Pertanika J. Sci. & Technol. 30 (4): 2727 - 2745 (2022)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

A Comparative Study of RNA-Seq Aligners Reveals Novoalign's Default Setting as an Optimal Setting for the Alignment of HeLa RNA-Seq Reads

Kristine Sandra Pey Adum and Hasni Arsad*

Integrative Medicine Cluster, Advanced Medicine and Dental Institute, Universiti Sains Malaysia, 13200 USM, Kepala Batas, Penang, Malaysia

ABSTRACT

The introduction of RNA-sequencing (RNA-Seq) technology into biological research has encouraged bioinformatics developers to build various analysis pipelines. The chosen bioinformatics pipeline mostly depends on the research goals and organisms of interest because a single pipeline may not be optimal for all cases. As the first step in most pipelines, alignment has become a crucial step that will affect the downstream analysis. Each alignment tool has its default and parameter settings to maximise the output. However, this poses great challenges for the researchers as they need to determine the alignment tool most compatible with the correct settings to analyse their samples accurately and efficiently. Therefore, in this study, the duplication of real data of the HeLa RNA-seq was used to evaluate the effects of data qualities on four commonly used RNA-Seq tools: HISAT2, Novoalign, TopHat and Subread. Furthermore, these data were also used to evaluate the optimal settings of each aligner for our sample. These tools' performances, precision, recall, F-measure, false discovery rate, error tolerance, parameter stability, runtime and memory requirements were measured. Our results showed significant differences between the settings of each alignment tool tested. Subread and TopHat exhibited the best performance when using optimised parameters setting. In contrast, the most reliable performance was observed for HISAT2 and Novoalign when the default setting was used. Although HISAT2 was the fastest alignment tool, the highest accuracy was achieved using Novoalign with the default setting.

Keywords: Alignment, HISAT2, novoalign, RNA-seq, subread, TopHat

ARTICLE INFO

Article history:
Received: 29 October 2021
Accepted: 14 February 2022
Published: 23 September 2022

DOI: https://doi.org/10.47836/pjst.30.4.24

E-mail addresses:
kristinesandra@gmail.com (Kristine Sandra Pey Adum)
hasniarsad@usm.my (Hasni Arsad)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

Next-generation sequencing (NGS) is a fast-growing technology that can fulfil efficient and highly sensitive sequencing demands. In contrast with the previous sequencing technologies, such as Sanger sequencing, NGS is much cheaper and faster (Križanović et al., 2018). Gaur and Chaturvedi (2017) stated that RNA-sequencing (RNA-Seq) is a powerful technique that enhances the understanding of complex transcriptomes by revealing insights into many biological phenomena, such as the underlying mechanisms and pathways of biological processes. Other applications of NGS also include whole-genome sequencing, followed by genetic variant detection in the whole-genome or the targeted region (Qin, 2019). However, the features and massive volume of NGS reads require the development of a new generation of computational algorithms and analysis pipelines equipped to handle such data (Koboldt, 2020).

Many researchers have developed more than 60 different algorithms for the sequence reads alignment to a reference genome tool, depending on various ranges of capabilities (Fonseca et al., 2012; Keel & Snelling, 2018; Schaarschmidt et al., 2020). As alignment is the first step in the RNA-Seq pipeline, it will drastically affect the downstream analyses. The read alignment in the RNA-Seq experiment can be conducted with or without a reference genome, but most studies would prefer mapping to a reference genome as the results have been proven to be more reliable and more accurate in quantifying lowly-expressed or small transcripts (Wu et al., 2018). The major challenge in handling eukaryotic transcriptomes is the alignment of spliced transcripts reads to the reference genome. Apart from being computationally challenging (Sahlin & Mäkinen, 2021), spliced-alignment read-lengths caused difficulties in detecting isoforms with complicated splicing structures and limiting the quantification of isoform abundance (Zhang et al., 2017).

Accordingly, many spliced aligners have been developed to overcome this problem. Depending on their algorithms, these aligners mapped the reads crossing the splice junction differently. There are two algorithm approaches for the alignment step: hash-tables and Burrow-Wheeler Transform (BWT) algorithms. Hash table-based aligners operate by rapid seeding of alignment candidates. These are then extended or discarded by using more precise alignment algorithms. Then, the reference genome or the reads are split and stored in a hash table to search for the exact match of the seed locations. This low space requirement algorithm builds an index for the positions of sequences rather than sequences themselves.

While the hash table algorithm is praised for its low space consumption, the BWT-based algorithm, on the other hand, loses error tolerance for high-speed retrieval of correct matches. The representation of the data structures by top-down paths in a tree structure are called prefixes/suffixes. Then, a rapid read searching of the substring matching is enabled by primarily beginning at the root. The requirement of vast memory for the uncompressed tree structure is the main drawback of using these algorithms. In order to overcome this problem, the Ferragina-Manzini index (FM-index) was developed by Ferragina and Manzini (2000) to reduce the memory occupied by the prefix/suffix tree. It is a compressed yet searchable suffix array-like structure based on the Burrows-Wheeler transform (BWT) (Keel & Snelling, 2018).

The selection of a suitable alignment tool for NGS data can be challenging due to the wide range of algorithms available. Therefore, various groups of researchers carried out benchmarking analyses to guide the users in choosing the correct aligners. For instance, a comprehensive benchmarking study of common splice-aware aligners was published by Baruzzo et al. (2017). The authors revealed that the aligners' performances varied by genome complexity. Unfortunately, although many benchmarking analyses had been carried out in guiding the users in choosing the best aligners, the problem is still plaguing the bioinformatics communities, while other solutions have not been derived (Donato et al., 2021; Grytten et al., 2020; Jain et al., 2020; Schilbert et al., 2020; Thankaswamy-Kosalai et al., 2017).

Comprehensive studies on alignment had been carried out, but most were using simulated data. We aimed to evaluate a more realistic setting on real data, so we chose the human cervical cancer cell line (HeLa) dataset for this study. Liu et al. (2019) proposed that HeLa cells present an essential example of human cancer cells that have broadly influenced biological studies. Furthermore, a large number of mutations and chromosol changes in HeLa cells makes it a complex genome dynamic ecosystem of the tumour genome (Hu et al., 2019). For an aligner to be viable for RNA-Seq, it must be able to (i) align reads across splice junctions, (ii) handle paired-end reads, (iii) handle strand-specific data, and (iv) run efficiently (Baruzzo et al., 2017). Four aligners that satisfy these four requirements are HISAT2 version 2.1.0 (Kim et al., 2015), Novoalign version 4.0 (http://www.novocraft.com/products/novoalign/), Subread version 2.0.1 (Liao et al., 2013) and TopHat version 2.1.1 (Trapnell et al., 2009). Based on the algorithms, Novoalign and Subread adopt a hash table algorithm, while HISAT2 and TopHat adopt an FM-index algorithm.

