deSALT: fast and accurate long transcriptomic read alignment with de Bruijn graph-based index

Bo Liu[1, 2] , Yadong Liu[1, 2], Junyi Li[1], Hongzhe Guo[1], Tianyi Zang[1, *], and Yadong Wang[1, *]

[1]Center for Bioinformatics, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China

[2] These authors contributed equally to this work.

Contact: bo.liu@hit.edu.cn, ydliu@hit.edu.cn, lijunyi@hit.edu.cn, hzguo@hit.edu.cn, tianyi.zang@hit.edu.cn, and ydwang@hit.edu.cn

Abstract

Long-read RNA sequencing (RNA-seq) is a promising approach in transcriptomics studies, however, the alignment of the long reads is a fundamental but still non-trivial task due to sequencing errors and complicated gene structures. We propose de Bruijn graph-based Spliced Aligner for Long Transcriptome read (deSALT), a tailored two-pass long RNA-seq read alignment approach, which constructs graph-based alignment skeletons to sensitively infer exons and uses them to generate high-quality spliced reference sequences to produce refined alignments. deSALT addresses several difficult technical issues, such as small exons and serious sequencing errors, which breakthroughs the bottlenecks of long RNA-seq read alignment. Benchmarks demonstrate that this approach has a greater ability to produce accurate and homogeneous full-length alignments and thus has enormous potentials in transcriptomics studies.

Introduction

RNA sequencing (RNA-seq) has become a fundamental approach to characterize transcriptomes. It reveals precise gene structures and quantifies gene/transcript expressions [1-5] in various applications, such as variant calling [6], RNA editing analysis [7, 8], and gene fusion detection [9, 10]. However, current widely used short read sequencing technologies have limited read length and systematic bias from library preparation. These drawbacks limit more accurate alignment [11] and precise gene isoform analysis [12], thus creating a bottleneck for transcriptomic studies.

Two kinds of long read sequencing technologies, i.e., single molecule real time (SMRT) sequencing produced by Pacific Biosciences (PacBio) [13] and nanopore sequencing produced by Oxford Nanopore Technologies (ONT) [14], are emerging and promising to breakthrough the bottleneck of short reads in transcriptomic analysis. Both of them enable the production of much longer reads, the mean and maximum lengths of the reads being over ten to hundreds of thousands of base pairs (bp) [15, 16], respectively. Taking this advantage, full-length transcripts can be sequenced by single reads, which is promising for substantially improving the accuracy of gene isoform reconstruction. Furthermore, there is less systematic bias in the sequencing procedure [17], which is also beneficial to gene/transcript expression quantification.

Besides their advantages, PacBio and ONT reads have much higher sequencing error rates than that of short reads. For PacBio SMRT sequencing, the sequencing error rate of raw reads (“subreads”) is about 10% to 20% [16]; for ONT nanopore sequencing, the sequencing error rates of 1D and 2D (also known as $1D^2$) reads are about 25% and 12% [18, 19], respectively. PacBio SMRT platforms can produce reads of inserts (ROIs) by sequencing circular fragments multiple times to largely reduce sequencing errors. However, this technology has lower sequencing yields and reduced read lengths. Therefore, these serious sequencing errors raise new technical challenges for RNA-seq data analysis. Read alignment could be the most affected one, and the effect may not be limited to the read alignment itself since it is fundamental to many downstream analyses.

Previous studies [20-22] have demonstrated that noisy DNA-seq long read alignment is a non-trivial task. Many technical issues, such as the serious sequencing errors, potential genome variants, and large read lengths, need to be handled well. For RNA-seq long read alignment, the task is even more difficult since the aligner has to deal with numerous splicing events besides the issues mentioned above. This requires the aligner to have a strong ability to implement a highly complicated split alignment (also called “spliced alignment”) to correctly recognize many splicing junctions and map the bases to the corresponding exons. Although most of the proposed DNA-seq long read alignment approaches have the ability to implement split alignment to handle genome structure variations (SVs) [21-23], tailored algorithms are still in high demand because splicing junctions occur more frequently and the lengths of exons are much shorter and divergent.

There have been several approaches supporting RNA-seq long read alignment, such as BBMap [24], GMAP [25], STAR [26], BLAT [27], and Minimap2 [28]. All of these approaches are based on the commonly used seed-and-extension strategy, by which various seeding and extension methods are implemented to address the technical issues. They all have the ability to handle splicing junctions. However, most of them are relatively slow speed [28], mainly due to the numerous short matches in the seeding step and the time-consuming local alignment in the extension step. Moreover, some of the algorithms have lower sensitivity [29], i.e., many reads are unaligned or only partially aligned, due to their relatively poor ability to handle sequencing errors. An outstanding algorithm is Minimap2, which simultaneously achieves tens of times faster speed and similar or higher sensitivity than other state-of-the-art aligners. This algorithm mainly benefits from its well-designed minimizer-based indexing [30] and SSE-based local alignment methods [31], which greatly improve the efficiency of the seeding and extension steps. Furthermore, its specifically designed local extension method is suited to handling splicing junctions.

In absolute terms, the ultimate goal of the task is to map all the bases for all the reads correctly. However, this could still be non-trivial to state-of-the-art aligners in several aspects. One problem is the alignment of the bases from relatively short exons (e.g., exons having only a few tens of bp). It is extremely hard to find seeds in the read parts from such short exons under the circumstances of serious sequencing errors and potential variants, so that the read parts are usually unaligned or mistakenly aligned. Another issue is that it is difficult to align the bases near the splicing junctions correctly. This problem also exists in short RNA-seq read alignment; however, it is more serious in the alignment of noisy long RNA-seq reads. Moreover, with the effect of sequencing errors, the alignments of the reads from the same gene isoform are usually divergent from each other, which is also misleading downstream analysis.

Herein, we propose the de Bruijn graph-based spliced aligner for long transcriptome reads (deSALT). deSALT is a fast and accurate RNA-seq long read alignment approach which takes the advantages of a novel two-pass read alignment strategy based on the de Bruijn graph-based index. It has a strong ability to handle complicated gene structures and serious sequencing errors to produce sensitive, accurate, and homogeneous alignments. For most of the reads, deSALT can produce full-length alignments to recover the exons and splicing junctions thoroughly along the entire reads. Moreover, the speed of deSALT is also faster than or comparable to state-of-the-art approaches. We believe that it has the potential to play an important role in many forthcoming transcriptomic studies.

Results

Motivation and overview of the deSALT approach

The seed-and-extension approach is suited to spliced alignment since it is able to match the short tokens of the read to its spanning exons first (i.e., seeding) and then implement base-level alignment between the read and the matched exons (i.e., extension). However, under the circumstances of frequent splicing events and serious sequencing errors, this task is non-trivial in practice. For a single read, it is usually difficult to find matches between the read and all of its spanning exons accurately, especially for the read parts with serious sequencing errors and relatively short exons. Since it is hard to compose a local reference sequence containing all the spanning exons of the read, the extension alignment would be less accurate, or some of the read parts could be unaligned or clipped. Moreover, due to the randomness of the sequencing errors, various mistakes could be made in the seeding and extension phases in regard to multiple reads from the same gene isoform, and the produced alignments of the reads could be divergent from each other.

Motivated by these technical problems and existing short RNA-seq read alignment algorithms [26, 32], deSALT uses a tailored two-pass approach to align the noisy long reads (a schematic illustration is in Figure 1). In the first pass, it employs a graph-based genome index [33] to find match blocks (MBs) between the read and the reference and uses a sparse dynamic programming (SDP) approach to compose the MBs into alignment skeletons (referred to as the “alignment skeleton generation” step). All the alignment skeletons of all the reads are then integrated to comprehensively detect the exon regions (referred to as the “exon inference” step). In the second pass, deSALT relocates the short matches between the read and the detected exons to compose a local spliced reference sequence (LSRS), which is expected to be a concatenation of all the spanning exons of the read. The read is aligned against the LSRS to produce a refined base-level alignment (referred to as the “refined alignment” step).

