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

Bo Liu[†], Yadong Liu[†], Tianyi Zang[*] & Yadong Wang[*]

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

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

[†] These authors contributed equally to this work.

[*] To whom correspondence should be addressed.

ABSTRACT

Long-read RNA sequencing (RNA-seq) is promising to transcriptomics studies, however, the alignment of the reads is still a fundamental but non-trivial task due to the sequencing errors and complicated gene structures. We propose deSALT, a tailored two-pass long RNA-seq read alignment approach, which constructs graph-based alignment skeletons to sensitively infer exons, and use them to generate spliced reference sequence to produce refined alignments. deSALT addresses several difficult issues, such as small exons, serious sequencing errors and consensus spliced alignment. Benchmarks demonstrate that this approach has a better ability to produce high-quality full-length alignments, which has enormous potentials to transcriptomics studies.

INTRODUCTION

RNA sequencing (RNA-seq) has become a fundamental approach to characterize transcriptomes. It enables to reveal precise gene structures and quantify gene/transcript expressions[1-5], as well as many other kinds of applications such as variant calling [6], RNA editing analysis [7, 8], gene fusion detection [9, 10]. However, due to the drawbacks of short read sequencing technologies, such as the limited read length and the systematic bias from library preparation, it is still non-trivial to accurately align reads [11], reconstruct gene isoforms [12], and quantify transcript-level expressions [5]. This has become a bottleneck to 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 transcriptome analysis. Both of them enable to produce much longer reads, i.e., the mean and maximum lengths of the reads have been over tens and a few hundreds of thousands of basepairs (bps) [15, 16], respectively. Taking this advantage, full length transcripts can be sequenced by single reads, which is promising to substantially improve the accuracy of gene isoform reconstruction. Moreover, there is also less systematic bias in the sequencing procedure [17], which is also beneficial to gene/transcript expression quantification.

Besides the 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%-20% [16]; and 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 also produce reads of interest (ROIs) by sequencing circular fragments multiple times to largely reduce sequencing errors, however, this technology also has much lower sequencing yields and reduced read lengths. The serious sequencing errors raise new technical challenges to RNA-seq data analysis. Read alignment could be the most affected one, and the affection could be not limited to 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, that many technical issues need to be well handled, such as the serious sequencing errors, potential genome variants and large read length. 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 has a strong ability to implement highly complicated split alignment (also called “spliced alignment”) to correctly recognize many splicing junctions and map the bases to corresponding exons. Although most of proposed DNA-seq long read alignment approaches have the ability to implement split alignment to handle genome structure variations (SVs)[21-23], splicing

junctions occur much more frequently and the lengths of exons are much shorter and divergent, so that tailored algorithms are still on demand.

There have been a couple of 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-extend strategy, and various seeding and extension methods are implemented to address the issues in RNA-seq long read alignment. All of these approaches have the ability to handle splicing junctions. However, most of these algorithms have relatively slow speed [28], mainly due to numerous short matches in the seeding step and time-consuming local alignment in the extension step. Moreover, some of the algorithms also have lower sensitivity [29], i.e., many reads are unaligned or only partially aligned, which is due to their relatively poor ability to handle sequencing errors. An outstanding algorithm is Minimap2, which simultaneously achieve tens of times faster speed and similar or higher sensitivity than other state-of-the-art aligners. This mainly benefits from its well-designed minimizer-based indexing [30] and SSE-based local alignment methods [31], which greatly improve the efficiency of seeding and extension steps, moreover, its specifically designed local extension method is suited to handle splicing junctions.

In absolute terms, the ultimate goal of the task should be to correctly map all the bases for all the reads, however, this could be still non-trivial to state-of-the-art aligners in several aspects. One is the alignment of the bases from relatively short exons, e.g., exons in only a few tens of bps. It is extremely hard to find seeds in the read parts from such short exons under the circumstance 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 correctly align the bases nearby the splicing junctions. 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 affection of sequencing errors, the alignments of the reads from the same gene isoform are usually divergent to each other, which is also misleading to downstream analysis.

Herein, we propose de Bruijn graph-based Spliced Aligner for Long Transcriptome read (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 de Bruijn graph-based index. It has the ability to well-handle the complicated gene structures as well as serious sequencing errors, to produce more sensitive, accurate and consensus alignments. For most of the reads, deSALT can produce full-length alignments to thoroughly recover the exons and splicing junctions along the entire reads. Moreover, the speed of deSALT is also faster or comparable to state-of-the-art approaches as well. We believe that it has the potential to play important roles in many forthcoming transcriptomic studies.

RESULTS

Motivation and overview of deSALT approach

Seed-and-extension approach is suited to spliced alignment since it can match the short tokens of the read to its spanning exons at first (i.e., seeding), and then implement base-level alignment between the read and the matched exons (i.e., extension). However, under the circumstance of frequent splicing events and serious sequencing errors, this task is non-trivial in practice. For a single read, it is usually difficult to accurately find matches between the read and all its spanning exons, especially for the read parts with serious sequencing errors and relatively short exons. Thus, it is hard to compose a local reference sequence containing all the spanning exons of the read, so that 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, both of the seeding and the extension phases could make various mistakes for the multiple reads from the same gene isoform, and the produced alignments of the reads are divergent to 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 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 to alignment skeletons (referred to “alignment skeleton generation” step). All the alignment skeletons of all the reads are then integrated to comprehensively detect the exon regions (referred to “exon inference” step). In the second pass, deSALT re-locates the short matches between the read and the detected exons to compose a local spliced reference sequence (LSRS), which is expected to be concatenation of all the spanning exons of the read, and the read is aligned against the LSRS to produce refined base-level alignment (referred to “refined alignment” step).

