QAlign: Aligning nanopore reads accurately using current-level modeling

Dhaivat Joshi, Shunfu Mao, Sreeram Kannan*, and Suhas Diggavi *†

Abstract

Efficient and accurate alignment of DNA / RNA sequence reads to each other or to a reference genome / transcriptome is an important problem in genomic analysis. Nanopore sequencing has emerged as a major sequencing technology and many long-read aligners have been designed for aligning nanopore reads. However, the high error rate makes accurate and efficient alignment difficult. Utilizing the noise and error characteristics inherent in the sequencing process properly can play a vital role in constructing an efficient aligner. In this paper, we design QAlign, a pre-processor that can be used with any long-read aligner for aligning long reads to a genome / transcriptome or to other long reads. The key idea in QAlign is to convert the nucleobase reads into discretized current levels that capture the error modes of the nanopore sequencer before running it through a sequence aligner. We show that QAlign improves alignment rates from around 80% to 90% with nanopore reads when aligning to the genome. We also show that QAlign improves the average overlap quality by $9.2\%, 2.5\%$ and 10.8% in three datasets for read-to-read alignment. Read to transcriptome alignment rates are improved from 50.8% to 86.3% and 82.3% to 95.3% in two datasets.

1 Introduction

In genomic data analysis, aligning DNA / RNA-seq reads to a genome / transcriptome is a key primitive, that precedes many downstream tasks, including variant calling [1][2][3], as well as genome and transcriptome assembly [4][5]. Getting accurate read alignment is difficult especially in repetitive regions of the genome, due to the short length of the reads obtainted via high throughput sequencing. Emerging sequencing technologies, particularly, nanopore sequencing [6][29] offers a potential solution

*SK and SD are corresponding authors: ksreeram@uw.edu and suhas@ee.ucla.edu

†DJ and SD are at the University of California, Los Angeles. SM and SK are at the University of Washington, Seattle

to this problem by providing long reads (spanning up to 100,000 bases) that can span these repetitive regions. However, these long reads are riddled with a high error rate thus making alignment of low accuracy [7] as well as the downstream taks difficult. For example, while nanopore sequencing has enabled fully automated assembly of some bacterial genomes, the assembly of human genome still produces many contigs that then have to be scaffolded manually [8]. Another important downstream task is structural variant calling, where long reads can play an important role. However, present structural variant calling algorithms have low precision and recall due to noise in the reads [9]. The assembly of long segmental duplications presents another important problem where long reads can bridge repeated regions but again becomes complicated due to read errors [10].

In this paper, we propose a novel method for aligning nanopore reads that takes into account the particular structure of errors that is inherent in the nanopore sequencing process. In many of the long read aligners, the read errors are modeled using insertions, deletions and substitutions which happen at differing rates. However, in nanopore sequencing, many errors induced have structure, which is missed by viewing the errors as independent insertions, deletions and substitutions. In the nanopore sequencer, the current level depends on a Q-mer (a set of Q bases which influence the current measurement in the nanopore). This is due to the physics of the nanopore sequencing, where a set of DNA base-pairs together influence the current output of the nanopore reader [12, 13, 14] (*e.g.,* occupying the nanopore width). Therefore, the output current depends on a set of DNA base-pairs (Q-mer) influencing it. The current reading, which is used by a de-novo base caller for decoding, therefore could cause structured errors, especially between Q-mers that have similar outputs. This confusability between different Q-mers, is captured by the so-called Q-mer map. In Figure 1b, the median current levels for various Q-mers are plotted and it is clear that there is significant overlap in the current levels observed when different Q-mers are passed through the nanopore. These overlaps are one source of structured errors in the sequencer, since they can be indistinguishable by any de-novo sequencer.

In this paper, we propose a novel alignment strategy that takes into account the structure of the Q-mer map in order to perform alignment. In Figure 1a , we give an example where a DNA sequence (GCATGACAGG) gets wrongly sequenced as a completely different sequence (CGGCAACCGA) due to this error mode of the nanopore sequencer. Ideally, we would like to maintain the list of "equivalent" Q-mers that could have plausibly caused the observed current readings. However, this is infeasible as this would entail changing the de-novo sequencing process itself to output either multiple possible

reads, or give soft information about different possiblities. This is difficult, as sophisticated de-novo sequencing have been developed using artificial neural networks, which have been optimized for read error-rate performance [11]. Moreover, for a modular approach, we would not want to change the de-novo sequencer for different downstream applications. Therefore, we take a *different* approach to resolve this problem, by using the de-novo sequenced read as the *input* to our strategy. We then deterministically *convert* this de-novo sequenced "ACGT" read into a current value using the Q-mer median current level of the corresponding Q-mer, using the Q-mer map as in Figure 1b. We further *quantize* (*i.e.,* represent the continuous valued conversion by a properly chosen *discrete* set) these resulting current values obtained through this conversion. This is illustrated in Figure 1c. The quantization is optimized with respect to the Q-mer map. In this work we use 2-4 levels of discrete values for the quantization. Now, given this new discrete representation of the de-novo reads, we develop the new alignment algorithm, whose workflow is illustrated in Figure 1a.

A natural question is why this should help, since we are processing the de-novo reads which are erroneous, and we are not using any additional soft information from the reads themselves. The basic insight is that this enables grouping together reads that are confusable given the structure of the Q-mer map of the nanopore reader. For example, when we have two reads illustrated in Figure 1c, if the de-novo sequencer has chosen one of the two as the "ACGT" read, it is clear that the alternate read, which has significant edit distance (in terms of "ACGT") is actually quite close when viewed from the lens of the Q-mer map, as captured by our quantized conversion process. Therefore, this process naturally groups together reads that could have been confused, and uses this as the input to our alignment algorithm, QAlign. Therefore, this reduces the effect of the errors by recognizing one structure in the error process. Note that QAlign builds an overlay layer on top of any alignment algorithm in order to align based on current levels implied by the reads instead of directly aligning the reads. Though we illustrate our ideas using the Minimap2 aligner [15], this principle can be implemented with *any* other long-read aligner such as GMAP [16].

We show that QAlign gives rise to significant performance improvements across a variety of alignment tasks including read-to-genome, read-to-read and read-to-transcriptome alignment as well as different datasets spanning from R7 nanopore sequenced data (Refer to Figure 7 in Supplementary material) to R9.4 data. QAlign shows significant improvement in the read-to-genome alignment rates for datasets where Minimap2 alignment rate is low (improving to around 90% from 80% in some data

sets). Furthermore, the alignments are also of higher quality: QAlign shows 15% lower normalized edit distance than Minimap2 as well as longer alignments.

For read-to-read alignments, QAlign is able to align around 3.5% more overlaps between read pairs with a high overlap quality (refer to Methods section 6.4 for a description of the overlap quality) where Minimap2 is either unable to align the read overlaps or aligns with a low overlap quality. We show that a hybrid alignment strategy which combines QAlign and Minimap2 can improve the accuracy even further to around 4.5%.

For aligning RNA to transcriptome, our method achieves 95.3 % alignment rate as opposed to 82.3% with mouse 2D reads and 86.3% as opposed to 50.8% with 1D reads. Furthermore, the alignments are also of higher quality: QAlign shows 41.15% lower normalized edit distance than Minimap2 as well as longer alignments.

In this study, we focus on the improvement of long read (in particular the Nanopore long read) alignment. To the best of our knowledge, there is no existing aligner, specifically designed to handle the error modes introduced in nanopore sequencing. There is, however, some work on incorporating the nanopore current levels in downstream tasks including post-processing of assembly by Nanopolish [17]. Nanopolish has demonstrated that utilizing the current levels can lead to assemblies with fewer errors. The major difference of our work with Nanopolish is the level at which the current-level information is taken into account. Since we take into account current-level information while performing alignment, we are able to get substantially more overlaps which can lead to potentially better assembly of contigs whereas Nanopolish is only able to correct fine errors.

1.1 Method Overview

The setting for the QAlign algorithm is described in this section. The alignment of a nucleobase (DNA / RNA) sequence, say s_1, onto another DNA sequence, s_2, can be determined using the sequence aligners such as Minimap2. These aligners can also be fine-tuned to use them for the alignment of quantized sequence s_1^Q onto another quantized sequence s_2^Q, and therefore these aligners can be used for QAlign as well.

