ExtendAlign: a computational algorithm for delivering multiple, local end-to-end alignments for short sequences

Mariana Flores-Torres,[1] Laura Gómez-Romero,[2] Joshua I. Haase-Hernández,[3] Israel Aguilar-Ordóñez,[4] Hugo Tovar,[2] S. Eréndira Avendaño-Vázquez[1]* and C. Fabián Flores-Jasso[1,5]*

[1]Consorcio de Metabolismo de RNA. Instituto Nacional de Medicina Genómica, INMEGEN, Ciudad de México, 14610, México.

[2]Unidad de Servicios Bioinformáticos. Instituto Nacional de Medicina Genómica, INMEGEN, Ciudad de México, 14610, México.

[3]Subdirección de Bioinformática. Instituto Nacional de Medicina Genómica, INMEGEN, Ciudad de México, 14610, México.

[4]Subdirección de Investigación Básica. Instituto Nacional de Medicina Genómica, INMEGEN, Ciudad de México, 14610, México.

[5]Lead contact.

*Correspondence:

S.E.A-V. +52 (55) 5350-1900 ext. 1138 (seavendano@inmegen.gob.mx),
C.F.F-J. +52 (55) 5350-1900 ext. 1138 (cfflores@inmegen.gob.mx)

ABSTRACT

The performance of local and global alignment algorithms circumscribe the alignment of sequences shorter than 30 nucleotides. Regardless of the computational approach applied, a series of diverse limitations accumulate as the length and similarity between the aligned sequences decreases, often resulting in alignment biases: local alignments have difficulties reporting correctly the number of matches and mismatches (m/mm) flanking the seed; global alignments lengthen the total alignment size and introduce gaps artificially. These biases compromise the accuracy of computational analysis of short sequences. Here we report ExtendAlign, a computational tool that overhauls and corrects the aforementioned bias generated by local and global alignments. ExtendAlign provides an end-to-end report of the accurate number of m/mm for all the nucleotides that flank a local alignment of short sequences, thus eliminating the artificial lengthening of the query size, the introduction of gaps, and the failure in reporting flanking m/mm. Since ExtendAlign combines the refinement and strength of global and local multiple sequence alignments, it delivers exceptional accuracy in correcting the alignment of dissimilar sequences in the range of ~35–50% similarity —also known as the twilight zone; indicating it can be adopted regularly whenever high accuracy is required for short-sequence alignments.

INTRODUCTION

Sequence alignment is usually the first step performed in bioinformatics to infer the molecular origin of an unknown sequence because it provides a similarity measure between two or more sequences (1–3). From the general types of computational alignment methods currently available (local and global), the tools that perform multiple sequence alignments locally are the most popular for the study, comparison and discovery of short-sequence nucleic acids (4–6); part because it is more straightforward to align multiple sequences to a subject at once, and part because it enables to locate multiple regions sharing certain similarity within a genome —a useful procedure to discover, for example, whether a small RNA has multiple copies (or hits) in a genome (7–11). In contrast, global alignments are typically employed to align sequences of similar length and whose most fundamental relationships have already been inferred, for example, to define or explore by hierarchical clustering the evolutionary function of several members of a microRNA family (1, 12–14); consequently, are not suitable for identifying multiple hits of a series of shorts queries over large sequences or genome subjects.

Albeit most multiple sequence alignment tools (MSATs) that perform local alignments handle robustly sequences above 100 nt long (like BLASTn, BWA, Bowtie, etc.), there is one limitation in the alignment of sequence queries shorter than 30 nt: as the size and similarity between two sequences declines, small stretches of sequences flanking the alignment achieved by the MSAT —that are shorter than the word size and that contain multiple mismatches or gaps— end up unreported (4, 6, 15, 16). This provokes a match and mismatch (m/mm) report bias for some alignments.

This situation affects primarily local alignments, but is not entirely solved by the sequence alignment tools that perform alignments globally. Although global alignments between dissimilar sequences are conducted in end-to-end manner (for instance, MSATs based on the Neddleman-Wunsch algorithm) (15, 17, 18), they typically work in and paired-way manner, may not report the number of m/mm encountered by a hit query accurately, and their alignments are usually found at one end on the subject when the two sequences compared have a large difference in size (2). This constrains global alignments to sequences of similar length that seek only for the best alignment hit. As a consequence, if one is interested in finding multiple alignment hits for a short sequence in a genome, the best option is to perform a local alignment, with the caveat that some alignment hits will have unreported nucleotides in the query that may still have matched with the subject. Another drawback of global alignments is that, in order to minimize mismatches, gaps are introduced artificially. The aforementioned limitations combined, influence the computational study of small RNAs. For example, one important criterion for microRNA annotation is conservation (19–21). In order to obtain the sequence identity, many algorithms used for microRNA discovery and annotation rely on other computational tools to perform the sequence alignments as the first step to then determine if a microRNA is to be considered homologous, or non-homologous —a task executed with difficulty if the detection of m/mm is not accurate (22). Hence, since current local and global alignments circumscribe the study of sequences shorter than 30 nt, there is a need for computational tools whose alignment results have the strength of a

multiple local alignment, and also the refinement provided similar to a global alignment in reporting accurately the number of m/mm for short queries.

Here we present ExtendAlign, a computational algorithm that addresses this problematic by extending the local alignments reported by an MSAT. ExtendAlign corrects the number of m/mm flanking a seed match in every hit provided by the multiple local alignments, and reports a rectified number of true m/mm in end-to-end fashion considering only the query sequence for all hits resulted from the MSAT without increasing the gap number in order to minimize mismatches. First, we evaluated the functionality of ExtendAlign by comparing its alignment results to local and global alignment tools using simulated data that included short sequence queries (22 nt) and sequence subjects with a wide range of sizes and dissimilarity. Then, we tested ExtendAlign to calculate the number of true m/mm after alignments of short sequences (human mature microRNAs) against slightly longer sequences (mouse pre-microRNAs), and against mouse, armadillo and platypus whole genomes. Our results show that ExtendAlign improves significantly the number of m/mm in all alignments tested, even at a sensitivity of a 7-mer seed size (high sensitivity BLAST). Remarkably, this improvement widens to dissimilar sequences in the range of ~35–50% similarity —also known as the twilight zone, the similarity frontier that halts phylogenetic analyses (23, 24). Although ExtendAlign is intended to identify the exact number of m/mm in short sequences (8–30 nt), it executes well with longer sequences and is also capable of improving the m/mm counts for fragments longer than 60 nt.

Finally, we use ExtendAlign to revisit a set of RNA sequences considered to have a bovine-specific origin contained within library datasets obtained from human samples of published literature —a recent biological controversy raised by the report of large amounts of contaminating RNAs from cow into cell lines due to their culture under fetal bovine serum conditions (25). Our pinpoint analysis with ExtendAlign shows that, despite being considered bovine-specific, more than 35% of all small RNA sequences obtained from human cell lines lie above 80% similarity between cow and human; opening the possibility that, because of the MSAT employed (Bowtie2), there was a high false-positive discovery rate of contaminating bovine-specific small RNAs that contributed to this controversy. In the light of theses analyses, we recommend the use of ExtendAlign every time there is a need to find the number of m/mm with the highest accuracy for short sequences, especially in studies that require a quantitative measure of dissimilarity level, or in which identity percent precision is key for determining homology between phylogenetically distant short sequences. ExtendAlign is available for download at https://github.com/Flores-JassoLab/ExtendAlign.

METHODS

Datasets

All mature and precursor microRNAs used as queries and subjects for ExtendAlign analysis were downloaded from miRBase (release 22). Mouse (*Mus musculus*, mm10p6), armadillo (*Dasypus novemcinctus*, v.3.0), and platypus (*Ornithorhynchus anatinus*, v.5.0.1) genomes used as long-sequence subjects were downloaded from NCBI.