The key point of deSALT is its comprehensive analysis on the local matches of all the reads through the generation and integration of the alignment skeletons. Since the sequencing errors are random[17], each of the alignment skeletons contains some distinct as well as complement information about exon regions, like puzzle pieces, and the integration of them can effectively filter the sequencing errors to implement a sensitive and noise-robust detection of exons. The detected exons then help in the later step 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 to effectively infer all the spanning exons of a given read and compose high-quality spliced reference sequence to produce accurate full-length alignment. This approach is robust to very short exons (e.g., exons < 30bp), frequent splicing events, potential single nucleotide variants (SNVs) and small indels, as well as sequencing errors. Furthermore, deSALT generates homogeneous LSRSs for the reads from

the same gene isoform with the integrated information, which enables to produce highly consensus alignments for them.

deSALT also has fast speed with its tailored design. Not like conventional two-pass alignment approaches[26, 32] that both of the two passes produce base-level alignment, deSALT only uses pseudo-alignment (alignment skeleton) in the first pass, whose operation 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, e.g, graph-index-based skeleton generation and SIMD-based local alignment[31], also help to accelerate the speed.

Results on simulated datasets

We simulated 36 RNA-seq long read datasets in various sequencing error rates and read lengths (Supplementary Table 1) to mimic the datasets from main stream 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 are 9 datasets respectively from 3 species (human, mouse and fruitfly) and in 3 sequencing depths (4X, 10X and 30X). All the datasets were produced by PBSim[34] based on Ensembl gene annotations[35] (human: GRCh38, version 94, mouse: GRCm38, version 94 and fruitfly: BDGP6, version 94), and the error models are 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 implemented on all the datasets for comparison. Refers to Methods section for more details on the implementation of the benchmarking.

The following five metrics were used to assess the sensitivity, accuracy and performance of the aligners, respectively.

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

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

Read80%: the proportion of Read80% reads. A read is considered as a Read80% read only if it meets two conditions that $N_T/N_G > 80\%$ 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 being considered as a true positive exon only if there is a ground truth exon in the read, that the distance between the corresponding boundaries of the predicted exon and the ground truth exon are within 5 bp.

Read100%: the proportion of Read100% reads. A read is considered as a Read100% read only if it meets two conditions that $N_T/N_G = 100\%$ and $N_T/N_P = 100\%$. It is worthnoting that a Read100% read indicates that the read has a highly correct full-length alignment.

#Bases/s: the 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.

Results on the thirty-six simulated datasets are in Figure 2 and Supplementary Tables 2-6. Mainly, four issues are observed as following.

1) deSALT has outstanding alignment yields.

deSALT has overall highest base% statistics, indicating that it mapped more bases to their correct positions. Especially, the advantage of deSALT is more obvious on the datasets in medium and high error rates (i.e., ONT 2D reads, PacBio subreads, and ONT 1D reads), which is preferable to handle real noisy long reads. Moreover, deSALT has larger advantages on exon% statistics, suggesting that it has stronger ability to recover the exons and splicing junctions within the reads.

deSALT also obviously outperforms other state-of-the-art aligners on Read80% and Read100% statistics, indicating that it has better ability to produce full length alignments. Especially, deSALT produces highest number of Read100% reads for all the datasets, i.e., for more reads deSALT can correctly aligns all their exons without introducing false positives. Such accurate full-length read alignments are valuable to downstream analysis.

In absolute terms, deSALT correctly aligns most of the bases as well as exons for the reads in 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%) is to some extent affected, although it outperforms. The most affected ones are the low coverage (4X) mammalian (human and mouse) datasets. This is mainly due to that the bases nearby exon boundaries are very difficult to confidently align under the circumstance of the serious noise. However, it is worthnoting that the yield on these error-prone reads improves with the increase of read depth. This is a feature of the two-pass approach, that the affection of sequencing errors can be better mitigated and the exon detection 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 highly consensus alignment. This is also an advantage of the two-pass alignment, that deSALT likes to compose homogeneous LSRSs for the reads from same gene isoforms, which is beneficial to simultaneously aligns them to correct positions. However, one-pass approaches are easier to be affected by sequencing errors as well as other factors such as very small exons and frequent splicing events, which usually produces more heterogeneous alignments with more mistakes.

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

A typical example describing the characteristics of deSALT is in Figure 3.

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. This derives 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. It helps to compose high quality LSRSs to correctly align the read parts spanning those short exons. However, other state-of-the-art aligners using one-pass strategy are more likely to be affected by the splicing events and sequencing errors, which results in reduced ability to find local matches on short exons and the corresponding read parts are mistakenly aligned. Two examples are 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 for state-of-the-art aligners to align reads having many splicing events and/or from the genes having multiple isoforms [11]. We assessed the alignment of the reads from the transcripts having various number of exons (2-5 exons, 6-9 exons and >9 exons). deSALT can produce equally good alignment for all the three read groups (Supplementary Table 4), indicating that it enables to handle the numerous splicing events within the reads (an example is in Supplementary Figure 2). Minimap2 has similar trend, but its Read80% and Read100% statistics are lower. GMAP has significant decreases on Read80% and Read100% statistics with the increase of exon numbers, indicating that it could be not good at handling the reads having many splicing events. We also assessed the alignment of the reads from the genes having multiple isoforms. The results of deSALT (Supplementary Table 5) demonstrate that there is no significant difference to that of the genes having single isoforms, suggesting that it has the ability to handle genes having multiple isoforms.

