for nucleotide alignment, and

$$\frac{j_1^q - i_1^q}{\text{len}(r^Q)} \tag{4}$$

for quantized alignment. A contiguous alignment tends to have this metric as 1. This metric gives a measure of the contiguity of the alignment.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

## 2.3 SV calling

The alignments from HQAlign and minimap2 in sorted *bam* format are used to detect structural variants using Sniffles2. These calls are benchmarked against a truth set using Truvari [16]. We have used F1 score, precision and recall as the metric to analyze the performance of HQAlign and compare them with minimap2. Precision ($P$) is defined as the fraction of SVs detected by the algorithm in the truth set among the total SVs detected by the algorithm. Recall ($R$) is the fraction of SVs detected by the algorithm in the truth set among the total SVs in the truth set. F1 score is the harmonic mean of precision and recall ($= \frac{2P \cdot R}{P+R}$). Further, we have observed that there are many complementary SV calls made by both minimap2 and $HQ3$ that are missed by the other method. Therefore, we have defined a union model which takes a union of the SV calls from both minimap2 and $HQ3$. The precision, recall, and F1 score of the union model are also computed and reported in Table 3.

Further, the quality of the SVs for the common calls in minimap2 and HQAlign is evaluated by comparing the following metrics w.r.t. the SVs in truth set

(i) **breakpoint accuracy:** Breakpoint accuracy is measured by taking an average of the difference in the in the start and end breakpoint of the SV w.r.t. the SV in truth set. For instance, as shown in Figure 2d, $i_1$ and $j_1$ are the start and the end point on genome of SV in the truth set, and $i_1^{'}$ and $j_1^{'}$ are the start and the end point of the same SV determined by any alignment method (minimap2/$HQ3$), then breakpoint score is calculated as

$$\frac{|i_1^{'} - i_1| + |j_1^{'} - j_1|}{2} \tag{5}$$

where $|\cdot|$ is absolute value function. Therefore, lower the score higher is the breakpoint accuracy of the SV determined by the alignment method.

(ii) **SV length similarity:** SV length similarity is measured as the ratio of minimum SV length in truth set and from algorithm to the maximum of two values. Mathematically,

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

it is

$$\frac{\min(j_1 - i_1, j_1^{'} - i_1^{'})}{\max(j_1 - i_1, j_1^{'} - i_1^{'})} \tag{6}$$

for the example shown in Figure 2d.

# 3 Results

In this section, we demonstrate the results for (1) comparison of alignments from HQ3 and minimap2 on real as well as simulated data, and (2) comparison of SV calls from HQ3 and minimap2 alignments using Sniffles2 as the variant caller on real and simulated data.

## 3.1 DNA read-to-genome alignment

### 3.1.1 Datasets

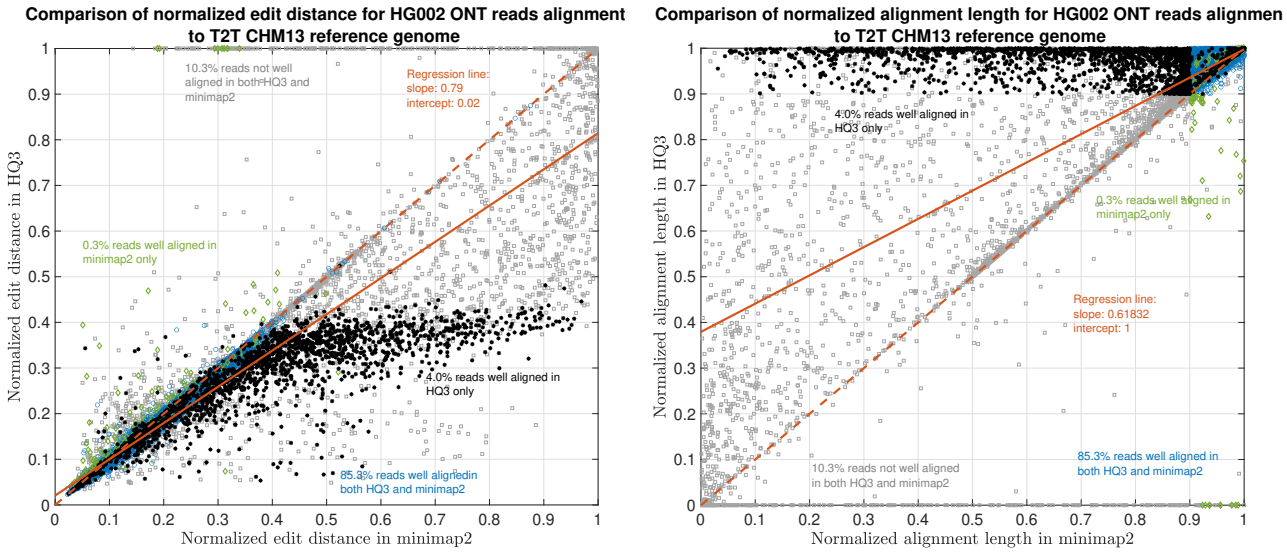
We have used the publicly available R9.4.1 ONT PromethION reads dataset from HG002 sample [17]. These reads are aligned to the recent telomere-to-telomere assembly CHM13 and to the human reference genome GRCh37. GRCh37 is used as the reference build to map the real data so that the curated variants can be used for accuracy analysis [18]. Further, we have also benchmarked the performance of HQAlign and minimap2 on simulated data for both alignments and SV calling.

### 3.1.2 Alignment results

The alignment of DNA reads to the genome is a primitive step in structural variant calling pipelines [19]. *HQ*3 alignments shows an improvement over minimap2 alignments in terms of contiguity measured by normalized alignment length and alignment quality measured by normalized edit distance.

The results are illustrated in the Figures 3, 4, 5, and Table 2. At a coarse level, the performance is measured by the fraction of the reads that are well-aligned by the algorithm. A read is well-aligned if at-least 90% of the read is aligned to genome and has a high mapping

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.