Simulations were performed by building 11,901 random 22 nt long query sequences following an uniform probability distribution for each nucleotide. A no variable seed of 7 nucleotides was fixed for every sequence. The start position for this constant region was chosen at random. To generate the subjects, randomly generated nucleotides were added based on the 11,901 query sequences. These extra nucleotides flanked a fixed 7mer position in each query (the seed) to yield a total size that ranged from 49 to 170 nt long. The start position for this constant region was chosen at random. One or up to fifteen nucleotide changes were introduced into the region corresponding to each query sequence. One insertion or deletion (indel) was introduced per every nine mismatches. The length of the indels followed an exponential distribution with lambda=2. The files containing the simulated database were analyzed using HSe-BLAST, ExtendAlign, as described below, and Needle (EMBOSS) with default parameters for command line application (gapextension, 0.5; gapopen, 10).

The dissimilarity percentage (Figure 7) between cow and human genomes (SRR515903; GEO, NCBI) was determined by first converting the SRA into a FASTQ file with the SRA toolkit (v. 2.9.1) (26). Bovine-specific sequences were obtained by aligning the reads to cow (*Bos taurus*, bosTau8, UCSC Genome Browser) and human (*Homo sapiens*, Hg38p12, NCBI) as reference genomes using Bowtie2 (27, 28), with the following parameters (Ref. 25): mode local; seed length, 25; mismatches in seed, 0. Queries that aligned (flag, 4 or 16) to cow, and not human (flag, 0) genomes where filtered from the dataset into a new FASTA file for further analysis with ExtendAlign.

BLASTn Alignments

The parameters of command line BLASTn were adjusted to achieve the highest sensitivity (HSe-BLAST) as follows: word size, 7; reward, 1; penalty, -1; gap open, 2; gap extend, 2; e-value, 10. In order to keep all query sequences even when scored as low complexity, the DUST filtering and soft masking options were disabled. For constructing the lookup table used by BLASTn, the BLAST extensions were performed without soft masking. BLASTn databases for subject sequences were constructed using the makeblastdb command-line tool including the parse_seqids option (in order to keep the original sequence identifiers) and the dbtype nucl (specific for input nucleotide sequences) parameter. BLASTn-short was used with default parameters for the command line application. BLASTn output files were adopted to serve as ExtendAlign input.

In the simulation analysis, in order to keep all alignments regardless of their score, the “eValue” and “max_target_seqs” parameters were change to 5000 and 500 respectively.

ExtendAlign Implementation

The ExtendAlign algorithm was implemented in four stages:

1) It starts by priming with a BLASTn alignment (either BLASTn-short or HSe-BLAST for short-sequence queries (sQ) and short-, long-sequence subjects (sS and lS, respectively).

2) From HSe-BLAST results, ExtendAlign determines if the reported MSAT´s query start/end positions ($_{MSAT}Q$) correspond to the full-length query ($_{FL}Q$). ExtendAlign identifies whether there exist unaligned nucleotides on the queries (extensions) for every hit reported by the BLASTn search according to its reported start/end positions.

3) ExtendAlign calculates first the length of extensions at each end: 5′ extension is calculated by simply subtracting 1 to the $_{MSAT}Q$ start end ($_{MSAT}Q$ –1) —for example, if 5′-$_{MSAT}Q$ = 4, there is a 3 nt 5′ extension; 3′ extension is calculated by subtracting the $_{FL}Q$ minus the $_{MSAT}Q$ end position —for example, if 3′-$_{MSAT}Q$ = 16 and $_{FL}Q$ = 21, there is a 5 nt 3′ extension. If extensions do not equal zero at either end, ExtendAlign fixes the $_{MSAT}Q$ coordinates and then uses Samtools to extract every sequence extension.

4) Finally, every nucleotide in the extension is aligned directly $_{MSAT}Q$ to $_{FL}Q$ to obtain the correct number of m/mm. As verification step, $m_{(total)} + mm_{(total)} + gap_{(total)} = {}_{FL}Q$, for every hit. In order to avoid mistakes while dealing with overhanging nucleotides, ExtendAlign inverts the extensions and performs the procedure described above. All extensions with true overhangs after ExtendAlign are considered mismatches. In sQ *vs*. sS, ExtendAlign allows BLASTn searches only for sense alignments; in sQ *vs*. lS, BLASTn searches are allowed for sense and antisense alignments. The output file of ExtendAlign contains the query name, subject name, $_{MSAT}Q$, $_{FL}Q$ and total match and mismatch values for MSAT and ExtendAlign.

Statistics

Exact data (no bins) were plotted as cumulative frequency distributions. Significance between datasets was obtained by using the Kolmogorov-Smirnov test for non-parametric data with a significance value of $p < 0.005$. Non-linear fit curve by least squares is shown for every dataset; coefficient of determination as measure of goodness of fit. Box plots are presented as min to max values shown by quartiles; significance was obtained by using the Wilcoxon test for paired, nonparametric data with a significance value of $p < 0.05$. Except for simulated data (plotted with R software), all plots were done using the GraphPad Prism software (v.7).

RESULTS

Motivation and Functionality of ExtendAlign

Common biases found after the alignment of short sequences are depicted in Figure 1A. Consider the default setting for BLAST (Basic Local Alignment Search Tool) and its version for aligning nucleotides (BLASTn) (29). When a short sequence query —i.e., 22 nt, is aligned to a subject with certain similarity, no hits will be reported with the default parameters (word size 11) (i). Adjusting a word size that performs better with short sequences —for example by using the default parameters for BLASTn-short (suitable for sequences shorter than 30 nt), allows the nucleotide alignments as long as the seed (word size, 7-mer) is not interrupted by a gap or mismatch (ii); in this example, BLASTn-short will report a hit with 13 matches out of 14 nucleotides aligned. If the parameters of BLASTn are further adjusted to achieve the highest sensitivity (hereinafter HSe-BLAST) —that is, a 7-mer seed size minimizing the penalties for gaps and mismatches (see methods), HSe-BLAST finds an alignment with 15 matches and 2 mismatches for the same pair of sequences (iii). However, the true number of matches and mismatches for an end-to-end alignment after finding a 7-mer seed for this particular hit is 17 and 5, respectively (iv). Even at the highest sensitivity, there are nucleotides beyond the seed size (red, bold letters) still unreported by the MSAT, thus biasing the alignment results; the query in the example for HSe-BLAST is reported as a 17 nt long sequence with 15 matches —instead of 17, and 2 mismatches —instead of 5. This type of bias is particularly common for alignments with dissimilar sequences that contain multiple gaps or mismatches near the 5′ or 3′ ends on the query, and that also fall within a word size that equals or exceeds the unreported sequence length. Notably, the alignment example shown in Figure 1 cannot be handled by other local alignment tools, such as Bowtie, Bowtie2 and BWA (data not shown), although it may be handled correctly by global alignment tools, such as Needle (EMBOSS) (Suppl. Fig. 1) (17, 18, 30). However, other intrinsic features of global alignments might cause a different type of bias in the reported results: the alignment achieved is reported as the subject length, and alignments will be reported as one hit only at one end if the subject is much larger than the query, making global alignments impractical when searching for multiple alignment hits to a genome. In addition, global alignments introduce gaps artificially in order to minimize mismatches and achieve the highest alignment score.

