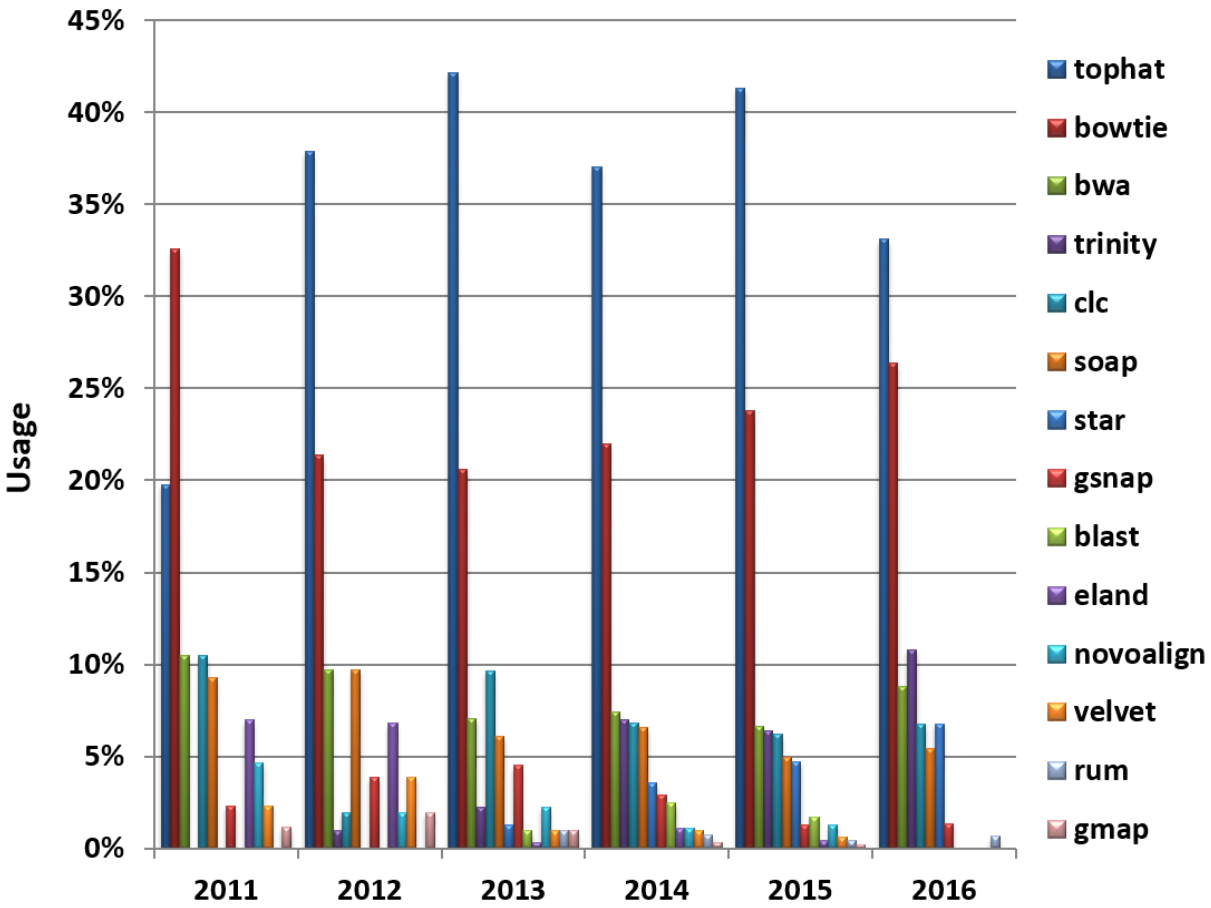
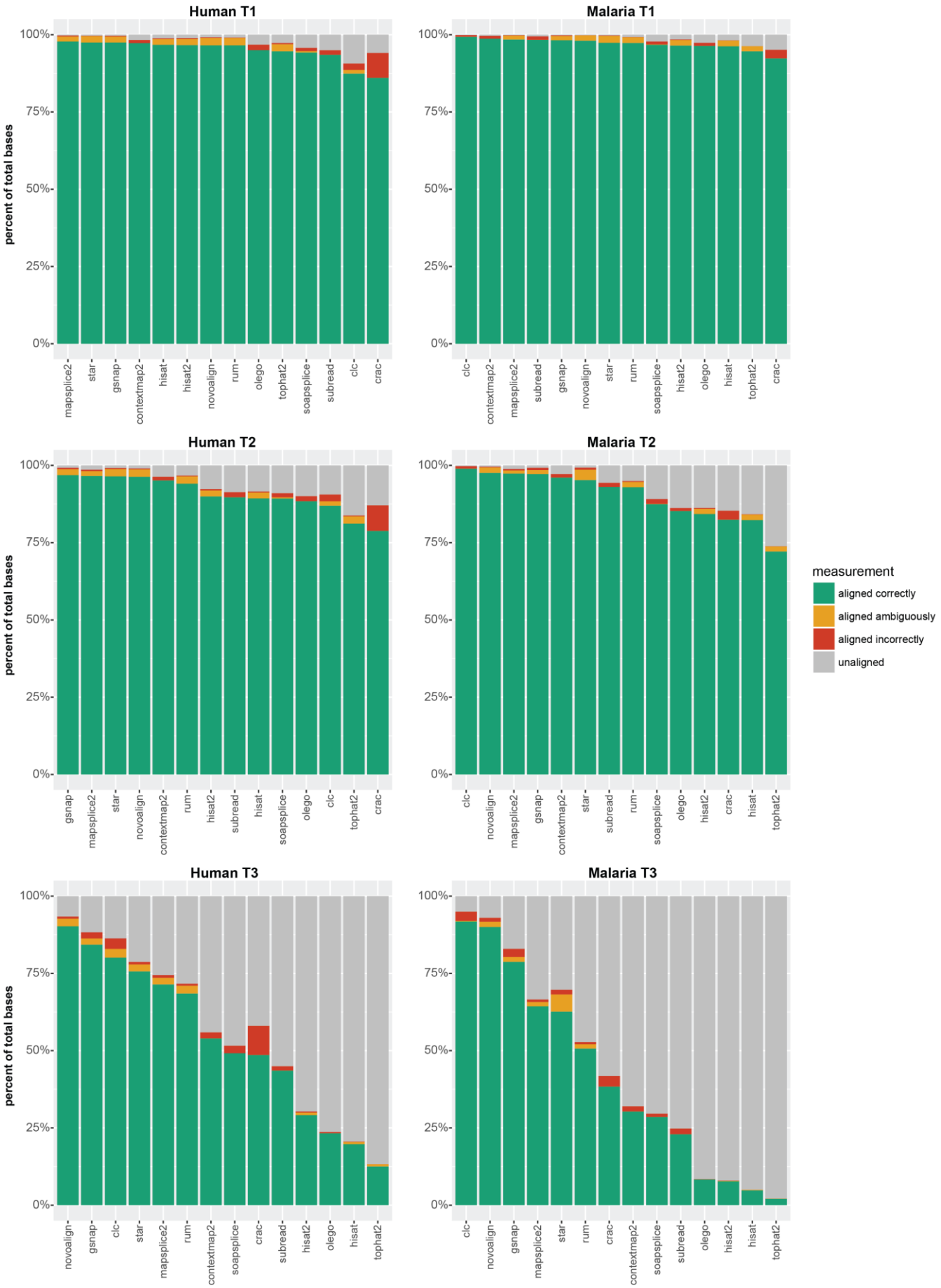
Supplementary Information

Table of contents

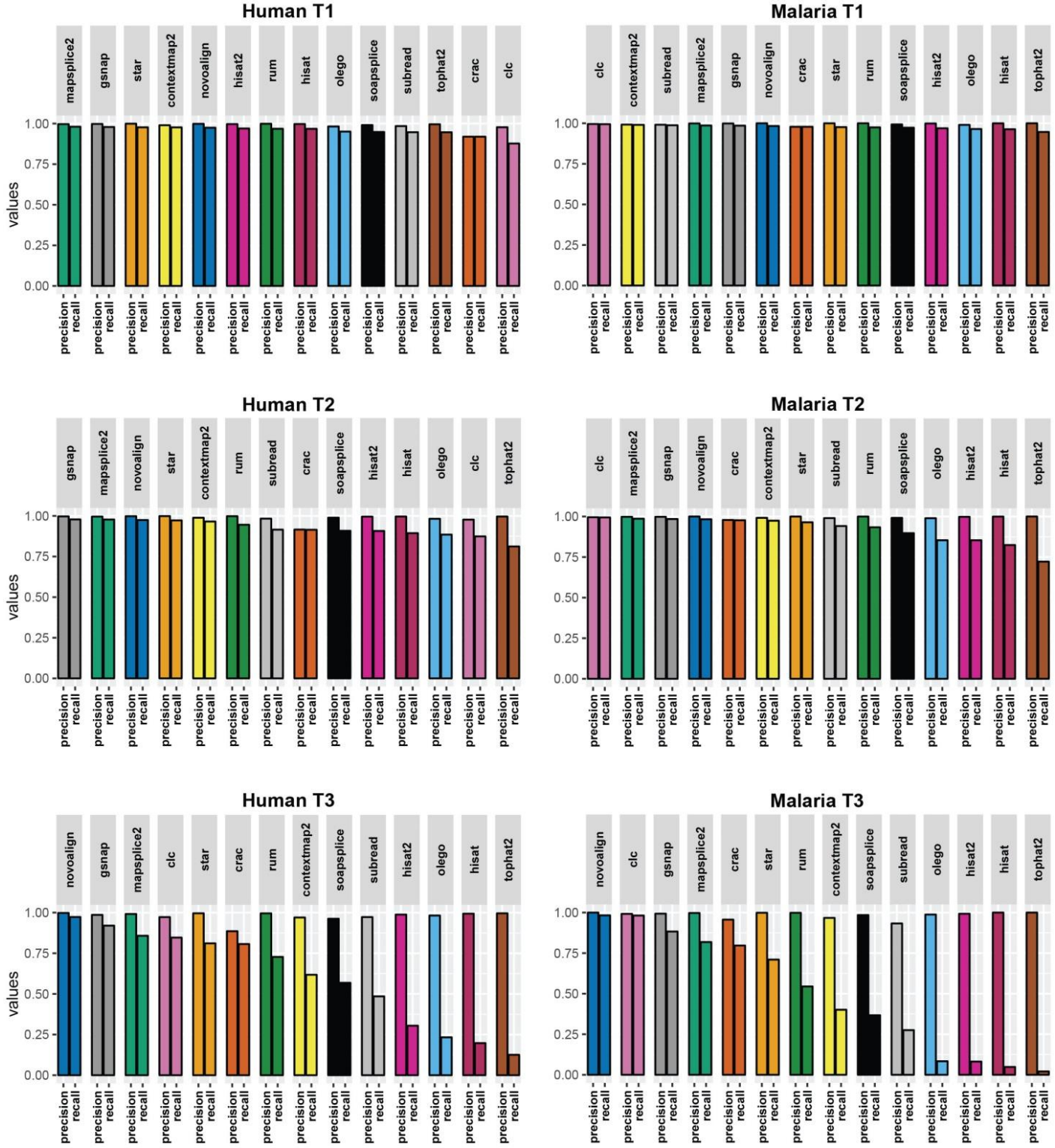
Figures



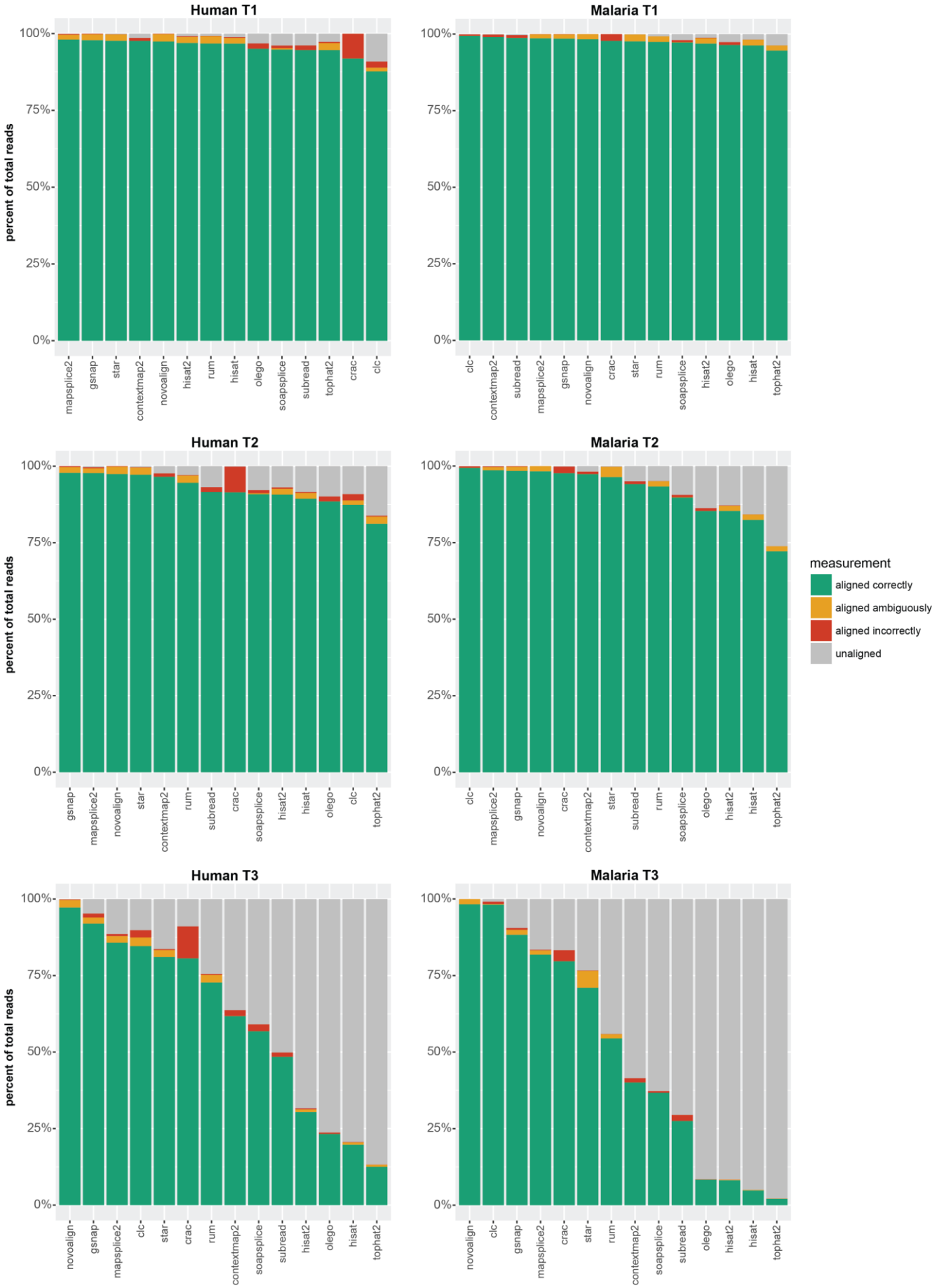
Supplementary Figure 1. Comparative usage of the 14 most popular algorithms. *Survey based on 2,000 randomly chosen peer reviewed papers, stratified by year.*



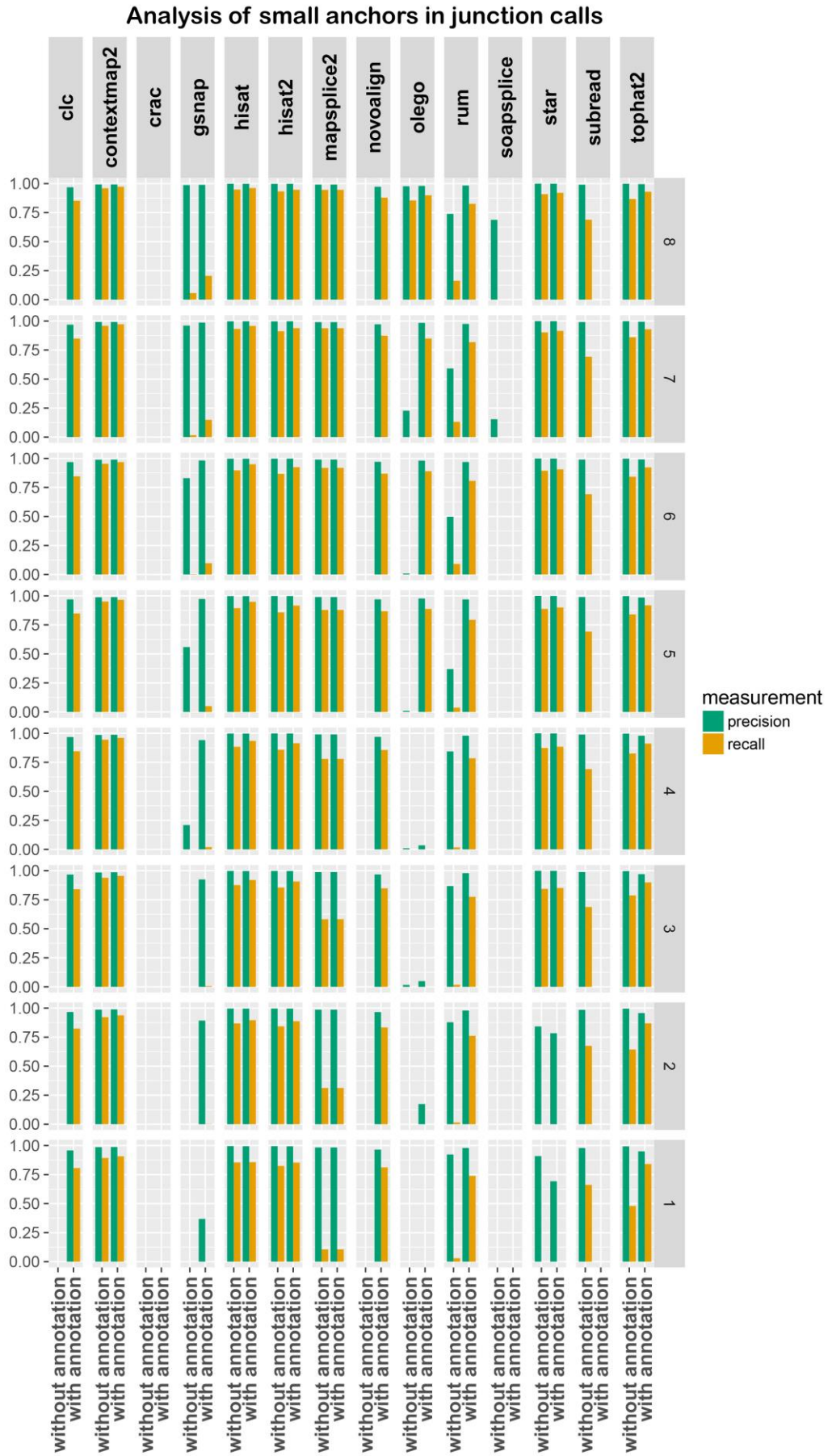
Supplementary Figure 2. Default parameters - Base level statistics for Human and Malaria datasets*. For each dataset, the bars show the percentage of bases aligned correctly, aligned ambiguously, aligned incorrectly and unaligned by each tool. The tools are sorted by descending percentage of bases aligned correctly. See how increasing the complexity of the dataset from T1 to T3, the aligning performances change drastically for the majority of the tools. Even on T1, the simplest dataset, the differences between the tool performances are visible. Except for the least polymorphic data, Malaria libraries are more difficult to align compared with Human data. Except for CLC Genomic Workbench, libraries with the same complexity show the same sets of best/worst performing tools on the two organisms. The results seem to be consistent between species. Moreover, CRAC shows the highest percentage of bases aligned incorrectly on all datasets. The figure highlights how much the different "default" could influence the accuracy of the alignment at base level.*



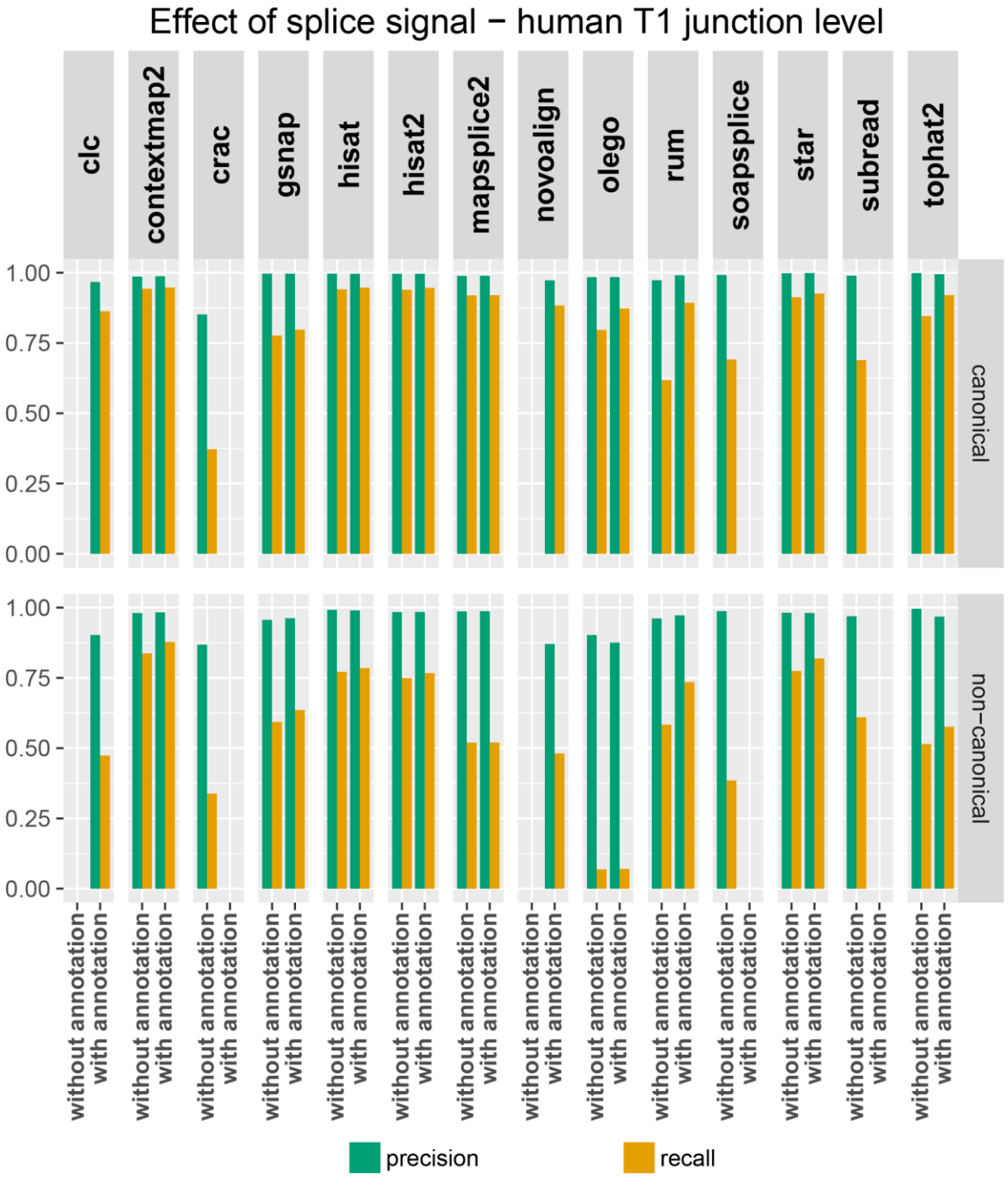
Supplementary Figure 3. Default parameters - Read level precision and recall for Human and Malaria datasets. *The tools are sorted by descending recall. CRAC, which does not perform very well at base level, here shows a different behavior. Olego, Tophat2, HISAT and HISAT2 are four of the worst performing tools on all datasets, except for Human T1. Curiously, CLC Genomic Workbench achieves one of the best recalls in all libraries, except for Human T1 and T2 where it is one of the worsts. As with the base level, the figure shows the important role of the different "default" settings on the quality of the alignment. With regards to precision, on T1 libraries, the tools perform generally very well (>97%) with CRAC as worst performer (~91.9% on Human T1).*



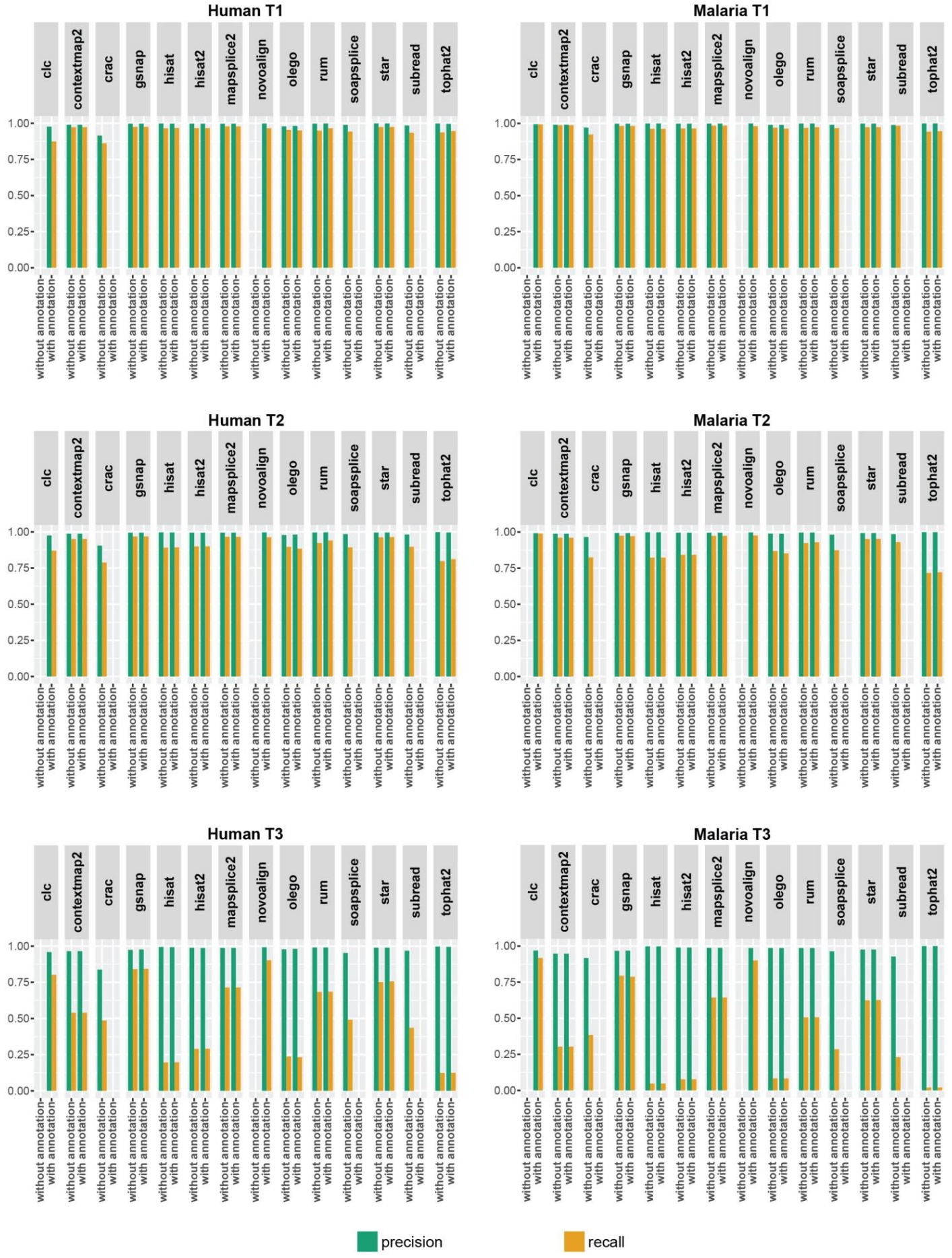
Supplementary Figure 4. Default parameters - Read level statistics for Human and Malaria datasets. *For each dataset, the bars show the percentage of reads aligned correctly, aligned ambiguously, aligned incorrectly and unaligned by each tool. The tool are sorted by descending percentage of reads aligned correctly. As for base level, increasing the complexity of the dataset shows a considerable change in the alignment accuracy for many tools. Surprising, the tools perform better on Malaria T1 than Human T1. Malaria is again the most challenging dataset on T2 and T3. As on base level, libraries having the same complexity show the same sets of best/worst performing tools across the organisms. CLC Genomic Workbench is the only exception, having different performances on Human and Malaria. On Malaria T2 and T3 libraries, STAR has one of the highest percentages of reads aligned ambiguously.*



