BMC Bioinformatics

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

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

Abstract

Background: 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. We introduce Magic-BLAST, a new aligner based on ideas from the Magic pipeline.

Results: Magic-BLAST 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.

Conclusions: We show that Magic-BLAST is the best at intron discovery over a wide range of conditions and the best at mapping reads longer than 250 bases, from any platform. It is versatile and robust to high levels of mismatches or extreme base composition, and 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.

Keywords: RNA-seq, BLAST, Alignment

Background

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. Different analysis approaches are available for some investigations. For example, gene expression can be studied with alignments or with alignment-free methods [1, 2]. On the other hand, the investigation of fine grain gene structure does require alignments or sequence assembly and may benefit from a specific sequencing technology. While relatively short reads, 50 or 100 bases paired end Illumina style reads are sufficient for 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 mapping and variant 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 [3–5], but a recent benchmark by Baruzzo and the Grant team showed that new aligners could still offer value to the community [6].

Here, we present Magic-BLAST, the RNA-seq splice aware and pair aware new member of the BLAST family of programs. We compare Magic-BLAST to five other popular aligners. We look at a variety of RNA-seq read lengths, between 100 bases and 100 kilobases, from Illumina, Roche 454 and PacBio as well as full length RefSeq transcripts and find that most aligners can accurately map short Illumina-style human paired reads with few mismatches, but 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,

* Correspondence: madden@ncbi.nlm.nih.gov
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA

Magic-BLAST performs well in the mapping of long reads and in identification of introns, even though it is single pass and is aligning RNA to the genome without knowledge of a transcriptome.

We chose the name Magic-BLAST to emphasize the ideas that went into the tool. Magic-BLAST derives its core algorithms from the Magic pipeline, described in detail in the Additional file 1 of [7]. It is implemented using the same C++ framework as the BLAST+ executables [8–10]. 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 has a more limited role than its namesake. Magic refers to an entire pipeline with high level functionalities, but Magic-BLAST implements de novo just a Magic-like aligner.

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. It was field-tested in several NCBI hackathons that provided substantial feedback on features and usability.

We compare Magic-BLAST (version 1.4) to three popular aligners, HISAT2 [11], STAR [12, 13], and TopHat2 [14, 15], also evaluated in [1, 6]. 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 for intron discovery, as measured on truth-bearing RefSeq [16] human transcripts (assembled 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 [6]. 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 well when mapping RNA to the genome or to the transcriptome. For RNA-seq, it works well with a variety of read lengths and data without requiring from the user any choice of parameters. It is versatile, easy to use, robustly precise and conservative in all circumstances.

Magic-BLAST can also align DNA-seq, but we examine only RNA-seq performance here.

Additionally, we performed a reduced set of tests on two aligners, Minimap2 [17] and Subread [18], that claim to handle long reads well.
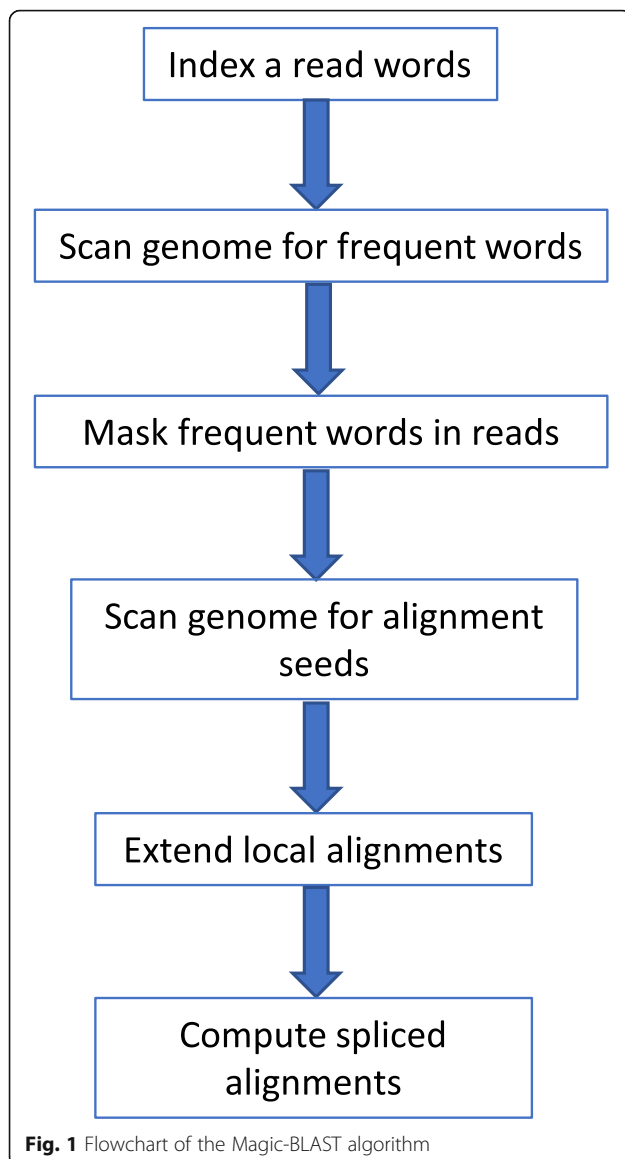
Implementation
Algorithm overview
The Magic-BLAST algorithm has a structure similar to that of other BLAST programs [10]. It reads the data in batches and builds a "lookup table", which is an index of word locations in the reads, 16-bases by default. It then scans the database sequences, usually a reference 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. In order to be used as a seed, the original 18 base match must be completely contained within one exon (i.e., cannot span two exons). Consequently, exons shorter than the seed length cannot be captured, but they are rare (less than 0.2% of RefSeq exons), 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. Figure 1 presents an overview of these 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 sequence 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 for RNA-seq. Additionally, Magic-BLAST specifically masks out 16-base words that contain at least 15 A's or 15 T's, effectively avoiding seeding on poly-A

Fig. 1 Flowchart of the Magic-BLAST algorithm

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

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 [7]. 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 (Table 1) 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.

Figure 2 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 [8] 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
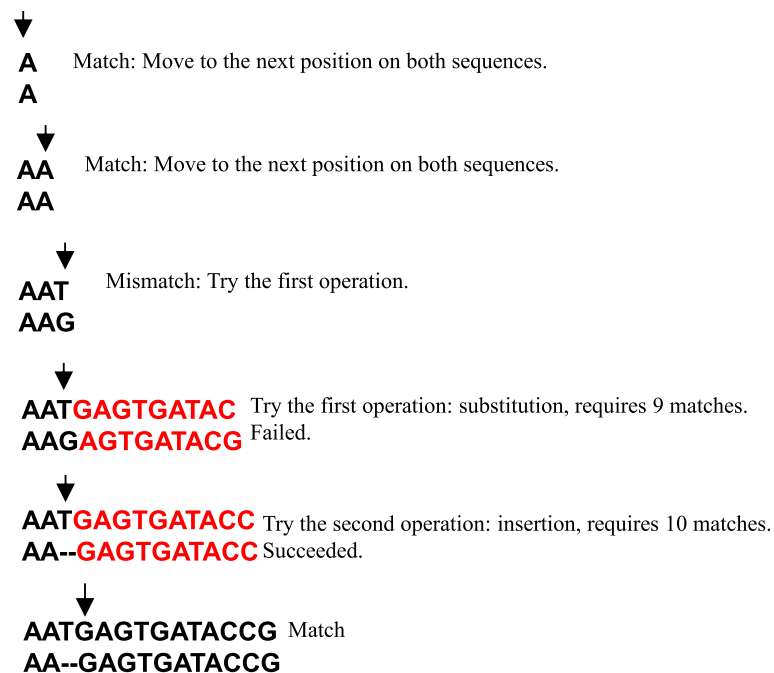
Fig. 2 An example alignment extension procedure. The arrows point at each step to the position currently considered in both sequences. First, there are two matches and the arrows move to the right by two positions on both sequences. When a mismatch (T-G) is encountered the algorithm tries alignment operations 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)

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 have no impact on the alignment score and are not exported in the SAM output.