Aiming for a solution for this problem, we developed ExtendAlign, a tool that reduces these report biases by extending the alignment achieved by an MSAT without missing m/mm (like local alignments), and without introducing gaps or lengthening the query size artificially (like global alignments). Primed by a local MSAT, ExtendAlign identifies first the number and identity of any flanking unaligned nucleotides in the query after the local alignment. After finding unreported nucleotides ExtendAlign recalculates and reports the total number of m/mm, in an end-to-end fashion for each query length (Fig. 1B). ExtendAlign corrects the alignments of short-, or long-sequence subjects (for example, whole genomes) whose typical examples of undetected m/mm nucleotides in query searches locate near the 5′ or 3′ ends of queries (Fig. 1C): (i), undetected m/mm; (ii), overhanging query nucleotides are considered mismatches; (iii) and m/mm are

extended up to the last query nucleotide at the 5′ or 3′ ends. Results are reported as a table containing the m/mm achieved before and after ExtendAlign for each query hit (Fig. 1C, right).

To exemplify how ExtendAlign positions at the intersection between local and global alignments, we aligned a simulated database of randomly generated query sequences of 22 nt against randomly generated longer subject sequences (49 to 170 nt long); every query was aligned to every subject sequence (see methods). Needle aligned 100% of each sequence for every hit achieved (5,950,500 hits), but also increased the total alignment size up to the longest subject (Fig. 2A), which is also accompanied by a similar increase in the number of gaps (Fig. 2B). This result implies that Needle: i) considers overhanging nucleotides in the query or subject as part of the alignment, and ii) introduces gaps interspersed along some sequences to achieve alignments. Conversely, HSe-BLAST increased marginally the number of gaps (up to 4), but does not achieve 100% nucleotide alignment coverage for all hits (6,717,736 hits) (Fig. 2A and 2B). Importantly, the percentage of nucleotides in each hit that did not aligned by HSe-BLAST should display a complementary representation in the mismatch number (Fig. 2C). For example, a 22 nt sequence query with 50% alignment contains 11 nt matches, and also 11 nt mismatches and/or gaps. However, neither the number of mismatches (up to 5, Fig. 2C), nor the number of gaps (up to 4, Fig. 2B) adds up to 11 nt, indicating that there are unreported mismatches as a result of the local alignment. As expected, the outcome delivered by ExtendAlign stands out from the local and global alignments because there is no increase in the total alignment size, nor a decrease in the number of mismatches, and it always achieves 100% coverage of the sequence query for every hit reported (Fig. 2A–C, in red). Other features that encompass the advantages of ExtendAlign applicability, like zero-gap introduction or allowance of indels at the extension phase, are summarized in Table 1; a comparison with other widely used alignment tools for these and several other features, including their recommended use, is also provided.

ExtendAlign improves the alignment of unreported matches and mismatches

To evaluate the performance of ExtendAlign we used human mature microRNAs as short-sequence queries. For short- and long-sequence subjects we used mouse precursor microRNAs and mouse whole genome. The rationale for this is that human microRNAs are used frequently as a reference in homology studies and it also contains the largest list of annotated microRNAs. Conversely, mouse genome contains the second largest list reported of precursor microRNAs, and most of its mature sequences are considered orthologs to human (miRBase release 22) (20). Based on their recommended use (Table 1), and because BLASTn delivers the highest hit number compared to other sequence alignment tools with similar word size and parameters (*i.e.*, BLAT, Bowtie, Bowtie 2, BWA, Suppl. Table 1) (15, 16, 31), we used BLASTn-short, and HSe-BLAST as local MSATs to prime the alignments in order to provide ExtendAlign with the largest datasets to test.

First, we initiated the alignment of 2,656 human mature microRNAs versus 1,234 mouse precursor microRNAs using BLASTn-short and its default parameters as MSAT; 2.66 $x10^4$ hits were found. The number of matches per hit achieved was plotted as a cumulative fraction histogram. Then we tested if

ExtendAlign could increase the number of matches per hit on nucleotides that may had been unreported by the MSAT. Our results show that ExtendAlign significantly increased the number of matches originally unreported by BLASTn-short ($p < 0.0001$) (Fig. 3A). Next we tested the alignments made by the MSAT using the highest sensitivity parameters. As expected, when alignments were done with HSe-BLAST, the number of hits increased (4.27×10^4 hits) compared to BLASTn-short, and also ExtendAlign detected unreported nucleotides that resulted in an increased number of matches even at the highest sensitivity MSAT alignment (Fig. 3B). Then, to test the performance of ExtendAlign with long-sequence subjects we aligned all human mature microRNAs against the mouse genome. Our results show that ExtendAlign also increased the number of matches found by BLASTn-short significantly ($p < 0.0001$) (Fig. 3C). When alignments were done with HSe-BLAST, we observed a marginal, but significant increase in the number of matches reported after the ExtendAlign correction ($p < 0.0001$, Kolmogorov-Smirnov $D = 0.005$) (Fig. 3D). The marginal difference achieved by ExtendAlign compared to HSe-BLAST cannot be attributable to the hit number, because it is larger than BLASTn-short due to its higher sensitivity (2.17×10^6 *vs*. 1.99×10^6). One explanation for this result is that, due to its higher sensitivity, HSe-BLAST is able to find a larger number of perfect, or near perfect sequence alignments when the subject is a whole genome. To test this idea, we removed all hits whose alignments found perfect match to the mouse genome, because they are not candidates for ExtendAlign correction. When the data was plotted again, we found a less marginal ($D = 0.69$) and significant difference ($p < 0.0001$) for alignments that were susceptible for correction by ExtendAlign (Fig. 3D, inset). This result supports the purpose of ExtendAlign in correcting the total number of matches contained in an alignment for dissimilar sequences.

Next, we analyzed the number of mismatches for each of the alignments described above (Fig. 4). In contrast to the improvement achieved on the number of matches, the difference in the number of mismatches unreported by the MSAT is more drastic. When the human mature microRNAs were aligned *vs*. mouse precursor microRNAs, ExtendAlign improved significantly both, BLASTn-short ($p < 0.0001$) and HSe-BLAST ($p < 0.0001$) (Fig. 4A and 4B, respectively). The large difference encountered in the mismatches reported by ExtendAlign, compared to matches, is due to the large amount of nucleotides that go unreported by the MSAT: all hits containing query overhangs at their 5′ or 3′ ends with unreported mismatches vastly exceed the number hits corresponding to matches (Fig. 1C, examples ii and iii; and Suppl. Table 2). When alignments were done against the mouse genome, ExtendAlign also increased the number of matches found by BLASTn-short significantly ($p < 0.0001$) (Fig. 3C). However, similarly to what occurred for unreported matches (Fig. 4D), the number of mismatches corrected by ExtendAlign had a significant, but marginal difference with respect to HSe-BLAST alone ($p < 0.0001$, Kolmogorov-Smirnov $D = 0.016$). Likewise, by removing the hits that found a perfect match, a less marginal and significant difference is observed (Kolmogorov-Smirnov $D = 0.89$, $p < 0.0001$) (Fig. 3D, inset). Even at the highest sensitivity, HSe-BLAST missed more than 2% of all mismatches corrected by ExtendAlign (Suppl. Table 2); this percentage increased considerably when BLASTn-short was used for the alignments, and also when precursor microRNAs served as sequence subjects. Taken together these results, we conclude that ExtendAlign is a

reliable algorithm that can be used to refine the initial alignment and minimize the biases caused by the unreported m/mm by the use of local MSATs alone.

In phylogenetic inference studies, genome sequencing has become important for establishing (or re-establishing) taxa in which is difficult to determine the genetic boundaries that comprise a taxonomic division; since a key determinant in homology searches is the identity at nucleotide level, having a precise number of m/mm between two contrasted sequences is of utmost importance. Some species are considered to be at a taxonomic border because many of their genes contain dissimilar sequences and it is difficult to assign the correct taxa (32). Examples of this are the clades of placental mammals (Afrotheria, Xenarthra, Boreoeutheria), whose sequence relationships are at the border of phylogenetic classification. Since ExtendAlign performed robustly in correcting number of m/mm between more dissimilar sequence alignments, we used the list of human precursor microRNAs as sequence queries against distant genomes among mammals in order to find candidates with dissimilar hit matches for correction. First, we used the genome of *Dasypus novemcinctus* (armadillo) because these animals have been considered to be at the border of taxonomic classification (32). To challenge and compare the results provided by ExtendAlign we used HSe-BLAST as MSAT. Our results revealed a significant increase in the number of matches per hit detected by after the ExtendAlign correction (Fig. 5A, $p < 0.0001$). The number of mismatches also increased significantly ($p < 0.0001$), albeit at much larger extent (Fig. 5B; $D = 0.91$, mismatches *vs*. $D = 0.43$, matches).