The quantized sequence, s^Q, is obtained from the DNA sequence, s, using the following pipeline (see Section 6.2 for details). All the 6-mers from the DNA sequence (s) are extracted. These 6-mers are then translated to a current level using the appropriate Q-mer map. Therefore, a sequence of

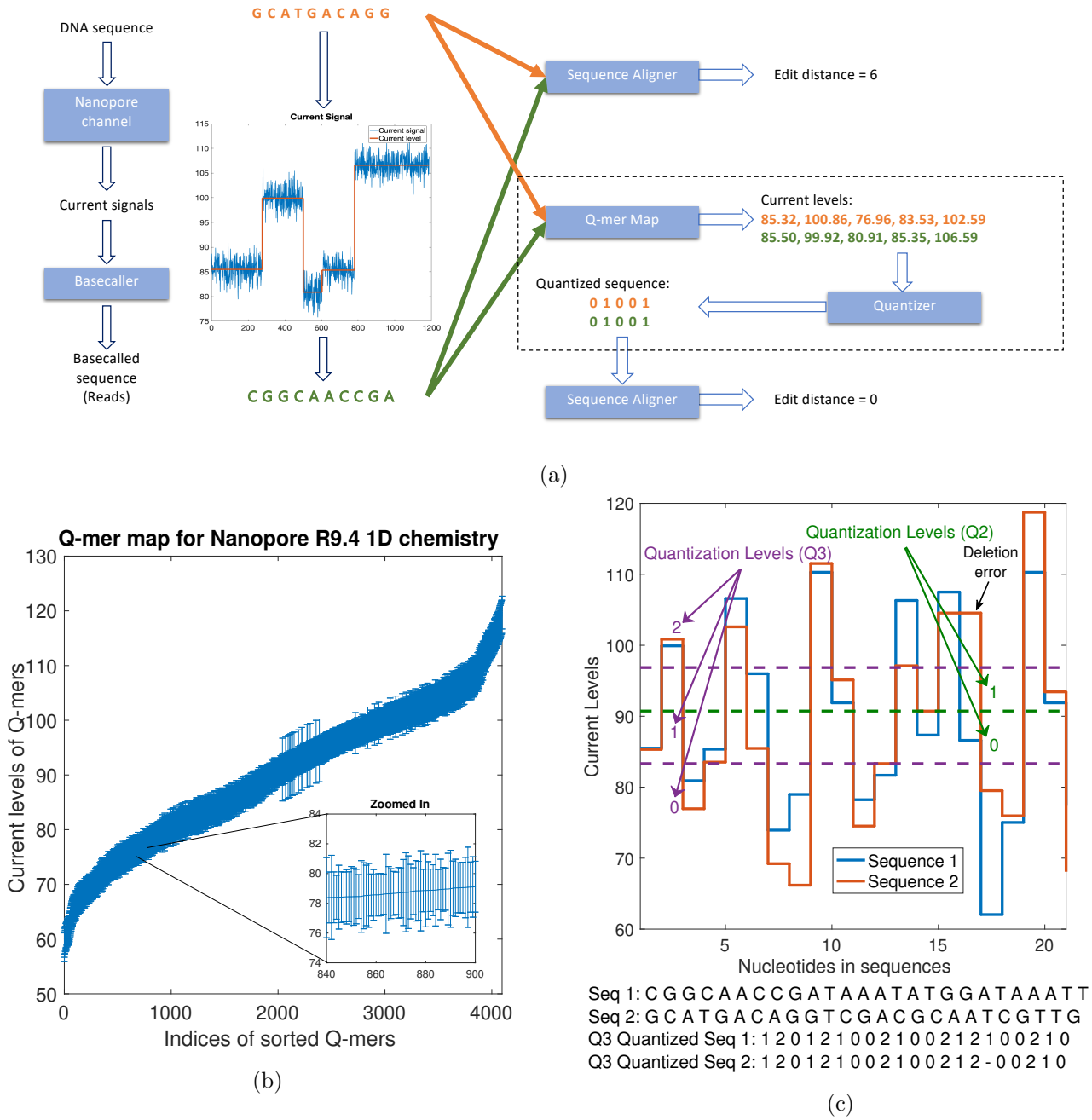


Figure 1: (a) Flowchart for the read alignment process and our approach (in the dotted box). (b) Q-mer Table for Nanopore R9.4 1D Flow cell (for $Q = 6$). (c) An example showing the two different sequences with similar current levels (ambiguous condition for the basecaller).

current level, c, is obtained from the DNA sequence (s). Each current level in the sequence, c, is then quantized to two levels in case of a binary quantizer (Q2), and to three levels, in case of a ternary quantizer (Q3). The sequence thus obtained by quantizing the current level sequence (c) is called the quantized sequence s^Q. Intutively, s^Q captures the list of all the "equivalent" nucleobase sequences

that could have been inferred by the base-caller upon the observation of the current signal variations in the nanopore.

1.2 Datasets

Datasets for DNA-Seq alignments: We use three datasets for DNA read-to-genome and read-to-read alignment: (1) MinION sequencing of K. Pneumoniae DNA R9.4 1D flow cell [18], (2) MinION sequencing of E. Coli DNA using R9 2D fow cell [19], and (3) MinION sequencing of E. Coli DNA R9.4 1D flow cell [20].

Datasets for RNA-Seq alignments: The experiments are based on the RNA reads from MinION sequencing of human cDNA using R9.4 1D flow cell [21], and from mouse using R9.4 2D flow cell [22] (SRA access No. SRR5286961). Compared to DNA-Seq datasets, RNA-Seq datasets carried out using nanopore sequencing are relatively rare. We select these datasets because they have relatively complete reference transcriptome (e.g. for human there are 200,401 annotated transcripts [23] and for mouse 46,415 annotated transcripts from UCSC Genome Browser [24]), as well as corresponding Q-mer map models [25] necessary for quantization.

2 DNA Read-to-Genome Alignment

The alignment of DNA reads to the genome is a task with wide-ranging applicability in sequencing experiements. It is a required step in variant calling pipelines [26], in particular structural variant calling can benefit significantly from long reads offered by the nanopore sequencing platform [9]. It is also useful in calling variants in long segemental duplications [10], where long duplications neccesiate long reads to resolve the repeat ambiguity. Another application for DNA read-to-genome alignment appears in reference matching - for example, in metagenomics, in estimating which reference species is present in the sample.

Methodology and performance metrics: In this section, we study the ability of QAlign to align DNA reads to a reference genome. In each experiment, we take the reads and align them to the genome both using QAlign and Minimap2. There are two different types of performance metrics: (1) Coarse performance is measured by the fraction of reads that have been well-aligned by the algorithm. A read is said to be well-aligned if at least 90% of the read is aligned to the genome. (2) Fine-grained performance is measured in our experiments by two metrics: the normalized alignment length and the

normalized edit-distance between the read and the genome region that the read aligns to. We note that we do not use the particular alignment returned by the different aligners since this is not directly comparable. The normalized-edit distance is defined as the ratio of the edit distance between the genomic region aligned by the read and the entire read to the length of the read - thus normalized edit distance of 0 corresponds to perfect alignment, and an upper bound is 1 if the read does not return an alignment. We note that the normalized edit distance between a random sequence of genome and a given read of the same length is 0.48 (Refer to Figure 1 of Supplementary material for the details on the normalized edit distance between random DNA sequences).

2.1 Results

The results are illustrated in the Figure 2. At a coarse level, QAlign is shown to significantly improve the fraction of well-aligned reads - in particular, in the K.Pneumoniae R9.4 dataset, this metric improves to 90.9% from 76.2%. In the E. Coli R9.4 1D dataset, it improves to 88.5% from 79.2%, and in the E. Coli R9 2D dataset, the numerics improves to 91.8% from 82.6% (see Figure 2a).

The results in Figure 2b-2c compares the quality of the alignments using Minimap2 and QAlign for the K. Pneumoniae dataset (plots for other datasets are available in Figure 3-7 Supplementary material).

For the fine-grained performance evaluation, Figure 2b compares the normalized edit distance for QAlign and Minimap2. The normalized edit distance is the edit distance between the entire read and the aligned section on the genome normalized with the length of the read. Intuitively, it gives a measure of how close are the two sequences. Therefore, the smaller the normalized edit distance, better is the alignment. In addition, the normalized edit distance for the reads that have normalized edit distance more than 0.48 is set to 1. Therefore, the figure represents only those alignments that are better than alignment of random DNA sequences. Figure 2b shows that the normalized edit distance is smaller for the alignments in Q2 than the alignments in ACGT. It is important to note here that the edit distance is computed on the nucleobase sequences in both the methods. In case of Q2, the information of the location of the alignment on the genome is leveraged from the alignment between the quantized read and the quantized current level representation of genome, but the edit distance is computed between the entire nucleobase read sequence and the aligned section on the nucleobase genome sequence. A detailed discussion on the metric is provided in the method section. The better alignment in Q2 is also

evident from the slope of the regression line in Figure 2b. It shows that on average the alignment in Q2 has 15.28% improvement in terms of the normalized edit distance than the alignments in ACGT.