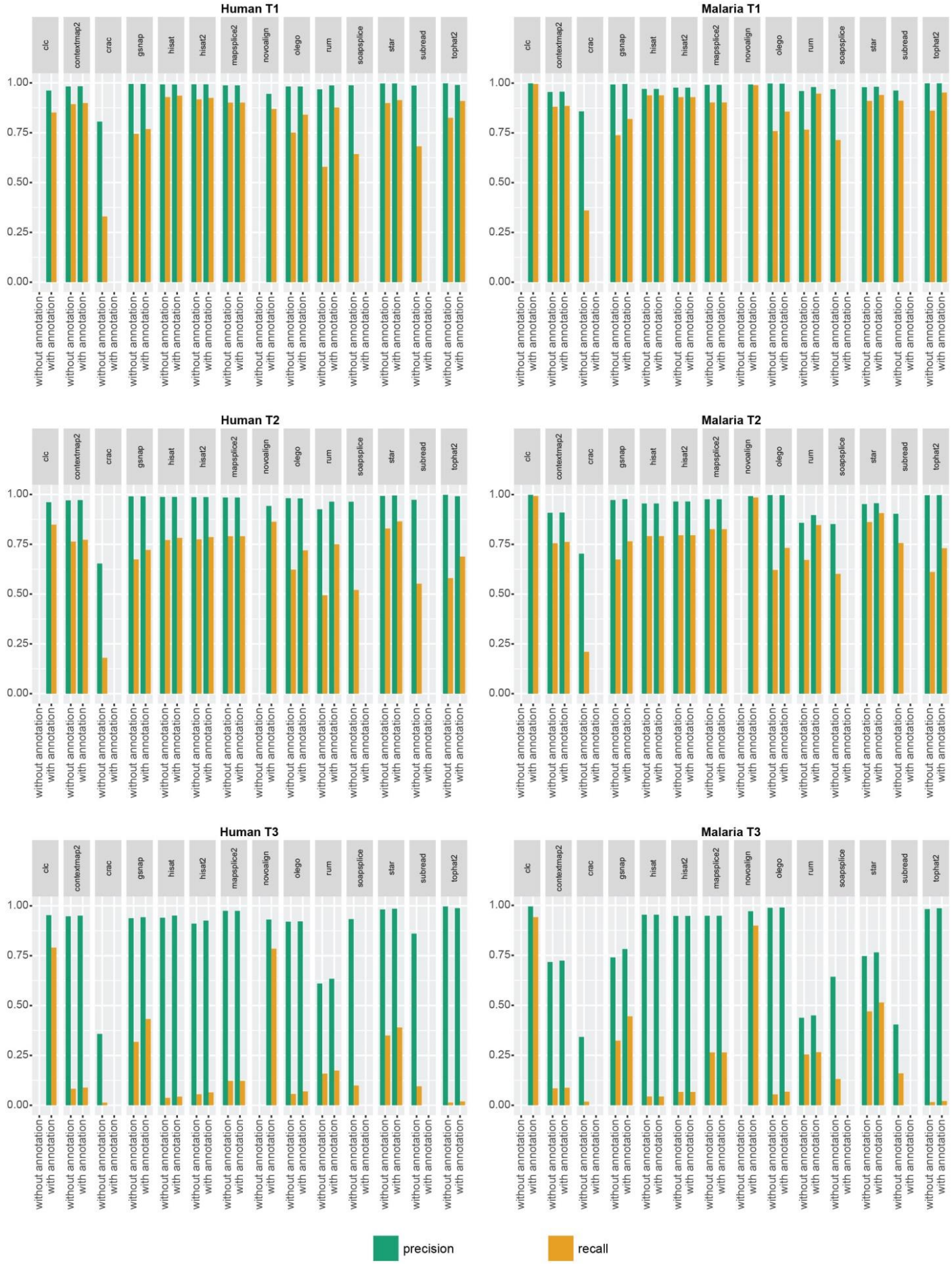
Supplementary Figure 5. Analysis of small anchors in junction calls. *Precision and recall are shown as a function of anchor size from one to eight bases in the human T1 dataset. An anchor is when only a few bases belong on one or the other side of a junction. The shorter the anchor the more difficult to align and the more beneficial annotation should be. Some methods require annotation and some cannot use annotation. Here we used RefSeq as generic annotation. Recall that the data were not generated using RefSeq, so this does not introduce bias. HiSAT, HiSAT2 and ContextMap2 perform very well on very short anchors even without annotation. Overall performance is highly variable. More discussion on this is given in the main paper Results section.*



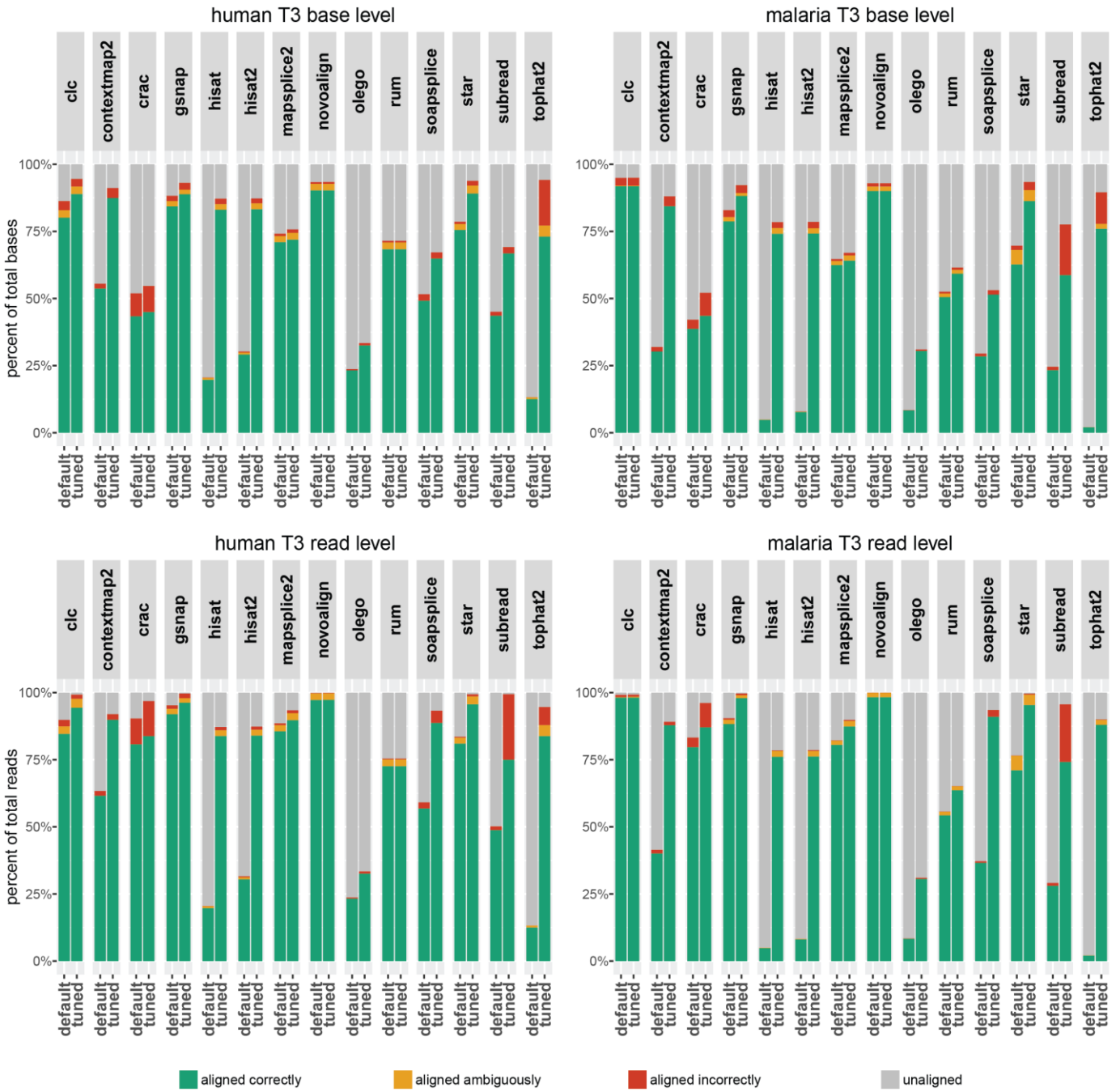
Supplementary Figure 6. Analysis of splice signal in junction calls. *Precision and recall are shown separately for canonical and non-canonical splice junctions. Evidently all algorithms have more trouble with non-canonical junctions than canonical. Surprisingly, annotation does not tend to help very much with this problem. ContextMap2, HiSAT, HiSAT2 and STAR do the best overall with GSNAP, RUM and TopHat2 showing moderate performance. OLEGO, CRAC and SoapSplice have the worst performance on non-canonical junctions.*



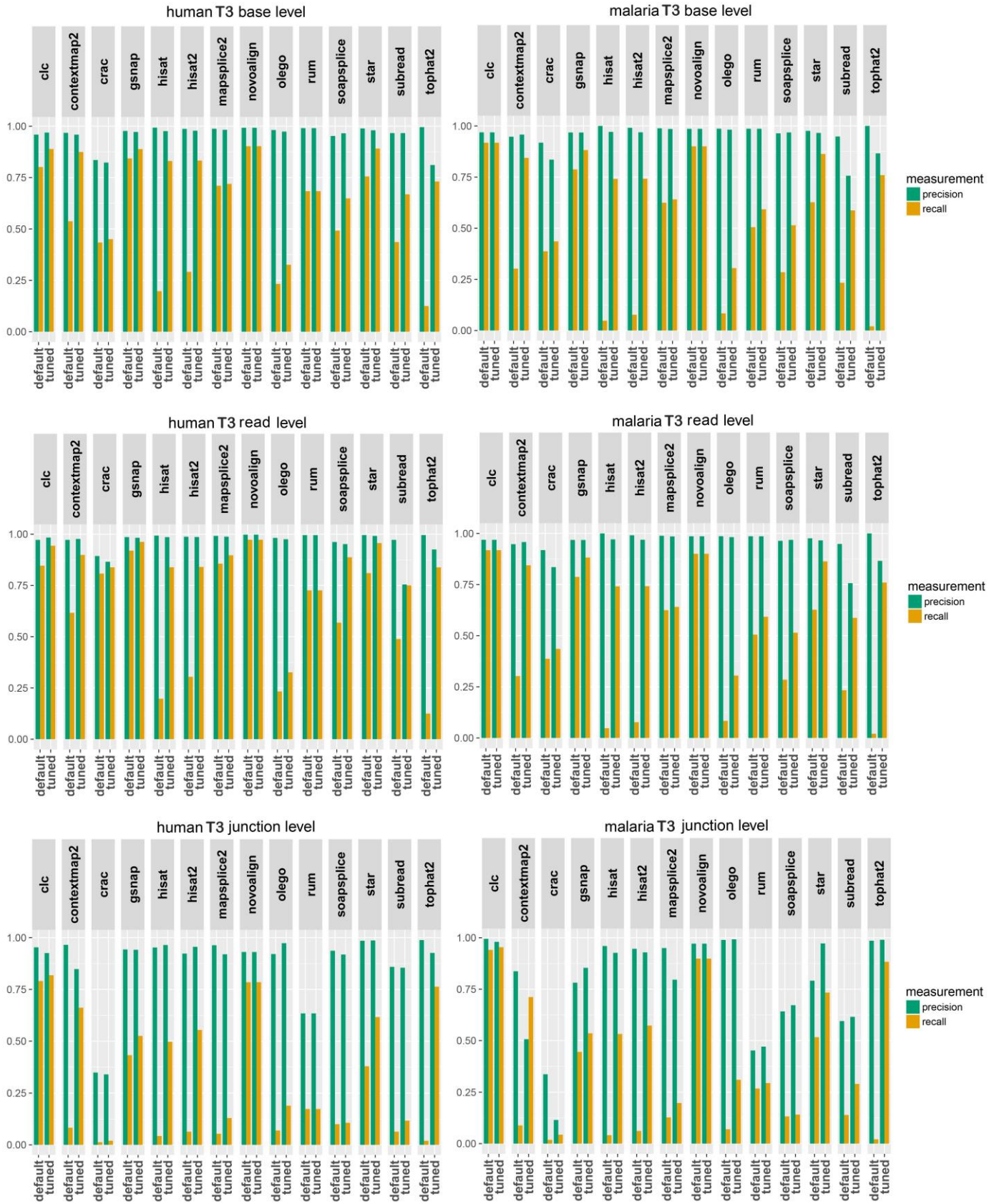
Supplementary Figure 7. Effect of the annotation at base level on precision and recall, for Human and Malaria datasets. *Note that using the annotation does not improve significantly the accuracy at base level..*



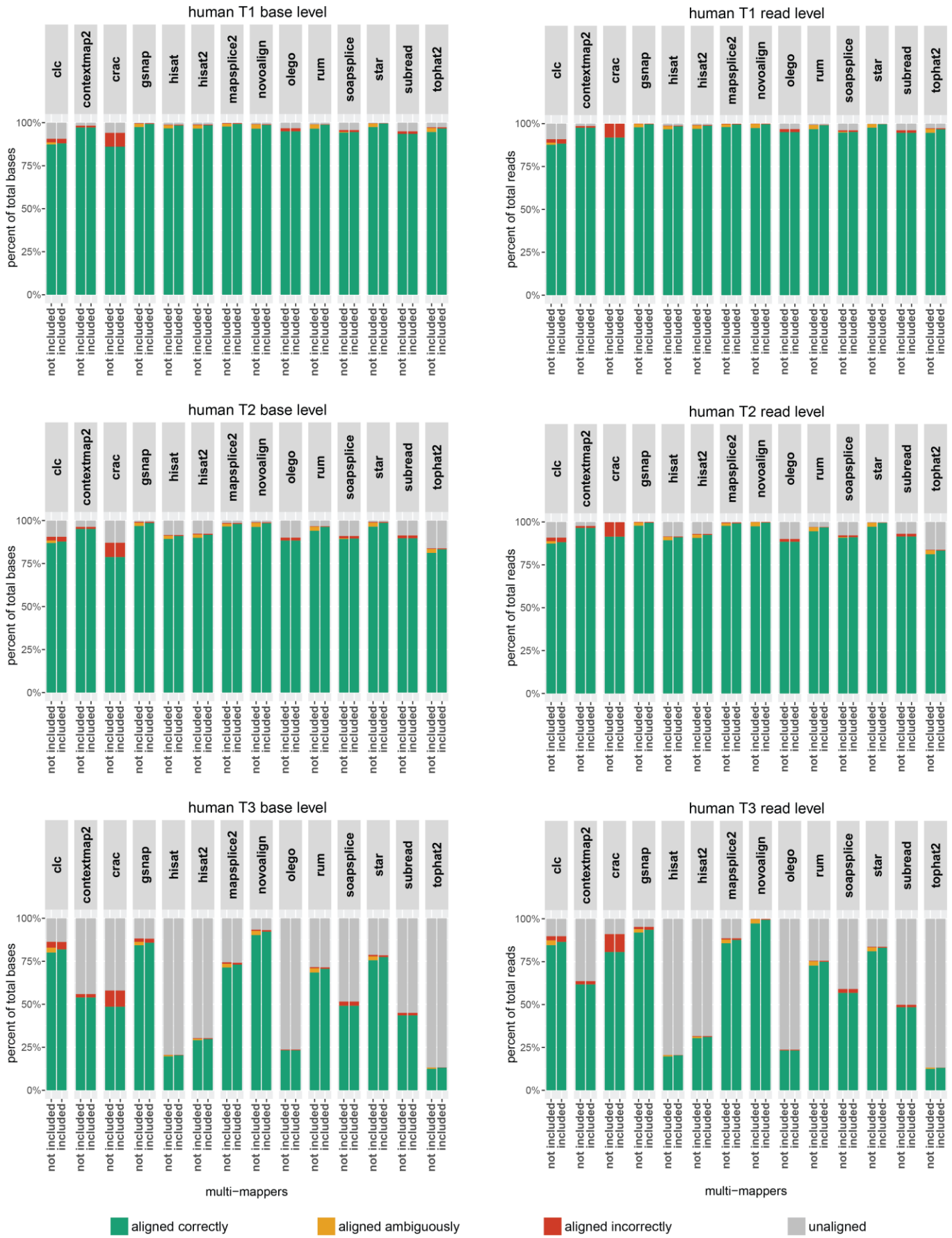
Supplementary Figure 8. Effect of the annotation at junction level for Human and Malaria datasets. *For each dataset, the bars show the precision and recall of each tool (without/with annotation). In contrast with base and read level, at the junction level the use of annotation brings some benefits in terms of recall. Almost every tool improves its recall using the annotation, while in a few cases the precision decreases.*



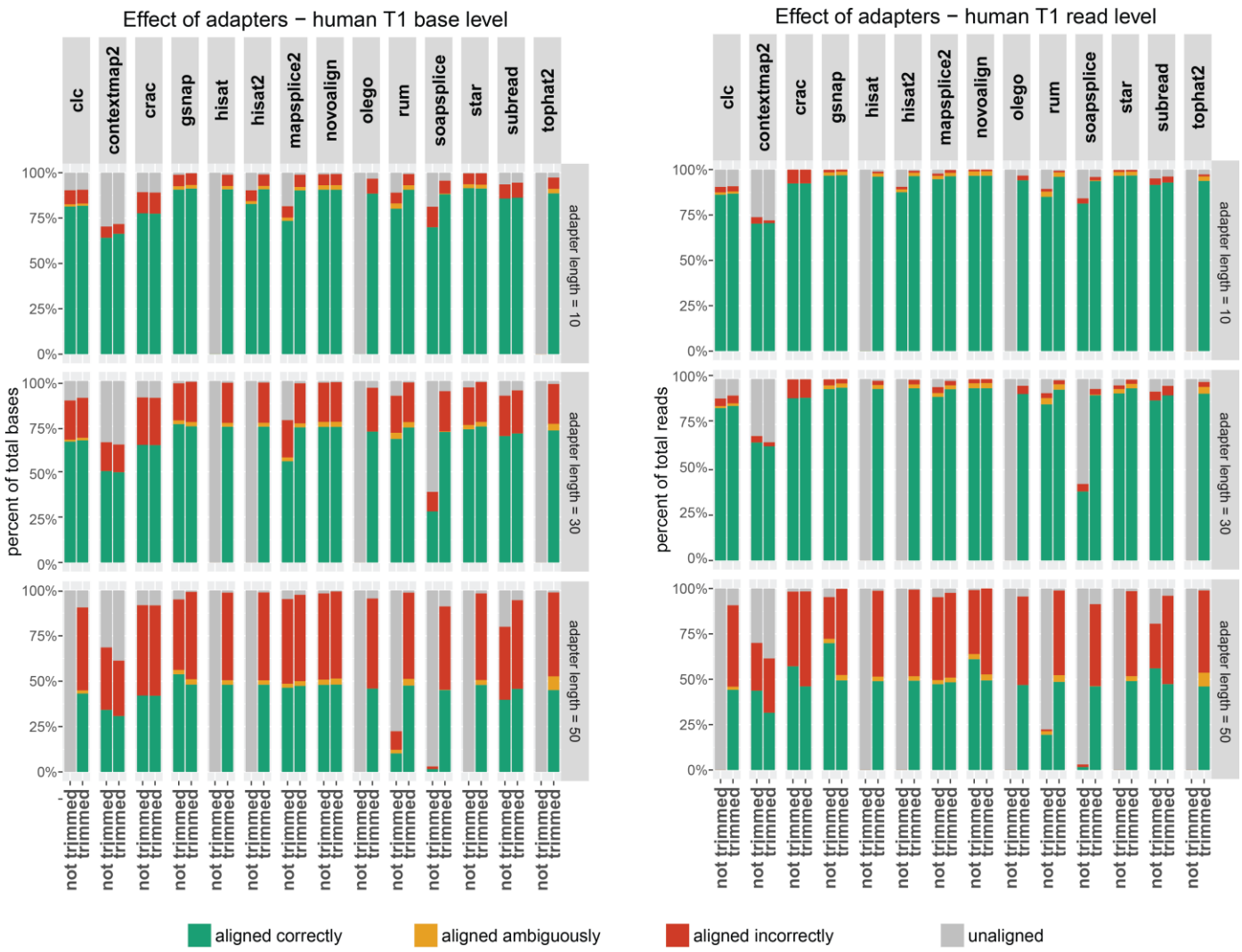
Supplementary Figure 9. Effect of parameter tweaking at base and read level for Human and Malaria datasets. *For each dataset, the bars show the percentage of bases (top) and reads (bottom) aligned correctly, aligned ambiguously, aligned incorrectly and unaligned by each tool. For each tool, the figure shows the alignment statistics for the "default" and the "tuned" alignments. The "tuned" alignment is the best configuration (in terms of base recall (top) and read recall (bottom)) achieved by the tweaking process. The figure highlights the difference between the default alignment and the best achievable mapping at base and read level. HISAT, HISAT2 and Tophat2 show the higher improvements in terms of percentage of bases aligned correctly, followed by Contextmap2, Soapsplice and Subread. For many tools, the greater number of bases aligned correctly brings also an increasing number of bases aligned incorrectly.*



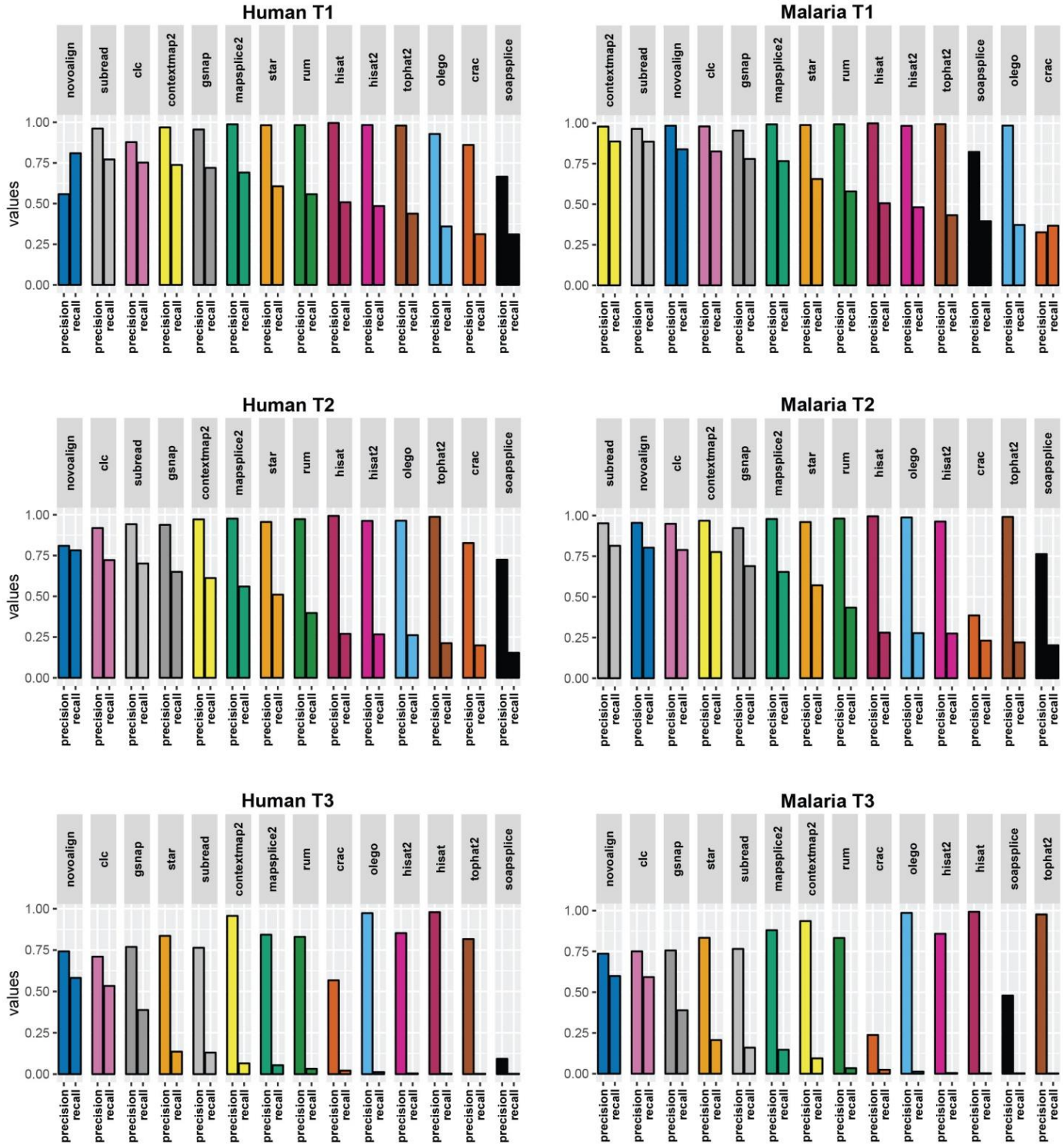
Supplementary Figure 10. Effect of parameter tweaking on precision and recall at base, read and junction level for Human and Malaria datasets. *For each tool, the figure shows the alignment statistics for the "default" and the "tuned" alignments. The "tuned" alignment is the best configuration (in terms of base recall(top), read recall(middle) and junction recall(bottom)) achieved by the tweaking process. HISAT, HISAT2 and Tophat2 show the higher improvements, followed by Contextmap2, Soapsplice and Subread. The "tuned" versions of CRAC, Subread and Tophat2 have often a considerable amount of reads aligned incorrectly, compared with their "default" versions.*



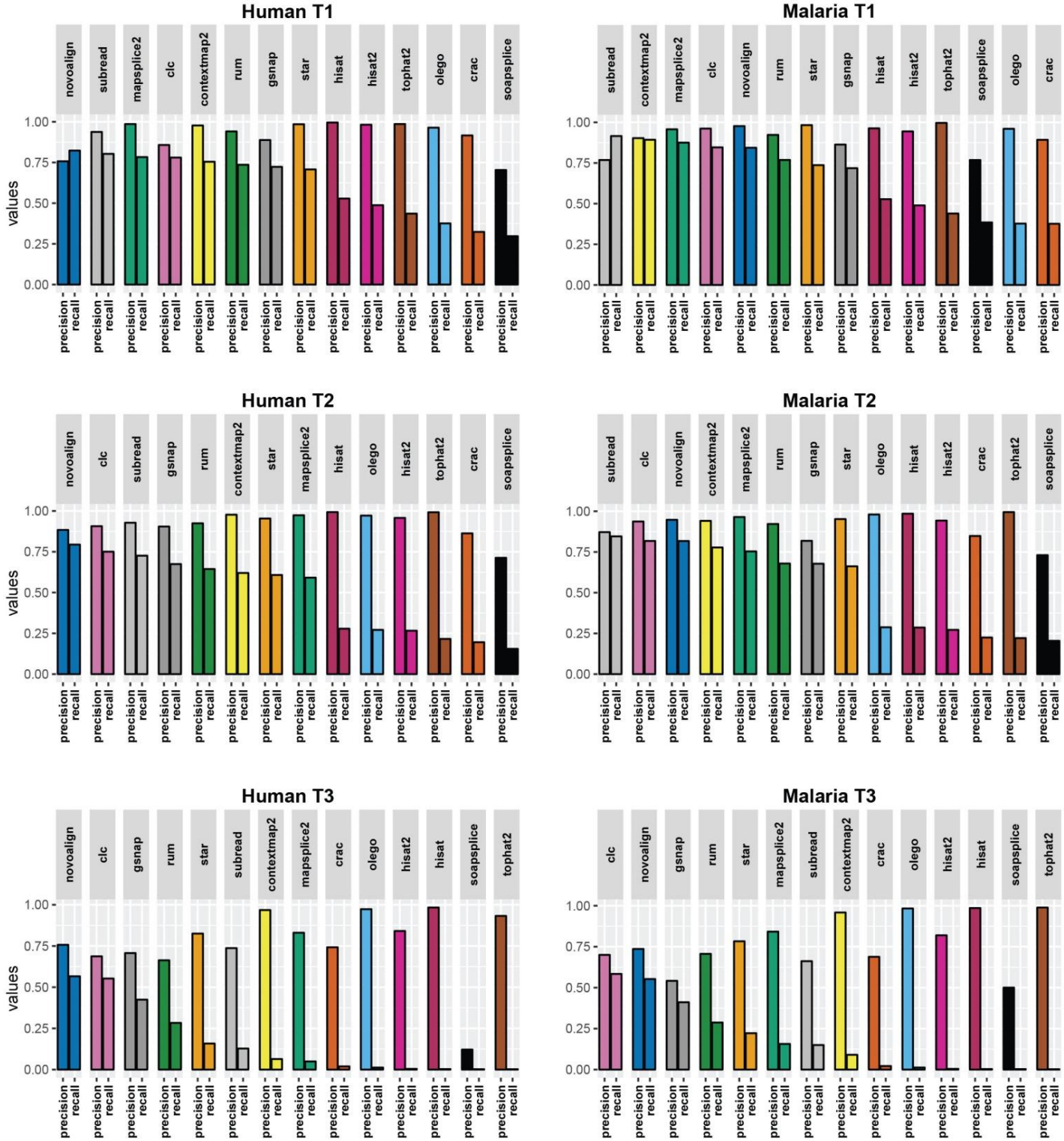
Supplementary Figure 11. Multi-Mapper analysis. *To identify the recall and precision in the case of multi-mapping fragments the record with the most correct bases aligned was chosen and any further calculations were based on this best alignment. Here the same statistics were calculated as introduced in the read and base level analysis section. The figure show the base (left) and read (right) level results on Human dataset. We observed that every tool shows a greater or equal recall and a lower or equal precision compared to not including multi-mappers.*