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

deSALT supports to input gene annotations to facilitate read alignment. Mainly, it combines the annotated exons with the alignment skeletons, to build 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 due to that, for most of 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 read parts (an example is in

Supplementary Figure 3) and is beneficial to produce full-length alignments (see the improvements on Read80% and Read100% metrics). With this feature, deSALT provides the opportunity to better use noisy long reads, which has many potentials in transcriptomics studies.

Overall, the simulation results demonstrate that deSALT is able to simultaneously achieve excellent sensitivity, accuracy and performance. Especially, it has the ability to address many difficult issues, such as serious sequencing errors, short exons, numerous splicing events, multiple isoforms, etc., which is promising 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 is from a well-studied CEPH sample (NA12878) produced by ONT platform (available at: https://github.com/nanopore-wgs-consortium/NA12878, containing 15152101 reads and 14134831170 bases in total), and the other one is from a mouse sample produced by PacBio platform [37] (Accession Number: SRR6238555, containing 2269795 reads and 3213849871 bases in total).

Due to lack of ground truth, we use a series of metrics based on gene annotations to evaluate the alignments, which are defined as following.

#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 as “predicted exons”). Here, the predicted exons in various reads are independently considered.

#ExonGO: the number of predicted exons being overlapped by annotated exons (also termed as “overlapped exons”). Herein, a predicted exon is considered as overlapped by annotated exons only if there is at least one annotated exon, that there is at least 10 pb overlapping between the predicted exon and the annotated exon.

#ExonGA: the number of predicted exons being exactly matched by annotated exons (also termed as “exactly matched exons”). Herein, a predicted exon is considered as exactly matched by annotated exons, only if there is an annotated exon, that the distance between the corresponding boundaries of the predicted exon and the annotated exon is within 5 bp.

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

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

Ensembl gene annotations (human: GRCh38, version 94 and mouse: GRCm38, version 94) are employed for the assessment.

The results are in Figure 4 and Supplementary Tables 7 and 8. Mainly, four issues are observed.

1) deSALT still has the best alignment yields.

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

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

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

Like that of simulated reads, deSALT also shows 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 deSALT enables to recover higher numbers of short exons for both of the two datasets. It is worthnoting that although only a small proportion of exons are short, they are important to study gene splicing, so that it is of great value to correctly align such read parts. However, this is still a difficult task to other state-of-the-art aligners. Furthermore, this advantage helps deSALT to produce better full-length alignment for the reads from the genes having small exons (an example is in Supplementary Figure 4), and achieve overall higher #ReadGA statistics.

3) deSALT produces more consensus alignment.

Another outstanding ability of deSALT is to produce more consensus alignment. It is 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 to each

other. An example is shown in Supplementary Figure 5, that the more consensus alignment of deSALT could be overall more accurate, especially for those bases nearby exon boundaries. The consensus alignments are also more useful to study splicing events since there is less noise than that of ambiguous alignments.

4) A proportion of bases are aligned to unannotated regions

According to Ensembl gene annotations, it is observed that about 10% of the bases are aligned by deSALT to the regions other than annotated exons: 1) a proportion of the bases (5.60% for the human ONT dataset and 5.13% for the mouse PacBio dataset) are 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) are aligned to intergenic regions. Minimap2 also has similar proportions of bases aligned to such regions as well. We found that the alignments of these read parts are highly clustered, i.e., in most cases, there are multiple reads aligned in a local region, indicating that there could be unannotated exons (in intragenic regions) or novel transcripts (in intergenic regions). Moreover, we compared the detailed alignments of deSALT and Minimap2, and found that they have similar outputs for these read parts, which also partially suggests that the alignment of deSALT is 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 to 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 that it is on wide demand to develop more advanced read alignment algorithm to breakthrough this bottleneck. Herein, we proposed deSALT, a novel read alignment algorithm using de Bruijn graph-based index and tailored two-pass strategy as a solution to this important open problem. Mainly, we show that how to build and integrate spliced alignment skeletons to handle sequencing errors and complex gene structures to generate high quality spliced reference sequences, and use them to produce highly accurate and consensus full-length alignments for long RNA-seq reads. To our knowledge, deSALT is the first long RNA-seq read alignment approach fully considering the intermediate results of all the reads and taking this advantage to produce refined spliced alignments.

On both of simulated and real datasets, the results of deSALT 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 with support of gene annotations. And its advantage on the recovery of exons and splicing junctions is more obvious, suggesting that deSALT has an excellent ability to spliced alignment. This is further demonstrated by several kinds of difficult scenarios, such as the alignment of the reads

having very short exons, having numerous splicing events and/or from the genes having multiple isoforms.

A more important feature of deSALT is its outstanding ability to produce accurate and consensus full-length alignment. With the ever-increasing read length, this feature is on widely demand since it provides the opportunity to directly investigate gene structures. However, this requires the employed aligner to well-handle many technical issues simultaneously. deSALT largely improved full-length alignment by using several key techniques, such as the sensitive exon detection, local exon matching and LSRS generation. For much more reads, deSALT can comprehensively and accurately recover their splicing junctions by single alignments, and the produced alignments are more consensus 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 [38], however, no successful alignment is produced (data not shown). In this situation, we realized that these clipped read parts could have extremely low quality.