The key point of deSALT is its comprehensive analysis of the local matches of all the reads through alignment skeletons. Since the sequencing errors are random [17], each of the alignment skeletons contains some distinct as well as some complementary information about exon regions, which look like puzzle pieces, and the integration of alignment skeletons can effectively filter the sequencing errors to implement a sensitive and noise-robust detection of exons. In the later step, the detected exons help to narrow down the searching space to find additional short matches which cannot be detected by the relatively longer seeds used in the initial step. With these local matches, deSALT enables the effective inference of all the spanning exons of a given read and the composition of a high-quality spliced reference sequence to produce accurate full-length alignment. This approach is robust to very short exons (e.g., exons < 30 bp), frequent splicing events, potential small variants, as well as sequencing errors. Furthermore, deSALT generates homogeneous LSRSs for the reads from the same gene isoform with the integrated information, which enables the production of higher homogeneous alignments.

deSALT also has fast speed with its tailored design. Unlike conventional two-pass alignment approaches [26, 32] where both of the two passes produce base-level alignment, deSALT only uses pseudo-alignment (alignment skeleton) in the first pass, the operation of which is similar to the seeding process. Thus, the whole process is just like a one-pass alignment plus a fast integration of the alignment skeletons. Moreover, some optimized implementations, for example, graph-index-based skeleton generation and SIMD-based local alignment [28, 31], also help to accelerate the speed.

Results on simulated datasets

We simulated 36 RNA-seq long read datasets with various sequencing error rates and read lengths (Supplementary Table 1) to mimic the datasets from mainstream platforms, i.e., ONT 1D reads (error rate: 25%; mean read length: 7800 bp), ONT 2D ($1D^2$) reads (error rate: 12%; mean read length: 7800 bp), PacBio subreads (error rate: 15%; mean read length: 8000 bp), and PacBio ROI reads (error rate: 2%; mean read length: 2000 bp). For each of the mimicked platforms, there were 9 datasets respectively from 3 species (human, mouse, and fruit fly) and 3 sequencing depths (4X, 10X, and 30X). All datasets were produced by PBSim [34] based on Ensembl gene annotations [35] (human: GRCh38, version 94; mouse: GRCm38, version 94; fruit fly: BDGP6, version 94), and the error models were configured by referring to previous studies on the characteristics of the sequencing platforms [17, 36]. deSALT and two state-of-the-art approaches, Minimap2 and GMAP, were applied to all the datasets for comparison. The Methods section provides more details of the implementation of the benchmarking.

Five metrics (i.e., Base%, Exon%, Read80%, Read100%, and #Bases/s) were used to assess the sensitivity, accuracy, and performance of the aligners, respectively (refer to the Methods section for the definitions). The results of the 36 simulated datasets are shown in Figure 2 and Supplementary Tables 2 to 6. Four main observations were made.

1) deSALT has outstanding alignment yields.

In our simulations, deSALT had the overall highest base% statistics, indicating that it mapped more bases to their correct positions. In particular, the advantage of deSALT was more obvious on the datasets with medium and high error rates (i.e., ONT 2D reads, PacBio subreads, and ONT 1D reads), which is preferable for handling real noisy long reads. Moreover, deSALT showed greater advantages in the exon% statistics, suggesting that it has a stronger ability to recover the exons and splicing junctions within the reads.

deSALT also obviously outperformed other state-of-the-art aligners on the Read80% and

Read100% statistics, indicating that it has better ability to produce full-length alignments. In particular, deSALT produced the highest number of Read100% reads for all the datasets (i.e., for larger number of reads, deSALT can correctly align all of their exons without introducing false positives). These accurate full-length read alignments are valuable to downstream analysis.

In absolute terms, deSALT correctly aligned most of the bases as well as the exons for the reads with low and medium error rates (i.e., PacBio ROIs and subreads and ONT 2D reads). For the error-prone ONT 1D reads, the exon% statistics (about 64%–90%) were affected to some extent, although it outperformed other aligners. The most affected datasets were the low coverage (4X) mammalian (human and mouse) datasets. This was mainly due to the fact that the bases near exon boundaries are very difficult to confidently align under the circumstance of serious noise. However, it is worth noting that the yield on these error-prone reads improved with the increase in read depth. It is a feature of the two-pass approach that the effect of sequencing errors can be better mitigated and the detection of exons can be improved with more available reads, and all the reads share this profit to compose more sensitive alignments.

Furthermore, deSALT has the ability to produce not only accurate, but also homogeneous alignments. This is also an advantage of the two-pass alignment, and deSALT tends to compose homogeneous LSRSs for the reads from same gene isoforms, which helps to align them to the correct positions simultaneously. However, one-pass approaches are more easily affected by sequencing errors and other factors, such as very small exons and frequent splicing events, which usually produce more heterogeneous alignments with more mistakes.

The speed of deSALT is similar to that of Minimap2, and both are tens of times faster than GMAP (Figure 1B and Supplementary Table 6). This speed is suited to large-scale datasets.

Figure 3 shows a typical example describing the characteristics of deSALT.

2) deSALT has good ability to align the reads spanning short exons.

We specifically assessed the alignment of the reads spanning short exons (exons < 31 bp), and the results (Supplementary Table 3) suggest that deSALT largely outperforms other aligners. It is derived from the two-pass approach that short exons can be better detected with the generation and integration of alignment skeletons, and the discovered exons are fully considered by the shorter local matches used in the second pass. deSALT helps to compose high-quality LSRSs to correctly align the read parts spanning those short exons. However, other state-of-the-art aligners that use a one-pass strategy are more likely to be affected by splicing events and sequencing errors, which results in reduced ability to find local matches on short exons. Also, the corresponding read parts are mistakenly aligned. Two examples are provided in Figure 3B and Supplementary Figure 1.

3) deSALT has good ability to handle multiple splicing events and multiple gene isoforms.

It is also non-trivial to align reads with many splicing events and/or from the genes with multiple isoforms [11]. We assessed the alignment of the reads from the transcripts with various number of exons (2–5 exons, 6–9 exons, and >9 exons). deSALT can produce equally good alignments for all three read groups (Supplementary Table 4), indicating that it enables the handling of numerous splicing events within the reads (an example is provided in Supplementary Figure 2). Minimap2 showed a similar trend, but its Read80% and Read100% statistics were lower. GMAP showed a significant decrease in the Read80% and Read100% statistics as the number of exons increased, indicating that it might not be good at handling reads with many splicing events. We also assessed the alignment of the reads from the genes with multiple isoforms. The results for deSALT (Supplementary Table 5) demonstrate that there is no significant difference between the alignment of reads from genes with multiple isoforms and that of genes with single isoforms, suggesting that deSALT has the ability to handle genes with multiple isoforms.

4) deSALT can further improve the alignment of error-prone reads with gene annotations.

deSALT supports to use gene annotations to facilitate read alignment. Basically, it combines the annotated exons with the alignment skeletons and builds a more comprehensive exon map. The results (Figure 2A and Supplementary Table 2) demonstrate that gene annotations are helpful to enhance the alignment of very noisy reads (i.e., ONT 1D reads). This is mainly because, for most such reads, deSALT only finds a few matches to build incomplete alignment skeletons which lower the sensitivity of exon detection. In this situation, gene annotations supply additional information to find matches for those read parts from missed exon regions. This solves many error-prone read parts (an example is provided in Supplementary Figure 3) and helps to produce full-length alignments (see the gains in the Read80% and Read100% statistics). With this feature, deSALT provides the opportunity to use noisy long reads better, which has many potentials in transcriptomics studies.

Overall, the simulation results demonstrate that deSALT is able to achieve excellent sensitivity, accuracy, and performance simultaneously. Especially, it has the ability to address many difficult issues, such as sequencing errors, short exons, frequent splicing events, multiple isoforms and so on, which makes it a promising approach to breakthrough the bottlenecks of long RNA-seq read alignment.