The red circles in the plot of Figure 2b represent the reads that are well aligned in both ACGT and in Q2 and at nearly the same location on the genome. Recall that the reads are said to be well aligned if at least 90% of the read is aligned onto the genome. The blue crosses represent well aligned reads in both Q2 and ACGT but at different location on the genome. The black astrisks are the reads that are well aligned in Q2 only, *i.e.*, in ACGT, these reads are either not aligned onto genome at all or the aligned length on the genome is less than 90% of the read length, whereas in Q2, the aligned length on genome is greater than 90% of the read length. The green diamonds are the reads that are well aligned in ACGT only. The pink square points are the reads that are not well aligned in either Q2 or ACGT, *i.e.*, either the reads are not aligned in both Q2 and ACGT or the aligned length is less than 90% in both ACGT and Q2. For each read, there could be multiple alignments on the genome because of the repeats in the genome, but we consider the alignment that has the minimum edit distance for the evaluation in these plots. It is evident from the plot that for the common alignments in ACGT and in Q2, *i.e.*, the red points and the blue points, the alignments in Q2 have smaller distance than the alignments in ACGT, which suggests that the alignment location on the genome determined with Q2 is more accurate than the location of alignment on genome determined with ACGT.

The results for another fine metric are shown in Figure 2c which compares the normalized align-length on genome in Q2 to the normalized align-length on genome in ACGT. The normalized align-length is the ratio of the length of the section on genome where a read aligns to the length of the read. There are several reads that are well aligned in Q2 only (the black asterisks), and the normalized align-length is close to 1 in Q2 but much less than 1 in ACGT, therefore representing several non-contiguous alignments in ACGT that are captured in Q2 ($\approx 15.0\%$). The normalized edit distance for such reads in Q2 is much less than the normalized edit distance in ACGT. Similar results are observed across different datasets as evident from the slope of the regression line for normalized edit distance comparison between Q2 and ACGT shown in Figure 2d.

3 Read-to-Read alignment

Alignment of genomic reads to other reads is a basic primitive useful in many settings. For example, this is a first step in many overlap-layout-consensus (OLC) assemblers [4]. A key challenge in read-

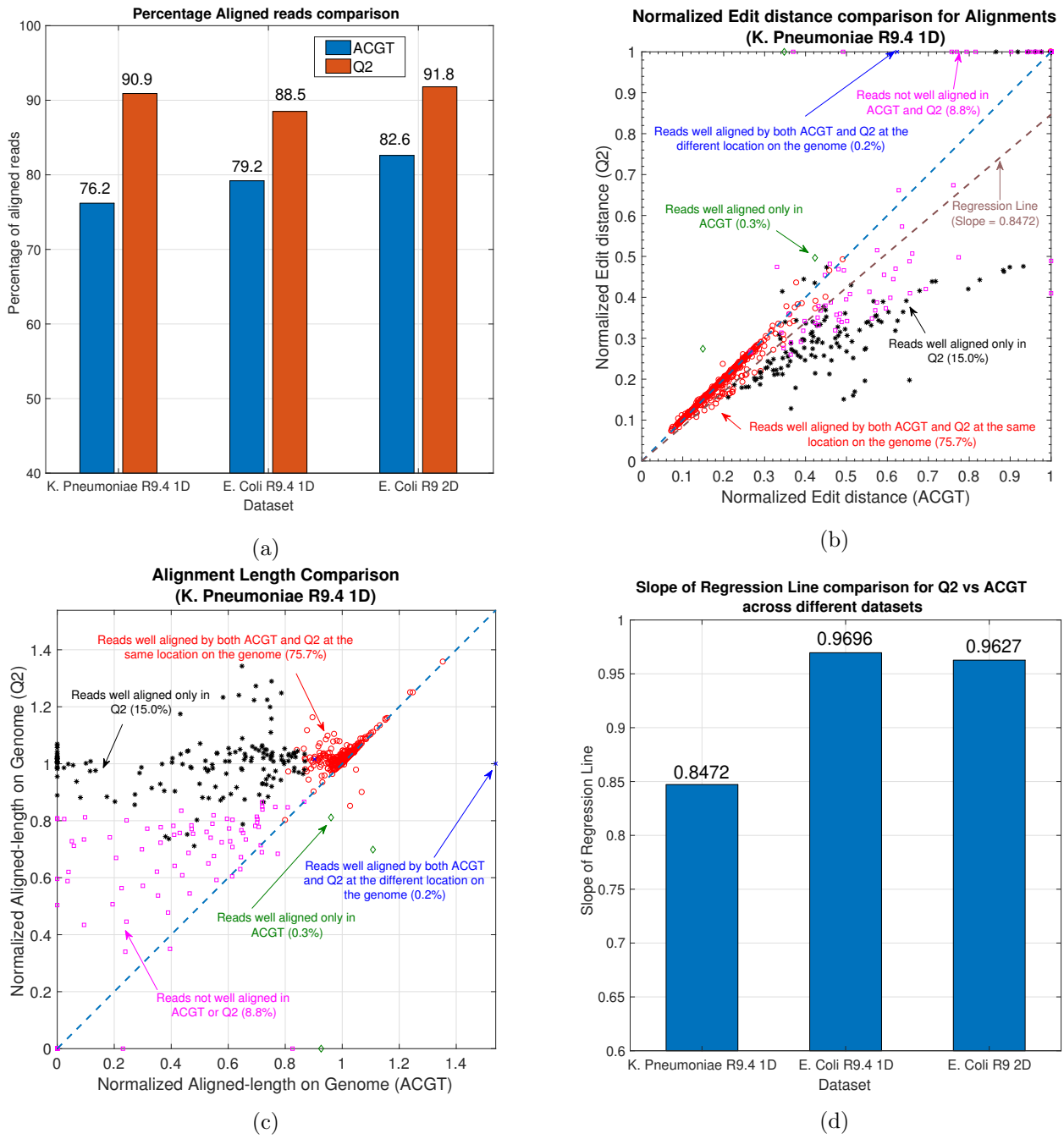


Figure 2: **Nanopore long DNA reads alignment onto Genome.** (a) Percentage of reads aligning onto genome for 1000 randomly chosen reads for each dataset. (b) Comparison of normalized edit distance for K. Pneumoniae R9.4 1D reads data. (c) Comparison of normalized align-length on genome for K. Pneumoniae R9.4 1D reads data. (d) Slope of the regression line (for normalized edit distance comparison plot for Q2 vs ACGT) for each dataset. The slope of the regression line shows the average gain in the normalized edit distance.

to-read assembly is the increased error rate that the aligner has to deal with. For example, if two reads R_1, R_2 are sampled from the same region of the genome, each may be within 15% edit-distance

of the reference genome (assuming a 15% error-rate), however, the edit distance between R_1 and R_2 can be up to 30% leading to an effective doubling of the error-rate. Long-reads hold the promise of fully-automated assembly but is currently feasible only when for bacterial genomes [27]. For complex mammalian genomes, long repeats fragment assembly and more accurate alignment can help alleviate this problem.

Methodology and Performance Metrics: In this section, we study the ability of QAlign to align DNA reads to other DNA reads. In each experiment, the DNA reads are aligned to other DNA reads using both QAlign and Minimap2. The read datasets from the same organisms are used for the experiments that we have used for read-to-genome alignment. We say that the algorithm is able to align the pair of reads (*i.e.*, the pair of reads has an overlap), if the 'Mapped Region' of overlap by the algorithm is atleast 90% of the estimated overlap between the same pair of reads in the ground truth (Please refer to the Method section for more details). Since we do not have a ground truth specifying what pairs of reads are aligning with each other with a head-to-tail alignment, in our evaluation, we use the read-to-genome alignment to leverage this information to estimate the ground truth for the read-to-read alignment. The overlap length between a pair of overlapping reads can be determined from the location of the alignments of the reads on the genome. We define the overlap quality as the ratio of the size of the intersection between the overlap region estimate (from the alignment algorithm) and the true overlap (from the ground truth), to the size of the union between the overlap region estimate and the true overlap. Thus, the overlap quality is equal to 1 if and only if the overlap region estimate is the same as true overlap. The expected overlap quality is the area under the complementary cumulative distribution curve. This expected value is also referred to as average overlap quality in this paper. Moreover, an ensemble model chooses the best overlap between a given pair of reads using the information of the overlap quality in ACGT and in Q2.