Next, we aligned the human precursor microRNAs against the genome of *Ornithorhynchus anatinus* (platypus). Platypus diverged from the mammalian lineages around 160 million years ago and share morphological and genomic features that represent a mix of mammals, reptiles, birds and fish (33). Therefore, the platypus genome provides a unique genomic blend as an attractive source for dissimilar alignments that could serve as candidate hits for ExtendAlign. We observed a significant increase in the number of matches and mismatches per hit, compared with the alignments achieved by HSe-BLAST alone ($p < 0.0001$; Fig. 5C and 5D, respectively). These results suggest that phylogenetic analyses executed using MSATs only may contain underrepresented alignments for all those hit matches for sequences with unreported nucleotides; suggesting that ExtendAlign might be a useful tool to potentially improve cladistics and reconstruct evolutionary trees based on a refined alignment.

Finding a correct sequence identity is not only important for phylogenetic analyses, but also to predict secondary structures for nucleic acids and proteins. In order to do so, multiple sequence alignments are performed followed by multiple folding events based on optimized thermodynamic scores and covariance (24). However, a conserved secondary structure prediction is highly limited for alignments of dissimilar sequences within the so-called “twilight zone” —the sequence alignment that ranges from ~35–50% similarity, because the degree of conservation cannot be estimated reliably (23, 24, 34). Notably, ~50% of all alignment hits achieved by HSe-BLAST in the armadillo and platypus genomes span across this range (Fig. 6A and 6B, HSe-BLAST; medians 43% and 41%, respectively). After ExtendAlign correction, the alignments within this zone improved above 50% of matches per hit (Fig. 6A and 5B, ExtendAlign; medians 44% and

45%, respectively; $p < 0.0001$). These results support the use of ExtendAlign as a regular procedure after alignments performed by local MSATs aimed to find an accurate identity percent, and reinforce the notion that unreported nucleotides, especially mismatches, may lead to inaccurate short sequence annotations.

Pinpoint classification of short sequences based on similarity percentage

Finally, we used ExtendAlign to revisit the sequence alignments performed in a recent work reporting that RNA sequences contained in fetal bovine serum (FBS), may be transferred into cell cultures (25); the extent at which this observation might affect the literature about microRNAs function is still unknown, and a matter of debate (35). The computational analysis performed by Wei and colleagues encompassed the search for bovine-specific RNA sequences contained within some public sequence read archives (SRAs) from human samples. By using Bowtie2, a common procedure to align sequences to genomes, the authors considered bovine-specific all those sequences that mapped the *Bos taurus* genome, and not the human genome (Fig. 7A). Since the analysis was made using SRAs aimed to sequence small RNAs, we used ExtendAlign to find out how dissimilar to human are indeed the sequences classified as bovine-specific (Fig. 7B). As expected, there is an ample percentage range at which the bovine-specific sequences redistribute, being the highest abundance at 75–77.5%; only a minor amount were not aligned at all (Fig. 7B, NA) —expected only for those sequences accurately classified as bovine-specific. However, we find that ~35% redistributed at 80–98% similarity range (Figure 6B, grey area); a 22-mer sequence with an 80% similarity has only 4 mismatches to its reference. Hence, this result suggests that an important amount of the bovine-specific sequences might be the result of a high false-positive discovery rate.

DISCUSSION

There has been an important effort by many research groups focused on the development of MSATs that reduce the processing time consumed for comparisons between two or more sequences that has yielded a number of reliable algorithms that perform robustly in the genomic era (6). Most tools, however, have focused on improving the recognition of highly similar sequences in order to find better and faster alignments for homology and functional studies (15, 16, 28, 36). As a result, the alignments for dissimilar short sequences have been left somewhat unattended. ExtendAlign is a tool that delivers local, end-to-end alignments of short sequences. Most MSATs perform robustly in identifying closed alignments with a high identity percent (>90%) for long sequences. However, as the size and similarity between two sequences declines, there is a bias originated by the MSAT word size that yields unreported nucleotides at the query's ends that affects the number of total m/mm for certain types of alignments. This bias is more acute for short-sequence queries (<30 nt). ExtendAlign corrects this bias and deliver the true number of m/mm, in end-to-end query hits contained in multiple sequence alignment without introducing gaps artificially. ExtendAlign was developed to incorporate the basic output of an MSAT and provide a list of all hits with their corrected m/mm for the total length of the query sequence. In this work we employed BLASTn because it provides the highest number of hits compared to other MSATs, but we envision it could be adapted to also accept the output from other local MSATs that allow more or less nucleotides in the seed word size. Since full-length, perfect match alignments are not prone for improvement, ExtendAlign is more robust in correcting the number of true m/mm in more dissimilar alignments; a useful feature for sequence alignments between more evolutionarily distant species. We showed that ExtendAlign refines the alignment of more dissimilar sequences, particularly those alignments situated within the ~30-50% of matching nucleotides, also called the twilight zone (23, 24, 34, 37). Examples of alignments that ExtendAlign may help to improve because of their similarity percent span across this range include the conservation and functionality of lncRNAs (which are subject to weak functional constraint and rapid turnover during evolution) (38). Although ExtendAlign was initially developed to satisfy the need for multiple local alignments in end-to-end manner of short sequences, it executes robustly with longer sequences, like precursor microRNAs, which are frequent instruments for the discovery of novel microRNAs.

In this work we show how ExtendAlign helps to pinpoint m/mm in alignments for short sequences with outstanding precision that may be otherwise go unnoticed by using only the commonly employed MSATs. The intriguing finding that RNAs from bovine origin may transfer into cell cultures as a result of incubation with FBS has raised an important controversy in the microRNA field of research (25, 35). Here we present evidence that at least some of the sequences considered bovine-specific by Wei and coworkers might be the result of a high false-positive discovery rate. Since the bovine-specific sequences were obtained with Bowtie2, the fact that still ~35% of them lie at a high identity as compared to human, validates the use of HSe-BLAST as local MSAT to prime ExtendAlign. As no other tool possesses the accuracy of ExtendAlign in identifying m/mm for the alignment of short sequences, the use of any other tool would present similar or

even higher false-positive rates. By using our tool we anticipate that future studies will tackle with much better precision how this problem truly affects the entire literature in the microRNA field.

Since multiple alignments are key in computational biology, ExtendAlign's utility can broaden up to other areas affected by the same type of bias of short sequences, like discovery of novel microRNAs, SNPs correction, tRNA-derived fragments expression regulation, and other small RNAs from emerging model organisms; or to pinpoint similarity among pathogenic or complex samples (*i.e*. host-pathogen metagenomics); or phylogeny of distant species; to name a few. ExtendAlign can be used to reevaluate an alignment data set containing all hits and then extracting the hits with an identity percent of interest (3). The sequencing era has given us the opportunity to analyze several species at genomic level; the prediction of conserved and non-conserved microRNAs among species has become a field of great interest in the last decade. Thus, novel tools like ExtendAlign can help in reducing false-positives commonly found in homology searches for small RNAs and also to increase the refinement of alignments between dissimilar sequences (4, 39).

