Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads

Grzegorz M Boratyn, Jean Thierry-Mieg*, Danielle Thierry-Mieg, Ben Busby and Thomas L Madden*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD, 20894, USA

* To whom correspondence should be addressed. Tel: 301-435-5987; Fax: 301-480-4023; Email: madden@ncbi.nlm.nih.gov. Correspondence may also be addressed to mieg@ncbi.nlm.nih.gov.

ABSTRACT

Next-generation sequencing technologies can produce tens of millions of reads, often paired-end, from transcripts or genomes. But few programs can align RNA on the genome and accurately discover introns, especially with long reads. To address these issues, we introduce Magic-BLAST, a new aligner based on ideas from the Magic pipeline. It uses innovative techniques that include the optimization of a spliced alignment score and selective masking during seed selection. We evaluate the performance of Magic-BLAST to accurately map short or long sequences and its ability to discover introns on real RNA-seq data sets from PacBio, Roche and Illumina runs, and on six benchmarks, and compare it to other popular aligners. Additionally, we look at alignments of human idealized RefSeq mRNA sequences perfectly matching the genome. We show that Magic-BLAST is the best at intron discovery over a wide range of conditions. It is versatile and robust to high levels of mismatches or extreme base composition and works well with very long reads. It is reasonably fast. It can align reads to a BLAST database or a FASTA file. It can accept a FASTQ file as input or automatically retrieve an accession from the SRA repository at the NCBI.

INTRODUCTION

RNA-seq and DNA-seq experiments generate tens of millions of reads sampled from transcripts or genomes. The resulting data allows investigations that include, but are not limited to, gene expression, gene structure, nucleotide and structural variations. It is possible to use different analysis approaches for some investigations. For example, coarse gene expression can be studied with alignments or with alignment free methods (1). On the other hand, the investigation of fine grain gene structure does require alignments or sequence assembly. In this case, an analysis may benefit from a specific sequencing technology. While short reads (Illumina style) may be sufficient for coarse gene expression profiling and most introns or single-nucleotide variant (SNV) calling, other studies such as SNV phasing, full length transcript description or structural genomic rearrangements are facilitated by longer reads. But not all aligners can handle longer reads. Towards comprehensive and reliable SNV discovery, the aligner should provide good tolerance to mismatches and robust discrimination of reads mapping ambiguously at multiple quasi-repeated sites. Several groups have written fast aligners (2,3), but a recent benchmark by Baruzzo and the Grant team showed that new aligners could still offer value to the community (4).

Here, we present Magic-BLAST. We compare Magic-BLAST to three other popular aligners (listed below). We look at a variety of read lengths from Illumina, Roche 454 and PacBio and find that most aligners can accurately map short Illumina-style human paired reads with few mismatches. We find that aligners do not work as well with longer reads. We also find that higher numbers of mismatches as well as compositionally biased genomes pose problems for some aligners. Magic-BLAST can handle the different sequencing technologies, error rates, and compositional bias without special tuning, and this should allow, in particular, to map an RNA-seq experiment to the genome of a related species when a good quality reference genome is not available. Additionally, Magic-BLAST outperforms all other aligners in identification of introns, including discovery of unannotated introns, under all circumstances, despite the fact that it is single pass.

We chose the name Magic-BLAST to emphasize the ideas and software that went into the tool. Magic-BLAST derives its core algorithms from the Magic pipeline, described in detail in the supplementary material of (5), and is implemented using the same C++ framework as the BLAST+ executables (6,7,8). The merger of these tools results in a versatile and robust aligner. Magic-BLAST implements the MAGIC ideas of checking for overrepresented target sequence fragments during seed selection as well as an innovative greedy alignment extension procedure. Magic-BLAST can produce spliced alignments when aligning RNA on genome and selects the best-matching position genome wide by optimizing a spliced alignment score for single and for paired reads.

Magic-BLAST takes advantage of the existing BLAST infrastructure. It aligns reads against a BLAST database, which is the same database used by other BLAST programs. It can also align reads against a FASTA file or even just an accession, with the actual sequences automatically retrieved from the NCBI. Sequence reads can be provided as SRA accessions or as files in FASTA or FASTQ format. Magic-BLAST can transparently gzip or gunzip the sequence reads and/or the reference FASTA or FASTQ files. Magic-BLAST was field-tested in several NCBI hackathons that provided substantial feedback on features and usability.

We compare Magic-BLAST to three popular aligners, HISAT2 (9), STAR (10,11), and TopHat2 (12), also evaluated in (1) and (4). We look at precision (percentage of results that are correct), recall (percentage of the total true positives returned) and F-score (harmonic mean of recall and precision) for alignments and intron discovery, as measured on truth-bearing RefSeq (13) human transcripts (modified to exactly match the genome), on experimental long or short reads, and on simulated benchmark data assessing the impact on the alignments of variable levels of polymorphisms and errors, up to mimicking an alignment to a related species (4). The last data set additionally tests the aligners on a genome with extremely biased base composition, using the malaria agent *Plasmodium falciparum*, which is 80.7% AT. Magic-BLAST is not the fastest tool, but is reasonably fast and works as well when mapping DNA or RNA to the genome or to the transcriptome. It auto-adapts to the type of data and read length without requiring from the user any choice of parameters. It is versatile, easy to use, robustly precise and conservative in all circumstances.

MATERIAL AND METHODS

Algorithm Overview

The Magic-BLAST algorithm has a structure similar to that of other BLAST programs (8). It reads a batch of RNA-seq or DNA-seq reads and builds a "lookup table", which is an index of word locations (16-bases by default) in the reads. It then scans the database sequences (usually a reference genome) for matches in the lookup table and attempts to extend selected initial matches to the length specified by the user (18 by default). The resulting matches form a seed for computation of local gapped alignments. Collinear local alignments are combined into spliced alignments. Exons shorter than the seed length cannot be captured, but they are rare (less than 0.2% of RefSeq introns), and most will be recognized by aligning in parallel on the known transcriptome. For paired reads, the best alignments are selected based on the alignment quality of the pair. For example, if one read of a pair maps equally well at two genomic sites, and the second read maps best at a single site, the read pair will be reported as mapping uniquely at the position dictated by the second read. In this way, the specificity of the mapping truly reflects the number of bases sequenced in the whole fragment, i.e. 200 bases specificity for 100+100 paired end reads. Below, we present a detailed description of the above steps.

Repeat filtering

Most genomes contain interspersed repeats and gene families that complicate correct placement of reads in a genome. To avoid seeding to ambiguous positions, Magic-BLAST scans the reference sequences and counts 16-base words. Those words that appear in the reference database more than a user-specified number of times (by default 60) are not indexed in the lookup table, so that they never form a seed alignment. To make this procedure more efficient, only words present in the reads are counted. The cut-off number 60 was selected experimentally as the best trade-off between sensitivity and runtime. Additionally, Magic-BLAST specifically masks out 16-base words that contain at least 15 A's or 15 T's, effectively avoiding seeding on polyA tails. This approach is similar to soft masking in other BLAST programs.

Local gapped alignment

Magic-BLAST computes a local alignment by extending exact word matches (18-bases by default) between a read and a reference sequence. We use a simplified greedy alignment extension procedure, previously used in Magic (5). Starting with the seed, the alignment is extended until the first mismatch. Next, we attempt to characterize the mismatch as a substitution, insertion or deletion of one to three bases by recursively testing the quality of the alignment of the following few bases. This is done by applying successively a table of candidate alignment operations until the associated requirement is met. A requirement is that a specific number of bases must match within a given number of bases following the applied operation. The first operation whose requirement is met is applied to the alignment and the algorithm proceeds to the next position on both sequences. A single substitution is reported if no requirement is satisfied. The list of alignment operations and their associated conditions used in Magic-BLAST is presented in Table 1.

Figure1 shows an example alignment extension. First, there are two matches and the algorithm moves to the right by two positions on both sequences. When a mismatch (T-G) is encountered the algorithm tries successively each alignment operation and checks its requirements. The first operation, a substitution which requires nine matching bases following the mismatch, fails. The second operation, an insertion which requires ten consecutive matches, succeeds and is applied to the alignment. In the last step there is a match (G-G).

We use the X-drop algorithm (6) to stop the extension. At each position, we record the running alignment score. The algorithm stops at the end of a sequence or when the current score is smaller than the best score found so far by more than the most penalized gapped alignment operation (three-base gap in Table 1). The algorithm then backtracks to the position with the best score.

Because most reads align to a reference with few or no mismatches, this method is faster and more memory efficient than the dynamic programming-based extension procedure used in other BLAST programs. Moreover, this approach facilitates collection of traceback information at little additional cost. This method can be tuned to a given sequencing technology for an expected rate of mismatches or gaps simply by adapting Table 1. For example, in Roche 454 or PacBio, where insertions and deletions are more frequent than substitutions, one could switch to a modified table.

We compute an alignment score using the following system: 1 for each matching pair of bases, -4 for a base substitution, zero for gap opening (either a read or reference sequence), and -4 for each base of gap extension (insertion or deletion). A user can modify the mismatch and gap penalties. The quality coefficients present in the FASTQ file , but have no impact on the alignment score.

About half the time, a matching base can be placed on either side of a gap, so the gap can slide at equal score. To avoid difficulties in SNP detection, Magic-BLAST shifts the sliding bases upstream of the gap, in the orientation of the target.

Spliced alignments

Spliced alignments are found by combining collinear local alignments on a read and a reference sequence. Magic-BLAST constructs a chain of local alignments that maximizes the combined alignment score. It then updates the alignment extents so that the spliced alignment is continuous on the read and the intron donor and acceptor sites are, whenever possible, consistent with the splice signals.

If two local alignments are continuous on a read (possibly with an overlap), then we first search for the canonical splice site (GT-AG or CT-AC) where the alignments meet. If this site is not found and each alignment has a score of at least 50, we search successively for the minor splice sites or their complement: GC-AG or CT-GC, AT-AC or GT-AT, then for any other site. The first site found is accepted. The alignment score threshold of 50 was chosen because real non-canonical splice sites are rare, but pairs of di-nucleotides are frequently found in the genome. As a result, for reads shorter than 100 bases, Magic-BLAST conservatively only calls GT-AG introns.

Magic-BLAST also attempts to produce spliced alignments if a read has several local alignments separated by one to ten unaligned bases. First, we look for a splice signal within four bases of the end of the left alignment and, if found, we fix the beginning of the candidate intron. Second, we search for the corresponding end of intron signal at offsets that ensure a continuous alignment on the read allowing at most one additional insertion or deletion. If this fails, the procedure is repeated with the end of the intron fixed and a search for the signal indicating the start of the intron. When the candidate splice signals are found, the alignments are trimmed or extended to the splice signals. The benefit of this method is to correctly identify introns even in the presence of a substitution or insertion or deletion close to the intron boundaries. Because this procedure is very sensitive and can produce many spurious alignments, Magic-BLAST only allows the GT-AG signal in this situation.

The spliced alignment is scored with the same scoring system as the local alignment. There is no reward or penalty for splice sites and no preference is given to continuous versus spliced alignments. When mapping RNA to the genome, Magic-BLAST does not use an annotation file or a two-pass method. We recommend instead to map in parallel on the genome and on the annotated transcriptome, then use the universal scoring system of Magic-BLAST to select the best alignment for each fragment.

Output

Magic-BLAST returns results in the Sequence Alignment/Map SAM/BAM format (14) or in a tab-delimited format similar to the tabular format in other BLAST programs, which is less standard but richer and easier to mine.

RESULTS

Datasets and programs

The ability of Magic-BLAST and other popular programs to map RNA-seq to genomes in a naïve fashion, without knowledge of an annotated transcriptome, and to find introns and their precise splice sites was assessed using seven truth-bearing datasets, one of which is new, and three experimental runs, from Illumina, Roche 454 and PacBio.

The new benchmark, called iRefSeq, is the image of the Human RefSeq (13) mRNAs exactly matching the genome. We selected the protein-coding RefSeq mRNAs, limiting to the 45,108 NM accessions that map to the main chromosomes and mitochondrial DNA of GRCh38. These mRNAs are transcribed from 19,269 protein coding genes. Using the mapping coordinates, as given in the RefSeq annotation, we assembled genomic sequences into transcript sequences so that they exactly match the genome (see https://github.com/ncbi/magicblast/tree/master/article for more information, including the scripts allowing to duplicate our analyses). iRefSeq mRNAs range in length from 147 bases to 109,224 bases. This perfect data set forms a useful benchmark for RNA-seq aligners, because it seems simple to align and each mismatch in the alignment indicates an imperfect mapping. Furthermore, the coordinates of the 210,509 distinct introns in iRefSeqs are known.

The benchmark set of RNA-seq reads, presented in (4), was also used. This set has some qualities that make it appealing for our analysis. The authors document their procedure well, they produce 10 million simulated (paired-end 100+100) Illumina-like reads at three vastly different error rates, nominally from 6.1 to 55 mismatches per kb, and they produce data for human and *Plasmodium falciparum*, a protozoan causing malaria in human (we refer to the latter sets as 'malaria'). Baruzzo name these three different error rates T1, T2, and T3. The variable error rates allow an analysis of how the aligners perform if the genome of the same species is available (T1), if only a poor-quality version of the genome or of the sequence data is available (T2), and if only the genome of a related organism is available (T3). The malaria sets allow an analysis of how the aligners perform under extreme genome base composition as the genome is 80.7% AT. The human and malaria sets have the same number of reads, so the malaria sets are at least 100 times deeper, a confounding effect to unravel. In practice, each set is provided as triplicate benchmark runs, but since the results are very similar, only the results of R1 are shown in the figures. Surprisingly, we noticed that the measured level of mismatches per kb actually differs between the human and malaria sets: T1 has 5.4 and 6.5 mismatches per kilobase aligned in human and malaria respectively, T2 11.86 and 16.6, and T3 60.2 and 86.5.