Other than only using reference genome, deSALT supports to use gene annotations to enhance the alignment. However, the benchmarking results are to some extent unexpected that there is no significant difference between the alignment with and without gene annotations, only except for low depth high error rate (ONT 1D) datasets. This is also reasonable since the two-pass strategy has strong ability to mitigate the affection of moderate sequencing errors even if read depth is low. Moreover, this ability can be further enhanced with the increase of 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 its own contributions to align those reads.

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

METHODS

Steps of deSALT

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

following.

1) Alignment skeleton generation (first-pass alignment): for each of the reads, deSALT uses RdBG-index [33] to find maximal exact matches between the unitigs of Reference de Buijn Graph (RdBG) and the read (termed as U-MEMs), and build one or more alignment skeletons in a sparse dynamic programming (SDP) approach.

2) Exon inference: deSALT maps all the alignment skeletons to the reference, and infer potential exons from the projections of the skeletons. A local sequence-based scoring system[39] 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 find additional local matches to 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 LSRS and implement 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 match them to the unitigs of RdBG with 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 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 pre-defined maximum intron length, $\mathrm{T}_{\mathrm{intron}}$ (default value: $\mathrm{T}_{\mathrm{intron}}$=200,000 bp). A weight is assigned to each of the edges based on the sizes of the two corresponding MBs and their distances (Supplementary Notes). A sparse dynamic programming (SDP) approach is then used to find the path having largest sum weight as the alignment skeleton. It is also worthnoting 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.

Also refer to Section 1.1-1.3 of Supplementary Notes of more implementation details.

Exon inference

deSALT maps all the alignment skeletons to reference genome and uses a set of pre-defined rules (Section 2.1 of Supplementary Notes) to iteratively combine the genomic regions covered by alignment skeletons from upstream to downstream. It is optional to introduce additional gene annotation file (in GTF format) into this process. deSALT treats known gene isoforms as a special kind of alignment skeletons, and also maps them to reference genome, so that the genomic regions covered by known gene isoforms and 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 lengths and too low coverages are then filtered out.

A local sequence-based scoring system [39] 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 pre-defined acceptor and donor scoring matrixes to score each of the positions in the upstream and downstream regions, respectively. The positions with highest scores in the two regions are respectively recognized as acceptor and donor splicing sites, and the region in between is determined as a refined exon.

Refined alignment (second-pass alignment)

Refined alignment is mainly implemented in two sub-steps as following.

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 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 as “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 part to its corresponding LSRS, using a SIMD-based implementation [31] of semi-global alignment (Section 3.2 of Supplementary Notes). Furthermore, deSALT checks if there are large deletion(s) in the CIGAR information. If so, deSALT removes the corresponding deletion part(s) in the LSRS, and re-align the read with the updated LSRS. This process is helpful to handle exons having 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 worthnoting that, for the reads having multiple alignment skeletons, deSALT separately processes each of the skeletons, and possibly produce multiple alignments for one read. In this situation, deSALT chooses the alignments having highest score as primary alignment, and outputs other alignments as secondary alignments.

Implementation of 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 is simulated in 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 are scanned, and three sets of genes were extracted. Each of the sets corresponds to a specific type, i.e., the genes having single splicing isoforms, the genes having multiple splicing isoforms and genes having short exons (<31 bp), respectively.

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

3) For each organism, four sequencing error models are 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 are configured by referring to previous studies [17, 36]. And for each model, three datasets in various depths (4X, 10X, 30X) are simulated. Thus, there are totally $3 \times 4 \times 3 = 36$ datasets generated. The availability of the simulated datasets is in Supplementary Notes.

Implementation of real data benchmark

The benchmarks were implemented with the same hardware environment to that of simulation benchmark. Two real datasets respectively produced by ONT and PacBio platforms are used. The ONT dataset is from NA12878 sample, which was sequenced by ONT MinION sequencer 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 in Supplementary Notes.

DECLARATIONS

Availability of data and material

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

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

Competing interests

The authors declare that they have no competing interests.

Funding

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

Authors' contributions

BL and YL designed the method, BL and YL implemented the method and YL performed the analysis. BL, YL, TZ and YW wrote the manuscript. BL and YL contributed equally to this work.

Acknowledgement

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

REFERENCE

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. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.
39. 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.