(a) (b)

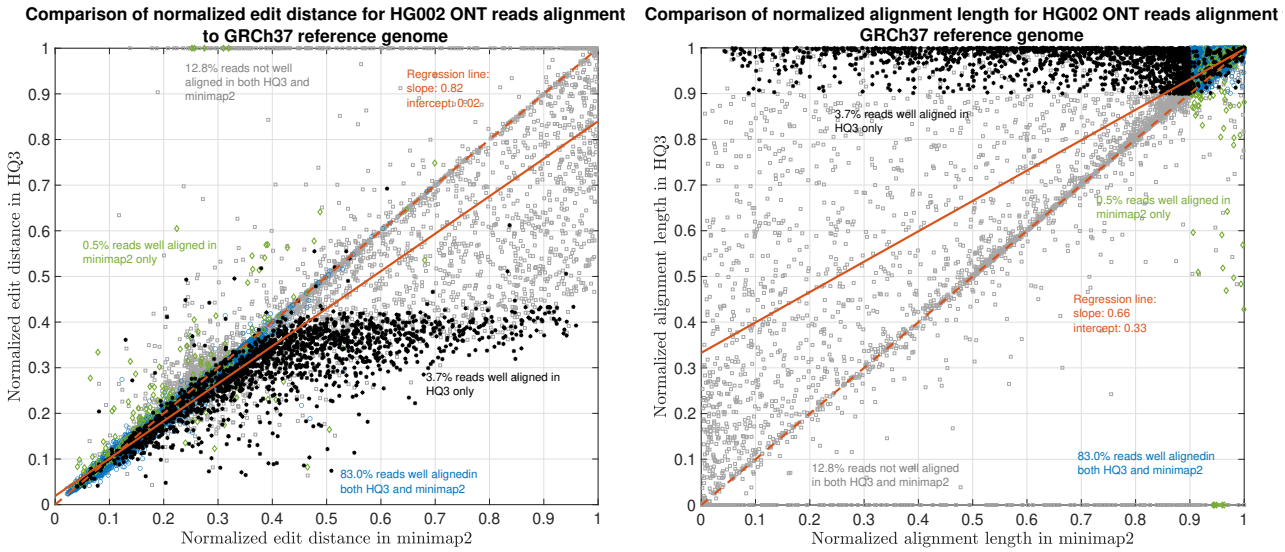
Figure 3: **HG002 nanopore long DNA reads alignment onto T2T CHM13 genome.** (a) Comparison of normalized edit distance for HG002 R9.4.1 PromethION reads data. Smaller values for normalized edit distance is desirable as it represents better alignment. The slope of the regression line is $0.79 < 1$, therefore, representing better alignments with $HQ3$ than minimap2 alignments for same reads on average. (b) Comparison of normalized alignment length for HG002 R9.4.1 PromethION reads data. Normalized alignment length of 1 is desirable as it represents that entire read is aligned. The majority of the reads are above $y = x$ line representing longer alignment length in $HQ3$ than minimap2 alignment.

Table 2: Comparison for the percentage of well-aligned reads onto genome, and slope of the regression line (for normalized edit distance comparison plot of $HQ3$ vs minimap2 alignments) with randomly sampled reads for each datasets. The slope of the regression line shows the average gain in the normalized edit distance.

| **Dataset (No. of sampled reads)** | **Method of alignment** | **Percentage well-aligned reads** | **Slope of regression line** | **Intercept** |
|---|---|---|---|---|
| HG002 R9.4.1 reads to CHM13 (50k) | minimap2 | 85.64 | 0.7940 | 0.0206 |
| | $HQ3$ | 89.35 | | |
| HG002 R9.4.1 reads to GRCh37 (50k) | minimap2 | 83.48 | 0.8301 | 0.0181 |
| | $HQ3$ | 86.65 | | |
| Simulated reads from chr 8 & X of CHM13 assembly (50k) | minimap2 | 81.01 | 0.9860 | 0.0028 |
| | $HQ3$ | 81.57 | | |

quality (see Methods). HQAlign improves the fraction of well-aligned reads than minimap2 - in particular, in the HG002 R9.4.1 reads alignment to T2T CHM13 reference, this metric

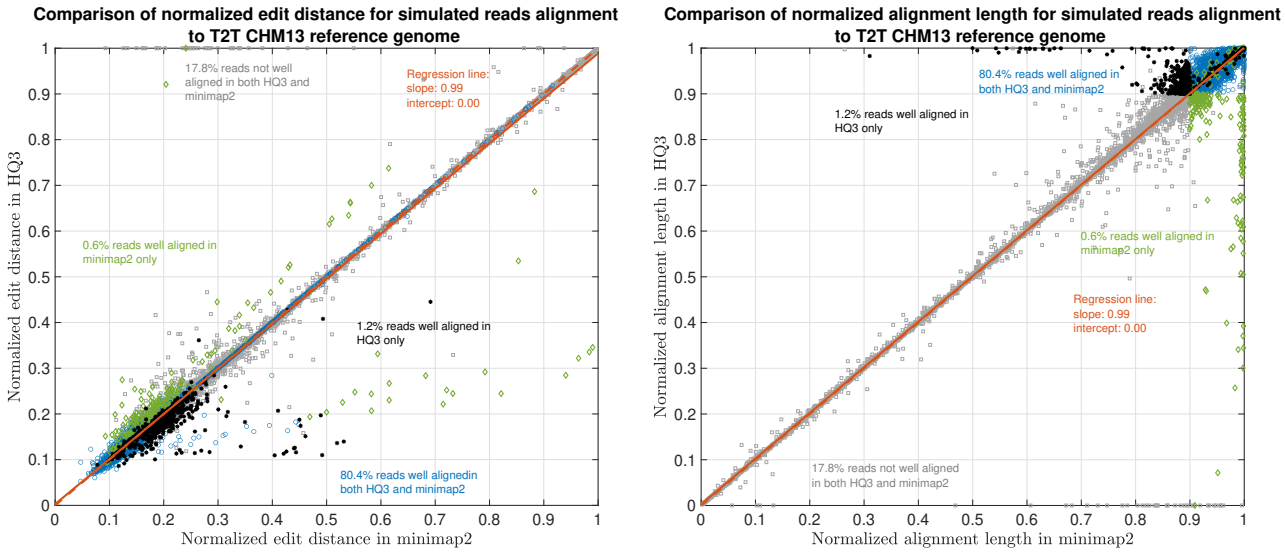
bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.