Furthermore, we selected three experimental RNA-seq data sets from the public Sequence Read Archive (SRA at NCBI) to represent three sequencing platforms, PacBio SRR5009494, with 8285 long reads (average 1854 bases) sequenced from colon carcinoma cells, Roche 454 Titanium SRR899420 with 416,045 reads (average 348 bases) sequenced from the MAQC/SEQC brain mRNA sample (15), and a deep Illumina HiSeq SRR534301 (also tested by (9)) with close to 109 million 101+101 bases paired-end reads sequenced from fetal lung. Here, we will refer to each RNA-seq set by its technology: PacBio, Roche and Illumina. Figure 2 presents a histogram of read lengths for the iRefSeq and experimental sets.

We examined the performance of several programs aligning RNA to the genome in the absence of a transcriptome (Human genome GRCh38 and *P. falciparum* genome provided in (4)). Magic-BLAST was compared to programs from 2013 to 2015: HISAT2 (9), STAR run in both one-pass and two-pass modes (10,11), and TopHat2 (12). The regular STAR is optimized for Illumina-type reads and cannot handle long reads. Hence, we ran STAR long, a version of STAR recommended for reads longer than 300 bases (10,1,16), in 2-pass mode. HISAT2 by default is two-pass; it was run with default parameters as well as a 'relaxed' mode which is more sensitive but less specific and slower: the HISAT2 default parameters left 4,663 iRefSeq unmapped while all aligned in relaxed mode. Magic-BLAST and TopHat2 do not have a two-pass mode and were run with default parameters.

Two analysis programs were used in this project: AliQC.py was developed in collaboration with Joe Meehan from the FDA for the SEQC project (15). It extracts, by comparing the SAM file to the genome, a detailed quality control on alignments, their length and multiplicity, and counts the mismatches by type and by position along the read (i.e. by sequencing cycle),. The number of mismatches were confirmed using the NM: optional tag present in the BAM files. Another program sam2gold.c was written in C to compare the SAM files to the specific format in which Baruzzo (4)

provided the benchmark truth. A master-script, deposited in GitHub, points to the input datasets and allows to download all the data, to realign the sequences and reproduce all the results.

We first measure how well the different aligners identify introns, then we examine the properties of the alignments.

Intron Discovery

To test how well the different aligners discover introns, the splice sites were extracted from the BAM output using the 'N' operation in the CIGAR string (14). For the iRefSeq and Baruzzo benchmark sets, the true position of each intron is known. The three experimental sets do not come with such a "Ground Truth", but a proxy for true and false positives are the introns annotated and not-annotated in iRefSeq. Of course, this is not strictly correct as some unannotated introns are no doubt real and just have not been discovered or annotated on the genome. On the other hand, it seems likely that all (or almost all) the annotated introns are real. This strategy allows a comparison of the results of all programs on all datasets and the measurement of precision, recall and F-score for intron discovery.

Magic-BLAST, HISAT2 in relaxed or standard mode, and STAR long are able to align very long reads on the genome and find introns. TopHat2 and the regular STAR failed to produce any results for very long reads, although TopHat marginally worked for Roche 454.

We first use a ROC curve approach (17) to precisely judge the quality of intron discovery, true versus false, as a function of minimal read coverage. For introns covered by at least 50 reads, 49 reads… to at least 1 read, (or 100, 99, … 1 for the ultra-deep runs Illumina and the malaria sets) the number of true positive (or annotated) introns are plotted on the Y-axis while the number of false positives (or unannotated) introns are on the X-axis. The resulting 50 or 100-points curves give us visual insight into how the different programs behave when the support, given by the number of reads mapped to each junction, decreases. The ROC curves are shown in Figure 3a-j and include the Truth for the seven benchmark sets (blue, on the Y axis). The best curve, of course, would have all the true positives before any false positives, meaning the steeper the slope, and ultimately the higher to the left the curve is, the better. In all ten cases, Magic-BLAST (red) is to the left and above all other curves and qualifies as the best intron finder in all conditions tested, from short to very long reads, with any level of mismatches, from perfect match to 8.6% mismatch, and this observation applies to benchmark as well as all real data sets tested from PacBio, Roche or Illumina.

Figure 4 summarizes this observation by quantifying intron precision, recall, and F-score for the iRefSeq and experimental sets. For the experimental sets where we use RefSeq annotated introns as truth, only the comparison of the scores of the various aligners is meaningful: on the same data, they should detect the same introns. But the precision and the recall depend on the tissue and on the depth of the experiment. The Illumina run is from fetal lung, a stage not well represented in the RefSeq collection, and this explains the apparently low precision (below 46%); many of the observed introns are probably real but not annotated. At the same time, RefSeq annotates introns and genes

from all tissues, and typically a sample derived from a single tissue and sequenced deeply will express 70 to 75% of all annotated introns, this explains the recall below 72% in this Illumina set. For the shallow data sets from Roche adult Brain and PacBio colon carcinoma cells, precision seems good (up to 92% in Magic-BLAST) because these tissues were used intensely in RefSeq annotation, and because at low coverage, one sees mainly the highly expressed genes, which are the best annotated. Yet the low depth of sequencing explains why Roche finds only 31% of the annotated iRefSeq introns, and the very small PacBio run find less than 6%.

The ROC curves make it apparent that the ability of the aligners to discover introns, with a good balance of true to false positives, changes as the read coverage for introns decreases. We use this insight to calculate a coverage dependent best F-score for the aligners. At the best F-score, Magic-BLAST has the highest F-score in all 10 experiments. It also reaches its best score at the lowest coverage of all aligners in almost all cases. In particular, the other aligners achieve optimal scores at much higher coverage than Magic-BLAST for the deep experimental Illumina and Baruzzo malaria sets. Magic-BLAST is more conservative than the other programs, and even the introns supported by a single read appear reasonably trustworthy.

Another notable feature in the ROC curves for the Baruzzo benchmark is that STAR produces the largest number of false positives in every case, followed by HISAT2 (Figure 3). In the deep malaria set, which has about 5500 annotated introns, STAR produces up to 600,000 false positive introns (Figure 3j). This greedy intron-finding behaviour fits with the observation that introns and splice sites found by STAR and HISAT2 at low coverage are mainly untrustworthy. If using these programs in an RNA-seq analysis, a good practise is to decide which coverage threshold gives the best ratio of novel to known introns (the best F-score), then to filter all new introns below this threshold.

It is worth noting that STAR 2-pass can produce worse results, as judged by pseudo-ROC curves, than the 1-pass version. This is especially apparent for deep sets such as the experimental Illumina and the Baruzzo malaria.

Alignment quality

Various metrics can be applied to characterize the ability of aligners to accurately map the reads. For simulated runs, where a truth is known by construction, one can compare the precise placement of each read by each program to the 'true' annotated position. Reads mapping uniquely can then be partitioned in one of four categories: completely correctly aligned (True positive type 1), partially but correctly aligned (True positive type 2), misaligned (False positive) or not aligned (False negative). In cases of multiple alignments, only the most favorable location is considered. This provides a direct measure of mapping precision, recall and F-score.

We also measure the detailed alignment statistics, independent of the truth: number of reads and bases aligned, unique versus multiple alignments, aligned length, mismatches and indels by type and position along the read (i.e. by sequencing cycle). These numbers are extracted from the BAM files using the AliQC.py program developed in collaboration with Joe Meehan from the FDA for the SEQC project (15). In cases of multiple alignments, only the 'primary' alignment is considered in this analysis. The number of mismatches were double-checked using the NM: optional tag present in the BAM files.

A simple and rich summary of the quality of the long read alignments is provided in Figure 5, which shows for each aligner the histogram of aligned length as compared to the length of the reads. Ideally the histogram of aligned length for each program should be superimposed on the curve for read lengths (blue). The closer the histogram of aligned length to the histogram of read length the better a program's performance. In the iRefSeq case, very long transcripts perfectly matching the genome (Figure 5a), TopHat2 fails, the curve for Magic-BLAST matches the read length histogram all the way from short to long transcripts. HISAT2 relaxed is better than HISAT2 and nearly as good as Magic-BLAST, but with and elbow of short alignments. STAR long is distinctly lower over the longest reads. The situation is different for the PacBio run which has a high rate of sequencing errors (Figure 5b). There, STAR long finds the longest alignments but maps less than 70% of the reads. Magic-BLAST finds more alignments than STAR long, but many are partial. The curves for HISAT2 and HISAT2 relaxed are much lower than the histogram of read length, especially for long reads. HISAT2 relaxed creates a very large number of alignments shorter than 200bp. The Roche reads are shorter (Figure 5c), and they are now in the range acceptable to TopHat2, which aligns about half of the reads up to 600bp. Alignment lengths for STAR long and Magic-BLAST are close to read lengths. The curve for HISAT2 is a little below the blue curve and HISAT2 relaxed generates a small number of partial alignments. Both Magic-BLAST and HISAT2 relaxed produce a small number of shorter alignments as is evident from the peaks on the very left end of the graph.

We now detail the results for the iRefSeq benchmark, then the six Baruzzo benchmarks, and finally the experimental runs from SRA.

The iRefSeq experiment evaluates the ability of the programs to align long spliced mRNA sequences exactly matching the genome, an idealized situation with no sequencing errors. Magic-BLAST, HISAT2, HISAT2_relaxed, and STAR long produced results for this experiment. There is no bias, as Magic-BLAST was not used to prepare the RefSeq annotation.

Figure 6 provides details on the iRefSeq experiment. Percentages aligned both at the base and sequence level are displayed in Figure 6a. Magic-BLAST performs the best. Measuring the mapping accuracy by comparison to the RefSeq annotation (the truth), displayed in Figure 6b, refines the picture: 97.2% of the 45,108 iRefSeq mRNAs are perfectly mapped over their entire length by Magic-

BLAST, 88.1% by HISAT2_relaxed, 84.1% by HISAT2, and 80.6% by STAR long (no clipped bases, substitutions or indels). There are also some correct but partial alignments. An important question is how often a program will misalign a read, since this would create noise in downstream analyses. This error happens 196 times (0.43%) with Magic-BLAST, but four to eight times more frequently in HISAT2 and STAR long. Most of the time, the misalignment is subtle, affecting just one or a few exons, but in 12 cases for Magic-BLAST, 103 for HISAT2_relaxed, 39 for HISAT2, and 100 for STAR long, the mRNA is wildly misaligned at a genomic site not overlapping the position annotated in RefSeq. Another functionally important case is the proper identification of the first and last exons, which reveal the location of promoters and 3' ends regulatory regions. In Magic-BLAST, five alignments overlap the truth but extend outside of the annotated gene, creating a new or incorrect first or last exon, but this problem happens much more frequently in HISAT2 (191 and 79 times) and in STAR long (58 times). In a gene reconstruction project, this type of defect is hard to fix and may have serious consequences, leading to incorrectly intertwined neighboring genes.

Similarly, at the intron level (Figure 6c), Magic-BLAST has more than 10 times fewer false positive introns than HISAT2 (158 versus 2336 or 2765) and four times fewer than STAR long. STAR long also misses the most introns overall by a wide margin. Magic-BLAST is superior to both HISAT2 variants and STAR long in all categories.

The iRefSeq mRNAs should match the genome exactly, and the plot of matches versus mismatches shows how well each aligner performs (Figure 6d). The perfect result (truth) is in the upper right corner of the plot. Magic-BLAST performs very well with only 771 mismatches. STAR long and HISAT2 are distant second, with 38018 and 66358 mismatches respectively. STAR long aligns fewer bases. HISAT2, with 87963 mismatches, is both less sensitive and less specific than HISAT2 relaxed, because HISAT2 with default parameters does not accept high levels of mismatches.

We also compared the ability of the programs to align simulated RNA-seq reads using benchmark data (4). Figure 7 presents a number of alignment statistics, including the percentage of bases and reads aligned as well as the alignment precision and recall. The human T1 set should be the easiest and all aligners do well in terms of bases aligned, though TopHat2 does noticeably worse than the others. For the human T2 set, STAR 2-pass does the best, and HISAT2 and TopHat2 show some degradation. Magic-BLAST maintains a strong performance at the T3 level, with STAR 1-pass and STAR 2-pass showing significant degradation, and HISAT2 and TopHat2 performing very poorly. The other statistics tell a similar story, with results degrading from T1 to T2 to T3.

The Plasmodium (malaria) benchmark sets should present a more challenging set owing to the biased composition of the genome, but the results mirror the human benchmark results. In terms of bases aligned (Figure 6), most programs do well at the T1 level though TopHat2 again is the worst. At the T2 level, STAR 1-pass and STAR 2-pass are the best. At the challenging T3 level, HISAT2 and TopHat2 produce almost no result, because these programs cannot cope with high levels of mismatches. Magic-BLAST maintains acceptable performance, and STAR 1-pass and STAR 2-pass

show significant degradation. The alignment precision and recall tell a similar story. All aligners do well at the T1 level with STAR the best, and Magic-BLAST, STAR 1-pass, and STAR 2-pass are able to maintain that performance at the T2 level. Magic-BLAST continues its strong performance at the T3 level.

Finally, we examine the ability of the programs to align real RNA-seq reads generated with different technologies. Figure 4 presents the statistics for the different aligners. For the PacBio set, Magic-BLAST aligns the largest percentage of bases followed by STAR long, with the two HISAT2 variants a distant third and fourth. For the Roche 454 set, Magic-BLAST again performs the best but the HISAT2 variants and STAR long come closer to matching the Magic-BLAST performance with TopHat2 aligning less than half the bases. For the Illumina reads, STAR 2-pass aligns the most bases, but Magic-BLAST aligns the most reads, and with the highest rate of unique (i.e. unambiguous) alignments.