In this study, we aimed to evaluate the effects of reading quality on alignment on four different aligners. Apart from that, we also targeted to compare the default and parameters settings of these aligners to obtain the optimal setting for HeLa RNA-Seq reads.

MATERIALS AND METHODS

Data Sets and Alignment Settings

In this experiment, we used two sets of paired-end, real Illumina sequencing data of human cervical cancer cells line (HeLa) treated with *C. Nutans*. In paired-end sequencing, a DNA fragment was selected and sequenced from both ends, producing high-quality data compared to only single-end sequencing. The raw sequenced data sets contained around 52.26 Mb and 53.89 Mb reads, respectively. In order to examine the performance of the aligners on the real sequencing data with varying quality, we compared the alignment before and after trimming off the low-quality bases. The trimming of raw data was processed by using the fastp trimming tool (Chen et al., 2018). We processed raw and trimmed reads using FastQC (Andrews, 2010) to evaluate the quality of the bases. The plot of the qualities suggests that the trimmed reads have better quality than the raw reads. The alignments were performed using the default setting of each of the four aligners (Appendix A-D). The human reference genome used in this experiment is the hg38 genome obtained from UCSC (http://hgdownload.soe.ucsc.edu/ downloads.html), and hg38 is chosen because this genome is the latest and most stable built of human reference genome now. In addition, this genome is the corrected and improvised version of the previous built, hg19.

For the second part of the study, an alignment of each tool was firstly performed using the default parameters using the trimmed data sets. Then, the specific parameter settings suggested by the tool were used to increase the quality of the alignments. In addition, four sets of parameter settings for each aligner were also used.

Evaluation of Precision and Recall

Alignment quality is perceived in the form of alignment precision and recall values. The precision determines which fraction of the aligned reads are being aligned correctly, while the recall value evaluates which fraction of the overall reads is correctly recovered. First, the number of true and false positive alignments was determined to estimate precision and recall values. Then, the mapping of any reader to a correct genomic location was defined as a true positive (TP), while the mapping of any read to an incorrect location was counted as a false positive (FP). Next, false positives were determined, including all the reads aligned to multiple locations. Apart from that, the reads failing to map to any correct position were considered false negative (FN) alignments. Since each read originated from one unique genome location, it should be mappable into a specific location after the alignment step. Thus, there was no such measurement for true negative alignments.

Precision, recall and false discovery rate (FDR) were calculated by using the following Equations 1 to 3:

$$Precision = \frac{TP}{TP+FP} \qquad [1]$$

$$Recall = \frac{TP}{TP+FN} \qquad [2]$$

$$FDR = \frac{FP}{TP+FN} \qquad [3]$$

F-measure that evaluates the trade-off between precision and recall was also calculated in this study using Equation 4:

$$F = \frac{2 * precision * recall}{precision + recall} \qquad [4]$$

If the multiple parameters set for one aligner resulted in equal F-measures, then the dimension of comparison would be based on runtime and memory requirements.

Impact of Parameter Choice on Alignment Quality

Four parameter combinations were tested to evaluate the impacts of each change on the alignment performances to investigate the optimised parameter setting further. In addition, the dispersion of F-measures by each alignment tool was used to determine the tool's sensitivity level to the tweaks of parameter values.

Runtime and Memory Requirements

Aligners were installed and ran on the check. If multi-threading was supported, then 12 cores were used. The memory usage was capped at 16 GB. Next, the total CPU time measurement and memory usage were extracted from the reports using the "time" command, especially the memory usage recorded from the maximum memory used during the job execution. The alignment jobs were run on Intel® Core ™ i7-8700 CPU @ 3.2GHz x 12 processors.

RESULTS AND DISCUSSIONS

Aligners' Performance on Sequencing Data with Different Qualities

The performance of the alignment on the different data qualities shows a slight difference in the results. For the trimmed data of both replicates, all aligners generally show a higher concordance compared to the untrimmed data, except for TopHat. It indicated that these aligners (Novoalign, HISAT2 and Subread) worked better with high-quality reads. These high-quality reads were obtained after processing our HeLa raw reads using a fastp trimming tool where the low-quality bases and the adapter had been trimmed off. On the other hand, TopHat was not affected by the quality of the aligned reads as the trimmed data showed less concordance compared to the untrimmed data. Nevertheless, the untrimmed data of our sample still show a good concordance but is slightly lower than the trimmed ones—the comparison is shown in Supplementary Data (Table 1A).

In this comparison, it was noticed that the difference among the results was more significant in recall compared to precision. It might be caused by the increasing number of errors that affected the precision values. On the other hand, it might be due to quality concerns which can significantly mislead analytical results and lead to inaccurate conclusions (Zhou et al., 2018). It explained why it is crucial to discard the low-quality bases and adapters that might contaminate the purity of our readings.

Accuracy and Efficiency of the Aligners

The alignment accuracy and efficiency were assessed in terms of the precision and recall values. Precision reveals which portion of the reads was correctly aligned, while recall reveals which portion of the overall reads was being recovered correctly. The aligners studied were built with two different algorithms: the hash table-based and FM-index algorithms. Between the two hash-table-based aligners, Novoalign has a much higher precision value than Subread (Supplementary Data - Table 2A). However, the recall values between these two aligners were equally high (>0.92), except for the Novoalign Tweak 2 parameter setting that obtained an extremely low recall value of 0.79419. A previous study by Donato et al. (2021) compared 17 aligners on simulated and empirical NGS data, and the findings revealed that Novoalign showed the highest accuracy in all alignments. In addition, the study also highlighted that Novoalign could detect a new transcript with greater ease than the other tools tested.