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 by convention 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 canonical 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 non-canonical site. The first site found is accepted. The alignment score threshold of 50 was chosen because minor and 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 that it correctly identifies introns even in the presence of a substitution, 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 support the use of an annotation file or a two-pass method. If desired, one can map both on the genome and on an annotated transcriptome, then use the universal scoring system of Magic-BLAST to select the best alignment, be it genomic or transcriptomic, for each fragment. In this paper, we mapped only to the genome.

Output
Magic-BLAST returns results in the Sequence Alignment/ Map SAM/BAM format [19] 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 several experimental runs, from Illumina, Roche 454 and PacBio. Table 2 summarizes the datasets used for this analysis.

The new benchmark, called iRefSeq, is the image of the Human RefSeq mRNAs [16] 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 (GCF_000001405.36). 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 Additional file 1: section 2.1). 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 iRefSeq are known.

The Baruzzo benchmark set of simulated RNA-seq reads, presented in [6], 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 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 names these three different error rates T1, T2, and T3. The variable error rates simulate how the aligners would perform if the genome of the same species was available (T1), if only a poor-quality version of the genome or the sequence data was available (T2), and if only the genome of a related organism was available

Table 2 Datasets used for analysis

Name	Type	Number of Reads	Average read length (bases)	Range of read lengths (bases)	Accession
iRefSeq	Real	45,108	3427	147–109224	N/A
PacBio	Real	8,285	1854	710–2703	SRR5009494
Roche	Real	416,045	348	36–808	SRR899420
Illumina 101 + 101	Real	109 million	202	202–202	SRR534301
Illumina 250 + 250	Real	23.3 million	500	500–500	SRR5438850
Illumina 300 + 300	Real	16.5 million	600	600–600	SRR5437876
PacBio Brain	Real	264,100	1323	46–6116	SRR5189652
PacBio testes	Real	408,801	1307	42–7359	SRR5189667
Human T1	Simulated	10 million	200	200–200	N/A
Human T2	Simulated	10 million	200	200–200	N/A
Human T3	Simulated	10 million	200	200–200	N/A
Malaria T1	Simulated	10 million	200	200–200	N/A
Malaria T2	Simulated	10 million	200	200–200	N/A
Malaria T3	Simulated	10 million	200	200–200	N/A

(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 (all 18 datasets are shown in Additional file 1: Tables S3, S4, S6). Surprisingly, we noticed that the measured level of mismatches per kb 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 experimental RNA-seq data sets from the public Sequence Read Archive (SRA at NCBI) to represent three sequencing platforms, PacBio (SRR5009494, with 8285 reads of average length 1854 bases, sequenced from colon carcinoma cells), Roche 454 Titanium (SRR899420, with 416,045 reads of average length 348 bases, sequenced from the MAQC/SEQC brain mRNA sample (16)), and a deep Illumina HiSeq run (SRR534301, with close to 109 million 101 + 101 bases paired-end reads sequenced from fetal lung, also tested in (9)). Here, we will refer to each of these RNA-seq sets by its technology: PacBio, Roche and Illumina. Figure 3 presents a histogram of read lengths for the

iRefSeq and these three experimental sets. To refine the conclusions, we later added two large PacBio hybrid runs (testes SRR5189667 and brain SRR5189652 from the GENCODE project [20], with non-native sequences edited using Illumina runs), two Illumina runs with longer read pairs (250 + 250 bases SRR5438850 from metastatic melanoma; 300 + 300 bases SRR5437876 from MCF7 cells, also studied in [1]).

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 [6]). Magic-BLAST was compared to programs from 2009 to 2015: HISAT2 [11], STAR [12, 13], STAR long [1, 12, 20] and TopHat2 [14, 15] (Details in Additional file 1: section 1.2). The standard STAR is optimized for Illumina-type reads while the authors recommend STAR long for reads longer than 300 bases, but we tried both versions on all runs. The two-pass mode is recommended if no transcriptome annotation is provided: STAR long was run only in two-pass mode, but both one-pass and two-pass modes were tried for standard STAR. We used a parameter set for STAR long based on information in [13] as well as https://groups.google.com/d/forum/rna-star. HISAT2 was run with default parameters as well as in a 'relaxed' mode which is more sensitive and works better on longer reads but is much slower: the HISAT2 default parameters left

Fig. 3 Actual read length histograms for 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 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 Y-axis presents the number of reads per 100-base bin (on the X-axis). The Y scale is reduced 100 times for Roche 454, and 50,000 times for Illumina, which has the highest throughput of all technologies

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. Additionally, we performed preliminary analysis on two aligners optimized for long reads (Minimap2 [17] and Subread [18]).

Two analysis programs were used in this project (Additional file 1: section 1): AliQC.py was developed in collaboration with Joe Meehan from the FDA for the SEQC project [21]. It extracts, by comparing the SAM file to the genome, a detailed quality control on alignments, their length and multiplicity, and mismatches by reads, by type and by position along the read (i.e. by sequencing cycle) (Additional file 1: section 7). In case of multiple alignments, only the 'primary' alignment is considered in this analysis. 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 format in which Baruzzo [6] provided the benchmark truth (Additional file 1: sections 2 to 6). A master-script, described in Additional file 1: section 1.5 and deposited in GitHub (https://github.com/ncbi/magicblast), can download all the data, realign the sequences, and reproduce Additional file 1: tables and sections 2 to 7, which support our entire analysis.

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 aligners discover introns, the splice sites were extracted from the BAM output using the 'N' operation in the CIGAR string [19]. For the iRefSeq and Baruzzo benchmark sets, the true position of each intron is known. The 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 RefSeq. 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. We require that the starting and ending positions of introns identified by an aligner exactly match either the Ground Truth in the benchmark sets or the RefSeq annotation for the experimental sets.

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 standard STAR failed to produce any results for very long reads, although TopHat2 marginally worked for Roche 454.

We first use a ROC curve approach [22] to precisely judge the quality of intron discovery, true versus false, as a function of minimal read coverage (Fig. 4 and Additional file 1: Figure S3). We group the introns by read coverage in up to 100 bins. For each bin, 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 curves have up to 100 points and give us visual insight into
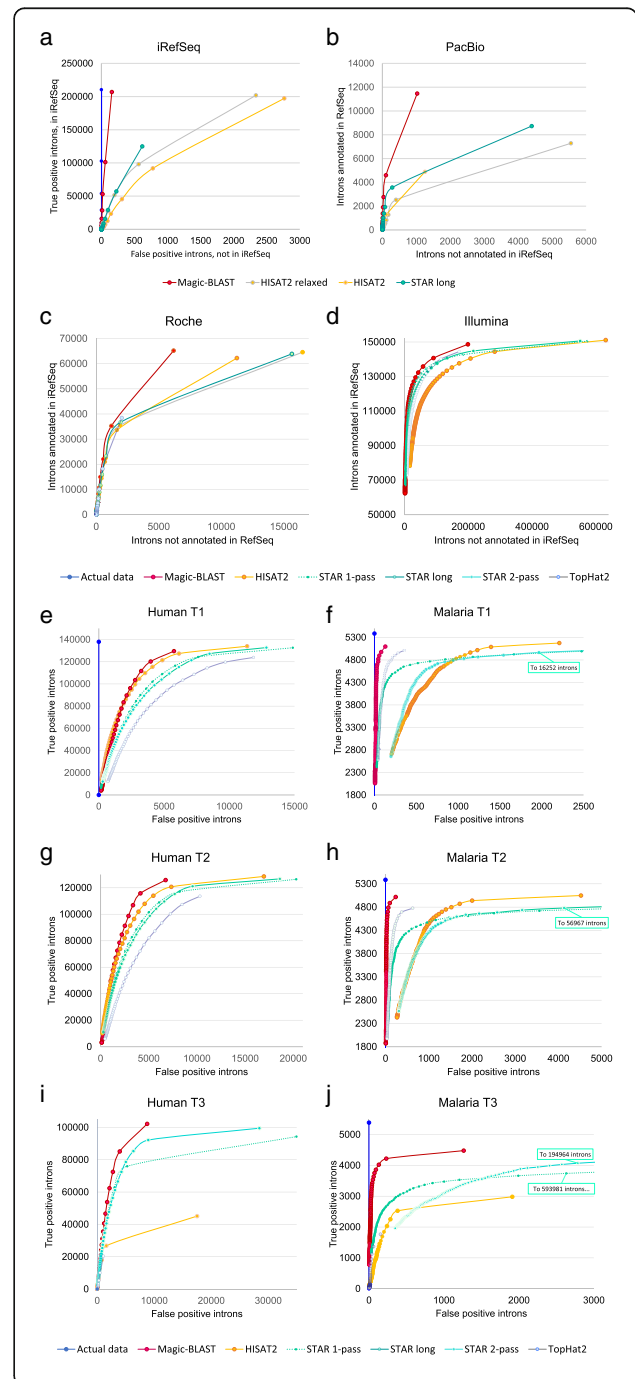
(See figure on previous page.)