3.1 Results

The results for read-to-read alignment are illustrated in the Figure 3a-3d. Figure 3a shows overlap quality comparison for the alignments in Q2 versus alignments in ACGT for the K. Pneumoniae dataset (the overlap quality comparison for the alignments in Q3 versus alignments in ACGT and with other datasets are provided in Figure 8 and 9 in Supplementary material). The blue circles in the figure represents the overlaps that are aligned in both QAlign and Minimap2. The black asterisk (along

the line $x = 0$) represents the overlaps that are aligned only in QAlign and not aligned in Minimap2, whereas the green diamonds (along the line $y = 0$) represents the overlaps that are aligned in Minimap2 and not aligned in QAlign. Recall that an overlap between a pair of reads is said to be aligned by the algorithm if the Mapped region by the algorithm is atleast 90% of the estimated region of overlap between the pair of reads in the ground truth. In Figure 3a, we observe that 4.62% of the read overlaps have overlap quality more than 90% in QAlign but not in Minimap2 whereas the opposite holds true in only 1.0% of the read overlaps. Thus QAlign gives a net performance improvement of 3.62% over Minimap2. Moreover, the read overlaps that are aligned only in Q2 is 7.3%, whereas the read overlaps that are aligned only in ACGT is 2.3%. Therefore, QAlign demonstrates a net gain of 5.0% in terms of the number of reads aligned by the algorithm. In addition to that, the slope of the regression line in the figure is 1.0089, therefore illustrating better overlap quality with QAlign than Minimap2 in case of common alignments by both the algorithms.

Figure 3b shows the Precision and Recall for different methods used (namely, ACGT, Q2, and Q3) to find the alignments between the overlapping reads across different datasets. It is evident from the plot that Q2 provides higher recall than ACGT at the cost of a bit lower precision. Q3, on the other hand, shows better recall than ACGT at the same precision.

A simple comparison between QAlign and Minimap2 is found in Figure 3c where the fraction of reads which have overlap quality greater than x is plotted for the two aligners - the performance gain is seen to hold across a wide range of threshold values x. The area under the curve (which equals the average overlap quality) is computed for ACGT, Q2, and Q3 across all the three datasets and is demonstrated in Figure 3d. The gain in the average overlap quality is observed using QAlign across all the three datasets as evident from this plot. As evident from the plot, there is a gain of 9.2% in K. Pneumoniae dataset, when we compute it as the ratio of the average overlap quality of Q2 to average overlap quality of ACGT. Similarly, there is a gain of 2.5% and 10.8% in the average overlap quality for the E. Coli R9.4 1D and E. Coli R9 2D dataset, respectively.

4 Read-to-transcriptome alignment

RNA-Seq is a popular sequencing technology with emerging applications including single-cell RNA-Seq [28]. While short high-throughput reads may suffice to assess rough gene expression estimates, isoform level analysis is better facilitated by long nanopore reads that can straddle several exons simultaneously

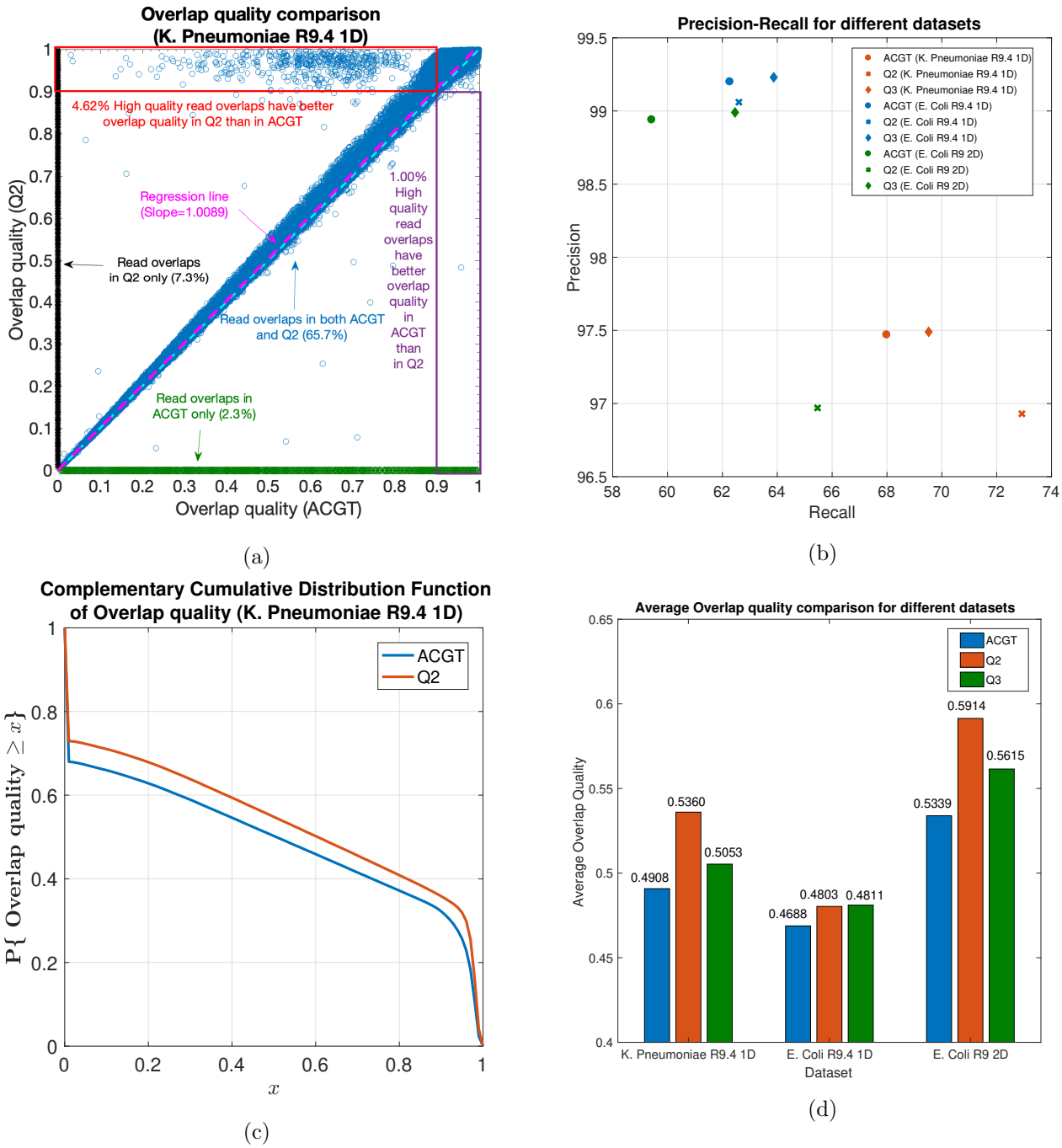


Figure 3: **Nanopore long DNA Read-to-read alignment.** (a) Comparison of overlap quality for K. Pneumoniae R9.4 1D reads dataset (Q2 vs ACGT). (b) Comparison of the Precision and Recall for different datasets. (c) CCDF of Overlap quality for K. Pneumoniae R9.4 1D reads dataset. (d) Comparison of the average overlap quality for different datasets.

[29]. Here we perform the alignment of cDNA reads (complementary DNA reads extracted from reverse transcription of RNA) onto a reference transcriptome.

Methodology and Performance Metrics: The methodology and the performance metrics are similar to

the DNA read-to-genome alignment experiments: we compare the fraction of well-aligned reads as well as the normalized alignment length and the normalized edit distance for each alignment algorithm.

4.1 Results

The results for read-to-transcriptome alignment are illustrated in Figure 4a-4d. At a coarse level, QAlign improves the fraction of the well-aligned reads significantly. For the Human R9.4 1D dataset, the metric improves to 86.3% from 50.8%, and for the Mouse R9.4 2D dataset, it improves to 95.3% from 82.3%, as shown in Figure 4a. Moreover, the normalized edit distance is set to 1 for the reads that have normalized edit distance greater than 0.48. Therefore, the figure represents the alignments that are not "equivalent" to the alignment of random DNA sequences.

Using QAlign not only increases the fraction of well aligned reads, but also the quality. To understand this, in Figure 4c, the lengths of aligned chunks are compared between ACGT and Q2 domain. Most of the reads gets larger aligned chunks using Q2 quantization. In addition, Figure 4b compares the normalized edit distance for Human R9.4 1D dataset, further demonstrates the gain of quantized alignment. Specifically, Q2 is able to align 36% more reads with 41.15% better quality than ACGT in read-to-transcriptome alignments. Additional comparisons between ACGT and Q2 for mouse data (Figure 4d) and between ACGT and Q3 for human and mouse datasets (refer to Figure 11 in Supplementary material) show similar trend.

