ExtendAlign: the post-analysis tool to correct and improve the alignment of dissimilar 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*] C. Fabián Flores-Jasso.[1,5 *]

AUTHORS' INFORMATION

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

[2]Genómica Computacional. 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.

MFT: mflores@inmegen.edu.mx

LGR: lgomez@inmegen.gob.mx

JHH: jihaase@inmegen.gob.mx

IAO: iaguilar@inmegen.edu.mx

HT: hatovar@inmegen.gob.mx

*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

In this work, we evaluated several tools used for the alignment of short sequences and found that most aligners execute reasonably well for identical sequences, whereas a variety of alignment errors emerge for dissimilar ones. Since alignments are essential in computational biology, we developed ExtendAlign, a post-analysis tool that corrects these errors and improves the alignment of dissimilar short sequences. We used simulated and biological data to show that ExtendAlign outperforms the other aligners in most metrics tested. ExtendAlign is useful for pinpointing the identity percentage for alignments of short sequences in the range of ~35–50% similarity.

KEWORDS

dissimilar alignment, short-sequence, local alignment, global alignment, distant reference, twilight-zone

BACKGROUND

Since the emergence of high-throughput sequencing technologies, the development of computational tools that aid in the assembling, comparison, and analysis of multiple genomic sequences has flourished tremendously [1,2]. The sequencing of genomes from novel organisms has led to important discoveries that have shaped our understanding of what makes species alike, and also unique, at the genomic level [2–6]. The analysis of gene expression by high-throughput sequencing has become the workhorse for many laboratories seeking to understand how cells respond to biochemical and metabolic changes or external stimuli [3,4,7,8]. For all those studies for which there exist a corresponding reference genome, one of the first steps of the computational analysis is the alignment of sequencing reads to its reference [3,9]; for those for which there is no corresponding reference yet, it is common to employ a related genome for the alignment of reads

[9–11]. Although this strategy has led to important discoveries over the last decade about the expression profile, mutation landscape, or sample diversity, for species whose genomes are still unknown, the computational analysis has been constrained primarily to those sequences for which there is a high similarity level compared to their references [10–14]. For dissimilar sequences, however, it is more challenging to assign or infer the biological context for which they function [11,12], and therefore it is also common to set those apart until there is an appropriate cognate reference.

Although sequencing technologies have evolved to yield longer sequencing reads compared to the early beginnings of the genomic era, the vast majority of studies rely on the use of sequencing platforms whose reads range ~50–150 nt in length [1,3,4]. There are a number of computational tools based primarily on two main types of pairwise algorithms which constitute the current and most popular methods for the alignment of sequences within this range: local and global algorithms [15–17]. Local alignments identify similar regions in sequences by determining homology in the presence of rearrangements [18–21]. Global alignments find similarity by transforming the aligned sequences into one another by a combination of simple edits [5,6,22–24]. While most aligners based on these two approaches execute reasonably well for alignments of identical, or nearly identical sequences, their accuracy and robustness decrease considerably as the similarity between sequences declines, particularly for sequences shorter than 30 nt.

In this work, we explored how the dissimilarity affects the alignment of short sequences by comparing the results of several computational tools widely used for the alignment of short sequences. We find that, while most aligners perform reasonably well with identical or nearly identical short sequences, they execute poorly in the alignment of dissimilar ones. Depending on the underlying algorithm, the aligners compared retrieve alignments that differ in the type of errors produced: local aligners tend to clip nucleotides flanking the seed substring, which results in alignment reports that miss matches or mismatches; global aligners do not fail in reporting

matches, but introduce gaps to achieve end-to-end alignments, which results primarily in a low precision score.

The accumulation of these errors sets a challenge for the computational analysis of short sequences and has an impact on all those alignments for which there is no corresponding reference available [10,12]. For example, studies aimed at discovering small RNAs from novel and distant species are inevitably forced to employ genome references other than their own — if the dissimilarity between sequences is extensive, the alignment results may bias the interpretation.

To address this issue, we developed ExtendAlign, a post-analysis tool that provides a significant improvement for the alignment of dissimilar short sequences. ExtendAlign quantifies the identity percentage in the alignment based on the accurate number of matches and mismatches that may initially be missed by a local alignment. Since it incorporates the output of a local algorithm and provides an end-to-end alignment report, ExtendAlign combines the strength of a multi-hit local alignment, with the refinement provided by a query-based global algorithm without applying a clipping strategy. Therefore, its reports position at the intersection between local and global alignments for short sequences. We evaluated the performance of ExtendAlign with simulated and biological data and show that it outperforms other computational tools commonly employed to align short sequences in most of the metrics tested.

By executing multiple alignments with short sequences against distant genomes as references, we show that ExtendAlign is particularly useful to recalculate and pinpoint the identity percentage of alignments that span across the "twilight zone" —the similarity that ranges ~35–50% [25,26]. Finally, we provide one practical example of the utility of ExtendAlign by revisiting the alignment 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 contaminating RNAs from cow into cell lines due to their culture with fetal bovine serum [27,28]. We found that more than 35% of all small RNA sequences considered bovine-

specific present in human cell lines are at least 80% identical to humans. This indicates that — because of the aligner and specific parameters employed— there was a high false-positive discovery rate of bovine-specific small RNAs that contributed to this controversy.

ExtendAlign is recommended for short sequence alignments that require the highest accuracy; or for studies that require a quantitative measure of the dissimilarity level; or for studies where precision in the identity percent cut-off is critical for determining homology between phylogenetically distant short sequences. ExtendAlign was developed as a Nextflow pipeline to guarantee reproducibility and scalability [29], and it is available for download at https://github.com/Flores-JassoLab/ExtendAlign.

RESULTS

Sequence dissimilarity impacts the alignment of short sequences

We first examined the results provided by tools commonly employed for the alignment of short sequences to get an insight into their alignment capabilities. Two identical sequences were aligned with Bowtie, Bowtie2, BWA, BWA-MEM, BLASTn, BLASTn-short, and Needle (Additonal file 1: Figure S1A, left) [18,22,30–33]. Except for BWA-MEM, all aligners retrieved a full-length alignment hit under default parameters. This simple comparison suggests that most of these aligners can handle short sequences if the purpose is the identification of perfect or near-perfect alignments; which is often the case for reads from high-throughput sequencing samples of small RNAs to a related reference genome. However, when two dissimilar sequences were aligned, only Needle and BLASTn-short retrieved alignment hits under default parameters (Suppl. Figure S1A, right). Needle is an aligner based on the Needleman-Wunch algorithm and performs end-to-end alignments [22]. The Bowtie, BWA and BLASTn versions tested are based on local algorithms, and therefore require a seed substring of a minimum fixed size to initiate alignments. The algorithmic approaches employed by Needle and BLASTn-short retrieve different alignments for the same two

130 sequences under default parameters (Suppl. Figure S1B). For example, Needle managed to find the
131 positions matched correctly, but reported overhanging positions as gaps.

132 The fact that only BLASTn-short retrieved alignment results compared to the other local
133 algorithms motivated us to examine in more detail whether this absence of hits was due to the
134 general intrinsic capabilities of local alignments. Hence, we adjusted the parameters of every tool
135 tested to make all seed sizes uniform, and also to maximize their alignment ability —which is
136 concomitantly associated with an increase in the alignment hits (see Suppl. Table S1). Under these
137 permissive parameters, BWA-MEM retrieved an alignment hit (Suppl. Figure S1A, right), albeit
138 different to those of BLASTn and BLASTn-short (Suppl. Figure S1C); Bowtie, Bowtie2, and BWA
139 did not retrieve alignments. Interestingly, regardless of the parameters employed, there were
140 positions with identical nucleotides not reported primarily outside the seed substring and near the
141 5′ or 3′ ends in the query (Suppl. Figure S1B and S1C, bold red letters). We rule out that the
142 permissive parameters chosen prevented the local aligners from performing efficiently since they
143 all found a full-length alignment for the identical sequences (Suppl. Figure S1A, left); arguing in
144 favor of their popularity for aligning sequences this short, despite the recommended use by their
145 developers (Suppl. Table S1) [18,22,30–33].