Results on real sequencing datasets

We assessed the aligners with two real sequencing datasets, one from a well-studied CEPH sample (NA12878) produced by the ONT platform (available at https://github.com/nanopore-wgs-consortium/NA12878; containing 15152101 reads and 14134831170 bases in total) and the other from a mouse sample produced by the PacBio platform [37] (Accession

Number: SRR6238555; containing 2269795 reads and 3213849871 bases in total).

We used a series of metrics based on gene annotations to evaluate the alignments (i.e., #BaseA, #BaseGA, #ExonP, #ExonGO, #ExonGA, #ExonGA(x), #ReadGA) due to a lack of ground truth (refer to Methods section for definitions). It is also worthnoting that Ensembl gene annotations (human: GRCh38, version 94 and mouse: GRCm38, version 94) were employed for the assessment.

The results are provided in Figure 4 and Supplementary Tables 7 and 8. Four observations were made.

1) deSALT still has the best alignment yields.

For both of the two real datasets, deSALT achieved the highest #BaseGA statistics (i.e., it aligned most bases to the annotated exon regions). Moreover, deSALT also had the highest numbers of predicted exons being overlapped by (#ExonGO) and exactly matched to (#ExonGA) annotated exons. These statistics indicate that deSALT achieves good sensitivity. Furthermore, deSALT had the highest #ReadGA statistics, indicating that it has better ability to produce correct full-length read alignments. The time cost with 24 and 32 CPU threads was also assessed (Supplementary Table 7), and the results suggest that deSALT is marginally (about 20%) faster than Minimap2.

It was also observed that the #BaseGA of Minimap2 on the two datasets was close to that of deSALT, indicating that the two approaches have similar alignment yields overall. However, deSALT outperformed Minimap2 on #ExonGO, #ExonGA, and #ReadGA statistics for both of the datasets. We investigated the detailed alignment results and found that, similar to that of the simulated reads, this outperformance derives from deSALT's better ability to deal with short exons and produce more homogeneous alignments (see below for details). A typical example of the alignment of real sequencing reads is shown in Figure 5.

2) deSALT has outstanding ability to handle relatively short exons.

deSALT also demonstrates outstanding ability to handle short exons. We assessed the alignment of the bases putatively from short exons by a series of #ExonGA(x) statistics (Figures 4B and 4D), i.e., ExonGA(20), ExonGA(30), ExonGA(40), ExonGA(50), and ExonGA(60). The results demonstrate that, for both datasets, deSALT enables the recovery of a higher number of short exons. It is worth noting that although only a small proportion of exons are short, they are important to the study of gene splicing, and so it is of great value to correctly align such read parts. However, this is still a difficult task for other state-of-the-art aligners. Furthermore, this advantage helps deSALT to produce better full-length alignments for reads from the genes with small exons (an example is shown in Supplementary Figure 4) and to achieve overall higher #ReadGA statistics.

3) deSALT produces homogeneous alignments.

Another outstanding ability of deSALT is the production of homogeneous alignments. It can be observed from the read alignments of deSALT that in local regions, various reads usually have highly similar alignments and exon boundary predictions which also coincide with gene annotations. However, for other aligners, the predicted exon boundaries of the same reads are usually more divergent from each other. As shown in the example in Supplementary Figure 5, the homogeneous alignments of deSALT could be more accurate overall, especially for those bases near exon boundaries. The homogeneous alignments are also more useful to the study of splicing events since there is less noise in these alignments than in ambiguous alignments.

4) A proportion of bases are aligned to unannotated regions

According to the Ensembl gene annotations, there were about 10% of the bases aligned by deSALT to regions other than the annotated exons: 1) a proportion of the bases (5.60% for the human ONT dataset and 5.13% for the mouse PacBio dataset) were aligned to intron regions; 2) a proportion of the bases (4.61% for the human ONT dataset and 4.02% for the mouse PacBio dataset) were aligned to intergenic regions. Minimap2 also had similar proportions of bases aligned to such regions. We found that the alignments of these read parts were highly clustered: i.e., in most cases, there were multiple reads aligned in a local region, indicating that there could be unannotated exons or novel transcripts. Furthermore, we found that deSALT and Minimap2 had similar outputs for these read parts, which also indicates that the alignments are plausible. Two examples in intragenic and intergenic regions are shown in Supplementary Figures 6 and 7, respectively.

Discussion and conclusion

Long read sequencing technologies provide the opportunity to break the limitations of short reads and improve transcriptomics studies. However, complex gene structures and serious sequencing errors make it still a non-trivial task to produce accurate full-length alignments to exert the advantages of long RNA-seq reads. So, there is wide demand for the development of more advanced read alignment algorithms to break through this bottleneck. Herein, we proposed deSALT, a novel read alignment algorithm using the de Bruijn graph-based index and a tailored two-pass strategy, as a solution to this important open problem. Mainly, we show how to build and integrate spliced alignment skeletons to handle sequencing errors and complex gene structures in order to generate high-quality spliced reference sequences and use them to produce accurate and homogeneous full-length alignments for long RNA-seq reads. To the best of our knowledge, deSALT is the first long RNA-seq read alignment approach that fully considers the intermediate results of all the reads and takes this advantage to produce refined spliced alignments.

On both the simulated and real datasets, the deSALT results demonstrate its good sensitivity and accuracy. For most of the datasets, it maps the highest number of bases to their ground truth positions or the positions supported by gene annotations. Its advantage with regard to the recovery of exons and splicing junctions is more obvious, suggesting that deSALT has excellent ability to produce spliced alignments. This is further demonstrated by several kinds of difficult scenarios, such as very short exons, numerous splicing events, and genes with multiple isoforms.

A more important feature of deSALT is its outstanding ability to produce accurate and homogeneous full-length alignments. With the ever-increasing length of reads, this feature is in great demand since it provides the opportunity to directly investigate gene structures. However, it requires the employed aligner to handle many technical issues well and simultaneously. deSALT largely improves full-length alignment by using several key techniques, such as sensitive exon detection, local exon matching, and LSRS generation. For larger numbers of reads, deSALT can comprehensively and accurately recover their splicing junctions by single alignments, and the produced alignments are homogeneous and confident. This contribution has the potential to facilitate many downstream analyses.

The real read alignments of deSALT highly coincide with gene annotations; however, there are still a proportion of reads and bases being mapped to intron and intergenic regions. Considering the similar results independently produced by deSALT and Minimap2, there could be some unknown transcripts being sequenced, and the alignments are plausible. Moreover, we also found that deSALT and Minimap2 similarly clipped a proportion of bases. We tried to extract some of the corresponding read parts and align them with BLAT [27]; however, no successful alignment was produced (data not shown). In this situation, we realized that these clipped read parts could be extremely low quality.

In addition to only use reference genome, deSALT supports the use of gene annotations to enhance the alignment. However, the benchmarking results were to some extent unexpected in that there was no significant difference between the alignment with and without gene annotations, except for in the low-depth, high-error rate (ONT 1D) datasets. This is also reasonable since the two-pass strategy has a strong ability to mitigate the effect of moderate sequencing errors even if the read depth is low. Moreover, this ability can be further enhanced with the increase of sequencing depth, so that high-coverage ONT 1D datasets can also be sensitively aligned without gene annotations. However, this function of deSALT is still useful since gene expression is uneven, i.e., there are always less expressed genes with fewer reads being sequenced, and gene annotations could make their own contributions to align those reads.

Overall, with its outstanding alignment yields and performance, deSALT is suited to aligning long RNA-seq reads. We believe it will be a useful alignment tool and play an important role in many cutting-edge transcriptomics studies.

Methods

Steps of the deSALT approach

deSALT supports long RNA-seq reads with either high (e.g., PacBio subreads and ONT reads) or low (e.g., PacBio ROI reads) error rates. Input reads are aligned in three major steps as follows:

1) Alignment skeleton generation (first-pass alignment): For each of the reads, deSALT uses the RdBG-index [33] to find the maximal exact matches between the unitigs of a reference de Buijn graph (RdBG) and the read (termed as U-MEMs) and to build one or more alignment skeletons using an SDP approach.