Fig. 4 ROC curves showing intron discovery as a function of minimal read coverage, from 1 to 100 (or to the maximal observed coverage). The point with coverage 1, corresponding to all introns found in the experiment, is the point closest to the top right corner. The plots show, for each minimal coverage, the true positives on the Y-axis and the false positive on the X-axis. In the 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 benchmark sets, iRefSeq and six Baruzzo, have a built-in truth (vertical blue line in some graphs). **a** For the iRefSeq set, because of the alternative splice variants, the truth has introns supported by 1, 2, and up to 51 RefSeqs, and Magic-BLAST (red) follows the truth remarkably closely, point by point. It finds slightly more true-positive introns than the HISAT2 programs, but the biggest difference is that HISAT2 finds 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 it finds the fewest unannotated introns. **c** The Roche presents a similar, though less extreme, result. **d** In Illumina (zoomed in Additional file 1: Figure S3.1), Magic-BLAST followed by STAR 1-pass have the steepest slopes. Then come STAR 2-pass and TopHat2, then HISAT2; these last three aligners call unannotated introns at high coverage. **e-j)** In the Baruzzo shallow human T1 (**e**) and T2 (**g**) benchmarks, Magic-BLAST then HISAT2 perform the best, followed by STAR 1- then 2-pass, then TopHat2. **i** In human T3, HISAT2 and TopHat2 drop considerably, only Magic-BLAST and STAR can find introns in the presence of a high level of mismatches. **f**, **h**, **j** In the ultra-deep malaria sets, Magic-BLAST remains best, STAR 2-pass and HISAT2 drop below TopHat2. At coverage 1, STAR has by far the largest number of false positives (Additional file 1: Figures S3)

how the different programs behave when the support, given by the number of reads mapped to each junction, increases. The truth for the benchmark sets is vertical and dark blue. 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 fifteen cases, Magic-BLAST (red) is to the left and above all other curves and qualifies as the best intron finder in all conditions tested, for reads from 100 bases to 100 kilobases, with any level of mismatches, from perfect match to 8.6% mismatch. This observation applies to benchmark as well as all seven real data sets tested from PacBio, Roche or Illumina.

This observation can be summarized by quantifying intron precision, recall, and F-score for the iRefSeq and experimental sets (Fig. 5a-c and Additional file 1: Figure S4). 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 101 + 101 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 genes and introns, this explains the recall around 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 finds 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 (Additional file 1: sections 3 and 4). At the best F-score, Magic-BLAST has the highest F-score in 15 experiments, plus the 12 duplicate benchmarks. The only exception is human T1 truncated at 50 + 50, where HISAT2 takes the lead. Magic-BLAST also reaches its best score at the lowest coverage of all aligners in almost all cases. 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 (Fig. 4, Additional file 1: section 3). In the deep malaria set, which has about 5500 annotated introns, STAR produces up to 600,000 false positive introns (Fig. 4j). This greedy intron-finding behavior fits with the observation that introns and splice sites found by STAR and HISAT2 at low coverage are mainly untrustworthy. It is worth noting that, as judged by the ROC curves, HISAT2 produces much worse results in the very deep malaria T1 than in the shallow human T1 (Fig. 4e and f), and that STAR 2-pass can produce worse results than the 1-pass version. This is especially apparent for deep sets such as the experimental Illumina and the Baruzzo malaria (Fig. 4f, h, j, Additional file 1: Figure S3.1). This hypothesis was tested and confirmed by subsampling 1% of malaria T1: STAR 2-pass became just as good as STAR 1-pass (compare Fig. 4f and Additional file 1: Figure S3.3) and the false positive introns went away much faster than the true positives in STAR and HISAT2.

Fig. 5 Performance of the aligners on intron discovery and alignment statistics measured on iRefSeq and seven experimental datasets. **a** Intron discovery precision p = (TP/TP + FP). **b** recall r = (TP/TP + FN) and **c**) F-score (2pr/p + r). All introns are counted, including those with a single support. An intron annotated in iRefSeq is counted as true, an unannotated intron as false. **d** percentage of reads aligned. **e** percentage of bases aligned. Average sequenced length per fragment are: iRefSeq 3427 bases, PacBio 1854 bases, Roche 348 bases, Illumina 202, 500 and 600 bases, PacBio 1300 and 1323 bases. The longer the reads, the more specific their alignments are

Figure 6g presents intron discovery F-scores for the Baruzzo set. Magic-BLAST has the best F-scores for all the sets, but really excels for the T3 human set and all the malaria sets.

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 alignment to the 'true' position of the read and partition all alignments 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). This provides a direct measure of alignment precision, recall and F-score (Figs 6c and 7a-c, Additional file 1: Figure S6.1). These categories do not treat true and false positives symmetrically, since a missing exon would be True positive type 2 but an extra exon would be a False positive. We take this cautious approach since we consider an incorrect exon a more serious issue than a missing exon.

In all cases, even when no truth is available, we characterized the quality of the alignments by using the alignment statistics derived from the AliQC program (Additional file 1: S1.3), which includes the number of

Fig. 6 Performance of the aligners and quality control characteristics measured on the Baruzzo Illumina-like benchmark.

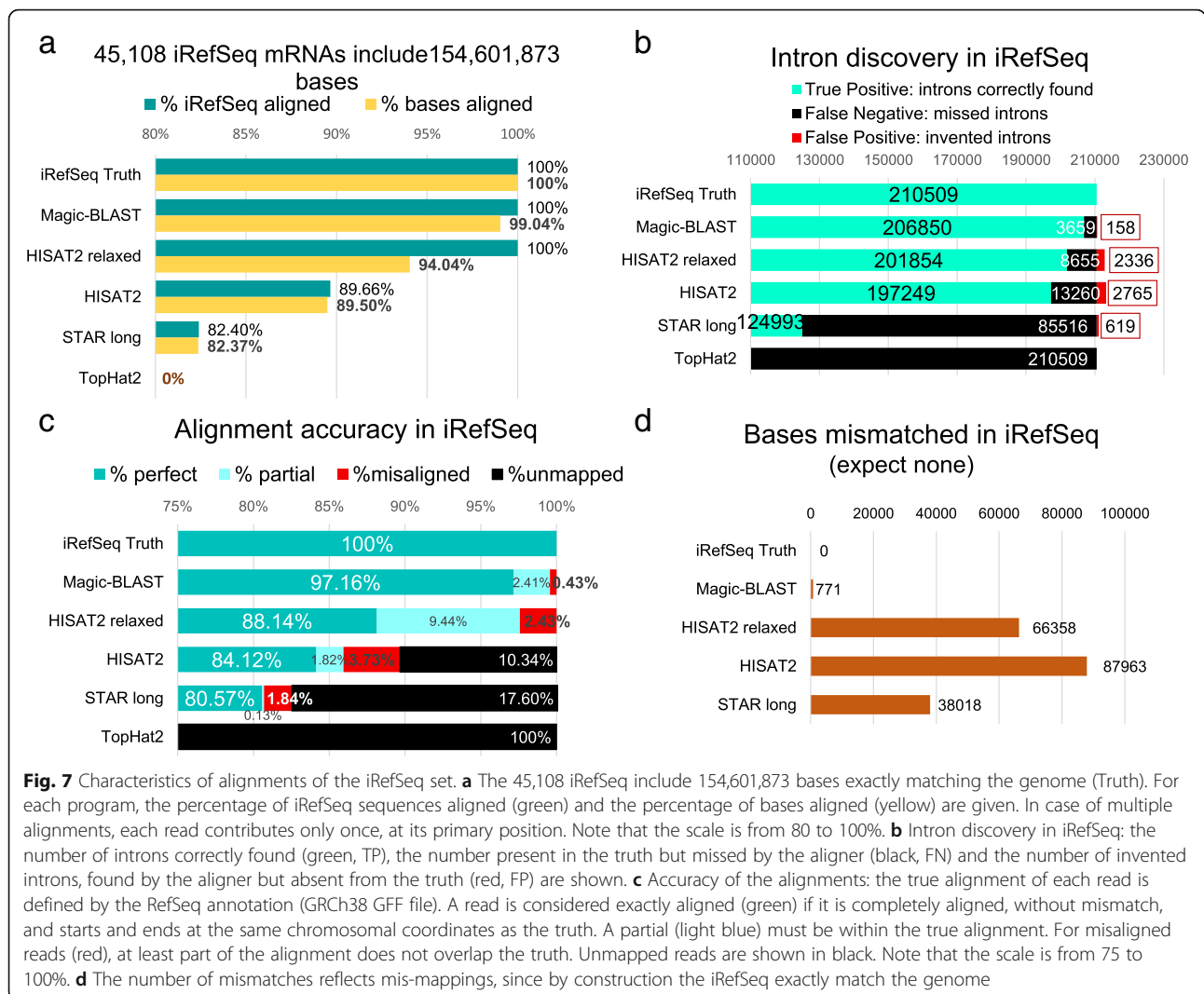| Dataset / Aligner	a Alignment precision	b Alignment recall	c % reads aligned	d % bases aligned	e % unique	f Mismatches per kilobase aligned	g Intron discovery F-score
Human T1 (truth f)						5.43	
Magic-BLAST	95.1%	99.3%	99.3	98.5	98.0	5.28	94.8%
HISAT2	96.4%	99.3%	99.3	99.2	97.8	5.26	94.6%
STAR 1-pass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