While for FM-index-based alignment, a significant difference between the precision values of HISAT2 and TopHat was observed. HISAT2 showed the highest precision at optimised parameters (labelled as Tweak 2) with the value of 0.80185. TopHat's highest precision was only at 0.70199 when the alignment was carried out at the default setting. Nonetheless, the recall values ranging from 0.91513 to 0.97998 were equally high for both aligners. The highest precision value (0.94773) in HISAT2 was shown in the sample with the tweaked parameters setting labelled HISAT2 Tweak 2. However, the recall value was just average. In contrast, the lowest precision value in HISAT2 was shown in the sample with HISAT Tweak 3 parameters setting, with the value of 0.76961 but with a significantly high recall value of 0.97998. These results showed that in HISAT2, the precision and recall values had a negative correlation. Besides the percentage of mapped reads, the alignment accuracy also depends on the correctness of the reads mapped to the reference genome or transcriptome. Schaarschmidt et al. (2020) revealed that alignment using HISAT2 resulted in high overlapping reads, mainly coming from the soft clipping of the first base of the reads. The failure of TopHat and HISAT2 to tolerate the soft-clipping and mismatches had caused a large fraction of reads to be left unmapped (Sahraeian et al., 2017). However, the setting can be turned off, directly eliminating the observed differences.

Likewise, the precision values of TopHat increased with the decrement of recall values. Amongst the default and tweaked parameters settings of TopHat, the default setting was measured with an outstandingly high value of precision (0.70199). However, on the contrary, the recall value was only 0.91513 and was the lowest among the other TopHat settings. On the other hand, the TopHat with parameter Tweak 3 set had the lowest precision value (0.62911), but the recall was at a seemingly high value of 0.97864. Although this aligner performed well in aligning a read onto the respective genomic locus, similar to our findings, the study by Raplee et al. (2019) also found notable discrepancies and deficiencies of TopHat in obtaining insufficient genomic alignment for reliable downstream analysis. Furthermore, TopHat prevented the truncation of the reads, which directly led to many unmapped reads (Sahraeian et al., 2017).

Meanwhile, for the F-measure, FM-index-based aligners showed a significant difference between the aligners, as the average of HISAT2 was 0.86355, while the average of TopHat was only 0.76479. Similarly, the F-measures between the two hash-table-based aligners also showed a vast difference. The F-measure values of Novoalign and Subread were 0.84247 and 0.76447, respectively. These results showed that the types of algorithms did not correlate with the F-measure. Overall, between these two FM-index-based aligners, it can be concluded that HISAT2 is a more reliable aligner with reasonable quality alignment.

By observing the F-measures for the overall alignment quality results, it was found that in most of the cases, the F-measures were reduced as the recall values were getting lower. It was notable in three Novoalign cases with almost similar precision values (around 0.72) but different recall values. The case with a high recall value (>0.97) showed a high F-measure, but the case with a low recall value showed an extreme drop in F-measure (0.76). Most of the time, the low recall value caused a reduction in the overall alignment quality in terms of the F-measure.

Performance of Aligners' Optimal Parameter Settings

The optimal parameter for each of the four evaluated aligners was determined by testing all permutations that appeared to have an impact on alignment quality. The optimal parameter sets run along with corresponding performance measures are shown in Table 1.

The recall metric for the optimal parameters was well balanced, and the values ranged between 0.948 (HISAT2) and 0.852 (Subread). While for the precision metric, the highest was shown by Novoalign (0.896) and the lowest shown in Subread (0.680). Novoalign displayed a significantly high value in precision and F-value metrics. FDR value also shows that Novoalign has the lowest value compared to the other aligners. The lower the FDR value defines the expected proportion of false positives among the declared significant results, so the lower the value, the better it will be. FDR is a useful approach to measure the false discoveries within a set of hypothesis tests called significant (Chen et al., 2021).

Table 1
Performance of aligners under different metrics

Metric	Novoalign	HISAT2	Tophat	Subread
Reads aligned	56252385	56049365	59413014	48916204
Recall	92.38	94.77	91.51	85.16
Precision	89.64	80.18	70.20	68.06
F- value	90.99	86.87	79.45	75.66
FDR	1.04	1.98	2.98	3.19

When testing many hypotheses, FDR is often employed to determine significance thresholds and quantify the overall error rate. We observed that Novoalign was the leading aligner with the highest accuracy.

Parameter Stability

Observing the parameters' effects on each aligner's performance is crucial and determining the aligner that performs well with the default settings. These evaluations will allow us to assess the robustness of the alignment qualities among parameter variations. While there are enormous spaces in manipulating the parameters of choice, the combinations may not necessarily produce a global optimum output. The main idea is that the parameter variation should allow the users to have a consistent precision and recall value to alter the runtime and memory properties without affecting the overall performance.

Parameter optimisation was performed on a duplicated data sample. Table 2 shows the dispersion values of the F-measure over the chosen parameter space. Low standard deviation, SD, as observed in HISAT2, indicated that the choice of parameters had little impact on the alignment performance. The high SD value indicated a wide alignment quality distribution, as shown by Novoalign. Meanwhile, both Subread and TopHat showed an average number of dispersions of 0.01142 and 0.02054, respectively. Different parameters had little impact on the alignment performance of HISAT2, whereas the alignment performance of Novoalign was widely affected. These results reflected that Novoalign was highly sensitive in terms of the choice of parameter settings. Hence, a precise setting must be carefully chosen when using Novoalign, as little change can cause a huge difference in the results.

Precision and recall values significantly affected the Novoalign aligner, as both of these values were widely distributed. Unlike Novoalign, the values of precision and recall in HISAT2 were consistent, with both lowly distributed. These results were not correlated with the Subread results. In Subread, there was a significantly high difference between the precision and recall values. In terms of precision, it was noticed that the values were consistent within each of the tools, regardless of the settings, except for the Novoalign default setting. The Novoalign default setting showed extremely high precision compared to the other parameter settings. However, for the recall values, the default setting of Novoalign was not as outstanding as the precision value since higher recall values were shown in the tweaked parameter settings. These results illustrated the importance of evaluating both the precision and recall values. Furthermore, the results

Table 2
Dispersion of F-measure across all parameter settings tested for each aligner

Aligner	Dispersion of F-measure
Novoalign	0.05433
HISAT2	0.00326
Tophat	0.02054
Subread	0.01142

were represented by the precision-recall analysis of HISAT2 and Novoalign, with the two aligners determined with extremely low and high F-measure dispersion, respectively (Figure 1).