To understand Figure 4b further, we group the dots (i.e. alignments) by coloring them based on whether the corresponding read is aligned to the same location by both ACGT and Q2 (red), or it is aligned to different locations by ACGT and Q2 (blue), or it is aligned by only Q2 (black), or it is aligned by only ACGT (green). Note there are 36% reads (black) able to be aligned by only Q2, and majority of them are well aligned as their normalized edit distance is below 0.48. In addition, there are 50.4% reads (red or blue) able to be aligned by both Q2 and ACGT, many of which lie below 45 degree line, indicating Q2 alignments overall show better alignment quality (lower normalized edit distance score) than ACGT. In particular, the regression line qualitatively indicates that among the red or blue alignments, 41.15% of them are better aligned in Q2.

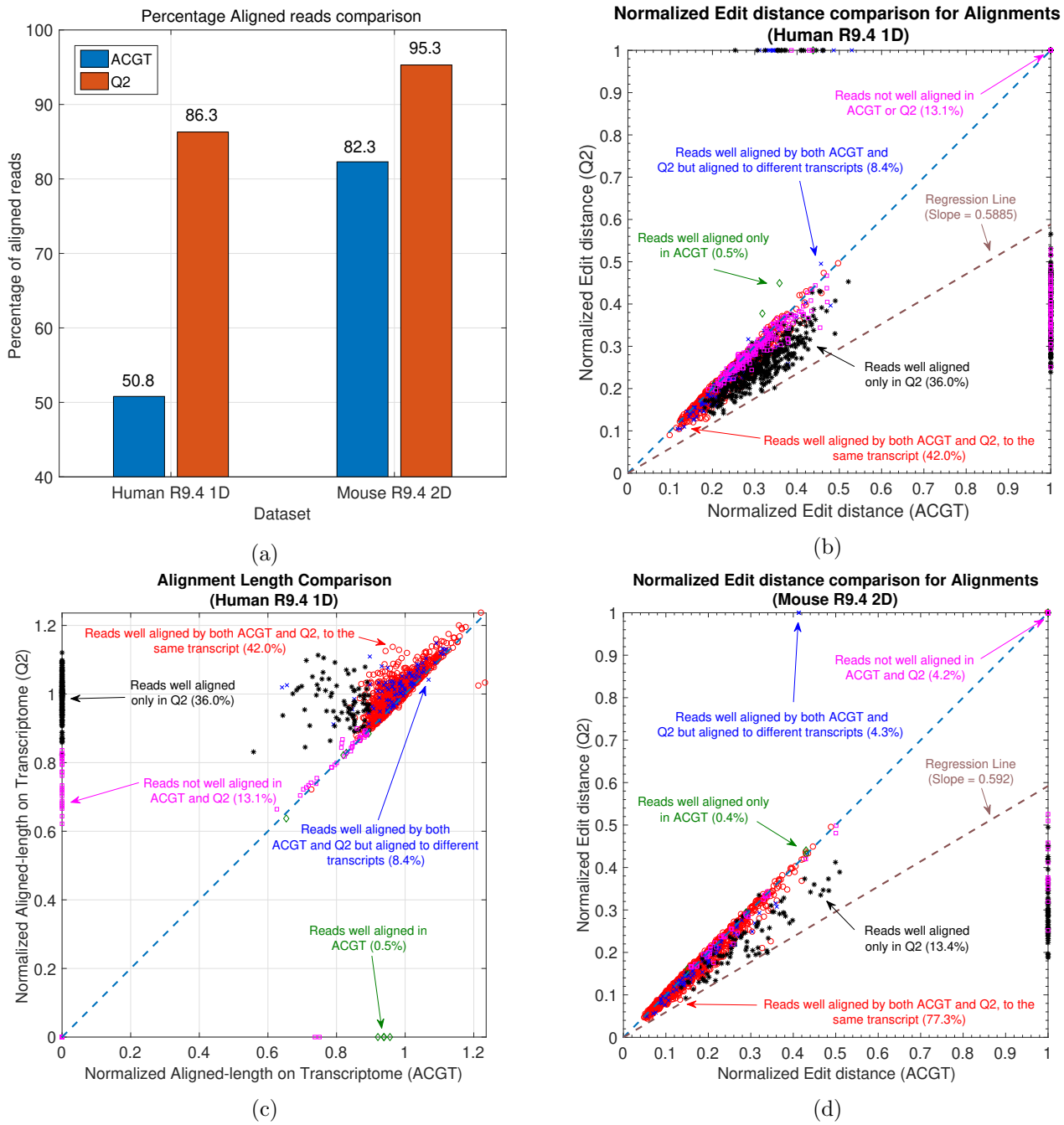


Figure 4: **Nanopore long RNA read to transcriptome alignment.** (a) Percentage of reads aligning onto transcriptome for 2000 randomly chosen reads for each dataset. (b) Comparison of normalized edit distance for Human R9.4 1D dataset. (c) Comparison of normalized align-lengths of the aligned sections on the transcriptome for Human R9.4 1D dataset. (d) Comparison of normalized edit distance for Mouse R9.4 2D dataset.

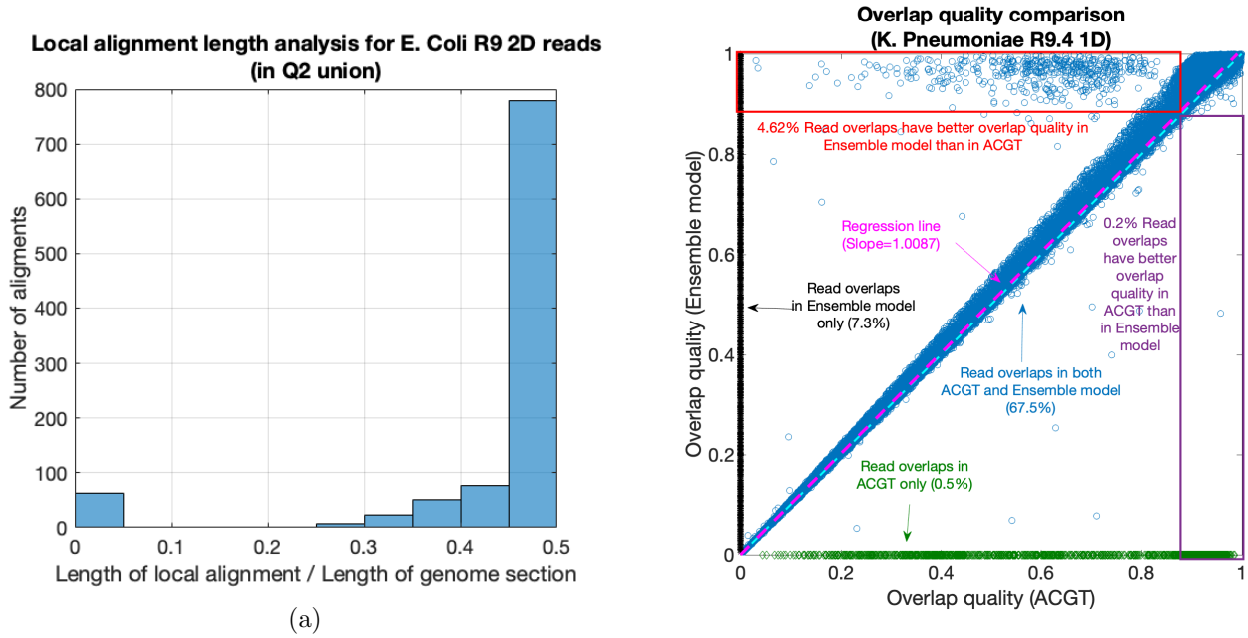
5 Discussion

Read-to-genome alignment for the 2D reads: An alternate pipeline for the alignment of the 2D consensus reads is using the corresponding 1D reads from both the template and the complement

strand and estimate the alignment location of the 2D consensus read onto the genome using the QAlign algorithm. This can be done by exploiting the nanopore physics with the 1D reads from each strands. The experiment pipeline is explained in detailed in the method section. The results in Figure 5a illustrates that the alignment length on the genome, which is the output of the QAlign algorithm is atleast twice the length aligned by the local alignment algorithm for the same read at no additional cost of the normalized edit distance (as shown in Figure 4 and 5 in the Supplementary paper).