(a) (b)

Figure 4: **HG002 nanopore long DNA reads alignment onto GRCh37 genome.** (a) Comparison of normalized edit distance for HG002 R9.4.1 PromethION reads data. Smaller values for normalized edit distance is desirable as it represents better alignment. The slope of the regression line is $0.82 < 1$, therefore, representing better alignments with $HQ3$ than minimap2 alignments for same reads on average. (b) Comparison of normalized alignment length for HG002 R9.4.1 PromethION reads data. Normalized alignment length of 1 is desirable as it represents that entire read is aligned. The majority of the reads are above $y = x$ line representing longer alignment length in $HQ3$ than minimap2 alignment.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.




(a) (b)

Figure 5: **Simulated nanopore reads alignment onto T2T CHM13 genome.** (a) Comparison of normalized edit distance for simulated nanopore reads data. Smaller values for normalized edit distance is desirable as it represents better alignment. The slope of the regression line is $0.99 < 1$, therefore, representing marginally better alignments with $HQ3$ than minimap2 alignments for same reads on average. (b) Comparison of normalized alignment length for simulated nanopore reads data. Normalized alignment length of 1 is desirable as it represents a contiguous alignment of the entire read.

improves to 89.35% from 85.64%, and for the alignments to GRCh37 reference, this metric improves to 86.65% from 83.48%. Furthermore, there are $310,036$ reads with at-least 1kb additional bases aligned using HQAlign compared to minimap2 alignments for T2T CHM13 reference, and there are $299,896$ reads with at-least 1kb additional bases aligned using HQAlign compared to minimap2 for GRCh37 reference.

The results in Figure 3 and 4 compares the quality of the alignments using minimap2 and HQAlign at a fine-grained level for HG002 ONT reads alignment to T2T CHM13 genome and GRCh37 genome, respectively. Figure 3a and 4a compares the normalized edit distance for HQAlign and minimap2. The normalized edit distance is the edit distance between the entire read and the aligned section on the genome normalized by the length of the read, in nucleotide domain for *both* minimap2 alignment and quantized alignment ($HQ3$). In case of $HQ3$, the information of the location of the alignment on the genome is leveraged from the quantized read and the quantized genome alignment, and the edit distance is computed between the

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

corresponding nucleotide read and the aligned region on the nucleotide genome (see Methods for details). Intuitively, normalized edit distance gives a measure of how close the two sequences are. Therefore, the smaller the normalized edit distance, better is the alignment.

Figure 3a shows that for alignments of the reads to T2T CHM13 reference, the normalized edit distance is on average smaller for $HQ3$ alignments than minimap2 alignments. The better alignment in $HQ3$ is also evident from the slope of the regression line in Figure 3a. It shows that on average $HQ3$ alignments has 21% improvement in terms of the normalized edit distance than the minimap2 alignments. Well aligned reads in both $HQ3$ and minimap2 are represented by blue circles in Figure 3, well aligned reads in $HQ3$ only are represented in black asterisks, well aligned in minimap2 only are represented in green diamonds and reads that are not well aligned in both are represented in grey squares. Further, it is important to note that for normalized edit distance less than 0.1, the alignments are marginally better in the DNA space, but for normalized edit distance higher than 0.1, the alignments are significantly better in $HQ3$ space, especially, the 4% reads that are well aligned in $HQ3$ and not well aligned in minimap2. This is because of higher contiguity of alignments in $HQ3$ space and signifies the improvement by $HQ3$ when the error rates is higher. For alignments to GRCh37 reference, $HQ3$ has an average improvement of 17%, as shown in Figure 4a.

The results for another fine-grained metric are shown in Figure 3b and 4b, which compares the normalized alignment length in $HQ3$ to the normalized alignment length in minimap2 alignments. The normalized alignment length is the ratio of the length of the section on genome where a read aligns to the length of the read. In Figure 3b, there are 4% reads that are well-aligned in $HQ3$ only, and the normalized alignment length is close to 1 in $HQ3$ but it is much less than 1 in minimap2, therefore representing several non-contiguous alignments in nucleotide domain that are captured as contiguous alignment in $HQ3$. In Figure 4b, there are 3.7% that are well-aligned in $HQ3$ only.

We have also benchmarked the performance of HQAlign with the simulated reads data and compared its alignment performance with minimap2 in Figure 5. The ONT reads are simulated from chromosome 8 and X of CHM13 T2T assembly using nanosim [20] with a coverage of

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

40x, median and mean read length 4.5 kb and 14 kb, respectively. The results shows that the alignment performance of both HQAlign and minimap2 are at par with each other.