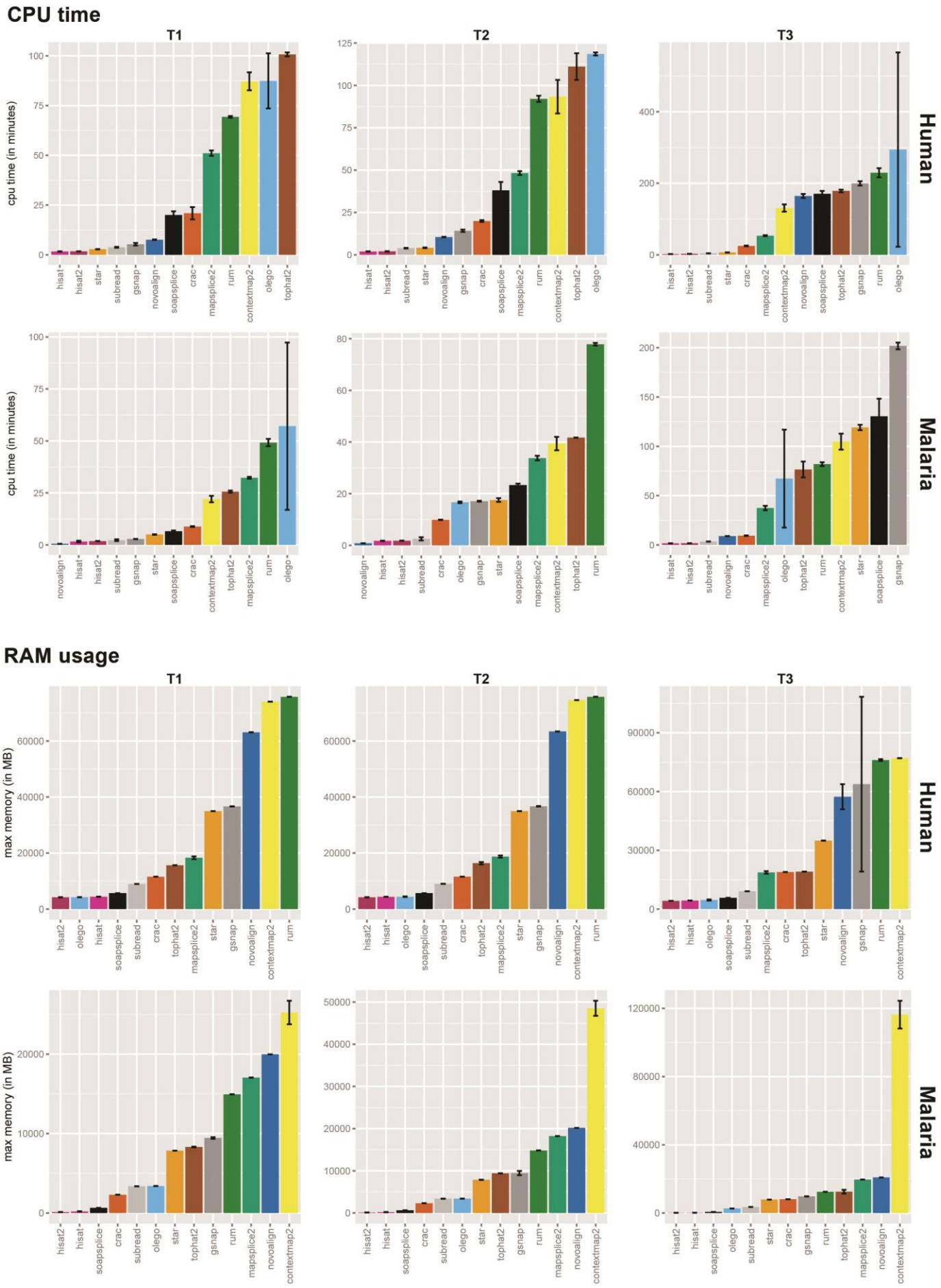
Supplementary Figure 12. Performance with varying lengths of adapter sequence. *Adapter sequence of lengths 10, 30 and 50 bases were added to reads and then they were aligned with and without trimming. The figure show the base (left) and read (right) level results on Human T1 dataset.*



Supplementary Figure 13. Default parameters – Insertion level precision and recall for Human and Malaria datasets. *The tools are sorted by descending recall. CLC Genomic Workbench, Contextmap2, GSNAP, Mapsplice2, Novoalign and Subread are the best performing tools on T1 and T2 libraries. On T3 libraries, only CLC Genomic Workbench, GSNAP and Novoalign have a recall greater than 30%. Comparing the recalls between the two organisms, many tools maintain the same position in the rank. HISAT has the highest precision on all libraries. CLC Genomic Workbench, CRAC, Soapsplice and Novoalign show always the worst precision on Human datasets. There is a similar trend on Malaria libraries, where also GSNAP and Subread show two of the lowest precisions.*



Supplementary Figure 14. Default parameters – Deletion level precision and recall for Human and Malaria datasets. *The tools are sorted by descending recall. Novoalign and CLC Genomic Workbench show the best recalls on T3 libraries, followed by GSNAP and RUM. Curiously, on Human and Malaria T1 and T2 libraries there is a significant difference between the performances of the first eight tools and the rest of the aligners. HISAT has the best precision on Human libraries, while Tophat2 has the best precision on Malaria dataset. Soapsplice has the worst performance on all libraries. One should use caution in interpreting a high precision in light of a low recall.*



Supplementary Figure 15. Performance in terms of CPU Time and RAM usage for Human and Malaria datasets. *Top: For each tool, the real CPU time was divided by the number of threads used (16). For each dataset and tool, the bars show the average CPU time (in minutes) computed on the three replicates. The error bars show the variability of this measure. The tools are sorted by ascending average CPU time. HISAT, HISAT2 and Subread are very often the fastest software. Except for the fastest tools, there is no clear trend between organisms and library complexities. Bottom: For each dataset and tool, the bars show the average maximum memory usage (in MB) computed on the three replicates. The error bars show the variability of this measure. The tools are sorted by ascending average memory usage. HISAT, HISAT2 and Soapsplice have the lowest maximum memory usage on Malaria dataset. On Human data, HISAT, HISAT2, Olego and Soapsplice have the lowest memory requirements.*

Supplementary Notes

Supplementary Note 1: Short Anchored Reads

Reads aligning across junctions represent perhaps the greatest challenge in the alignment problem. In particular, reads extended by only few bases in one of the spanned exons (less than 10bp), could be very difficult to align, keeping in mind that introns typically are thousands if not tens of thousands of bases in length. Indeed, these short sequences (heretofore referred to as "*anchors*") could be easily aligned to the adjacent next intron, soft/hard clipped, or aligned in a wrong position, depending on the algorithm. If an anchor has length one, then there is in fact a 25% change it will align perfectly to the adjacent intron and in general an anchor of length n has a $1/4^n$ chance of aligning to the adjacent intron perfectly.

Some aligners utilize information provided by an external annotation source to aid in defining the exon boundaries. This procedure allows then to improve the ability to accurately align short anchors. In order to study the ability of the tools to handle short anchors, we collected metrics specifically for the reads containing such anchors. First, reads having an anchor of length lower or equal than 8bp were identified in the R1 datasets (that is, replicate one, recall that there are three replicates of each). Only reads that span two exons and that had no indels were chosen, which should present one of the easiest scenario to the aligners. In this way, the results are mainly affected by the ability of each tool to handle the short anchored reads, more than their ability to manage other alignment issues.

The lists of reads filtering the .cig files were obtained, looking for reads having the CIGAR string in the format (100-y)MNNN...NNN(y)M or (y)MNNN...NNN(100-y)M, where is "y" is an integer value in the range [1, 8]. So for example 95M4142N5M and 8M4567N92M are valid CIGAR strings, while 90M4142N10M or 90M5D5M4142N5M are not valid. In this way, for each R1 library eight lists were obtained containing only the reads having an anchor of length 1,2, …, 8 bp respectively. This was done both for Human and Malaria datasets, resulting in a total of 48 lists.

For each tool, the reads contained in these lists were then extracted from the corresponding alignment (SAM) files. Since each aligner was run both providing and omitting the annotation as input, the results of both versions were collected. This allowed for the analysis of differential performance due to the information provided by the annotation. The results are available in **Supplementary Data 13**.

Supplementary Note 2: Alignment with and without Annotation

The impact of the annotation on the quality of the alignment was tested. In order to study this effect, for each tool the first replicate of each library (R1) was aligned using the default settings and providing/omitting the annotation as input. The two conditions were not compared on CLC Genomic Workbench and Novoalign, because these tools are able to perform spliced alignment only with annotation. Conversely, CRAC, SOAPsplice and Subread do not allow to provide any kind of annotation as input. The results of this comparison are collected in **Supplementary Data 9** and **Supplementary Data 10**, for human and malaria respectively.

The use of the annotation rarely provides a significant improvement at read and base level, both on Malaria and Human. On Human T1, the most significant difference at read and base level accuracy is shown by RUM and Tophat2. Providing the annotation as input to the two tools brings an improvement of ~1.5% and ~1% respectively. Similarly, on Malaria T1 the biggest difference is an improvement of 0.5% on RUM, providing the annotation. On T2 and T3 libraries, again the read and base level metrics show negligible differences. Using the annotation there is a relevant improvement only on Human T2, where the precision of RUM and Tophat2 increases of ~1.5%. On Human T3, the annotation helps STAR (in the 1-pass mode) improving the read and base recall of ~1.8%.

The use of annotation shows its most relevant effects at the junction level, both on Malaria and Human. Often, using the annotation allows to increase the junction recall while has a minor effect on the precision. The tools showing the greatest improvements are GSNAP, RUM, STAR and Tophat2. **Supplementary Table 1** summarizes the differences on these four tools.

Supplementary Note 3: Tweaking of Alignment Parameters

An extensive tweaking of the input parameters was performed for each tool. The first goal was to understand the full potential of each tool and determine how different optimal performance can be from the default. A tool having the smarter and more flexible defaults is preferable to one which requires changing the parameters significantly. In real data where no ground truth is available, a tuning process can be only performed in terms of percentage of mapped reads. However, mapping more reads does not necessary mean that they are mapped correctly. Moreover, often it is difficult to understand how to change the parameter values, especially when they relate to scoring/penalty functions. Usually, when a change in the default setting is performed, it is based on the output percentage of mapped reads or on some experience/intuition about the best value of the parameters. Again, without a ground truth this process is highly sensitive to error. For this reason, the defaults are very often used.

The second purpose of the tweaking tests is to develop practical suggestions to the users. The importance of some parameters is shown, which seem to be independent from the particular dataset/organism. When possible, quantitative suggestions are given about the best value/range for these parameters. More generally, some qualitative indications are given about the role of the parameters and the trend they show.

Tweaking was performed on T3 complexity data, since that is where there is the greatest room for improvement - in T1 some algorithms align 90-95% of the reads correctly so these would not serve to illustrate the relative impact of the parameters effectively. For this reason, the malaria library T3R1 was used, and in particular 1M reads were randomly sampled for the analysis. The subset of reads was used in order to expand the feasible number of alignments. Then the best configurations obtained for malaria T3R1 were applied to the human T3R1 library to make sure the results are not species specific. Since the data complexity is the same (they are both T3 libraries), this test allows for the evaluation of which parameters are more dependent on the quality of data and which are more related to the organism.

Since the number of available parameters and values for each tool defines an exponential search space, an exhaustive search of all the possible configurations would be unfeasible. Therefore, for each tool, we chose the subset of parameters and values to test following these criteria:

i) Suggestions/indications found on the official tool manual/website/user guide about the role and importance of each parameter.
ii) Use of common parameters that often have a major role on the alignment (e.g. number of allowed mismatches, seed length, etc.).

Moreover, we queried each tool's authors for suggestion about the most important set of parameters. For those who responded (about half) their suggestions were included.

The different number of parameters and configurations tested for each tool depends on:

i) the available number of parameters (and value ranges) provided by each tool
ii) the complexity in the individuation of the most influential parameters/values
iii) the computational effort related to each configuration
iv) the suggestions provided by the tools' authors

To the best of our knowledge, these tests make this the most comprehensive RNA-Seq aligner tuning study performed to date.

The results of the "*tweaked*" alignments on human and malaria are presented in **Supplementary Data 2** and **Supplementary Data 3** respectively.

The best "*tweaked*" configurations achieved at the end of the tuning process are summarized in **Supplementary Data 7** and **Supplementary Data 8**, for human and malaria respectively. When more than one configuration achieves the best results, one was chosen to be reported in these summary files.

CLC Genomic Workbench

Alignment - tweaking:

For the tweaking, the parameters in **Supplementary Table 2** were tested. About 20 different combinations of these values were used during the tweaking process. Few examples are presented in **Supplementary Table 3** and **Supplementary Table 4**.

Malaria

At read, base and junction level the mapping option *"Also map to inter-genic regions"* slightly decreases the alignment accuracy. The *"maximum number of hits for a read"* is one of the most important parameters since a value greater or equal than 10 improves the recall. At read level, the default is one of the best configurations. The cost parameters and the similarity/length fractions do not affect significantly the results. At base level and junction level (junction precision), again the default is one of the best configurations. However, here the cost parameters have a major role: the defaults perform significantly better than any other tested values. On the other hand, the junction recall is improved using cost parameters lower than the defaults.

Human

As with Malaria, the *"maximum number of hits for a read"* is one of the most important parameters and a value greater or equal than 10 improves the recall. At read and base level, the mapping option "*Also map to inter-genic regions"* helps to improve the recall of a value between 5% and 7% compared with the default setting *"Map to gene regions only (fast)"*. At read level, the best configuration is about 10% more accurate than the default (recall ~94% compared with recall ~84%). Moreover, using a *"maximum number of hits for a read"* of 30 instead of 10 (default) significantly improves the read level recall. At base level and junction level (junction precision), the cost parameters play a major role: the defaults perform significantly better than any other tested values. Again, using cost parameters lower than the defaults improves the junction recall.

Conclusions

The adoption of a *"maximum number of hits for a read"* greater or equal than the default (10) seems to improve the results. The main difference between Malaria and Human is related to the mapping options: the option "*Also map to inter-genic regions"* makes the results slightly worse on Malaria while improves them on Human. At base level and junction level (junction precision), the default values for the cost parameters perform significantly better than any other tested values. On the other hand, at read level and junction level (junction recall) the adoption of cost parameters lower than default results in an improvement.

Summarizing, the default on Malaria is very close to the best achieved. On human, the default is good (compared with the other tools), but the tweaking could still improve the alignment results.
In order to balance the performances at base, read and junction level the default setting plus *"maximum number of hits for a read"* set at 30 seems a good choice. On human, the option "*Also map to inter-genic regions"* could increase the quality of the results.

Contextmap2

Alignment - tweaking:

```
java -Xms16000M -Xmx128000m -XX:+UseConcMarkSweepGC -XX:NewSize=300M -
XX:MaxNewSize=300M -jar ContextMap_v2.6.0.jar mapper -reads <reads file> --pairedend -
gtf <gtf file> --noncanonicaljunctions -aligner_name bwa -aligner_bin <bwa path> -
indexer_bin <bwa path> -indices <bwa index> -genome <genome directory> -o <output
path> -t 16 -seed <SEED> -seedmismatches <SEED_MISMATCHES> -mismatches <MISMATCHES> -
mmdiff <MMDIFF> -maxhits <MAXHITS> -minsize <MINSIZE>
```

For the tweaking, the parameters in **Supplementary Table 5** were tested. About 950 different combinations of these values were used during the tweaking process. Few examples are presented in **Supplementary Table 6** and **Supplementary Table 7**.

Malaria

At read and base level, *SEED* and *MISMATCHES* are the most important parameters. Changing the default *SEED* parameter results in decreased recall in our tests. On the other hand, a value greater than the default for *MISMATCHES* can significantly improve the recall. At read level, starting from the default configuration and progressively changing only *MISMATCHES* from 4 (default) to 30, changes the read level recall from ~40% to ~87%. At base level, the same progressive change results in increasing the base level recall from ~30% to ~84%. At both levels, values greater than 30 for *MISMATCHES* does not bring any further improvements. At the junction level, the recall shows the same trend of base and read level with regards to *SEED* and *MISMATCHES*. However, here increasing the *MISMATCHES* over 30 slightly improves the results.
On the contrary, the junction level precision increases when *MISMATCHES* is lower or equal to the default: the junction precision is ~70% with the default *MISMATCHES* while it becomes ~9% using *MISMATCHES* set to 30. Moreover, here a *SEED* greater than the default improves the junction recall.

Human

Applying on Human the best configurations found for Malaria, the same trends were observed: increasing *MISMATCHES* and leaving the default for *SEED* allows for improving the recall at the read and base level. The read level recall increases by ~28% while the base level recall increases by ~33%. Moreover, there is an improvement in the junction level recall of ~11%. The junction level precision shows the same opposite behavior already seen on human.

Conclusions

The parameters *SEED_MISMATCHES*, *MMDIFF*, *MAXHITS_VALUES* and *MINSIZE_VALUES* do not affect significantly the results at any level. The default values seem to be a good choice for these parameters. Both organisms highlight the important role of *MISMATCHES* and *SEED*. Increasing the first one and not changing the second one is very beneficial for the majority of metrics. Determining a value that balances the junction level precision and recall seems difficult.

CRAC

Alignment - tweaking:

```
crac -i <crac index> -k <K> -r <reads file 1> <reads file 2> --sam <output sam file> -
-reads-length <read length> --no-ambiguity --max-locs <MAX_LOCS> --min-percent-single-
loc <MIN_PERCENT_SINGLE_LOC> --min-percent-multiple-loc <MIN_PERCENT_MULTIPLE_LOC> --
summary <summary file> --nb-threads 16
```

For the tweaking, the parameters in **Supplementary Table 8** were tested. About 440 different combinations of these values were tested during the tweaking process. Few examples are presented in **Supplementary Table 9** and **Supplementary Table 10**.

Malaria

First, several values of *K* were tested, since there is no default for Malaria.

At the read (base) level, increasing *K* from 20 to 27 one unit at time causes a decrease in the recall of 3-4% (1-2%) at each step. Conversely, decreasing *K* from 20 to 18 improves the recall by ~2%. The same trend is shown by the junction level recall but with a smaller effect: the increasing/decreasing of the recall is lower than 0.5 for each unit of *K*. From these tests, the optimal value for malaria could be around 18. The junction precision shows the opposite trend, providing better results when the value of *K* is increased. From *K*=18 to *K*=27, the junction precision increases from ~21% to ~48%. Increasing *MAX_LOCS* and leaving the option `--no-ambiguity` unset provides very small benefits in terms of recall (<0.5%). The parameters *MIN_PERCENT_SINGLE_LOC* and *MIN_PERCENT_MULTIPLE_LOC* do not affect the results.

Human

At the read, base and junction levels a value of *K*=19 or *K*=20 is the best choice on human, even if *K*=18 (the best value on malaria) and *K*=22 (suggested value for human) still provides decent results. Again, only on the junction level precision does a bigger *K* seem to improve the results: with *K*=22 the junction precision is ~35%, with *K*=27 it goes to ~37.5%.

Conclusions

As stated in the manual, the most important parameter is *K*. The right value for this parameter depends on the organism more than on the quality of the data. The value *K*=22 proposed for Human seems to be a good choice, even if we achieved slightly better results using *K*=19 or *K*=20. On the Malaria datasets, we suggest a value around 18. The other tested parameters show negligible effects.

GSNAP

Alignment - tweaking:

```
gsnap -D <index output path> -d <index name> -A sam --max-mismatches <MAX_MISMATCHES>
--indel-penalty <INDEL_PENALITY> --gmap-min-match-length <GMAP_MIN_MATCH_LENGTH> --
pairexpect <PAIR_EXPECT> --pairdev <PAIR_DEV> --merge-distant-samechr --ordered --
novelsplicing 1 --use-splicing <index name>.splicesites --nthreads 16 --batch 5 --
expand-offsets 1 <read file 1> <read file 2> > <output sam file>
```

For the tweaking, the parameters in **Supplementary Table 11** were tested. About 630 different combinations of these values were tested during the tweaking process. Few examples are presented in **Supplementary Table 12** and **Supplementary Table 13**.

Malaria

The most important parameter is *GMAP_MIN_MATCH_LENGTH*, followed by *MAX_MISMATCHES*. A *GMAP_MIN_MATCH_LENGTH* lower than default (20) with a *MAX_MISMATCHES* greater or equal than 15 results in an increase in the read and base recall of ~9% compared to the default. The effects of *GMAP_MIN_MATCH_LENGTH* and *MAX_MISMATCHES* seem related, so changing only one of the two parameters did not bring any improvements. In addition, a small *INDEL_PENALITY* (lower or equal than default) slightly improves the base and read level recall. At the junction level there is not a clear trend. Again, a high *MAX_MISMATCHES* (>15) helps to improve the precision, but the role of *GMAP_MIN_MATCH_LENGTH* here is not so clear. About the junction recall, a *GMAP_MIN_MATCH_LENGTH* equal or lower than default still improves the results. The role of *MAX_MISMATCHES* is unclear (the best configurations have a *MAX_MISMATCHES* of 2-3 or around 10). The parameters related to fragment length (*PAIR_EXPECT* and *PAIR_DEV*) do not affect the results.

Human

Applying to Human the best base/read recall configurations found on Malaria, we see an improvement of ~4% on the same metrics. So the importance of *GMAP_MIN_MATCH_LENGTH* and *MAX_MISMATCHES* seems to be confirmed also on this organism. Moreover, the best Malaria configurations at the junction level (best recall) show an improvement in the same metrics when applied to Human.

Conclusions

The important role of *GMAP_MIN_MATCH_LENGTH* and *MAX_MISMATCHES* is confirmed on both Malaria and Human. On T3 library, using a *GMAP_MIN_MATCH_LENGTH* lower than default with a *MAX_MISMATCHES* $\geq$ 15 allows increasing all the metrics. On Malaria for example, all the metrics have an improvement between 3% and 9% simply increasing *MAX_MISMATCHES* to 15 and decreasing *GMAP_MIN_MATCH_LENGTH* to 15.

HISAT

Alignment - tweaking:

`hisat --threads 16 --end-to-end -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> --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>`

For the tweaking, the parameters in **Supplementary Table 14** were tested. About 820 different combinations of these values were tested during the tweaking process. Few examples are presented in **Supplementary Table 15** and **Supplementary Table 16**.

Malaria

At base and read level, the most important parameters are *MAX_MISMATCH_PENALITY* and *MIN_MISMATCH_PENALITY*. Using a small value for both parameters (1 and 0/1, respectively) results in a recall of greater than 70%, while the default recall is ~5%. Moreover, a value of *PENALITY_NONCANONICAL* greater than the default seems to slightly improve the results. A value between 12 and 20 brings a recall improvement, while after 20 there is a saturation effect. The same considerations for read and base level are still valid for the junction level recall. The default recall is ~4%, while using the previous values it becomes ~52%. With regards to *PENALITY_NONCANONICAL*, increasing this value results in a slight improvement in the junction precision. On the other hand, using a value lower than default decreases the precision (96% using the default, 79% using *PENALITY_NONCANONICAL* = 3 and 66% using *PENALITY_NONCANONICAL* = 0).

Human

The importance of *MAX_MISMATCH_PENALITY*, *MIN_MISMATCH_PENALITY* and *PENALITY_NONCANONICAL* is highlighted also in Human. In particular, the best configurations of these parameters found on Malaria (1, 0 and 20 respectively) bring a significant improvement on Human. The read and base level recall increases more than 62% while the junction recall has an improvement of ~45%. The junction precision shows a lower improvement, increasing only by 1.2%.

Conclusions

Tests suggest that *MAX_MISMATCH_PENALITY*, *MIN_MISMATCH_PENALITY* and *PENALITY_NONCANONICAL* have a very important role in the quality of alignment. Both on Human and Malaria, there is a significant improvement on all metrics setting these parameters properly. The Bowtie2-like parameters (*NUM_MISMATCH*, *SEED_LENGTH*, *SEED_INTERVAL*, *SEED_EXTENSION* and *RE_SEED*) seem have no effects on the quality of the results. The use of Bowtie2-like parameter `--end-to-end` does not change the results.

HISAT2

Alignment - tweaking:

```
hisat2 --threads 16 --end-to-end -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>--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>
```

For the tweaking, the parameters in **Supplementary Table 17** were tested. About 830 different combinations of these values were tested during the tweaking process. Few examples are presented in **Supplementary Table 18** and **Supplementary Table 19**.

Malaria

The behavior of HISAT2 parameters is very similar to HISAT: *MAX_MISMATCH_PENALITY* and *MIN_MISMATCH_PENALITY* are the most important parameters at base and read level. A small value for both parameters (1 and 0/1, respectively) results in a recall greater than 70%, while the recall achieved using the default is only ~8%. Again, a value of *PENALITY_NONCANONICAL* greater than the default seems to slightly improve the results. A value between 12 and 20 brings an improvement while after 20 there is a saturation effect. At the junction level, the default recall is ~6%, while using the previous values gives ~60%. Similar to HISAT, the parameter *PENALITY_NONCANONICAL* has an important role in the junction level precision, followed by *MAX_MISMATCH_PENALITY* and *MIN_MISMATCH_PENALITY*. Using a value of *PENALITY_NONCANONICAL* lower than default decreases the precision (~95% using the default, ~75% using *PENALITY_NONCANONICAL* = 3 and ~56% using *PENALITY_NONCANONICAL* = 0). In contrast with read and base level, increasing *MAX_MISMATCH_PENALITY* from 1 to 6 brings an improvement in the junction precision. At any level, using a *MAX_SOFTCLIPPING_PENALITY* greater or equal than default (like 2 or 3) slightly improves the results. The junction precision shows the biggest improvement, gaining more than 1% for some configurations. On the other hand, *MIN_SOFTCLIPPING_PENALITY* seems to have no effect on the results.

Human

On Human, *MAX_MISMATCH_PENALITY*, *MIN_MISMATCH_PENALITY* and *PENALITY_NONCANONICAL* are still the most important parameters. Using one of the best Malaria settings of these parameters (1, 0 and 20 respectively) results in a significant improvement on Human. The read and base level recall increases more than 52% and the junction recall has an improvement of ~49%. The junction precision shows a low improvement, increasing only ~3%. Again, using a *MAX_SOFTCLIPPING_PENALITY* greater than or equal to the defaults results in some benefits.

Conclusions

In conclusion, the most important parameters seem to be *MAX_MISMATCH_PENALITY*, *MIN_MISMATCH_PENALITY* and *PENALITY_NONCANONICAL*, both on Human and Malaria. Changing these parameters can result in a huge improvement on the results, especially in terms of recall. Moreover, *MAX_SOFTCLIPPING_PENALITY* shows a secondary role in the quality of the alignments.

The Bowtie2 parameters (*NUM_MISMATCH*, *SEED_LENGTH*, *SEED_INTERVAL*, *SEED_EXTENSION* and *RE_SEED*) seem have no effects on the alignment. The use of Bowtie2 parameter `--end-to-end` does not change the results.

Mapsplice2

Alignment - tweaking:

```
python mapsplice.py --threads 16 --min-map-len <MIN_MAP_LENGTH> --splice-mis <SPLICE_MISMATCHES> --max-append-mis <APPEND_MISMATCHES> --ins <INSERTION_LENGTH> --del <DELETION_LENGTH> --filtering <FILTER> --non-canonical-double-anchor --output <output path> -c <genome fasta files> -x <index name> -1 <read file 1> -2 <read file 2>
```

For the tweaking, the parameters in **Supplementary Table 20** were tested. About 1080 different combinations of these values were used during the tweaking process. Few examples are presented in **Supplementary Table 21** and **Supplementary Table 22**.

Malaria

At read level, the most important parameter is *MIN_MAP_LENGTH*. Using a value of 25 or lower (15 and 20 in our tests) increases the recall by ~7%. A saturation effect is seen using values lower than 25. Other important parameters are *APPEND_MISMATCHES* and *SPLICE_MISMATCHES*. The first brings an improvement when set to 2 or 3, while the second one is best when set to 0. At base level, the tweaking of the parameters does not result in any significant improvement. The difference between the best achieved recall and the default recall is ~1.5%. The most influent parameter is *APPEND_MISMATCHES*, where a value of 2-3 helps to improve the recall. Again, a value of *MIN_MAP_LENGTH* lower or equal to 25 results in some benefits. In contrast to the read level, a high value of *SPLICE_MISMATCHES* (2) gives a slight improvement in the recall. At junction level, the recall is affected mainly by *APPEND_MISMATCHES* and *MIN_MAP_LENGTH*. A small value of *APPEND_MISMATCHES* (0 or 1), and a value of *MIN_MAP_LENGTH* lower than default (best achieved using 25), results in an improvement of ~7%. The junction level precision shows an improvement when adopting a high *MIN_MAP_LENGTH* (66 or 77) together with a low *SPLICE_MISMATCHES* (0) and the default value for *APPEND_MISMATCHES*. *INSERTION_LENGTH*, *DELETION_LENGTH* and *FILTER* seem to have negligible effects at all levels.

Human

Similar to Malaria, at the read level a *MIN_MAP_LENGTH* lower than the default seems to improve the recall. Moreover, *SPLICE_MISMATCHES* and *APPEND_MISMATCHES* show the same trends previously observed. Also at base level the trends seen in Malaria are preserved on Human: the difference in terms of recall between the best tweaking configuration and the default is small. Again, a high *APPEND_MISMATCHES* helps to increase the recall. The junction level precision still benefits from a high *MIN_MAP_LENGTH* and *APPEND_MISMATCHES* with a low *SPLICE_MISMATCHES*. On the other hand, the junction recall shows better results using a low *MIN_MAP_LENGTH* and a small *APPEND_MISMATCHES*.

Conclusions

There is a significant consistency between the result of the tweaking on Human and Malaria. *MIN_MAP_LENGTH*, *SPLICE_MISMATCHES* and *APPEND_MISMATCHES* are the most impactful parameters and show the same trends on both organisms.

NOVOALIGN

```
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
```

For the tweaking, the parameters in **Supplementary Table 23** were used. 16 different combinations of these values were used during the tweaking process. Few examples are presented in **Supplementary Table 24** and **Supplementary Table 25**.

Malaria
The default setting performs very well at all levels, compared with the other tools. Decreasing *B_SCORE* from the default value results in an improvement in the junction precision while all the other metrics result in worse performance. Given a fixed value of *B_SCORE*, decreasing *A_SCORE* seems to slightly improve the results at all metrics at any levels. The other tested options show negligible effects on the results.

Human
On human, the behavior observed for *B_SCORE* is confirmed: decreasing *B_SCORE* only helps the junction precision while has negative effects on the other metrics.

Conclusions
The default setting achieves very good results, comparable with the best tested configuration. The *B_SCORE* parameter plays a major role in our tests, while *A_SCORE* has a secondary role.