Ensemble model for read-to-read alignment: In case of the read-to-read alignment, there are several read overlaps that have better overlap quality in QAlign than in ACGT and several other read overlaps that have better overlap quality in ACGT than in QAlign. Therefore, in order to cover all the overlaps between the reads that are estimated by the ground truth, an ensemble model is used, and the overlap quality comparison for the ensemble model against the ACGT read overlaps is shown in Figure 5b.

Read-to-transcriptome alignment: The significant gains of using quantization in read to transcriptome alignments open up new possibilities, which we discuss here and leave as future exploration. One important application of long RNA read to transcriptome alignment is to detect novel variants in expressed coding regions. The variants (especially Single-nucleotide polymorphism, or SNP) in expressed coding regions may affect protein generations and consequently cause severe diseases [30]. Utilizing the RNA-Seq to directly detects such kind of variants are promising economic methods, but they are not thoroughly explored due to several challenges. One challenge is there are many repeats among RNA transcripts due to alternative splicing - a gene could have its exons combined to generate several RNA transcripts. As a result, reads can be multiply mapped to these regions. Existing methods including the popular GATK [31] choose to discard these reads, and lose the variant information in multiply mapped reads. Although there are recent efforts to tackle this [3], long RNA read alignment have the potential to fundamentally remedy this. However, it suffers from high error rate, and some of the errors could come from base calling. Quantized long RNA read alignments could bring better alignments (partially due to bypassing the errors accumulated from base calling) and at the same time reduce multiply mapped reads, and therefore could offer better signals for us to call variants especially in repetitive regions.



(a) (b)

Figure 5: (a) Hybrid approach for the 2D reads using the alignments if the 1D reads from each strand of the DNA. (b) Ensemble model for read-to-read alignment for K. Pneumoniae R9.4 1D reads.

6 Methods

6.1 Nanopore Sequencer

The Oxford Nanopore Technology is based on DNA transmigrated through a nanopore, which results in the changes in the ionic currents through the pore [29]. The current changes are caused by the different nucleotides that are partially blocking the pore as the (single-strand) DNA sequence to be measured is migrated through the nanopore. An enzyme slows down the motion of the DNA through the pore, so that the variations in the current signals can be measured accurately [29]. Base calling algorithms [32] are developed to infer the nucleobases from the measured current changes.

Ideally, the DNA is migrated through the nanopore at a constant rate and per time observed current is only affected by a single nucleobase. Consequently, the DNA sequence can be decoded unambiguously with high probability.

In reality, there are several non-idealities due to the physics of the nanopore and the enzyme. (i) *Inter-symbol interference:* Since the nanopore is bigger than a single nucleotide, the observed current at a given time is influenced by multiple (neighbour) bases or Q-mer (where Q could be 4,5, or 6). (ii) *Random dwelling time:* The amount of time spent by each Q-mer of the DNA sequence

in the nanopore may vary. So the rate at which the DNA sequence migrates through the nanopore is a stochastic process. (iii) *Segment insertion and deletion:* There are segments that are repeated as well as segments that migrate through the nanopore without registering a current reading. This results in redundant and missing segments. (iv) *Q-mer map fading:* The measured current level is non deterministic with respect to the corresponding Q-mer (see Fig. 1b). (v) *Noisy samples:* Each observed current level is also subject to a random noise [29].

6.2 Quantization

The nucleotide base sequences are inferred from the observed current in the nanopore, therefore, using a Q-mer map to translate the base-called sequences to the current levels implicitly maintains the list of all "equivalent" basecalled sequences that could be inferred from the observed current levels. These current levels can be quantized to an alphabet of finite size. The quantization has two folds advantage: (i) It relaxes the reliability on the current levels which is advantageous because of the *Q-mer map fading effect* discussed in the previous section. (ii) Quantization to finite level enable us to exploit the state-of-the-art aligners which are based on hash-extension methodology rather than using the dynamic programming techniques. This is advantageous in terms of the computation time as the performance can be improved at no extra cost of the computation time by using the optimized sequence aligners.

Mathematically, the quantization process can be described as follows. Let $\Sigma = \{A, C, G, T\}$ be the alphabet of nucleotides. For a symbol $x \in \Sigma$, $\bar{x}$ is the Watson-Crick complement of x. A string $s = x_1 x_2 \ldots x_n$ over Σ is called a *DNA sequence*, such that the length of the string is $|s| = n$ and its *reverse complement* is $\overline{s} = \overline{x_1 x_2 \ldots x_n} = \overline{x}_n \overline{x}_{n-1} \ldots \overline{x}_1$. Let q be a list of all Q-mers (e.g. Q=6) in the string s, such that the list is sorted by the occurance of the Q-mer in the string s, i.e., let $q = k_1 k_2 \ldots k_{n-Q+1}$ and $k_i = x_i x_{i+1} \ldots x_{i+Q-1}$ for $i = 1, 2, \ldots, n - Q + 1$. Now, we define $f : \Sigma^Q \to \mathbb{R}$ as the Q-mer map, which is a deterministic function that translate each Q-mer to the (mean) current level. The Q-mer map is determined by chemistry of the nanopore flow cell as shown in Figure 1b. Proper Q-mer map needs to be chosen for individual datasets based on their flow cells. Now, let $C = c_1 c_2 \ldots c_{n-Q+1}$ be the sequence of the current levels, such that, $c_i = f(k_i)$ for $i = 1, 2, \ldots, n - Q + 1$. The current sequence $C = c_1 c_2 \ldots c_{n-Q+1}$ can be further quantized into $w = q_1 q_2 \ldots q_{n-Q+1}$ by applying a thresholding function $q_i = g(c_i)$. The quantization can be binary ($q_i \in \{0, 1\}$) or ternary ($q_i \in \{0, 1, 2\}$), as illustrated in Fig. 1c.

We can now use the aligners such as Minimap2 to align the quantized sequences instead of the nucleotide sequences. It is important to note that these aligners inherently performs the alignment of the query sequence (e.g. s_1) to the reference sequence (e.g. s_2) and also aligns the *reverse complement* ($\overline{s}_1$) of the query sequence to the reference sequence (s_2). For the corresponding quantized sequences, aligners need to align the query sequence (e.g. w_1) and its reverse complementary (e.g. $\overline{w}_1$) *explicitly* to the reference (e.g. w_2). Therefore, a similar setting is also required for the quantized sequence of the *reverse complement* sequence $\overline{s}$, denoted by $\overline{w}$. The same pipeline as described above is used to obtain the quantized sequence for $\overline{s}$, with a minor change of replacing the neucleotide sequence s with $\overline{s}$.

The performance of such an aligner can be evaluated by comparing the alignments of the nucleotide sequences s_1 and s_2 to the alignments of their quantized sequences w_1 and w_2, respectively, using appropriate performance evaluation metrics.

6.3 DNA Read-to-Genome Alignment

In this experiment, the alignments of a read r to the reference genome G is determined using Minimap2. These alignments can also be determined with the quantized read r^{Q_2} and the quantized genome G^{Q_2} using Minimap2. The Minimap2 command for the read to reference alignment with nucleotide sequence is:

```
minimap2 -c -k 9 reference.fasta reads.fasta
```

The Minimap2 command for the alignment with the quantized sequences is:

```
minimap2 -c -k 23 --for-only reference_q2.fasta reads_q2.fasta
minimap2 -c -k 23 --for-only reference_q2.fasta rc_reads_q2.fasta
```

The two separate commands for the alignment with the quantized sequences are for the alignment of the quantized *template* strand of the reads to the quantized genome and the alignment of the quantized *reverse complement* strand of the reads to the quantized reference genome, respectively. We can then aggregate the results from both the outputs to determine the best alignment for each read. The different seed lengths (minimizer length `k` in Minimap2) are used for the alignments with nucleotide

sequences versus the alignments with the quantized sequences. This is to ensure that the computation time is the same in both regimes. The detail analysis of the computation time as we vary the seed length is provided in Table 1 of the Supplementary material. The comparison can now be made for the quality of the alignments with the two methods given the computation time for both the methods is the same.

The output from Minimap2 is the location of the alignment of read r onto the genome G. As shown in Fig. 6a, *Read 1* aligns at location i_1 through j_1 on the genome. Similarly, we have the location of the alignment of the quantized read r^{Q_2} onto the quantized genome G^{Q_2} as from i_1^q to j_1^q. In order to compare the quality of the alignments with the two methods, we define a performance metric called *Normalized edit distance* as folows:

$$\textit{Normalized edit distance for nucleotide sequence} = \frac{\textit{edit_distance}\{r; G[i_1 : j_1]\}}{\textit{length}(r)}$$

$$\textit{Normalized edit distance for Quantized sequence} = \frac{\textit{edit_distance}\{r; G[i_1^q : j_1^q]\}}{\textit{length}(r)}$$

As evident from the metric definition above, we are only leveraging the information about the alignment location with the quantized sequences and then computing the edit distance of the nucleotide read r with the section on the nucleotide genome G, where the genome section is found using the alignment of the quantized reads and quantized genome.