146 To investigate further how sequence dissimilarity impacts the alignment of short sequences, we
147 simulated query and subject databases and aligned them with all the aligners mentioned above.
148 Inaccuracies in the alignment of short sequences might affect the study of several classes of RNAs,
149 for example, small RNAs [34]. An abundant class of small RNAs are microRNAs, and since their
150 mature sequence sizes peak at ~22 nt in plants and animals [35–38], the simulated query database
151 consisted of 8,500 sequences randomly generated of 22 nt long. The subject database consisted of
152 sequence sizes ranging from 50–170 nucleotides with fixed 7mer seed, each with a randomly
153 chosen position 1, and one or up to fifteen mismatches, insertions or deletions (see Methods for
154 details).

155 Under default parameters, all aligners showed a balanced performance among the metrics tested
156 (except for BWA-MEM, which did not retrieve alignment hits) (Table 1). For example, BLASTn-
157 short retrieved the largest number of hits, at the cost of Sensitivity and Specificity; Needle retrieved
158 the highest number of True Positive Hits and the highest Recovery Rate (7,038 and 0.83,
159 respectively), but was also the least precise of all tools tested; the Bowtie versions tested and BWA
160 retrieved a small number of alignment hits, and their Recovery Rate was also low but showed high
161 Sensitivity compared to the other aligners. Importantly, Needle retrieved an equal number of Hits
162 to queries, while the BLAST versions exceeded this number, reflecting that there was more than
163 one alignment hit per each query. Under permissive parameters, in contrast, all local aligners
164 increased the number of Hits, being Bowtie2 the highest. Both BLASTn and BLASTn-short
165 retrieved the highest number of True Positives, as well as Recovery Rate and Precision (7,202, 0.85,
166 and 0.98, respectively); albeit also presented the least Sensitivity and Specificity (0.26 and 0.001,
167 respectively). BWA-MEM showed a drastic improvement in several metrics, indicating that a
168 parameter adjustment might have a severe impact on sequence alignments.

169 Overall, the results of this simulation imply that there are fundamental limitations for the
170 alignment of dissimilar short sequences with current methods —while some metrics performance
171 improved under specific parameters, others decrease; regardless of the parameters, no aligner
172 excelled in all the metrics tested. Primarily, the local approaches miss identical nucleotides near the
173 5′ or 3′ ends on the alignment, or provide erroneous reports on the number of matches, particularly
174 for dissimilar sequences with several gaps or mismatches flanking the seed substring (Suppl.
175 Figure S1B and S1C, and data not shown). Needle, on the contrary, showed the least precision of
176 all aligners because its alignment report is based on the farthest 5′ and 3′ ends within each
177 alignment (Suppl. Figure S2A and S2B, and data not shown). The cumulative errors caused by
178 either algorithm might give rise to alignment biases, particularly while aligning thousands of
179 dissimilar sequences to establish similarity between short sequences obtained from high-
180 throughput sequencing data. Compromising the alignment accuracy of short sequences may have

181 an impact on the understanding of small RNAs; for example, by restraining the identification of
182 evolutionary relationships that might exist among microRNAs of distant species [12,38].

183 To overcome this problem, we developed ExtendAlign, a post-analysis tool that improves the
184 alignment results of dissimilar short sequences by correcting the errors mentioned above (Figure
185 1A). ExtendAlign identifies the number and identity of all nucleotides flanking a seed substring in
186 a query that might have gone unaligned. After finding unreported nucleotides, ExtendAlign
187 recalculates and reports the total number of matches and mismatches (m/mm) in an end-to-end
188 manner for each query. Therefore, alignment biases are diminished by: i) accounting for all
189 undetected m/mm, ii) considering overhanging nucleotides in the query as mismatches, and iii)
190 extending the alignment to the 5′ and 3′ ends in the query (not the subject). Common examples of
191 alignment errors and their corrections are shown in Figure 1B. Because of its low sensitivity and
192 specificity, but also because of its high recovery rate and precision under default and permissive
193 parameters, ExtendAlign was developed to function after priming the alignments with the
194 BLASTn versions tested in this work. Importantly, BLASTn-short is a version of BLASTn aimed to
195 align short sequences [33,39]; however, the permissive parameters of BLASTn-short and BLASTn
196 of this work are equivalent, and hereinafter are referred to as *high sensitivity*-BLASTn (HSe-
197 BLASTn).

198

199 **ExtendAlign increases the sensitivity and specificity of local alignments**

200 We examined the performance of ExtendAlign using the previous simulated databases and
201 measured all the metrics tested (Table 2). In general, ExtendAlign showed an improved or similar
202 performance in most metrics compared to the other aligners tested under default and permissive
203 parameters. For example, Sensitivity and Specificity increased more than 3-fold and 30-fold,
204 respectively, compared to HSe-BLASTn; which is comparable with the performance score of other
205 aligners. Despite being the highest-ranked compared to other aligners, Precision was the only

metric for which ExtendAlign did not improve in comparison to its priming base. We attribute this result to how the alignment correction takes place: ExtendAlign identifies nucleotides flanking the seed substring that might have gone unnoticed by the local alignments (Figure 1B); thus, the apparent lower Precision score might be a consequence of increasing the alignment length and also decreasing the number of False Negatives.

Next, we analyzed how the alignment correction by ExtendAlign compares to local and global approaches. For this, we chose HSe-BLASTn and Needle as they provide the highest Number of True Positive Hits in our simulation analysis. Since Needle alignments are paired-wised, only 8,500 hits were retrieved, whereas HSe-BLASTn retrieved 4,715,796 hits due to its multiple-hit capabilities (Table 1 and Suppl. Figure S3A). Hence, to compare the results for each alignment approach, all the alignment pairs located by HSe-BLASTn with one-hit only were presented to Needle (4,561,877 pairs) (Suppl. Figure S3B, and Suppl. Info). Finally, the total number of hits for each aligner was plotted as a function of sequence coverage and total alignment size, gaps, or mismatches (Figure 2). HSe-BLASTn retrieved a variety of query sequence coverage that ranged from ~30–100%, and whose total alignment length ranged from 7–22 nt, with a maximum of four gaps, and up to six mismatches per query (Figure 2A–C). Needle, on the contrary, retrieved 100% coverage for every alignment pair, but the alignment size and the number of gaps were raised to 150 nt (Figure 2A and 2B). This 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 (Suppl. Figure S2A and S2B). This result makes Needle impractical for aligning sequences that vary in size, for instance, finding multiple loci for short sequences against whole chromosomes or genomes, which would result in a different type of bias in comparison with local algorithms (Suppl. Figure S3A). Conversely, ExtendAlign does not increase the total alignment size, nor introduces gaps, and most importantly, since it always achieves 100% query coverage, does not miss mismatches flanking the seed substring. Due to these results, we conclude that ExtendAlign is a robust post-analysis tool that refines the alignment of

232 dissimilar short sequences and provides exceptional accuracy because of its end-to-end correction
233 capability. The features and recommended use of ExtendAlign are listed in the Supplementary
234 Table S1.

235 **ExtendAlign increases the number of total matches and mismatches for dissimilar alignments**

236 We further evaluated ExtendAlign to get a better understanding of its precision. If increasing the
237 alignment coverage impacts the Precision score, it should be reflected directly in a concomitant
238 increase in the number of m/mm identified, presumably at positions located near the 5′ or 3′ ends.
239 In a biological context, it is expected to find multiple examples of dissimilar alignments if different
240 short sequences are aligned to a large reference, for example, a whole genome. Hence, we aligned
241 all human microRNAs against the mouse genome and measured directly the number of m/mm for
242 all alignments. Being conserved among mammals, the mouse genome contains several identical
243 loci to most human microRNAs [40], but also a plethora of dissimilar loci to look for m/mm. After
244 the alignments of microRNAs against a related genome reference, however, we observed only a
245 marginal increase in the total number of m/mm after the correction by ExtendAlign compared to
246 those of HSe-BLASTn (Figures 3A and 3B, main graphs, $p < 0.0001$; Kolmogorov-Smirnov $D =$
247 0.005, matches; and $D = 0.016$, mismatches).