OLEGO

Alignment - tweaking:

```
olego --output-file output_1.sam --num-threads 16 --regression-model <regression model> --verbose --junction-file <junction file> --max-total-diff <TOTAL_DIFF> --word-size <WORD_SIZE> --max-word-diff <MAX_WORD_DIFF> --word-max-overlap <WORD_MAX_OVERLAP> --min-anchor <MIN_ANCHOR> --allow-rep-anchor <olego index> <read file 1>

olego --output-file output_2.sam --num-threads 16 --regression-model <regression model> --verbose --junction-file <junction file> --max-total-diff <TOTAL_DIFF> --word-size <WORD_SIZE> --max-word-diff <MAX_WORD_DIFF> --word-max-overlap <WORD_MAX_OVERLAP> --min-anchor <MIN_ANCHOR> --allow-rep-anchor <olego index> <read file 2>

perl mergePEsam.pl -v output_1.sam output_2.sam output.sam
```

For the tweaking, the parameters in **Supplementary Table 26** were tested. About 1000 different combinations of these values were used during the tweaking process. Few examples are presented in **Supplementary Table 27** and **Supplementary Table 28**.

Malaria
At read and base level, the most important parameter is *MAX_WORD_DIFF*. Using a value of *MAX_WORD_DIFF* greater than default significantly improves the recall: a value of 2 allows one to achieve a recall of ~25% while the default value (0) results in a recall of ~8%. Other important parameters are *TOTAL_DIFF*, *WORD_MAX_OVERLAP* and *MIN_ANCHOR*. Increasing *TOTAL_DIFF* and *MIN_ANCHOR* gives some benefits in terms of recall. About *WORD_MAX_OVERLAP*, the default value gives the best results. The option `--allow-rep-anchor` does not affect the results in our tests. Regarding *WORD_SIZE*, the best configurations were achieved with a *WORD_SIZE* of 13 or 14 but the trend is unclear. At the junction level it is difficult to understand the behavior of some parameters. The junction level precision and recall are positively influenced by a *MIN_ANCHOR* greater or equal than default. A *MAX_WORD_DIFF* low (0 or 1 in our tests) seems to bring some benefits at the junction level precision, while a value of 2 shows a better recall. The effects of the other parameters are not clear.

Human
The default configuration on Human achieves a better results compared to the default on Malaria. However, the improvements obtained by the parameters tweaking reach the same level on both organisms. Again, increasing *MAX_WORD_DIFF*, *TOTAL_DIFF* and *MIN_ANCHOR* results in an improvement on all metrics.

Conclusions

Both Human and Malaria take advantage from the tuning of some parameters. *MAX_WORD_DIFF*, *TOTAL_DIFF* and *MIN_ANCHOR* show an important role in the quality of the alignments in both organisms.
At read and base level, the parameter tweaking on Olego seems to not bring a significant improvement. The best results achieved after the tweaking are the worse compared with the tuning of the other tools. Moreover, the high junction precision shown by Olego is related to the low number of mapped reads so is a bit misleading. In conclusion, Olego seems not able to manage very low quality data very well, or otherwise we were not able to find the right parameters/values during the tweaking process.

RUM

Alignment - tweaking:

```
rum_runner align --index-dir <index directory> --name <job name> --output <output path> --chunks 16 <read file 1> <read file 2> --verbose --preserve-names --blat-min-identity <BLAT_MIN_IDENTITY> --blat-rep-match <BLAT_REP_MATCH> --blat-step-size <BLAT_STEP_SIZE> --blat-tile-size <BLAT_TILE_SIZE>
```

For the tweaking, the parameters in **Supplementary Table 29** were tested. About 750 different combinations of these values were tested during the tweaking process. Few examples are presented in **Supplementary Table 30** and **Supplementary Table 31**.

Malaria

At read and base level, the most important parameter is *BLAT_MIN_IDENTITY*. Using a value lower than 90 shows an improvement in the recall. Conversely, increasing this parameter causes a significant recall drop. For example, at read (base) level a value of 90 results in an recall of 63% (~58.8%), while with a value of 95 the recall is ~35% (~33%). Moreover, our tests suggest that setting *BLAT_STEP_SIZE* and *BLAT_TILE_SIZE* lower than 12 is preferable. Increasing *BLAT_REP_MATCH* seems to slightly improve the recall. The junction recall follows all the previously described trends. With regards to the junction precision, the most influent parameter is a high (98) *BLAT_MIN_IDENTITY*. The junction precision still benefits from a low *BLAT_STEP_SIZE* and a *BLAT_TILE_SIZE*.

Human

There are no improvements trying to apply the best Malaria configurations on Human. Indeed all the metrics show worse results (except for the junction precision). Probably the configurations are more related to the kind of genome than the quality of the data.

Conclusions

The parameter *BLAT_MIN_IDENTITY* has a major role in the quality of the alignment. On Malaria, a value lower than default shows better results in terms of recall. The parameter values seem more related to the particular genome than the quality of the data. Indeed, applying the best Malaria configurations on Human there is no improvement on the metrics.

SOAPsplice

Alignment - tweaking:

```
soapsplice -d <index> -1 <read file 1> -2 <read file 2> -o <output file> -p 16 -f 2 -l 0 -I <FRAGMENT_LENGTH_MEAN> -m <MISMATCHES> -g <INDEL> -i <TAIL> -a <SHORT_LENGTH>
```

For the tweaking, the parameters in **Supplementary Table 32** were tested. About 460 different combinations of these values were tested during the tweaking process. Few examples are presented in **Supplementary Table 33** and **Supplementary Table 34**.

Malaria

At read and base level, the key parameter is *MISMATCHES*. Using a value greater than default gives improved results in terms of recall. Compared with the default setting, a value of 5 improves the read level recall of ~22% and the base level recall of ~19%. On the other hand, a lower value decreases the quality of the alignment. Similarly, increasing *TAIL* has a positive effect on the alignment. However, the best value for this parameter seems to be different at read and base level (66 and 33 respectively). Values around the default are probably the best choices for *SHORT_LENGTH* (between 6 and 10 in our test). With regards to *INDEL*, the default value shows the best results. At junction level it is more difficult to figure out the role of each parameter. The junction recall seems to be mainly influenced by *SHORT_LENGTH*. A value of *SHORT_LENGTH* around the default is advisable. Moreover, setting *TAIL* around the default value and increasing *MISMATCHES* results in improved recall. With regards to the junction precision, there are no identifiable trends.

Human

The previous trends are confirmed on Human. Increasing the value of *MISMATCHES* and setting a high *TAIL* value results in an increase in the read and base recall. At the junction level, the best Malaria configurations do not significantly improve the precision or recall.

Conclusions

MISMATCHES seems to be the most influential parameter. Increasing this parameter results in better results at every level. Additionally, the parameter TAIL plays an important role, especially at read and base level.

STAR

Alignment - tweaking:

```
STAR --runThreadN 16 --genomeDir <index path> --readFilesIn <read file 1> <read file
2> --outFileNamePrefix <output alignment prefix> --twopassMode Basic --outSAMunmapped
Within --limitOutSJcollapsed <NUM_COLLAPSED_JUNCTIONS> --limitSjdbInsertNsj
<NUM_INSERTED_JUNCTIONS> --outFilterMultimapNmax <NUM_MULTIMAPPER> --
outFilterMismatchNmax <NUM_FILTER_MISMATCHES> --outFilterMismatchNoverLmax
<RATIO_FILTER_MISMATCHES> --seedSearchStartLmax <SEED_LENGTH> --alignSJoverhangMin
<OVERHANG> --alignEndsType <END_ALIGNMENT_TYPE> --outFilterMatchNminOverLread
<NUM_FILTER_MATCHES> --outFilterScoreMinOverLread <NUM_FILTER_SCORE> --
winAnchorMultimapNmax <NUM_ANCHOR> --alignSJDBoverhangMin <OVERHANG_ANNOTATED> --
outFilterType <OUT_FILTER>
```

For the tweaking, the parameters in **Supplementary Table 35** were tested. About 2600 different combinations of these values were tested during the tweaking process. Few examples are presented in **Supplementary Table 36** and **Supplementary Table 37**.

Malaria

At read and base level, the most important parameter is *NUM_FILTER_MISMATCHES*. Increasing the default value (values of 20, 25 or 33 in our tests), results in an improvement of ~16% in the read and base level recall. In addition, *END_ALIGNMENT_TYPE* plays an important role in the alignment. The value "*Local*" gives the best results, followed by "*Extend5pOfRead1*", "*Extend3pOfRead1*" and last "*EndToEnd*". A value of *OVERHANG* greater or equal to the default and a value of *NUM_FILTER_SCORE* less than the default (0.3 in our tests) results in an increase in the recall rate of up to ~5%, as compared to the defaults. Moreover, often a lower *SEED_LENGTH* (12, 30 and 33 in our tests) performs better than the default values. (note: the author suggested to us in private communication not to tweak another similar parameter, *--seedSearchLmax*, because it does not work robustly). If one puts *NUM_ANCHOR* = 50 (default) together with *OUT_FILTER* = *BySJout* and *NUM_FILTER_MATCHES* < 0.66 (default), *t*hen this slightly improves the results. The parameters *RATIO_FILTER_MISMATCHES*, *NUM_COLLAPSED_JUNCTIONS*, *NUM_INSERTED_JUNCTIONS*,

OVERHANG_ANNOTATED and *NUM_MULTIMAPPER* seem to have negligible effects. The junction level recall follows the same trends as above. The only small differences are that a low value of *OVERHANG* (1 in our tests) and the smallest *SEED_LENGTH* tested (12) help to further improve the junction recall. On the other hand, the junction precision has a different behavior. Here a low *NUM_FILTER_MISMATCHES* (≤10) with an *END_ALIGNMENT_TYPE* "*EndToEnd*" or a high *NUM_FILTER_MISMATCHES* with a *END_ALIGNMENT_TYPE* different from "*EndToEnd*" gives better results. Moreover, a high NUM_ANCHOR and the default value of NUM_FILTER_SCORE gives a higher junction level precision.

Human

On Human, using the best Malaria configurations improves all metrics. The base and read level recall gives an improvement of ~14% while the junction recall is increased of ~24%.

Conclusions

The default options of STAR achieve one of the best results on the T3 complexity dataset. However, tweaking some parameters is possible to further improve the metrics. The important roles of *NUM_FILTER_MISMATCHES*, *END_ALIGNMENT_TYPE*, *OVERHANG*, *NUM_FILTER_SCORE*, *SEED_LENGTH* are confirmed on both Human and Malaria. Increasing the number of allowed mismatches and leaving the default value for *END_ALIGNMENT_TYPE* improve the results. At the same time, increasing *OVERHANG* and decreasing *SEED_LENGTH* and *NUM_FILTER_SCORE* increase the recall at all levels.

Subread-Subjunc

Alignment - tweaking:

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

For the tweaking, the parameters in **Supplementary Table 38** were tested. About 1060 different combinations of these values were used during the tweaking process. Few examples are presented in **Supplementary Table 39** and **Supplementary Table S40**.

Malaria

The most important parameters are *MISMATCHES*, *NUM_HIT_SUBREADS* and *NUM_HIT_PAIR_SUBREADS*. For example, changing the value of *MISMATCHES* alone already has a large effect on the read (base) level recall: the default value of 3 results in a recall of ~28% (~23%) while a value of 20 gives a recall of ~63% (~55%). Similarly, the setting of *NUM_HIT_SUBREADS* and *NUM_HIT_PAIR_SUBREADS* plays an important role in the quality of the alignment. From our tests, setting *NUM_HIT_SUBREADS* and *NUM_HIT_PAIR_SUBREADS* to one improves the recall considerably when a high (≥8) value for *MISMATCHES* is used. With regards to *NUM_EXTRACTED_SUBREADS*, a value different from the default seems to be helpful in many cases. Moreover, the option --complexIndels often improves the results. The parameters *INDEL* and *MIN_FRAGMENT_LENGTH* do not affect the results. The junction recall seems to have the same behavior of the read and base accuracy, except for *NUM_EXTRACTED_SUBREADS*. Indeed, for the junction level recall the best value of *NUM_EXTRACTED_SUBREADS* is 15. The junction precision has no clear trend. In many configurations, the best results are achieved with a small value for *MISMATCHES*, with *NUM_HIT_SUBREADS* = 3 and *NUM_HIT_PAIR_SUBREADS* = 3. Moreover, at the junction level we have noticed a particular behavior using *NUM_HIT_SUBREADS* and *NUM_EXTRACTED_SUBREAD*. Indeed, using *NUM_HIT_SUBREADS* = 5 and *NUM_EXTRACTED_SUBREAD* = 5 the tool is not able to map any reads across splice-junctions.

Human

The best Malaria configurations benefit Human comparably. The read and base level recall have an improvement of 25% and 23% respectively, while the junction level recall increases by ~5%.

Conclusions
The tweaking of the parameters allows to considerably increase the alignment quality, both on Malaria and Human. The most influential parameters are *MISMATCHES*, *NUM_HIT_SUBREADS* and *NUM_HIT_PAIR_SUBREADS*.

Tophat2

Alignment - tweaking:

```
tophat2 --output-dir <output path> --num-threads 16 --mate-inner-dist
<INNER_MATE_MEAN> --mate-std-dev <INNER_MATE_SD> --b2-very-sensitive --GTF <gtf file>
--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> <index> <reads file 1> <reads file 2>
```

For the tweaking, the parameters in **Supplementary Table 41** were tested. About 230 different combinations of these values were used during the tweaking process. Few examples are presented in **Supplementary Table 42** and **Supplementary Table 43**.

Malaria
The most important parameter is *NUM_MISMATCHES*. Increasing the value of this parameter has a huge effect on the quality of the alignment. At read and base level, the best configuration results in an improvement in the recall rate of ~86% and ~74% respectively. Similarly, the junction level recall improves by ~81% with the same configuration. With regards to the value of *NUM_MISMATCHES*, there is a progressive improvement between the default value (2) and 17, at which point there is a saturation effect. The junction precision has behaves differently: values of *NUM_MISMATCHES* between 6 and 10 give the best results. The other parameters have negligible effects on the quality of the alignment at all levels. Surprisingly, *NUM_REALIGN_EDIT_DIST* = 0 does not improve the results.

Human
Using a high *NUM_MISMATCHES* on human significantly improves the result, similar to what is observed in Malaria. A value of 18 increases the read and base level recall by ~70%, while the junction recall has an improvement of ~74%.

Conclusions
Tophat2 displays both the worst default results and the highest improvement by the tweaking of the parameters. This behavior suggests that performing the tweaking of the parameters (especially *NUM_MISMATCHES*) is very important and highly recommended for this tool. The junction level precision/recall achieved by tweaking the parameters is one of the best, compared with the other tools.

Supplementary Note 4: Multi-Mappers

Unique mappers have largely been the focus. With read lengths of 100 bases typically around 10% of reads align ambiguously. However, many post alignment applications make us of multi-mappers in their calculations, e.g RSEM, Cufflinks and eXpress. To identify the recall and precision in the case of multi-mapping fragments the record with the most correct bases aligned was chosen and any further calculations were based on this best alignment. Including multi-mappers does increase the recall marginally in all tools, but often at the cost of precision. The results are shown in **Supplementary Figure 4**. The alignment statistics and the accuracy metrics can be found in **Supplementary Data 5-6**, **Supplementary Data 9-10** and again in **Supplementary Data 13-15**.

Supplementary Note 5: Adapters Simulation

In our simulated data, the fragment length distribution has a minimum of 100bp. In practice the minimum fragment length can be lower than this in which case reads will contain a part of the adapters employed during the library preparation. In this case adapter sequence will be added to the ends of both forward and reverse reads; and should not occur at the start of the reads.

Short adapter sequence may be harmless; however the longer they get the more they must affect the performance of the alignment process. Some aligners have the ability to manage the presence of adapter sequences, identifying and trimming them or simply ignoring these parts of the reads. Other aligners do not have this feature, in which case it is advisable to preprocess the reads using an adapter trimming utility.

In order to simulate this scenario, some of the simulated data was modified to contain adapter sequences of varying lengths. Starting from the Human T1R1 dataset, five sets of 100K read-pairs each were selected. In the first chuck, the last (at 3' end) 10bp of each read was substituted with the first (5' end) 10bp of the adapter sequence. For the second chuck, the last 20bp of each read was substituted with the first 20bp of the adapter sequence. Similarly, for 30, 40 and 50 bp for the remaining chunks respectively. Human T1R1 was chosen since it is the easiest dataset to align, so the main challenge for the aligners would be handling the adapters. Moreover, we simulated errors in the adapter sequences as happens in the real case. We employed the same rate adopted in the T1 level complexity datasets.

The Illumina Universal Adapter:

- http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-adapter-sequences_1000000002694-01.pdf

was used as template for the adapter sequence:
`5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT`
The initial bases of the above sequence were added for the forward reads, and their reverse complement was used for the reverse reads.

In addition, a copies of these five datasets were created and a preprocessing step was performed on them using an external adapter removing tool. The widely employed software *CutAdapt* (http://cutadapt.readthedocs.io/) was used, with the following command:
`cutadapt -a` *`AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT`* `-A` *`AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT -o <output forward read file> -p <output reverse read file> <input forward read file> <input reverse read file>`*
The software was capable of always identifying and correctly removing the adapter sequence (in both the forward and reverse reads); as a result, there was no more adapter sequence in the preprocessed datasets.

The scripts to add the adapter sequences into the read files and to perform the adapter removing step are available at: https://bitbucket.org/baruz/aligner-benchmark

At the end, for each of the five datasets a raw version (containing the adapter sequences) and a preprocessed version (containing no adapter sequences) were created, for a total of 10 datasets. Due to the trimming, the datasets containing adapters of lengths 10, 20, 30, 40 and 50 bp had in their preprocessed versions read lengths of 90, 80, 70, 60 and 50 bp respectively. On the other hands, the raw datasets maintained a read length of 100bp.

We aligned each of these ten datasets using the default alignment command of each aligner tool, as described in detail above in the *READS ALIGNMENT* section.

Results are shown in **Supplementary Figure 12**. Evidently most algorithms are robust to short adapters. Once residual adapter sequence reaches 50 bases they cause considerable problems for all algorithms with or without trimming, while trimming is always necessary for HISAT, OLEGO and TopHat2. For short residual adapter sequence CLC and ContextMap2 are the methods which have the most trouble, with or without trimming. For medium sized adapters the story is similar, except that now SoapSplice requires trimming for reasonable performance. Analysis on the adapter sequences highlights the importance of soft clipping and local alignment. Indeed, the poor performances of HISAT, Olego and TopHat2 on the

untrimmed datasets may be due to the inability to perform soft-clipping. Overall, GSNAP displays the best robustness to adapters.

Supplementary Note 6: Insertions and Deletions

In addition to base, read and junction level analysis separate statistics were collected on insertions and deletions. Performance on insertions and deletions is similar -- algorithms that have trouble with insertions tend to also have trouble with deletions. NOVOALIGN and CLC are the most consistently accurate in Human and among the best performer on Malaria. Subread achieves good performance on T1 and T2 datasets, while in T3 library GSNAP is the best option after NOVOALIGN and CLC. Insertion and deletion graphs for all data sets are given in **Supplementary Figures 13-14**.

Supplementary Note 7: Alignment Using a 2-Pass Mode

Some tools perform an automatic second alignment step in order to rescue more reads and/or identify more novel junctions. Other tools allow the user to specify this option explicitly: STAR provides the alignment option `--twopassMode <None|Basic>` while Olego describes a procedure to perform a second alignment step. The STAR and Olego two pass options were tested on all human and malaria R1 libraries. The results of this comparison are included in **Supplementary Data 9** and **Supplementary Data 10**, for human and malaria respectively.

The read and base level recall of both tools show only minor or negligible differences between the two modes. However, the 2-pass mode allows for increasing the junction recall, especially when the annotation is omitted as input.

Supplementary Note 8: Hardware and Performance Metrics

Computational performance refers to how long it takes the alignment to run and how much memory it requires. Run time is critical as the difference between the fast and slow aligners can mean a difference of hundreds to thousands of dollars of compute time in a typical study. Run time and RAM trade off in that you can usually lower one by raising the other. At the beginning it was considered necessary to stay below 8 Gb of RAM, but RAM has undergone a dramatic decrease in price since the inception of high throughput sequencing. Thus the trend over the last couple years, first introduced in STAR in 2013, has been towards the use of 30-60 Gb of RAM, which has decreased run-time dramatically. Therefore, one should not necessarily demerit an algorithm for utilizing a lot of RAM, if the tradeoff is much less CPU and run time.

All tests were performed on a HPC cluster (http://www.med.upenn.edu/hpc/) with a LSF job scheduling system. The cluster consists in 144 IBM iDataPlex Nodes (16 physical cores per node; 192 or 256 GB of RAM per node) using a Red Hat Linux 6.4. A job was designed for each alignment run and 16 threads were reserved for each job.

Some performance metrics were collected using the LSF tools provided by the system. In the analysis "*Total CPU time*", the "*Total run time"* and the "*Max memory usage*" were used. The *Total CPU time* is the total time the CPU spent running each tool; the *Total run time* is the time from the dispatch to completion of a job; the *Max memory usage* is the maximum resident memory used by a job.

Figure 4 and **Supplementary Figure 15** show the average values of each metric and the error bars (s.d.). The averages were computed between the three replicates of each library, on the "*default*" alignments.

In the “CPU time” charts, the CPU time value was used, divided by the number of threads used (16). In this way the average execution time is provided for each thread, with the ideal assumption of a full degree of parallelism during the the execution of the program.

In the “run time” charts, the Novoalign run time was divided by the number of threads used (16). Since Novoalign has no multithreading in its free license version, we performed this simple normalization in order to achieve a comparable result. However, this is an underestimation since the real scalability could be lower than the ideal one.

Olego shows a high variability on both CPU and RAM metrics on some libraries. We performed additional tests repeating the alignment jobs several times. However, the variability was confirmed also by these additional runs.

The variability in the memory usage for GSNAP on Human T3 was due to a high usage on T3R1 (115GB) respect to T3R2 (37GB) and T3R3 (39GB). The high usage on T3R1 was confirmed after additional tests.

For the complete dataset of performance metrics see **Supplementary Data 11** and **Supplementary Data 12**.

Supplementary Note 9: Alignment Notes for Each Aligner

The complete list of tested aligners is:

1. CLC Genomics Workbench v8.5 (http://www.clcbio.com/products/clc-genomics-workbench/)
2. ContextMap2 v2.6.0[2]
3. CRAC v2.4.0[3]
4. GSNAP v2015-9-29[4]
5. HISAT v0.1.6beta[5]
6. HISAT2 v2.0.0beta[5]
7. MapSplice2 v2.2.0[6]
8. Novoalign v3.02.13 (http://www.novocraft.com/products/novoalign/)
9. Olego v1.1.6[7]
10. RUM v2.0.5_06[8]
11. SOAPSPLICE v1.10[9]
12. STAR v2.5.0a[10]
13. SUBREAD v1.5.0[11]
14. TOPHAT v2.1.0[12]

More information about each method is available in **Supplementary Data 1** (http://bioinf.itmat.upenn.edu/BEERS/bp1/supplementaldata.php)

Each tool was downloaded from its website and installed, as according to the relevant documentation. In the alignment process we did not provide any information related to the simulations (error rate, indel rate, substitution rate) which would not normally be available in practice. We provided only information available in practice on real data, like generic annotation. For human e RefSeq was used throughout as provided annotation. For malaria the standard annotation provided by plasmoDB[20] was used. However the data were generated by a random selection of gene models selected from ten different annotation efforts, so as not to introduce bias. These are provided in GTF, BED, etc. format as required by the algorithm. Also provided are read length and fragment length distribution, when requested by the method, as typically done in practice.

In order to perform a fair comparison, indexes were created fresh for each aligner even if an index was already available. Thus, we are sure than all the aligners used the exact same version of the genome and the same annotation.

For each tool, we designed the alignment command starting from the default parameters. When the tool provided specific parameter presets or precise suggestions to increase the quality of the alignment, we followed these suggestions. Then, we set the parameters related to the read (read length, fragment length,

inner mate distance …), or related to the kind of genome (suggested seed size, *k*-mer size …), if the default was not compatible with our data. We called this set of alignments "*default*": they were the best alignment possible following the manual and having the standard knowledge of your dataset.
In order to determine the role of annotation, we performed alignments both including and omitting this information. Annotation in the required format was created for each tool, as necessary. Moreover, we explored the parameter space by performing tweaking the parameters of each tool and performing many alignments, in order to try to raise the accuracies as much as possible. Following the suggestions in the tool documentation, a set of parameters to test in the tweaking process were defined. Suggestions which are not precise (qualitative and not quantitative), were not included in the "*default*" but were explored in the tweaking analysis. Where these indications were not available, we tried the most exhaustive and reasonable set of parameters. We called this set of alignments "*tweaked*" throughout the document.
Both the "*default*" and the "*tweaked*" alignments were performed using 16 threads on an HPC cluster. When provided by the tool, we used the performance parameters that guarantee the shortest execution time (without any loss of precision). These options may use more RAM than the default. However, in the real scenario the available amount of RAM is usually a small issue as compared to execution time.
The results of the "*default*" alignments on human and malaria are available in **Supplementary Data 5** and **Supplementary Data 6**, respectively.
The results of the "*tweaked*" alignments on human and malaria are presented respectively in **Supplementary Data 2** and **Supplementary Data 3**.
The best "*tweaked*" configurations achieved at the end of the tuning process are summarized in **Supplementary Data 7** and **Supplementary Data 8**, for human and malaria respectively.
The scripts used to align the dataset, both in the "*default*" and the "*tweaking*" mode, can be found here: https://bitbucket.org/baruz/aligner-benchmark
In the following sections, we briefly describe the alignment protocol for each tool.

CLC Genomic Workbench
Version: 8.5 9/8/2015
Note: This is commercial software, with a 15 day free license for academic use. Therefore, the number of the "tweaking" alignments tested is smaller than the other tools. We used the software with its GUI, since it is the standard and better documented version. Due to the use of a GUI, we were not able to provide a precise set of performance metrics in terms of CPU and RAM usages.
Annotation: We provided a GTF file to the program, which it converts to a native format. The results were the "Gene track" and a "mRNA track", both used during the alignment process.
Genome: We provided the genome to the program, which it converts to a native format.
Index: The program build the index for each alignment run; there is not a standalone indexing step.
Alignment - default:
For the option "References", we used the default "Genome annotated with genes and transcripts" for the annotation, providing the "Gene track" and a "mRNA track". For the "Mapping" option we set to the default "Map to gene regions only (fast)".
For the "Read alignment" option we used the defaults: mismatch cost = 2; insertion cost = 3;
deletion cost = 3; length fraction = 0.8; similarity fraction = 0.8; global alignment = not selected; autodetect paired distance = true; strand specific = both; maximum number of hits for a read = 10.
Alignment - tweaking:
For the tweaking, we tested this set of parameters and values: Mapping, mismatch cost, insertion cost, deletion cost, length fraction, similarity fraction, global alignment and maximum number of hits for a read.
Post-processing: The output was converted to a SAM format with the tools provided by the software.