DATA AVAILABILITY AND COMPUTATIONAL REQUIREMENTS

Working requirements: ExtendAlign runs in command line, all input files must be FASTA format; BLASTn, v. 2.2.31+; Samtools, v. 1.5 (40); EMBOSS v. 6.6.0.0; Execline, v. 2.5.0.1, and 9base/plan9port for Unix. Executable, instructions to run ExtendAlign and all datasets related to this work are available at: https://github.com/Flores-JassoLab/ExtendAlign.

AUTHOR CONTRIBUTIONS

M.FT., J.I.HH., and I.AO., constructed EA and set GitHub repository; M.FT., L.GR, H.T., S.E.AV. and C.F.FJ., validated EA; M.FT., S.E.AV. and C.F.FJ., envisioned the project and wrote the manuscript.

SUPPLEMENTARY DATA

Supplementary Data are available at online .

ACKNOWLEDGEMENT

We thank all members of the Avendaño-Vázquez and Flores-Jasso laboratories for critical comments on the manuscript. We also thank Wayne Matten (NCBI-BLAST crew) for valuable help for implementing the command line HSe-BLAST.

FUNDING

This work was supported in part by the Instituto Nacional de Medicina Genomica [08/2017/I-322] and SS/IMSS/ISSSTE-CONACyT [289862] to S.E.A-V; and the Instituto Nacional de Medicina Genomica [05/2017/I-321] to C.F.F-J.

REFERENCES

1. Corpet,F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.*, **16**, 10881–10890.

2. Chowdhury,B. and Garai,G. (2017) A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, **109**, 419–431.

3. Shah,N., Nute,M.G., Warnow,T. and Pop,M. (2018) Misunderstood parameter of NCBI BLAST im-pacts the correctness of bioinformatics work- flows. *Bioinformatics*, **10**, 1093/bioinformatics/xxxxx.

4. Polyanovsky,V.O., Roytberg,M.A. and Tumanyan,V.G. (2011) Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. *Algorithms Mol. Biol.*, **6**, 1–12.

5. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A Greedy Algorithm for Aligning DNA Sequences. *J. Comput. Biol.*, **7**, 203–214.

6. Li,H. and Homer,N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.*, **11**, 473–483.

7. Tanzer,A. and Stadler,P.F. (2004) Molecular evolution of a microRNA cluster. *J. Mol. Biol.*, **339**, 327–335.

8. Pfeffer,S., Sewer,A., Lagos-Quintana,M., Sheridan,R., Sander,C., Grässer,F.A., van Dyk,L.F., Kiong Ho,C., Shuman,S., Chien,M., *et al.* (2005) Identification of microRNAs of the herpesvirus family. *Nat. Methods*, **2**, 269–276.

9. Smalheiser,N.R. and Torvik,V.I. (2005) Mammalian microRNAs derived from genomic repeats. *Trends Genet.*, **21**, 322–326.

10. Meunier,J., Lemoine,F., Soumillon,M., Liechti,A., Weier,M., Guschanski,K., Hu,H., Khaitovich,P. and Kaessmann,H. (2013) Birth and expression evolution of mammalian microRNA genes. *Genome Res.*, **23**, 34–35.

11. Guerra-Assunção,J.A. and Enright,A.J. (2010) MapMi: Automated mapping of microRNA loci. *BMC Bioinformatics*, **11**.

12. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443-453.

13. Edgar,R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

14. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

15. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.

16. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

17. Li,W., Cowley,A., Uludag,M., Gur,T., McWilliam,H., Squizzato,S., Park,Y.M., Buso,N. and Lopez,R. (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.*, **43**, W580–W584.

18. Mcwilliam,H., Valentin,F., Goujon,M., Li,W., Narayanasamy,M., Martin,J., Miyar,T. and Lopez,R. (2009) Web services at the European Bioinformatics Institute-2009. *Nucleic Acids Res.*, **37**, W6-W10

19. Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res.*, **32**, D109–D111.

20. Kozomara,A. and Griffiths-Jones,S. (2014) MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, 68–73.

21. Kozomara,A. and Griffiths-Jones,S. (2011) MiRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, 152–157.

22. Güler,P., Bozcuk,S., Mutlu,F. and Sorkun,K. (2014) miRNomics: MicroRNA Biology and Computational Analysis Yousef,M., Allmer,J. (eds) Humana Press, Totowa, NJ.

23. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng. Des. Sel.*, **12**, 85–94.

24. Bremges,A., Schirmer,S. and Giegerich,R. (2010) Fine-tuning structural RNA alignments in the twilight zone. *BMC Bioinformatics*, **11**, 1-8.

25. Wei,Z., Batagov,A.O., Carter,D.R.F. and Krichevsky,A.M. (2016) Fetal Bovine Serum RNA Interferes with the Cell Culture derived Extracellular RNA. *Sci. Rep.*, **6**, 31175; doi: 10.1038/srep31175 (2016).

26. Guduric-Fuchs,J., O'Connor,A., Camp,B., O'Neill,C.L., Medina,R.J. and Simpson,D.A. (2012) Selective extracellular vesicle-mediated export of an overlapping set of microRNAs from multiple cell types. *BMC Genomics*, **13**, 1-14.

27. Langmead,B., Wilks,C., Antonescu,V. and Charles,R. (2018) Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*, 10.1093/bioinformatics/bty648.

28. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 10.1038/nmeth.1923.

29. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

30. McWilliam,H., Li,W., Uludag,M., Squizzato,S., Park,Y.M., Buso,N., Cowley,A.P. and Lopez,R. (2013) Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.*, **41**, W597-W600.

31. Kent,W.J. (2002) BLAT - The BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

32. Tarver,J.E., Dos Reis,M., Mirarab,S., Moran,R.J., Parker,S., O'Reilly,J.E., King,B.L., O'Connell,M.J.,

Asher,R.J., Warnow,T., *et al.* (2016) The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biol. Evol.*, **8**, 330–344.

33. Warren,W.C., Hillier,L.D.W., Marshall Graves,J.A., Birney,E., Ponting,C.P., Grützner,F., Belov,K., Miller,W., Clarke,L., Chinwalla,A.T., *et al.* (2008) Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, **453**, 175–183.

34. Wilm,A., Mainz,I. and Steger,G. (2006) An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol. Biol.*, **1**, 19.

35. Tosar,J.P., Cayota,A., Eitan,E., Halushka,M.K. and Witwer,K.W. (2017) Ribonucleic artefacts: Are some extracellular RNA discoveries driven by cell culture medium components? *J. Extracell. Vesicles*, **6**:1, 1272832, DOI: 10.1080/20013078.2016.1272832.

36. Chen,Y., Ye,W., Zhang,Y. and Xu,Y. (2015) High speed BLASTN: An accelerated MegaBLAST search tool. *Nucleic Acids Res.*, **43**, 7762–7768.

37. Ponting,C.P. (2017) Biological function in the twilight zone of sequence conservation. *BMC Biol.*, **15**, 1–9.

38. Kapusta,A. and Feschotte,C. (2014) Volatile evolution of long noncoding RNA repertoires: Mechanisms and biological implications. *Trends Genet.*, **30**, 439–452.

39. Brudno,M., Malde,S., Poliakov,A., Do,C.B., Couronne,O., Dubchak,I. and Batzoglou,S. (2003) Glocal alignment: Finding rearrangements during alignment. *Bioinformatics*, **19**, i54-62.