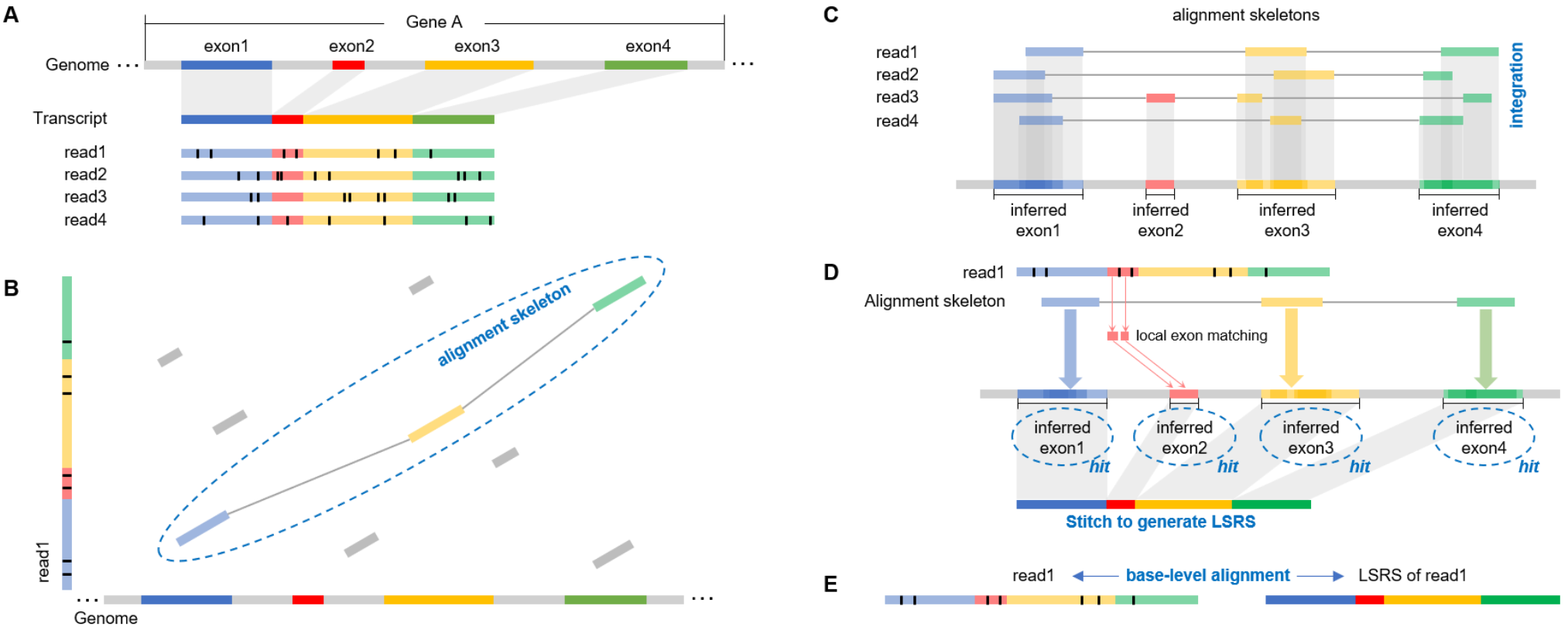

Figure 1. A schematic illustration of deSALT approach

(A) There is a gene (termed as “Gene A”) having four exons (respectively marked by blue, red, yellow and green colors, and introns are marked by grey color), and four reads that all of them sequence through the whole transcript (assuming that Gene A has only one isoform). Moreover, each of the reads have some sequencing errors (marked by the short black lines in the reads). **(B)** Alignment skeleton generation (first pass alignment): for each of the reads (read1 is employed as an example), deSALT find the MBs between it and the reference genome (marked as colored bars) and connects them to build an optimized alignment skeleton in a sparse dynamic programming (SDP) approach. **(C)** Exon inference: deSALT integrates all 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 reference genome. **(D-E)** Refined alignment (second pass alignment): for each of the reads, deSALT finds additional local matches on the exons between or nearby the exons involved in 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 stitch all them to generate LSRS. (It is shown in the figure that there are two newly found matches on exon 2 and they help to recuse this exon to build correct LSRS.) The read is then aligned with LSRS to produce refined alignment.