Contextmap2
Version: 2.6.0 11/10/2015
Note: The software was developed in Java, we used Java SDK 1.7.0 to run the program. Following the suggestion in the manual, some performance parameters for the Java multithreading environment were set.
Annotation: Annotation was provided in GTF format.
Index: The tool uses the index built by one of these programs: Bowtie, Bowtie2, BWA. Following the suggestion contained in the publication, we used BWA.

bwa index -a bwtsw -p *<index name>* *<genome fasta files>*
Genome: A single fasta file must be provided for each sequence to align. Therefore, the genome file was split into several files, one for each chromosome.
Reads: The input reads should be provided in one file, with a specific convention for the read name. A script was developed to adapt the read names to those required by the tools.
Alignment - default:
java -Xms4000M -Xmx64000m -XX:+UseConcMarkSweepGC -XX:NewSize=300M -XX:MaxNewSize=300M -jar ContextMap_v2.6.0.jar mapper -reads *<reads file>* --pairedend -gtf *<gtf file>* --noncanonicaljunctions -aligner_name bwa -aligner_bin *<bwa path>* -indexer_bin *<bwa path>* -indices *<bwa index>* -genome *<genome directory>* -o *<output path>* -t 16
Alignment - annotation:
In order to test the role of the annotation, we performed a run without the parameter -gtf *<gtf file>*
Alignment - tweaking:
java -Xms16000M -Xmx128000m -XX:+UseConcMarkSweepGC -XX:NewSize=300M -XX:MaxNewSize=300M -jar ContextMap_v2.6.0.jar mapper -reads *<reads file>* --pairedend -gtf *<gtf file>* --noncanonicaljunctions -aligner_name bwa -aligner_bin *<bwa path>* -indexer_bin *<bwa path>* -indices *<bwa index>* -genome *<genome directory>* -o *<output path>* -t 16 -seed *<SEED>* -seedmismatches *<SEED_MISMATCHES>* -mismatches *<MISMATCHES>* -mmdiff *<MMDIFF>* -maxhits *<MAXHITS>* -minsize *<MINSIZE>*

CRAC
Version: 2.4.0 10/2/2015
Index:
crac-index -v index *<crac index>* *<genome fasta file>*
Alignment - default:
crac -i *<crac index>* -k 22 -r *<reads file 1>* *<reads file 2>* --sam *<output sam file>* --reads-length *<read length>* --summary *<summary file>* --nb-threads 16
Alignment - annotation:
The tool does not allow the user to provide annotation.
Alignment - tweaking:
crac -i *<crac index>* -k *<K>* -r *<reads file 1>* *<reads file 2>* --sam *<output sam file>* --reads-length *<read length>* --no-ambiguity --max-locs *<MAX_LOCS>* --min-percent-single-loc *<MIN_PERCENT_SINGLE_LOC>* --min-percent-multiple-loc *<MIN_PERCENT_MULTIPLE_LOC>* --summary *<summary file>* --nb-threads 16

GSNAP
Version: 2015-9-29
Note: GSNAP provides several performance options. We used the most advanced performance option in order to minimize the execution time (--batch 5 --expand-offsets 1)
Annotation: Starting from the GTF file, we created the annotation in the required format following the instructions on the manual
cat *<gtf file>* | gtf_splicesites > foo.splicesites
cat foo.splicesites | iit_store -o genome.splicesites
cp genome.splicesites.iit *<index output path>*/*<index name>*/*<index name>*.maps/
Index:
gmap_build -D *<index output path>* -d *<index name>* *<genome fasta file>*
Alignment - default:
gsnap -D *<index output path>* -d *<index name>* -A sam --merge-distant-samechr --ordered --novelsplicing 1 --use-splicing *<index name>.splicesites* --nthreads 16 --batch 5 --expand-offsets 1 *<read file 1>* *<read file 2>* > *<output sam file>*
Alignment - annotation:
In order to test the role of annotation, a run without the following parameter was performed -use-splicing *<index name>.splicesites*
Alignment - tweaking:

gsnap -D *<index output path>* -d *<index name>* -A sam --max-mismatches *<MAX_MISMATCHES>* --indel-penalty *<INDEL_PENALITY>* --gmap-min-match-length *<GMAP_MIN_MATCH_LENGTH>* --pairexpect *<PAIR_EXPECT>* --pairdev *<PAIR_DEV>* --merge-distant-samechr --ordered --novelsplicing 1 --use-splicing *<index name>.splicesites* --nthreads 16 --batch 5 --expand-offsets 1 *<read file 1> <read file 2>* > *<output sam file>*

HISAT

Version: 0.1.6beta 4/7/2015

Annotation: Starting from the GTF file, the annotation file was created in the required format, following the instructions in the manual

extract_splice_sites.py *<gtf file>* > *<genome name>.splicesites.txt*

mv *<genome name>.splicesites.txt <output index path>*

Index:

hisat-build *<genome fasta file> <index name>*

Alignment - default:

hisat --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>*

Alignment - annotation:

In order to test the role of annotation, a run was performed without the following parameter --known-splicesite-infile *<output index path>/<genome name>.splicesites.txt*

Alignment - tweaking:

hisat --threads 16 --end-to-end -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>* --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>*

HISAT2

Version: 2.0.0beta 9/7/2015

Note: HISAT2 is a successor to both HISAT and Tophat2

Annotation: Starting from the GTF file, the annotation file was created in the required format, following the instructions on the manual

extract_splice_sites.py *<gtf file>* > *<genome name>.splicesites.txt*

extract_exons.py *<gtf file>* > *<genome name>.exons.txt*

mv *<genome name>.splicesites.txt <output index path>/<genome name>.splicesites.txt*

mv *<genome name>.exons.txt <output index path>/<genome name>.exons.txt*

Index:

hisat2-build -p 16 *<genome fasta file> <index name>*

It is possible to provide the annotation information into the index using these two parameters: --ss *<output index path>/<genome name>.splicesites.txt*

--exon *<output index path>/<genome name>.exons.txt*

With our data, these options work on malaria dataset. On the human dataset, due to crashing of the software, several issues had to be reported when these options were used.

Alignment - default:

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>*

Alignment - annotation:

In order to test the role of the annotation, a run was performed without the following parameter `--known-splicesite-infile <output index path>/<genome name>.splicesites.txt`

Alignment - tweaking:

```
hisat2 --threads 16 --end-to-end -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>--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>
```

Mapsplice2

Version: 2.2.0 8/14/2015

Annotation: In MapSplice annotation is used only to annotate fusion junctions.

Index: The tool uses the index build with Bowtie (1.1.2).

```
bowtie-build <genome fasta files> <index name>
```

Genome: A single fasta file should be provided for each sequence to align. Therefore, the genome file was split into several files, one for each chromosome. Moreover, all files must have the ".fa" extension and the header line related to the file name (header ">chr1" for file chr1.fa)

Alignment - default:

```
python mapsplice.py --threads 16 --non-canonical-double-anchor --output <output path> -c <genome fasta files> -x <index name> -1 <read file 1> -2 <read file 2>
```

Alignment - annotation:

Annotation is used only to annotate fusion junctions, so should not affect the results. However, annotation was provided to ascertain if there is any difference (`--gene-gtf <gtf file>`).

Alignment - tweaking:

```
python mapsplice.py --threads 16 --min-map-len <MIN_MAP_LENGTH> --splice-mis <SPLICE_MISMATCHES> --max-append-mis <APPEND_MISMATCHES> --ins <INSERTION_LENGTH> --del <DELETION_LENGTH> --filtering <FILTER> --non-canonical-double-anchor --output <output path> -c <genome fasta files> -x <index name> -1 <read file 1> -2 <read file 2>
```

Novoalign

Version: 3.02.13 6/29/2015

Note: Novoalign is commercial software. Free licenses with limited functionality are available for academic purposes. One of the limitations of these free licenses is the absence of multithreading for the alignment process. In the context of RNA-Seq, the use of USEQ (8.9.5) (http://useq.sourceforge.net/) to create the reference is suggested in the guide.

Annotation: Annotation, in UCSC refFlat format, is used for the creation of the index.

First, the guide suggests creating transcripts and splicing junction sequences with:

- Useq MakeTranscriptome

This command was used:

```
java -Xmx64G -jar <Useq path>/USeq_8.9.5/Apps/MakeTranscriptome -f <genome fasta files> -u <refFlat annotation> -r <N>
```

where *<N>* is the value read length - 4. For example for 100 bp paired-end read *<N>* is 96.

The output consisted of two files:

refFlatRad96Num100kMin10Splices.fasta.gz

refFlatRad96Num100kMin10Transcripts.fasta.gz

Second, the guide suggests adding the full genome with genes masked with "n" characters, using:

- USeq MaskExonsInFastaFiles

This command was used:

java -jar *<Useq path>*/USeq_8.9.5/Apps/MaskExonsInFastaFiles -f *<genome fasta files>* -u *<refFlat annotation>* -s *<masked exons output path>*
Then all of the fastq files in the directory *<masked exons output path> were concatenated* into a single file *geneMaskedGenome.fasta*
Index:
novoindex -t 16 *<output index file> refFlatRad96Num100kMin10Splices.fasta refFlatRad96Num100kMin10Transcripts.fasta geneMaskedGenome.fasta*
Alignment - default:
novoalign -d *<output index file>* -f *<read file 1> <read file 2>* -F FA -o SAM -r All 10 -i PE *<FRAGMENT_LENGTH_MEAN>,<FRAGMENT_LENGTH_SD>* -v 0 70 70 "[>]([^:]*)" > *<output sam file> 2>alignment.log*
Alignment - annotation:
For aligning RNA-Seq data, Novoalign requires annotation.
Alignment - tweaking:
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*
Post-processing:
After the alignment, the guide suggests to convert the splice junction coordinates in *<output sam file>* back to genome coordinates using:

- USeq SamTranscriptomeParser (http://useq.sourceforge.net/cmdLnMenus.html#SamTranscriptomeParser).

However, SamTranscriptomeParser requires that the field "QUAL" in the SAM file is different from "*" even if this field is not used by the script. When the input reads are fasta files, like our data, some random qualities should be put in the "QUAL" field.
Fixed the missing qualities, the conversion of the coordinate was performed with
java -Xmx64G -jar *<Useq path>*/USeq_8.9.5/Apps/SamTranscriptomeParser -f *<output sam file>* -a 50000 -n 100 -u -s *<output sam file with fixed coordinates>*

Olego
Version: 1.1.6 9/14/2015
Annotation: The tool suggests to provide a junction database to improve the sensitivity. A perl script is provided to create this database starting from a BED file. So first converted the GTF file was converted into a BED file and then the junction database was created with this command
perl bed2junc.pl *<bed file> <output junction file>*
Moreover, the tool uses a regression model during the alignment step. The provided script was used to generate the regression model with this command
perl *<olego path>*/regression_model_gen/OLego_regression.pl -g *<genome fasta files>* -a *<bed file>* -o *<output regression model prefix>*
Index:
olegoindex -p *<olego output index> <genome fasta file>*
Alignment - default:
olego --output-file output_1.sam --num-threads 16 --regression-model *<regression model>* --verbose --junction-file *<junction file> <olego index> <read file 1>*
olego --output-file output_2.sam --num-threads 16 --regression-model *<regression model>* --verbose --junction-file *<junction file> <olego index> <read file 2>*
perl mergePEsam.pl -v *output_1.sam output_2.sam output.sam*
In order to rescue more reads, an optional procedure is perform another mapping step without de novo mapping, since the junction annotations was already used. We tried also this option, we called it "2-pass". In the "2-pass" mapping, the following commands were added to the previous ones:
convert the SAM file to BED file
perl sam2bed.pl -v --use-RNA-strand *output.sam output.bed*
find the junctions in the BED file

```
perl bed2junc.pl output.bed output.junc
# second pass with option --non-denovo
olego --output-file output_1.twopass.sam --num-threads 16 --regression-model
<regression model> --verbose --junction-file output.junc --non-denovo <olego index>
<read file 1>
olego --output-file output_2.twopass.sam --num-threads 16 --regression-model
<regression model> --verbose --junction-file output.junc --non-denovo <olego index>
<read file 2>
perl mergePEsam.pl -v output_1.twopass.sam output_2.twopass.sam output.twopass.sam
```

Alignment - annotation:

In order to test the role of the annotation, a run was performed without the following parameter:

```
--junction-file <junction file>
```

For the “twopass” version, the annotation was not provided in the first pass.

Alignment - tweaking:

```
olego --output-file output_1.sam --num-threads 16 --regression-model <regression
model> --verbose --junction-file <junction file> --max-total-diff <TOTAL_DIFF> --word-
size <WORD_SIZE> --max-word-diff <MAX_WORD_DIFF> --word-max-overlap <WORD_MAX_OVERLAP>
--min-anchor <MIN_ANCHOR> --allow-rep-anchor <olego index> <read file 1>

olego --output-file output_2.sam --num-threads 16 --regression-model <regression
model> --verbose --junction-file <junction file> --max-total-diff <TOTAL_DIFF> --word-
size <WORD_SIZE> --max-word-diff <MAX_WORD_DIFF> --word-max-overlap <WORD_MAX_OVERLAP>
--min-anchor <MIN_ANCHOR> --allow-rep-anchor <olego index> <read file 2>
perl mergePEsam.pl -v output_1.sam output_2.sam output.sam
```

RUM

Version: 2.0.5_06 1/27/2015

Annotation: During the index creation the tool requires annotation. The GTF file was converted into the required format with a custom script.

Index:

```
perl create_indexes_from_ucsc.pl <genome txt file> <annotation files>
```

Alignment - default:

```
rum_runner align --index-dir <index directory> --name <job name> --output <output
path> --chunks 16 <read file 1> <read file 2> --verbose --preserve-names
```

Alignment - annotation:

In order to test the role of annotation, a run was performed with the parameter `--genome-only`

Alignment - tweaking:

```
rum_runner align --index-dir <index directory> --name <job name> --output <output
path> --chunks 16 <read file 1> <read file 2> --verbose --preserve-names --blat-min-
identity <BLAT_MIN_IDENTITY> --blat-rep-match <BLAT_REP_MATCH> --blat-step-size
<BLAT_STEP_SIZE> --blat-tile-size <BLAT_TILE_SIZE>
```

SOAPsplice

Version: 4/24/2013 1.10

Index:

```
2bwt-builder <genome fasta file> <output index>
```

Alignment - default:

```
soapsplice -d <index> -1 <read file 1> -2 <read file 2> -o <output file> -p 16 -f 2 -l
0 -I <FRAGMENT_LENGTH_MEAN>
```

Alignment - annotation:

The tool does not allow the user to provide any annotation.

Alignment - tweaking:

```
soapsplice -d <index> -1 <read file 1> -2 <read file 2> -o <output file> -p 16 -f 2 -l 0 -I <FRAGMENT_LENGTH_MEAN> -m <MISMATCHES> -g <INDEL> -i <TAIL> -a <SHORT_LENGTH>
```

STAR

Version: 2.5.0a 11/7/2015

Index:

```
STAR --runThreadN 16 --runMode genomeGenerate --genomeDir <index path> --genomeFastaFiles <genome fasta file> --sjdbGTFfile <gtf file> --sjdbOverhang <read length -1>
```

Alignment - default:

```
STAR --runThreadN 16 --genomeDir <index path> --readFilesIn <read file 1> <read file 2> --outFileNamePrefix <output alignment prefix> --twopassMode Basic --outSAMunmapped Within
```

Moreover, some alignments were performed without the “two-pass mode” to observe if there is a difference:

```
STAR --runThreadN 16 --genomeDir <index path> --readFilesIn <read file 1> <read file 2> --outFileNamePrefix <output alignment prefix> --outSAMunmapped Within
```

Alignment - annotation:

In order to test the role of the annotation, another index was created without the annotation removing the following parameters from the command:

```
--sjdbGTFfile <gtf file>
--sjdbOverhang <read length -1>
```

The alignment commands were the same.

Alignment - tweaking:

```
STAR --runThreadN 16 --genomeDir <index path> --readFilesIn <read file 1> <read file 2> --outFileNamePrefix <output alignment prefix> --twopassMode Basic --outSAMunmapped Within --limitOutSJcollapsed <NUM_COLLAPSED_JUNCTIONS> --limitSjdbInsertNsj <NUM_INSERTED_JUNCTIONS> --outFilterMultimapNmax <NUM_MULTIMAPPER> --outFilterMismatchNmax <NUM_FILTER_MISMATCHES> --outFilterMismatchNoverLmax <RATIO_FILTER_MISMATCHES> --seedSearchStartLmax <SEED_LENGTH> --alignSJoverhangMin <OVERHANG> --alignEndsType <END_ALIGNMENT_TYPE> --outFilterMatchNminOverLread <NUM_FILTER_MATCHES> --outFilterScoreMinOverLread <NUM_FILTER_SCORE> --winAnchorMultimapNmax <NUM_ANCHOR> --alignSJDBoverhangMin <OVERHANG_ANNOTATED> --outFilterType <OUT_FILTER>
```

Subread-Subjunc

Version: 1.5.0 10/29/2015

Note: Subjunc is the tool in the Subread package specific for RNA-Seq alignment. Subread could be used on RNA-Seq but is limited to the purpose of expression analysis, for this reason Subjunc was used.

Index:

```
subread-buildindex <genome fasta file> -B -o <output index>
```

Alignment - default:

```
subjunc -i <index> -r <read file 1> -R <read file 2> -T 16 --allJunctions --SAMoutput -o <output alignment>
```

Alignment - annotation:

The tool does not allow the user to provide any kind of annotation.

Alignment - tweaking:

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

Tophat2

Version: 2.1.0 6/29/2015

Index:

bowtie2-build -f *<genome fasta file> <output index>*

Alignment - default:

tophat2 --output-dir *<output path>* --num-threads 16 --mate-inner-dist *<INNER_MATE_MEAN>* --mate-std-dev *<INNER_MATE_SD>* --b2-very-sensitive --GTF *<gtf file> <index> <reads file 1> <reads file 2>*

Moreover, alignments were performed with and without the options --b2-very-sensitive and --coverage-search

Alignment - annotation:

In order to test the role of annotation, a run was performed without the parameter --GTF *<gtf file>*

Alignment - tweaking:

tophat2 --output-dir *<output path>* --num-threads 16 --mate-inner-dist *<INNER_MATE_MEAN>* --mate-std-dev *<INNER_MATE_SD>* --b2-very-sensitive --GTF *<gtf file>* --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> <index> <reads file 1> <reads file 2>*

Supplementary Note 10: Tests on the Latest Tool Versions

Since finishing the comparison, some tools released updates and bug fixes. We performed some tests on the latest available versions of each tool to look for any significant changes in performance. The list of new versions tested was:

- CLC Genomic Workbench 9.0.1 6/7/2016
- Contextmap2 2.07.06 5/31/2016
- CRAC 2.5.0 01/29/2016
- GSNAP 2016-06-30 06/30/2016
- HISAT2 2.0.4 5/18/2016
- Mapsplice2 3.0.1 Beta 7/19/2016
- Novoalign 3.04.06 5/18/2016
- STAR 2.5.2a 5/10/2016
- Subread 1.5.0-p3 5/27/2016
- Tophat2 2.1.1 2/23/2016

HISAT, Olego, RUM and SOAPsplice have no updated versions. The new version of Mapsplice2 is a beta version, while the new Mapsplice3; is an internal release provided by the author.

These latest versions were tested in the "*default*" mode over the first million read-pairs for data sets T1R1, T2R1 and T3R1, both for malaria and human. The results are given in **Supplementary Data 4.** We observe a slight improvement in GSNAP at the base level, a slight increase in junctions recall in ContextMap2 and MapSplice2 and curiously an increase in base level recall for MapSplice2 in human while a decrease in base level recall for MapSplice2 in malaria.

Tables

Supplementary Table 1 - Most relevant effects of including/omitting the annotation at junction level for the default alignments

Library	Tool	Human		Malaria	
		Junction recall (omitting annotation)	Junction recall (including annotation)	Junction recall (omitting annotation)	Junction recall (including annotation)
T1R1	GSNAP	74.42%	76.86%	73.80%	82.00%
T1R1	RUM	57.91%	87.66%	76.64%	94.73%
T1R1	STAR (1-pass)	71.63%	89.32%	78.22%	94.06%
T1R1	STAR (2-pass)	89.87%	91.38%	91.06%	94.06%
T1R1	Tophat2	82.51%	90.93%	86.18%	95.29%
T2R1	GSNAP	67.39%	72.14%	67.39%	76.52%
T2R1	RUM	49.42%	75.01%	72.58%	73.47%
T2R1	STAR (1-pass)	62.69%	84.20%	71.30%	90.72%
T2R1	STAR (2-pass)	82.92%	86.45%	86.18%	90.69%
T2R1	Tophat2	58.00%	68.82%	61.14%	73.07%
T3R1	GSNAP	31.80%	43.24%	32.33%	44.54%
T3R1	RUM	15.88%	17.43%	25.43%	26.59%
T3R1	STAR (1-pass)	19.42%	37.44%	31.84%	51.48%
T3R1	STAR (2-pass)	34.99%	39.02%	46.97%	51.44%
T3R1	Tophat2	1.40%	1.95%	1.62%	2.13%

Supplementary Table 2 - CLC Genomic Workbench tweaking parameters and values

Parameters	**Tested values**	**Note**
Mapping	“Map to gene regions only (fast)”; “Also map to inter-genic regions”	Default = “Map to gene regions only (fast)”
mismatch cost	1; 2	Integer, Default = 2, Range [1,3]
insertion cost	1; 3	Integer, Default = 3, Range [1,3]
deletion cost	1; 3	Integer, Default = 3, Range [1,3]
length fraction	0.5; 0.8	Default = 0.8, Range [0,1]
similarity fraction	0.5; 0.8	Default = 0.8, Range [0,1]
global alignment	not selected	Default = not selected
autodetect paired distance	true	Default = true
strand specific	both	Default = both
maximum number of hits for a read	1; 10; 30	Integer, Default = 10, Range [1,30]

Supplementary Table 3 - CLC Genomic Workbench tweaking examples on Malaria T3R1 dataset

Parameters:	number of multihits - map to only gene or also intergenic region - mismatch cost - insertion cost - deletion cost - length fraction - similarity fraction			
Configuration:	10-onlyGene-2-3-3-0.8-0.8	30-onlyGene-2-3-3-0.8-0.8	30-onlyGene-1-1-1-0.8-0.8	10-onlyGene-1-1-1-0.8-0.5
Note:	Default	Best recall at base level	Best recall at read level	Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	20000000	20000000	20000000	20000000
% reads aligned correctly [RECALL]:	98.15%	98.16%	98.17%	98.09%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	99.19%	99.17%	99.19%	99.26%
% reads aligned incorrectly:	0.79%	0.81%	0.79%	0.73%
% reads aligned ambiguously:	0.21%	0.21%	0.32%	0.59%
% reads unaligned:	0.85%	0.82%	0.72%	0.59%
% reads aligned:	99.15%	99.18%	99.28%	99.41%
% of reads with true introns:	4.31%	4.31%	4.31%	4.31%
---------- BASE LEVEL ----------				
total number of bases of reads:	2000000000	2000000000	2000000000	2000000000
% bases aligned correctly [RECALL]:	91.81%	91.82%	77.66%	77.61%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	96.88%	96.86%	78.62%	78.68%
% bases aligned incorrectly:	2.95%	2.97%	21.11%	21.02%
% bases aligned ambiguously:	0.21%	0.21%	0.32%	0.59%
% bases unaligned:	5.03%	5.00%	0.91%	0.78%
% bases aligned:	94.97%	95.00%	99.09%	99.22%
% of bases in true insertions:	0.41%	0.41%	0.41%	0.41%
% of bases in true deletions:	0.37%	0.37%	0.37%	0.37%
insertions FD rate [1 - PRECISION]:	24.98%	24.99%	65.37%	65.39%
insertions FN rate [1 - RECALL]:	40.75%	40.75%	33.20%	33.24%
deletions FD rate [1 - PRECISION]:	30.01%	30.01%	76.40%	76.43%
deletions FN rate [1 - RECALL]:	41.63%	41.62%	32.37%	32.42%
skipping FD rate [1 - PRECISION]:	0.39%	0.39%	0.82%	0.79%
skipping FN rate [1 - RECALL]:	5.75%	5.76%	3.52%	3.49%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	0.52%	0.53%	1.99%	1.96%
junctions FN rate [1 - RECALL]:	5.87%	5.87%	4.59%	4.58%
Junction Sides none	3793	3802	7774	7475
Junction Sides left	661	659	5564	5564
Junction Sides right	261	262	5228	5230
Junction Sides both	903335	903250	915545	915686
Junction Sides none	0.41%	0.41%	0.83%	0.80%
Junction Sides left	0.07%	0.07%	0.59%	0.59%
Junction Sides right	0.02%	0.02%	0.55%	0.55%
Junction Sides both	99.48%	99.47%	98.01%	98.04%

Supplementary Table 4 - CLC Genomic Workbench tweaking examples on Human T3R1 dataset

Parameters:	number of multihits - map to only gene or also intergenic region - mismatch cost - insertion cost - deletion cost - length fraction - similarity fraction			
Configuration:	10-onlyGene-2-3-3-0.8-0.8	30-alsoIntergenic-2-3-3-0.8-0.8	30-alsoIntergenic-1-1-1-0.5-0.8	30-onlyGene-1-1-1-0.5-0.8
Note:	Default	Best recall at base level	Best recall at read level	Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	20000000	20000000	20000000	20000000
% reads aligned correctly [RECALL]:	84.63%	93.88%	94.37%	87.32%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	97.21%	98.15%	98.36%	97.01%
% reads aligned incorrectly:	2.42%	1.76%	1.56%	2.68%
% reads aligned ambiguously:	2.80%	2.83%	3.36%	4.38%
% reads unaligned:	10.15%	1.53%	0.71%	5.62%
% reads aligned:	89.85%	98.47%	99.29%	94.38%
% of reads with true introns:	13.46%	13.46%	13.46%	13.46%
---------- BASE LEVEL ----------				
total number of bases of reads:	2000000000	2000000000	2000000000	2000000000
% bases aligned correctly [RECALL]:	80.13%	88.91%	76.94%	71.20%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	95.91%	96.90%	80.32%	79.22%
% bases aligned incorrectly:	3.40%	2.83%	18.84%	18.66%
% bases aligned ambiguously:	2.80%	2.83%	3.36%	4.38%
% bases unaligned:	13.67%	5.43%	0.86%	5.76%
% bases aligned:	86.33%	94.57%	99.14%	94.24%
% of bases in true insertions:	0.18%	0.18%	0.18%	0.18%
% of bases in true deletions:	0.16%	0.16%	0.16%	0.16%
insertions FD rate [1 - PRECISION]:	28.97%	29.07%	78.84%	78.43%
insertions FN rate [1 - RECALL]:	46.67%	43.43%	35.23%	37.64%
deletions FD rate [1 - PRECISION]:	31.26%	31.28%	87.02%	86.70%
deletions FN rate [1 - RECALL]:	44.73%	41.35%	31.51%	34.10%
skipping FD rate [1 - PRECISION]:	3.89%	3.67%	5.11%	5.36%
skipping FN rate [1 - RECALL]:	27.82%	26.34%	22.24%	22.23%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	4.67%	4.55%	7.32%	7.44%
junctions FN rate [1 - RECALL]:	20.94%	20.09%	18.28%	18.17%
Junction Sides none	75223	72728	83157	86579
Junction Sides left	19543	19848	51284	51247
Junction Sides right	18542	18754	54539	54581
Junction Sides both	2315657	2340431	2393640	2396900
Junction Sides none	3.09%	2.96%	3.21%	3.34%
Junction Sides left	0.80%	0.80%	1.98%	1.97%
Junction Sides right	0.76%	0.76%	2.11%	2.10%
Junction Sides both	95.33%	95.45%	92.68%	92.56%

Supplementary Table 5 - Contextmap2 tweaking parameters and values

Parameters	Tested values	Note
`-seed <SEED>`	10; 20; 30	Default = 20 (or 30 using Bowtie1)
`-seedmismatches <SEED_MISMATCHES>`	0; 1; 2	Default = 0 (or 1 using Bowtie1)
`-mismatches <MISMATCHES>`	3; 4; 5; 6; 7; 8; 9; 10; 12; 15; 17; 20; 25; 30; 35; 40; 45; 50; 55; 60	Default = 4
`-mmdiff <MMDIFF>`	0; 1; 2	Default = 0
`-maxhits <MAXHITS>`	10; 20; 50	Default = 10 (or 3 using Bowtie2)
`-minsize <MINSIZE>`	5; 10; 15	Default = 10

Supplementary Table 6 - Contextmap2 tweaking examples on Malaria T3R1 dataset

Command:	java -Xms16000M -Xmx128000m -XX:+UseConcMarkSweepGC -XX:NewSize=300M -XX:MaxNewSize=300M -jar -jar ContextMap_v2.6.0.jar mapper -reads <reads file> --pairedend -gtf <gtf file> --noncanonicaljunctions -aligner_name bwa -aligner_bin <bwa path> -indexer_bin <bwa path> -indices <bwa index> -genome <genome directory> -o <output path> -t 16 -seed <SEED> -seedmismatches <SEED_MISMATCHES> -mismatches <MISMATCHES> -mmdiff <MMDIFF> -maxhits <MAXHITS> -minsize <MINSIZE>			
Parameters:	SEED - SEED_MISMATCHES - MISMATCHES - MMDIFF - MAXHITS - MINSIZE			
Configuration:	20-0-4-0-10-10	20-0-35-0-10-10	20-0-35-1-10-10	20-0-60-0-10-10
Note:	Default	Best recall at base level	Best recall at read level	Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	40.13%	87.84%	87.89%	86.96%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	96.68%	98.66%	98.58%	98.04%
% reads aligned incorrectly:	1.37%	1.18%	1.26%	1.73%
% reads aligned ambiguously:	0.00%	0.00%	0.00%	0.00%
% reads unaligned:	58.50%	10.98%	10.85%	11.31%
% reads aligned:	41.50%	89.02%	89.15%	88.69%
% of reads with true introns:	4.31%	4.31%	4.31%	4.31%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	30.27%	84.39%	84.34%	83.33%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	94.74%	95.78%	95.49%	94.20%
% bases aligned incorrectly:	1.67%	3.71%	3.97%	5.12%
% bases aligned ambiguously:	0.00%	0.00%	0.00%	0.00%
% bases unaligned:	68.06%	11.90%	11.69%	11.55%
% bases aligned:	31.94%	88.10%	88.31%	88.45%
% of bases in true insertions:	0.41%	0.41%	0.41%	0.41%
% of bases in true deletions:	0.37%	0.37%	0.37%	0.37%
insertions FD rate [1 - PRECISION]:	4.08%	18.71%	23.31%	21.31%
insertions FN rate [1 - RECALL]:	91.30%	44.77%	46.59%	47.13%
deletions FD rate [1 - PRECISION]:	2.84%	14.90%	18.11%	16.71%
deletions FN rate [1 - RECALL]:	91.45%	38.91%	40.76%	40.66%
skipping FD rate [1 - PRECISION]:	98.59%	99.70%	99.70%	99.81%
skipping FN rate [1 - RECALL]:	90.92%	20.80%	21.00%	21.19%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	16.29%	44.02%	45.15%	49.32%
junctions FN rate [1 - RECALL]:	91.14%	30.06%	30.09%	28.80%
Junction Sides none	1623	50850	53137	64877
Junction Sides left	21	1071	1114	857
Junction Sides right	7	764	881	648
Junction Sides both	8490	67009	66986	68221
Junction Sides none	16.00%	42.48%	43.51%	48.19%
Junction Sides left	0.20%	0.89%	0.91%	0.63%
Junction Sides right	0.06%	0.63%	0.72%	0.48%
Junction Sides both	83.71%	55.98%	54.85%	50.68%