First, the results of the comparison of all alignments to the truth are shown: **a** alignment precision, **b** alignment recall, counting as true an alignment which is identical or included in the truth (perfect or partial); each alignment for multiply aligned reads is counted. Second, the statistics of the alignment independently of the truth are shown: **c** percentage of reads aligned. **d** percentage of bases aligned. Note that the aligners differ on their view of what can be reported as an 'alignment' and on their thresholds for minimal aligned length and maximal level of mismatches. **e** Percentage of reads aligned uniquely. A higher rate of unique alignments is desirable, but the true number of ambiguous reads is unknown. **f** The number of mismatches per kilobase aligned. The true rate (first line of each dataset) was measured using the official benchmark mapping. A rate below the truth, correlated with lower alignment recall, characterizes an aligner which cannot deal with high levels of mismatches. **g** the intron discovery F-score, measured for all introns (coverage at least 1). Magic-BLAST reaches the highest F-score for intron discovery in all datasets

reads and bases aligned, unique versus multiple alignments, aligned length histograms, mismatches per read, type and position (Additional file 1: section 7 presents more information on the topics covered in this paragraph).

A simple and rich summary of the quality of the long reads alignments is provided in Fig. 8 (and Additional file 1: Figures S2), which shows for each aligner the histogram of aligned length as compared to the length of the reads. The best performance an aligner could possibly achieve would be to map each read along its entire length, so ideally the histogram of aligned length should be superimposed on the histogram of the read lengths (blue), especially for the longest reads (to the right). In the iRefSeq case of very long transcripts perfectly matching the genome (Fig. 8a), TopHat2 fails, Magic-BLAST matches the read length histogram all the way from long to short transcripts. HISAT2 relaxed is better than HISAT2 and nearly as good as Magic-BLAST, but with an elbow of very short alignments. STAR long is

distinctly lower over the longest reads. The situation is different for PacBio (Fig. 8b) where the bulk of the sequences are between 1000 and 2700 bases long and show a high rate of sequencing errors, close to 20 mismatches per kb. There, STAR long finds some of the longest alignments but fails to align close to 30% of the reads. Magic-BLAST maps all reads, but often as partials because if an alignment presents a gap, Magic-BLAST reports only the highest scoring partial alignment. The curves for HISAT2 and HISAT2 relaxed are much lower, especially for long reads. In addition, HISAT2 relaxed creates a very large number of alignments shorter than 200 bases. Again, TopHat2 fails. The Roche reads are shorter, between 33 and 808 bases (Fig. 8c), and some are now in the range acceptable to TopHat2, which aligns about half of the reads up to 600 bases. Alignment lengths for STAR long and Magic-BLAST are close to read lengths. The curve for HISAT2 is a little below. Both Magic-BLAST and HISAT2 relaxed produce a small number of shorter alignments, as is evident from

Fig. 7 Characteristics of alignments of the iRefSeq set. **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. Note that the scale is from 80 to 100%. **b** Intron discovery in iRefSeq: the number of introns correctly found (green, TP), the number present in the truth but missed by the aligner (black, FN) and the number of invented introns, found by the aligner but absent from the truth (red, FP) are shown. **c** 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 (green) if it is completely aligned, without mismatch, and starts and ends at the same chromosomal coordinates as the truth. A partial (light blue) must be within the true alignment. For misaligned reads (red), at least part of the alignment does not overlap the truth. Unmapped reads are shown in black. Note that the scale is from 75 to 100%. **d** The number of mismatches reflects mis-mappings, since by construction the iRefSeq exactly match the genome

the peaks on the very left end of the graph. Magic-BLAST has the fewest complete alignments, but the situation is reversed for the longer Illumina reads 250 + 250 and 300 + 300 (Additional file 1: Figure S2.2) where Magic-BLAST aligns 50% more bases than STAR and twice as many as HISAT2.

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 (Fig. 7, Additional file 1: section 5). Magic-BLAST, HISAT2, HISAT2 relaxed, STAR long, and Minimap2 produced results for this experiment. There is no bias, as Magic-BLAST was not used to prepare the RefSeq annotation.

Figure 7a displays the percentage of reads and bases aligned. Magic-BLAST performs the best. Similarly, at

the intron level (Fig. 7b), 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 (false negative) by a wide margin while Magic-BLAST misses the least.

Measuring the mapping accuracy by comparison to the iRefSeq annotation taken as the truth (Fig. 7c) shows that 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 misaligns 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 misalignments are subtle, affecting just one or a few exons, but
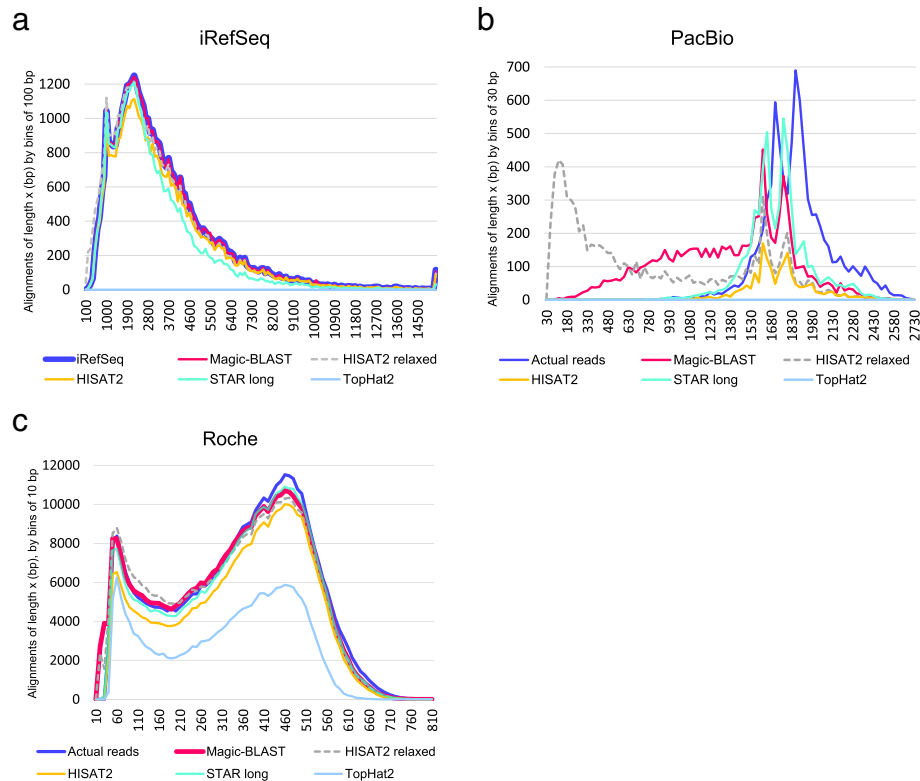
Fig. 8 Histograms of read and alignment lengths for iRefSeq (**a**), PacBio (**b**), and Roche 454 (**c**). The X-axis shows the length in bases, the Y-axis the number of reads in each length bin, in the original dataset (blue) or after alignment. The counts are binned every 100 bases for iRefSeq, every 30 bases for PacBio, and every 10 bases for Roche