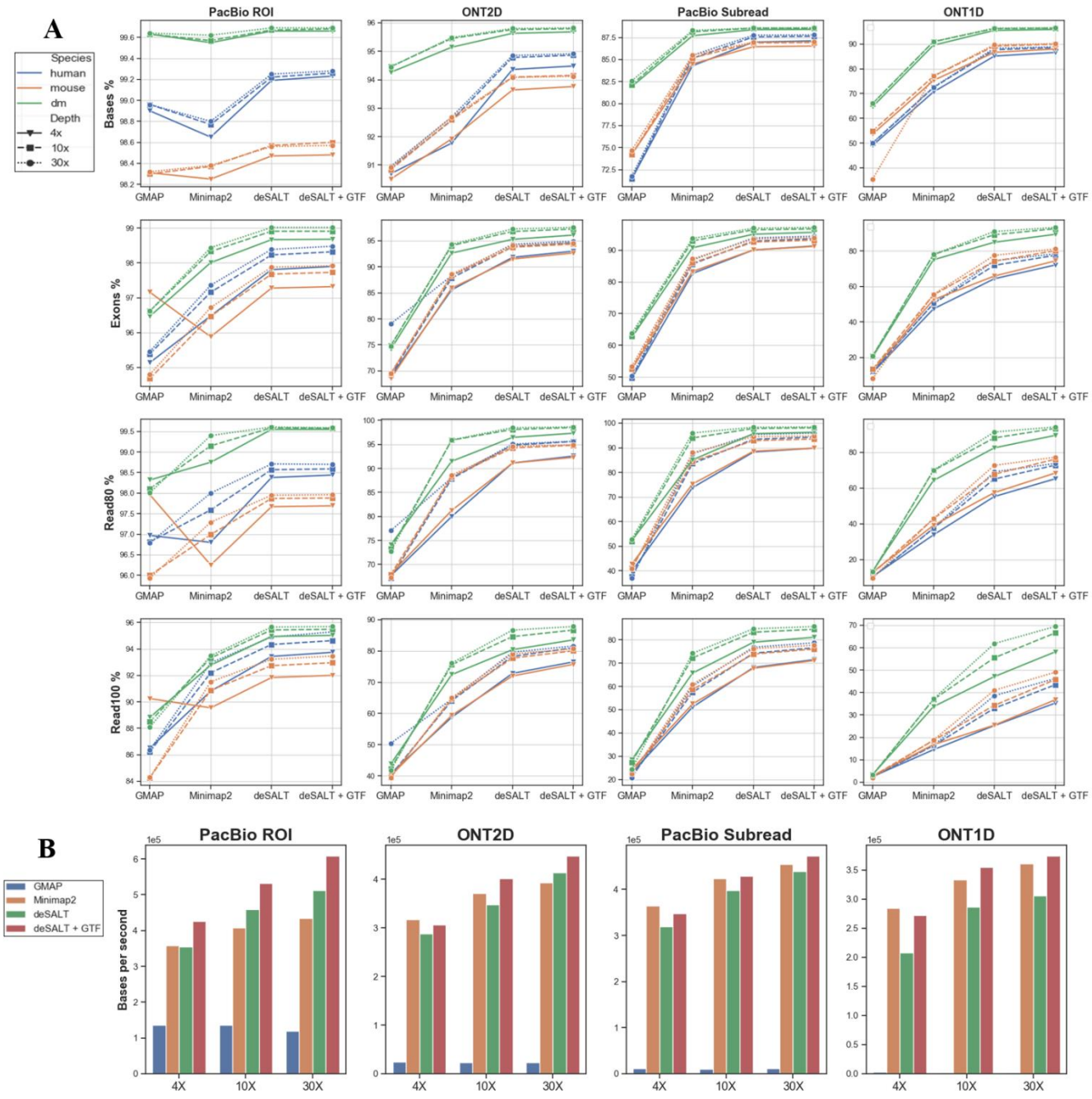

Figure 2. Results on simulated datasets

The figure depicts the yields (A) and speed (B) of the aligners on the simulated datasets, worthnoting that both of the results of deSALT without and with gene annotations are shown (indicated as “deSALT” and “deSALT+GTF”, respectively). **(A)** Each of the subplots indicates one of the four metrics (Base%, Exon%, Read80%, 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 fruitfly. Moreover, the shapes (reverse triangles, rectangles and rounds) indicates 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 (PacBio CCS, ONT2D, PacBio subreads and ONT1D, respectively). The datasets in various sequencing depths are separately shown, and the bars in different colors refer to various aligners. Also refer to Supplementary Table 6 for more comprehensive assessment of the alignment speeds in various numbers of CPU threads (1, 4, 8 and 16 CPU threads).