Supplementary Table 7 - Contextmap2 tweaking examples on Human T3R1 dataset

Command:	java -Xms16000M -Xmx128000m -XX:+UseConcMarkSweepGC -XX:NewSize=300M -XX:MaxNewSize=300M -jar -jar ContextMap_v2.6.0.jar mapper -reads <reads file> --pairedend -gtf <gtf file> --noncanonicaljunctions -aligner_name bwa -aligner_bin <bwa path> -indexer_bin <bwa path> -indices <bwa index> -genome <genome directory> -o <output path> -t 16 -seed <SEED> -seedmismatches <SEED_MISMATCHES> -mismatches <MISMATCHES> -mmdiff <MMDIFF> -maxhits <MAXHITS> -minsize <MINSIZE>			
Parameters:	SEED - SEED_MISMATCHES - MISMATCHES - MMDIFF - MAXHITS - MINSIZE			
Configuration:	20-0-4-0-10-10	20-0-35-0-10-10	20-0-35-1-10-10	30-0-3-0-10-10
Note:	Default	Best recall at base, read and junction level		
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	61.63%	89.88%	89.45%	52.05%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	97.21%	97.67%	97.47%	97.27%
% reads aligned incorrectly:	1.76%	2.13%	2.31%	1.45%
% reads aligned ambiguously:	0.00%	0.00%	0.00%	0.00%
% reads unaligned:	36.61%	7.99%	8.24%	46.50%
% reads aligned:	63.39%	92.01%	91.76%	53.50%
% of reads with true introns:	13.47%	13.46%	13.47%	13.47%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	53.76%	87.46%	87.02%	46.84%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	96.75%	95.89%	95.64%	96.88%
% bases aligned incorrectly:	1.80%	3.73%	3.95%	1.50%
% bases aligned ambiguously:	0.00%	0.00%	0.00%	0.00%
% bases unaligned:	44.44%	8.81%	9.03%	51.66%
% bases aligned:	55.56%	91.19%	90.97%	48.34%
% of bases in true insertions:	0.18%	0.18%	0.18%	0.18%
% of bases in true deletions:	0.17%	0.17%	0.17%	0.17%
insertions FD rate [1 - PRECISION]:	3.61%	16.01%	19.21%	3.89%
insertions FN rate [1 - RECALL]:	95.58%	66.82%	69.02%	97.99%
deletions FD rate [1 - PRECISION]:	3.30%	15.82%	17.80%	2.99%
deletions FN rate [1 - RECALL]:	95.33%	60.66%	63.44%	97.62%
skipping FD rate [1 - PRECISION]:	11.48%	27.36%	29.11%	13.45%
skipping FN rate [1 - RECALL]:	91.75%	27.86%	28.37%	95.82%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	3.46%	15.19%	15.43%	3.18%
junctions FN rate [1 - RECALL]:	91.72%	33.84%	33.98%	95.69%
Junction Sides none	733	28087	28065	349
Junction Sides left	61	3474	3749	29
Junction Sides right	74	3177	3501	37
Junction Sides both	24288	193990	193585	12637
Junction Sides none	2.91%	12.27%	12.26%	2.67%
Junction Sides left	0.24%	1.51%	1.63%	0.22%
Junction Sides right	0.29%	1.38%	1.52%	0.28%
Junction Sides both	96.54%	84.81%	84.57%	96.82%

Supplementary Table 8 - CRAC tweaking parameters and values

Parameters	Tested values	Note
`-k <K>`	16; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27	Integer, Default = 22 (for Human)
`--no-ambiguity`	with and without this option	Default = without this option
`--max-locs <MAX_LOCS>`	300; 400; 1000	Default = 300
`--min-percent-single-loc <MIN_PERCENT_SINGLE_LOC>`	0.1; 0.15; 0.2	Default = 0.15
`--min-percent-multiple-loc <MIN_PERCENT_MULTIPLE_LOC>`	0.4; 0.5; 0.6	Default = 0.5

Supplementary Table 9 - CRAC tweaking examples on Malaria T3R1 dataset

Command:	crac -i <crac index> -k <K> -r <reads file 1> <reads file 2> --sam <output sam file> --reads-length <read length> --no-ambiguity --max-locs <MAX_LOCS> --min-percent-single-loc <MIN_PERCENT_SINGLE_LOC> --min-percent-multiple-loc <MIN_PERCENT_MULTIPLE_LOC> --summary <summary file> --nb-threads 16			
Parameters:	K - NO_AMBIGUITY - MAX_LOCS - MIN_PERCENT_SINGLE_LOC - MIN_PERCENT_MULTIPLE_LOC			
Configuration:	22-off-300-0.15-0.5	18-off-1000-0.15-0.5	16-off-1000-0.15-0.5	25-off-1000-0.15-0.5
Note:	Default	Best recall at base and read level	Best recall at junction level	
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	79.69%	87.08%	79.57%	69.68%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	95.71%	90.58%	80.52%	96.44%
% reads aligned incorrectly:	3.57%	9.04%	19.25%	2.56%
% reads aligned ambiguously:	0.00%	0.03%	0.24%	0.00%
% reads unaligned:	16.74%	3.85%	0.94%	27.76%
% reads aligned:	83.26%	96.15%	99.06%	72.24%
% of reads with true introns:	4.31%	4.31%	4.31%	4.31%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	38.73%	43.56%	39.10%	33.40%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	91.83%	83.53%	68.20%	93.02%
% bases aligned incorrectly:	3.44%	8.58%	18.22%	2.50%
% bases aligned ambiguously:	0.00%	0.03%	0.24%	0.00%
% bases unaligned:	57.83%	47.83%	42.44%	64.10%
% bases aligned:	42.17%	52.17%	57.56%	35.90%
% of bases in true insertions:	0.41%	0.41%	0.41%	0.41%
% of bases in true deletions:	0.37%	0.37%	0.37%	0.37%
insertions FD rate [1 - PRECISION]:	79.07%	76.71%	74.85%	77.98%
insertions FN rate [1 - RECALL]:	97.50%	95.54%	95.62%	98.53%
deletions FD rate [1 - PRECISION]:	30.97%	32.13%	31.25%	30.59%
deletions FN rate [1 - RECALL]:	97.88%	96.65%	96.90%	98.71%
skipping FD rate [1 - PRECISION]:	99.10%	99.84%	99.96%	97.33%
skipping FN rate [1 - RECALL]:	96.43%	92.72%	90.75%	97.98%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	66.27%	79.37%	88.49%	56.39%
junctions FN rate [1 - RECALL]:	98.16%	96.37%	95.66%	98.92%
Junction Sides none	3472	13337	31870	1343
Junction Sides left	2	14	61	0
Junction Sides right	6	34	62	3
Junction Sides both	1772	3480	4165	1041
Junction Sides none	66.10%	79.08%	88.14%	56.26%
Junction Sides left	0.03%	0.08%	0.16%	0.00%
Junction Sides right	0.11%	0.20%	0.17%	0.12%
Junction Sides both	33.73%	20.63%	11.51%	43.61%

Supplementary Table 10 - CRAC tweaking examples on Human T3R1 dataset

Command:	crac -i <crac index> -k <K> -r <reads file 1> <reads file 2> --sam <output sam file> --reads-length <read length> --no-ambiguity --max-locs <MAX_LOCS> --min-percent-single-loc <MIN_PERCENT_SINGLE_LOC> --min-percent-multiple-loc <MIN_PERCENT_MULTIPLE_LOC> --summary <summary file> --nb-threads 16			
Parameters:	K - NO_AMBIGUITY - MAX_LOCS - MIN_PERCENT_SINGLE_LOC - MIN_PERCENT_MULTIPLE_LOC			
Configuration:	22-off-300-0.15-0.5	20-off-1000-0.15-0.5	19-off-1000-0.15-0.5	18-off-1000-0.15-0.5
Note:	Default	Best recall at base level	Best recall at read and junction level	
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	80.79%	83.70%	83.84%	81.24%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	89.35%	88.59%	86.54%	82.30%
% reads aligned incorrectly:	9.62%	10.78%	13.03%	17.46%
% reads aligned ambiguously:	0.00%	0.00%	0.00%	0.03%
% reads unaligned:	9.59%	5.52%	3.13%	1.27%
% reads aligned:	90.41%	94.48%	96.87%	98.73%
% of reads with true introns:	13.47%	13.46%	13.46%	13.47%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	43.41%	45.02%	44.91%	43.37%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	83.54%	82.30%	79.13%	72.85%
% bases aligned incorrectly:	8.55%	9.68%	11.84%	16.16%
% bases aligned ambiguously:	0.00%	0.00%	0.00%	0.03%
% bases unaligned:	48.04%	45.30%	43.25%	40.44%
% bases aligned:	51.96%	54.70%	56.75%	59.56%
% of bases in true insertions:	0.18%	0.18%	0.18%	0.18%
% of bases in true deletions:	0.17%	0.17%	0.17%	0.17%
insertions FD rate [1 - PRECISION]:	46.99%	44.85%	43.11%	43.26%
insertions FN rate [1 - RECALL]:	98.59%	97.98%	97.78%	97.96%
deletions FD rate [1 - PRECISION]:	26.51%	26.79%	26.42%	26.48%
deletions FN rate [1 - RECALL]:	98.76%	98.29%	98.16%	98.36%
skipping FD rate [1 - PRECISION]:	60.37%	78.62%	91.84%	97.46%
skipping FN rate [1 - RECALL]:	96.70%	95.37%	94.90%	95.43%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	65.07%	65.49%	65.96%	66.05%
junctions FN rate [1 - RECALL]:	98.72%	98.17%	98.00%	98.20%
Junction Sides none	6998	10154	11314	10229
Junction Sides left	13	24	41	37
Junction Sides right	21	42	54	51
Junction Sides both	3775	5386	5889	5304
Junction Sides none	64.75%	65.06%	65.40%	65.48%
Junction Sides left	0.12%	0.15%	0.23%	0.23%
Junction Sides right	0.19%	0.26%	0.31%	0.32%
Junction Sides both	34.93%	34.51%	34.04%	33.95%

Supplementary Table 11 - GSNAP tweaking parameters and values

Parameters	Tested values	Note
`--max-mismatches <MAX_MISMATCHES>`	2; 3; 4; 5; 6; 7; 8; 9; 10; 15; 20; 25; 30; 35	Default = (readlength+index_interval-1)/kmer - 2 ; should be around 4 for both human and malaria using this formula
`--indel-penalty <INDEL_PENALITY>`	1; 2; 3; 4; 5	Default = 2
`--gmap-min-match-length <GMAP_MIN_MATCH_LENGTH>`	7; 10; 12; 15; 17; 20; 25	Default = 20
`--pairexpect <PAIR_EXPECT>`	default; <*fragment length mean*>	Default = 200
`--pairdev <PAIR_DEV>`	default; <*fragment length SD*>	Default = 100

Supplementary Table 12 - GSNAP tweaking examples on Malaria T3R1 dataset

Command:	gsnap -D <index output path> -d <index name> -A sam --max-mismatches <MAX_MISMATCHES> --indel-penalty <INDEL_PENALITY> --gmap-min-match-length <GMAP_MIN_MATCH_LENGTH> --pairexpect <PAIR_EXPECT> --pairdev <PAIR_DEV> --merge-distant-samechr --ordered --novelsplicing 1 --use-splicing <index name>.splicesites --nthreads 16 --batch 5 --expand-offsets 1 <read file 1> <read file 2> > <output sam file>			
Parameters:	MAX_MISMATCHES - INDEL_PENALITY - GMAP_MIN_MATCH_LENGTH - PAIR_EXPECT - PAIR_DEV			
Configuration:	4-2-20-default-default	15-1-10-222-42	20-1-7-222-42	2-5-15-222-42
Note:	Default	Best recall at base level	Best recall at read level	Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	20000000	20000000	20000000	20000000
% reads aligned correctly [RECALL]:	88.31%	97.92%	97.96%	85.69%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	99.33%	99.38%	99.33%	99.46%
% reads aligned incorrectly:	0.59%	0.61%	0.65%	0.46%
% reads aligned ambiguously:	1.62%	1.10%	1.09%	1.76%
% reads unaligned:	9.48%	0.37%	0.30%	12.09%
% reads aligned:	90.52%	99.63%	99.70%	87.91%
% of reads with true introns:	4.31%	4.31%	4.31%	4.31%
---------- BASE LEVEL ----------				
total number of bases of reads:	2000000000	2000000000	2000000000	2000000000
% bases aligned correctly [RECALL]:	78.74%	88.21%	88.23%	77.32%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	96.81%	96.78%	96.67%	97.17%
% bases aligned incorrectly:	2.59%	2.93%	3.03%	2.25%
% bases aligned ambiguously:	1.62%	1.10%	1.09%	1.76%
% bases unaligned:	17.05%	7.76%	7.65%	18.67%
% bases aligned:	82.95%	92.24%	92.35%	81.33%
% of bases in true insertions:	0.41%	0.41%	0.41%	0.41%
% of bases in true deletions:	0.37%	0.37%	0.37%	0.37%
insertions FD rate [1 - PRECISION]:	24.42%	27.36%	28.31%	21.84%
insertions FN rate [1 - RECALL]:	61.07%	53.84%	54.66%	57.06%
deletions FD rate [1 - PRECISION]:	45.84%	41.74%	42.07%	42.61%
deletions FN rate [1 - RECALL]:	58.90%	57.13%	58.43%	56.85%
skipping FD rate [1 - PRECISION]:	98.56%	97.91%	97.71%	97.10%
skipping FN rate [1 - RECALL]:	54.08%	51.03%	52.39%	44.65%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	21.83%	11.32%	10.21%	14.61%
junctions FN rate [1 - RECALL]:	55.46%	52.01%	53.29%	46.45%
Junction Sides none	116877	56280	48547	84979
Junction Sides left	1135	1159	1103	1361
Junction Sides right	1335	1311	1281	1560
Junction Sides both	427403	460572	448300	513935
Junction Sides none	21.37%	10.83%	9.72%	14.11%
Junction Sides left	0.20%	0.22%	0.22%	0.22%
Junction Sides right	0.24%	0.25%	0.25%	0.25%
Junction Sides both	78.17%	88.68%	89.79%	85.39%

Supplementary Table 13 - GSNAP tweaking examples on Human T3R1 dataset

Command:	gsnap -D <index output path> -d <index name> -A sam --max-mismatches <MAX_MISMATCHES> --indel-penalty <INDEL_PENALITY> --gmap-min-match-length <GMAP_MIN_MATCH_LENGTH> --pairexpect <PAIR_EXPECT> --pairdev <PAIR_DEV> --merge-distant-samechr --ordered --novelsplicing 1 --use-splicing <index name>.splicesites --nthreads 16 --batch 5 --expand-offsets 1 <read file 1> <read file 2> > <output sam file>			
Parameters:	MAX_MISMATCHES - INDEL_PENALITY - GMAP_MIN_MATCH_LENGTH - PAIR_EXPECT - PAIR_DEV			
Configuration:	4-2-20-default-default	15-1-10-221-41	20-1-7-default-default	2-5-15-221-41
Note:	Default	Best recall at base and read level		Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	20000000	20000000	20000000	20000000
% reads aligned correctly [RECALL]:	91.98%	96.24%	96.14%	84.12%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	98.58%	98.23%	98.07%	98.19%
% reads aligned incorrectly:	1.31%	1.72%	1.88%	1.54%
% reads aligned ambiguously:	2.00%	1.71%	1.71%	1.86%
% reads unaligned:	4.71%	0.33%	0.27%	12.48%
% reads aligned:	95.29%	99.67%	99.73%	87.52%
% of reads with true introns:	13.46%	13.46%	13.46%	13.46%
---------- BASE LEVEL ----------				
total number of bases of reads:	2000000000	2000000000	2000000000	2000000000
% bases aligned correctly [RECALL]:	84.32%	88.86%	88.75%	77.51%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	97.69%	97.18%	96.98%	97.30%
% bases aligned incorrectly:	1.98%	2.56%	2.75%	2.14%
% bases aligned ambiguously:	2.00%	1.71%	1.71%	1.86%
% bases unaligned:	11.70%	6.87%	6.79%	18.49%
% bases aligned:	88.30%	93.13%	93.21%	81.51%
% of bases in true insertions:	0.18%	0.18%	0.18%	0.18%
% of bases in true deletions:	0.16%	0.16%	0.16%	0.16%
insertions FD rate [1 - PRECISION]:	23.07%	24.48%	25.40%	21.81%
insertions FN rate [1 - RECALL]:	61.21%	56.17%	57.16%	57.66%
deletions FD rate [1 - PRECISION]:	29.29%	30.32%	31.14%	28.52%
deletions FN rate [1 - RECALL]:	57.56%	55.09%	56.37%	55.44%
skipping FD rate [1 - PRECISION]:	28.42%	26.98%	28.29%	27.83%
skipping FN rate [1 - RECALL]:	61.72%	58.18%	59.42%	54.77%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	5.70%	4.39%	4.47%	5.86%
junctions FN rate [1 - RECALL]:	56.76%	56.45%	58.03%	47.45%
Junction Sides none	64400	50808	49555	79362
Junction Sides left	6234	4072	4102	8344
Junction Sides right	5804	3679	3790	8013
Junction Sides both	1266475	1275708	1229323	1539302
Junction Sides none	4.79%	3.80%	3.85%	4.85%
Junction Sides left	0.46%	0.30%	0.31%	0.51%
Junction Sides right	0.43%	0.27%	0.29%	0.49%
Junction Sides both	94.30%	95.61%	95.53%	94.14%

Supplementary Table 14 - HISAT tweaking parameters and values

Parameters	Tested values	Note
`--end-to-end`	with and without this option	Bowtie2-like parameter, Default = without this option
`-N <NUM_MISMATCH>`	0; 1	Bowtie2-like parameter, Integer, Default = 0, Range = [0,1]
`-L <SEED_LENGTH>`	15; 18; 20; 22	Bowtie2-like parameter, Integer, Default = 22
`-i S,1,<SEED_INTERVAL>`	0.25; 0.5; 0.75; 1.15	Bowtie2-like parameter, Default = 1.15
`-D <SEED_EXTENSION>`	15; 20; 25	Bowtie2-like parameter, Default = 15
`-R <RE_SEED>`	2; 3; 5	Bowtie2-like parameter, Default = 2
`--pen-noncansplice <PENALITY_NONCANONICAL>`	0; 3; 12; 20; 30	Default = 3
`--mp <MAX_MISMATCH_PENALITY>, …`	1; 3; 6	Default = 6
`--mp …, <MIN_MISMATCH_PENALITY>`	0; 1; 2	Default = 2

Supplementary Table 15 - HISAT tweaking examples on Malaria T3R1 dataset

Command:	hisat --threads 16 --end-to-end -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> --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>			
Parameters:	ENDTOEND_MODE - NUM_MISMATCH - SEED_LENGTH - SEED_INTERVAL - SEED_EXTENSION - RE_SEED - PENALITY_NONCANONICAL - MAX_MISMATCH_PENALITY - MIN_MISMATCH_PENALITY			
Configuration:	default-0-22-1.15-15-2-3-default-default	default-0-20-1.15-15-2-20-1-0	default-1-20-0.5-25-5-3-3-0	default-1-20-0.5-25-5-3-1-0
Note:	Default	Best recall at base, read and junction level		
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	4.82%	76.11%	21.25%	75.56%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	100.00%	99.72%	99.83%	99.66%
% reads aligned incorrectly:	0.00%	0.20%	0.03%	0.25%
% reads aligned ambiguously:	0.17%	2.17%	0.68%	2.17%
% reads unaligned:	95.01%	21.52%	78.04%	22.02%
% reads aligned:	4.99%	78.48%	21.96%	77.98%
% of reads with true introns:	4.31%	4.31%	4.31%	4.31%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	4.81%	74.11%	21.14%	73.26%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	100.00%	97.08%	99.29%	96.60%
% bases aligned incorrectly:	0.00%	2.22%	0.15%	2.57%
% bases aligned ambiguously:	0.17%	2.17%	0.68%	2.17%
% bases unaligned:	95.02%	21.50%	78.03%	22.00%
% bases aligned:	4.98%	78.50%	21.97%	78.00%
% of bases in true insertions:	0.41%	0.41%	0.41%	0.41%
% of bases in true deletions:	0.37%	0.37%	0.37%	0.37%
insertions FD rate [1 - PRECISION]:	0.79%	2.71%	0.96%	2.79%
insertions FN rate [1 - RECALL]:	99.68%	84.69%	98.06%	86.45%
deletions FD rate [1 - PRECISION]:	1.01%	3.52%	1.21%	3.63%
deletions FN rate [1 - RECALL]:	99.69%	84.91%	98.12%	86.52%
skipping FD rate [1 - PRECISION]:	77.95%	88.31%	97.45%	97.90%
skipping FN rate [1 - RECALL]:	95.98%	45.91%	82.67%	43.54%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	3.99%	7.32%	21.35%	30.99%
junctions FN rate [1 - RECALL]:	95.88%	46.76%	82.91%	47.78%
Junction Sides none	162	3941	4333	21569
Junction Sides left	0	26	78	650
Junction Sides right	2	59	33	247
Junction Sides both	3950	51016	16377	50038
Junction Sides none	3.93%	7.15%	20.81%	29.74%
Junction Sides left	0.00%	0.04%	0.37%	0.89%
Junction Sides right	0.04%	0.10%	0.15%	0.34%
Junction Sides both	96.01%	92.68%	78.65%	69.01%

Supplementary Table 16 - HISAT tweaking examples on Human T3R1 dataset

Command:	hisat --threads 16 --end-to-end -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> --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>			
Parameters:	ENDTOEND_MODE - NUM_MISMATCH - SEED_LENGTH - SEED_INTERVAL - SEED_EXTENSION - RE_SEED - PENALITY_NONCANONICAL - MAX_MISMATCH_PENALITY - MIN_MISMATCH_PENALITY			
Configuration:	default-0-22-1.15-15-2-3-default-default	default-0-20-1.15-15-2-3-1-0	default-0-15-0.5-20-5-20-1-0	default-0-20-1.15-15-2-20-1-0
Note:	Default	Best recall at base and read level		Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	19.75%	83.88%	83.61%	83.59%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	99.32%	98.55%	98.59%	98.60%
% reads aligned incorrectly:	0.13%	1.23%	1.19%	1.18%
% reads aligned ambiguously:	0.72%	2.11%	2.08%	2.14%
% reads unaligned:	79.40%	12.78%	13.12%	13.09%
% reads aligned:	20.60%	87.22%	86.88%	86.91%
% of reads with true introns:	13.47%	13.46%	13.47%	13.46%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	19.75%	83.08%	82.88%	82.87%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	99.30%	97.60%	97.72%	97.74%
% bases aligned incorrectly:	0.13%	2.03%	1.92%	1.91%
% bases aligned ambiguously:	0.72%	2.11%	2.08%	2.14%
% bases unaligned:	79.40%	12.78%	13.12%	13.08%
% bases aligned:	20.60%	87.22%	86.88%	86.92%
% of bases in true insertions:	0.18%	0.18%	0.18%	0.18%
% of bases in true deletions:	0.17%	0.17%	0.17%	0.17%
insertions FD rate [1 - PRECISION]:	2.51%	3.39%	3.34%	3.31%
insertions FN rate [1 - RECALL]:	99.69%	87.57%	87.03%	87.03%
deletions FD rate [1 - PRECISION]:	1.81%	3.66%	3.58%	3.53%
deletions FN rate [1 - RECALL]:	99.69%	87.34%	86.84%	86.84%
skipping FD rate [1 - PRECISION]:	12.74%	12.98%	7.46%	7.62%
skipping FN rate [1 - RECALL]:	96.09%	51.16%	53.06%	52.44%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	4.76%	8.78%	3.62%	3.56%
junctions FN rate [1 - RECALL]:	95.73%	50.32%	50.55%	50.27%
Junction Sides none	589	12279	4945	4874
Junction Sides left	20	1228	230	233
Junction Sides right	17	506	266	262
Junction Sides both	12548	145672	144984	145810
Junction Sides none	4.47%	7.68%	3.28%	3.22%
Junction Sides left	0.15%	0.76%	0.15%	0.15%
Junction Sides right	0.12%	0.31%	0.17%	0.17%
Junction Sides both	95.24%	91.22%	96.38%	96.44%

Supplementary Table 17 - HISAT2 tweaking parameters and values

Parameters	Tested values	Note
`--end-to-end`	with and without this option	Bowtie2-like parameter, Default = without this option
`-N <NUM_MISMATCH>`	0; 1	Bowtie2-like parameter, Integer, Default = 0, Range = [0,1]
`-L <SEED_LENGTH>`	15; 18; 20; 22	Bowtie2-like parameter, Integer, Default = 22
`-i S,1,<SEED_INTERVAL>`	0.25; 0.5; 0.75; 1.15	Bowtie2-like parameter, Default = 1.15
`-D <SEED_EXTENSION>`	15; 20; 25	Bowtie2-like parameter, Default = 15
`-R <RE_SEED>`	2; 3; 5	Bowtie2-like parameter, Default = 2
`--pen-noncansplice <PENALITY_NONCANONICAL>`	0; 3; 12; 20	Default = 3
`--mp <MAX_MISMATCH_PENALITY>, …`	1; 2; 3; 6	Default = 6
`--mp …, <MIN_MISMATCH_PENALITY>`	0; 1; 2	Default = 2
`--sp <MAX_SOFTCLIPPING_PENALITY>, …`	1; 2; 3	Default = 2
`--sp …, <MIN_SOFTCLIPPING_PENALITY>`	0; 1; 2	Default = 1

Supplementary Table 18 – HISAT2 tweaking examples on Malaria T3R1 dataset

Command:	hisat2 --threads 16 --end-to-end -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>--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>			
Parameters:	ENDTOEND_MODE - NUM_MISMATCH - SEED_LENGTH - SEED_INTERVAL - SEED_EXTENSION - RE_SEED - PENALITY_NONCANONICAL - MAX_MISMATCH_PENALITY - MIN_MISMATHC_PENALITY - MAX_SOFTCLIPPING_PENALITY - MIN_SOFTCLIPPING_PENALITY			
Configuration:	default-0-22-1.15-15-2-3-6-2-2-1	default-1-20-0.5-25-5-20-1-0-3-0	default-1-20-0.5-25-5-20-3-0-2-1	default-0-22-1.15-15-2-20-1-0-2-1
Note:	Default	Best recall at base and read level		Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	8.13%	76.24%	21.80%	76.11%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	99.39%	99.53%	99.30%	99.51%
% reads aligned incorrectly:	0.04%	0.35%	0.15%	0.37%
% reads aligned ambiguously:	0.17%	2.00%	0.39%	1.94%
% reads unaligned:	91.66%	21.41%	77.66%	21.58%
% reads aligned:	8.34%	78.59%	22.34%	78.42%
% of reads with true introns:	4.31%	4.31%	4.31%	4.31%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	7.72%	74.20%	21.64%	73.99%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	99.09%	96.90%	98.97%	96.88%
% bases aligned incorrectly:	0.07%	2.36%	0.22%	2.37%
% bases aligned ambiguously:	0.17%	2.00%	0.39%	1.94%
% bases unaligned:	92.04%	21.44%	77.75%	21.70%
% bases aligned:	7.96%	78.56%	22.25%	78.30%
% of bases in true insertions:	0.41%	0.41%	0.41%	0.41%
% of bases in true deletions:	0.37%	0.37%	0.37%	0.37%
insertions FD rate [1 - PRECISION]:	13.74%	3.37%	4.73%	4.01%
insertions FN rate [1 - RECALL]:	99.53%	85.04%	97.95%	85.31%
deletions FD rate [1 - PRECISION]:	17.72%	4.23%	5.84%	5.05%
deletions FN rate [1 - RECALL]:	99.57%	85.22%	98.04%	85.48%
skipping FD rate [1 - PRECISION]:	87.64%	87.44%	84.04%	88.99%
skipping FN rate [1 - RECALL]:	93.90%	41.29%	82.35%	41.60%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	5.36%	6.57%	5.15%	7.08%
junctions FN rate [1 - RECALL]:	93.81%	42.37%	82.41%	42.70%
Junction Sides none	323	3765	894	4057
Junction Sides left	6	38	7	43
Junction Sides right	7	79	14	82

Junction Sides both	5934	55224	16856	54903
Junction Sides none	5.15%	6.36%	5.03%	6.86%
Junction Sides left	0.09%	0.06%	0.03%	0.07%
Junction Sides right	0.11%	0.13%	0.07%	0.13%
Junction Sides both	94.64%	93.43%	94.85%	92.92%

Supplementary Table 19 – HISAT2 tweaking examples on Human T3R1 dataset

Command:	hisat2 --threads 16 --end-to-end -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>--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>			
Parameters:	ENDTOEND_MODE - NUM_MISMATCH - SEED_LENGTH - SEED_INTERVAL - SEED_EXTENSION - RE_SEED - PENALITY_NONCANONICAL - MAX_MISMATCH_PENALITY - MIN_MISMATHC_PENALITY - MAX_SOFTCLIPPING_PENALITY - MIN_SOFTCLIPPING_PENALITY			
Configuration:	default-0-22-1.15-15-2-3-6-2-2-1	default-1-20-0.5-25-5-12-1-0-3-0	default-1-20-0.5-25-5-20-1-0-2-0	default-1-20-0.5-25-5-20-6-0-2-0
Note:	Default	Best recall at base, read and junction level		
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	30.46%	84.03%	83.94%	21.09%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	98.76%	98.62%	98.56%	99.03%
% reads aligned incorrectly:	0.38%	1.17%	1.21%	0.20%
% reads aligned ambiguously:	0.80%	2.15%	2.08%	0.68%
% reads unaligned:	68.36%	12.65%	12.77%	78.03%
% reads aligned:	31.64%	87.35%	87.23%	21.97%
% of reads with true introns:	13.47%	13.46%	13.47%	13.47%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	29.15%	83.29%	83.09%	20.94%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	98.69%	97.80%	97.75%	99.00%
% bases aligned incorrectly:	0.38%	1.86%	1.91%	0.21%
% bases aligned ambiguously:	0.80%	2.15%	2.08%	0.68%
% bases unaligned:	69.67%	12.70%	12.92%	78.17%
% bases aligned:	30.33%	87.30%	87.08%	21.83%
% of bases in true insertions:	0.18%	0.18%	0.18%	0.18%
% of bases in true deletions:	0.17%	0.17%	0.17%	0.17%
insertions FD rate [1 - PRECISION]:	15.21%	3.77%	4.39%	8.21%
insertions FN rate [1 - RECALL]:	99.55%	87.17%	87.37%	99.66%
deletions FD rate [1 - PRECISION]:	16.50%	3.94%	4.68%	8.01%
deletions FN rate [1 - RECALL]:	99.57%	86.97%	87.14%	99.65%
skipping FD rate [1 - PRECISION]:	24.94%	7.20%	8.98%	17.40%
skipping FN rate [1 - RECALL]:	94.30%	47.90%	48.27%	95.75%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	7.69%	4.42%	4.42%	5.94%
junctions FN rate [1 - RECALL]:	93.65%	44.59%	44.93%	95.37%
Junction Sides none	1490	6033	6094	826
Junction Sides left	24	756	657	18
Junction Sides right	38	720	700	14