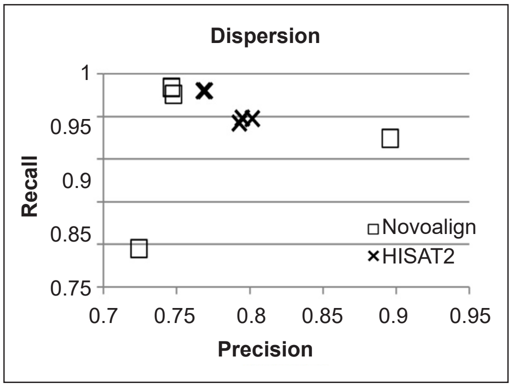


Figure 1. Influence of parameter selection. Precision (x-axis) and recall (y-axis) are shown for Novoalign (squares) and HISAT2 (X signs) aligners

Remarkably, the high dispersion value of Novoalign also resulted in a dramatic drop in the recall, especially if the value of A in the alignment scoring threshold was set at the highest score acceptable for alignment, which was 20. These results were out of the expectation, as one would expect many true positives output from this alignment scoring setting. In contrast, the false negative value was extremely high compared to the other parameters. A low accuracy discovered in Novoalign might be due to its over-mapping at both ends of short reads (Nodehi et al., 2021; Shang et al., 2014). Thus, although dispersion does not state the alignment software's overall performance, it indicates whether the optimal performance can be achieved without an in-depth understanding of algorithmic details.

Runtime and Memory Requirements

The runtime evaluation of each aligner was based on two main steps in the alignment process. The first is the indexing and the second one is the alignment. By comparing the indexing runtime for each of the aligners, it was found that the relationship between the indexing time correlated with the types of algorithms used. For example, HISAT2 and TopHat, FM-index-based aligners require a longer time (more than 60 mins) to build a genome index. In contrast, hash-table-based aligners, like Subread and Novoalign, can build genome index in less time (< 15mins).

Most of these four aligners were designed with a trade-off between the indexing and alignment runtime (Table 3). For example, Subread and Novoalign were able to build an index of the genome within a short duration. Still, they required plenty of alignment time at the default setting, with 47.30 mins and 244.8 mins recorded, respectively. Conversely,

Table 3
Indexing and alignment runtime of the aligners

Aligner	Algorithm	Indexing runtime (mins)	Alignment runtime (mins)
HISAT2	FM-index based	61.13	12.02
TopHat		78.39	602.19
Subread	Hash-table based	13.30	47.3
Novoalign		10.49	244.8

HISAT2 required a short duration to build an index but took 12.02 mins of alignment time when using the default setting. It was approximately five times faster than the indexing runtime (Figure 2).

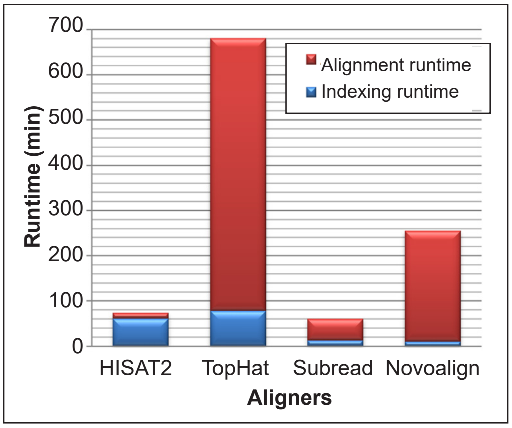


Figure 2. The measurements of indexing and alignment runtime of each aligner on the default setting

On the other hand, TopHat required an extremely long time for indexing and aligning compared to other tools. The TopHat's default setting required 602.19 mins to complete the alignment. However, TopHat required the least memory compared to the other three tools to compensate for the lengthy runtime. TopHat only required 4.0 Gb RAM, while Novoalign, Subread and HISAT2 required 8.0 Gb, 10.0 Gb and 6.7 Gb of RAM, respectively.

By observing the runtime for the default and optimised parameters settings of each aligner, we found that the alignment runtime of each aligner barely depended on the types of algorithms but more on the individual settings. Although the runtime differs significantly among the hash table-based aligners, the runtime among Subread's optimised parameters and default settings was more consistent than Novoalign. Subread generally required 1 hour of runtime and even less for the default setting. In contrast to Subread, the default setting of Novoalign required more than 2 hours of runtime to obtain a comparable precision and recall. In addition, Novoalign had the longest computational time of reads mapping, probably due to its predispositions toward other parameters (Donato et al., 2021).

As for the FM-index-based approach, the alignment runtime of the aligners varied vastly. HISAT2 outperformed TopHat in terms of the runtime. The average runtime of HISAT2 was only 12 mins and 43 sec, while TopHat needed at least 3.45 hours of runtime. In terms of memory consumption, HISAT2 required 6.7 Gb RAM, while TopHat required the least memory at only 4.0 Gb RAM. These results were consistent with the previous findings by Keel and Snelling (2018), who found that HISAT2 was significantly faster and used less memory through simulated data sets. Despite the very low memory consumption, TopHat could still achieve a reasonable alignment quality, thus supporting the widespread use of TopHat within many of the RNA-seq mapping approaches and as the most cited aligner. On the other hand, no significant correlation was observed between the F-measure with the runtime and memory requirements.