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

TABLE AND FIGURE LEGENDS

Figure 1. (*A*) Common biases caused by unreported matches and mismatches and ExtendAlign correction examples. Unreported nucleotides; red, bold letters. (*B*), ExtendAlign (EA) takes the output file of an MSAT after the alignment of long-, or short-sequence subjects and short-sequence queries. (*C*) i–iii, examples of unreported nucleotides comprising matches or mismatches that originate alignment biases. Nucleotides flanking the seed-word size after MSAT alignment (red, bold) are extended and the true number of m/mm is reported by ExtendAlign according to the query length. Typical examples of nucleotides unreported by commonly used MSATs locate near the 5′ or 3′ ends on the queries aligned to long- (*C*, i), or short-sequence (*C*, ii and iii) subjects. The nucleotides that overhang on queries either at their 5′ or 3′ ends (underlined, grey; example ii) are interpreted as mismatches by ExtendAlign; overhanging nucleotides on subjects are not reported.

Figure 2. ExtendAlign achieves 100% query alignment without increasing the alignment size or introducing gaps. 11,901 randomly generated query and subject sequences (22mers and 49–170, respectively) were aligned using HSe-BLAST, Needle, and ExtendAlign. Percent of query aligned (*A*), Number of gaps per sequence pair (*B*), and Number of mismatches per query (*C*) were plotted as a function of total alignment size for each alignment hit; the frequencies for each intersection were plotted in *z*. Hits per alignment: HSe-BLAST (grey) 6,717,736 hits; Needle (cyan) 5,950,500 hits; ExtendAlign (red) 6,717,736 hits.

Figure 3. ExtendAlign corrects the number of matches per hit reported by MSAT in short-sequence queries. (*A* and *B*) Human mature microRNAs *vs*. mouse precursor microRNAs. (*A*) BLASTn-short *vs.* ExtendAlign (EA). (*B*) High Sensitivity BLASTn (HSe-BLAST) *vs.* ExtendAlign. (*C* and *D)* Human mature microRNAs *vs*. mouse genome. (*C*) BLASTn-short. (*D*) HSe-BLAST. BLASTn-short (grey), HSe-BLAST (grey) and EA (red) are plotted as cumulative fraction histogram of the number of matches found in each alignment. Significance values were calculated using the Kolmogorov-Smirnov [K-S] test for unpaired nonparametric data; *D* = K-S distance.

Figure 4. ExtendAlign corrects the number of mismatches per hit unreported by MSATs in short-sequence queries. (*A* and *B*) Human mature microRNAs *vs*. mouse precursor microRNAs. (*A*) BLASTn-short *vs.* ExtendAlign (EA). (*B*) High Sensitivity BLASTn (HSe-BLAST). (*C* and *D*) Human mature microRNAs *vs*. mouse genome. (*C*) BLASTn-short. (*D*) HSe-BLAST. The number of mismatches per hit achieved by BLASTn-short (grey), HSe-BLAST (grey) and EA (red) are plotted as cumulative fraction histogram. Significance values were calculated using the Kolmogorov-Smirnov [K-S] test for unpaired nonparametric data; *D* = K-S distance.

Figure 5. ExtendAlign improves the number of matches and mismatches for undetected nucleotides in pre-microRNA as queries vs armadillo and platypus genome. (*A* and *B*) Comparison of HSe-BLAST and ExtendAlign (EA), human precursor microRNAs (query), *vs*. armadillo genome (subject). (*C* and *D*) Comparison between HSe-BLAST and ExtendAlign (EA), human precursor microRNAs (query), *vs*. platypus genome (subject). (*A* and *C*) Detected number of matches; (*B* and *D*) Detected number of mismatches. The

number of matches and mismatches per hit achieved by HSe-BLAST (grey) and ExtendAlign (red) are plotted as cumulative fraction histogram. Significance values were calculated using the Kolmogorov-Smirnov [K-S] test for unpaired nonparametric data; *D* = K-S distance.

Figure 6. ExtendAlign improves the alignment percent of dissimilar sequences that span across the twilight zone. (*A* and *B*) Comparison of the identity percent for all matches per hit made by HSe-BLAST and after ExtendAlign correction. (*A*) Human precursor microRNAs vs. armadillo genome. (*B*) Human precursor microRNAs vs. platypus genome. Significance was obtained by Wilcoxon test for paired, non-parametric data.

Figure 7. ExtendAlign categorizes with pinpoint accuracy the similarity percentage non-detected by commonly employed MSATs. (*A*) Diagram for the alignment strategy followed by Wei et al. to define the bovine-specific RNA sequences found on public SRA data. (*B*) Bovine-specific sequences distribute along a wide range of similarity percent. NA, Not aligned. Gray area = ~35% of all reads (1.1×10^5).

Table 1. Comparison of ExtendAlign and other common MSATs.

	ExtendAlign	Bowtie	Bowtie2	BWA	BLAST	BLASTn-short	NEEDLE
Local, end-to-end alignment	✓	✗	✗	✓	✗	✗	✗
Alignment report based on	query	query	query	query	query	query	longest sequence (query + subject)
performs multiple pairwise alignment	✓	✓	✗	✓	✓	✓	✗
Multiple alignment hits against subject sequence	✓	✓	✗	✓	✓	✓	✗
Reports similarity percentage	✓	✗	✗	✗	✓	✓	✓
Minimum seed size	7 seed (default), customizable	5 nt (28 nt default)	5 nt (20 nt default)	BWA-mem, 19 nt	11 nt	7 nt	NA
Allows imperfect seed	✓	✓	✓	✓	✗	✗	NA
Recomended minimum query size	8 nt (default)	30–1,024 nt range	50 nt	35 nt (BWA-backtrack)	50 nt	7–50 nt range	NA
Allows indels at extension phase	✗	✗	✓	✓	✓	✓	✓
Recommended use	For multiple alignment of highly dissimilar and similar to very short, and long sequences.	For the alignment of large sets of short DNA sequences to large genomes.	For the alignment of long DNA sequences (above 50 nt) to large genomes.	For mapping low-divergent sequences against a large reference genome.	For inter-species comparisons.	For inter-species comparisons. Optimized for sequences shorter than 50 nt.	For alignment of closely related sequences.

Flores-Torres et al., Figure 1

A

(i) 5′-UGGAUUCUUUAGAGUGCAUCUGCGUC-3′
5′-GCUGCUGUAGAGGGCAUCUGCU-3′

BLASTn (word size 11):
No hits
Query size 22 nt

(ii) 5′-UGGAUUCUUUAGAGUGCAUCUGCGUC-3′
||||| ||||||||
5′-**GCUGCUG**UAGAGGGCAUCUGC**U**-3′

BLASTn-short (word size 7):
13 matches, 1 mismatch
Query size reported 14 nt
Unreported nt: 8

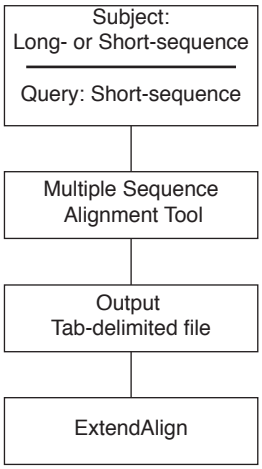
(iii) 5′-UGGAUUCUUUAGAGUGCAUCUGCGUC-3′
|| ||||| ||||||||
5′-**GCUG**CUGUAGAGGGCAUCUGC**U**-3′

HSe-BLAST (word size 7):
15 matches, 2 mismatches
Query size reported 17 nt
Unreported nt: 5