248 One explanation for this result is that the amount of perfect, or near-perfect loci spotted by HSe-
249 BLASTn alone, vastly outnumbers the net correction by ExtendAlign (Suppl. Table S2). This is
250 further supported by the results obtained when mature microRNAs were aligned only against the
251 mouse precursor microRNAs (Figure 3C and 3D, Kolmogorov-Smirnov $D = 0.46$, matches; and $D =$
252 0.83, mismatches; $p < 0.0001$). Precursor microRNAs (pre-microRNAs) size range is ~60–75 nt
253 [41,42]. Consequently, an eight-fold increase in the total number of mismatches imply that their
254 alignment to mature sequences largely constrain the chance of finding perfect match hits compared
255 to an entire genome that could mask the correction with ExtendAlign (Suppl. Table S2,
256 mismatches). Therefore, since perfect-match hits are not candidates for correction, we removed

257 them from the m/mm counts. Accordingly, the correction showed a markedly less marginal
258 difference compared to HSe-BLASTn, for both matches and mismatches (Figures 3A and 3B, inset
259 graphs; $p < 0.0001$).

260 Regardless of the subject reference or the exclusion of perfect-matched hits, we noticed a more
261 drastic difference between the total number of matches *vs*. mismatches in all cases —*e.g.*,
262 microRNAs against genome: $D = 0.005$ *vs*. $D = 0.016$, and $D = 0.69$ *vs*. $D = 0.89$; microRNAs against
263 pre-microRNAs: $D = 0.46$ *vs*. $D = 0.83$. This strongly suggested that the correction by ExtendAlign
264 is more effective if more divergent sequences are aligned because mismatches are a direct
265 measurement of similarity [9,43]. For this reason, we tested whether the alignment of short
266 sequences against less conserved genomes showed similar behavior. For this, we used as examples
267 the genome of *Dasypus novemcinctus* (armadillo) because it is at the border of taxonomic
268 classification, and due to the large content of dissimilar sequences in its genes it is difficult to
269 assign the correct taxa [44]; and also the *Ornithorhynchus anatinus* (platypus) genome because it
270 diverged from the mammalian lineage 160 million years ago, and thus possesses a unique blend of
271 morphological and genomic features of mammals, reptiles, birds and fish [45]. The alignment of
272 human pre-microRNAs revealed a significant increase in the total number of m/mm for both
273 genomes after the ExtendAlign correction (Figure 4A–D). Importantly, in contrast to the alignment
274 results observed for mouse genome (Figure 3A and 3B), there is a noticeable difference in the total
275 m/mm compared to HSe-BLASTn in all cases without the need of removing perfect-matched
276 sequences from the alignment, further supporting the utility of ExtendAlign to align dissimilar
277 short sequences. The difference observed is also more drastic in matches compared to mismatches
278 (armadillo: $D = 0.43$ *vs.* $D = 0.91$; platypus: $D = 0.12$ *vs.* $D = 0.36$). Since Precision is calculated based
279 on the number of mismatches (true and false positives), their increase has a direct effect on this
280 metric. Taking together all these results, we conclude that the cumulative bias observed in the
281 alignment of dissimilar short sequences mainly results from mismatched nucleotides not covered

282 during the local alignment; identifying such missing positions reflects therefore an apparent
283 decrease in the precision score.

284

285

286

287 **Pinpointing the similarity percentage of short sequences with ExtendAlign**

288 Next, we evaluated the utility of ExtendAlign to classify short sequences according to their
289 similarity percent against a reference. The lack of similarity between distant short sequences has an
290 impact on predicting secondary structures in nucleic acids [25,26]. As the similarity between
291 sequences decreases and approaches the so-called "twilight zone" —the similarity that ranges
292 from ~35–50% identity, it is generally assumed that secondary structure predictions might not be
293 reliable because alignments at this range tend to obscure the covariance signal [25,26,46]. From the
294 alignments of pre-microRNAs *vs*. the armadillo and platypus genomes, we classified all alignment
295 hits according to their identity percentage to each reference. Fifty percent of all hits achieved by
296 HSe-BLASTn to the armadillo and platypus genomes span across this range, while 25% of hits
297 reached ~35% or less similarity (Figure 5A and 5B). After the ExtendAlign correction, the median
298 identity to each reference increased from 43% (armadillo) and 41% (platypus), to more than 50%
299 each ($p < 0.0001$). The increase above the twilight zone of half the hits indicates an improvement in
300 the alignment of dissimilar short sequences.

301 Finally, we used ExtendAlign to revisit the alignments performed in a recent work that reported
302 that small RNA contained in fetal bovine serum (FBS) might transfer into cell cultures [27]. The
303 extent at which this observation might affect the entire literature about microRNAs function is
304 unknown, and still a matter of debate [28]. In their work, Wei and colleagues performed a
305 computational analysis that encompassed the search for "bovine-specific" RNA sequences
306 contained within public sequence read archives (SRAs) from human samples. By using Bowtie2,

307 the authors considered all the sequences that aligned to the cow genome, but did not to the human
308 genome, as bovine-specific (Figure 6A). Since the analysis was made using SRAs aimed to
309 sequence small RNAs, we employed ExtendAlign to pinpoint the identity percent of all sequences
310 classified as bovine-specific with respect to human (Figure 6B). As expected, there is an ample
311 percentage range at which the bovine-specific sequences redistribute, with the highest abundance
312 peak at 75–77.5%; only a minor proportion was not aligned at all (Figure 6B, NA) —which would
313 be expected only for those sequences accurately classified as bovine-specific. However, we find
314 that ~35% redistributed at 80–98% similarity range (Figure 6B, grey area). Hence, this suggests that
315 a substantial amount of the bovine-specific sequences might have been the result of a high false-
316 positive discovery rate. For example, a bovine a 22mer RNA sequence with an 80% identity to
317 human —that is, four mismatches to its reference— is not necessarily unique to bovine.

318 **DISCUSSION**

319 There has been a substantial effort by many research groups that focused on the development of
320 aligners that reduce the processing time consumed for comparisons between sequences, which has
321 yielded several reliable algorithms that perform robustly in the genomic era [47]. Most tools,
322 however, have focused on improving the recognition of highly similar sequences to find
323 alignments faster for homology and functional studies (27–29,40,45). In parallel to sequencing
324 technologies that have increased the length of sequencing reads, most aligners also have adapted
325 to handle longer sequences with the advantage of being also memory-efficient [4,30]. As a result,
326 the alignments for dissimilar short sequences have been left somewhat unattended, and most
327 algorithms have difficulties in providing accurate alignments for them. In this work, we show that
328 most popular aligners handle perfect- or near-perfect match alignments well, which is not
329 concerning when the interest is to align or map short sequences to a related reference, *e.g.*, human
330 small RNAs to the human genome [12]. However, this situation changes drastically when the
331 similarity between sequences decreases. Although some aligners improved their results by
332 modifying their parameters, others do not retrieve alignment results at all when presented with

333 dissimilar short sequences. For those alignments that do take place, we found that aligners based
334 on local algorithms have errors and tend to miss nucleotides at either 5′ or 3′ ends at regions that
335 flank the seed substring. In contrast, global algorithms do not miss nucleotides but introduce gaps
336 interspersed along the alignment to achieve end-to-end alignments —a useful feature to find *indels*
337 [22,49]. Besides, global algorithms are impractical for finding imperfect alignments for sequences
338 that differ in size, as they do not perform multiple hit alignments to a reference (Suppl. Figure
339 S3A); this limits their utility to alignments with relationships already inferred. Thus, the problem
340 of obtaining accurate reports for the alignment of short sequences has persisted over the years.