2) Exon inference: deSALT maps all the alignment skeletons to the reference and infers potential exons from the projections of the skeletons. A local sequence-based scoring system [38] is employed to refine the inferred exons. Moreover, it is optional to introduce gene annotations as additional information to enhance exon detection.

3) Refined alignment (second-pass alignment): For each of the reads, deSALT finds additional local matches to the inferred exons with shorter tokens (seeds) than the ones used in the first step. Further, it combines the newly found matches and the alignment skeleton to retrieve and stich all the spanning exons to build an LSRS and implement a base-level read alignment.

Alignment skeleton generation (first-pass alignment)

The RdBG-index is built in advance by the indexing module of deBGA [33] (Supplementary Notes), and the *k*-mer size of the index is set as the default value (*k*=22) if not specifically mentioned.

For a certain read, deSALT extracts *l*-mers (*l*<*k*, default value: *l*=15) at every *m* bp (default value: *m*=5) as seeds and matches them to the unitigs of RdBG with the RdBG-index. The matches are extended in both directions to generate U-MEMs. deSALT then merges co-linear U-MEMs on the same unitigs as super U-MEMs (SU-MEMs) and maps the SU-MEMs, as well as the U-MEMs that cannot be merged to reference genome, as MBs to build alignment skeletons.

deSALT uses the MBs as vertices to build a direct acyclic graph (DAG). The edges of the DAG are defined by the pairs of MBs whose distances are no longer than a predefined maximum intron length, $\mathrm{T_{intron}}$ (default value: $\mathrm{T_{intron}}$=200,000 bp). A weight is assigned to each of the edges on the basis of the sizes of the two corresponding MBs and their distances (Supplementary Notes). An SDP approach is then used to find the path with the largest sum weight as the alignment skeleton. It is also worth noting that deSALT could produce multiple alignment skeletons with very similar scores (sum weights) for some of the reads, considering that such reads possibly have multiple “equally best” alignments.

Please refer to Section 1.1–1.3 of the Supplementary Notes for more implementation details.

Exon inference

deSALT maps all the alignment skeletons to the reference genome and uses a set of predefined rules (Section 2.1 of the Supplementary Notes) to iteratively combine the genomic regions covered by alignment skeletons from upstream to downstream. It is optional to introduce a gene annotation file (in GTF format) into this process. deSALT treats known gene isoforms as a special kind of alignment skeletons, and it also maps them to the reference genome so that the genomic regions covered by known gene isoforms and the alignment skeletons generated from reads are combined together. The combined regions are then recognized as draft exons, and their lengths and alignment skeleton coverages are calculated. The ones with too short a length and too low coverage are then filtered out.

A local sequence-based scoring system [38] is then employed to refine the draft exons (Section 2.2 of Supplementary Notes). For each of the draft exons, deSALT selects two small flanking regions. The scoring system uses predefined acceptor and donor scoring matrixes to score each of the positions in the upstream and downstream regions respectively. The positions with the highest scores in the two regions are recognized as acceptor and donor splicing sites, and the region in between is determined to be a refined exon.

Refined alignment (second-pass alignment)

Refined alignment is mainly implemented in two sub-steps as follows:

1) LSRS generation: deSALT splits the read into a series of parts and separately composes partial LSRSs for each of them (Section 3.1 of the Supplementary Notes). Here, each read part is defined as a specific substring of the read within two neighboring MBs of its alignment skeleton. For a read part, deSALT detects a set of exons (termed “spanning exons”) which are placed in between or nearby the two corresponding MBs and have short matches to the read part. The spanning exons are then stitched together as the whole LSRS.

2) Base-level alignment: deSALT aligns each of the read parts to its corresponding LSRS using a SIMD-based implementation [28, 31] of semi-global alignment (Section 3.2 of the Supplementary Notes). Furthermore, deSALT checks if there are large deletions in the CIGAR information; if there are, deSALT removes the corresponding deletion part(s) in the LSRS and realigns the read with the updated LSRS. This process is helpful for handling exons with alternative splicing sites (i.e., the read part is only from a part of some inferred exon, but the whole exon is fully included in the LSRS).

It is also worth noting that for the reads with multiple alignment skeletons, deSALT processes each of the skeletons separately and possibly produces multiple alignments for one read. In this situation, deSALT chooses the alignment with the highest score as the primary alignment and outputs other alignments as secondary alignments.

Implementation of the simulation benchmark

All the benchmarks were implemented on a server with Intel Xeon E4280 CPU at 2.0GHZ and 1 Terabytes RAM, running Linux Ubuntu 16.04. The simulated datasets were generated from the reference of the three organisms: Homo sapiens GRCh38 (human), Mus musculus GRCm38 (mouse), and Drosophila melanogaster r6 (fruit fly), with corresponding Ensembl gene annotations [35].

More precisely, each of the datasets was simulated using the following three steps, which are similar to a previous study on the evaluation of long RNA-seq read aligners [29]:

1) Given a gene annotation file, the recorded gene annotations on scaffolds, assembly patches, and alternative loci were scanned and three sets of genes were extracted. Each of the sets corresponded to a specific type (i.e., genes with single splicing isoforms, genes with multiple splicing isoforms, and genes with short exons (<31 bp), respectively).

2) The three sets of genes were separately used to generate *in silico* transcript sequences. For a certain gene in a specific set, the transcript sequences were generated according to all its isoforms. All the generated transcript sequences were integrated together, and the transcript sequences shorter than 200 bp were filtered out. The remaining transcript sequences were inputted into PBSim [34]. The numbers of the employed transcript sequences from the various organisms (gene annotations) are listed in Supplementary Table 9.

3) For each organism, four sequencing error models were used for the simulation:

"PacBio ROI reads": sequencing error rate = 2%, mean read length = 2000 bp;

"PacBio subreads": sequencing error rate = 15%, mean read length = 8000 bp;

"ONT 2D (1D^2) reads": sequencing error rate = 13%, mean read length = 7800 bp;

"ONT 1D reads": sequencing error rate = 25%, mean read length = 7800 bp.

The models were configured by referring to previous studies [17, 36]. For each model, three datasets of various sequencing depths (4X, 10X, 30X) were simulated, i.e., a total of $3 \times 4 \times 3 = 36$ datasets were generated. The availability of the simulated datasets is provided in the Supplementary Notes.

The following five metrics were used to evaluate the alignment results of the simulated reads.

Base%: the proportion of bases being correctly aligned to their ground truth positions (i.e., the mapped positions of the bases were within 5 bp of their ground truth positions).

Exon%: the proportion of exons being correctly mapped. An exon in a certain read was considered to be correctly mapped only if its two boundaries were mapped within 5 bp of their ground truth positions.

Read80%: the proportion of Read80% reads. A read was considered to be a Read80% read only if it met two conditions, namely $N_T/N_G > 80\%$ and $N_T/N_P > 80\%$, where N_G is the number of ground truth exons within the read, N_P is the number of exons predicted by the alignment, and N_T is the number of true positive exons. Herein, a predicted exon is considered to be a true positive exon only if there was a ground truth exon in the read, and the distance between the corresponding boundaries of the predicted exon and the ground truth exon were within 5 bp.

Read100%: the proportion of Read100% reads. A read was considered to be a Read100% read only if it met two conditions, namely $N_T/N_G = 100\%$ and $N_T/N_P = 100\%$. It is worth noting that a Read100% read indicates that the read has a highly correct full-length alignment.

#Bases/s: the number of bases aligned per second, which depicts the alignment speed and is computed by N_{base}/T_{aln}, where N_{base} is the total number of bases in the dataset and T_{aln} is the elapsed time.

Implementation of the real data benchmark

The benchmarks were implemented with the same hardware environment as that used for the simulated datasets. Two real datasets respectively produced by ONT and PacBio platforms were used. The ONT dataset was from the NA12878 sample, which was sequenced by the ONT MinION sequencer by using direct RNA sequencing kits (30 flowcells) and the 1D ligation kit (SQK-LSK108) on R9.4 flowcells with R9.4 chemistry (FLO-MIN106). More detailed information about this dataset is available at https://github.com/nanopore-wgs-consortium/NA12878. The PacBio dataset (Accession Number: SRR6238555) is a full-length isoform sequencing of total mouse RNA using standard PacBio-seq protocols [37]. The availability of the two real datasets is provided in the Supplementary Notes.