Read-to-genome alignment for 2D reads

The 2D reads (e.g. r) are consensus calling using the 1D reads from both the forward strand (e.g. r_t) and the reverse strand (e.g. r_c). For the read-to-genome alignment algorithm of the 2D reads using QAlign, the experiment pipeline has been modified as described in this section so that the error profile introduced in the sequencing of the 1D reads can be mitigated using QAlign: The quantized reads from both the template strand (e.g. r_t^q) and the reverse strand strand (e.g. r_c^q) are aligned to the quantized genome (e.g. G^q) individually using Minimap2. The union of the two alignments[1] is considered as the output of the QAlign algorithm, i.e., the algorithm provides the genome section based on the alignments of the 1D reads where the 2D consensus will align to; In case there is no overlap in the alignments of the reads from the individual strands, both the genome sections are given as the output of the QAlign algorithm since the 2D consensus read might align to either of these sections.

[1]For example, alignment regions are [0, 2] and [1, 3], and union is [0, 3].

Since QAlign yields the genome section as the union of the two alignments, it is much larger (nearly twice) than the read length. Therefore, the genome section is further refined by local alignment of the consensus read onto the section. The performance evaluation of QAlign can be done by computing the normalized edit distance between the 2D consensus read and the refined genome section.

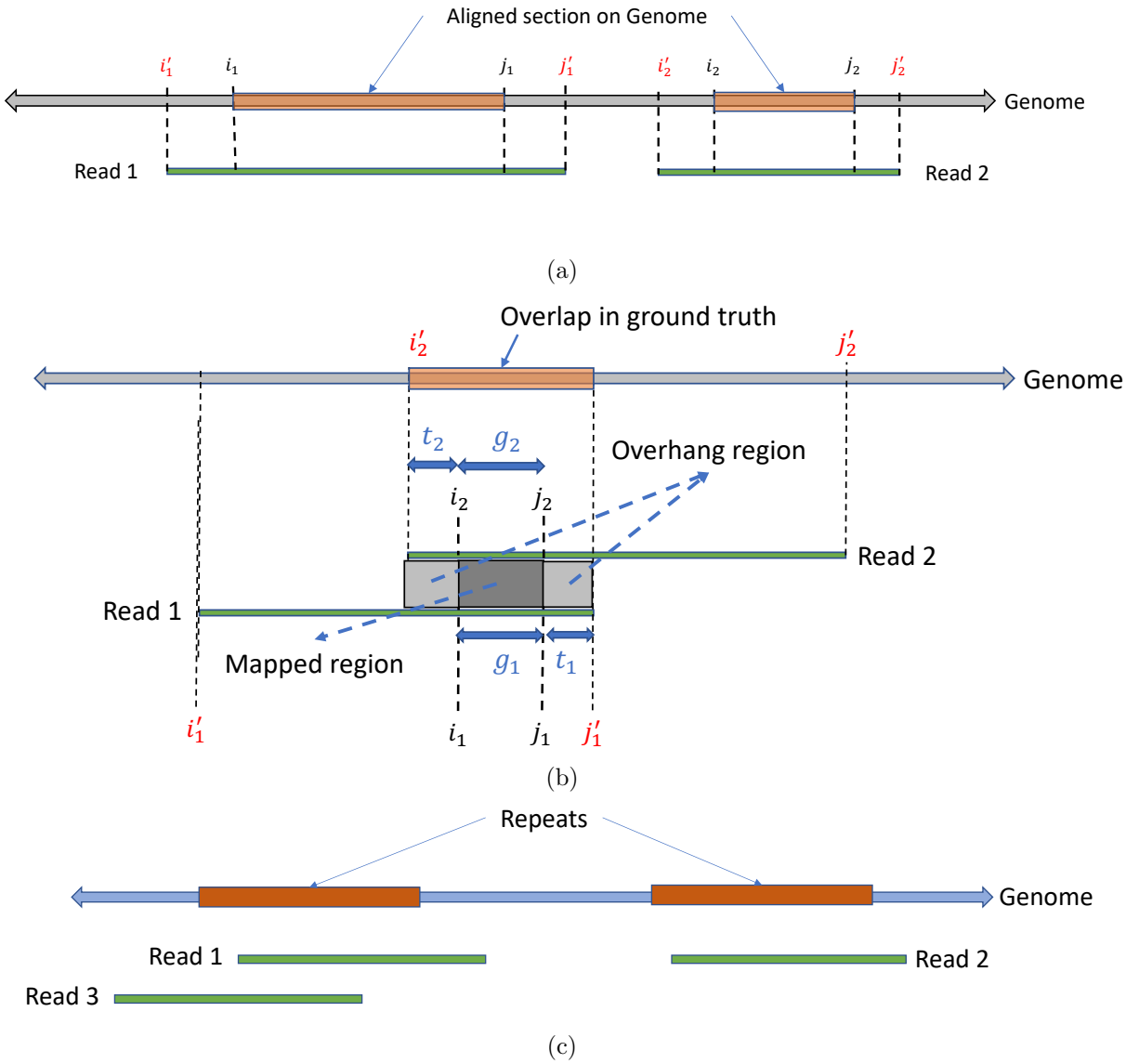


Figure 6: (a) An example of read-to-genome alignment. (b) An example of read-to-read alignment. (c) An example for Head-to-Tail alignments between reads.

6.4 Read-to-Read Alignment

The genome assembly requires the information about the read overlaps. These overlaps between the reads can be determined by read-to-read alignment settings of the aligners. In this experiment, the

alignments between the nucleotide reads r_1 and r_2 are obtained using Minimap2. Then the alignments between the quantized reads r_1^q and r_2^q are also determined. The Minimap2 command used for the alignment with nucleotide sequences is:

```
minimap2 -cx ava-ont -k 10 reads.fasta reads.fasta
```

The Minimap2 command used for the alignments with the quantized sequences is:

```
minimap2 -cx ava-ont -k 20 --for-only reads_q2.fasta reads_q2.fasta
minimap2 -cx ava-ont -k 20 --for-only reads_q2.fasta rc_reads_q2.fasta
```

The two separate commands used for the alignment with the quantized sequence also aims to *explicitly* account for the *template* and *reverse complement* strands of the reads, and the best alignment between a pair of reads is determined by the maximum overlap length. Similar to Read-to-Genome alignment, different seed lengths (minimizer lengths '`k`' in Minimap2) are also used so that the computation time in both regimes is maintained to be the same. The details on the change in computation time as we change the minimizer length (`k`) is provided in Table 2 of the Supplementary material.

Estimation of the ground truth: The ground truth for the read-to-read alignment is unknown. One way to judge the quality is to compute the normalized edit distance of alignment overlap. However, this is not only expensive but also suffers from false alignments between reads from repeated regions. Instead, we leverage the read-to-genome alignments to build the ground truth for read-to-read alignment. Specifically, all of the reads are firstly aligned to the genome using the nucleotide and the quantized sequences. A section G_1 on genome G (i.e., $G_1 = G[1 : 1000000]$), is chosen in which we are interested to find all the read-to-read alignments. Assume there are n_1 number of reads aligned to G_1 in ACGT with normalized edit distance less than 0.48, and n_2 number of reads aligned to G_1 in Q2 with normalized edit distance less than 0.48. Now, we randomly choose some n reads from a union of n_1 and n_2 reads such that $n \approx \frac{Nc}{L}$, where c is the required coverage depth, N is the length of the genome section G_1 ($= 1000000$ for the experiments), L is the average length of the reads in $n_1 \cup n_2$. The union of the alignments from ACGT and Q2 is taken so that the missed alignments or non-contiguous alignments in ACGT with a normalized edit distance greater than 0.48 but with normalized edit distance less than 0.48 in Q2 are taken into account, and vice-versa.

The ground truth is estimated as described below: Consider the alignment of *Read 1* (r_1) and *Read 2* (r_2) onto genome as shown in Fig. 6b, where the alignment location of the reads on the genome are (i'_1, j'_1) and (i'_2, j'_2), respectively. We say that the reads are overlapping in the ground truth if there is an overlap of at least 100 bases, i.e., $l \geq 100$, where $l = \min(j'_2 - i'_1, j'_1 - i'_2)$ In case where the overlap length between a pair of reads using the two methods are different: read-to-genome alignment in ACGT (l^{ACGT}) and read-to-genome alignment in Q2 (l^{Q2}), the larger of the two is chosen (e.g. $l = \max(l^{ACGT}, l^{Q2})$).