Run times

Table 2 presents the CPU times and maximal RAM requirements for the alignment experiments. We ran the aligners both with standard (i.e., default or recommended) options as well as with some experimental options. We used standard options for HISAT2, STAR using 2-pass (since no transcriptome annotation was provided) and STAR long for reads longer than 300 bases. Magic-BLAST and TopHat2 were run with default options. Note that TopHat2 does not produce results on long reads. We also tested STAR in 1-pass mode. For long reads, we also tested a relaxed version of HISAT2 that is more sensitive but less specific, yet this option is 3 to 40 times slower than HISAT2. The standard HISAT2 is the least memory greedy (4 GB of RAM) and the fastest aligner on long reads and malaria T3. STAR 1-pass is the fastest on short reads (except for malaria T3). STAR 2-pass comes next (long or short), then Magic-BLAST. TopHat2 is the slowest. For HISAT2, as explained in (9), the CPU only depends on the amount of data to map and not on the size of the genome: it is as fast on human as it is on the 130 times smaller malaria genome. In contrast, STAR uses a large amount of memory, it is slower on the deep malaria set than on the shallow human in all benchmark runs, and it requires increasing amounts of CPU time as the number of mismatches increases: for example, malaria T3 consumed 16 times more CPU than malaria T1 and 10 times more than human T3. We observed that when STAR becomes slow, the quality of its results also decays. STAR 1-pass is 2 to 4 times faster than the 2-pass version, sometimes faster than HISAT2, but with the same features as STAR 2-pass. Magic-BLAST works for all datasets, produces better results on introns, long read alignments, and high-level mismatches, and is stable. It is three to ten times slower than HISAT2 and on average twice slower than STAR 2-pass, but it is twice faster than TopHat2

The timings were performed on a 2.8 GHz Intel Xeon X5660 processor with 49 GB of RAM with a CentOS7 LINUX operating system. The time was measured with the LINUX time command by

summing the reported user and system times. Before each run, the database and index files were cached in memory, to minimize influence of network and disk access on run time

DISCUSSION

We have examined the performance of four aligners with a wide variety of sequence data. First, we examined the performance of all programs with three experimental test sets with different lengths and characteristics. Second, we presented a new benchmark designed to test the ability of the programs to align very long sequences that have no mismatches to the genome. Finally, we looked at an artificial benchmark of Illumina type reads, for human as well as a 100-fold deeper malaria runs, with three levels of mismatches. The aligners have different strengths and weaknesses, which reflect in part the strategic choices of the authors (e.g. favouring complete alignments or limiting the number of mismatches per read) and the characteristics of the implementations (e.g. second pass intron validation).

On the real 101 bases Illumina runs, STAR 2-pass aligns the most bases. On human T1, and even T2, STAR-2 pass aligns the largest number of bases and reads and maintains an excellent alignment F-score. However, its performance drops relative to Magic-BLAST for T3, and Magic-BLAST maintains good performance for the T3 sets.

Only Magic-BLAST, HISAT2 and STAR long can align very long sequences, even if there are no mismatches as demonstrated by the iRefSeq set. Both Magic-BLAST and HISAT2 (with non-default parameters) were able to align all reads in the set, but Magic-BLAST had significantly fewer partial alignments. We found the same trends with the experimental sets. TopHat2, STAR 1-pass, and STAR 2-pass produced no usable results for the PacBio reads, and STAR 1-pass and STAR 2-pass produced no results for the Roche 454 reads. Magic-BLAST produced the best overall results and excelled with longer reads.

Intron discovery posed different challenges for the aligners, as any false positive alignments are magnified by the process of intron discovery. We looked at pseudo-ROC curves, using the provided results for the benchmark sets and the annotations on the human genome as a guide for the experimental sets. It is clear from the pseudo-ROC curves that the cautious Magic-BLAST approach to intron discovery pays dividends, with far fewer false positives produced for a given number of true positives. The intron discovery F-score tells a similar story. Magic-BLAST had the best results and its results are trustworthy, even at very low coverage. For the T1 human Illumina type reads with few mismatches, the difference between Magic-BLAST, HISAT2, STAR 1-pass, and STAR 2-pass was relatively small, but Magic-BLAST excelled for more distant matches or the compositionally biased malaria sets. For the T3 sets, Magic-BLAST had much better intron-finding F-scores than the other programs, consistent with its read mapping F-scores. For the iRefSeq and PacBio sequences, Magic-BLAST, HISAT2, HISAT2 relaxed, and STAR long produced results, with Magic-BLAST producing the

best results. As discussed, the representation of introns in the RefSeq annotation may be uneven. Certain tissue types may be underrepresented, while highly expressed genes from other tissues are certain to be included. This result points out the need to measure the performance of an aligner on real data, with all the messiness of biology, but also on benchmark data with known results.

We also found that the intron discovery ROC curves for STAR 2-pass were worse than STAR 1-pass for deep sets such as the Baruzzo malaria sets as well as the experimental Illumina sets, even though STAR 2-pass is supposed to improve upon STAR 1-pass. For the Illumina run, the 1-pass curve lies just below Magic-BLAST, but the 2-pass curve is strongly shifted towards the unannotated/false positive. HISAT2, which also uses a 2-pass technique, is twice further towards the noise. One could argue that most introns discovered by HISAT2 or STAR 2-pass in Illumina are real and missing from the annotation. However, our design contains seven controls out of 10 sets, where the true curve is vertical: in iRefSeq and in the Baruzzo sets, especially malaria, there is no doubt that the highly covered unannotated introns of STAR and HISAT2 are false positives, despite their high coverage. Both STAR 2-pass and HISAT2 perform better for the shallow T1 human set. The optimal F-scores, calculated for the different aligners and experiments, are consistent with the ROC-curves in this regard. It is likely that in both aligners, the second pass reinforces the false discoveries of the greedy first pass, and this tendency is also visible in PacBio and Roche. Sadly, this type of noise cannot be erased, and if used in a gene reconstruction project, as was done in (16), those well supported but false positive introns will generate alternative splice variants that do not exist in the biological sample and will durably pollute reference gene models.

No single aligner is always the fastest. STAR 1-pass is the fastest for the experimental Illumina reads as well as the T1 benchmarks but slows down dramatically for the T2 and T3 cases. For the malaria T3 case, STAR 1-pass and STAR 2-pass are the two slowest programs and exceedingly memory greedy. STAR long is the fastest for the longest reads such as PacBio and iRefSeq but produces much worse results (in terms of alignment accuracy and intron finding) than Magic-BLAST.

Magic-BLAST is the only aligner to produce reliable results for a wide variety of sequence lengths, compositions, or error rates without changes to the command-line or even the necessity to use a different version of the aligner. The other aligners require the user to know, in advance, details of their experiments and decide which options or version to use. We also field-tested Magic-BLAST in several hackathons, allowing us to identify problems, hear user suggestions, and improve usability

It is also instructive to examine how the standard BLASTN algorithm handles spliced alignments. We search an mRNA (u00001.1) against the human genome reference assembly (GRCh38.p12) with BLASTN and find two problems. First, BLASTN identifies (apparently) strong matches on chromosomes 2, 14, 17, 20, 21, and 22 as well as two unplaced genomic scaffolds. A quick examination of the BLASTN alignments shows that all the matches are processed pseudogenes, except for the one on chromosome 17. Second, BLASTN does correctly identify the genomic exons

on chromosome 17 but gives an imprecise result as shown in Figure 8. A spliced aligner, like Magic-BLAST, correctly finds that the first exon ends at 85 on the query, and the second exon starts at 86 on the query. Additionally, BLASTN has no facility for recognizing paired reads so that it aligns and scores each read independently.

BLAST toolkit integration

Magic-BLAST takes full advantage of the BLAST Toolkit. It uses a BLAST database for reference sequences thatcompresses sequences 4-to-1, so it does not take up much disk space or memory. The same database can also hold reference sequence metadata such as taxonomy, length, identifiers, and titles. The sequences and the metadata can be retrieved with the blastdbcmd executable (see https://www.ncbi.nlm.nih.gov/books/NBK279690/). Additionally, the databases can also hold user-supplied masking information for the reference sequences that can be selectively enabled. Similar to the BLAST+ programs, one can use the -seqid_list option to map sequences only to selected reference sequences in a larger BLAST database.

For bioinformatics workflows, a significant advantage of Magic-BLAST is the flexibility in obtaining the sequence reads and the ease of setting up the reference set. It can map both RNA and DNA sequencing runs to a reference genome or transcriptome. Reference sequences can be given as FASTA, BLAST database, or as a list of NCBI accessions. Searches with a BLAST database are the fastest. Magic-BLAST has been used in several rapidly prototyped workflows (https://github.com/NCBI-Hackathons). Examples provided on project web pages show how users can download NCBI reference genomes, create BLAST databases and map SRA experiments with Magic-BLAST. For instance, Magic-BLAST was used for the fast estimation of the expression of a selected transcript in an SRA run (https://github.com/NCBI-Hackathons/deSRA). This project compared the expression of the transcripts in two NGS sets by quickly creating a small BLAST database for selected transcripts. Magic-BLAST then searched the two NGS projects against the BLAST database.

Conclusion and next steps

We presented Magic-BLAST, a new tool for mapping next generation sequencing runs against a genome or a transcriptome. Its performance was compared with that of similar popular programs:HISAT2, STAR, and TopHat2. Magic-BLAST is the best intron finder with the highest precision and highest recall on all the tested sets, real or simulated. We showed that Magic-BLAST works with sequences of any length: short Illumina reads, long Roche 454, PacBio reads, or even full-length mRNA sequences.

Magic-BLAST integrates very well with other NCBI tools and services and is convenient to use since it recognizes NCBI accessions for SRA reads, mRNA or genomic sequences, and uses BLAST databases.

We are exploring ways to improve Magic-BLAST, such as improving the run time and adapter detection, shortening the required exon length, and identifying repeats. We are also working closely with users to address their needs.

AVAILABILITY

Magic-BLAST executables and source code are available at ftp://ftp.ncbi.nlm.nih.gov/blast/executables/magicblast/LATEST. Command line executables are provided for 64-bit Linux, Mac-OS and Windows systems. The package includes the makeblastdb program for creation of a BLAST database. The makeblastdb program is the same as distributed with the BLAST+ package and added here only for convenience. Basic operation instructions and examples are provided in the Magic-BLAST cookbook: https://ncbi.github.io/magicblast .

A master script used for the current analysis is available at https://github.com/ncbi/magicblast/tree/master/article. The reads and genome sequences for the benchmark sets were published in (4). For other experiments, the human genome assembly GRCh38 (GCF_000001405.36) and its RefSeq annotation were downloaded from NCBI Human Genome Resources page: https://www.ncbi.nlm.nih.gov/genome/guide/human. The experimental sequences are accessible by their SRR accession numbers from the NCBI SRA repository. The GRCh38 GFF file is available at:
ftp://ftp.ncbi.nlm.nih.gov/blast/demo/magicblast_article/Fasta/iRefSeq/GRCh38_genome.gff.gz

The alignments produced in our experiments are available at ftp://ftp.ncbi.nlm.nih.gov/blast/demo/magicblast_article/SAM.

ACKNOWLEDGEMENT

We would like to thank David Lipman for proposing and encouraging this project. We would also like to thank Richa Agarwala, Michael DiCuccio, Alex Morgulis, Terence Murphy, Eugene Yaschenko, Peter Cooper and Tao Tao for useful discussions, experiments, and feedback, and Joe Meehan for his contribution to the AliQC python program.

FUNDING

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine. Funding for open access charge: National Institutes of Health.

CONFLICT OF INTEREST

The authors declare no competing interests.

REFERENCES

1. Sahraeian, S.M.E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P.T., Au, K.F., Bani Asadi, N., Gerstein, M.B., Wong, W.H., Snyder, M.P. et al. (2017) Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat Commun, 8, 59
2. Mielczarek, M. and Szyda, J. (2016) Review of alignment and SNP calling algorithms for next-generation sequencing data. J Appl-Genet, 57, 71-79.
3. Reinert, K., Langmead, B., Weese, D. and Evers, D.J. (2015) Alignment of Next-Generation Sequencing Reads. Annu Rev Genomics Hum Genet, 16, 133-151.
4. Baruzzo, G., Hayer, K.E., Kim, E.J., Di Camillo, B., FitzGerald, G.A. and Grant, G.R. (2017) Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat Methods, 14, 135-139.
5. Zhang, W., Yu, Y., Hertwig, F., Thierry-Mieg, J., Thierry-Mieg, D., Wang, J., Furlanello, C., Devanarayan, V., Cheng, J., Deng, Y. et al. (2015) Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome Biol, 16, 133
6. Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) Basic local alignment search tool. J Mol Biol, 215, 403-410.
7. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25, 3389-3402.
8. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. BMC Bioinformatics, 10, 421.
9. Kim, D., Langmead, B. and Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods, 12, 357-360
10. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29, 15-21.
11. Dobin, A. and Gingeras, T.R. (2016) Optimizing RNA-Seq Mapping with STAR. Methods Mol Biol, 1415, 245-262
12. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol, 14, R36.
13. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y.,

Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res, 44(D1):D733-45

14. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Subgroup, G.P.D.P. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2078-2079.

15. SEQC/MAQC-III Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol, 32, 903-914.

16. Lagarde, J., Uszczynska-Ratajczak, B., Carbonell, S., PÈrez-Lluch, S., Abad, A., Davis, C., Gingeras, T.R., Frankish, A., Harrow, J., Guigo, R. et al. (2017) High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. Nat Genet, 49, 1731-1740.

17. Gribskov, M. and Robinson, N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput Chem, 20, 25-33.

.

TABLES

Table 1. List of alignment operations used in the Magic-BLAST alignment extension

Operation	Length	Required number of matching bases	Allowed number of mismatches
Substitution	1	9	0
Insertion	1	10	0
Deletion	1	10	0
Insertion	2	10	0
Deletion	2	10	0
Insertion	3	13	0
Deletion	3	13	0
Substitution	2	12	0
Insertion	1	10	2
Deletion	1	10	2
Insertion	2	10	2
Deletion	2	10	2
Insertion	3	13	2
Deletion	3	13	2

Table 2. CPU (user and system) time in minutes for the searches discussed. A blank cell indicates that search was not performed.