The following metrics were used to evaluate the alignment results of the real sequencing reads.

#BaseA: the number of bases being aligned.

#BaseGA: the number of bases aligned to the positions within annotated exons.

#ExonP: the number of exons predicted by the alignments (also termed “predicted exons”). Here, the predicted exons in various reads were independently considered.

#ExonGO: the number of predicted exons being overlapped by annotated exons (also termed "overlapped exons"). Herein, a predicted exon was considered to be overlapped by annotated exons only if there was at least one annotated exon and at least 10 bp overlapping between the predicted exon and the annotated exon.

#ExonGA: the number of predicted exons being exactly matched by annotated exons (also termed "exactly matched exons"). Herein, a predicted exon was considered to be exactly matched by annotated exons only if there was an annotated exon and the distance between the corresponding boundaries of the predicted exon and the annotated exon were within 5 bp.

#ExonGA(x): the number of exactly matched exons whose lengths were shorter than x bp.

#ReadGA: the number of ReadGA reads. A read was considered to be a ReadGA read only if each of the intron boundaries implied by its alignment was within 5 bp of an annotated exon. Herein, a ReadGA read indicates that the read could has a correct full-length alignment.

Availability of data and material

The source code of deSALT and the data simulations and benchmarking scripts are available at https://github.com/hitbc/deSALT.

Please refer to the Supplementary Notes for the availability of the simulated and real sequencing datasets.

Author contributions

BL designed the method, YL implemented the method, and BL, YL, JL and HG performed the analysis. All of the authors wrote the manuscript. BL and YL contributed equally to this work.

Acknowledgement

We are very grateful to Prof. Yi Xing of the University of Pennsylvania and the Children's Hospital of Philadelphia for his helpful discussion.

This work has been supported by the National Key Research and Development Program of China (Nos: 2018YFC0910504, 2017YFC0907503, and 2017YFC1201201).

Additional information

Competing interests

The authors declare that they have no competing interests.

References

1. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**:57-63.

2. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**:621-628.

3. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.** *Nat Biotechnol* 2015, **33**:290-295.

4. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc* 2012, **7**:562-578.

5. Teng M, Love MI, Davis CA, Djebali S, Dobin A, Graveley BR, Li S, Mason CE, Olson S, Pervouchine D, et al: **A benchmark for RNA-seq quantification pipelines.** *Genome Biol* 2016, **17**:74.

6. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS: **SNP discovery via 454 transcriptome sequencing.** *Plant J* 2007, **51**:910-918.

7. Peng Z, Cheng Y, Tan BC, Kang L, Tian Z, Zhu Y, Zhang W, Liang Y, Hu X, Tan X, et al: **Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome.** *Nat Biotechnol* 2012, **30**:253-260.

8. Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, Li JB: **Identifying RNA editing sites using RNA sequencing data alone.** *Nat Methods* 2013, **10**:128-132.

9. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T,

Palanisamy N, Chinnaiyan AM: **Transcriptome sequencing to detect gene fusions in cancer.** *Nature* 2009, **458:**97-101.

10. Kim D, Salzberg SL: **TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.** *Genome Biol* 2011, **12:**R72.

11. Engstrom PG, Steijger T, Sipos B, Grant GR, Kahles A, Ratsch G, Goldman N, Hubbard TJ, Harrow J, Guigo R, Bertone P: **Systematic evaluation of spliced alignment programs for RNA-seq data.** *Nat Methods* 2013, **10:**1185-1191.

12. Steijger T, Abril JF, Engstrom PG, Kokocinski F, Hubbard TJ, Guigo R, Harrow J, Bertone P: **Assessment of transcript reconstruction methods for RNA-seq.** *Nat Methods* 2013, **10:**1177-1184.

13. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al: **Real-time DNA sequencing from single polymerase molecules.** *Science* 2009, **323:**133-138.

14. Mikheyev AS, Tin MM: **A first look at the Oxford Nanopore MinION sequencer.** *Mol Ecol Resour* 2014, **14:**1097-1102.

15. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al: **Nanopore sequencing and assembly of a human genome with ultra-long reads.** *Nat Biotechnol* 2018, **36:**338-345.

16. Rhoads A, Au KF: **PacBio Sequencing and Its Applications.** *Genomics Proteomics Bioinformatics* 2015, **13:**278-289.

17. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA: **Pacific biosciences sequencing technology for genotyping and variation discovery in human data.** *BMC Genomics*

2012, **13**:375.

18. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ: **Assessing the performance of the Oxford Nanopore Technologies MinION.** *Biomol Detect Quantif* 2015, **3**:1-8.

19. Sovic I, Krizanovic K, Skala K, Sikic M: **Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads.** *Bioinformatics* 2016, **32**:2582-2589.

20. Li H, Homer N: **A survey of sequence alignment algorithms for next-generation sequencing.** *Brief Bioinform* 2010, **11**:473-483.

21. Chaisson MJ, Tesler GJBb: **Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory.** 2012, **13**:238.

22. Liu B, Gao Y, Wang YJB: **LAMSA: fast split read alignment with long approximate matches.** 2017, **33**:192-201.

23. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MCJNM: **Accurate detection of complex structural variations using single-molecule sequencing.** 2018, **15**:461-468.

24. Bushnell B: **BBMap: a fast, accurate, splice-aware aligner.** Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States); 2014.

25. Wu TD, Watanabe CKJB: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences.** 2005, **21**:1859-1875.

26. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29**:15-21.

27. Kent WJJGr: **BLAT—the BLAST-like alignment tool.** 2002, **12**:656-664.

28. Li HJB: **Minimap2: pairwise alignment for nucleotide sequences.** 2018, **34**:3094-3100.

29. Križanovic K, Echchiki A, Roux J, Šikic M: **Evaluation of tools for long read RNA-seq splice-aware alignment.** *Bioinformatics* 2018, **34**:748-754.

30. Roberts M, Hayes W, Hunt BR, Mount SM, Yorke JA: **Reducing storage requirements for biological sequence comparison.** *Bioinformatics* 2004, **20**:3363-3369.

31. Suzuki H, Kasahara MJb: **Acceleration Of Nucleotide Semi-Global Alignment With Adaptive Banded Dynamic Programming.** 2017:130633.

32. Trapnell C, Pachter L, Salzberg SLJB: **TopHat: discovering splice junctions with RNA-Seq.** 2009, **25**:1105-1111.

33. Liu B, Guo H, Brudno M, Wang YJB: **deBGA: read alignment with de Bruijn graph-based seed and extension.** 2016, **32**:3224-3232.

34. Ono Y, Asai K, Hamada MJB: **PBSIM: PacBio reads simulator—toward accurate genome assembly.** 2012, **29**:119-121.

35. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CGJNar: **Ensembl 2018.** 2017, **46**:D754-D761.

36. Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang XJ, Buck D, Au KF: **Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis.** *F1000Res* 2017, **6**:100.

37. Jiang M, Zhang S, Yang Z, Lin H, Zhu J, Liu L, Wang W, Liu S, Liu W, Ma Y, et al: **Self-Recognition of an Inducible Host lncRNA by RIG-I Feedback Restricts Innate Immune Response.** *Cell* 2018, **173**:906-919.e913.

38. Kim K-B, Park K, Kong EBJJob, biology m: **A method for identifying splice sites and translation**

start sites in human genomic sequences. 2002, **35:**513-517.