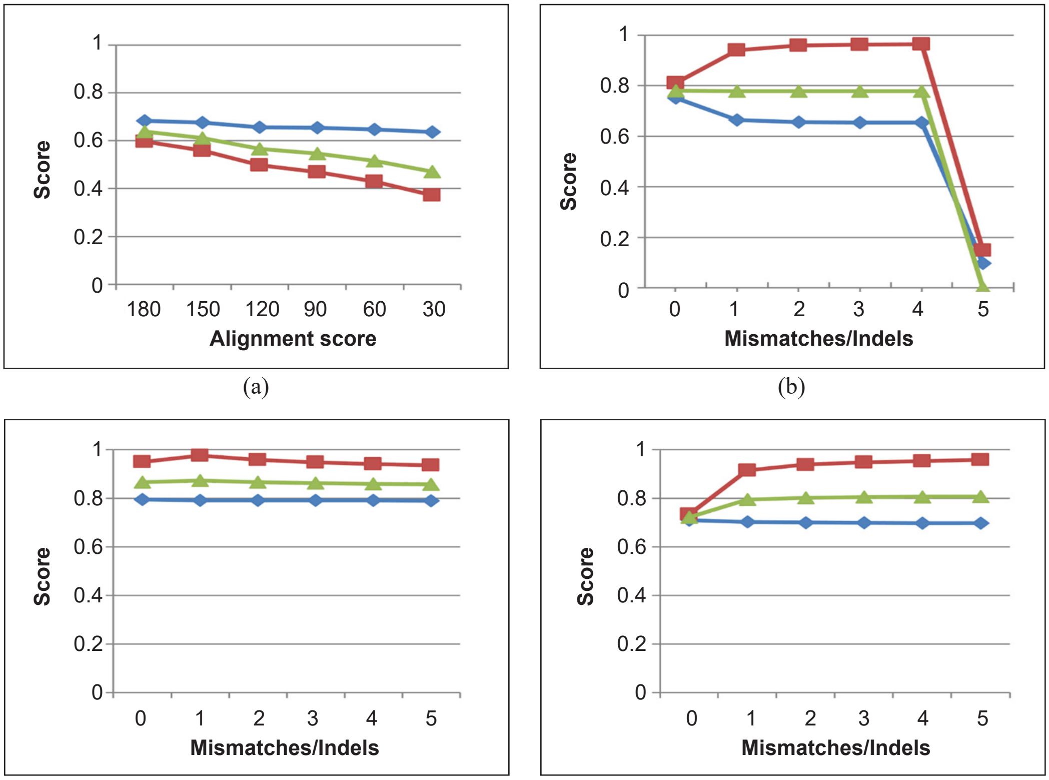
HISAT2 performed extremely fast alignment with comparable accuracy to the other aligners. In contrast, the alignment runtime for Subread was the second-fastest but achieved poor alignment quality. The runtime for Novoalign was acceptable and at an average rate, while the alignment with TopHat was the slowest. The alignment with TopHat was thus considered inefficient, especially when working with multiple data sets.

Error Tolerance

This section aims to analyse each aligner's sensitivity in response to a specific number of errors allowed in the reads. Goodwin et al. (2016) believed that NGS platforms provide a massive amount of data, while each platform is associated with error rates ranging from 0.1 to 15%. Therefore, a good alignment algorithm used in mapping sequence data must be able to compensate for these inevitable raw data errors (Keel & Snelling, 2018). A previous study by Sun et al. (2017) found that alignment is a critical step for intermediate indel detection. Therefore, each read was measured to determine the number of mismatches or indels for this purpose. The highest number of mismatches and indels in a correctly aligned read was 10 and 5, respectively, for all the aligners tested, except for Novoalign. Unlike other aligners, Novoalign first searches the candidate alignment positions from the reference genome using the Needleman-Wunsch algorithm based on the alignment score. Due to this alignment-score-based search algorithm, the users cannot define the number of allowed mismatches in each alignment, but the users can still set up a threshold of an alignment score. The mapping quality scores define the accuracy of alignment, meaning that the higher the alignment quality score, the more accurate an alignment is. Thus, the alignment score threshold is from 30 to 180 for Novoalign.

Figure 3 shows the impacts of errors on the alignment quality for each aligner by using its default setting. Generally, it was noticed that the precision of the alignments was barely affected by the number of errors. As illustrated in Figures 3(c) and 3(d), the precision values of HISAT2 and TopHat showed a flat line. Sun et al. (2017) believed that most variant calling programs would miss the intermediate indels from these aligners, except when the soft-clipped reads were sufficiently triggered. Furthermore, the study also discovered that the TopHat family RNA-seq mapping programs do not align the reads with intermediate indels, or the reads were minimally aligned when HISAT2 was used.

In contrast with precision readings, the recall values obtained showed that more significant changes were recorded with increasing errors. Interestingly, we also determined a drastic drop in Subread once the mismatches and indels were set at 5 and 10, respectively, as shown in Figure 3(b). It was possibly due to the general design of the alignment algorithm itself, as the algorithm is robust enough to detect a small number of single-based mismatches, depending on the parameter setting. As a result, most of the algorithm's recall stayed relatively constant while still being within the tolerated range of mismatches but dropped significantly as soon as this range was exceeded, as shown in the Subread aligner. In addition, the other aligners could tolerate up to five mismatches. For Novoalign, as illustrated in Figure 3(a), the recall rate gradually decreased as the alignment score decreased.



(a) (b)

(c) (d)

Figure 3. The impacts of errors on the alignment quality for each of the aligners tested: (a) Novoalign; (b) Subread; (c) HISAT2; and (d) TopHat. The dependencies of alignment precision (blue lines), recall (red lines) and F-measure (green lines) on each of the alignment algorithms based on the default setting were analysed in this study.

Some aligners tolerate indels as these tools are designed to handle gapped alignments. As a result, most algorithms' recall values were considerably impaired by indels' presence. A remarkable tolerance to indels was shown by HISAT2 and TopHat, with a near-constant performance observed, even as the indel counts increase. Even though the baseline recall of Novoalign was already rather low at 0.6, the other aligners were shown to tolerate indels, with significantly higher values of baseline recall recorded. Subread was found to be the most vulnerable to indels among these aligners. The main limitation in the gapped alignment-based indels detection method is the need for indels that are entirely contained within a read and correctly detected during the initial read mapping step (Donato et al., 2021; Li et al., 2009). In the case of small indel detection, the supporting reads are frequently presented with too few bases that can match with the reference genome, or the reads may contain only one end that can map correctly to the reference genome. However, the remaining bases following the indel may be trimmed or soft-clipped by the NGS aligner (Donato et al., 2021; Landman et al., 2014).

As our study was using RNA-Seq data of HeLa cells, we needed to have the tools that could tolerate indels and mismatches. It was because, according to Bottomley et al. (1969), Liu et al. (2019), and Rutledge (2014), HeLa contained a huge number of genomic variants. Furthermore, it was reported that the variation in the karyotype of HeLa had been stated inconsistently and proven that such changes had affected the phenotype of the cells (Fasterius & Al-Khalili Szigyarto, 2018). Indeed, Hela has been purposely established to be genetically unstable (Fasterius & Al-Khalili Szigyarto, 2018; Yoo et al., 2017). Therefore, by allowing the mismatches in the alignment, we still managed to handle these biological characteristics of Hela whilst obtaining the maximum alignment precision.

CONCLUSION

In this study, we concluded that the alignment procedure's accuracy depended on two aspects. The first was the quality of the reads, and the second was the parameter settings. A high-quality read is free from any adapter contamination and low-quality bases. The alignments of the high-quality reads show significant increment in all aligners tested in this study except for TopHat data 1. It indicated that the TopHat alignment quality was not affected by the quality of the reads. While for the parameter settings, the number of mismatches and indels allowed displayed a great impact on alignment accuracy.