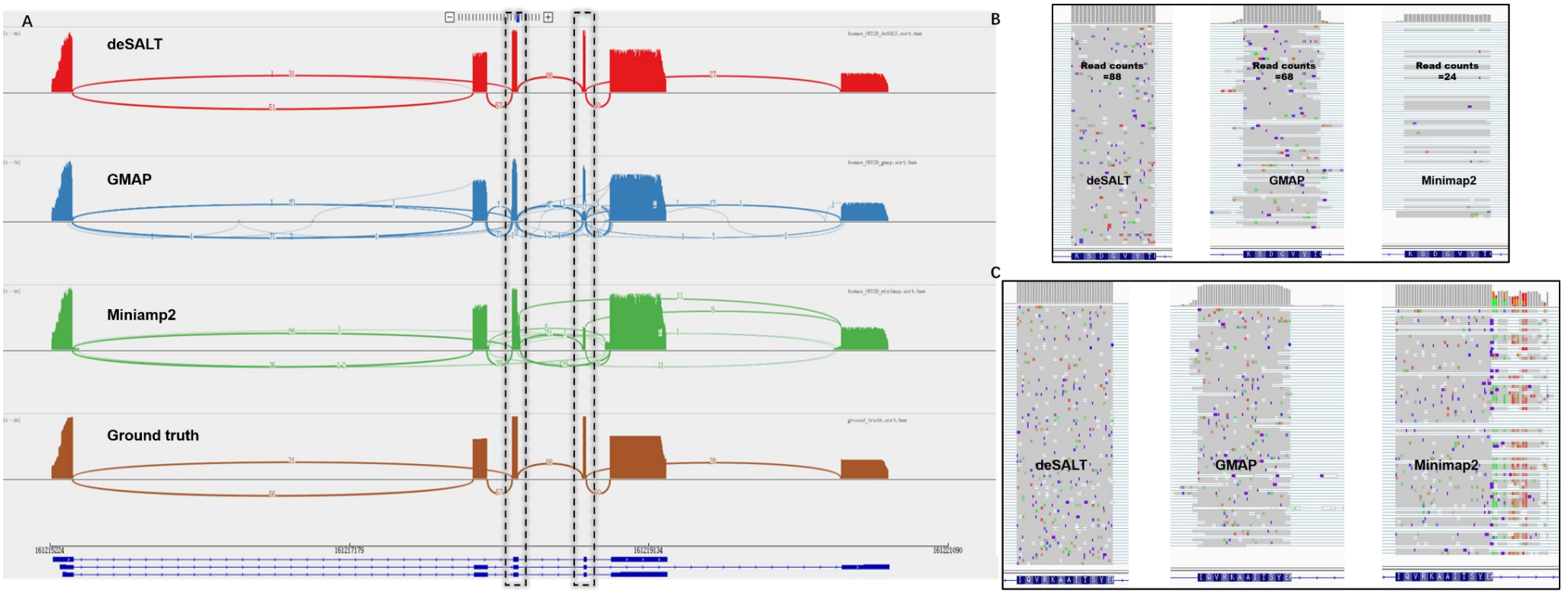

Figure 3. An example of the alignments of simulated reads by various aligners

This figure represents the snapshots of the alignments of the reads from the simulated 30X ONT 2D ($1D^2$) human dataset around FCER1G gene (Chr1: 161215297 - 161219248) of reference GRCh38. FCER1G gene has 6 exons and 3 isoforms (according to Ensembl gene annotation). According to the ground truth of the dataset, there are totally 88 reads in this region. In this region, the numbers of Read100 and Read80 reads of deSALT are 84 and 88, respectively, which are much higher than that of GMAP (#Read100: 45 and #Read80: 54) and Minimap2 (#Read100: 19and #Read80: 23). This suggests that deSALT has strong ability to produce accurate full-length alignments. The subfigures depict several features of deSALT which results in the better yield. **(A)** The Sashimi plots of the three aligners represent the overall views of their alignments. Comparing to the ground truth (the bottom track), it is observed that all the three aligners have good ability to produce split-alignments to handle the multiple splicing events in the reads. However, the alignments of deSALT are more consensus, i.e., at each splicing site, most of the reads have highly similar and correct breakpoints, and overall these consensus alignments coincide with the ground truth better. The more heterogenous alignments of GMAP and Minimap2 are usually due to some less accurate alignments at small exons and exon boundaries, which are more precisely depicted in subfigures (B) and (C). **(B)** A detailed view at the 4th exon of FCER1G gene (length: 21bp). deSALT correctly aligns all the 88 reads spanning this exon, however, the corresponding numbers of GMAP (68) and Minimap2 (24) are lower. This indicates that the two-pass approach of deSALT has better ability to handle small exons. **(C)** A detailed view at the 3rd exon of FCER1G gene (length: 36bp). It is observed that the reads have nearly the same breakpoints with the alignments of deSALT. However, for that of the other two aligners, the breakpoints of the reads are more divergent to each other and some of them are less accurate, which could be due to the affection of serious sequencing errors as well as the nearby small exons.

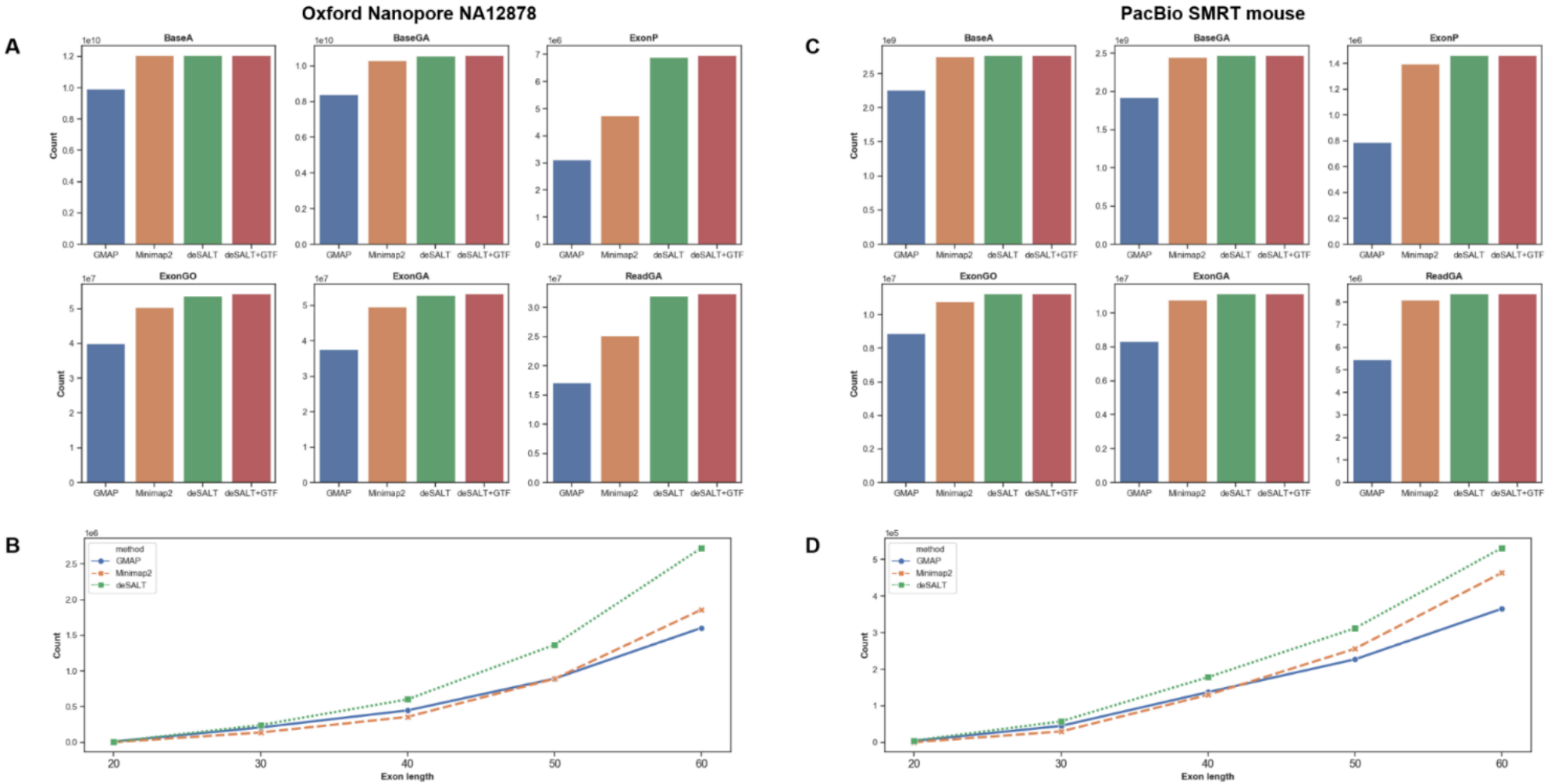

Figure 4. Results on real datasets.