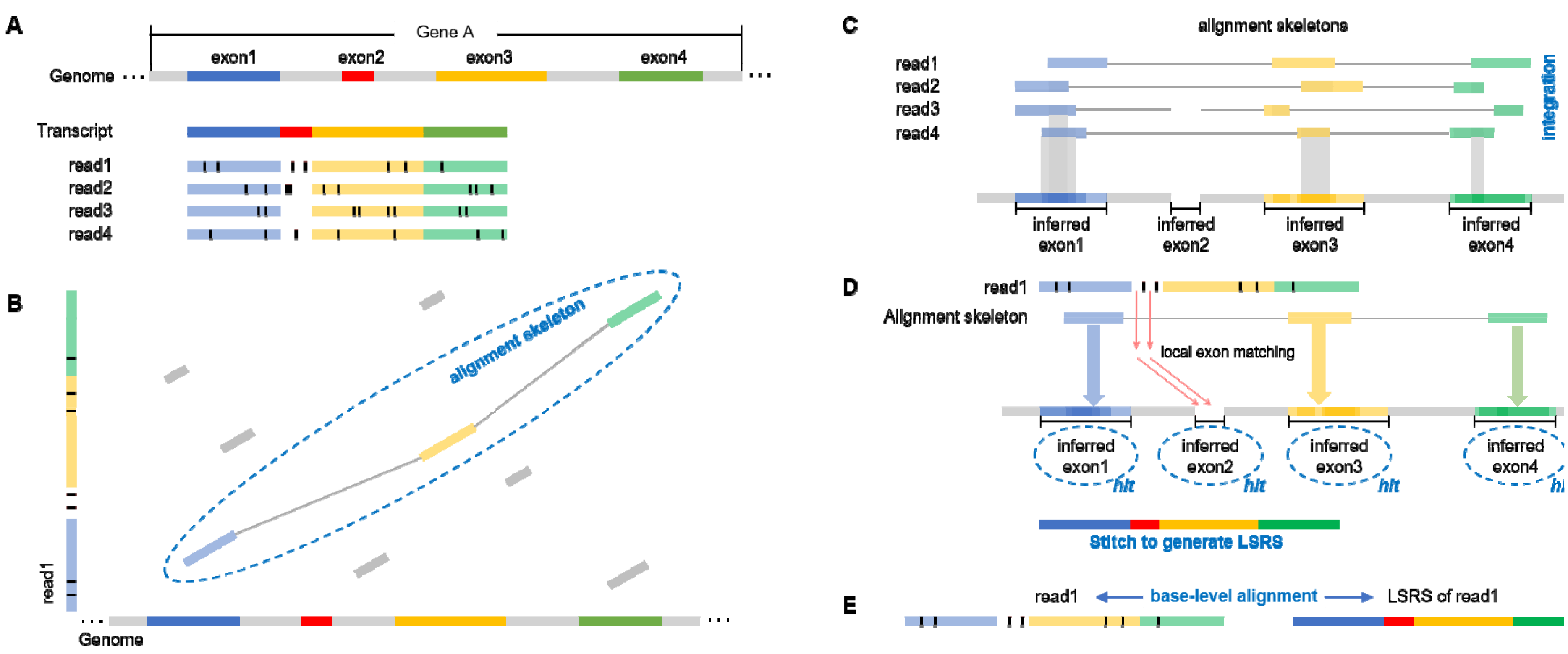

Figure 1. A schematic illustration of the deSALT approach

(A) There is a gene (termed "Gene A") with four exons (respectively marked by the colors blue, red, yellow, and green; introns are marked by grey color) and four reads th sequence throughout the whole transcript (assuming that Gene A has only one isoform). Moreover, each of the reads has some sequencing errors (marked by the short bars in the reads). **(B)** Alignment skeleton generation (first-pass alignment): for each of the reads (read1 is employed as an example), deSALT finds the MBs between it an reference genome (marked as colored bars) and connects them to build an optimized alignment skeleton using an SDP approach. **(C)** Exon inference: deSALT integrates the generated alignment skeletons by mapping their involved MBs to the reference genome. The projections of the MBs are analyzed to infer exon regions in the refer genome. **(D-E)** Refined alignment (second-pass alignment): for each of the reads, deSALT finds additional local matches on the exons between or near the exons involv the alignment skeleton. Further, it recognizes all the inferred exons related to the alignment skeleton or the newly found local matches as "hit exons" and stitches all of the generate an LSRS. (The figure shows that there are two newly found matches on exon 2, and they help to recuse this exon to build a correct LSRS.) The read is then ali with the LSRS to produce a refined alignment.