Apart from the accuracy, the notable difference between these aligners was the runtime. After comparing each aligner's runtime and alignment qualities, we can summarise that the fast runtime did not guarantee a high alignment quality. So, the choice of the aligner depends on the target of the researchers. Nevertheless, our study found that HISAT2 performed extremely fast alignment with comparable accuracy to the other aligners.

In conclusion, the default setting of Novoalign was the most reliable setting that suited our data sample. In addition, our study provides a systematic comparison between the commonly used alignment programmes in RNA-Seq studies. Furthermore, the approach from this study can be applied to future research that deals with human cancer cell datasets generated from different platforms. Moreover, this study can also be utilised as guidance in selecting the best alignment tool for various sources of samples.

In this study, we focused mainly on comparing the aligners from two specific angles, using real data of HeLa cells reads of varying qualities. Thus, there were a few limitations in this study. Firstly, rather than the data from real cases of cancer, we only used a sample from a cell line. Secondly, our sequencing data sets were only sequenced from the Illumina sequencer. Third, we were only using pair-end data in this study without using the single-pair end to compare the results. Lastly, while there are many available alignment tools, we only compared a small number of alignment tools in this study. Although this article has these limitations, our study applies to the pair-end whole genome real sequencing data and data generated from other sources. The main challenge we faced was to set up a fair comparison of the chosen parameters and evaluation metrics among these different aligners. Regardless of this challenge, we managed to achieve the objectives of this study, and we expect that this study can guide other researchers in choosing the optimised settings of the aligners of interest.

ACKNOWLEDGEMENTS

The authors want to thank Universiti Sains Malaysia, Malaysia, for the support.

REFERENCES

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Babraham Bioinformatics. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Baruzzo, G., Hayer, K. E., Kim, E. J., Di Camillo, B., Fitzgerald, G. A., & Grant, G. R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods*, *14*(2), 135-139. https://doi.org/10.1038/nmeth.4106

Bottomley, R. H., Trainer, A. L., & Griffin, M. J. (1969). Enzymatic and chromosomal characterization of HeLa variants. *The Journal of Cell Biology*, *41*(3), 806-815. https://doi.org/10.1083/jcb.41.3.806

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

Chen, X., Robinson, D. G., & Storey, J. D. (2021). The functional false discovery rate with applications to genomics. *Biostatistics, 22*(1), 68-81. https://doi.org/10.1093/biostatistics/kxz010

Donato, L., Scimone, C., Rinaldi, C., D'Angelo, R., & Sidoti, A. (2021). New evaluation methods of read mapping by 17 aligners on simulated and empirical NGS data: An updated comparison of DNA- and RNA-seq data from Illumina and Ion Torrent technologies. *Neural Computing and Applications*, *33*(22), 15669-15692. https://doi.org/10.1007/s00521-021-06188-z

Fasterius, E., & Al-Khalili Szigyarto, C. (2018). Analysis of public RNA-sequencing data reveals biological consequences of genetic heterogeneity in cell line populations. *Scientific Reports, 8*(1), 1-11. https://doi.org/10.1038/s41598-018-29506-3

Ferragina, P., & Manzini, G. (2000). Opportunistic data structures with applications. In *Proceedings 41st Annual Symposium on Foundations of Computer Science* (pp. 390-398). IEEE Publishing. https://doi.org/10.1109/sfcs.2000.892127

Fonseca, N. A., Rung, J., Brazma, A., & Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics*, *28*(24), 3169-3177. https://doi.org/10.1093/bioinformatics/bts605

Gaur, P., & Chaturvedi, A. (2017). *A survey of bioinformatics-based tools in RNA-sequencing (RNA-seq) data analysis*. In *Translational Bioinformatics and its Application* (pp. 223-248). Springer. https://doi.org/10.1007/978-94-024-1045-7_10

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333-351. https://doi.org/10.1038/nrg.2016.49

Grytten, I., Rand, K. D., Nederbragt, A. J., & Sandve, G. K. (2020). Assessing graph-based read mappers against a novel baseline approach highlights strengths and weaknesses of the current generation of methods. *BMC Genomics, 21*, Article 282. https://doi.org/10.1186/s12864-020-6685-y

Hu, W. E., Zhang, X., Guo, Q. F., Yang, J. W., Yang, Y., Wei, S. C., & Su, X. D. (2019). HeLa-CCL2 cell heterogeneity studied by single-cell DNA and RNA sequencing. *PLoS One*, *14*(12), Article e0225466. https://doi.org/10.1371/journal.pone.0225466

Jain, C., Rhie, A., Zhang, H., Chu, C., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Weighted minimizer sampling improves long read mapping. *Bioinformatics*, *36*, I111-I118. https://doi.org/10.1093/BIOINFORMATICS/BTAA435

Keel, B. N., & Snelling, W. M. (2018). Comparison of Burrows-Wheeler transform-based mapping algorithms used in high-throughput whole-genome sequencing: Application to illumina data for livestock genomes 1. *Frontiers in Genetics*, *9*, 1-6. https://doi.org/10.3389/fgene.2018.00035

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357-360. https://doi.org/10.1038/nmeth.3317

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine*, *12*(1), 1-13. https://doi.org/10.1186/s13073-020-00791-w

Križanović, K., Echchiki, A., Roux, J., & Šikić, M. (2018). Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics*, *34*(5), 748-754. https://doi.org/10.1093/bioinformatics/btx668

Landman, S. R., Hwang, T. H., Silverstein, K. A. T., Li, Y., Dehm, S. M., Steinbach, M., & Kumar, V. (2014). SHEAR: Sample heterogeneity estimation and assembly by reference. *BMC Genomics*, *15*(1), 1-12. https://doi.org/10.1186/1471-2164-15-84

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Liao, Y., Smyth, G. K., & Shi, W. (2013). The subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, *41*(10), e108-e108. https://doi.org/10.1093/nar/gkt214

Liu, Y., Mi, Y., Mueller, T., Kreibich, S., Williams, E. G., Van Drogen, A., Borel, C., Frank, M., Germain, P. L., Bludau, I., Mehnert, M., Seifert, M., Emmenlauer, M., Sorg, I., Bezrukov, F., Bena, F. S., Zhou, H., Dehio, C., Testa, G., & Aebersold, R. (2019). Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nature Biotechnology*, *37*(3), 314-322. https://doi.org/10.1038/s41587-019-0037-y