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 true position. 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 lead to incorrectly intertwined neighboring genes. We also performed some preliminary analysis of Minimap2 and Subread with respect to the iRefSeq mapping accuracy (Additional file 1: Table S6). Subread was only able to align 533 of the 45,108 reads. Minimap2 performed comparably to Magic-BLAST on this criteria.

The iRefSeq mRNAs should match the genome exactly, and the number of mismatches shows how well each aligner performs (Fig. 7d). The truth has no mismatch. Magic-BLAST has 771 mismatches, STAR long and HISAT2 relaxed 38018 and 66358 mismatches respectively, and STAR long aligns many fewer bases. On

iRefSeq, Magic-BLAST alignments are superior to both HISAT2 and STAR long by all criteria.

We then assessed the accuracy of the alignments and measured the mapping statistics using 100 + 100 base Illumina-like simulated RNA-seq data from the Baruzzo benchmark [6] (Fig. 6). The first two columns (6a and 6b) use the built-in truth to measure the precision and recall of the mapping: an alignment is counted as true if it overlaps the benchmark position, even if it is partial; otherwise it is counted as false (Additional file 1: section 6). In case of multiply aligned reads (Fig. 6e), each alignment contributes. The human T1 set should be the easiest. STAR does the best, followed by HISAT, and then Magic-BLAST: their precision is above 95% and recall above 99%. TopHat2 is noticeably lower. 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, STAR shows a significant degradation of the recall, and HISAT2 and TopHat2 perform very poorly. The percentage of reads and bases aligned tells a similar story (Fig. 6c, d), with results degrading from T1 to T2 to T3, especially for HISAT2 and TopHat2 probably because the rate of mismatches in the benchmark, 60 per kb in Human T3 (Fig. 6f), exceeds the thresholds of these programs.

The Plasmodium (malaria) benchmark sets should be more challenging owing to the AT-rich strong bias in the composition of the genome; also, the coverage is 100-fold deeper. The results are similar to the human benchmark; the main difference is that STAR is very slow and inefficient on malaria and produces huge numbers of false positive introns (Additional file 1: Figure S3) and Magic-BLAST now has the best alignment precision and recall for T1, T2 and T3 (Fig. 6).

Finally, we examine the ability of the programs to align experimental RNA-seq reads generated with different technologies. We do not have a ground truth for these sets, but, as mentioned above, we may use alignment statistics to judge quality. Figure 5d-e shows the percentage of reads and bases aligned for the experimental sets and iRefSeq. Focusing on the percentage of bases aligned, Magic-BLAST, HISAT2, and STAR all do relatively well for Illumina 101 + 101 and Roche, with STAR aligning the most reads on the short Illumina set and Magic-BLAST aligning the most on the Roche set. For longer Illumina reads (250 + 250 and 300 + 300 bases), Magic-BLAST aligns 50% more bases than STAR long or STAR and 70% more than HISAT2. Magic-BLAST also aligns the most bases on the other longer sets.

Additionally, we performed some preliminary tests on Minimap2 and Subread shown in Additional file 1: Table S7. We briefly discuss the matching and aligned bases columns from that table here, concentrating on Minimap2, Subread, and Magic-BLAST. For the iRefSeq sets, both Magic-BLAST and Minimap2 align more than 99% of the bases, though Minimap2 does slightly better. Subread is only able to align around 22% of the bases, performing worse than STAR long or HISAT2 relaxed. In general, the percentage of bases aligned by Subread is much lower than Minimap2 or Magic-BLAST. For PacBio (SRR5009494), Minimap2 aligns substantially more bases (96.5%) than Magic-BLAST (73.5%). For the two hybrid PacBio runs (SRR5189652 and SRR5189667), both Minimap2 and Magic-BLAST can align 70–78% of the bases, though Minimap2 aligns more in both cases. For the Illumina 300 + 300 (SRR5437876) and 250 + 250 (SRR5438850), Magic-BLAST and Minimap2 both align around 74–77% of the bases, with Magic-BLAST aligning slightly more. In experiments with truncated versions of the Illumina 250 + 250 set (with lengths 150 + 150, 120 + 120, 100 + 100, 75 + 75, and 50 + 50), Magic-BLAST consistently outperforms Minimap2. Minimap2 returned no results for the 75 + 75 and 50 + 50 truncated sets. Minimap2 did not produce proper SAM output for paired-reads.

It is also worthwhile to examine the percentages of bases aligned for all aligners with the truncated Illumina 250 + 250 reads (Additional file 1: Table S7). Most aligners were able t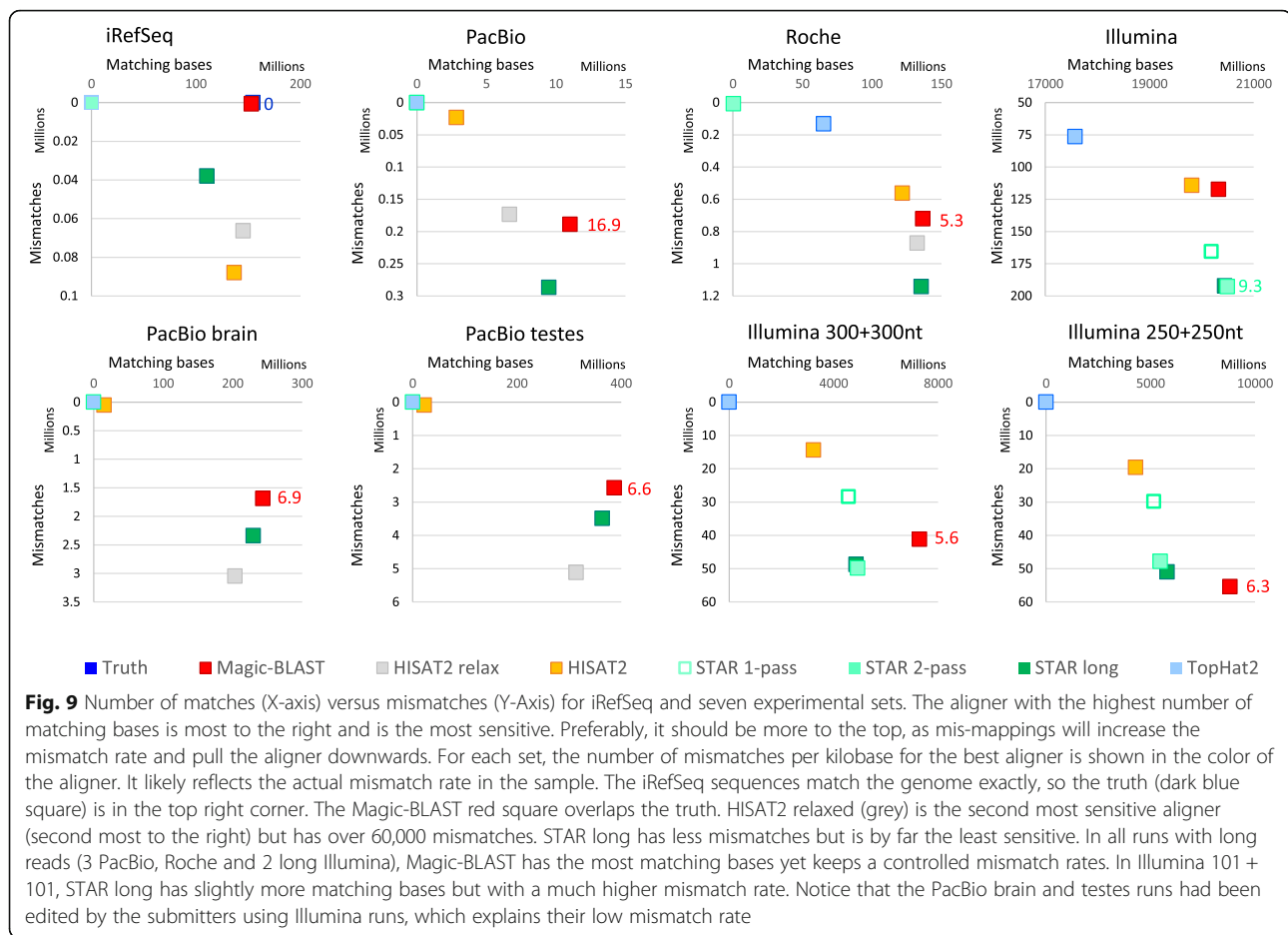o align around 95% of the shorter reads with Magic-BLAST and Minimap2 aligning relatively more bases for the longer reads.