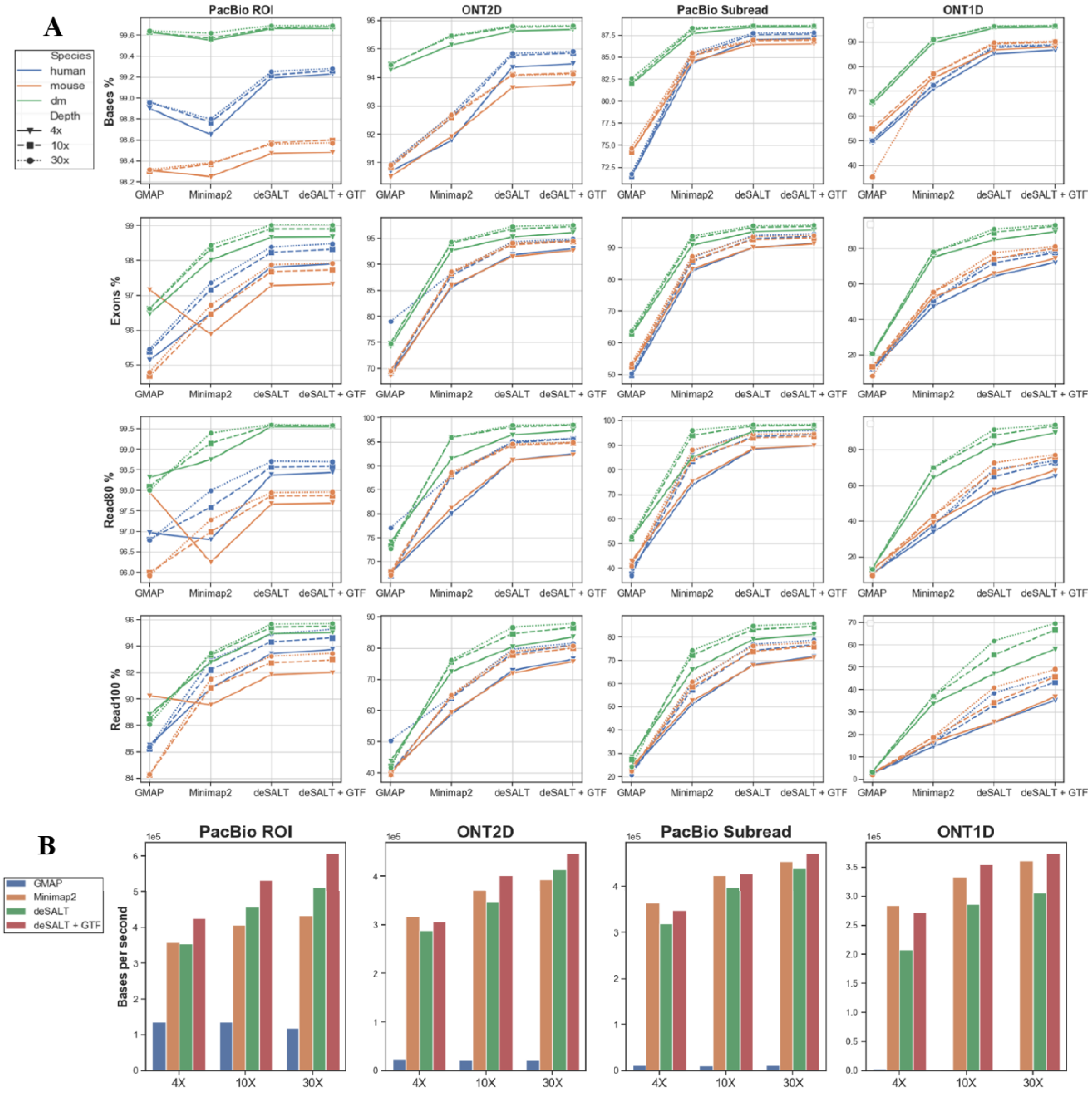

Figure 2. Results on simulated datasets

The figure depicts the yields (A) and speed (B) of the aligners on the simulated datasets. It is worth noting that the deSALT results both with and without gene annotations are shown (indicated as “deSALT+GTF” and “deSALT”, respectively). **(A)** Each of the subplots indicates one of the four metrics (Base%, Exon%, Read80%, and Read100%, respectively) of the aligners on the datasets in a specific error model (PacBio ROI, ONT2D, PacBio subreads, and ONT1D, respectively). In each subplot, the blue, orange, and green lines respectively correspond to the results of the datasets from various kinds of species (i.e., human, mouse, and fruit fly). Moreover, the shapes (reverse triangles, rectangles, and circles) indicate the datasets in various sequencing depths (4X, 10X, and 30X, respectively). **(B)** Each of the subplots indicates the speed (#Base/s) of the aligners (in 8 CPU threads) on the simulated human datasets in a specific error model. The datasets in various sequencing depths are shown separately, and the bars in different colors refer to various aligners. Also refer to Supplementary Table 6 for a more comprehensive assessment of the alignment speeds in various numbers of CPU threads (1, 4, 8, and 16 CPU threads).

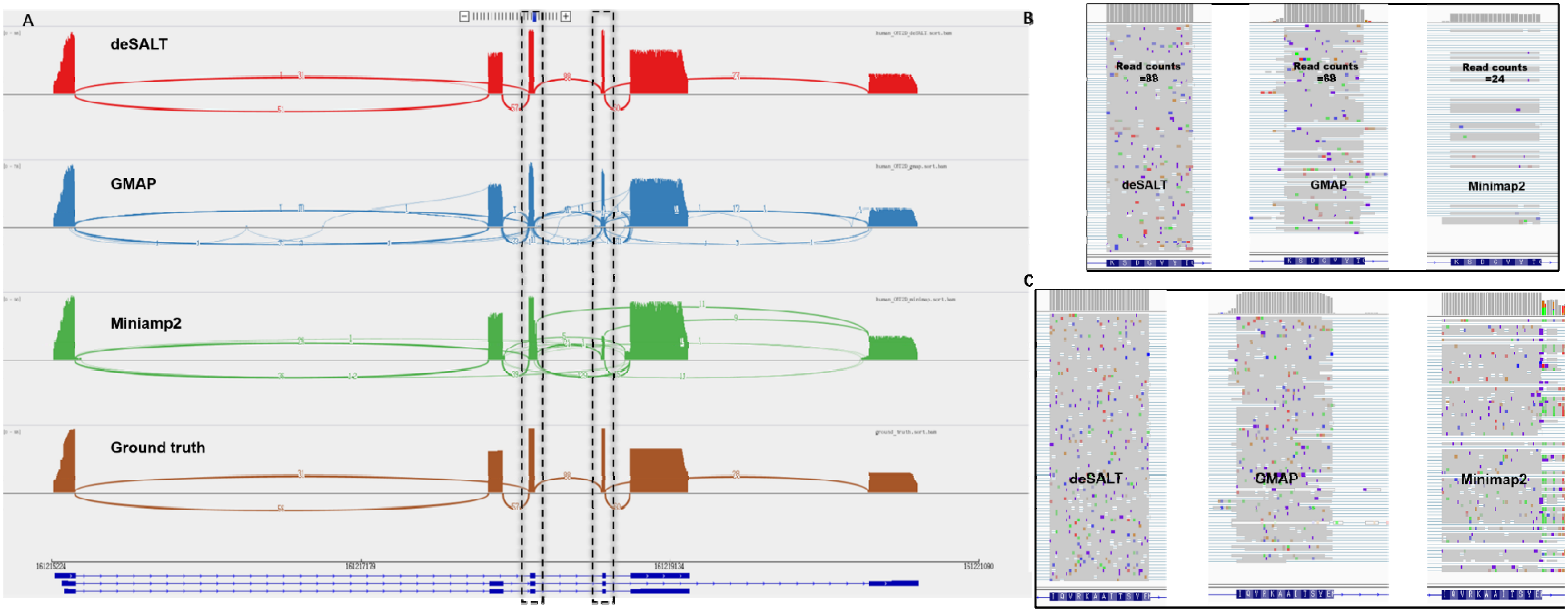

Figure 3. An example of the alignments of simulated reads by various aligners

This figure represents the snapshots of the alignments of the reads from the simulated 30X ONT 2D (1D^2) human dataset around the FCER1G gene ((161215297–161219248) of GRCh38. The FCER1G gene has 6 exons and 3 isoforms (according to Ensembl gene annotation). According to the ground truth, there a reads in this region. The numbers of Read100 and Read80 reads of deSALT are 84 and 88, respectively, much higher than those of GMAP (#Read100: 45 and #Read8(and Minimap2 (#Read100: 19 and #Read80: 23). This suggests that deSALT has a strong ability to produce accurate full-length alignments. The sub-figures depict se features of deSALT which result in a better yield. **(A)** The Sashimi plots represent the overall views of the alignments. Compared to the ground truth (the bottom track) observed that all three aligners have good ability to handle the multiple splicing events in the reads. However, the deSALT alignments are more homogenous (i.e., at splicing site, most of the reads have similar breakpoints, which also coincide with the ground truth. The more heterogeneous alignments of GMAP and Minimap2 are us due to some less accurate alignments at small exons and exon boundaries. **(B)** A detailed view at the fourth exon of the FCER1G gene (length: 21 bp). deSALT correctly a all of the 88 reads spanning this exon; however, the corresponding numbers for GMAP (68) and Minimap2 (24) are lower. This indicates that the two-pass approach of de: has better ability to handle small exons. **(C)** A detailed view at the third exon of the FCER1G gene (length: 36 bp). It is observed that the reads have nearly the : breakpoints with the homogeneous alignments of deSALT. However, for the other two aligners, the breakpoints of the reads are more divergent from each other, and sor them are less accurate, which could be due to the effect of sequencing errors as well as to the nearby small exons.