	Magic-BLAST	HISAT2	HISAT2_relaxed	STAR-1pass	STAR-2pass	STAR long	TopHat2
iRefSeq	72.9	2.0	85.3			16.6	
PacBio	5.2	0.6	30.2			4.8	
Roche 454	32.2	2.8	8.9			9.4	60.7
Illumina	2951.1	563.5		252.9	861		4978.3
Human T1	501.2	49.1		20.6	53.5		572.7
Human T2	497.2	51.6		27.5	70.6		658.9
Human T3	468.4	49.0		45.6	116.5		846.8
Malaria T1	139.6	44.7		29.7	83.0		163.2
Malaria T2	138.1	46.2		95.7	285.3		232.2
Malaria T3	107.9	40.6		643.5	2000		370.3

FIGURE LEGENDS

Figure 1. An example alignment extension procedure. The arrows point at each step to the position currently considered in both sequences. First, there are two matches and the arrows move to the right by two positions on both sequences. When a mismatch (T-G) is encountered the algorithm tries alignment operations and conditions. The first operation: a substitution, which requires nine matching bases following the mismatch, fails. The second operation, an insertion which requires ten matches succeeds and is applied to the alignment. In the last step there is a match (G-G).

Figure 2: Actual read length distributions for the three experimental sets as well as iRefSeq. All Illumina reads have the same length (101+101 bases, paired end) but other sets have non-trivial length distributions. Newer sequencing technologies tend to produce longer sequences. PacBio has the longest reads of the experimental sets, ranging from 710 to 2703 bases (average 1854 bases). Roche 454 Titanium reads span from 33 to 808 bases, with an average of 348 bases. iRefSeq has the longest reads of all the sets: these full-length mRNAs range in length from 147 bases to 109,224 bases, with an average of 3427 bases. 9900 are longer than 10 kb. The scale in y is reduced 100

times for Roche 454, and 50,000 times for Illumina, which has the highest throughput of all technologies.

Figure 3: ROC curves showing intron discovery as a function of minimal read coverage, from 50 or 100 to 1. The plot shows for each minimal coverage, the true positive on the y axis and the false positive on the x axis. In the three experimental sets, annotated and unannotated introns are used as a proxy for true and false positives. The best curve would have all the true positives before any false positives, meaning the steeper the slope the better. The seven benchmark sets, iRefSeq and six Baruzzo, have a built-in truth. a) For the iRefSeq set, because of the alternative splice variants, the truth has introns supported by 1, 2, … 51 RefSeqs, and Magic-BLAST (red) follows the truth remarkably closely. It finds slightly more true positive introns than the HISAT2 programs, but the biggest difference is that HISAT2 find fifteen to seventeen times as many false positive introns. STAR long finds only 60% of the introns with some false positives. b) For PacBio, with less than 9000 reads, Magic-BLAST already finds 11464 annotated introns, many more than the two HISAT2 versions and STAR long, yet fewer unannotated introns. c) The Roche 454 presents a similar, though less extreme, result. d) In Illumina (zoomed in Figure S3.1), Magic-BLAST followed by STAR 1-pass have the steepest slopes. Then come STAR 2-pass and TopHat2, then HISAT2, these three aligners call unannotated introns at high coverage. e to j) In the Baruzzo Human T1 and T2 benchmarks, Magic-BLAST then HISAT2 perform the best, followed by STAR 1- then 2-pass, then TopHat2. In Human T3, HISAT2 and TopHat drop considerably, only Magic-BLAST and STAR can cope with a high level of mismatches. In the ultra-deep malaria sets, magic-BLAST remains best, but STAR 2-pass and HISAT2 drop way below TopHat. At coverage 1, STAR has by far the largest number of false positives (Figure S3.2 to S3.7).

Figure 4. Performance of the aligners on intron discovery and alignment statistics for the experimental datasets and iRefSeq. iRefSeq is used as a reference for intron discovery.

Figure 5. Read and alignment length distributions for iRefSeq (a), PacBio (b), Roche 454 (c), and Illumina (d) data sets.

Figure 6. Characteristics of alignments of the iRefSeq. a) The 45,108 iRefSeq include 154,601,873 bases exactly matching the genome (Truth). For each program, the percentage of iRefSeq sequences aligned (green) and the percentage of bases aligned (yellow) are given. In case of multiple alignments, each read contributes only once, at its primary position. b) Accuracy of the alignments: the true alignment of each read is defined by the RefSeq annotation (GRCh38 GFF file). A read is considered exactly aligned if the aligner recovered the exact chromosomal coordinates of the first and last base of the read. c.) Results for intron discovery in iRefSeq. d.) Number of matches versus mismatches: the iRefSeq sequences should match the genome exactly, so the Truth is in the upper right corner.

Figure 7. Performance of the aligners and quality control characteristics measured on the Baruzzo Illumina like benchmark: successively the intron discovery F-score; the statistics of the alignments

(percentage of reads and bases aligned, percentage of compatible pairs and percentage of reads aligned uniquely); the precision and recall of the alignments, as compared to the benchmark; finally the number of mismatches per kilobase aligned, for each aligner and as measured in the original dataset according to the benchmark mapping.

Figure 8. The first two BLASTN alignments of U00001 on human chromosome 17. U00001 is the query and is shown on the top of each row. The first exon really ends at base 85 of the mRNA, but BLASTN aligns the mRNA query to the first exon as well as an extra base in the intron. The second alignment starts three bases before the beginning of the second exon. BLASTN is not splice aware and aligns beyond the splice site.

Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads

Grzegorz M Boratyn, Jean Thierry-Mieg*, Danielle Thierry-Mieg, Ben Busby and Thomas L Madden*

Main Figures, to be submitted with the article

Figure 1. An example alignment extension procedure.

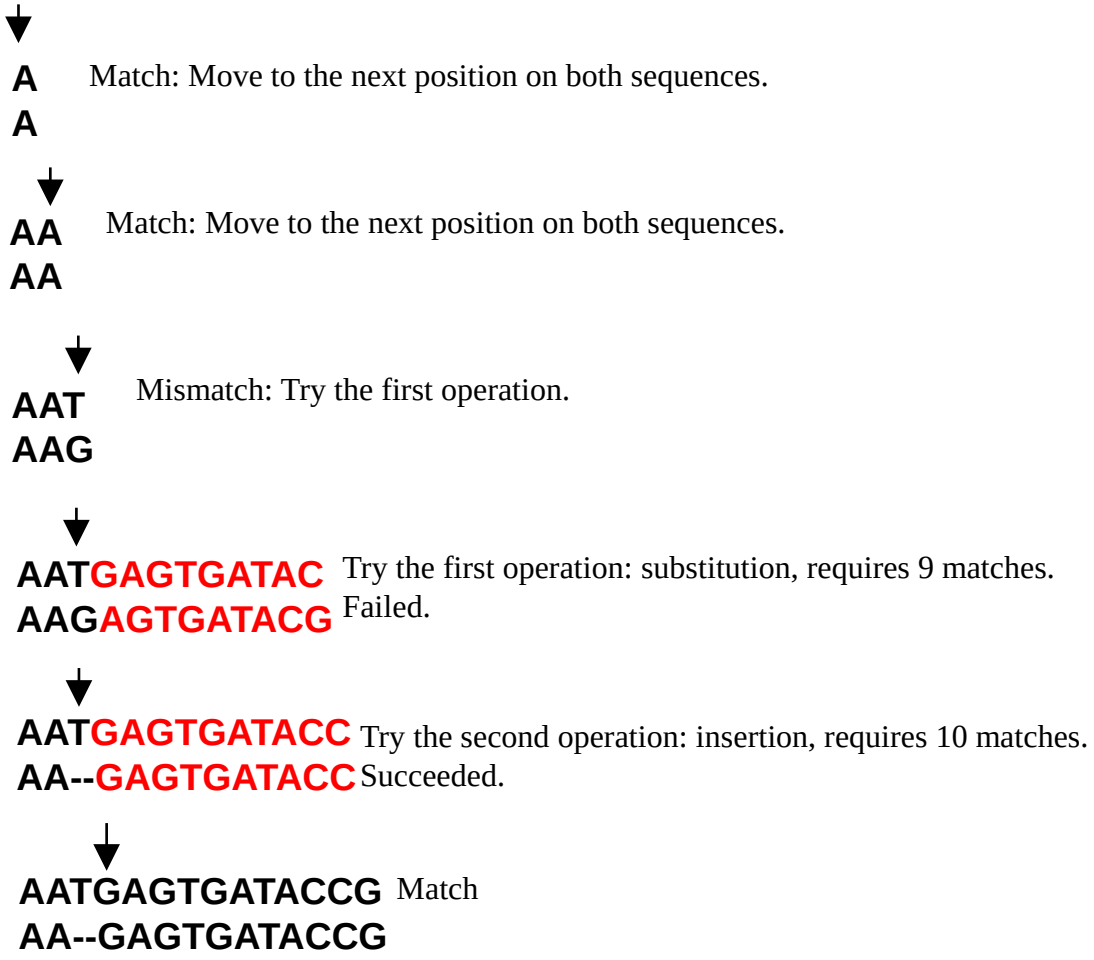


Figure 2

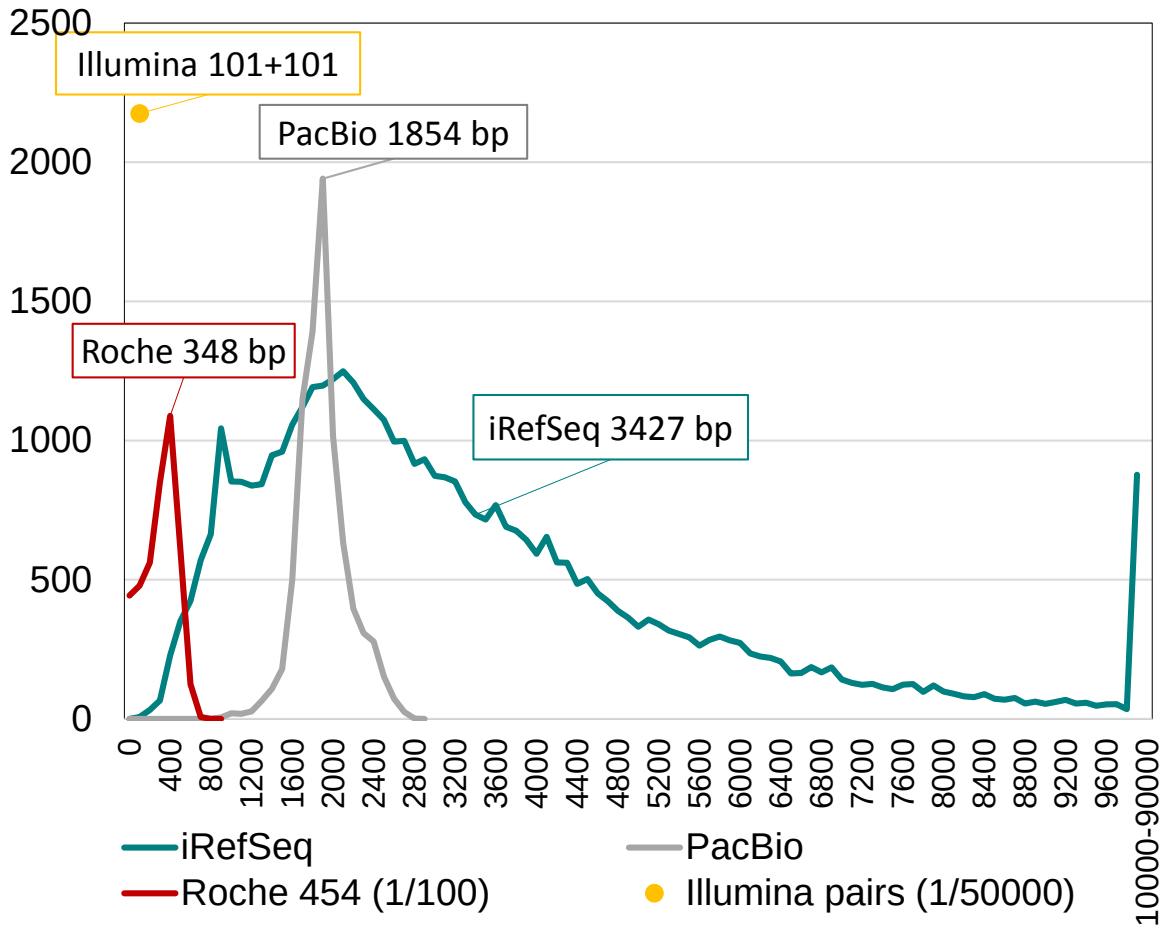

2500
2000
1500
1000
500
0
Illumina 101+101
PacBio 1854 bp
Roche 348 bp
iRefSeq 3427 bp
0
400
800
1200
1600
2000
2400
2800
3200
3600
4000
4400
4800
5200
5600
6000
6400
6800
7200
7600
8000
8400
8800
9200
9600
10000-90000
iRefSeq
PacBio
Roche 454 (1/100)
Illumina pairs (1/50000)