341 ExtendAlign is a post-analysis tool that identifies and corrects the errors originated by the
342 alignment of dissimilar short sequences. It incorporates the tabular output of a local aligner and
343 provides a list of all hits with their corrected m/mm for the total length of each query sequence.
344 ExtendAlign improves the alignments by i) extending local alignments in an end-to-end manner;
345 ii) including all nucleotides flanking the seed substring and therefore always fulfills 100% query
346 coverage; iii) not increasing the total alignment size, nor introducing gaps artificially; and iv),
347 increasing the sensitivity and specificity (Table 2). Thus, the results provided in the post-analysis
348 lie at the intersection between those of local and global algorithms. Other algorithms combine local
349 and global approaches (*i.e.*, *Glocal*), but their recommended use and functionality is constrained to
350 sequences longer than 100 nt (data not shown), eliminating its utility for the alignment of small
351 RNAs [50]. In contrast, the recommended minimum size of the query of ExtendAlign is 8 nt, if
352 primed with HSe-BLASTn (Suppl. Table S1).

353 We employed BLASTn as priming base because it provides the highest number of hits compared to
354 the other aligners tested in this work, but we envision ExtendAlign could be adapted to be
355 compatible with other local aligners too. Although ExtendAlign was initially developed to satisfy
356 the need for multiple-hits pairwise alignments in an end-to-end manner of short sequences, it
357 executes robustly with longer sequences, like precursor microRNAs, which are instruments
358 frequently used for the discovery of novel microRNAs.

359 Since ExtendAlign is not an aligner itself, its recovery rate will match that of its priming base. For
360 this reason, it will only correct alignments that took place by the aligner. Selecting HSe-BLASTn as
361 priming base allowed to correct alignments resulting because of the high recovery rate, which also
362 helped to pinpoint m/mm in alignments for short sequences with outstanding specificity and
363 sensitivity that may otherwise go unnoticed by using other commonly employed aligners alone.
364 Since full-length and perfect-match alignments are not prone to improvement, ExtendAlign is
365 more robust in correcting the number of m/mm in more dissimilar alignments —a useful feature
366 for alignments between evolutionarily distant species. Particularly, ExtendAlign refines the
367 alignments of dissimilar short sequences situated within the twilight-zone, the identity percent
368 barrier that impedes to estimate conservation reliably [25,26,46]. Since lncRNAs are subject to weak
369 functional constraint and rapid turnover during evolution [51], their similarity percent spans
370 across this range too [26], and therefore the use of ExtendAlign may aid in the study of their
371 conservation and functionality if long sequences are aligned in shorter segments to a reference.

372 The intriguing finding that RNAs from bovine origin may transfer into cell cultures as a result of
373 incubation with FBS has raised a controversy in the field of research of microRNAs [27,28]. Here
374 we present evidence that at least some of the sequences considered bovine-specific by Wei and
375 colleagues might be the result of a high false-positive discovery rate. Since those bovine-specific
376 sequences were obtained with Bowtie2, the fact that upon a detailed analysis ~35% of them
377 showed high identity when aligned to human, validates the use of ExtendAlign to report the
378 sequence identity for short sequences accurately. As no other tool combines the high accuracy and
379 discovery rate of ExtendAlign in identifying m/mm for the alignment of short sequences, its use
380 reduces the false discovery rate importantly. We anticipate that future studies will tackle with
381 much better precision how this problem truly affects the entire literature in the microRNA field.

382 Since aligners are essential in computational biology, we anticipate that the utility of ExtendAlign
383 can broaden up to other areas affected by the same type of bias in short sequence alignments, like
384 the discovery of novel microRNAs [52,53], tRNA-derived fragments expression regulation [54],

385 and other small RNAs from emerging model organisms [11]; to study homology of lncRNAs [55];
386 or to pinpoint similarity among pathogenic or complex samples (*e.g.*, host-pathogen
387 metagenomics) [56]; to name a few. The sequencing era has allowed us to analyze several species
388 at the genomic level; the prediction of conserved and non-conserved microRNAs among species
389 has become a field of great interest in the last decade. Thus, ExtendAlign can help in reducing
390 false-positives commonly found in homology searches for small RNAs and also to increase the
391 refinement of alignments between dissimilar sequences when high precision is needed.

392

393 **METHODS**

394 **Datasets**

395 Simulations were performed by building 8,500 random 22 nt long query sequences following a
396 uniform probability distribution to model nucleotide content. The subjects were generated by a
397 fixed seed of 7 nt for every sequence, with a start position chosen at random. One, or up to fifteen-
398 nucleotide changes were introduced into the region corresponding to each query sequence. One
399 insertion or deletion (indel) was introduced per every nine mismatches. The length of indels
400 followed an exponential distribution with lambda = 2. Finally, randomly generated nucleotides
401 were added to both ends of the previously mutated query sequences to yield a total length that
402 ranged from 50 to 170 nt.

403 The datasets of mature and precursor microRNAs used as queries and subjects were downloaded
404 from miRBase (release 22) [57,58]. The mouse (*Mus musculus*, mm10p6), armadillo (*Dasypus*
405 *novemcinctus*, v.3.0), and platypus (*Ornithorhynchus anatinus,* v.5.0.1) genomes used as long-
406 sequence subjects were downloaded from NCBI. The SRR515903 dataset for bovine-specific
407 sequence analysis was downloaded from the Gene Expression Omnibus [27,59].

408

409 **HSe-BLASTn setup**

410 BLASTn v2.8.1 [18,39] was used to prime and implement ExtendAlign seeds (Table 1). The
411 parameters of the high sensitivity version of BLASTn, referred to as HSe-BLASTn in this work are
412 as follows: word size: 7; reward: 1; penalty: -1; gap open: 2; gap extend: 2; e-value: 10; DUST: false;
413 soft masking: false. The DUST filtering and soft masking options were disabled in order to keep all
414 query sequences, even when scored as low complexity. HSe-BLASTn databases for subject
415 sequences were built using the makeblastdb command line tool including the parse_seqids option
416 (to keep the original sequence identifiers) and the dbtype nucl (specific for input nucleotide
417 sequences) parameter.

418

419 **ExtendAlign Implementation**

420 ExtendAlign was developed as a sub-modular project wrapped in a Nextflow environment [29]. At
421 its core, multiple sub-modules combine in-house scripts (one for each pipeline stage), and
422 implementations of BLAST+ v2.8.1 [18,39], bedtools v2.27.0 [60] and SeqKit v0.10.1 [61]. The sub-
423 module scripting follows the mk syntax to establish input-output file dependency control [62].
424 ExtendAlign receives DNA/RNA sequences for query and subject in FASTA or multi-FASTA file
425 format. The algorithm runs through the following general stages:

426 1. FASTA formatting. Sequence length is appended to FASTA headers.

427 2. Construction of BLAST database. Subject FASTA file is used to produce a BLAST database.

428 3. HSe-BLASTn alignment. Queries are aligned against subjects using the high sensitivity
429 parameters of the HSe-BLASTn setup.

430 4. Hit filtering. HSe-BLASTn may have reported many alignment hits per query. At this stage,
431 ExtendAlign provides the option to keep "all-hits," or to find the "best-hit." Best-hit was defined
432 as the longest alignment with the least mismatches (including the query mismatches in the gaps).

433 5. Extension coordinates calculation. For each HSe-BLASTn hit, the length of the query and subject
434 extensions is calculated as follows:

435 $len(5'\ ext) = \min\ (aln_{start}(Q)\ , aln_{start}(S)\ \ ) - 1$

436 $len(3'ext) = min(len(Q) - aln_{end}(Q), len(S) - aln_{end}(S))$

437 where Q and S refer to the query and subject sequences, respectively; aln_{start} and aln_{end} refer to HSe-
438 BLASTn alignment start and end position, respectively; and *len* refers to the sequence length.