(iv) 5′-UGGAUUCUUUAGAGUGCAUCUGCGUC-3′
| | || ||||| ||||||||
5′-GCUGCUGUAGAGGGCAUCUGCU-3′

Total query matches: 17
Total query mismatches: 5

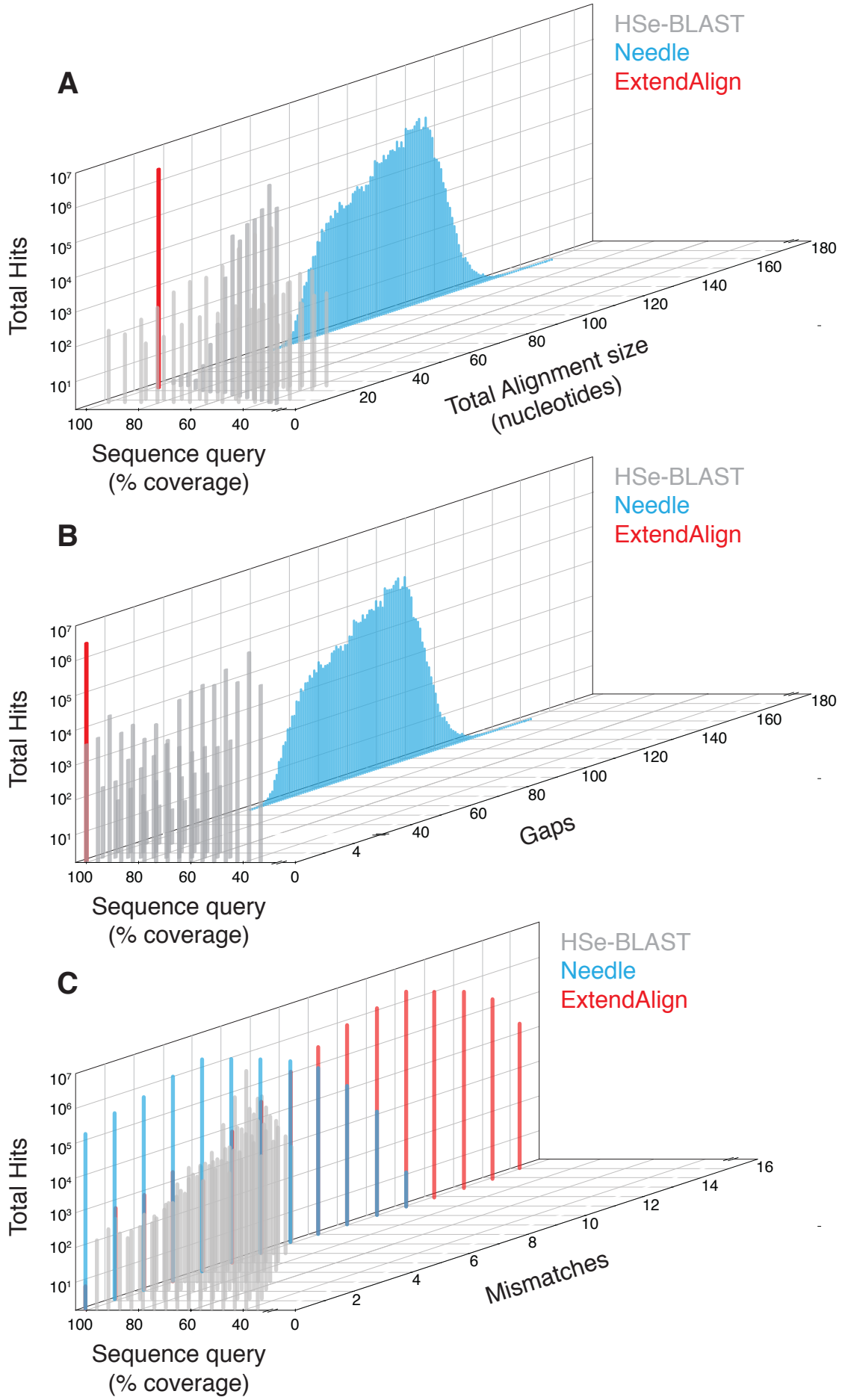
B



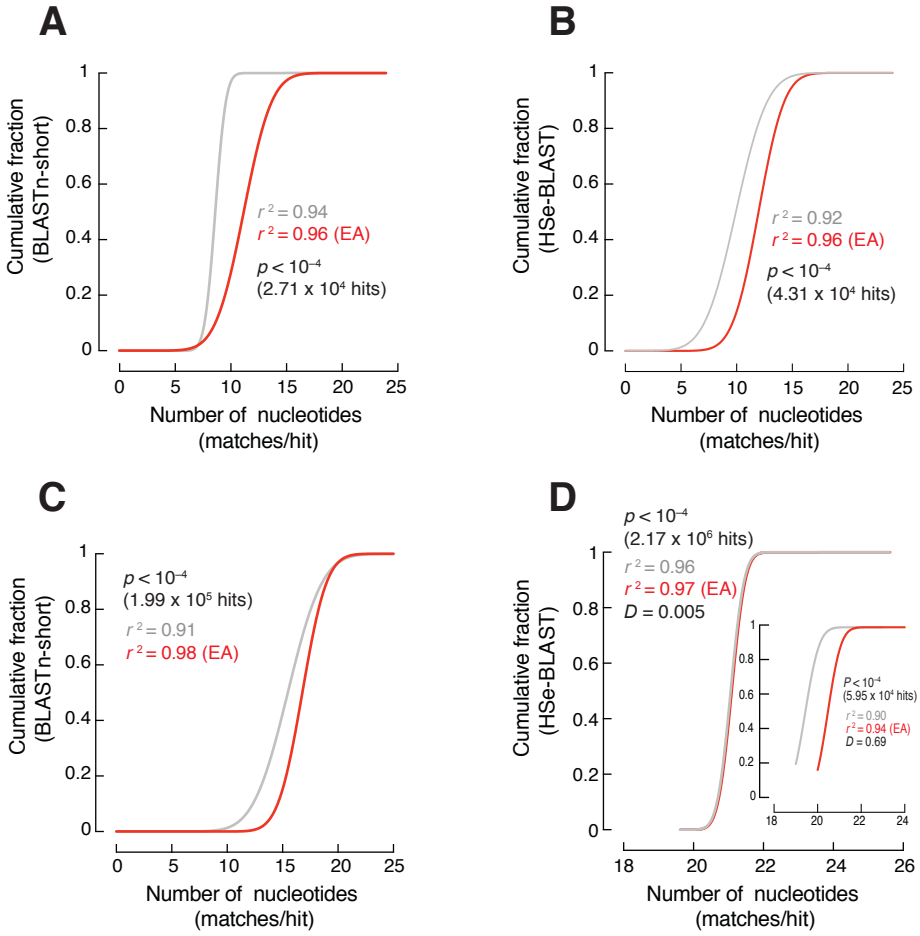
C

	MSAT	MSAT + ExtendAlign	
(i)	14	17	Matches
	2	4	Mismatches
	16	21	Query length
(ii)	7	9	Matches
	0	7	Mismatches
	7	16	Query length
(iii)	19	21	Matches
	0	5	Mismatches
	19	26	Query length