We can combine information about the number of bases aligned with the mismatch rate. The mismatch rate depends on the sample and sequencing quality, its true level is unknown for experimental sets, but mismapping by the aligner will tend to increase the apparent rate. We assume that the lower the mismatch rate, the better. In Fig. 9, we plot the number of matches against the number of mismatches summed over all 'primary' alignments (maximum one alignment per read). The most efficient aligner will align the largest number of bases – hence be most to the right - yet limit its mismatch rate to a minimum – hence be higher than other aligners with a similar number of matching bases. We indicate next to the aligner that maps the most bases, and in the same color, the number of mismatches per kilobase, which is our best estimate of the true level of mismatches in the sample (Fig. 9 and Additional file 1: Figure S7.1). Again, it is apparent that Magic-BLAST (red) has the largest number of matching bases in all cases where read length is larger than 101 bases. Additionally, Magic-BLAST has either the lowest mismatch rate, or clearly maps more bases. For Illumina 101 + 101, STAR 2-pass (green) aligns 1% more matching bases, but with a very high level of mismatches by Illumina standards, 9.3 per kb in STAR to be compared to 5.7 in Magic-BLAST. In general, the STAR family of programs tends to align with a much higher level of mismatches than Magic-BLAST, and as this is often coupled with a lower number of matches, one may suspect more suboptimal mappings in STAR compared to Magic-BLAST. TopHat2 (light blue), and to a lesser extent HISAT2 (yellow) allow fewer mismatches hence stay near the left top corner.

Additional file 1: sections 6 and 7 provide more details on all these aspects.

Run times

Figure 10 presents the CPU times and peak memory requirements for the alignments. A priori, the comparison seems straightforward, however, in presenting a fair assessment, the quality of the results matters. The table is colour coded by the percentage of matching bases relative to the best aligner for the dataset. For example, TopHat2 fails on PacBio (red) and only marginally succeeds on Roche (beige, 30 to 50% of top). Similarly, HISAT2 nearly fails on PacBio and Malaria T3 (brown, 5 to 10% of top). In such conditions, the speed of the code is less relevant. Looking at the colours in the table, Magic-BLAST is most often in the top aligner category (within 1% in 13 of 15 cases); it is less than 3% below the top in the two remaining cases, human benchmark T1 100 + 100 (− 1.2%), and T2 (− 1.1%). The second best
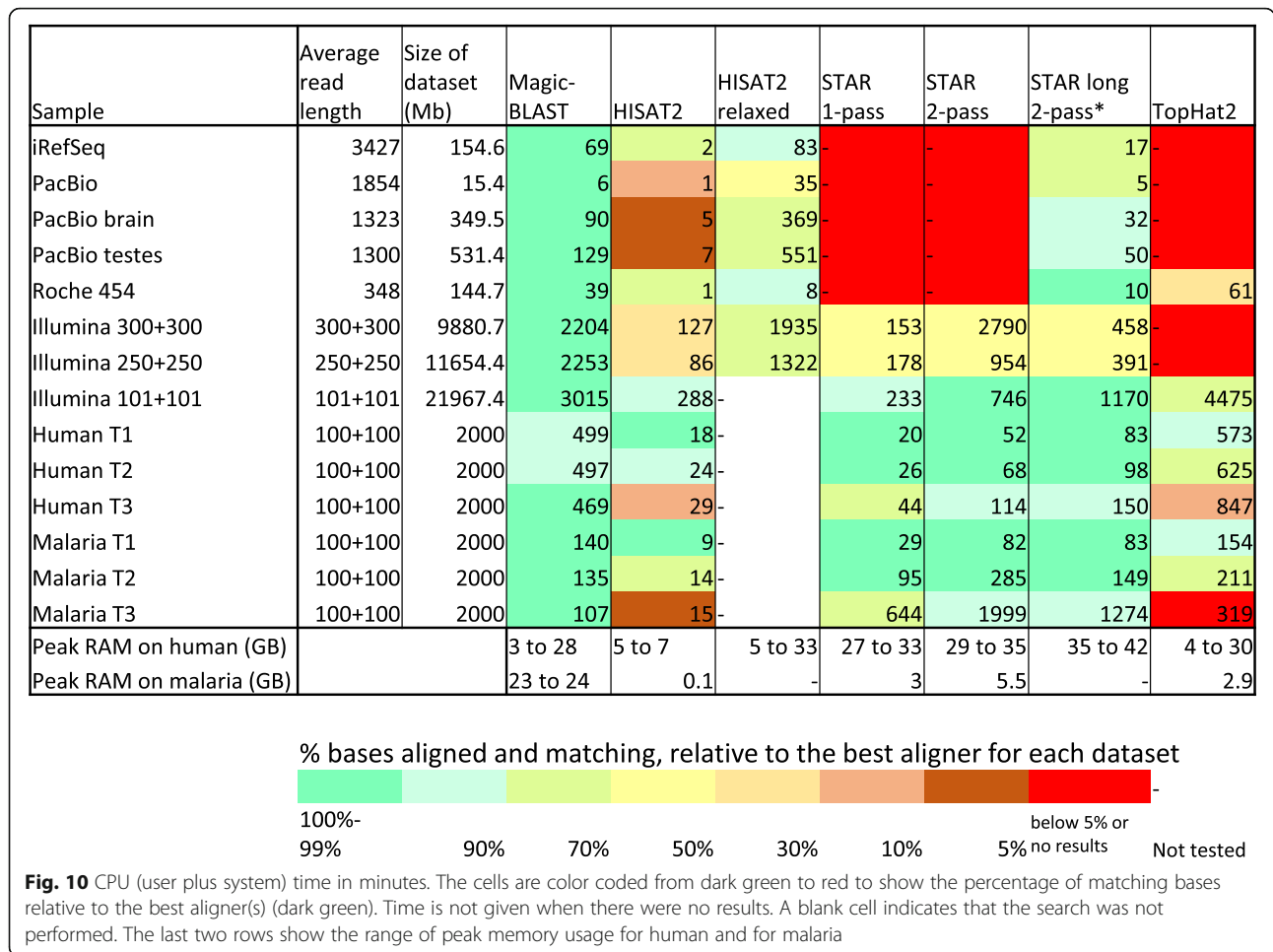
Fig. 9 Number of matches (X-axis) versus mismatches (Y-Axis) for iRefSeq and seven experimental sets. The aligner with the highest number of matching bases is most to the right and is the most sensitive. Preferably, it should be more to the top, as mis-mappings will increase the mismatch rate and pull the aligner downwards. For each set, the number of mismatches per kilobase for the best aligner is shown in the color of the aligner. It likely reflects the actual mismatch rate in the sample. The iRefSeq sequences match the genome exactly, so the truth (dark blue square) is in the top right corner. The Magic-BLAST red square overlaps the truth. HISAT2 relaxed (grey) is the second most sensitive aligner (second most to the right) but has over 60,000 mismatches. STAR long has less mismatches but is by far the least sensitive. In all runs with long reads (3 PacBio, Roche and 2 long Illumina), Magic-BLAST has the most matching bases yet keeps a controlled mismatch rates. In Illumina 101 + 101, STAR long has slightly more matching bases but with a much higher mismatch rate. Notice that the PacBio brain and testes runs had been edited by the submitters using Illumina runs, which explains their low mismatch rate

aligner overall is STAR long 2-pass, which, as we discovered, can be used on any read length, but is slower on short reads than STAR 2-pass and less efficient than Magic-BLAST for long reads, from Illumina 250 + 250 paired ends to PacBio and iRefSeq. HISAT2 and Tophat2 rarely achieve top category (3 cases for HISAT2, none for HISAT2 relaxed, and none for TopHat2).