439

440 The 5′ and 3′ extension regions are described by the vector defined by their start and end
441 coordinates, determined by:

442 $\overrightarrow{5'region\ (x)} = [aln_{start}(x) - len(5'ext), aln_{start}(x)]\ |\ x\ \in \{S, Q\}$

443 $\overrightarrow{3'region\ (x)} = [aln_{end}(x), len(3'ext) + aln_{end}(x)]\ |\ x\ \in \{S, Q\}$

444

445 This pipeline stage yields a set of query and subject sequence coordinates from which nucleotides
446 will be extracted based on each HSe-BLASTn hit. The outlined procedure self-adjusts when dealing
447 with minus strand hits to enable ExtendAlign to work with any strandness configuration of a
448 BLASTn run.

449

450 6. Extraction of extended nucleotides. Using the coordinates from the previous step and the
451 original FASTA inputs for each HSe-BLASTn hit, the nucleotide sequences are appended at the
452 query and subject 5′ and 3′ ends.

453 7. Percent identity recalculation. To allow RNA vs DNA comparison extended "U" nucleotides are
454 transformed to "T". Then, the corresponding (5′ *vs*. 5′ and 3′ *vs*. 3′) extended nucleotides are
455 compared positionally to calculate extension mismatches; no gaps are allowed.

456 The amount of the query covered by the HSe-BLASTn alignment and the extension phase (effective
457 length = eff_{len}) is calculated as follows:

458 $$eff_{len} = len(aln) - aln_{gaps} + len(5'ext) + len(3'ext)$$

459 where *len(aln)* refers to the length of the alignment reported by HSe-BLASTn, and aln_{gaps} is the
460 number of gaps introduced into the query sequence by HSe-BLASTn.

461

462 The total number of mismatches found in the query is calculated as follows:

463 $$total_{mm} = ext_{mm} + aln_{mm}$$

464 where ext_{mm} is the number of mismatches introduced during the extension phase and aln_{mm} is the
465 number of mismatches reported by HSe-BLASTn.

466 The identity percent calculation depends on the difference between the effective length and the
467 query length:

468 $$EA_{pident} = \frac{min(eff_{len}, len(Q) - total_{mm})}{len(Q) + aln_{gaps}}$$

469

470 8. No-hit query appendage. Query names for which HSe-BLASTn did not find hits are appended
471 to the correction table results as "NO_HIT" in the subject field.

472 9. Generation of alignment report. Results are gathered in a tab-separated file that reports all hits
473 with a list of query and subject names, with percent identity before and after the ExtendAlign
474 correction. NO_HIT queries are included in the report.

475

476 **Performance tests**

477 BWA-MEM, BWA [32], BLASTn [18], BLASTn-short[33,39], Bowtie2 (end-to-end) [31,63] and
478 ExtendAlign (all-hits mode) were used to align every simulated query sequence against the subject
479 database. For Needle [22], we designed a computing cycle to align each query sequence against its
480 corresponding subject sequence. Each algorithm was run twice, first with default parameters, then
481 with permissive parameters. Permissive parameters were chosen to maximize the number of
482 alignments to increase the chance of finding the true positive one. Alignments were considered
483 true positive when queries aligned with their respective subjects and when the position of the fixed
484 seed region (established during the generation of the data) matched the expected seed according to
485 our simulated databases (see Datasets above). Recovery percentage was defined as the number of
486 true positive alignments versus the total number of query sequences. The modified parameters for
487 each algorithm are listed in the Suppl. Information. The information contains the corresponding
488 command line used. Each nucleotide from the true positive alignment was assigned to one of four
489 categories: true positive (TP), when the simulation and the aligner agreed in the existence of a
490 change; true negative (TN), when the simulation and the aligner agreed in the absence of a change;
491 false-positive (FP), when the aligner found a change that was not recorded in the simulation; and
492 false-negative (FN), when the aligner did not find a change that was introduced by the simulation.
493 Query nucleotides not included in the alignment span were considered as FN. Subject overhangs
494 were considered as FP. Specificity was calculated as TN / (TN + FP), recall was calculated as TP /
495 (TP + FN) and precision was calculated as TP / (TP + FP).

496

497 **Bovine-specific dissimilarity percentage calculation**

498 The SRR515903 was converted into a FASTQ file with the SRA toolkit (v. 2.9.1). Bovine-specific
499 sequences were extracted by aligning the library to cow (*Bos taurus*, bosTau8, UCSC Genome

Browser) and human (*Homo sapiens*, Hg38p12, NCBI) as reference genomes with Bowtie2 [31,63], using the parameters defined in Wei *et al*. [27]: mode, local; seed length, 25; mismatches in seed, 0. Queries that aligned to the cow genome, and did not to the human genome were extracted into a new FASTA file for downstream analysis with ExtendAlign.

Statistical Analysis

Data were plotted with no bins as cumulative frequency distributions. The 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 with the coefficient of determination as a 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, non-parametric data with a significance value of $p < 0.05$. Simulated data was plotted with the R software. All other plots were done using GraphPad Prism software (v.7).

REFERENCES

1. Su Z, Łabaj PP, Li S, Thierry-Mieg J, Thierry-Mieg D, Shi W, et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol. 2014;32:903–14.

2. Zhang J. The impact of next-generation on genomics. J Genet Genomics. 2011;38:95–109.

3. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet. 2011;52:413–35.

4. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. Springer US; 2019;

523 5. Delcher A, L, Kasif S, Fleischmann RD, Peterson J, White O, et al. Alignment of whole genomes.
524 Nucleic Acids Res. 1999;27:2369–76.

525 6. Delcher AL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids
526 Res. 2002;30:2478–83.

527 7. Guo M, Du Y, Gokey JJ, Ray S, Bell SM, Adam M, et al. Single cell RNA analysis identifies
528 cellular heterogeneity and adaptive responses of the lung at birth. Nat Commun. Springer US;
529 2019;10.

530 8. Manuscript A, Dong C, Zhao G, Zhong M, Yue Y, Wu L, et al. RNA sequencing and
531 transcriptomal analysis of human monocyte to macrophage differentiation. Gene. 2013;519:279–87.

532 9. Liliana GH. A Computer Program for Aligning a cDNA Sequence with a Genomic DNA
533 Sequence. Genome Res. 1998;8:967–74.

534 10. Pai TW, Li KH, Yang CH, Hu CH, Lin HJ, Wang W Der, et al. Multiple model species selection
535 for transcriptomics analysis of non-model organisms. BMC Bioinformatics. 2018;19.

536 11. Ockendon NF, O'Connell LA, Bush SJ, Monzón-Sandoval J, Barnes H, Székely T, et al.
537 Optimization of next-generation sequencing transcriptome annotation for species lacking
538 sequenced genomes. Mol Ecol Resour. 2016;16:446–58.

539 12. Jha A, Shankar R. miReader: Discovering Novel miRNAs in Species without Sequenced
540 Genome. PLoS One. 2013;8.

541 13. Wei Y, Chen S, Yang P, Ma Z, Kang L. Characterization and comparative profiling of the small
542 RNA transcriptomes in two phases of locust. Genome Biol. 2009;10:R6.

543 14. Valdés-López O, Yang SS, Aparicio-Fabre R, Graham PH, Reyes JL, Vance CP, et al. MicroRNA
544 expression profile in common bean (Phaseolus vulgaris) under nutrient deficiency stresses and
545 manganese toxicity. New Phytol. 2010;187:805–18.

546 15. Zhang Z, Schwartz S, Wagner L, Miller W. A Greedy Algorithm for Aligning DNA Sequences. J
547 Comput Biol [Internet]. 2000;7:203–14. Available from:
548 http://www.liebertonline.com/doi/abs/10.1089/10665270050081478

549 16. Polyanovsky VO, Roytberg MA, Tumanyan VG. Comparative analysis of the quality of a global
550 algorithm and a local algorithm for alignment of two sequences. Algorithms Mol Biol. 2011;6:1–12.