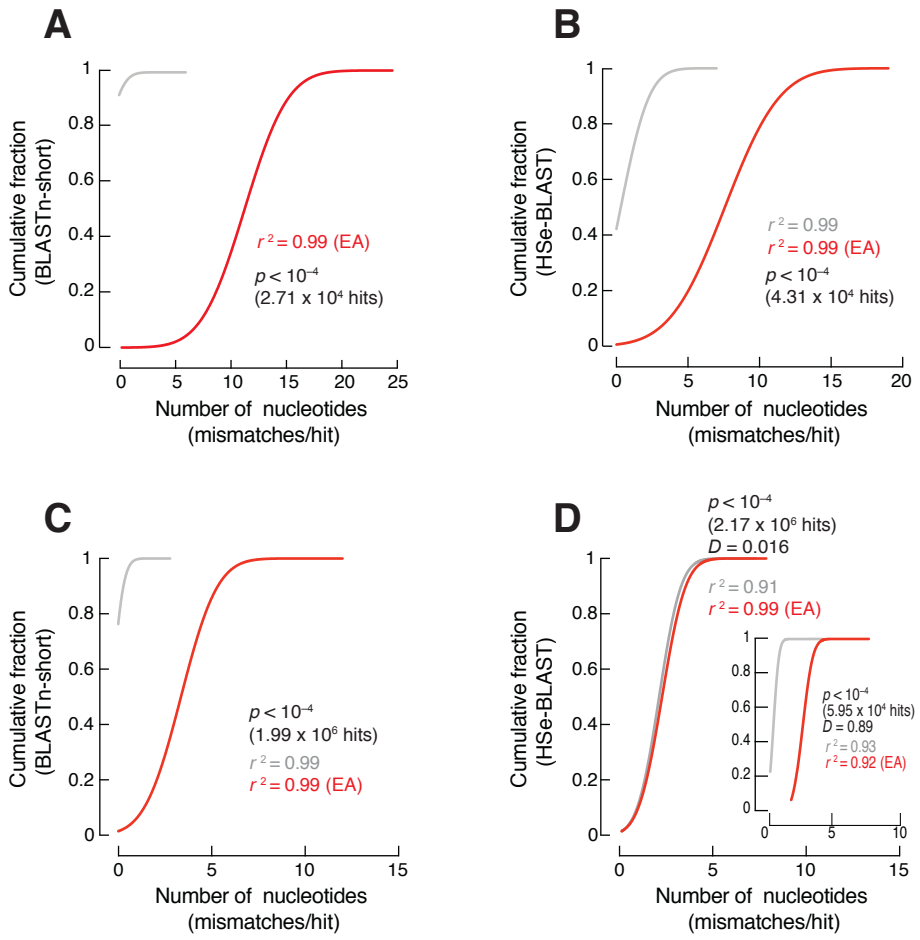
Flores-Torres et al., Figure 2



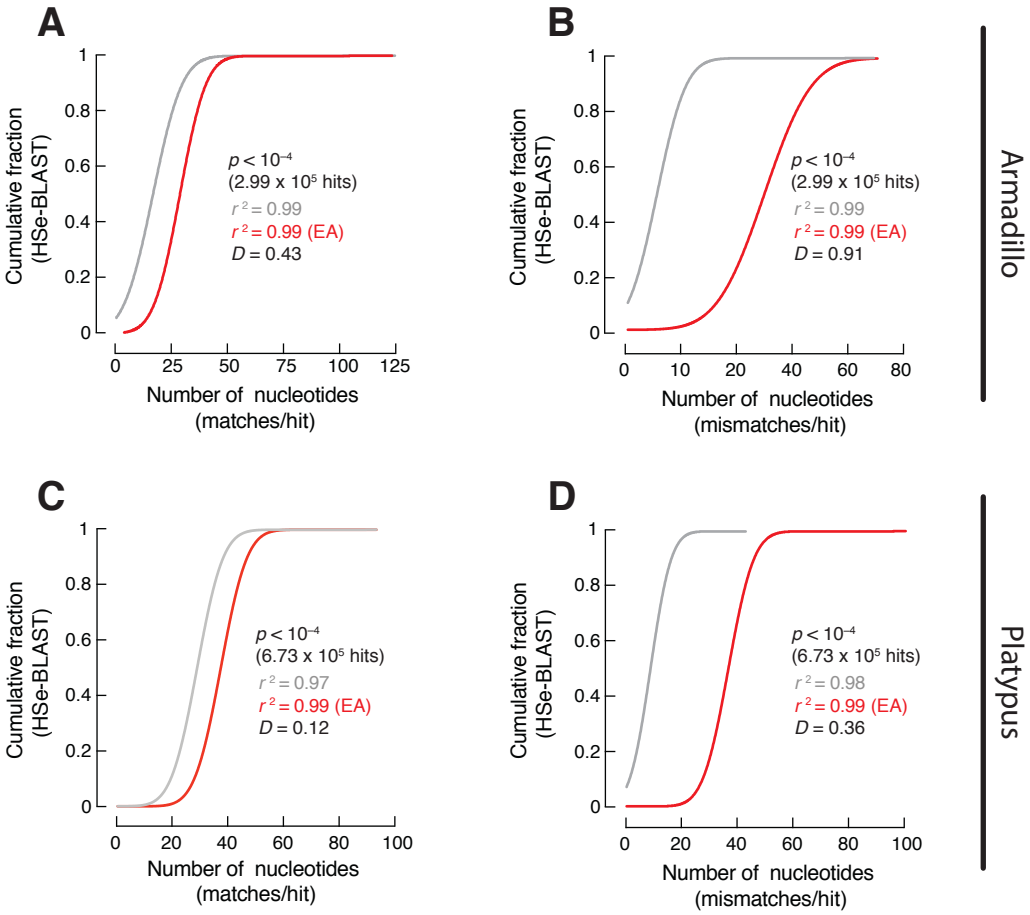
Flores-Torres et al., Figure 3



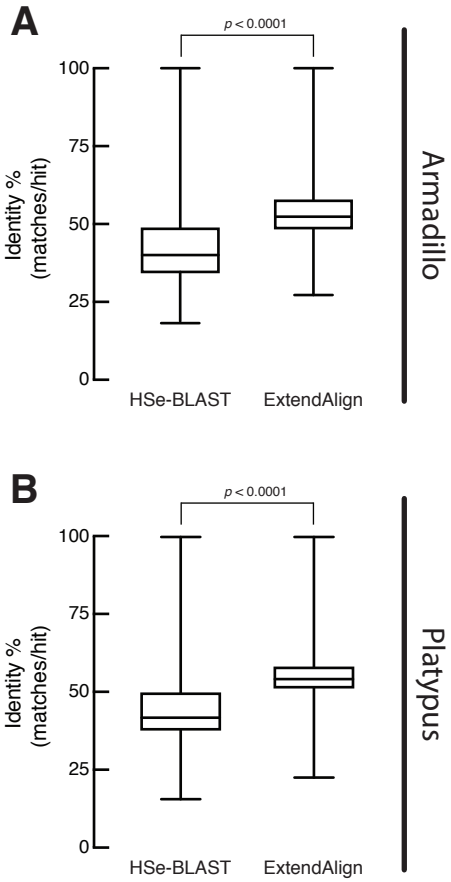
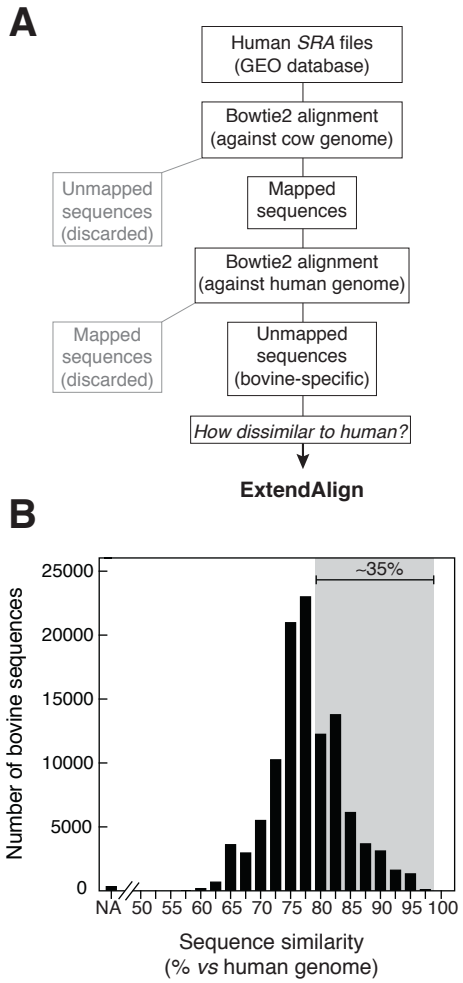
Flores-Torres et al., Figure 4



Flores-Torres et al., Figure 5



Flores-Torres et al., Figure 6



Flores-Torres et al., Figure 7