Nodehi, H. M., Tabatabaiefar, M. A., & Sehhati, M. (2021). Selection of optimal bioinformatic tools and proper reference for reducing the alignment error in targeted sequencing data. *Journal of Medical Signals and Sensors*, *11*(1), 37-44. https://doi.org/10.4103/jmss.JMSS-7-20

Qin, D. (2019). Next-generation sequencing and its clinical application. *Cancer Biology and Medicine*, *16*(1), 4-10. https://doi.org/10.20892/j.issn.2095-3941.2018.0055

Raplee, I. D., Evsikov, A. V., & De Evsikova, C. M. (2019). Aligning the aligners: Comparison of rna sequencing data alignment and gene expression quantification tools for clinical breast cancer research. *Journal of Personalized Medicine*, *9*(2), Article 18. https://doi.org/10.3390/jpm9020018

Rutledge, S. (2014). What HeLa cells are you using? *The Winnower, 9*, 1-9. https://doi.org/10.15200/winn.143896.65158

Sahlin, K., & Mäkinen, V. (2021). Accurate spliced alignment of long RNA sequencing reads. *Bioinformatics, 37*(24), 4643-4651. https://doi.org/10.1093/bioinformatics/btab540

Sahraeian, S. M. E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P. T., Au, K. F., Bani Asadi, N., Gerstein, M. B., Wong, W. H., Snyder, M. P., Schadt, E., & Lam, H. Y. K. (2017). Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nature Communications*, *8*(1), 1-14. https://doi.org/10.1038/s41467-017-00050-4

Schaarschmidt, S., Fischer, A., Zuther, E., & Hincha, D. K. (2020). Evaluation of seven different RNA-seq alignment tools based on experimental data from the model plant *Arabidopsis thaliana*. *International Journal of Molecular Sciences*, *21*(5), Article 1720. https://doi.org/10.3390/ijms21051720

Schilbert, H. M., Rempel, A., & Pucker, B. (2020). Comparison of read mapping and variant calling tools for the analysis of plant NGS data. *Plants*, *9*(4), Article 439. https://doi.org/10.3390/plants9040439

Shang, J., Zhu, F., Vongsangnak, W., Tang, Y., Zhang, W., & Shen, B. (2014). Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *BioMed Research International*, *2014*, Article 309650. https://doi.org/10.1155/2014/309650

Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P., & Kocher, J. P. A. (2017). Indel detection from RNA-seq data: Tool evaluation and strategies for accurate detection of actionable mutations. *Briefings in Bioinformatics*, *18*(6), 973-983. https://doi.org/10.1093/bib/bbw069

Thankaswamy-Kosalai, S., Sen, P., & Nookaew, I. (2017). Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics*, *109*(3-4), 186-191. https://doi.org/10.1016/j.ygeno.2017.03.001

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-seq. *Bioinformatics*, *25*(9), 1105-1111. https://doi.org/10.1093/bioinformatics/btp120

Wu, D. C., Yao, J., Ho, K. S., Lambowitz, A. M., & Wilke, C. O. (2018). Limitation of alignment-free tools in total RNA-seq quantification. *BMC Genomics, 19*(1), 1-14. https://doi.org/10.1101/246967

Yoo, Y. S., Han, H. G., & Jeon, Y. J. (2017). Unfolded protein response of the endoplasmic reticulum in tumor progression and immunogenicity. *Oxidative Medicine and Cellular Longevity*, *2017*, Article 2969271. https://doi.org/10.1155/2017/2969271

Zhang, C., Zhang, B., Lin, L. L., & Zhao, S. (2017). Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*, *18*(1), 1-11. https://doi.org/10.1186/s12864-017-4002-1

Zhou, Q., Su, X., Jing, G., Chen, S., & Ning, K. (2018). RNA-QC-chain: Comprehensive and fast quality control for RNA-Seq data. *BMC Genomics*, *19*(1), 1-10. https://doi.org/10.1186/s12864-018-4503-6

SUPPLEMENTARY DATA

Table 1A
Percentage of reads uniquely aligned in two data sets by four aligners under default setting

Aligners	**Metrics**	**Dataset 1**		**Dataset 2**	
		Untrimmed data	**Trimmed data**	**Untrimmed data**	**Trimmed data**
Novoalign	Precision	0.8678	0.8964	0.8262	0.8618
	Recall	0.98886	0.9238	0.9526	0.8693
	F	0.9242	0.9099	0.8849	0.8655
	FDR	0.1323	0.1036	0.1738	0.1382
HISAT2	Precision	0.7554	0.7947	0.7851	0.8018
	Recall	0.7941	0.9484	0.9816	0.9477
	F	0.7743	0.8648	0.8724	0.8687
	FDR	0.2446	0.2053	0.2149	0.1982
Tophat	Precision	0.7087	0.6286	0.6348	0.6291
	Recall	0.8620	0.8898	0.9451	0.9786
	F	0.7779	0.7367	0.7608	0.7659
	FDR	0.2913	0.314	0.3652	0.3709
Subread	Precision	0.6028	0.6806	0.5794	0.6330
	Recall	0.8689	0.8516	0.8648	0.9401
	F	0.7118	0.7566	0.6940	0.7566
	FDR	0.3972	0.3194	0.4205	0.3670

Table 2A
The alignment summaries of default and four sets of parameters settings by using Novoalign, HISAT2, TopHat and Subread aligners