The memory requirement is also important (Additional file 1: Table S1.2). HISAT2 uses by far the least memory (maximum 7 GB of memory in human, 0.1 GB in malaria) while STAR is the most greedy (27 to 42 GB). As a result, on an 8 core machines with 64 GB of RAM, one can run eight human samples in parallel using HISAT2, but only one using STAR long. Magic-BLAST uses slightly less memory than STAR on human but requires as much memory when aligning on malaria than on human. TopHat2, like HISAT2, generally uses little memory (less than 5 GB).

Consider now the speed of the programs and let us limit the comparison to good aligners (> 90% of the top aligner, green and light green classes). On the benchmark human T1 or T2 reads, 100 + 100, STAR 2-pass or STAR long (and HISAT2 for T1) are the top aligners

and the fastest codes, an order of magnitude faster than Magic-BLAST in human and three times faster in malaria. For the T3 cases which mimics cross-species mapping, Magic-BLAST is the only top aligner. Compared to STAR 2-pass, it is 3 times slower on human T3 and 10 times faster in malaria T3. STAR on malaria T3 consumed 24 times more CPU than on malaria T1 and 17 times more than on human T3. On the experimental Illumina 101 + 101, Magic-BLAST and STAR 2-pass are top aligners, and STAR is 2.5 times faster than Magic-BLAST. On datasets with longer reads, iRefSeq, PacBio, and Illumina 250 or 300, Magic-BLAST is the only top aligner, it is three to seven times slower than STAR long and one to three times faster than HISAT2 relaxed.

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. Additional file 1: section 1.1 discusses how to control the memory requirements when treating very large datasets with Magic-BLAST.

Sample	Average read length	Size of dataset (Mb)	Magic-BLAST	HISAT2	HISAT2 relaxed	STAR 1-pass	STAR 2-pass	STAR long 2-pass*	TopHat2
iRefSeq	3427	154.6	69	2	83	-	-	17	-
PacBio	1854	15.4	6	1	35	-	-	5	-
PacBio brain	1323	349.5	90	5	369	-	-	32	-
PacBio testes	1300	531.4	129	7	551	-	-	50	-
Roche 454	348	144.7	39	1	8	-	-	10	61
Illumina 300+300	300+300	9880.7	2204	127	1935	153	2790	458	-
Illumina 250+250	250+250	11654.0	2253	86	1322	178	954	391	-
Illumina 101+101	101+101	21967.4	3015	288	-	233	746	1170	4475
Human T1	100+100	2000	499	18	-	20	52	83	573
Human T2	100+100	2000	497	24	-	26	68	98	625
Human T3	100+100	2000	469	29	-	44	114	150	847
Malaria T1	100+100	2000	140	9	-	29	82	83	154
Malaria T2	100+100	2000	135	14	-	95	285	149	211
Malaria T3	100+100	2000	107	15	-	644	1999	1274	319
Peak RAM on human (GB)			3 to 28	5 to 7	5 to 33	27 to 33	29 to 35	35 to 42	4 to 30
Peak RAM on malaria (GB)			23 to 24	0.1	-	3	5.5	-	2.9

% bases aligned and matching, relative to the best aligner for each dataset

| 100%-99% | 90% | 70% | 50% | 30% | 10% | 5% | below 5% or no results | - Not tested |

Fig. 10 CPU (user plus system) time in minutes. The cells are color coded from dark green to red to show the percentage of matching bases relative to the best aligner(s) (dark green). Time is not given when there were no results. A blank cell indicates that the search was not performed. The last two rows show the range of peak memory usage for human and for malaria

Discussion

We have examined the performance of four aligners (Magic-BLAST, HISAT2, STAR, and TopHat2) with a wide variety of read lengths between 100 bases and 100 kilobases. First, we examined the performance of all programs with several experimental test sets from different platforms, 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 short 100 + 100 base Illumina type reads, for human as well as a 100-fold deeper malaria runs, with three levels of mismatches. Owing to time constraints, we did not try longer simulated read sets. The aligners have different strengths and weaknesses, which reflect in part the strategic choices of the authors (e.g. favoring complete alignments or limiting the number of mismatches per read) and the characteristics of the implementations (e.g. second pass intron validation). Additionally, we looked at a more limited set of tests for two aligners, Minimap2 and Subread, optimized for long reads.

Magic-BLAST works for all datasets, produces good results on introns, long read alignments, and high-levels of mismatches, and is stable. It tries to align all reads, preferring a partial alignment to an unmapped read. As a consequence, some short alignments may be reported. For reads longer than 100 bases, it is a winning strategy. It aligns more bases and with less mismatches, and reliably discovers more introns. It is more exhaustive and more precise.

As demonstrated by the iRefSeq set, only Magic-BLAST, HISAT2 with non-default parameters, STAR long and Minimap2 could align very long sequences, even if there were no mismatches. Magic-BLAST, HISAT2 relaxed, and Minimap2 were able to align all reads in the set while STAR long aligned 82%. In terms of bases aligned, Magic-BLAST and Minimap2 aligned around 99% of the bases, while HISAT2 relaxed and STAR long aligned 94 and 82% respectively. While we expect zero mismatches, Magic-BLAST had

68 to 130 times less mismatches than HISAT2 relaxed and STAR long.

The same general trends were found for the long experimental runs. For one PacBio run, Minimap2 aligned 96.5% of the bases while Magic-BLAST aligned 73.5%, but Minimap2 and Magic-BLAST aligned comparable numbers of bases (70–77%) for the two hybrid PacBio runs and the long Illumina runs. The next best aligner aligned 65 to 68% of the bases. Minimap2 did not perform as well for shorter reads and did not report which alignments belonged to paired reads. Subread did not perform well for the longest reads.

We truncated the Illumina 250 + 250 run to a variety of shorter lengths (down to 50 bases) and found that most aligners could align most of the bases in the truncated runs, though Minimap2 was unable to align the 50 and 75 base reads at all.

Intron discovery posed different challenges for the aligners. We looked at ROC curves, using the provided results for the benchmark sets and the annotations on the human genome as a guide for the experimental sets. 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 remark explains 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. It is clear from the ROC curves that our cautious approach to intron discovery pays dividends: Magic-BLAST is the best intron finder in all datasets, with far fewer false positives produced for a given number of true positives. The intron discovery F-score tells a similar story. Magic-BLAST has the best results in all datasets and its results are trustworthy, even at very low coverage. For the T1 or T2 human Illumina type reads with few mismatches, the difference in ROC curves between Magic-BLAST and HISAT2 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, PacBio, Roche and long Illumina sequences, Magic-BLAST produced the best introns 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 malaria sets or the Illumina 101 + 101, even though STAR 2-pass is expected to improve upon STAR 1-pass. For the Illumina run (zoom in Additional file 1: Figure S3.1), 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 than STAR 2-pass 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 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 than the deep malaria T1. The optimal F-scores (Additional file 1: Figure S4.1), calculated for the different aligners and experiments, are consistent with the ROC-curves in this regard. We have shown by subsampling that in both aligners, the second pass reinforces the false discoveries of the greedy first pass, and this tendency is also visible in all PacBio runs and Roche. Sadly, this type of noise cannot be erased, and if used in a gene reconstruction project, as was done in [20], 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.