Figure 3a 3b

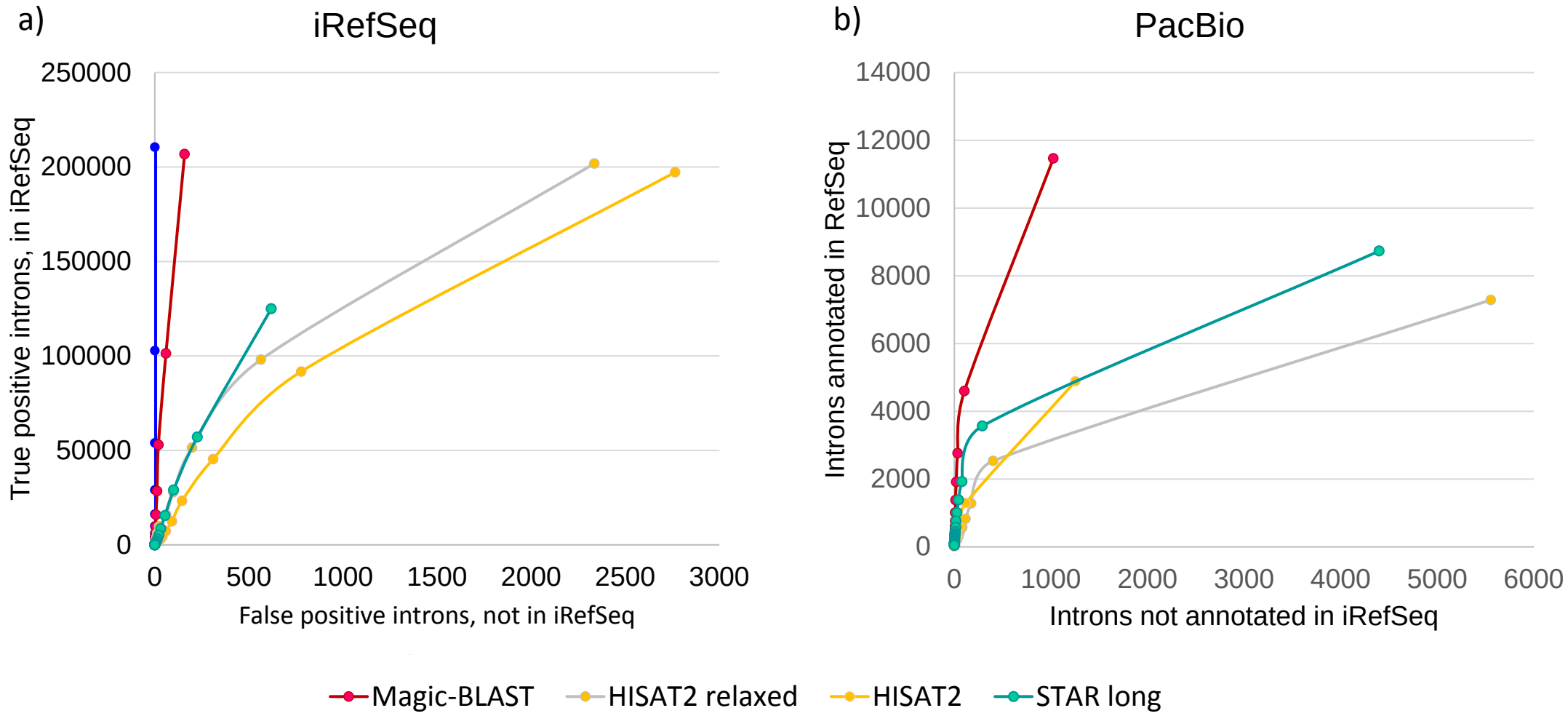

a)
iRefSeq
250000
200000
150000
100000
50000
0
True positive introns, in iRefSeq
0
500
1000
1500
2000
2500
3000
False positive introns, not in iRefSeq
b)
PacBio
14000
12000
10000
8000
6000
4000
2000
0
Introns annotated in RefSeq
0
1000
2000
3000
4000
5000
6000
Introns not annotated in iRefSeq
Magic-BLAST
HISAT2 relaxed
HISAT2
STAR long

Figure 3c, 3d

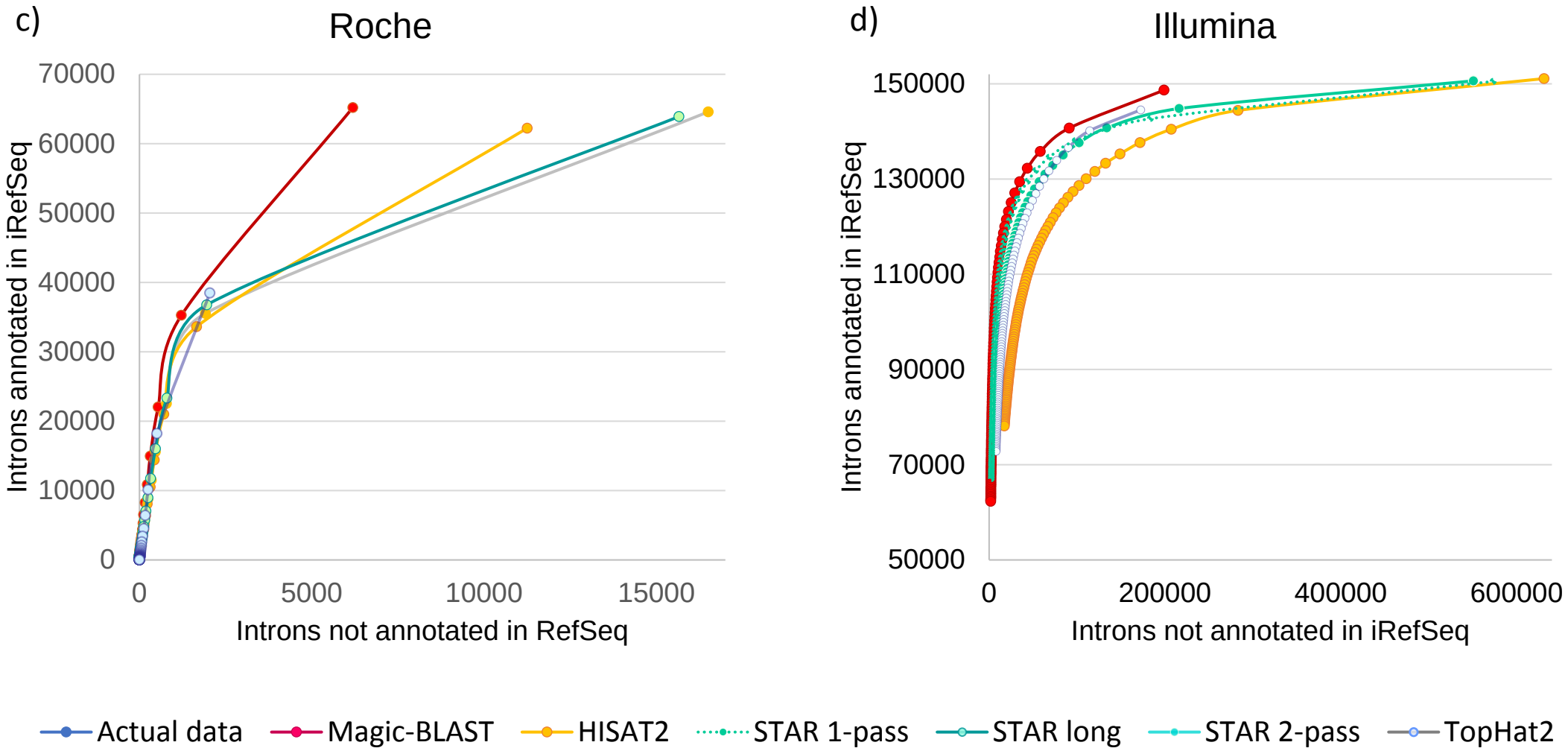
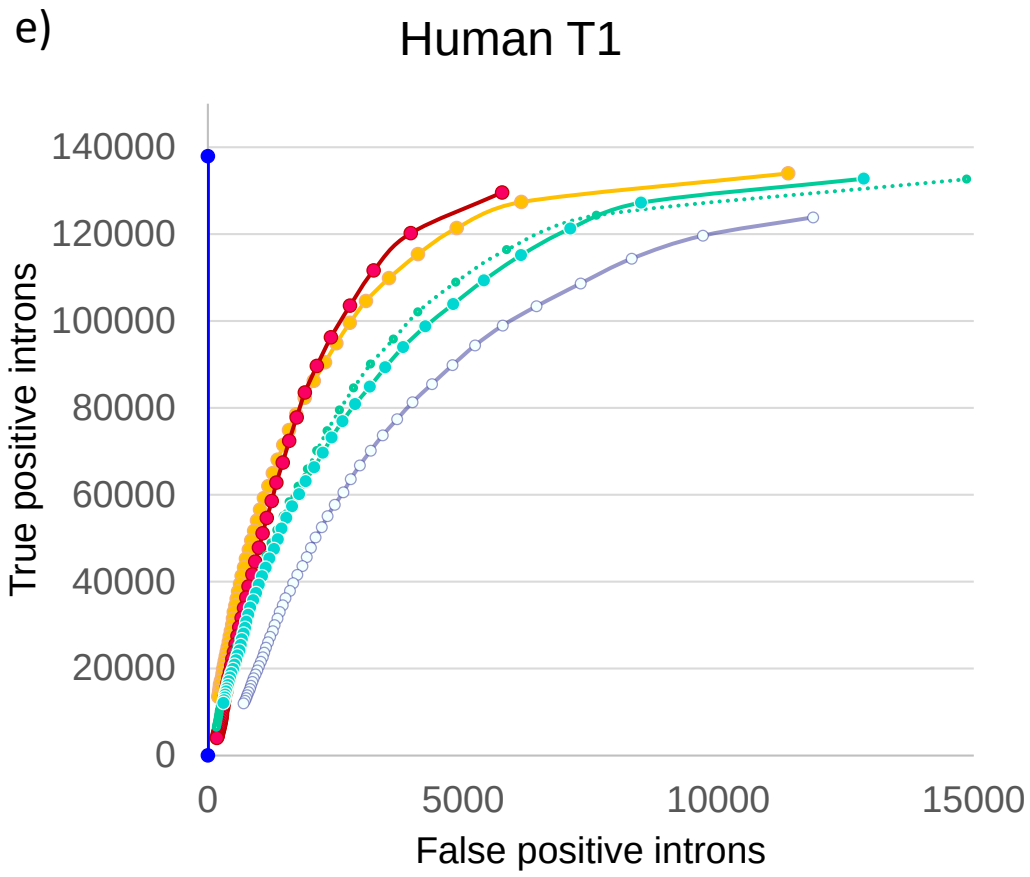

c)
Roche
70000
60000
50000
40000
30000
20000
10000
0
Introns annotated in iRefSeq
0
5000
10000
15000
Introns not annotated in RefSeq
d)
Illumina
150000
130000
110000
90000
70000
50000
Introns annotated in iRefSeq
0
200000
400000
600000
Introns not annotated in iRefSeq
Actual data
Magic-BLAST
HISAT2
STAR 1-pass
STAR long
STAR 2-pass
TopHat2

Figure 3e 3f


e)
Human T1
140000
120000
100000
80000
60000
40000
20000
0
True positive introns
0
5000
10000
15000
False positive introns

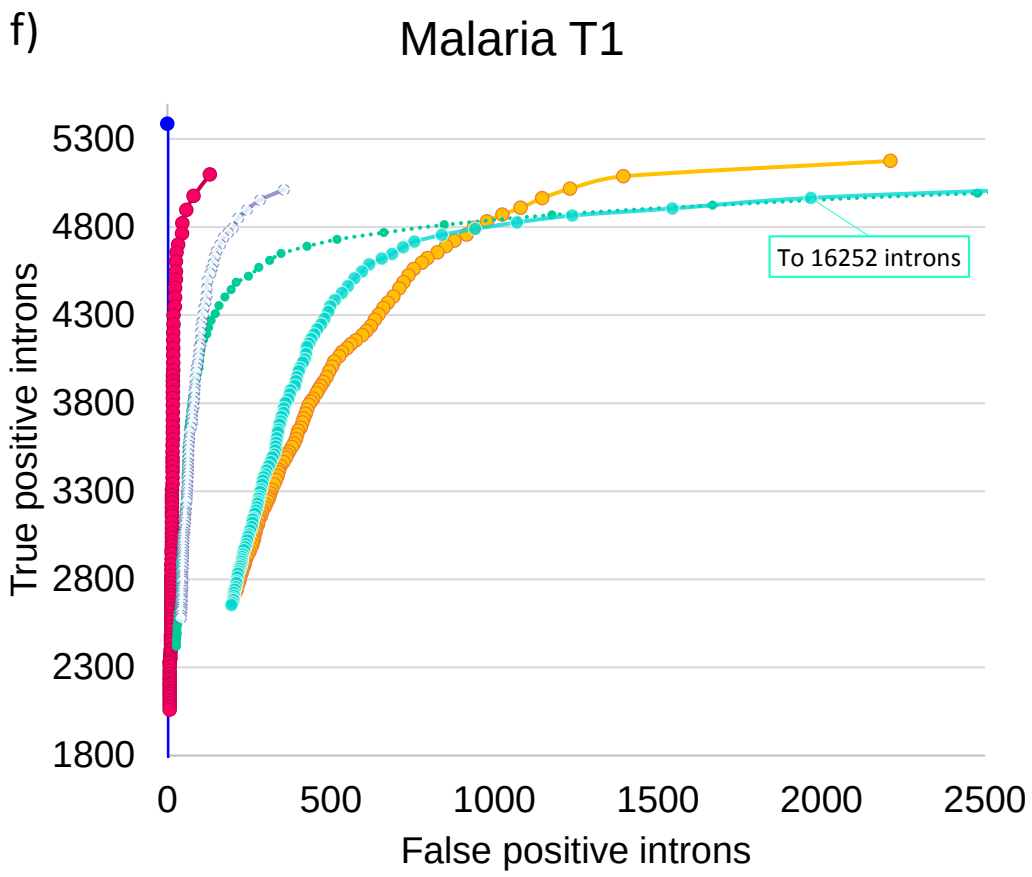
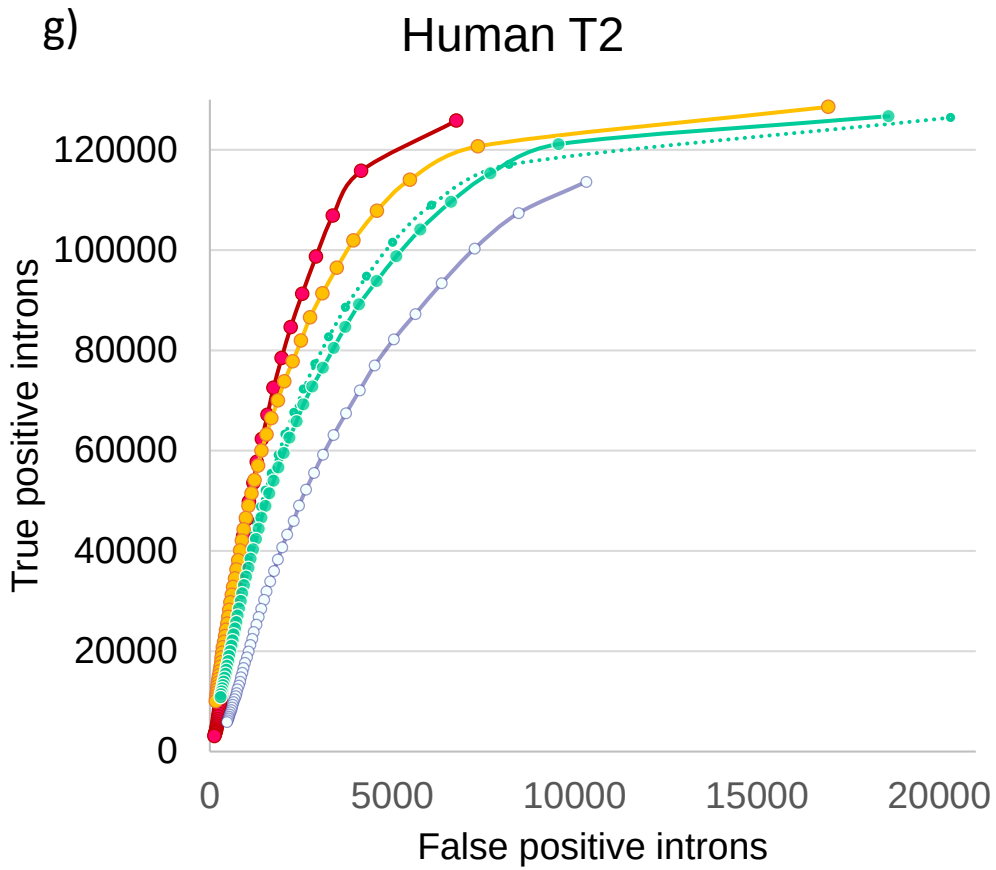