Tools	Sample	Total reads	Precision	Recall	F	FDR	Runtime	Memory
Novoalign	NOVOA default	56252385	0.89636	0.92376	0.90985	0.10364	138m 39s	8.0 Gb
	NOVOA tweak1	68796275	0.74681	0.98410	0.84919	0.25319	270m 04s	8.0 Gb
	NOVOA tweak2	64061997	0.72463	0.79419	0.75782	0.27537	57m 59s	8.0 Gb
	NOVOA tweak3	68769479	0.74690	0.98297	0.84883	0.25310	272m 22s	8.0 Gb
	NOVOA tweak4	68497483	0.74798	0.97529	0.84664	0.25202	244m 08s	8.0 Gb
HISAT2	HISAT2 default	56525790	0.79269	0.94209	0.86095	0.20731	11m 08s	6.7 Gb
	HISAT2 tweak1	57616454	0.76800	0.97996	0.86113	0.23200	14m 38s	6.7 Gb
	HISAT2 tweak2	56049365	0.80185	0.94773	0.86870	0.19816	11m 44s	6.7 Gb
	HISAT2 tweak3	57530440	0.76961	0.97998	0.86215	0.23039	13m 23s	6.7 Gb
	HISAT2 tweak4	56519361	0.79474	0.94844	0.86481	0.20526	12m 02s	6.7 Gb
TopHat	TOPHAT default	59413014	0.70199	0.91513	0.79452	0.29801	600m 30s	4.0 Gb
	TOPHAT tweak1	70379105	0.62912	0.97838	0.76580	0.37089	274m 14s	4.0 Gb
	TOPHAT tweak2	67973298	0.63477	0.94999	0.76103	0.36523	207m 53s	4.0 Gb
	TOPHAT tweak3	70390636	0.62911	0.97864	0.76588	0.37089	284m 12s	4.0 Gb
	TOPHAT tweak4	62577636	0.62861	0.88975	0.73673	0.37139	602m 19s	4.0 Gb
Subread	Subread default	48916204	0.65845	0.95194	0.77845	0.34155	45m 45s	10.0 Gb
	Subread tweak1	48916204	0.64358	0.97506	0.77538	0.35642	49m 30s	10.0 Gb
	Subread tweak2	48916204	0.61826	0.97060	0.75537	0.38174	69m 51s	10 .0Gb
	Subread tweak3	48916204	0.63300	0.94011	0.75658	0.36700	68m 46s	10.0 Gb
	Subread tweak4	48916204	0.68062	0.85162	0.75658	0.31938	47m 30s	10 .0 Gb

APPENDICES

Aligners Command and Parameters Setting

Appendix A

Alignment with HISAT2

Read alignment was performed using command as follows:

```
hisat2 --threads 16 --time --reorder --known-splicesite-infile <output index
path>/<genome name>.splicesites.txt --novel-splicesite-outfile splicesites.novel.txt -
-novel-splicesite-infile splicesites.novel.txt -f -x <index name> -1 <read file 1> -2
<read file 2> -S <output sam file>
-N <NUM_MISMATCH>
-L <SEED_LENGTH>
-i S,1,<SEED_INTERVAL>
-D <SEED_EXTENSION>
-R <RE_SEED>
--pen-noncansplice<PENALITY_NONCANONICAL>
--mp <MAX_MISMATCH_PENALITY>,<MIN_MISMATCH_PENALITY>
--sp <MAX_SOFTCLIPPING_PENALITY>,<MIN_SOFTCLIPPING_PENALITY>
```

Sample	-N <NUM_MISMATCH> -L <SEED_LENGTH> -i S,1,<SEED_INTERVAL> -D <SEED_EXTENSION> -R <RE_SEED> --pen noncansplice<PENALITY_NONCANONICAL> --mp <MAX_MISMATCH_PENALITY>,<MIN_MISMATCH_PENALITY> --sp<MAX_SOFTCLIPPING_PENALITY>, <MIN_SOFTCLIPPING_PENALITY>
default	0-22-1.15-15-2-3-6,2-2,1
Tweak 1	0-20-0.5-25-5-20-1,0-3,0
Tweak 2	1-20-0.5-25-5-20-3,0-2,1
Tweak 3	0-22-1.5-15-2-20-1,0-2,1
Tweak 4	OFF-OFF- OFF-OFF- OFF-OFF- OFF-OFF

Appendix B

Alignment with Novoalign

Read alignment was performed using command as follows:

```
novoalign -d <output index file> -f <read file 1> <read file 2> -F FA -o SAM -r All 10
-t <A_SCORE>,<B_SCORE> -h -1 -1 -i PE <FRAGMENT_LENGTH_MEAN>,<FRAGMENT_
LENGTH_SD> -v 0
70 70 “[>]([^:]*)” > <output sam file> 2>alignment.log
```

Sample	REPEAT_FILTER - -t <A_SCORE>,<B_SCORE>
default	Default-Default-default
tweak 1	OFF-10-4.5
tweak 2	OFF-20-2
tweak 3	OFF-12-4.5
tweak 4	default-20-4.5

Appendix C

Alignment with Subread

Read alignment was performed using command as follows:

```
subjunc -i <index> -r <read file 1> -R <read file 2> -T 16
--SAMoutput -o <output alignment>  --allJunctions
                                   -d <MIN_FRAGMENT_LENGTH>
                                   -I <INDEL>
                                   -m <NUM_HIT_SUBREADS>
                                   -M <MISMATCHES>
                                    -n <NUM_EXTRACTED_SUBREADS>
                                   -p <NUM_HIT_PAIR_SUBREADS>
                                   --complexIndels
```

Sample	-d <MIN_FRAGMENT_LENGTH> -I <INDEL> -m <NUM_HIT_SUBREADS> -M <MISMATCHES> -n <NUM_EXTRACTED_SUBREADS> -p <NUM_HIT_PAIR_SUBREADS> --complexIndels
default	50-5-3-3-10-1-off
tweak 1	0-10-1-20-15-1-off
tweak 2	0-10-1-20-5-1-on
tweak 3	0-10-3-3-10-3-on
tweak 4	0-5-3-3-10-1-off

Appendix D

Alignment with TopHat

Read alignment was performed using command as follows:

```
tophat2 --output-dir <output path> <index> <reads file 1> <reads file 2> --num-threads 16 --GTF <gtf
file> --mate-inner-dist <INNER_MATE_MEAN>
      --mate-std-dev <INNER_MATE_SD>
      --b2-very-sensitive
      --read-mismatches <NUM_MISMATCHES>
      --read-gap-length <NUM_GAP_LENGTH>
      --read-edit-dist <NUM_EDIT_DIST>
      --read-realign-edit-dist <NUM_REALIGN_EDIT_DIST>
      --max-insertion-length <NUM_INSERTION_LENGTH>
      --max-deletion-length <NUM_DELETION_LENGTH>
      --max-multihits <NUM_MULTIHITS>
```

Sample	B2_VERY_SENSITIVE - NUM_MISMATCHES - NUM_GAP_LENGTH -NUM_EDIT_DIST - NUM_REALIGN_EDIT_DIST - NUM_INSERTION_LENGTH - NUM_DELETION_LENGTH - NUM_MULTIHITS
default	On-2-2-2-3-3-3-20
tweak 1	On-18-25-25-26-24-24-100
tweak 2	On-7-6-7-8-4-4-100
tweak 3	On-25-25-25-26-24-24-100
tweak 4	OFF-OFF- OFF-OFF- OFF-OFF- OFF-OFF