In Fig. 10, we presented the run times for Magic-BLAST, HISAT2, STAR and TopHat2 and their relative performance on each dataset. For the longer sequences (first seven rows), from Illumina 250 + 250 to iRefSeq, Magic-BLAST was consistently in the best category. In some cases, Magic-BLAST could be 3–7 times slower than the next best aligner (STAR long or HISAT2 relaxed), but the quality of the results, both in terms of the alignments and intron finding, made it the best choice. For shorter sequences, such as Illumina 101 + 101 or the Baruzzo 100 base T1 and T2 benchmark sets, most aligners performed relatively well. The other aligners did not perform as well for the more distant T3 Baruzzo sets and STAR slowed down dramatically for the T3 malaria case, becoming 10 times slower than Magic-BLAST.

For the shorter Illumina reads or the 100 + 100 simulated datasets with few mismatches (T1 or T2), other programs produced alignments comparable to or better than those of Magic-BLAST. Here, STAR 2-pass aligned the largest number of bases and reads and maintained an excellent alignment F-score. HISAT2 was memory efficient and did rather well on short reads with few mismatches. We also truncated the human T1 sequences to 50 bases and found that Magic-BLAST produced a lower intron discovery F-score than the other aligners (Additional file 1: Figure S4.3). For sequences that are 100 bases or shorter, similar to the target, most other aligners were faster than Magic-BLAST and therefore a reasonable choice even though Magic-BLAST would also produce reasonable results for these shorter reads.

The Magic-BLAST procedure is different than the other aligners (HISAT2, STAR, and TopHat2) in a number of ways. Magic-BLAST creates an in-memory index

of reads and scans the target genome, where the target genome can be in the form of a BLAST database, FASTA file or NCBI accession. The other programs create an index of the genome. It has a new alignment extension algorithm, based on a simple state machine, where each alignment operation requires a specific number of matching bases to follow it in order to succeed. The extensions stop quickly when read and target sequences diverge, which prevents extending an alignment into an intron or an adapter. This procedure provides a good tradeoff between the expensive dynamic-programming-based optimal alignment methods in STAR and TopHat2 and heuristic approach used in HISAT2, which is fast but locked to specific read lengths and error levels. Magic-BLAST does not penalize splice sites when computing alignment scores, whereas HISAT2, STAR, and TopHat2 use a splice penalty. Its alignment scoring system is universal with respect to introns. This means that Magic-BLAST produces the same score for either the spliced alignment of a read to a genomic region or the corresponding pseudogene or transcript, facilitating score-by-score comparisons of these alignments. Magic-BLAST also identifies introns based on high quality alignments and the presence of known splice signals, instead of using a splice site as an alignment operation when computing an optimal alignment. This approach can reduce the number of false positive introns caused by misaligned read segments or too many errors in the read.

Magic-BLAST produced reliable results for a wide variety of sequence lengths, from 100 bases to 100 kilobases, compositions, or error rates without changes to the command-line whereas the standard options of other programs are often suboptimal [6]. We also field-tested Magic-BLAST in several hackathons, allowing us to identify problems, hear user suggestions, and improve usability.

It is finally 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 Fig. 11. A spliced aligner, like Magic-BLAST, aligns this read only on chromosome 17 and 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 that compresses sequences 4-to-1, saving disk space and 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 -seqidlist option to map sequences only to selected reference sequences in a larger BLAST database. There is a caveat though: not mapping the reads competitively on the entire genome will gather all the reads coming from the selected target, but also some from its cognate genes, homologous sequences and eventual pseudogenes.

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, but 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

We presented Magic-BLAST, a new tool for mapping next generation RNA-seq runs with read length between 100 nucleotides and 100 kilobases against a genome or a transcriptome. Its performance was compared with that of similar popular programs: HISAT2, STAR, and TopHat2. Magic-BLAST was the best intron finder on all tested sets, real or simulated, and the good aligner for all long reads, including Illumina 250 + 250 or 300 + 300, Roche 454, PacBio and full-length mRNA sequences. We also performed some preliminary comparisons with Minimap2 and Subread.

Magic-BLAST integrates very well with other NCBI tools and services and is convenient to use since it

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      CCGCTACAgggggggCCTGAGGCACTGCAGAAAGTGGGCCTGAGCCTCGAGGATGACGGT  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
```

Fig. 11 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 and is not aware of paired end sequencing

recognizes NCBI accessions for SRA runs, mRNA or genomic sequences, and uses BLAST databases.

We are exploring ways to improve Magic-BLAST, such as exporting lists of introns and support, improving adapter detection, shortening the required exon length, and identifying repeats. Additionally, we are examining how to eliminate very short alignments, how to further extend alignments for long reads, and improve the speed of Magic-BLAST. We are also working closely with users to address their needs.

Additional file

Additional file 1: The supplement consists of a document with seven sections, successively 1- Bioinformatics (aligners and analysis programs), 2- datasets and length histograms, 3- introns discovered per coverage (ROC curves), 4- introns precision and recall, 5- iRefSeq results, 6- alignment mapping precision and recall measured by comparison to the benchmark, 7- alignment statistics and quality control. Additionally, a number of excel spread sheets with underlying data have been provided. All comments are welcome. (ZIP 3341 kb)

Abbreviations
BAM: Binary Alignment Map (binary version of SAM); BLAST: Basic Local Alignment Search Tool; CIGAR: Concise Idiosyncratic Gapped Alignment Report; DNA-seq: DNA sequencing; FN: False Negative; FP: False Positive; GB: Gigabyte; HISAT: Hierarchical Indexing for Spliced Alignment of Transcripts; NCBI: National Center for Biotechnology Information; RNA-Seq: RNA sequencing; ROC: Receiver Operating Characteristics; SAM: Sequence Alignment Map; SNP: Single Nucleotide Polymorphism; SNV: Single Nucleotide Variant; SRA: Sequence Read Archive; STAR: Spliced Transcripts Alignment to a Reference; TP: True Positive

Availability and requirements
Project name: Magic-BLAST

Authors' contributions

Funding

Availability of data and materials

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 and described in the supplementary material section 1. All data come from external sources and can be downloaded by the master script, together with all our alignment SAM files (ftp://ftp.ncbi.nlm.nih.gov/blast/demo/magicblast_article).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Sahraeian SME, Mohiyuddin M, Sebra R, Tilgner H, Afshar PT, Au KF, Bani Asadi N, Gerstein MB, Wong WH, Snyder MP, et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat Commun. 2017;8(1):59.
2. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621–8.
3. Mielczarek M, Szyda J. Review of alignment and SNP calling algorithms for next-generation sequencing data. J Appl Genet. 2016;57(1):71–9.
4. Reinert K, Langmead B, Weese D, Evers DJ. Alignment of next-generation sequencing reads. Annu Rev Genomics Hum Genet. 2015;16:133–51.
5. Križanovic K, Echchiki A, Roux J, Šikic M. Evaluation of tools for long read RNA-seq splice-aware alignment. Bioinformatics. 2018;34(5):748–54.
6. Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat Methods. 2017;14(2):135–9.
7. Zhang W, Yu Y, Hertwig F, Thierry-Mieg J, Thierry-Mieg D, Wang J, Furlanello C, Devanarayan V, Cheng J, Deng Y, et al. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome Biol. 2015;16:133.
8. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
9. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
10. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinf. 2009;10:421.
11. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
12. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
13. Dobin A, Gingeras TR. Optimizing RNA-Seq mapping with STAR. Methods Mol Biol. 2016;1415:245–62.
14. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11.
15. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36.
16. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45.
17. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
18. Liao Y, Smyth GK, Shi W. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 2013;41(10):e108.
19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
20. Lagarde J, Uszczynska-Ratajczak B, Carbonell S, PÈrez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R, et al. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. Nat Genet. 2017;49(12):1731–40.
21. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nat Biotechnol. 2014;32(9):903–14.
22. Gribskov M, Robinson NL. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput Chem. 1996;20(1):25–33.
23. PUBLIC DOMAIN NOTICE for NCBI: https://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/lxr/source/doc/public/LICENSE. Accessed 13 June 2019.

Publisher's Note