f)
Malaria T1
5300
4800
4300
3800
3300
2800
2300
1800
True positive introns
To 16252 introns
0
500
1000
1500
2000
2500
False positive introns

Figure 3g 3h


g)
Human T2
True positive introns
False positive introns
0
20000
40000
60000
80000
100000
120000
0
5000
10000
15000
20000

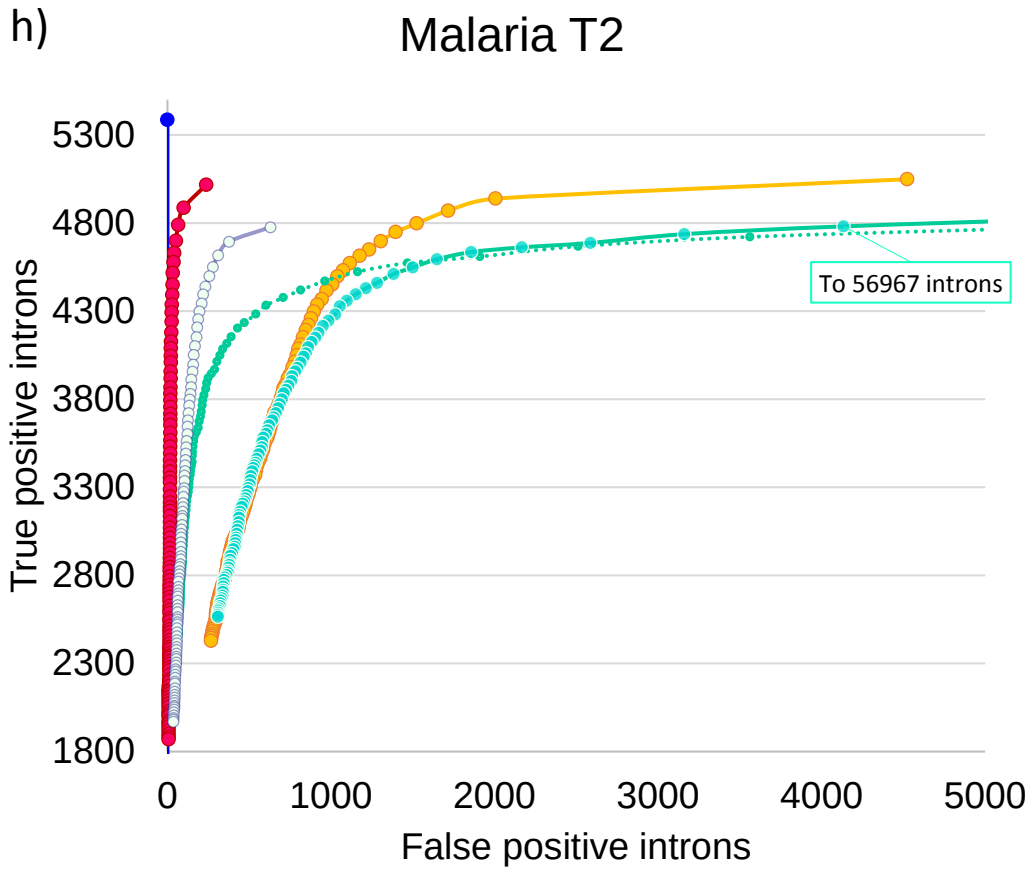

h)
Malaria T2
True positive introns
False positive introns
1800
2300
2800
3300
3800
4300
4800
5300
0
1000
2000
3000
4000
5000
To 56967 introns

Figure 3i then 3j

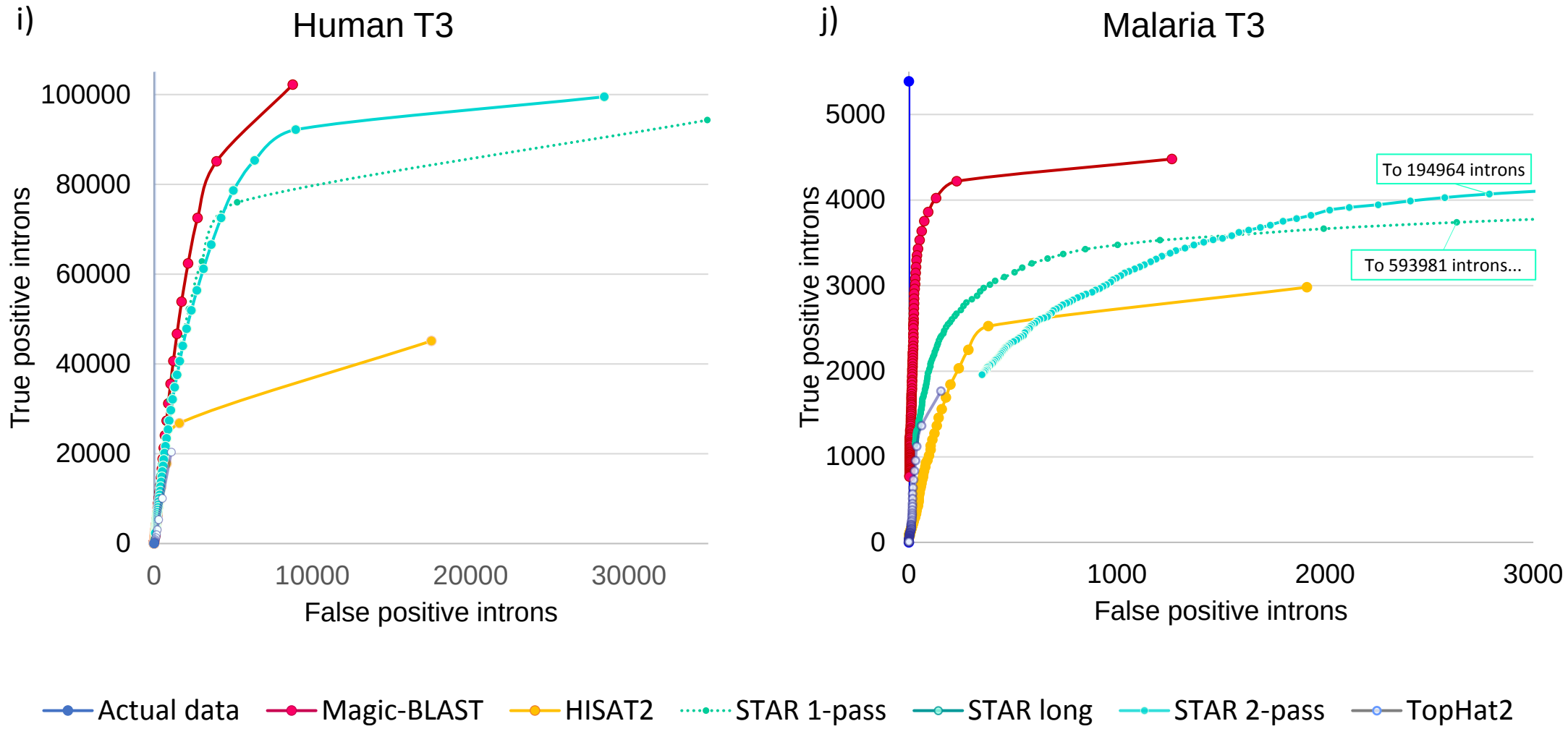
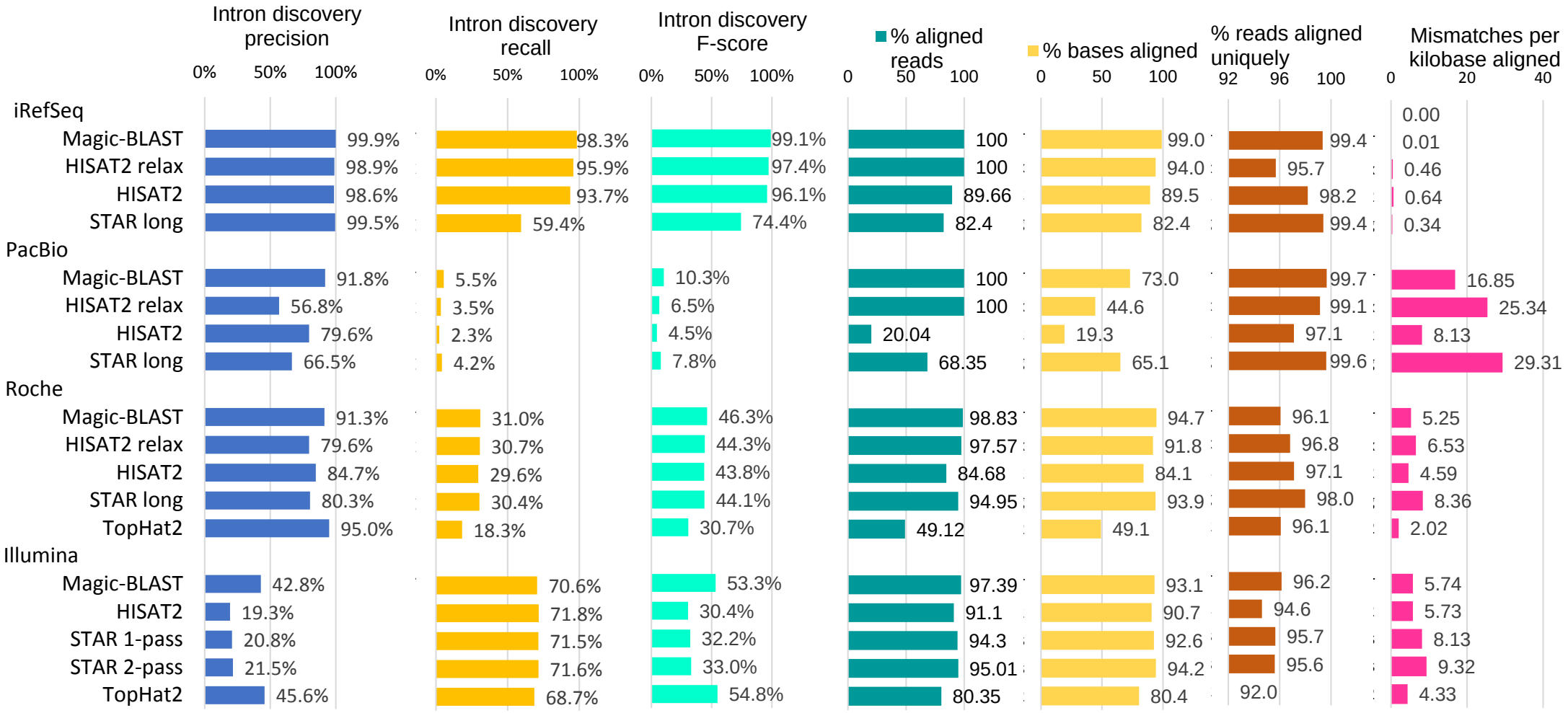
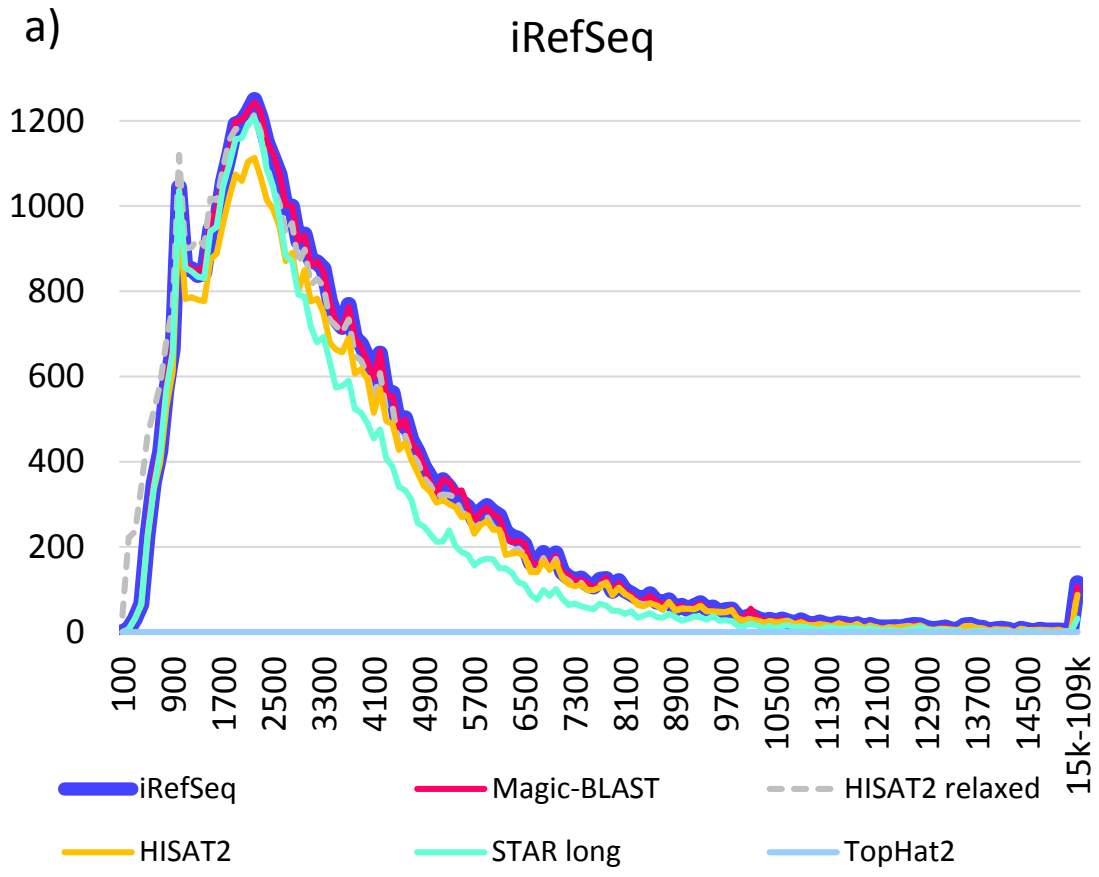