Junction Sides both	18641	162470	161470	13596
Junction Sides none	7.37%	3.54%	3.60%	5.71%
Junction Sides left	0.11%	0.44%	0.38%	0.12%
Junction Sides right	0.18%	0.42%	0.41%	0.09%
Junction Sides both	92.31%	95.58%	95.58%	94.06%

Supplementary Table 20 - Mapsplice2 tweaking parameters and values

Parameters	Tested values	Note
`--min-map-len <MIN_MAP_LENGTH>`	15; 20; 25; 33; 50; 66; 75	Default = 50
`--splice-mis <SPLICE_MISMATCHES>`	0; 1; 2	Integer, Range = [0,2], Default = 1
`--max-append-mis <APPEND_MISMATCHES>`	0; 1; 2; 3	Integer, Range = [0,3], Default = 3
`--ins <INSERTION_LENGTH>`	6; 8; 10	Integer, Range = [0,10], Default = 6
`--del <DELETION_LENGTH>`	6; 8; 10	Integer, Range = [0,10], Default = 6
`--filtering <FILTER>`	1; 2	Integer, Range = [1,2], Default = 1

Supplementary Table 21 - Mapsplice2 tweaking examples on Malaria T3R1 dataset

Command:	python mapsplice.py --threads 16 --min-map-len <MIN_MAP_LENGTH> --splice-mis <SPLICE_MISMATCHES> --max-append-mis <APPEND_MISMATCHES> --ins <INSERTION_LENGTH> --del <DELETION_LENGTH> --filtering <FILTER> --non-canonical-double-anchor --output <output path> -c <genome fasta files> -x <index name> -1 <read file 1> -2 <read file 2>			
Parameters:	MIN_MAP_LENGTH - SPLICE_MISMATCHES - APPEND_MISMATCHES - INSERTION_LENGTH - DELETION_LENGTH - FILTER			
Configuration:	50-1-3-6-6-1	25-2-3-6-6-1	25-0-3-6-6-1	25-2-1-10-6-1
Note:	Default	Best recall at base level	Best recall at read level	Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	80.56%	86.96%	87.41%	85.63%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	99.80%	99.51%	99.55%	99.33%
% reads aligned incorrectly:	0.15%	0.42%	0.39%	0.57%
% reads aligned ambiguously:	1.54%	1.97%	2.06%	3.18%
% reads unaligned:	17.75%	10.65%	10.14%	10.62%
% reads aligned:	82.25%	89.35%	89.86%	89.38%
% of reads with true introns:	4.31%	4.31%	4.31%	4.31%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	62.48%	64.10%	63.83%	42.55%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	98.85%	98.51%	98.58%	98.37%
% bases aligned incorrectly:	0.72%	0.96%	0.91%	0.70%
% bases aligned ambiguously:	1.54%	1.97%	2.06%	3.18%
% bases unaligned:	35.26%	32.97%	33.20%	53.57%
% bases aligned:	64.74%	67.03%	66.80%	46.43%
% of bases in true insertions:	0.41%	0.41%	0.41%	0.41%
% of bases in true deletions:	0.37%	0.37%	0.37%	0.37%
insertions FD rate [1 - PRECISION]:	13.81%	13.73%	14.01%	25.60%
insertions FN rate [1 - RECALL]:	92.07%	92.25%	91.93%	88.87%
deletions FD rate [1 - PRECISION]:	17.63%	17.57%	17.16%	27.45%
deletions FN rate [1 - RECALL]:	92.42%	91.94%	94.54%	89.64%
skipping FD rate [1 - PRECISION]:	97.71%	97.54%	95.59%	99.54%
skipping FN rate [1 - RECALL]:	87.19%	86.60%	87.65%	80.18%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	4.99%	4.85%	2.60%	20.37%
junctions FN rate [1 - RECALL]:	87.23%	86.69%	87.64%	80.24%
Junction Sides none	595	618	289	4781
Junction Sides left	18	16	21	21
Junction Sides right	30	16	6	38
Junction Sides both	12243	12757	11844	18932
Junction Sides none	4.61%	4.60%	2.37%	20.11%
Junction Sides left	0.13%	0.11%	0.17%	0.08%
Junction Sides right	0.23%	0.11%	0.04%	0.15%
Junction Sides both	95.01%	95.15%	97.40%	79.63%

Supplementary Table 22 - Mapsplice2 tweaking examples on Human T3R1 dataset

Command:	python mapsplice.py --threads 16 --min-map-len <MIN_MAP_LENGTH> --splice-mis <SPLICE_MISMATCHES> --max-append-mis <APPEND_MISMATCHES> --ins <INSERTION_LENGTH> --del <DELETION_LENGTH> --filtering <FILTER> --non-canonical-double-anchor --output <output path> -c <genome fasta files> -x <index name> -1 <read file 1> -2 <read file 2>			
Parameters:	MIN_MAP_LENGTH - SPLICE_MISMATCHES - APPEND_MISMATCHES - INSERTION_LENGTH - DELETION_LENGTH - FILTER			
Configuration:	50-1-3-6-6-1	25-2-3-10-6-1	25-0-3-6-10-2	25-2-1-10-6-1
Note:	Default	Best recall at base level	Best recall at read level	Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	85.62%	89.29%	89.73%	88.48%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	99.16%	98.73%	98.79%	98.73%
% reads aligned incorrectly:	0.71%	1.14%	1.09%	1.13%
% reads aligned ambiguously:	2.24%	2.55%	2.64%	3.34%
% reads unaligned:	11.43%	7.02%	6.54%	7.05%
% reads aligned:	88.57%	92.98%	93.46%	92.95%
% of reads with true introns:	13.47%	13.46%	13.46%	13.46%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	71.03%	71.93%	71.86%	56.40%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	98.75%	98.24%	98.30%	98.10%
% bases aligned incorrectly:	0.89%	1.28%	1.24%	1.09%
% bases aligned ambiguously:	2.24%	2.55%	2.64%	3.34%
% bases unaligned:	25.84%	24.24%	24.26%	39.17%
% bases aligned:	74.16%	75.76%	75.74%	60.83%
% of bases in true insertions:	0.18%	0.18%	0.18%	0.18%
% of bases in true deletions:	0.17%	0.17%	0.17%	0.17%
insertions FD rate [1 - PRECISION]:	17.60%	17.64%	17.76%	22.36%
insertions FN rate [1 - RECALL]:	94.83%	94.90%	94.68%	91.27%
deletions FD rate [1 - PRECISION]:	19.58%	19.82%	19.01%	23.94%
deletions FN rate [1 - RECALL]:	95.24%	95.23%	96.06%	91.77%
skipping FD rate [1 - PRECISION]:	39.98%	42.62%	31.83%	63.18%
skipping FN rate [1 - RECALL]:	94.99%	94.65%	95.16%	87.69%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	3.66%	4.00%	3.16%	8.03%
junctions FN rate [1 - RECALL]:	94.63%	94.18%	94.80%	87.03%
Junction Sides none	539	644	459	3195
Junction Sides left	35	40	16	69
Junction Sides right	23	27	23	57
Junction Sides both	15748	17065	15265	38054
Junction Sides none	3.29%	3.62%	2.91%	7.72%
Junction Sides left	0.21%	0.22%	0.10%	0.16%
Junction Sides right	0.14%	0.15%	0.14%	0.13%
Junction Sides both	96.34%	96.00%	96.84%	91.97%

Supplementary Table 23 - Novoalign tweaking parameters and values

Parameters	Tested values	Note
`-t <A_SCORE>,...`	default; 10; 12; 13; 20	Default = log4(*N*), where *N* is the reference genome length
`-t ...,<B_SCORE>`	2; 3; 4.5	Default = 4.5
`-h -1 -1`	with and without this option	Default = without this option
`-i PE <FRAGMENT_LENGTH_MEAN>,<FRAGMENT_LENGTH_SD>`	with and without this option	Default *<FRAGMENT_LENGTH_MEAN>* = 250; Default *<FRAGMENT_LENGTH_SD>* = 50

Supplementary Table 24 - Novoalign tweaking examples on Malaria T3R1 dataset

Command:	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			
Parameters:	REPEAT_FILTER - PAIREND_DEFAULT - A_SCORE - B_SCORE			
Configuration:	off-off-default-default	off-off-default-default	on-off-12-3	off-on-default-default
Note:	Default	Best recall at base and read level		Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	98.29%	98.29%	92.78%	98.15%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	99.98%	99.98%	99.98%	99.98%
% reads aligned incorrectly:	0.01%	0.01%	0.01%	0.01%
% reads aligned ambiguously:	1.67%	1.67%	1.67%	1.81%
% reads unaligned:	0.03%	0.03%	5.54%	0.03%
% reads aligned:	99.97%	99.97%	94.46%	99.97%
% of reads with true introns:	4.31%	4.31%	4.31%	4.31%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	90.04%	90.04%	85.43%	89.92%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	98.59%	98.59%	98.69%	98.59%
% bases aligned incorrectly:	1.28%	1.28%	1.12%	1.27%
% bases aligned ambiguously:	1.67%	1.67%	1.67%	1.81%
% bases unaligned:	7.01%	7.01%	11.78%	7.00%
% bases aligned:	92.99%	92.99%	88.22%	93.00%
% of bases in true insertions:	0.41%	0.41%	0.41%	0.41%
% of bases in true deletions:	0.37%	0.37%	0.37%	0.37%
insertions FD rate [1 - PRECISION]:	26.55%	26.55%	25.09%	26.52%
insertions FN rate [1 - RECALL]:	40.03%	40.03%	45.90%	40.11%
deletions FD rate [1 - PRECISION]:	26.43%	26.43%	25.31%	26.43%
deletions FN rate [1 - RECALL]:	44.82%	44.82%	50.34%	44.89%
skipping FD rate [1 - PRECISION]:	0.83%	0.83%	1.02%	0.81%
skipping FN rate [1 - RECALL]:	8.36%	8.36%	12.75%	8.37%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	2.86%	2.86%	2.63%	2.85%
junctions FN rate [1 - RECALL]:	10.12%	10.12%	14.27%	10.11%
Junction Sides none	2517	2517	2187	2517
Junction Sides left	3	3	10	4
Junction Sides right	8	8	16	4
Junction Sides both	86121	86121	82139	86129
Junction Sides none	2.83%	2.83%	2.59%	2.83%
Junction Sides left	0.00%	0.00%	0.01%	0.00%
Junction Sides right	0.00%	0.00%	0.01%	0.00%
Junction Sides both	97.14%	97.14%	97.37%	97.15%

Supplementary Table 25 - Novoalign tweaking examples on Human T3R1 dataset

Command:	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			
Parameters:	REPEAT_FILTER - PAIREND_DEFAULT - A_SCORE - B_SCORE			
Configuration:	off-off-default-default	on-off-10-4.5	on-off-20-2	on-off-12-4.5
Note:	Default	Best recall at base level, Best recall at read level, Best recall at junction level		
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	97.25%	97.28%	75.49%	97.27%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	99.78%	99.78%	99.83%	99.78%
% reads aligned incorrectly:	0.21%	0.21%	0.12%	0.21%
% reads aligned ambiguously:	2.46%	2.46%	2.22%	2.46%
% reads unaligned:	0.08%	0.05%	22.17%	0.06%
% reads aligned:	99.92%	99.95%	77.83%	99.94%
% of reads with true introns:	13.47%	13.46%	13.47%	13.47%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	90.27%	90.29%	71.39%	90.28%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	99.24%	99.24%	99.57%	99.24%
% bases aligned incorrectly:	0.68%	0.68%	0.30%	0.68%
% bases aligned ambiguously:	2.46%	2.46%	2.22%	2.46%
% bases unaligned:	6.59%	6.57%	26.09%	6.58%
% bases aligned:	93.41%	93.43%	73.91%	93.42%
% of bases in true insertions:	0.18%	0.18%	0.18%	0.18%
% of bases in true deletions:	0.17%	0.17%	0.17%	0.17%
insertions FD rate [1 - PRECISION]:	26.10%	26.08%	18.88%	26.07%
insertions FN rate [1 - RECALL]:	41.81%	41.78%	76.09%	41.78%
deletions FD rate [1 - PRECISION]:	24.21%	24.22%	18.85%	24.22%
deletions FN rate [1 - RECALL]:	43.34%	43.30%	77.23%	43.31%
skipping FD rate [1 - PRECISION]:	9.84%	9.93%	8.27%	9.92%
skipping FN rate [1 - RECALL]:	23.41%	23.36%	58.80%	23.39%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	6.93%	6.94%	3.96%	6.93%
junctions FN rate [1 - RECALL]:	21.53%	21.51%	57.00%	21.51%
Junction Sides none	13106	13140	3879	13128
Junction Sides left	2026	2033	652	2028
Junction Sides right	1975	1977	661	1977
Junction Sides both	230078	230145	126093	230132
Junction Sides none	5.30%	5.31%	2.95%	5.30%
Junction Sides left	0.81%	0.82%	0.49%	0.82%
Junction Sides right	0.79%	0.79%	0.50%	0.79%
Junction Sides both	93.07%	93.06%	96.04%	93.07%

Supplementary Table 26 - Olego tweaking parameters and values

Parameters	Tested values	Note
`--max-total-diff <TOTAL_DIFF>`	4; 5; 6; 7; 8; 9; 10	Float or Integer; Default = 0.06
`--word-size <WORD_SIZE>`	13; 14; 15; 16	Default = 15
`--max-word-diff <MAX_WORD_DIFF>`	0; 1; 2	Default = 0
`--word-max-overlap <WORD_MAX_OVERLAP>`	0; 1	Default = 1
`--allow-rep-anchor`	with or without this option	Default = without this option
`--min-anchor <MIN_ANCHOR>`	5; 8; 11	Default = 8

Supplementary Table 27 - Olego tweaking examples on Malaria T3R1 dataset

Command:	olego --output-file output_1.sam --num-threads 16 --regression-model <regression model> --verbose --junction-file <junction file> --max-total-diff <TOTAL_DIFF> --word-size <WORD_SIZE> --max-word-diff <MAX_WORD_DIFF> --word-max-overlap <WORD_MAX_OVERLAP> --min-anchor <MIN_ANCHOR> --allow-rep-anchor <olego index> <read file 1> olego --output-file output_2.sam --num-threads 16 --regression-model <regression model> --verbose --junction-file <junction file> --max-total-diff <TOTAL_DIFF> --word-size <WORD_SIZE> --max-word-diff <MAX_WORD_DIFF> --word-max-overlap <WORD_MAX_OVERLAP> --min-anchor <MIN_ANCHOR> --allow-rep-anchor <olego index> <read file 2>			
Parameters:	TOTAL_DIFF - WORD_SIZE - MAX_WORD_DIFF - WORD_MAX_OVERLAP - REP_ANCHOR - MIN_ANCHOR			
Configuration:	0.06-15-0-1-off-8	10-14-2-1-off-11	10-15-2-1-off-11	10-13-2-1-off-11
Note:	Default	Best recall at base and read level		Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	8.38%	30.64%	29.33%	30.60%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	98.81%	98.59%	98.55%	98.62%
% reads aligned incorrectly:	0.10%	0.43%	0.43%	0.42%
% reads aligned ambiguously:	0.00%	0.00%	0.00%	0.00%
% reads unaligned:	91.52%	68.93%	70.24%	68.98%
% reads aligned:	8.48%	31.07%	29.76%	31.02%
% of reads with true introns:	4.31%	4.31%	4.31%	4.31%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	8.37%	30.52%	29.22%	30.49%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	98.65%	98.18%	98.14%	98.21%
% bases aligned incorrectly:	0.11%	0.56%	0.55%	0.55%
% bases aligned ambiguously:	0.00%	0.00%	0.00%	0.00%
% bases unaligned:	91.52%	68.92%	70.23%	68.96%
% bases aligned:	8.48%	31.08%	29.77%	31.04%
% of bases in true insertions:	0.41%	0.41%	0.41%	0.41%
% of bases in true deletions:	0.37%	0.37%	0.37%	0.37%
insertions FD rate [1 - PRECISION]:	1.47%	6.05%	6.44%	6.06%
insertions FN rate [1 - RECALL]:	98.68%	89.98%	90.87%	89.74%
deletions FD rate [1 - PRECISION]:	1.45%	6.44%	6.85%	6.45%
deletions FN rate [1 - RECALL]:	98.76%	89.62%	90.56%	89.39%
skipping FD rate [1 - PRECISION]:	34.90%	9.95%	20.89%	13.29%
skipping FN rate [1 - RECALL]:	93.12%	69.41%	70.55%	68.64%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	1.05%	0.68%	0.75%	0.76%
junctions FN rate [1 - RECALL]:	93.00%	69.61%	70.77%	68.96%
Junction Sides none	71	197	204	224
Junction Sides left	0	2	4	2
Junction Sides right	0	0	1	1
Junction Sides both	6708	29117	28010	29741
Junction Sides none	1.04%	0.67%	0.72%	0.74%
Junction Sides left	0.00%	0.00%	0.01%	0.00%
Junction Sides right	0.00%	0.00%	0.00%	0.00%
Junction Sides both	98.95%	99.32%	99.25%	99.24%

Supplementary Table 28 - Olego tweaking examples on Human T3R1 dataset

Command:	olego --output-file output_1.sam --num-threads 16 --regression-model <regression model> --verbose --junction-file <junction file> --max-total-diff <TOTAL_DIFF> --word-size <WORD_SIZE> --max-word-diff <MAX_WORD_DIFF> --word-max-overlap <WORD_MAX_OVERLAP> --min-anchor <MIN_ANCHOR> --allow-rep-anchor <olego index> <read file 1> olego --output-file output_2.sam --num-threads 16 --regression-model <regression model> --verbose --junction-file <junction file> --max-total-diff <TOTAL_DIFF> --word-size <WORD_SIZE> --max-word-diff <MAX_WORD_DIFF> --word-max-overlap <WORD_MAX_OVERLAP> --min-anchor <MIN_ANCHOR> --allow-rep-anchor <olego index> <read file 2>			
Parameters:	TOTAL_DIFF - WORD_SIZE - MAX_WORD_DIFF - WORD_MAX_OVERLAP - REP_ANCHOR - MIN_ANCHOR			
Configuration:	0.06-15-0-1-off-8	10-13-2-1-off-11	10-14-2-0-off-11	6-13-1-1-off-11
Note:	Default	Best recall at base, read and junction level		
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	23.30%	32.63%	31.79%	30.99%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	98.18%	97.52%	97.67%	97.70%
% reads aligned incorrectly:	0.43%	0.82%	0.75%	0.72%
% reads aligned ambiguously:	0.00%	0.00%	0.00%	0.00%
% reads unaligned:	76.27%	66.55%	67.46%	68.29%
% reads aligned:	23.73%	33.45%	32.54%	31.71%
% of reads with true introns:	13.47%	13.46%	13.47%	13.47%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	23.29%	32.58%	31.75%	30.97%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	98.13%	97.37%	97.51%	97.64%
% bases aligned incorrectly:	0.44%	0.87%	0.80%	0.74%
% bases aligned ambiguously:	0.00%	0.00%	0.00%	0.00%
% bases unaligned:	76.27%	66.55%	67.45%	68.29%
% bases aligned:	23.73%	33.45%	32.55%	31.71%
% of bases in true insertions:	0.18%	0.18%	0.18%	0.18%
% of bases in true deletions:	0.17%	0.17%	0.17%	0.17%
insertions FD rate [1 - PRECISION]:	2.49%	14.60%	16.68%	3.97%
insertions FN rate [1 - RECALL]:	98.80%	93.81%	94.29%	95.49%
deletions FD rate [1 - PRECISION]:	2.54%	15.67%	17.26%	4.36%
deletions FN rate [1 - RECALL]:	98.76%	93.62%	94.01%	95.39%
skipping FD rate [1 - PRECISION]:	37.20%	7.28%	6.17%	8.11%
skipping FN rate [1 - RECALL]:	93.03%	81.51%	82.38%	85.43%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	7.92%	2.67%	2.72%	2.64%
junctions FN rate [1 - RECALL]:	93.09%	81.06%	81.76%	85.06%
Junction Sides none	1656	1426	1389	1115
Junction Sides left	39	38	45	32
Junction Sides right	47	54	57	39
Junction Sides both	20260	55538	53505	43820
Junction Sides none	7.52%	2.49%	2.52%	2.47%
Junction Sides left	0.17%	0.06%	0.08%	0.07%
Junction Sides right	0.21%	0.09%	0.10%	0.08%
Junction Sides both	92.08%	97.33%	97.28%	97.36%

Supplementary Table 29 - RUM tweaking parameters and values

Parameters	Tested values	Note
`--blat-min-identity <BLAT_MIN_IDENTITY>`	75; 88; 90; 93; 95; 98	Default = 93
`--blat-rep-match <BLAT_REP_MATCH>`	256; 1024; 4096	Default = 256
`--blat-step-size <BLAT_STEP_SIZE>`	6; 8; 10; 11; 12; 13; 14	Default = 6
`--blat-tile-size <BLAT_TILE_SIZE>`	8; 10; 11; 12; 13; 14	Default = 12

Supplementary Table 30 - RUM tweaking examples on Malaria T3R1 dataset

Command:	rum_runner align --index-dir <index directory> --name <job name> --output <output path> --chunks 16 <read file 1> <read file 2> --verbose --preserve-names --blat-min-identity <BLAT_MIN_IDENTITY> --blat-rep-match <BLAT_REP_MATCH> --blat-step-size <BLAT_STEP_SIZE> --blat-tile-size <BLAT_TILE_SIZE>			
Parameters:	BLAT_MIN_IDENTITY - BLAT_REP_MATCH - BLAT_STEP_SIZE - BLAT_TILE_SIZE			
Configuration:	93-256-6-12	88-4096-10-8	95-256-6-12	75-4096-6-8
Note:	Default	Best recall at base and read level		Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	54.30%	63.67%	35.33%	63.67%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	99.83%	99.82%	99.79%	99.82%
% reads aligned incorrectly:	0.09%	0.11%	0.07%	0.11%
% reads aligned ambiguously:	1.37%	1.45%	1.08%	1.46%
% reads unaligned:	44.24%	34.77%	63.52%	34.76%
% reads aligned:	55.76%	65.23%	36.48%	65.24%
% of reads with true introns:	4.31%	4.31%	4.31%	4.31%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	50.53%	59.26%	32.86%	59.25%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	98.62%	98.61%	98.52%	98.61%
% bases aligned incorrectly:	0.70%	0.83%	0.49%	0.83%
% bases aligned ambiguously:	1.37%	1.45%	1.08%	1.46%
% bases unaligned:	47.40%	38.46%	65.57%	38.46%
% bases aligned:	52.60%	61.54%	34.43%	61.54%
% of bases in true insertions:	0.41%	0.41%	0.41%	0.41%
% of bases in true deletions:	0.37%	0.37%	0.37%	0.37%
insertions FD rate [1 - PRECISION]:	17.23%	17.24%	17.28%	17.23%
insertions FN rate [1 - RECALL]:	96.60%	96.60%	96.59%	96.61%
deletions FD rate [1 - PRECISION]:	29.56%	29.74%	29.24%	29.76%
deletions FN rate [1 - RECALL]:	71.31%	67.95%	80.19%	67.77%
skipping FD rate [1 - PRECISION]:	74.38%	76.51%	78.05%	82.02%
skipping FN rate [1 - RECALL]:	71.57%	69.46%	78.53%	68.67%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	54.78%	53.61%	56.94%	52.93%
junctions FN rate [1 - RECALL]:	73.18%	71.15%	79.73%	70.54%
Junction Sides none	30642	31428	25308	31206
Junction Sides left	317	342	249	337
Junction Sides right	168	183	127	192
Junction Sides both	25705	27650	19427	28230
Junction Sides none	53.91%	52.72%	56.10%	52.04%
Junction Sides left	0.55%	0.57%	0.55%	0.56%
Junction Sides right	0.29%	0.30%	0.28%	0.32%
Junction Sides both	45.22%	46.39%	43.06%	47.07%

Supplementary Table 31 - RUM tweaking examples on Human T3R1 dataset

Command:	rum_runner align --index-dir <index directory> --name <job name> --output <output path> --chunks 16 <read file 1> <read file 2> --verbose --preserve-names --blat-min-identity <BLAT_MIN_IDENTITY> --blat-rep-match <BLAT_REP_MATCH> --blat-step-size <BLAT_STEP_SIZE> --blat-tile-size <BLAT_TILE_SIZE>			
Parameters:	BLAT_MIN_IDENTITY - BLAT_REP_MATCH - BLAT_STEP_SIZE - BLAT_TILE_SIZE			
Configuration:	93-256-6-12	93-256-6-12	75-4096-10-8	98-256-14-8
Note:	Default	Best recall at base, read and junction level		
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	72.60%	72.60%	41.17%	20.35%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	99.53%	99.53%	97.48%	99.63%
% reads aligned incorrectly:	0.33%	0.33%	1.06%	0.07%
% reads aligned ambiguously:	2.52%	2.52%	1.82%	0.11%
% reads unaligned:	24.55%	24.55%	55.95%	79.47%
% reads aligned:	75.45%	75.45%	44.05%	20.53%
% of reads with true introns:	13.47%	13.47%	13.47%	13.47%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	68.37%	68.37%	39.37%	20.30%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	99.06%	99.06%	97.09%	99.62%
% bases aligned incorrectly:	0.64%	0.64%	1.17%	0.07%
% bases aligned ambiguously:	2.52%	2.52%	1.82%	0.11%
% bases unaligned:	28.47%	28.47%	57.64%	79.52%
% bases aligned:	71.53%	71.53%	42.36%	20.48%
% of bases in true insertions:	0.18%	0.18%	0.18%	0.18%
% of bases in true deletions:	0.17%	0.17%	0.17%	0.17%
insertions FD rate [1 - PRECISION]:	17.16%	17.16%	17.98%	0.00%
insertions FN rate [1 - RECALL]:	96.86%	96.86%	98.51%	100.00%
deletions FD rate [1 - PRECISION]:	33.92%	33.92%	34.74%	17.81%
deletions FN rate [1 - RECALL]:	71.69%	71.69%	88.18%	99.99%
skipping FD rate [1 - PRECISION]:	10.55%	10.55%	5.50%	0.88%
skipping FN rate [1 - RECALL]:	84.13%	84.13%	95.90%	96.30%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	36.59%	36.59%	33.31%	0.78%
junctions FN rate [1 - RECALL]:	82.67%	82.67%	93.34%	95.91%
Junction Sides none	27738	27738	9481	74
Junction Sides left	1321	1321	187	13
Junction Sides right	274	274	96	7
Junction Sides both	50834	50834	19552	11994
Junction Sides none	34.60%	34.60%	32.34%	0.61%
Junction Sides left	1.64%	1.64%	0.63%	0.10%
Junction Sides right	0.34%	0.34%	0.32%	0.05%
Junction Sides both	63.41%	63.41%	66.69%	99.22%

Supplementary Table 32 - SOAPsplice tweaking parameters and values

Parameters	Tested values	Note
`-m <MISMATCHES>`	0; 1; 2; 3; 4; 5	Range = [0, 5]; Default = 3
`-g <INDEL>`	0; 1; 2	Range = [0, 2]; Default = 2
`-i <TAIL>`	5; 7; 10; 25; 33; 42; 50; 58; 66; 75	Default = 7
`-a <SHORT_LENGTH>`	6; 8; 10; 20; 25	Default = 8