551 17. Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC. Cross-species sequence
552 comparisons: a review of methods and available resources. Genome Res [Internet]. Cold Spring
553 Harbor Laboratory Press; 2003 [cited 2019 Aug 10];13:1–12. Available from:
554 http://www.ncbi.nlm.nih.gov/pubmed/12529301

555 18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol
556 Biol. 1990;215:403–10.

557 19. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol.
558 1981;147:195–7.

559 20. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, et al. Human-mouse
560 alignments with BLASTZ. Genome Res. 2003;13:103–7.

561 21. Brudno M, Morgenstern B. Fast and sensitive alignment of large genomic sequences. Proc IEEE.
562 2002;

563 22. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the
564 amino acid sequence of two proteins. J Mol Biol. 1970;48:443–53.

565 23. Morgenstern B. DIALIGN 2: Improvement of the segment-to-segment approach to multiple
566 sequence alignment. Bioinformatics. 1999;15:211–8.

567 24. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, et al. LAGAN and Multi-
568 LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. Genome Res.
569 2003;13:721–31.

570 25. Bremges A, Schirmer S, Giegerich R. Fine-tuning structural RNA alignments in the twilight
571 zone. BMC Bioinformatics. 2010;11.

572 26. Ponting CP. Biological function in the twilight zone of sequence conservation. BMC Biol. BMC
573 Biology; 2017;15:1–9.

574 27. Wei Z, Batagov AO, Carter DRF, Krichevsky AM. Fetal Bovine Serum RNA Interferes with the
575 Cell Culture derived Extracellular RNA. Sci Rep. Nature Publishing Group; 2016;6:31175.

576 28. Tosar JP, Cayota A, Eitan E, Halushka MK, Witwer KW. Ribonucleic artefacts: Are some
577 extracellular RNA discoveries driven by cell culture medium components? J. Extracell. Vesicles.
578 2017.

579 29. DI Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables
580 reproducible computational workflows. Nat Biotechnol. 2017;35:316–9.

581 30. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of
582 short DNA sequences to the human genome. Genome Biol [Internet]. 2009;10:R25. Available from:
583 http://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25

584 31. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;

585 32. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform.
586 Bioinformatics. 2010;26:589–95.

587 33. Bethesda (MD): National Center for Biotechnology Information (US). BLAST® Command Line
588 Applications User Manual. 2008.

589 34. Cech TR, Steitz JA. The noncoding RNA revolution - Trashing old rules to forge new ones. Cell.
590 2014.

591 35. Voinnet O. Origin, Biogenesis, and Activity of Plant MicroRNAs. Cell. Elsevier Inc.;
592 2009;136:669–87.

593 36. Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP. MicroRNAs in plants. Trends
594 Plant Sci. 2002;7:1616–26.

595 37. DP B. MicroRNAs : Genomics, Biogenesis, Mechanism, and Function.pdf. Cell. 2004;116:281–
596 97.

597 38. Arteaga-Vázquez M, Caballero-Pérez J, Vielle-Calzada J-P. A Family of MicroRNAs Present in
598 Plants and Animals. Plant Cell. 2006;18:3355–69.

599 39. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:
600 Architecture and applications. BMC Bioinformatics. 2009;10:1–9.

601 40. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of
602 DNA elements in the mouse genome. Nature. Nature Publishing Group; 2014;515:355–64.

603 41. Provost P, Kim S, Lee Y, Ahn C, Han J, Choi H, et al. The nuclear RNase III Drosha initiates
604 microRNA processing. Nature. 2003;425:415–9.

605 42. Lee Y, Jeon K, Lee J-T, Kim S, Kim VN. MicroRNA maturation: stepwise processing and
606 subcellular localization. EMBO J. European Molecular Biology Organization; 2002;21:4663–70.

607 43. Kent WJ. BLAT - The BLAST-like alignment tool. Genome Res. 2002;12:656–64.

608 44. Tarver JE, Dos Reis M, Mirarab S, Moran RJ, Parker S, O'Reilly JE, et al. The interrelationships
609 of placental mammals and the limits of phylogenetic inference. Genome Biol Evol. 2016;8:330–44.

610 45. Warren WC, Hillier LDW, Marshall Graves JA, Birney E, Ponting CP, Grützner F, et al. Genome
611 analysis of the platypus reveals unique signatures of evolution. Nature. 2008;453:175–83.

612 46. Wilm A, Mainz I, Steger G. An enhanced RNA alignment benchmark for sequence alignment
613 programs. Algorithms Mol Biol. 2006;1:1–11.

614 47. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing.
615 Brief Bioinform. 2010;11:473–83.

616 48. Chen Y, Ye W, Zhang Y, Xu Y. High speed BLASTN: An accelerated MegaBLAST search tool.
617 Nucleic Acids Res. 2015;43:7762–8.

618 49. Sankoff D. Matching Sequences under Deletion/Insertion Constraints. Proc Natl Acad Sci.
619 1972;69:4–6.

620 50. Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, et al. Glocal alignment:
621 Finding rearrangements during alignment. Bioinformatics. 2003;19.

622 51. Kapusta A, Feschotte C. Volatile evolution of long noncoding RNA repertoires: Mechanisms
623 and biological implications. Trends Genet. 2014;30:439–52.

624 52. Friedländer MR, Lizano E, Houben AJS, Bezdan D, Báñez-Coronel M, Kudla G, et al. Evidence
625 for the biogenesis of more than 1,000 novel human microRNAs. Genome Biol. 2014;15:1–17.

626 53. Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, et al. Birth and
627 expression evolution of mammalian microRNA genes. Genome Res. 2013;23:34–35.

628 54. Lee YS, Shibata Y, Malhotra A, Dutta A. A novel class of small RNAs: tRNA-derived RNA
629 fragments (tRFs). Genes Dev. 2009;

630 55. Kapusta A, Feschotte C. Volatile evolution of long noncoding RNA repertoires: Mechanisms
631 and biological implications. Trends Genet. 2014;30:439–52.

632 56. Walker AW, Duncan SH, Louis P, Flint HJ. Phylogeny, culturing, and metagenomics of the
633 human gut microbiota. Trends Microbiol. 2014.

634 57. Kozomara A, Griffiths-Jones S. MiRBase: Annotating high confidence microRNAs using deep
635 sequencing data. Nucleic Acids Res. 2014;42:68–73.

636 58. Kozomara A, Birgaoanu M, Griffiths-Jones S. MiRBase: From microRNA sequences to function.
637 Nucleic Acids Res. 2019;47.

638 59. Guduric-Fuchs J, O'Connor A, Camp B, O'Neill CL, Medina RJ, Simpson DA. Selective

639 extracellular vesicle-mediated export of an overlapping set of microRNAs from multiple cell types.
640 BMC Genomics. 2012;13.

641 60. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features.
642 Bioinformatics. 2010;26:841–2.

643 61. Shen W, Le S, Li Y, Hu F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file
644 manipulation. PLoS One. 2016;11:1–10.

645 62. Hume A. Mk: a succesor to make. T Bell Lab Comput Sci. 1987;445–57.

646 63. Langmead B, Wilks C, Antonescu V, Charles R. Scaling read aligners to hundreds of threads on
647 general-purpose processors. Bioinformatics. 2019;35:421–32.

648

649

650

651 **DECLARATIONS**

652 **ETHICS APPROVAL AND CONSENT TO PARTICIPATE**

653 Not applicable

654

655 **CONSENT FOR PUBLICATION**

656 Not applicable

657

658 **AVAILABILITY OF DATA AND MATERIAL**

659 The datasets generated and/or analyzed during the current study are available in the Flores-
660 JassoLab Github repository: https://github.com/Flores-JassoLab/ExtendAlign.