## 3.2 SV calling

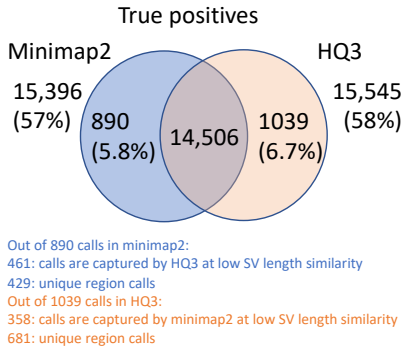
### 3.2.1 Dataset

Long read sequencing plays an important role in detecting structural variations. We evaluated SV detection using minimap2 and HQAlign with Sniffles2 as the variant calling algorithm on both real and simulated data. We simulated 2000 INDELS and 200 Inversion SVs on chromosome 8 and X of T2T CHM13 reference genome using SURVIVOR [21] with SV length uniformly distributed between 50 and 10000, and the ONT reads are simulated using nanosim with average length of 14k, median length of 4.5k and maximum length 2.5Mbp at a coverage of 40x. We have used Truvari to benchmark the calls against the truth set. For real data alignment with GRCh37 as the reference genome, the SV calls are compared against the ground truth sets from (1) Genome In A Bottle (GIAB) Tier 1 calls [18] and (2) another truth set is constructed by comparing the haplotype-resolved assembly of HG002 against GRCh37 reference genome using dipcall [22]. For T2T CHM13 reference genome, since the ground truth for SVs is not available, we have constructed the truth set by comparing the haplotype-resolved assembly of HG002 against CHM13 reference using dipcall. However, it is hard to establish ground truth for the SV calls that are made in the centromere regions, even though the assembly is likely to be correct. Therefore, we have provided both the analysis including the SV calls in centromere regions (in Figures 6 and 11) and the analysis for SV calls excluding the centromere regions (in Figures 10 and 14).

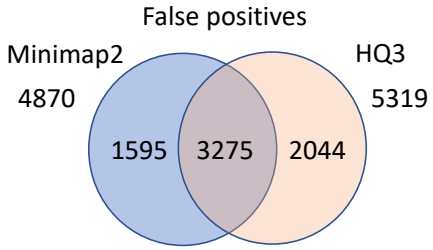
### 3.2.2 SV calling results

The standalone performance of both $HQ3$ and minimap2 is at par with each other across different references and truth set used in this study for real data as well as for the simulated data in terms of the F1 score. However, both $HQ3$ and minimap2 detects complementary SV

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

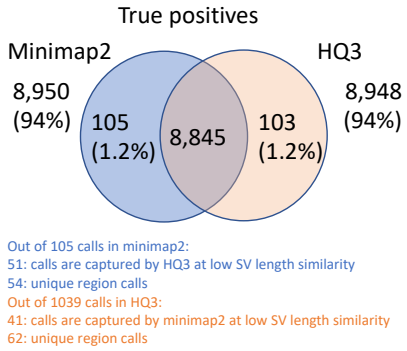



(a)

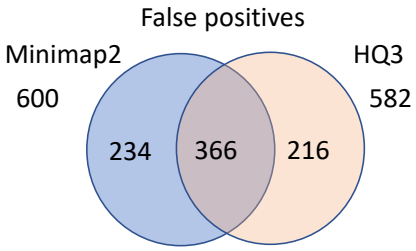



(b)

Figure 6: **Comparison of SV calls from $HQ3$ and minimap2 with HG002-to-CHM13 dipcall as truth set.** (a) Comparison of true positive calls. (b) Comparison of false positive calls.

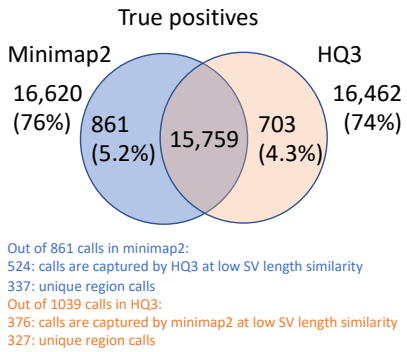



(a)

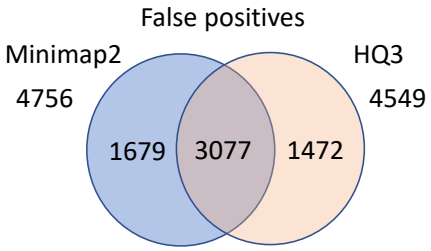



(b)

Figure 7: **Comparison of SV calls from $HQ3$ and minimap2 with Genome in a Bottle (GIAB) Tier 1 truth set for GRCh37 build.** (a) Comparison of true positive calls. (b) Comparison of false positive calls.



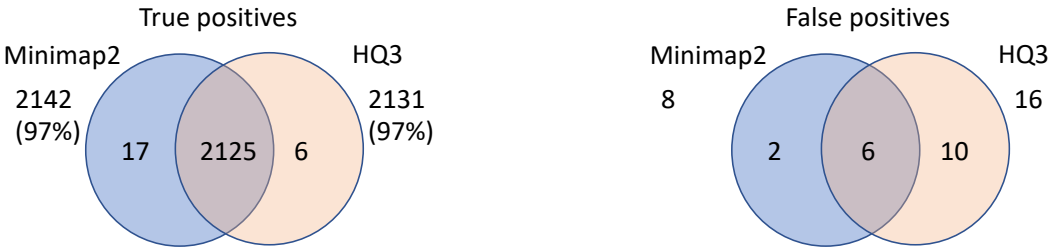

(a)




(b)

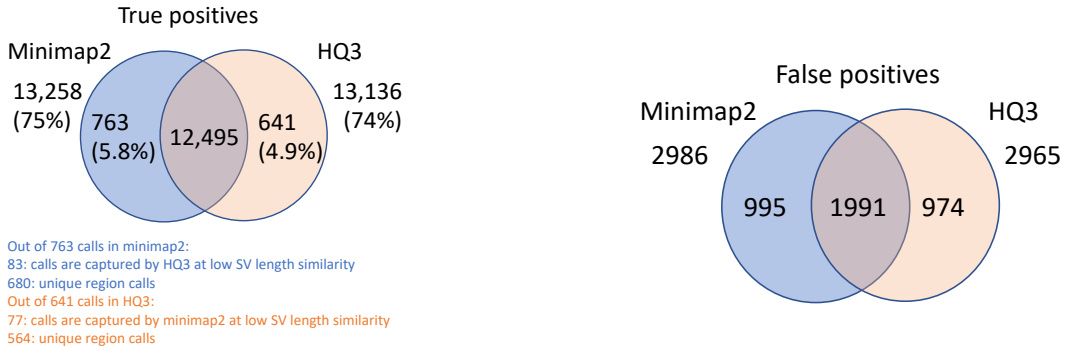
Figure 8: **Comparison of SV calls from $HQ3$ and minimap2 with HG002-to-GRCh37 dipcall as truth set.** (a) Comparison of true positive calls. (b) Comparison of false positive calls.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.