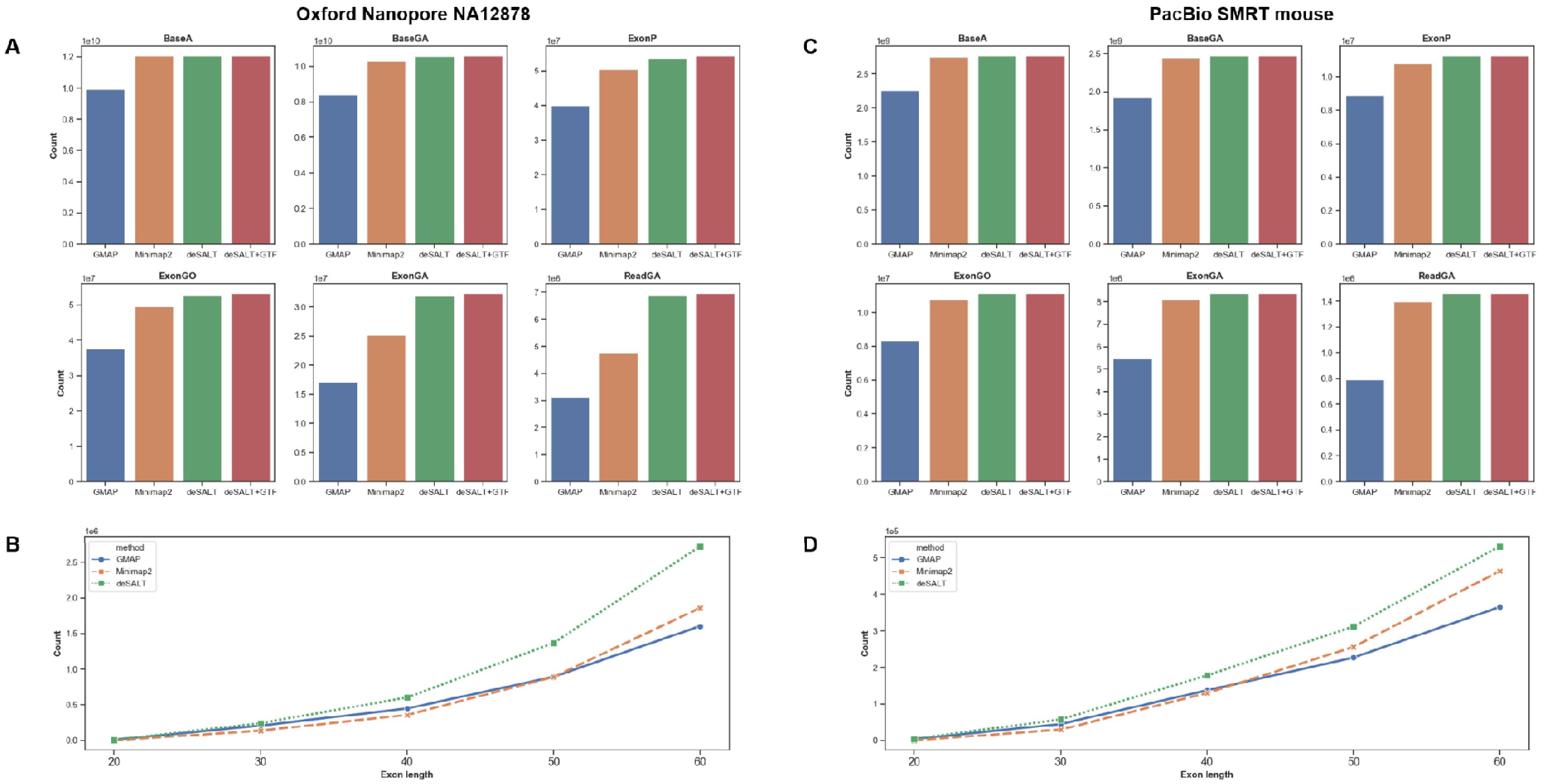

Figure 4. Results on real datasets

The figure depicts the yields of the aligners and the sensitivity of the aligners on short exons. Sub-figures (A) and (C) indicate the six metrics (BaseA, BaseGA, ExonP, Exc ExonGA, and ReadGA, respectively) of the aligners on the human ONT dataset (A) and the mouse PacBio dataset (C). Each bar in a subplot indicates the result of a sp aligner. Sub-figures (B) and (D) indicate the ExonGA(x) metrics of the aligners on the human ONT dataset (B) and the mouse PacBio dataset (D). More precisel ExonGA(20), ExonGA(30), ExonGA(40), ExonGA(50), and ExonGA(60) of the aligners are shown in sub-figures (B) and (D), depicting the sensitivities of the aligne relatively short (i.e., up to 60 bp) exons.

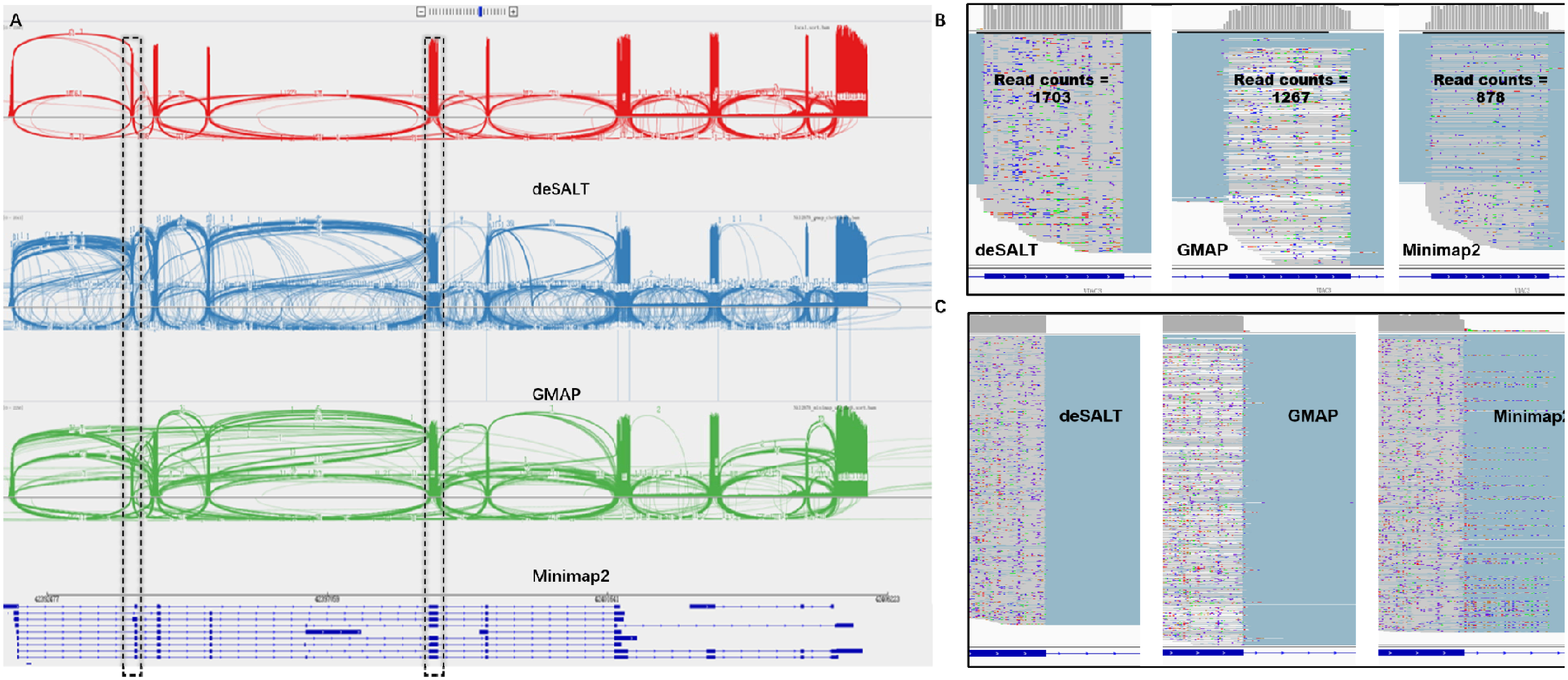

Figure 5. An example of the alignments of real sequencing reads by various aligners

This figure represents the snapshots of the alignments of the reads from the human ONT dataset around the VDAC3 gene (Chr8: 42391761–42405937) of reference GRC The VDAC3 gene has 10 exons and 12 isoforms (according to Ensembl gene annotation). deSALT, GMAP, and Minimap2 respectively mapped 2652, 2639, and 2462 rea this region. The ratios #BaseGA/#BaseT of the aligners are respectively 78.93% (deSALT), 74.2% (GMAP), and 72.69% (Minimap2), where #BaseT is the total numb bases aligned to the VDAC3 region by the corresponding aligner. This indicates that deSALT produces overall more accurate split alignments. Moreover, the #Rea statistics of the aligners are respectively 1630 (deSALT), 889 (GMAP), and 751 (Minimap2), also indicating that deSALT produces better full-length alignments. **(A)** The o views (sashimi plots) of the alignments indicate that deSALT produces more homogenous alignments, just as it does in the case of simulated reads. Considering the h #BaseGA/# BaseT and #ReadGA statistics, such alignments could be more plausible. **(B)** A detailed view at the second exon of the VDAC3 gene (exon length: 40 bp). de aligns much more (i.e., 1703 reads) to this short exon than GMAP (1267 reads) and Minimap2 (878 reads); moreover, the proportion of the matched bases also coincides the common sequencing error rate of the ONT datasets. This indicates that deSALT potentially handles this exon better. **(C)** A detailed view at the 3' splicing site of the fifth of the VDAC3 gene (exon length: 153 bp). It is obvious that the alignments of deSALT near the splicing site are more homogeneous, and the breakpoints of the reads coi with the annotation. However, the alignments of GMAP and Minimap2 are more heterogeneous and less accurate.