Figure 4


Intron discovery precision
0%
50%
100%
Intron discovery recall
0%
50%
100%
Intron discovery F-score
0%
50%
100%
% aligned reads
0
50
100
% bases aligned
0
50
100
% reads aligned uniquely
92
96
100
Mismatches per kilobase aligned
0
20
40
iRefSeq
Magic-BLAST 99.9% 98.3% 99.1% 100 99.0 99.4 0.00
HISAT2 relax 98.9% 95.9% 97.4% 100 94.0 95.7 0.01
HISAT2 98.6% 93.7% 96.1% 89.66 89.5 98.2 0.46
STAR long 99.5% 59.4% 74.4% 82.4 82.4 99.4 0.64
0.34
PacBio
Magic-BLAST 91.8% 5.5% 10.3% 100 73.0 99.7 16.85
HISAT2 relax 56.8% 3.5% 6.5% 100 44.6 99.1 25.34
HISAT2 79.6% 2.3% 4.5% 20.04 19.3 97.1 8.13
STAR long 66.5% 4.2% 7.8% 68.35 65.1 99.6 29.31
Roche
Magic-BLAST 91.3% 31.0% 46.3% 98.83 94.7 96.1 5.25
HISAT2 relax 79.6% 30.7% 44.3% 97.57 91.8 96.8 6.53
HISAT2 84.7% 29.6% 43.8% 84.68 84.1 97.1 4.59
STAR long 80.3% 30.4% 44.1% 94.95 93.9 98.0 8.36
TopHat2 95.0% 18.3% 30.7% 49.12 49.1 96.1 2.02
Illumina
Magic-BLAST 42.8% 70.6% 53.3% 97.39 93.1 96.2 5.74
HISAT2 19.3% 71.8% 30.4% 91.1 90.7 94.6 5.73
STAR 1-pass 20.8% 71.5% 32.2% 94.3 92.6 95.7 8.13
STAR 2-pass 21.5% 71.6% 33.0% 95.01 94.2 95.6 9.32
TopHat2 45.6% 68.7% 54.8% 80.35 80.4 92.0 4.33

Figures 5a, 5b



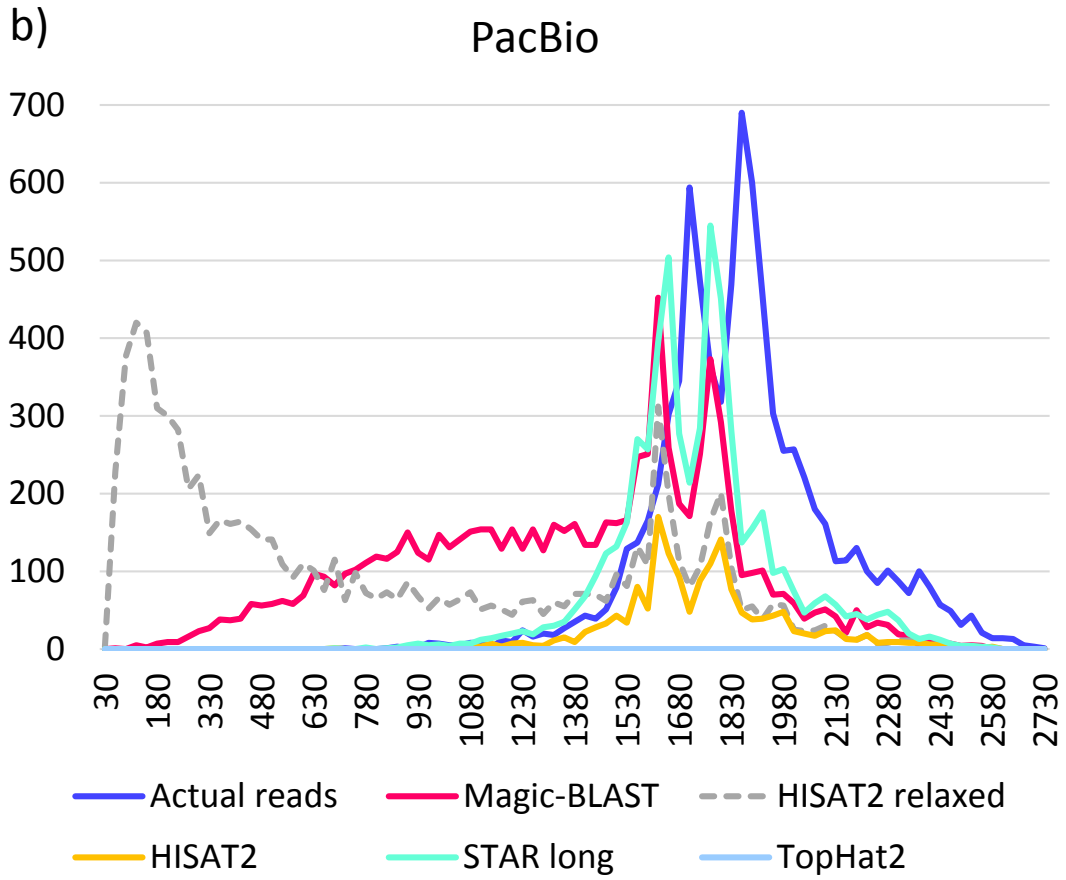
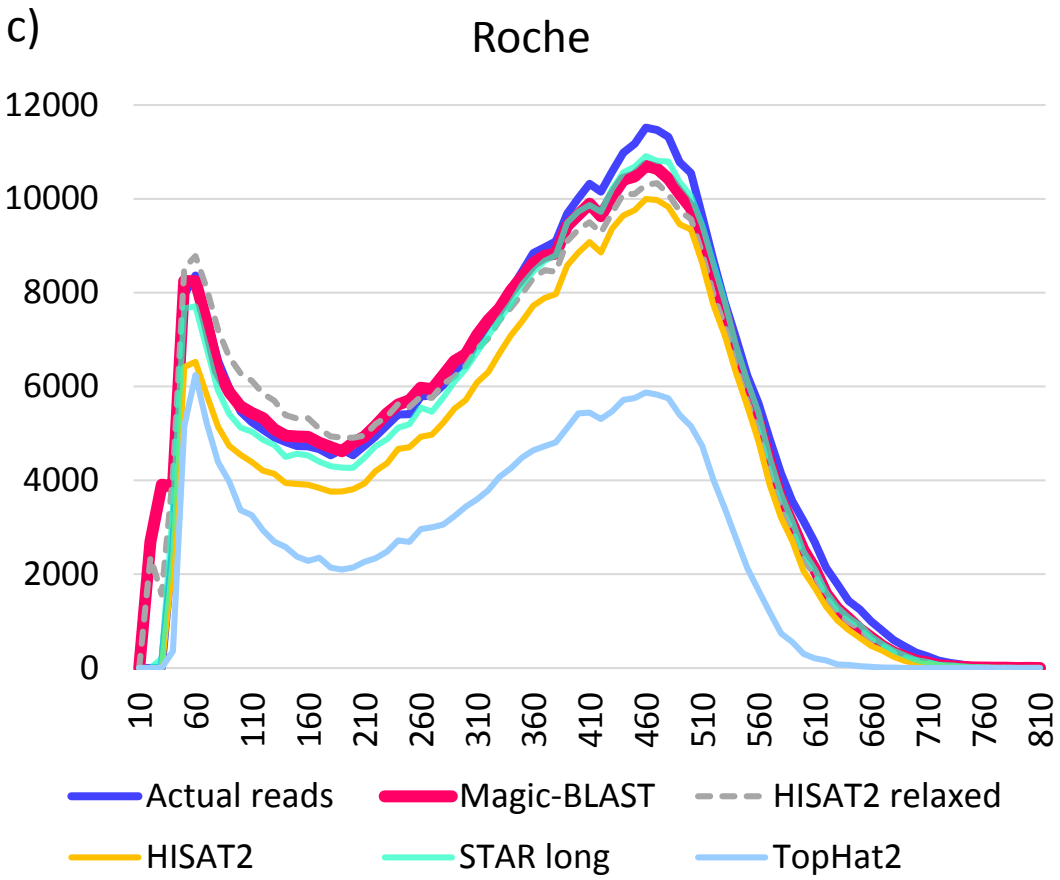
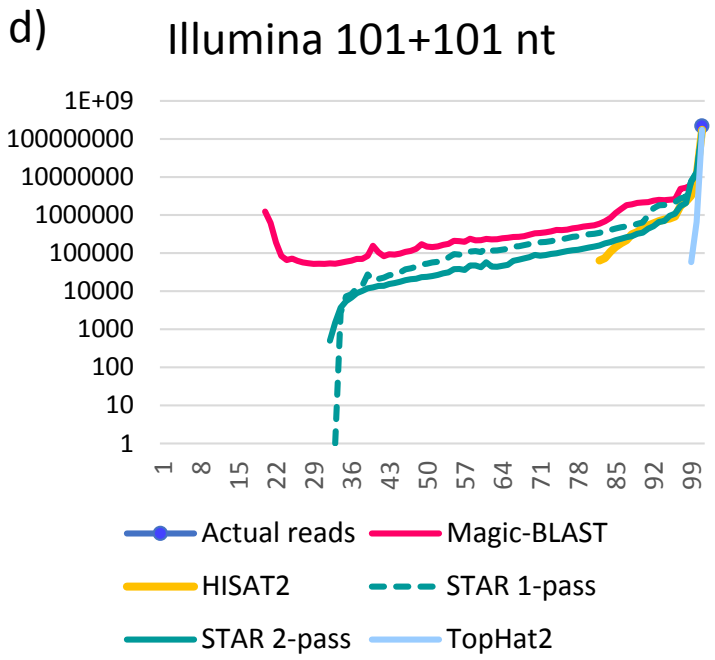


Figure 5c 5d


c)
Roche
12000
10000
8000
6000
4000
2000
0
10
60
110
160
210
260
310
360
410
460
510
560
610
660
710
760
810
Actual reads
Magic-BLAST
HISAT2 relaxed
HISAT2
STAR long
TopHat2


d)
Illumina 101+101 nt
1E+09
100000000
10000000
1000000
100000
10000
1000
100
10
1
1
8
15
22
29
36
43
50
57
64
71
78
85
92
99
Actual reads
Magic-BLAST
HISAT2
STAR 1-pass
STAR 2-pass
TopHat2