(a) (b)

Figure 9: **Comparison of SV calls from $HQ3$ and minimap2 with simulated data.** (a) Comparison of true positive calls. (b) Comparison of false positive calls.




(a) (b)

Figure 10: **Comparison of SV calls from $HQ3$ and minimap2 with HG002-to-CHM13 dipcall as truth set and excluding the calls from the centromere regions.** (a) Comparison of true positive calls. (b) Comparison of false positive calls.

calls most likely in the repeat regions where accurate alignment is difficult and therefore, leads to many broken calls.

The analysis with comparison of SV calls from $HQ3$ and minimap2 with GIAB Tier 1 truth set gives a precision, recall and F1 score of 0.94, 0.94, and 0.94, respectively for both minimap2 and $HQ3$. A union model of minimap2 and $HQ3$ can improve the recall rate at the same F1 score, and the union model has a precision, recall, and F1 score of 0.93, 0.95, and 0.94, respectively. Moreover, out of 103 SV calls that are made by $HQ3$ only (Figure 7), 41 calls are made by minimap2 alignments at a lower SV length similarity and 62 calls are unique region calls. Out of 105 SV calls made by minimap2 only, 51 are captured by $HQ3$ at a lower SV length similarity and 54 are unique region calls. $HQ3$ improves the breakpoint accuracy for 14.11%

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

calls that have higher than 50 difference in breakpoints and it improves the length similarity of 19.97% calls that have SV length similarity lower than 0.95 (Figure 12).

We have compared the SV calls made by HG002 reads against T2T CHM13 reference genome using both minimap2 and $HQ3$ and benchmarking them against truth set generated by comparing HG002 haplotype-resolved assembly to T2T CHM13 assembly. The standalone performance have precision, recall and F1 score of 0.77, 0.57 and 0.66, respectively for minimap2 and 0.75, 0.58 and 0.65, respectively for $HQ3$. However, because of high number of complementary true positive calls in minimap2 and $HQ3$, the union model has a significant improved recall at the same F1 score with precision, recall and F1 score of 0.71, 0.61 and 0.66, respectively. Out of 1039 (6.7%) calls that are made in $HQ3$ only, 358 are captured by minimap2 at a lower SV length similarity threshold and 681 are unique calls, whereas out of 890 (5.8%) calls that are made by minimap2 only, 461 are captured by $HQ3$ at a lower SV length similarity threshold and 429 are unique (as shown in Figure 6a). Further, for the common true positive calls in both minimap2 and $HQ3$, we observe a similar pattern as the other datasets in improvement of breakpoint accuracy with $HQ3$ for 18.66% calls that have difference in breakpoint greater than 50, and improvement in SV length similarity for 19.76% calls with similarity less than 0.95 (Figure 11a-b).

For SV calls from HG002 reads alignment to GRCh37 and benchmarking them against truth set generated by comparing HG002 haplotype-resolved assembly to GRCh37 build, minimap2 has precision 0.78, recall 0.76, and F1 score 0.77 while $HQ3$ has precision 0.79, recall 0.75, and F1 score 0.77. Out of 16462 true positive calls in $HQ3$, 703 (4.27%) are made only in $HQ3$ with SV length similarity to the truth set greater than 0.7 (default parameter in Truvari). However, 376/703 calls that are captured by minimap2 with SV length similarity less than 0.7 and 327/703 calls that are uniquely made by $HQ3$. Likewise, out of 16620 true positive calls in minimap2, 861 (5.18%) are made only in minimap2 with SV length similarity greater than 0.7. However, 524/861 are captured by $HQ3$ with SV length similarity less than 0.7 and 337/861 are uniquely made by minimap2. A fine-grain analysis of the common true positive calls by minimap2 and $HQ3$ in Figure 13a, shows that a major density of SV calls (81.85%) have

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 3: Comparison for precision, recall and and F1 score for SV calls made by $HQ3$, minimap2, and the Union model.