The read-to-read alignment will consider two reads to have an overlap if the length of the 'Mapped Region' is at least 90% of the 'Mapped Region' ∪ 'Overhang Region' (Fig. 6b, i.e., $g_1 \geq 0.9(g_1+t_1+t_2)$ and $g_2 \geq 0.9(g_2 + t_1 + t_2)$, or equivalently $t_1 + t_2 \leq 0.1(\min(g_1, g_2))$. For quality evaluation, we define another metric called the 'overlap quality' as:

$$\textit{Overlap quality} = \frac{(g_1 - d_1) + (g_2 - d_2)}{2l + d_1 + d_2}$$

where[2] $d_1 = \max(0, g_1 + t_1 - l)$ and $d_2 = \max(0, g_2 + t_2 - l)$. It measures how well are the reads aligned (in terms of the overlap length of the reads) compared to the alignment in the ground truth. A larger value of the overlap quality represents better alignment. Consider X to be a random variable that represents the quality of the overlaps, then the complementary cumulative distribution function is given by $\mathbb{P}\{X > x\}$. and the average overlap quality is the expected value of the overlap quality, $\mathbb{E}[X] = \int \mathbb{P}\{X > x\}dx$, which is the area under the complementary CDF of X.

It is possible that two reads will be falsely aligned especially when they are from repetitive regions. To remedy this, we only consider head-to-tail alignment between reads. For example in Fig. 6c, three reads *Read 1*, *Read2* and *Read 3* have been sequenced where *Read 1* and *Read 2* are from repetitive regions. Consequently after read-to-read alignment, there will be an overlap between *Read 1* and *Read 2* that can be filtered out since it is not a head-to-tail alignment. However, there will also be another false overlap between *Read 3* and *Read 2*, which will not be removed as it satisfies the head-to-tail alignment condition. In order to further reduce the number of false positives of read-to-read alignments, the (approximate) normalized edit distance between the quantized sequences provided by Minimap2 is used for further filtering (refer to Figure 2 of the Supplementary material for the threshold values).

[2]Empirically d_1 and d_2 tend to be simply zero.

The read-to-read alignment results can be further improved by implementing an Ensemble model that captures the best alignment (e.g. longer length of 'Mapped Region') from both QAlign and alignment with nucleotide sequences. Moreover, the ensemble method also incorporates the alignments that are complementary in either method.

6.5 RNA Read-to-Transcriptome Alignment

The pipeline for RNA read-to-transcriptome alignment is analogous to read-to-genome alignment. It is not a spliced alignment of the reads onto genome but the entire read is aligned to the transcriptome. The commands for Minimap2 are the same where the reference genome is replaced with the reference transcriptome. Since the ground truth is unknown for the alignments, the normalized edit distance is used as the metric to compare the quality of the alignments using the nucleotide sequences and the alignments using the quantized sequences.

Data Access

The dataset for DNA-Seq alignments are:

- MinION sequencing of K. Pneumoniae DNA using R9.4 1D flow cell (`https://figshare.com/articles/Trimmed_ONT_reads/5170852`)
- MinION sequencing of E. Coli K12 DNA using R9 2D flow cell (`http://lab.loman.net/2016/07/30/nanopore-r9-data-release/`)
- MinION sequencing of E. Coli UTI89 DNA using R9.4 1D flow cell (`https://www.ncbi.nlm.nih.gov/sra/SRX4387499%5Baccn`)

The dataset for RNA-Seq alignments are:

- MinION sequencing of Human cDNA using R9.4 1D flow cell (`https://github.com/nanopore-wgs-consortium/NA12878`)
- MionION sequencing of house Mouse using R9.4 2D flow cell (`https://www.ncbi.nlm.nih.gov/sra/SRR5286961`)

Acknowledgements

SD and DJ were supported in part by NSF grant 1705077. SM and SK were supported in part by NIH grant 1R01HG008164 and NSF grants 1651236 and 1703403.

Author's contributions

DJ, SK, and SD conceived the original idea and developed the project. DJ led the development of the software tool. DJ and SM performed the analysis on the various datasets. All the authors wrote the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

[1] DePristo, M. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491–498 (2011).

[2] Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

[3] Mao, S., Mohajer, S., Ramachandran, K., Tse, D. & Kannan, S. absnp: Rna-seq snp calling in repetitive regions via abundance estimation (2017). URL `http://drops.dagstuhl.de/opus/volltexte/2017/7658/`.

[4] Pevzner, P. A., Tang, H. & Waterman, M. S. An eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences* **98**, 9748–9753 (2001). URL `https://doi.org/10.1073/pnas.171285098`.

[5] Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nature Protocols* **8**, 1494–1512 (2013). URL `https://doi.org/10.1038/nprot.2013.084`.

[6] Mikheyev, A. S. & Tin, M. M. Y. A first look at the oxford nanopore MinION sequencer. *Molecular Ecology Resources* **14**, 1097–1102 (2014). URL https://doi.org/10.1111/1755-0998.12324.

[7] Križanović, K., Echchiki, A., Roux, J. & Šikić, M. Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics* **34**, 748–754 (2017). URL https://doi.org/10.1093/bioinformatics/btx668.

[8] Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**, 338–345 (2018). URL https://doi.org/10.1038/nbt.4060.

[9] Stancu, M. C. *et al.* Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature Communications* **8** (2017). URL https://doi.org/10.1038/s41467-017-01343-4.

[10] Chaisson, M. J., Mukherjee, S., Kannan, S. & Eichler, E. E. Resolving multicopy duplications de novo using polyploid phasing. In *Lecture Notes in Computer Science*, 117–133 (Springer International Publishing, 2017). URL https://doi.org/10.1007/978-3-319-56970-3_8.

[11] Tran NH, Zhang X, Xin L, Shan B & Li M De novo peptide sequencing by deep learning. **114 31**, 8247–8252 (Proceedings of the National Academy of Sciences of the United States of America, 2017). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5547637/.

[12] Laszlo, A. H. *et al.* Decoding long nanopore sequencing reads of natural DNA. *Nature Biotechnology* **32**, 829–833 (2014). URL https://doi.org/10.1038/nbt.2950.

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

[14] Mao, W., Diggavi, S. & Kannan, S. Models and information-theoretic bounds for nanopore sequencing (2017). arXiv:1705.11154.

[15] Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018). URL https://doi.org/10.1093/bioinformatics/bty191.

[16] Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005). URL https://doi.org/10.1093/bioinformatics/bti310.

[17] Oxford Nanopore Technologies. *Nanopolish*. https://github.com/jts/nanopolish.

[18] Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics* **3** (2017). URL https://doi.org/10.1099/mgen.0.000132.

[19] Ecoli r9 dataset. http://lab.loman.net/2016/07/30/nanopore-r9-data-release/.

[20] Ecoli r9.4 dataset. https://www.ncbi.nlm.nih.gov/sra/SRX4387499[accn].

[21] Nanopore WGS Consortium. *Oxford Nanopore Human Reference Datasets*. https://github.com/nanopore-wgs-consortium/NA12878.

[22] Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual b cells. *Nature Communications* **8** (2017). URL https://doi.org/10.1038/ncomms16027.

[23] The GENCODE Project. *GENCODE v27*. https://www.gencodegenes.org/human/release_27.html.

[24] Haeussler, M. *et al.* The UCSC genome browser database: 2019 update. *Nucleic Acids Research* **47**, D853–D858 (2018). URL https://doi.org/10.1093/nar/gky1095.

[25] Nanopore Technology. *Oxford Nanopore Kmer Models*. https://github.com/nanoporetech/kmer_models.

[26] DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491–498 (2011). URL https://doi.org/10.1038/ng.806.

[27] Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods* **12**, 733–735 (2015). URL https://doi.org/10.1038/nmeth.3444.

[28] Tang, F. *et al.* mRNA-seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009). URL https://doi.org/10.1038/nmeth.1315.

[29] Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nature Biotechnology* **34**, 518–524 (2016). URL https://doi.org/10.1038/nbt.3423.

[30] Wikipedia. *Single-nucleotide polymorphism, Disease.* https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism#Disease.

[31] Broad Institute. *GATK Best Practices workflow for SNP and indel calling on RNAseq data.* https://software.broadinstitute.org/gatk/guide/article?id=3891.

[32] Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for oxford nanopore sequencing. *Genome Biology* **20** (2019). URL https://doi.org/10.1186/s13059-019-1727-y.