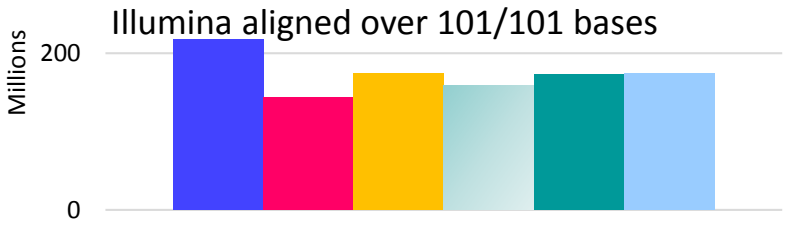

Illumina aligned over 101/101 bases
Millions
200
0

Figure 6

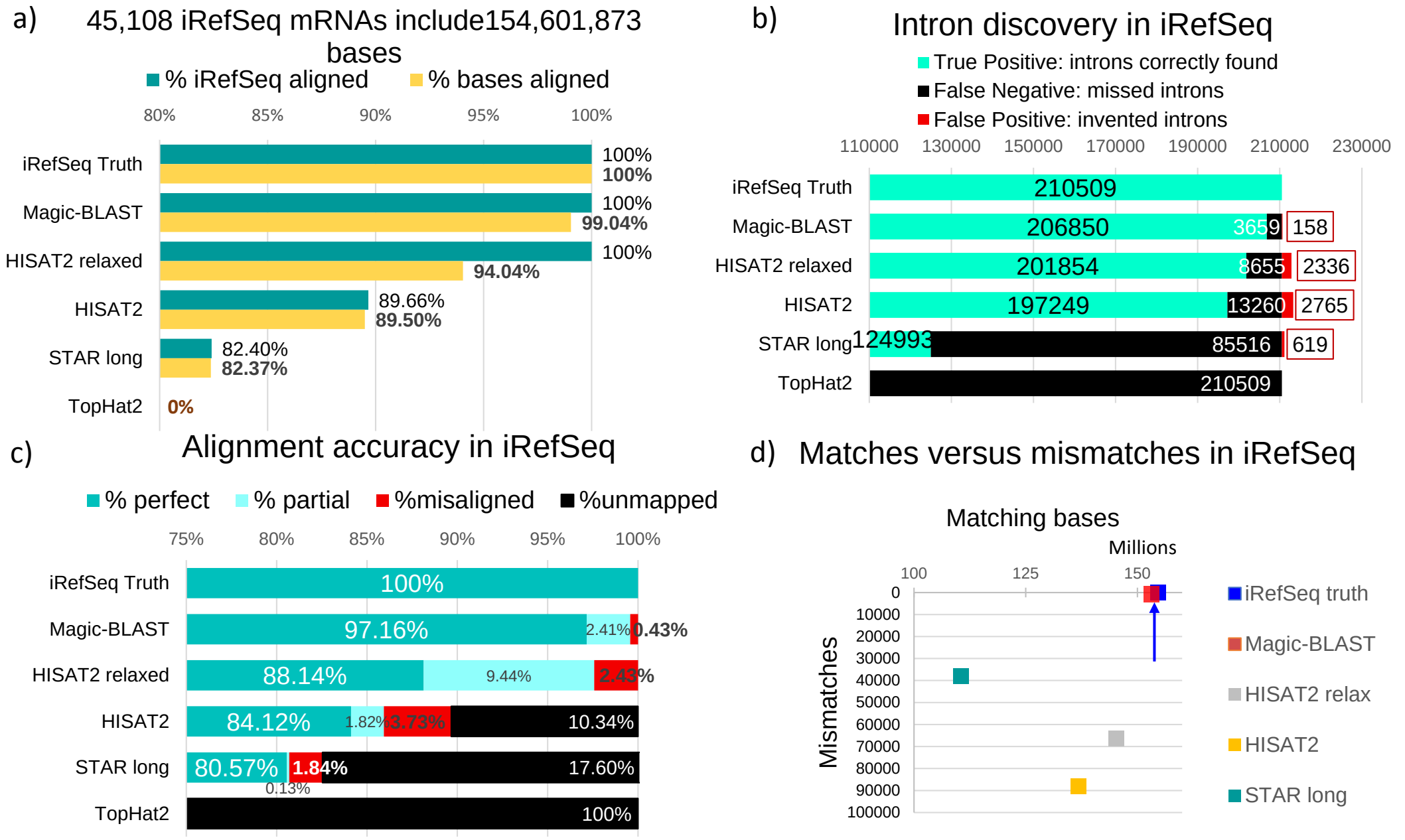

a)
45,108 iRefSeq mRNAs include154,601,873 bases
% iRefSeq aligned
% bases aligned
80%
85%
90%
95%
100%
iRefSeq Truth
100%
100%
Magic-BLAST
100%
99.04%
HISAT2 relaxed
100%
94.04%
HISAT2
89.66%
89.50%
STAR long
82.40%
82.37%
TopHat2
0%
b)
Intron discovery in iRefSeq
True Positive: introns correctly found
False Negative: missed introns
False Positive: invented introns
110000
130000
150000
170000
190000
210000
230000
iRefSeq Truth
210509
Magic-BLAST
206850
3659
158
HISAT2 relaxed
201854
8655
2336
HISAT2
197249
13260
2765
STAR long
124993
85516
619
TopHat2
210509
c)
Alignment accuracy in iRefSeq
% perfect
% partial
%misaligned
%unmapped
75%
80%
85%
90%
95%
100%
iRefSeq Truth
100%
Magic-BLAST
97.16%
2.41%
0.43%
HISAT2 relaxed
88.14%
9.44%
2.43%
HISAT2
84.12%
1.82%
3.73%
10.34%
STAR long
80.57%
1.84%
0.13%
17.60%
TopHat2
100%
d)
Matches versus mismatches in iRefSeq
Matching bases
Millions
100
125
150
Mismatches
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
iRefSeq truth
Magic-BLAST
HISAT2 relax
HISAT2
STAR long

Figure 7

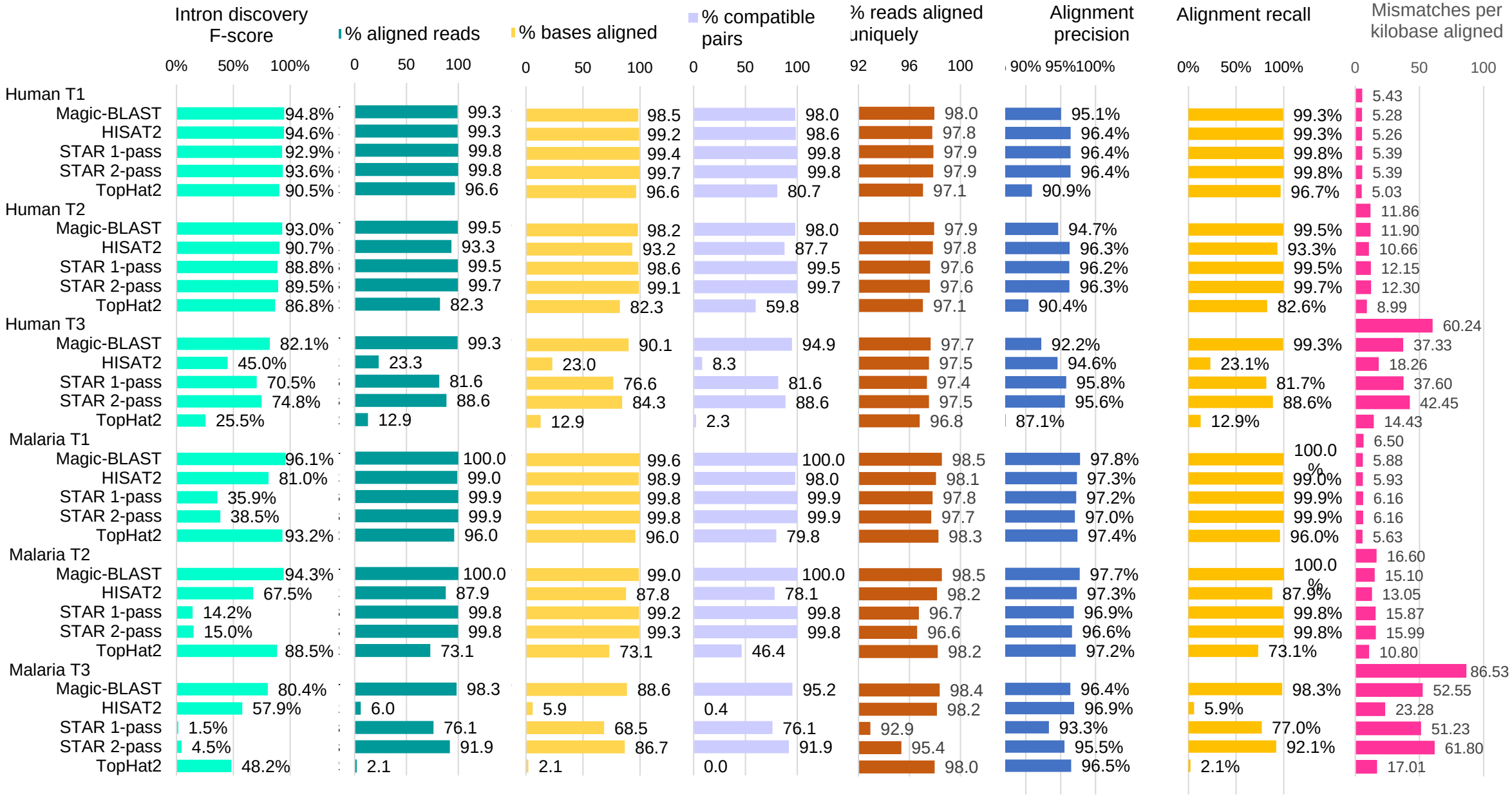

Intron discovery F-score
% aligned reads
% bases aligned
% compatible pairs
% reads aligned uniquely
Alignment precision
Alignment recall
Mismatches per kilobase aligned
Human T1
Magic-BLAST 94.8% 99.3 98.5 98.0 98.0 95.1% 99.3% 5.43
HISAT2 94.6% 99.3 99.2 98.6 97.8 96.4% 99.3% 5.28
STAR 1-pass 92.9% 99.8 99.4 99.8 97.9 96.4% 99.8% 5.26
STAR 2-pass 93.6% 99.8 99.7 99.8 97.9 96.4% 99.8% 5.39
TopHat2 90.5% 96.6 96.6 80.7 97.1 90.9% 96.7% 5.03
Human T2
Magic-BLAST 93.0% 99.5 98.2 98.0 97.9 94.7% 99.5% 11.86
HISAT2 90.7% 93.3 93.2 87.7 97.8 96.3% 93.3% 11.90
STAR 1-pass 88.8% 99.5 98.6 99.5 97.6 96.2% 99.5% 10.66
STAR 2-pass 89.5% 99.7 99.1 99.7 97.6 96.3% 99.7% 12.15
TopHat2 86.8% 82.3 82.3 59.8 97.1 90.4% 82.6% 12.30
Human T3
Magic-BLAST 82.1% 99.3 90.1 94.9 97.7 92.2% 99.3% 8.99
HISAT2 45.0% 23.3 23.0 8.3 97.5 94.6% 23.1% 60.24
STAR 1-pass 70.5% 81.6 76.6 81.6 97.4 95.8% 81.7% 37.33
STAR 2-pass 74.8% 88.6 84.3 88.6 97.5 95.6% 88.6% 18.26
TopHat2 25.5% 12.9 12.9 2.3 96.8 87.1% 12.9% 37.60
Malaria T1
Magic-BLAST 96.1% 100.0 99.6 100.0 98.5 97.8% 100.0% 42.45
HISAT2 81.0% 99.0 98.9 98.0 98.1 97.3% 99.0% 14.43
STAR 1-pass 35.9% 99.9 99.8 99.9 97.8 97.2% 99.9% 6.50
STAR 2-pass 38.5% 99.9 99.8 99.9 97.7 97.0% 99.9% 5.88
TopHat2 93.2% 96.0 96.0 79.8 98.3 97.4% 96.0% 5.93
Malaria T2
Magic-BLAST 94.3% 100.0 99.0 100.0 98.5 97.7% 100.0% 6.16
HISAT2 67.5% 87.9 87.8 78.1 98.2 97.3% 87.9% 6.16
STAR 1-pass 14.2% 99.8 99.2 99.8 96.7 96.9% 99.8% 5.63
STAR 2-pass 15.0% 99.8 99.3 99.8 96.6 96.6% 99.8% 16.60
TopHat2 88.5% 73.1 73.1 46.4 98.2 97.2% 73.1% 15.10
Malaria T3
Magic-BLAST 80.4% 98.3 88.6 95.2 98.4 96.4% 98.3% 13.05
HISAT2 57.9% 6.0 5.9 0.4 98.2 96.9% 5.9% 15.87
STAR 1-pass 1.5% 76.1 68.5 76.1 92.9 93.3% 77.0% 15.99
STAR 2-pass 4.5% 91.9 86.7 91.9 95.4 95.5% 92.1% 10.80
TopHat2 48.2% 2.1 2.1 0.0 98.0 96.5% 2.1% 86.53
52.55
23.28
51.23
61.80
17.01

Figure 8

Homo sapiens chromosome 17, GRCh38.p12 Primary Assembly

Sequence ID: NC_000017.11 **Length:** 83257441 **Number of Matches:** 21

Range 1: 47189145 to 47189224 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
145 bits(160)	1e-31	80/80(100%)	0/80(0%)	Plus/Minus

Features: cell division cycle protein 27 homolog isoform 3
cell division cycle protein 27 homolog isoform 2

```
Query  7         CCGCTACAggggggCCTGAGGCACTGCAGAAAGTGGGCCTGAGCCTCGAGGATGACGGT  66
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  47189224  CCGCTACAGGGGGGGCCTGAGGCACTGCAGAAAGTGGGCCTGAGCCTCGAGGATGACGGT  47189165

Query  67        GCTGCAGGAACCCGTCCAGG  86
                 ||||||||||||||||||||
Sbjct  47189164  GCTGCAGGAACCCGTCCAGG  47189145
```

Range 2: 47181562 to 47181640 GenBank Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Identities	Gaps	Strand
143 bits(158)	4e-31	79/79(100%)	0/79(0%)	Plus/Minus

Features: cell division cycle protein 27 homolog isoform 4
cell division cycle protein 27 homolog isoform X6

```
Query  83        CAGGCTGCTATATGGCAAGCACTAAACCACTATGCTTACCGAGATGCGGTTTTCCTCGCA  142
                 |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  47181640  CAGGCTGCTATATGGCAAGCACTAAACCACTATGCTTACCGAGATGCGGTTTTCCTCGCA  47181581

Query  143       GAACGCCTTTATGCAGAAG  161
                 |||||||||||||||||||
Sbjct  47181580  GAACGCCTTTATGCAGAAG  47181562
```