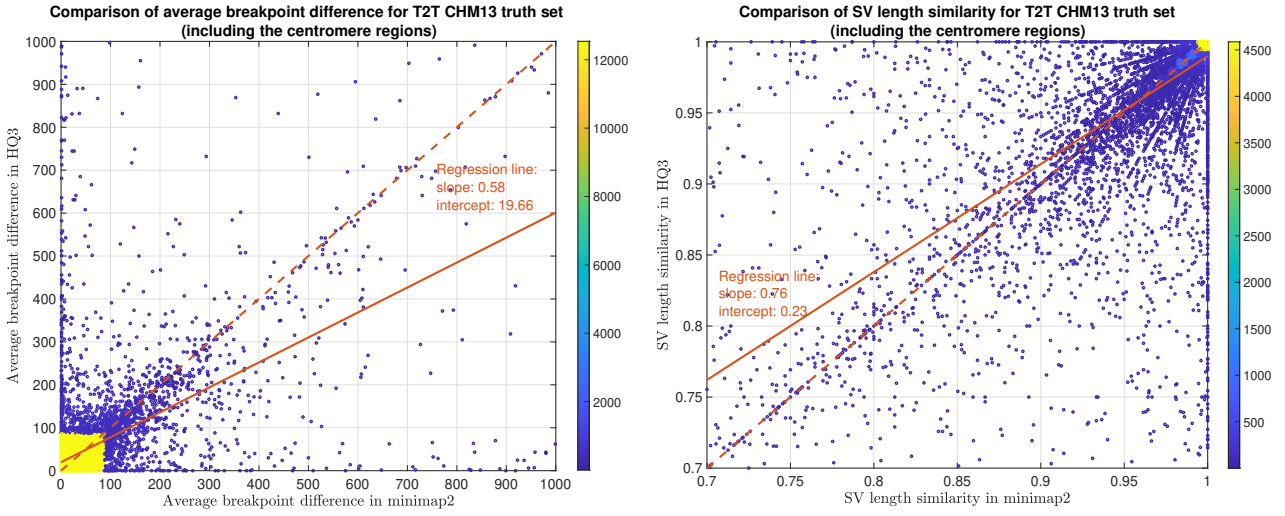
| **Dataset** | **Truth set** | **Method of alignment** | **Precision** | **Recall** | **F1 score** |
|---|---|---|---|---|---|
| HG002 reads to GRCh37 | GIAB Tier 1 calls | minimap2 | 0.94 | 0.94 | 0.94 |
| | | $HQ3$ | 0.94 | 0.94 | 0.94 |
| | | Union | 0.93 | 0.95 | 0.94 |
| HG002 reads to CHM13 | comparing HG002 assembly to CHM13 (including centromere calls) | minimap2 | 0.77 | 0.57 | 0.66 |
| | | $HQ3$ | 0.75 | 0.58 | 0.65 |
| | | Union | 0.71 | 0.61 | 0.66 |
| HG002 reads to CHM13 | comparing HG002 assembly to CHM13 (excluding centromere calls) | minimap2 | 0.82 | 0.75 | 0.78 |
| | | $HQ3$ | 0.82 | 0.74 | 0.78 |
| | | Union | 0.78 | 0.79 | 0.78 |
| HG002 reads to GRCh37 | comparing HG002 assembly to GRCh37 | minimap2 | 0.78 | 0.76 | 0.77 |
| | | $HQ3$ | 0.79 | 0.74 | 0.77 |
| | | Union | 0.74 | 0.79 | 0.77 |
| Simulated reads to CHM13 | Simulated SVs on chr 8 and X | minimap2 | 0.99 | 0.97 | 0.98 |
| | | $HQ3$ | 0.99 | 0.97 | 0.98 |
| | | Union | 0.99 | 0.98 | 0.98 |

difference in breakpoint below 50 in both minimap2 and $HQ3$, and minimap2 has marginally better performance in terms of lower difference in breakpoint of SVs for difference in breakpoint below 50. Whereas, for a large difference in the SV breakpoint (greater than 50), $HQ3$ is better in terms of the breakpoint accuracy of the SV calls (on average across all SV calls). Therefore, $HQ3$ improves the SV breakpoint for the rest 18.15% calls that have high difference in breakpoints. Further, Figure 13b demonstrate that $HQ3$ has better SV length similarity when the length similarity is below 0.95 which corresponds to 21.82% calls.

# 4 Discussion

HQAlign method is an alignment method designed for the detection of structural variants for nanopore sequencing reads. HQAlign provides alignment that outperforms the recent minimap2 aligner in terms of the accuracy and quality of the alignments. The SV calling from HQAlign is also at par with minimap2 in terms of F1 score and it outperforms minimap2 SV calls in terms

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



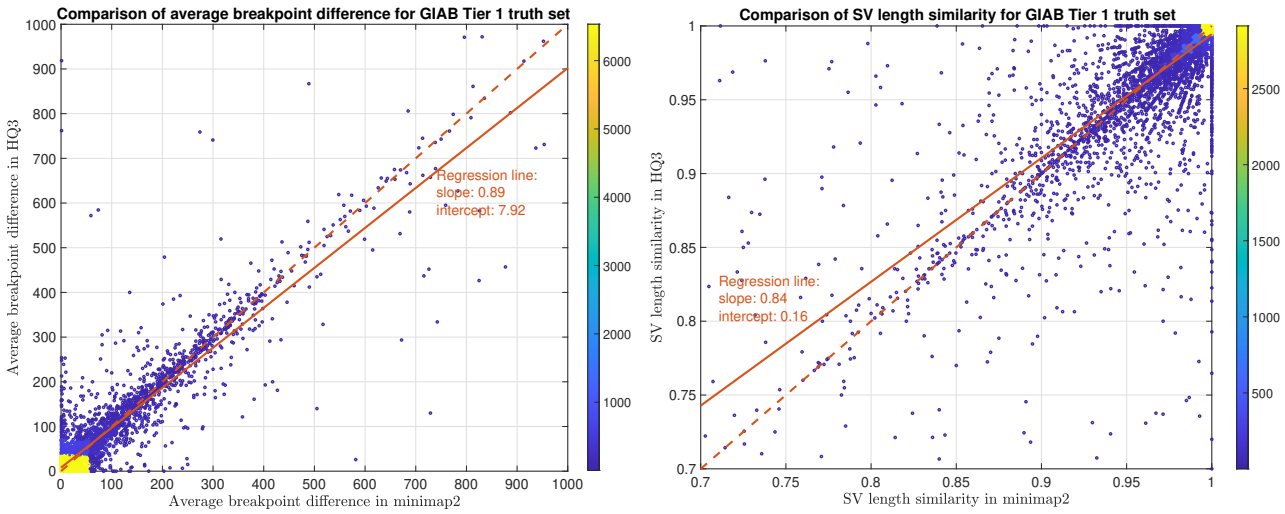

(a) (b)

Figure 11: **SV quality comparison for common true positive calls in $HQ3$ and minimap2 against HG002-to-CHM13 dipcall truth set.** (a) Comparison of SV breakpoint accuracy in $HQ3$ and minimap2 for common true positive calls. The difference of SV breakpoint is compared to the truth set generated from comparing HG002 haplotype-resolved assembly to T2T CHM13 build. A smaller difference represents better breakpoint accuracy. Therefore, slope of the regression line $0.58 < 1$ represents better accuracy of $HQ3$ than minimap2 on average. (b) Comparison of SV length similarity in $HQ3$ and minimap2 for common true positive calls. The slope of the regression line $0.76 < 1$ represents better SV length in minimap2 than $HQ3$ on average, but the intercept is high (0.23). However, this is due to a large density of SVs with length similarity $\geq 0.95$ in both minimap2 and $HQ3$. For length similarity less than 0.95, $HQ3$ has better performance than minimap2.

of the quality of SVs measured in breakpoint accuracy and SV length similarity. Moreover, there are many complementary SVs captured by HQAlign that are missed by minimap2 alignments.