661

662 **COMPETING INTERESTS**

663 The authors declare that they have no competing interests.

664

665 **FUNDING**

666 This work was supported in part by the Instituto Nacional de Medicina Genómica [08/2017/I-322]
667 and SS/IMSS/ISSSTE-CONACyT [289862] to SEAV; and the Instituto Nacional de Medicina
668 Genómica [05/2017/I-321] and [08/2019/I-407] to CFFJ.

669

670 **AUTHORS' CONTRIBUTIONS**

671 MFT, JIHH, and IAO, constructed ExtendAlign and set GitHub repository; MFT, LGR, HT, SEAV
672 and CFFJ, validated ExtendAlign; MFT, SEAV and CFFJ, envisioned the project and MFT, LGR,
673 HT, SEAV and CFFJ wrote the manuscript. All authors read and approved the final manuscript.

674 **ACKNOWLEDGEMENTS**

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

678

679

680

681 **TABLE LEGENDS**

682 **Table 1.** Alignment performance of commonly used aligners for simulated short sequences.

683 **Table 2.** Alignment performance of short sequences with ExtendAlign.

684

685 **FIGURE LEGENDS**

686 **Figure 1.** ExtendAlign is a post-analysis tool to correct for errors created during the alignment of
687 dissimilar short sequences. A, ExtendAlign (EA) is designed to incorporate the output of a local
688 pairwise aligner as tab-delimited file and improve the alignment results. B, Examples of
689 unreported nucleotides by pairwise local aligners comprising matches or mismatches that
690 originate alignment biases and the correction that takes place after EA. Unreported nucleotides
691 (bold red letters) are identified based on the query length and extended to identify possible m/mm
692 (i–iii). Overhanging positions are reported as mismatches only for the query (ii, underlined grey
693 letters); alignments are based solely on the query length.

694 **Figure 2.** ExtendAlign correction positions at the intersection between local and global approaches.
695 A–C, The one-hit per alignment sequence pairs obtained with HSe-BLASTn (grey) from the
696 simulated databases analysis were presented to Needle (cyan), and compared to ExtendAlign
697 (red). The coverage percent of query aligned was plotted as a function of the total number of hits
698 for the total alignment size (A), the number of gaps (B), and the number of mismatches per
699 alignment (C).

700 **Figure 3.** ExtendAlign improves alignments by increasing significantly the number of matches and
701 mismatches. The total number of matches (A and C), and mismatches per hit (B and D), were
702 plotted as a cumulative fraction for the alignments of human mature microRNAs against the
703 mouse genome (A and B), or against pre-microRNAs (C and D). Insets in A and B correspond to
704 the number of m/mm from alignments whose perfect-matched hits were removed. Grey, HSe-

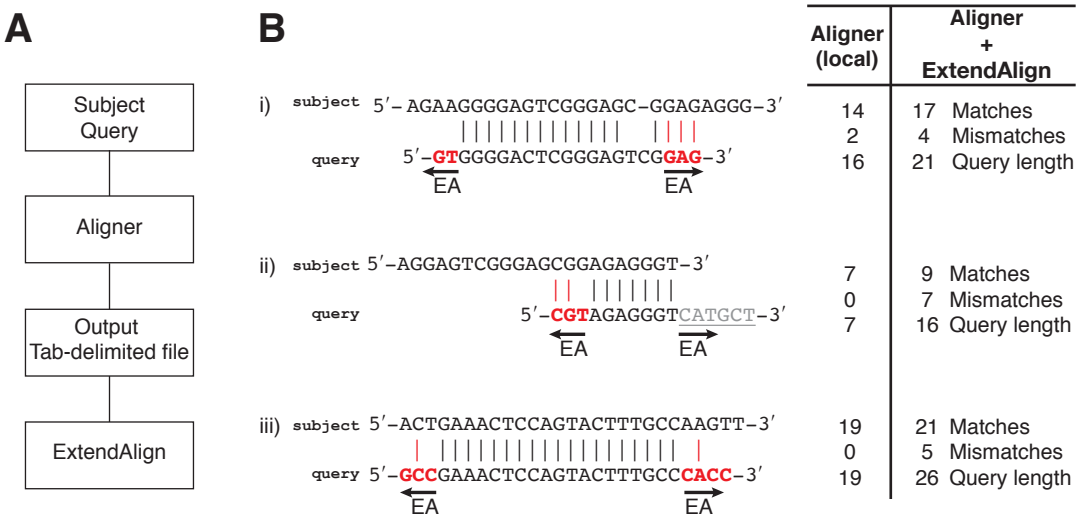
705 BLASTn; red, ExtendAlign. Significance values were calculated using the Kolmogorov-Smirnov
706 test (K-S) for unpaired nonparametric data; D = K-S distance. The non-linear fit curve by least
707 squares as used as measure of goodness of fit.

708 **Figure 4.** Dissimilar alignments are suitable targets for correction. The total number of matches (A
709 and C), and mismatches per hit (B and D), were plotted as a cumulative fraction for the alignments
710 of human pre-microRNAs against the armadillo genome (A and B), or against the platypus
711 genome (C and D). Grey, HSe-BLASTn; red, ExtendAlign. Significance values, K-S distance, and
712 goodness of fit, were measured as in Figure 3.

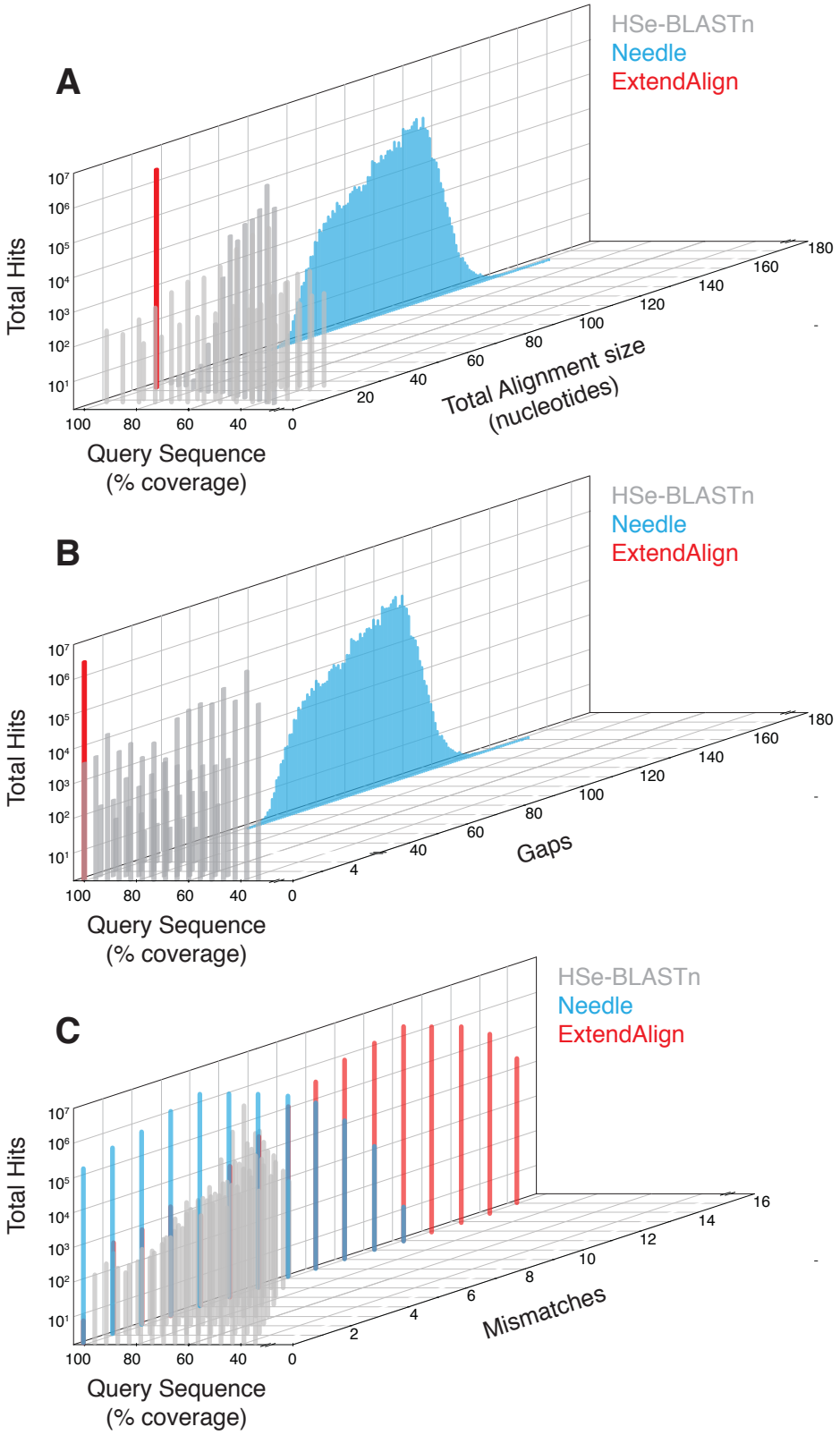
713 **Figure 5.** ExtendAlign is useful to correct alignments that span across the twilight zone. The
714 identity percentage of all human pre-microRNAs was measured as matches per query length and
715 compared for HSe-BASTn and ExtendAlign. A, armadillo genome; B, platypus genome. Box plots
716 are min to max values shown by quartiles; significance was obtained by using the Wilcoxon test for
717 paired, non-parametric data.