The figure depicts the yields of the aligners (subfigure A and C), and the sensitivity of the aligners on short exons (subfigure B and D). Subfigures (A) and (C) indicate the six metrics (BaseA, BaseGA, ExonP, ExonGO, ExonGA, 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 specific aligner. Subfigures (B) and (D) indicate the ExonGA(x) metrics of the aligners on the human ONT dataset (B) and the mouse PacBio dataset (D). More precisely, ExonGA(20), ExonGA(30), ExonGA(40), ExonGA(50) and ExonGA(60) of the aligners are shown in subfigures (B) and (D), depicting the sensitivities of the aligners for relatively short (i.e., up to 60 bp) exons.

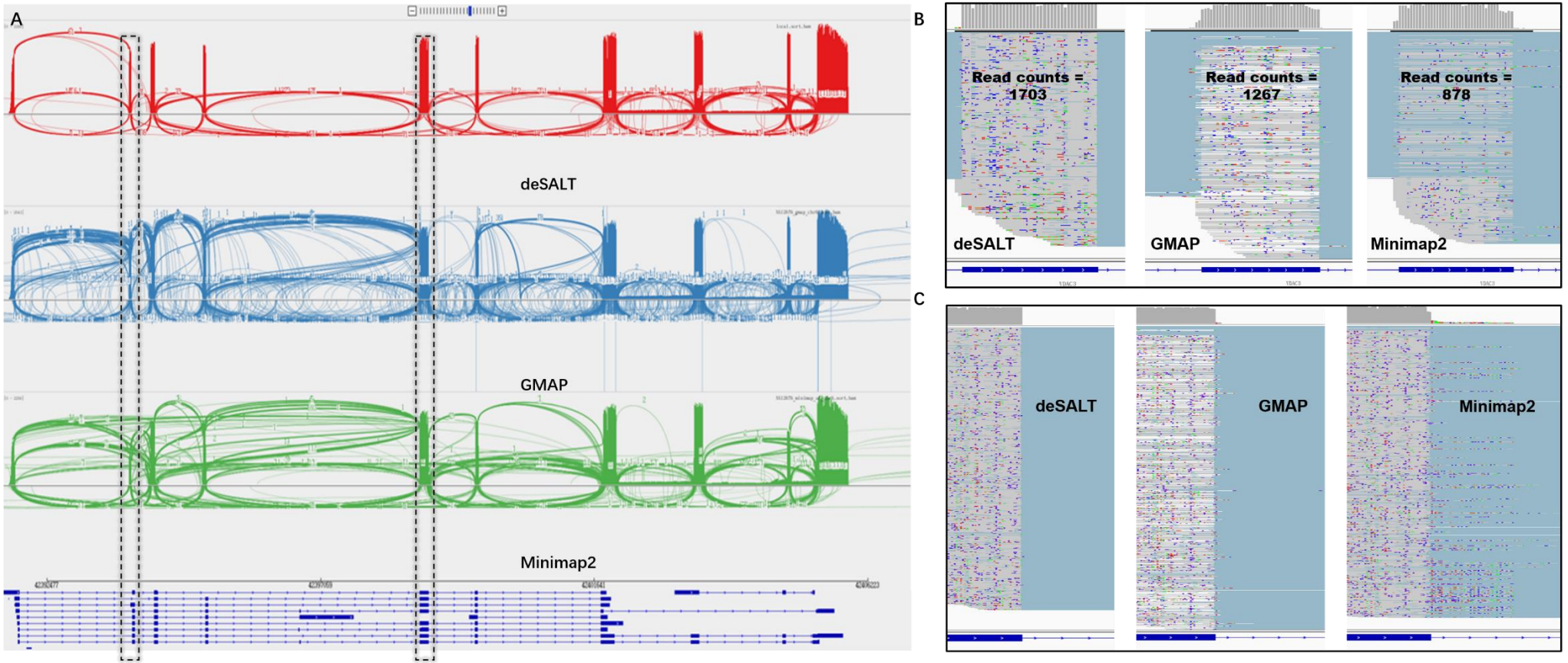

Figure 5. An example of the alignments of real sequencing reads by various aligners

This figure represents the snapshots of the alignments of the reads from the human ONT dataset around VDAC3 gene (Chr8: 42391761-42405937) of reference GRCh38. VDAC3 gene has 10 exons and 12 isoforms (according to Ensembl gene annotation). deSALT, GMAP and Minimap2 respectively mapped 2652, 2639 and 2462 reads to 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 number of bases aligned to VDAC3 region by the corresponding aligner. This indicates that deSALT produces overall more accurate split alignments. Moreover, the #ReadGA metrics of the aligners are respectively 1630 (deSALT), 889 (GMAP) and 751 (Minimap2), also indicating that deSALT produces better full-length alignments. **(A)** The overall views (sashimi plots) of the alignments indicate that deSALT produces more consensus alignments, like that of simulated reads. Considering the higher #BaseGA/# ratio and #ReadGA metrics, such alignments could be more plausible. **(B)** A detailed view at the 2nd exon of VDAC3 gene (exon length: 40 bp). deSALT aligns much more (i.e., 1703 reads) to this short exon than that of the GAMP (1267 reads) and Minimap2 (878 reads), moreover, the proportion of the matched bases also coincide with the common sequencing error rate of ONT datasets. This indicates that deSALT potentially handles this exon better. **(C)** A detailed view at the 3' splicing site of 5th exon of VDAC3 gene (exon length: 153 bp). It is obvious that the alignments of deSALT near the splicing site is highly consensus, and the breakpoints of the reads coincide with the annotation. However, the alignments of GMAP and Minimap2 are more heterogenous and to less accurate.