The reason for this improvement in the performance of alignment and SV calling with HQAlign is that it takes into account the underlying physics of nanopore sequencer through the $Q$-mer map, which could be one of the major causes of the high error rates in nanopore sequencing, and also it focuses on a narrow region of the genome (where the read aligns in nucleotide domain) for alignment with quantized sequences. Further, this pipeline is adapted specifically for the detection of SVs. We demonstrated how HQAlign utilizes the bias of $Q$-mer map without accessing the raw current signal of nanopore sequencer by translating the basecalled nucleotide sequences to quantized current level (of finite alphabet size) sequences. This improvement help in detecting several SVs that are missed by minimap2 due to high error

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.




(a) (b)

Figure 12: **SV quality comparison for common true positive calls in $HQ3$ and minimap2 against GIAB Tier 1 truth set.** (a) Comparison of SV breakpoint accuracy in $HQ3$ and minimap2 for common true positive calls. The difference of SV breakpoint is compared to the GIAB Tier 1 truth set. A smaller difference represents better breakpoint accuracy. Therefore, slope of the regression line $< 1$ represents better accuracy of $HQ3$ than minimap2 on average. (b) Comparison of SV length similarity in $HQ3$ and minimap2 for common true positive calls. The slope of the regression line $< 1$ represents better SV length in minimap2 than $HQ3$ on average. However, this is due to a large density of SVs with length similarity $\geq 0.95$ in both minimap2 and $HQ3$. For length similarity less than 0.95, $HQ3$ has better performance than minimap2.

rates in the nanopore reads. Further, the recall rate for SV detection can be improved by combining the complementary calls from both $HQ3$ and minimap2 in the union model at the same F1 score.

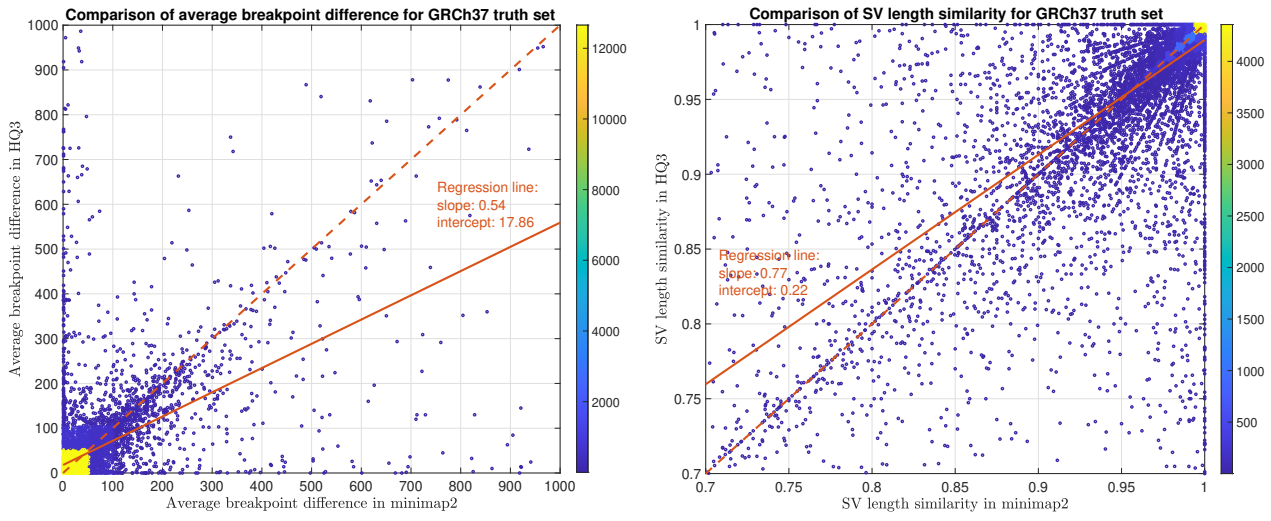
# Competing interests

The authors declare that they have no competing interests.

# Author's contributions

DJ, SK, and SD conceived the original idea and developed the project. DJ led the development of the software tool and its open-source development. MC helped with SV metrics, datasets, and

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.




(a) (b)