Supplementary Table 33 - SOAPsplice tweaking examples on Malaria T3R1 dataset

Command:	soapsplice -d <index> -1 <read file 1> -2 <read file 2> -o <output file> -p 16 -f 2 -l 0 -I <FRAGMENT_LENGTH_MEAN> -m <MISMATCHES> -g <INDEL> -i <TAIL> -a <SHORT_LENGTH>			
Parameters:	MISMATCHES - INDEL - TAIL - SHORT_LENGTH			
Configuration:	3-2-7-8	5-2-33-8	5-2-66-8	5-1-5-8
Note:	Default	Best recall at base level	Best recall at read level	Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	36.65%	69.63%	91.04%	55.74%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	98.45%	98.59%	97.31%	98.50%
% reads aligned incorrectly:	0.57%	0.99%	2.51%	0.84%
% reads aligned ambiguously:	0.00%	0.00%	0.00%	0.00%
% reads unaligned:	62.78%	29.38%	6.45%	43.42%
% reads aligned:	37.22%	70.62%	93.55%	56.58%
% of reads with true introns:	4.31%	4.31%	4.31%	4.31%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	28.50%	51.46%	41.78%	45.81%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	96.38%	96.83%	94.17%	96.63%
% bases aligned incorrectly:	1.06%	1.68%	2.58%	1.59%
% bases aligned ambiguously:	0.00%	0.00%	0.00%	0.00%
% bases unaligned:	70.44%	46.86%	55.64%	52.60%
% bases aligned:	29.56%	53.14%	44.36%	47.40%
% of bases in true insertions:	0.41%	0.41%	0.41%	0.41%
% of bases in true deletions:	0.37%	0.37%	0.37%	0.37%
insertions FD rate [1 - PRECISION]:	51.96%	46.76%	47.86%	12.58%
insertions FN rate [1 - RECALL]:	99.70%	98.69%	96.52%	99.97%
deletions FD rate [1 - PRECISION]:	48.31%	54.22%	56.81%	15.99%
deletions FN rate [1 - RECALL]:	99.72%	98.88%	97.37%	99.97%
skipping FD rate [1 - PRECISION]:	99.71%	99.00%	99.23%	99.51%
skipping FN rate [1 - RECALL]:	87.04%	88.41%	95.89%	86.19%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	35.78%	23.06%	18.23%	32.78%
junctions FN rate [1 - RECALL]:	86.77%	88.08%	95.79%	85.87%
Junction Sides none	7012	3394	882	6553
Junction Sides left	34	20	12	33
Junction Sides right	16	9	5	15
Junction Sides both	12679	11421	4034	13541
Junction Sides none	35.51%	22.86%	17.87%	32.53%
Junction Sides left	0.17%	0.13%	0.24%	0.16%
Junction Sides right	0.08%	0.06%	0.10%	0.07%
Junction Sides both	64.22%	76.94%	81.77%	67.22%

Supplementary Table 34 - SOAPsplice tweaking examples on Human T3R1 dataset

Command:	soapsplice -d <index> -1 <read file 1> -2 <read file 2> -o <output file> -p 16 -f 2 -l 0 -I <FRAGMENT_LENGTH_MEAN> -m <MISMATCHES> -g <INDEL> -i <TAIL> -a <SHORT_LENGTH>			
Parameters:	MISMATCHES - INDEL - TAIL - SHORT_LENGTH			
Configuration:	5-2-66-8	3-2-7-8	5-2-33-8	5-1-5-8
Note:	Best recall at read level	Default	Best recall at base level	Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	88.73%	56.86%	79.92%	71.72%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	95.16%	96.18%	97.28%	96.99%
% reads aligned incorrectly:	4.50%	2.25%	2.22%	2.22%
% reads aligned ambiguously:	0.04%	0.02%	0.04%	0.03%
% reads unaligned:	6.73%	40.87%	17.82%	26.03%
% reads aligned:	93.27%	59.13%	82.18%	73.97%
% of reads with true introns:	13.46%	13.47%	13.46%	13.46%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	53.41%	49.19%	64.88%	64.77%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	92.86%	95.24%	96.56%	96.51%
% bases aligned incorrectly:	4.10%	2.45%	2.31%	2.33%
% bases aligned ambiguously:	0.04%	0.02%	0.04%	0.03%
% bases unaligned:	42.45%	48.34%	32.77%	32.87%
% bases aligned:	57.55%	51.66%	67.23%	67.13%
% of bases in true insertions:	0.18%	0.18%	0.18%	0.18%
% of bases in true deletions:	0.17%	0.17%	0.17%	0.17%
insertions FD rate [1 - PRECISION]:	50.47%	90.84%	52.07%	54.14%
insertions FN rate [1 - RECALL]:	96.92%	99.78%	98.95%	99.97%
deletions FD rate [1 - PRECISION]:	56.16%	87.88%	55.00%	41.33%
deletions FN rate [1 - RECALL]:	97.39%	99.78%	99.00%	99.98%
skipping FD rate [1 - PRECISION]:	58.57%	24.46%	30.51%	32.90%
skipping FN rate [1 - RECALL]:	96.72%	90.03%	90.85%	89.26%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	8.96%	6.33%	6.03%	8.16%
junctions FN rate [1 - RECALL]:	96.73%	90.00%	90.91%	89.33%
Junction Sides none	931	1947	1679	2741
Junction Sides left	4	15	12	15
Junction Sides right	9	20	19	21
Junction Sides both	9595	29344	26680	31294
Junction Sides none	8.83%	6.21%	5.91%	8.04%
Junction Sides left	0.03%	0.04%	0.04%	0.04%
Junction Sides right	0.08%	0.06%	0.06%	0.06%
Junction Sides both	91.04%	93.67%	93.97%	91.84%

Supplementary Table 35 - STAR tweaking parameters and values

Parameters	Tested values	Note
--limitOutSJcollapsed *<NUM_COLLAPSED_JUNCTIONS>*	1000000; 5000000	Default = 1000000
--limitSjdbInsertNsj *<NUM_INSERTED_JUNCTIONS>*	1000000; 5000000	Default = 1000000
--outFilterMultimapNmax *<NUM_MULTIMAPPER>*	10; 100	Default = 10
--outFilterMismatchNmax *<NUM_FILTER_MISMATCHES>*	3; 5; 8; 10; 20; 25; 33	Default = 10
--outFilterMismatchNoverLmax *<RATIO_FILTER_MISMATCHES>*	0.3; 1	Default = 0.3
--seedSearchStartLmax *<SEED_LENGTH>*	12; 30; 33; 50	Default = 50
--alignSJoverhangMin *<OVERHANG>*	3; 5; 8; 15	Default = 5
--alignEndsType *<END_ALIGNMENT_TYPE>*	“Local”; “EndToEnd”; “Extend5pOfRead1”; “Extend3pOfRead1”	Default = “Local”
--outFilterMatchNminOverLread <NUM_FILTER_MATCHES>	0; 0.66	Default = 0.66
--outFilterScoreMinOverLread <NUM_FILTER_SCORE>	0.3; 0.66	Default = 0.66
--winAnchorMultimapNmax <NUM_ANCHOR>	50; 200	Default = 50
--alignSJDBoverhangMin <OVERHANG_ANNOTATED>	1; 3	Default = 3
--outFilterType <OUT_FILTER>	“Normal”; “BySJout”	Default = “Normal”

Supplementary Table 36 - STAR tweaking examples on Malaria T3R1 dataset

Command:	STAR --runThreadN 16 --genomeDir <index path> --readFilesIn <read file 1> <read file 2> --outFileNamePrefix <output alignment prefix> --twopassMode Basic --outSAMunmapped Within --limitOutSJcollapsed <NUM_COLLAPSED_JUNCTIONS> --limitSjdbInsertNsj <NUM_INSERTED_JUNCTIONS> --outFilterMultimapNmax <NUM_MULTIMAPPER> --outFilterMismatchNmax <NUM_FILTER_MISMATCHES> --outFilterMismatchNoverLmax <RATIO_FILTER_MISMATCHES> --seedSearchStartLmax <SEED_LENGTH> --alignSJoverhangMin <OVERHANG> --alignEndsType <END_ALIGNMENT_TYPE> --outFilterMatchNminOverLread <NUM_FILTER_MATCHES> --outFilterScoreMinOverLread <NUM_FILTER_SCORE> --winAnchorMultimapNmax <NUM_ANCHOR> --alignSJDBoverhangMin <OVERHANG_ANNOTATED> --outFilterType <OUT_FILTER>			
Parameters:	NUM_COLLAPSED_JUNCTIONS - NUM_INSERTED_JUNCTIONS - NUM_MULTIMAPPER - NUM_FILTER_MISMATCHES - RATIO_FILTER_MISMATCHES - SEED_LENGTH - OVERHANG - END_ALIGNMENT_TYPE - NUM_FILTER_MATCHES - NUM_FILTER_SCORE - NUM_ANCHOR - OVERHANG_ANNOTATED - OUT_FILTER			
Configuration:	1000000-1000000-10-10-0.3-50-5-Local-0.66-0.66-50-3-Normal	1000000-1000000-100-33-0.3-12-15-Local-0-0.3-50-3-BySJout	1000000-1000000-100-33-0.3-30-15-Local-0-0.3-50-3-BySJout	1000000-1000000-100-33-0.3-12-15-Local-0-0.3-50-1-BySJout
Note:	Default	Best recall at base level	Best recall at read level	Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	71.08%	95.29%	95.39%	95.29%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	99.87%	99.76%	99.60%	99.76%
% reads aligned incorrectly:	0.08%	0.22%	0.37%	0.22%
% reads aligned ambiguously:	5.42%	4.10%	3.86%	4.10%
% reads unaligned:	23.42%	0.39%	0.38%	0.39%
% reads aligned:	76.58%	99.61%	99.62%	99.61%
% of reads with true introns:	4.31%	4.31%	4.31%	4.31%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	62.71%	86.30%	85.80%	86.30%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	97.60%	96.59%	96.40%	96.59%
% bases aligned incorrectly:	1.54%	3.04%	3.19%	3.03%
% bases aligned ambiguously:	5.42%	4.10%	3.86%	4.10%
% bases unaligned:	30.33%	6.56%	7.15%	6.57%
% bases aligned:	69.67%	93.44%	92.85%	93.43%
% of bases in true insertions:	0.41%	0.41%	0.41%	0.41%
% of bases in true deletions:	0.37%	0.37%	0.37%	0.37%
insertions FD rate [1 - PRECISION]:	16.64%	18.70%	19.04%	18.69%
insertions FN rate [1 - RECALL]:	79.26%	66.28%	68.81%	66.28%
deletions FD rate [1 - PRECISION]:	21.71%	23.51%	23.41%	23.50%
deletions FN rate [1 - RECALL]:	77.77%	64.22%	66.59%	64.22%
skipping FD rate [1 - PRECISION]:	99.81%	90.69%	86.26%	90.69%
skipping FN rate [1 - RECALL]:	48.56%	28.51%	31.24%	26.75%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	20.86%	2.70%	2.60%	2.74%
junctions FN rate [1 - RECALL]:	48.34%	28.42%	31.16%	26.65%

Junction Sides none	12960	1832	1688	1899
Junction Sides left	38	23	21	26
Junction Sides right	44	47	50	49
Junction Sides both	49502	68583	65963	70281
Junction Sides none	20.72%	2.59%	2.49%	2.62%
Junction Sides left	0.06%	0.03%	0.03%	0.03%
Junction Sides right	0.07%	0.06%	0.07%	0.06%
Junction Sides both	79.14%	97.30%	97.40%	97.26%

Supplementary Table 37 - STAR tweaking examples on Human T3R1 dataset

Command:	STAR --runThreadN 16 --genomeDir <index path> --readFilesIn <read file 1> <read file 2> --outFileNamePrefix <output alignment prefix> --twopassMode Basic --outSAMunmapped Within --limitOutSJcollapsed <NUM_COLLAPSED_JUNCTIONS> --limitSjdbInsertNsj <NUM_INSERTED_JUNCTIONS> --outFilterMultimapNmax <NUM_MULTIMAPPER> --outFilterMismatchNmax <NUM_FILTER_MISMATCHES> --outFilterMismatchNoverLmax <RATIO_FILTER_MISMATCHES> --seedSearchStartLmax <SEED_LENGTH> --alignSJoverhangMin <OVERHANG> --alignEndsType <END_ALIGNMENT_TYPE> --outFilterMatchNminOverLread <NUM_FILTER_MATCHES> --outFilterScoreMinOverLread <NUM_FILTER_SCORE> --winAnchorMultimapNmax <NUM_ANCHOR> --alignSJDBoverhangMin <OVERHANG_ANNOTATED> --outFilterType <OUT_FILTER>			
Parameters:	NUM_COLLAPSED_JUNCTIONS - NUM_INSERTED_JUNCTIONS - NUM_MULTIMAPPER - NUM_FILTER_MISMATCHES - RATIO_FILTER_MISMATCHES - SEED_LENGTH - OVERHANG - END_ALIGNMENT_TYPE - NUM_FILTER_MATCHES - NUM_FILTER_SCORE - NUM_ANCHOR - OVERHANG_ANNOTATED - OUT_FILTER			
Configuration:	1000000-1000000-10-10-0.3-50-5-Local-0.66-0.66-50-3-Normal	1000000-1000000-100-33-0.3-12-15-Local-0-0.3-50-3-BySJout	1000000-1000000-100-20-0.3-30-15-Local-0-0.66-200-3-BySJout	1000000-1000000-100-33-0.3-12-15-Local-0-0.3-50-1-BySJout
Note:	Default	Best recall at base and read level		Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	81.00%	95.65%	88.98%	95.56%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	99.57%	99.15%	99.56%	99.16%
% reads aligned incorrectly:	0.34%	0.81%	0.39%	0.80%
% reads aligned ambiguously:	2.24%	2.95%	2.31%	3.04%
% reads unaligned:	16.42%	0.59%	8.32%	0.60%
% reads aligned:	83.58%	99.41%	91.68%	99.40%
% of reads with true introns:	13.47%	13.46%	13.47%	13.46%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	75.56%	89.11%	84.00%	89.03%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	98.90%	98.00%	98.60%	98.01%
% bases aligned incorrectly:	0.83%	1.81%	1.18%	1.80%
% bases aligned ambiguously:	2.24%	2.95%	2.31%	3.04%
% bases unaligned:	21.37%	6.13%	12.51%	6.13%
% bases aligned:	78.63%	93.87%	87.49%	93.87%
% of bases in true insertions:	0.18%	0.18%	0.18%	0.18%
% of bases in true deletions:	0.17%	0.17%	0.17%	0.17%
insertions FD rate [1 - PRECISION]:	16.58%	19.08%	17.63%	19.04%
insertions FN rate [1 - RECALL]:	86.52%	72.67%	79.19%	72.72%
deletions FD rate [1 - PRECISION]:	17.31%	20.44%	18.75%	20.36%
deletions FN rate [1 - RECALL]:	84.13%	69.90%	76.48%	69.95%
skipping FD rate [1 - PRECISION]:	16.03%	0.91%	0.48%	0.97%
skipping FN rate [1 - RECALL]:	63.69%	42.62%	52.06%	41.52%
---------- JUNC LEVEL ----------				

junctions FD rate [1 - PRECISION]:	1.50%	1.20%	0.83%	1.36%
junctions FN rate [1 - RECALL]:	62.13%	39.61%	48.84%	38.37%
Junction Sides none	1471	1642	926	1730
Junction Sides left	101	236	151	362
Junction Sides right	116	271	174	395
Junction Sides both	111051	177069	150002	180704
Junction Sides none	1.30%	0.91%	0.61%	0.94%
Junction Sides left	0.08%	0.13%	0.09%	0.19%
Junction Sides right	0.10%	0.15%	0.11%	0.21%
Junction Sides both	98.50%	98.80%	99.17%	98.64%

Supplementary Table 38 - Subread tweaking parameters and values

Parameters	Tested values	Note
`-d <MIN_FRAGMENT_LENGTH>`	0; 50	Default = 50
`-I <INDEL>`	3; 5; 8; 10; 15; 20	Default = 5
`-m <NUM_HIT_SUBREADS>`	1; 3; 5	Default = 3
`-M <MISMATCHES>`	3; 5; 8; 10; 20; 30	Default = 3
`-n <NUM_EXTRACTED_SUBREADS>`	5; 10; 15	Default = 10
`-p <NUM_HIT_PAIR_SUBREADS>`	1; 3	Default = 1
`--complexIndels`	With and without this option	Default = without this option

Supplementary Table 39 - Subread tweaking examples on Malaria T3R1 dataset

Command:	subjunc -i <index> -r <read file 1> -R <read file 2> -T 16 --allJunctions --SAMoutput -o <output alignment> -d <MIN_FRAGMENT_LENGTH> -I <INDEL> -m <NUM_HIT_SUBREADS> -M <MISMATCHES> -n <NUM_EXTRACTED_SUBREADS> -p <NUM_HIT_PAIR_SUBREADS> --complexIndels			
Parameters:	MIN_FRAGMENT_LENGTH - INDEL - NUM_HIT_SUBREADS - MISMATCHES - NUM_EXTRACTED_SUBREADS - NUM_HIT_PAIR_SUBREADS - COMPLEX_INDELS			
Configuration:	50-5-3-3-10-1-off	0-10-1-20-5-1-on	50-5-1-30-5-1-off	50-5-1-30-15-1-off
Note:	Default	Best recall at base level	Best recall at read level	Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	28.07%	74.16%	74.22%	66.18%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	96.44%	77.76%	77.57%	97.52%
% reads aligned incorrectly:	1.03%	21.20%	21.46%	1.68%
% reads aligned ambiguously:	0.00%	0.00%	0.00%	0.00%
% reads unaligned:	70.90%	4.64%	4.32%	32.14%
% reads aligned:	29.10%	95.36%	95.68%	67.86%
% of reads with true introns:	4.31%	4.31%	4.31%	4.31%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	23.35%	58.73%	58.73%	57.99%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	94.84%	75.68%	75.79%	95.78%
% bases aligned incorrectly:	1.27%	18.87%	18.75%	2.55%
% bases aligned ambiguously:	0.00%	0.00%	0.00%	0.00%
% bases unaligned:	75.38%	22.40%	22.52%	39.46%
% bases aligned:	24.62%	77.60%	77.48%	60.54%
% of bases in true insertions:	0.41%	0.41%	0.41%	0.41%
% of bases in true deletions:	0.37%	0.37%	0.37%	0.37%
insertions FD rate [1 - PRECISION]:	22.47%	29.63%	22.33%	21.62%
insertions FN rate [1 - RECALL]:	84.39%	62.71%	64.28%	65.42%
deletions FD rate [1 - PRECISION]:	30.60%	49.35%	30.33%	30.20%
deletions FN rate [1 - RECALL]:	85.50%	63.22%	64.75%	64.96%
skipping FD rate [1 - PRECISION]:	99.85%	99.68%	99.65%	99.83%
skipping FN rate [1 - RECALL]:	84.96%	78.50%	78.58%	68.49%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	40.49%	25.44%	25.72%	38.43%
junctions FN rate [1 - RECALL]:	86.06%	79.97%	80.01%	70.95%
Junction Sides none	9018	6456	6552	17236
Junction Sides left	17	64	58	62
Junction Sides right	52	28	22	69
Junction Sides both	13360	19196	19158	27833
Junction Sides none	40.17%	25.07%	25.40%	38.13%
Junction Sides left	0.07%	0.24%	0.22%	0.13%
Junction Sides right	0.23%	0.10%	0.08%	0.15%
Junction Sides both	59.51%	74.56%	74.28%	61.57%

Supplementary Table 40 - Subread tweaking examples on Human T3R1 dataset

Command:	subjunc -i <index> -r <read file 1> -R <read file 2> -T 16 --allJunctions --SAMoutput -o <output alignment> -d <MIN_FRAGMENT_LENGTH> -I <INDEL> -m <NUM_HIT_SUBREADS> -M <MISMATCHES> -n <NUM_EXTRACTED_SUBREADS> -p <NUM_HIT_PAIR_SUBREADS> --complexIndels			
Parameters:	MIN_FRAGMENT_LENGTH - INDEL - NUM_HIT_SUBREADS - MISMATCHES - NUM_EXTRACTED_SUBREADS - NUM_HIT_PAIR_SUBREADS - COMPLEX_INDELS			
Configuration:	50-5-3-3-10-1-off	0-10-1-20-15-1-off	0-10-1-20-5-1-on	0-10-3-3-10-3-on
Note:	Default	Best recall at base and junction level	Best recall at read level	
---------- READ LEVEL ----------				
total number of reads:	2000000	2000000	2000000	2000000
% reads aligned correctly [RECALL]:	48.83%	74.19%	75.04%	48.79%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	97.22%	97.39%	75.47%	96.79%
% reads aligned incorrectly:	1.39%	1.98%	24.38%	1.61%
% reads aligned ambiguously:	0.00%	0.00%	0.00%	0.00%
% reads unaligned:	49.78%	23.83%	0.58%	49.60%
% reads aligned:	50.22%	76.17%	99.42%	50.40%
% of reads with true introns:	13.47%	13.46%	13.46%	13.47%
---------- BASE LEVEL ----------				
total number of bases of reads:	200000000	200000000	200000000	200000000
% bases aligned correctly [RECALL]:	43.62%	66.85%	64.08%	43.54%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	96.64%	96.64%	76.79%	96.18%
% bases aligned incorrectly:	1.51%	2.32%	19.36%	1.72%
% bases aligned ambiguously:	0.00%	0.00%	0.00%	0.00%
% bases unaligned:	54.87%	30.83%	16.56%	54.74%
% bases aligned:	45.13%	69.17%	83.44%	45.26%
% of bases in true insertions:	0.18%	0.18%	0.18%	0.18%
% of bases in true deletions:	0.17%	0.17%	0.17%	0.17%
insertions FD rate [1 - PRECISION]:	25.42%	24.32%	26.73%	27.10%
insertions FN rate [1 - RECALL]:	90.10%	81.13%	87.91%	90.51%
deletions FD rate [1 - PRECISION]:	27.28%	26.49%	29.42%	29.70%
deletions FN rate [1 - RECALL]:	90.03%	79.82%	87.31%	90.42%
skipping FD rate [1 - PRECISION]:	24.85%	18.01%	16.57%	9.06%
skipping FN rate [1 - RECALL]:	92.92%	86.75%	96.24%	95.96%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	14.12%	14.53%	13.82%	9.73%
junctions FN rate [1 - RECALL]:	93.61%	88.30%	96.38%	96.30%
Junction Sides none	2725	5107	1465	1073
Junction Sides left	172	345	121	50
Junction Sides right	184	378	116	48
Junction Sides both	18746	34307	10614	10872
Junction Sides none	12.48%	12.72%	11.89%	8.90%
Junction Sides left	0.78%	0.85%	0.98%	0.41%
Junction Sides right	0.84%	0.94%	0.94%	0.39%
Junction Sides both	85.88%	85.47%	86.18%	90.27%

Supplementary Table 41 - Tophat2 tweaking parameters and values

Parameters	Tested values	Note
`--b2-very-sensitive`	With and without this option	Default = without this option
`--coverage-search`	With and without this option	Default = without this option
`--read-mismatches <NUM_MISMATCHES>`	3; 4; 5; 6; 7; 8; 9; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35	Default = 2
`--read-gap-length <NUM_GAP_LENGTH>`	3; 4; 5; 6; 7; 8; 9; 10; 16; 25; 26; 35	Default = 2
`--read-edit-dist <NUM_EDIT_DIST>`	5; 7; 10; 16; 25; 26; 35	Default = 2
`--read-realign-edit-dist <NUM_REALIGN_EDIT_DIST>`	0; autoset-default	Default = value such that the tool will not try to realign reads already mapped in earlier steps.
`--max-insertion-length <NUM_INSERTION_LENGTH>`	4; 5; 9; 10; 16; 24	Default = 3
`--max-deletion-length <NUM_DELETION_LENGTH>`	4; 5; 9; 10; 16; 24	Default = 3
`--max-multihits <NUM_MULTIHITS>`	100	Default = 20

Supplementary Table 42 - Tophat2 tweaking examples on Malaria T3R1 dataset

Command:	tophat2 --output-dir <output path> --num-threads 16 --mate-inner-dist <INNER_MATE_MEAN> --mate-std-dev <INNER_MATE_SD> --b2-very-sensitive --GTF <gtf file> --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> <index> <reads file 1> <reads file 2>			
Parameters:	COVERAGE_SEARCH - B2_VERY_SENSITIVE - NUM_MISMATCHES - NUM_GAP_LENGTH -NUM_EDIT_DIST - NUM_REALIGN_EDIT_DIST - NUM_INSERTION_LENGTH - NUM_DELETION_LENGTH - NUM_MULTIHITS			
Configuration:	off-on-2-2-2-default-3-3-20	off-on-18-25-25-default-24-24-100	off-on-5-5-5-default-5-5-100	off-on-27-35-35-default-24-24-100
Note:	Default	Best recall at base and read level		Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	20000000	20000000	20000000	20000000
% reads aligned correctly [RECALL]:	2.08%	88.00%	18.74%	87.78%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	100.00%	99.80%	100.00%	99.13%
% reads aligned incorrectly:	0.00%	0.17%	0.00%	0.76%
% reads aligned ambiguously:	0.08%	1.85%	0.73%	1.92%
% reads unaligned:	97.84%	9.98%	80.53%	9.54%
% reads aligned:	2.16%	90.02%	19.47%	90.46%
% of reads with true introns:	4.31%	4.31%	4.31%	4.31%
---------- BASE LEVEL ----------				
total number of bases of reads:	2000000000	2000000000	2000000000	2000000000
% bases aligned correctly [RECALL]:	2.07%	75.98%	18.60%	75.79%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	100.00%	86.58%	99.17%	86.00%
% bases aligned incorrectly:	0.00%	11.77%	0.15%	12.33%
% bases aligned ambiguously:	0.08%	1.85%	0.73%	1.92%
% bases unaligned:	97.85%	10.40%	80.52%	9.96%
% bases aligned:	2.15%	89.60%	19.48%	90.04%
% of bases in true insertions:	0.41%	0.41%	0.41%	0.41%
% of bases in true deletions:	0.37%	0.37%	0.37%	0.37%
insertions FD rate [1 - PRECISION]:	2.34%	76.00%	14.41%	76.26%
insertions FN rate [1 - RECALL]:	99.78%	34.87%	94.02%	35.08%
deletions FD rate [1 - PRECISION]:	1.12%	30.24%	6.03%	30.42%
deletions FN rate [1 - RECALL]:	99.78%	38.18%	94.29%	38.34%
skipping FD rate [1 - PRECISION]:	89.38%	4.24%	83.97%	5.00%
skipping FN rate [1 - RECALL]:	97.94%	12.23%	81.47%	12.24%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	1.44%	0.73%	0.77%	0.98%
junctions FN rate [1 - RECALL]:	97.87%	11.67%	81.12%	11.66%
Junction Sides none	298	6119	1388	8263
Junction Sides left	0	46	1	46
Junction Sides right	0	44	1	45
Junction Sides both	20474	847642	181231	847733
Junction Sides none	1.43%	0.71%	0.76%	0.96%

Junction Sides left	0.00%	0.00%	0.00%	0.00%
Junction Sides right	0.00%	0.00%	0.00%	0.00%
Junction Sides both	98.56%	99.27%	99.23%	99.02%

Supplementary Table 43 - Tophat2 tweaking examples on Human T3R1 dataset

Command:	tophat2 --output-dir <output path> --num-threads 16 --mate-inner-dist <INNER_MATE_MEAN> --mate-std-dev <INNER_MATE_SD> --b2-very-sensitive --GTF <gtf file> --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> <index> <reads file 1> <reads file 2>			
Parameters:	COVERAGE_SEARCH - B2_VERY_SENSITIVE - NUM_MISMATCHES - NUM_GAP_LENGTH -NUM_EDIT_DIST - NUM_REALIGN_EDIT_DIST - NUM_INSERTION_LENGTH - NUM_DELETION_LENGTH - NUM_MULTIHITS			
Configuration:	off-on-2-2-2-default-3-3-20	off-on-18-25-25-default-24-24-100	off-on-7-6-7-default-4-4-100	off-on-25-25-25-default-24-24-100
Note:	Default	Best recall at base and read level		Best recall at junction level
---------- READ LEVEL ----------				
total number of reads:	20000000	20000000	20000000	20000000
% reads aligned correctly [RECALL]:	12.53%	83.81%	55.99%	83.80%
% reads aligned correctly (over uniquely aligned reads) [PRECISION]:	99.57%	92.55%	99.04%	92.50%
% reads aligned incorrectly:	0.05%	6.74%	0.54%	6.79%
% reads aligned ambiguously:	0.70%	4.10%	2.34%	4.13%
% reads unaligned:	86.72%	5.35%	41.13%	5.28%
% reads aligned:	13.28%	94.65%	58.87%	94.72%
% of reads with true introns:	13.46%	13.46%	13.46%	13.46%
---------- BASE LEVEL ----------				
total number of bases of reads:	2000000000	2000000000	2000000000	2000000000
% bases aligned correctly [RECALL]:	12.53%	73.09%	53.24%	73.08%
% bases aligned correctly (over uniquely aligned bases) [PRECISION]:	99.53%	81.10%	94.22%	81.06%
% bases aligned incorrectly:	0.05%	17.02%	3.26%	17.07%
% bases aligned ambiguously:	0.70%	4.10%	2.34%	4.13%
% bases unaligned:	86.72%	5.79%	41.16%	5.72%
% bases aligned:	13.28%	94.21%	58.84%	94.28%
% of bases in true insertions:	0.18%	0.18%	0.18%	0.18%
% of bases in true deletions:	0.16%	0.16%	0.16%	0.16%
insertions FD rate [1 - PRECISION]:	18.34%	87.60%	79.28%	87.60%
insertions FN rate [1 - RECALL]:	99.77%	36.56%	83.33%	36.57%
deletions FD rate [1 - PRECISION]:	6.76%	46.08%	43.90%	46.07%
deletions FN rate [1 - RECALL]:	99.78%	38.13%	83.78%	38.13%
skipping FD rate [1 - PRECISION]:	1.72%	7.28%	1.66%	7.28%
skipping FN rate [1 - RECALL]:	98.17%	28.64%	69.81%	28.64%
---------- JUNC LEVEL ----------				
junctions FD rate [1 - PRECISION]:	1.17%	7.36%	1.52%	7.37%
junctions FN rate [1 - RECALL]:	98.05%	23.69%	67.04%	23.68%
Junction Sides none	558	100360	9766	100478
Junction Sides left	69	39798	2545	39825
Junction Sides right	47	37393	2536	37428
Junction Sides both	57351	2235240	965592	2235264
Junction Sides none	0.96%	4.15%	0.99%	4.16%

Junction Sides left	0.11%	1.64%	0.25%	1.65%
Junction Sides right	0.08%	1.54%	0.25%	1.55%
Junction Sides both	98.83%	92.64%	98.48%	92.63%