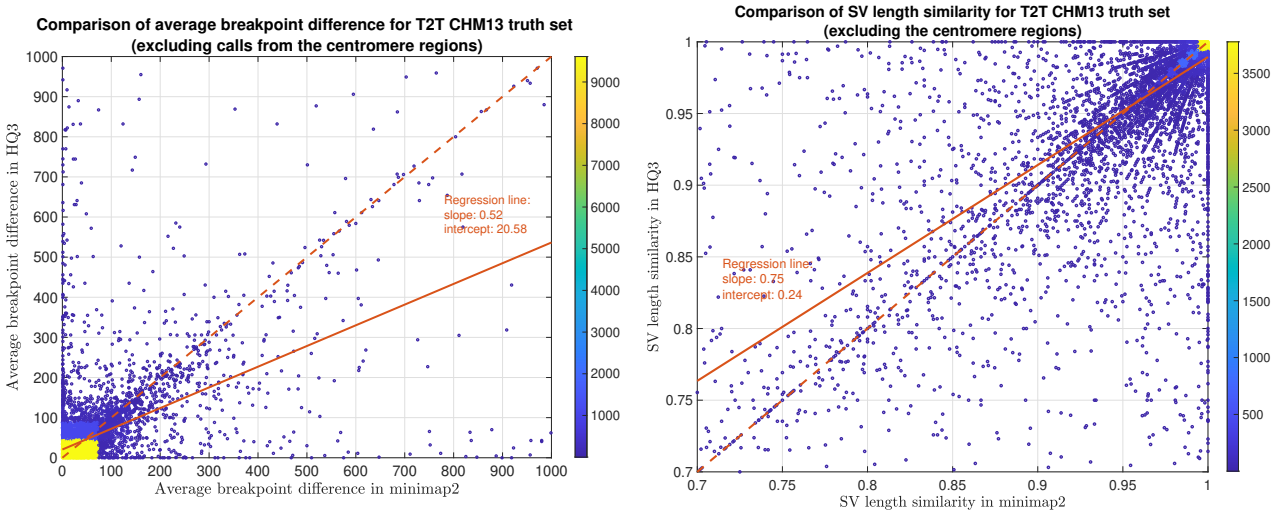
Figure 13: **SV quality comparison for common true positive calls in $HQ3$ and minimap2 against HG002-to-GRCh37 dipcall truth set.** (a) Comparison of SV breakpoint accuracy in $HQ3$ and minimap2 for common true positive calls. The difference of SV breakpoint is compared to the truth set generated from comparing HG002 haplotype-resolved assembly to GRCh37 build. A smaller difference represents better breakpoint accuracy. Therefore, slope of the regression line $< 1$ represents better accuracy of $HQ3$ than minimap2 on average. (b) Comparison of SV length similarity in $HQ3$ and minimap2 for common true positive calls. The slope of the regression line $< 1$ represents better SV length in minimap2 than $HQ3$ on average. However, this is due to a large density of SVs with length similarity $\geq 0.95$ in both minimap2 and $HQ3$. For length similarity less than 0.95, $HQ3$ has better performance than minimap2.

SV comparison analysis between methods. DJ performed the analysis on the various datasets for both alignment and SV calling. All the authors wrote the manuscript.

# Acknowledgements

SD and DJ were supported in part by National Science Foundation grant 1705077. MC was supported by National Institutes of Health grant R01HG011649. SK was supported in part by National Institutes of Health grant 1R01HG008164 and National Science Foundation grants 1651236 and 1703403.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.




(a) (b)

Figure 14: **SV quality comparison for common true positive calls in $HQ3$ and minimap2 against HG002-to-CHM13 dipcall truth set (excluding the centromere region).** (a) Comparison of SV breakpoint accuracy in $HQ3$ and minimap2 for common true positive calls. The difference of SV breakpoint is compared to the truth set generated from comparing HG002 haplotype-resolved assembly to T2T CHM13 build and the SV calls in the centromere region are excluded in this truth set. A smaller difference represents better breakpoint accuracy. Therefore, slope of the regression line $< 1$ represents better accuracy of $HQ3$ than minimap2 on average. (b) Comparison of SV length similarity in $HQ3$ and minimap2 for common true positive calls. The slope of the regression line $< 1$ represents better SV length in minimap2 than $HQ3$ on average. However, this is due to a large density of SVs with length similarity $\geq 0.95$ in both minimap2 and $HQ3$. For length similarity less than 0.95, $HQ3$ has better performance than minimap2.

# References

[1] Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nature reviews genetics* **12**, 363–376 (2011).

[2] Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

[3] Rowell, W. J. *et al.* Comprehensive variant detection in a human genome with highly accurate long reads. *EUROPEAN JOURNAL OF HUMAN GENETICS* **27**, 1723–1723 (2019).

[4] Mikheyev, A. S. & Tin, M. M. A first look at the oxford nanopore minion sequencer. *Molecular ecology resources* **14**, 1097–1102 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[5] Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nature biotechnology* **34**, 518–524 (2016).

[6] Križanović, K., Echchiki, A., Roux, J. & Šikić, M. Evaluation of tools for long read rna-seq splice-aware alignment. *Bioinformatics* **34**, 748–754 (2018).

[7] Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).

[8] Chaisson, M. J. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature communications* **10**, 1–16 (2019).

[9] Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).

[10] Laszlo, A. H. *et al.* Decoding long nanopore sequencing reads of natural dna. *Nature biotechnology* **32**, 829 (2014).

[11] Nanopore - how it works. `https://nanoporetech.com/how-it-works`.

[12] Mao, W., Diggavi, S. N. & Kannan, S. Models and information-theoretic bounds for nanopore sequencing. *IEEE Transactions on Information Theory* **64**, 3216–3236 (2018).

[13] Joshi, D., Mao, S., Kannan, S. & Diggavi, S. Qalign: aligning nanopore reads accurately using current-level modeling. *Bioinformatics* **37**, 625–633 (2021).

[14] Rhie, A. *et al.* The complete sequence of a human y chromosome. *bioRxiv* (2022).

[15] Smolka, M. *et al.* Comprehensive structural variant detection: from mosaic to population-level. *BioRxiv* (2022).

[16] English, A. C., Menon, V. K., Gibbs, R., Metcalf, G. A. & Sedlazeck, F. J. Truvari: Refined structural variant comparison preserves allelic diversity. *bioRxiv* (2022).

[17] Ren, J. & Chaisson, M. J. lra: A long read aligner for sequences and contigs. *PLOS Computational Biology* **17**, e1009078 (2021).

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.08.523172; this version posted January 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[18] Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nature biotechnology* **38**, 1347–1355 (2020).

[19] DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics* **43**, 491 (2011).

[20] Yang, C., Chu, J., Warren, R. L. & Birol, I. Nanosim: nanopore sequence read simulator based on statistical characterization. *GigaScience* **6**, gix010 (2017).

[21] Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature communications* **8**, 1–11 (2017).

[22] Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature methods* **15**, 595–597 (2018).