718 **Figure 6.** ExtendAlign pinpoints the similarity percentage of dissimilar short sequences. A,
719 Diagram of pipeline followed to revisit the assignment of bovine-specific sequences from public
720 databases with ExtendAlign. B, Histogram showing the abundance per sequence similarity as
721 percentage of identity with the human genome. The grey area corresponds to sequences regarded
722 as bovine-specific but have more than 75% similarity with the human genome. NA, not aligned.

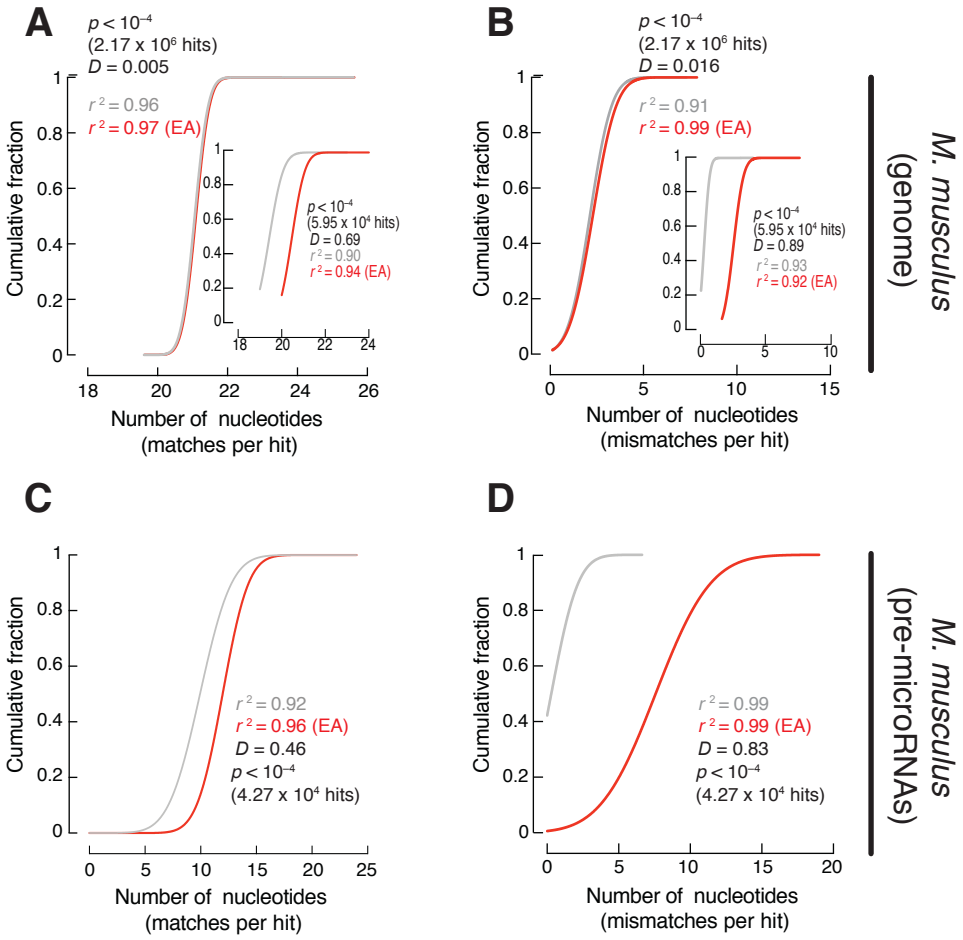
Flores-Torres *et al.*, Figure 1



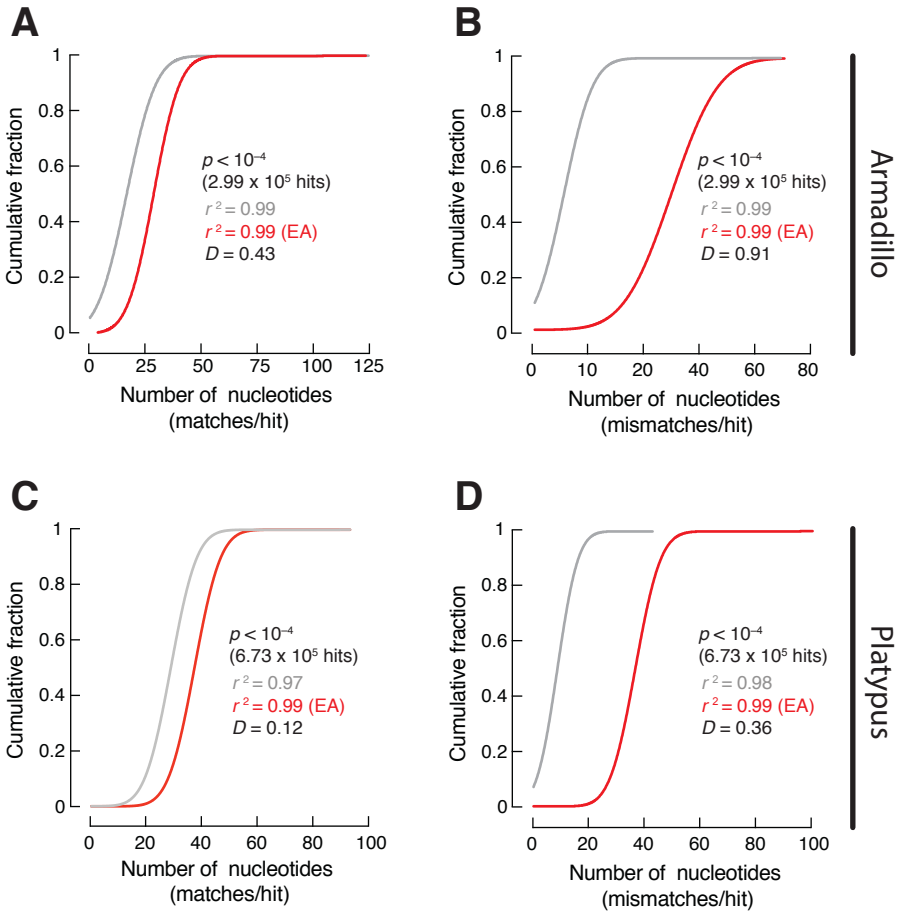
Flores-Torres *et al.*, Figure 2



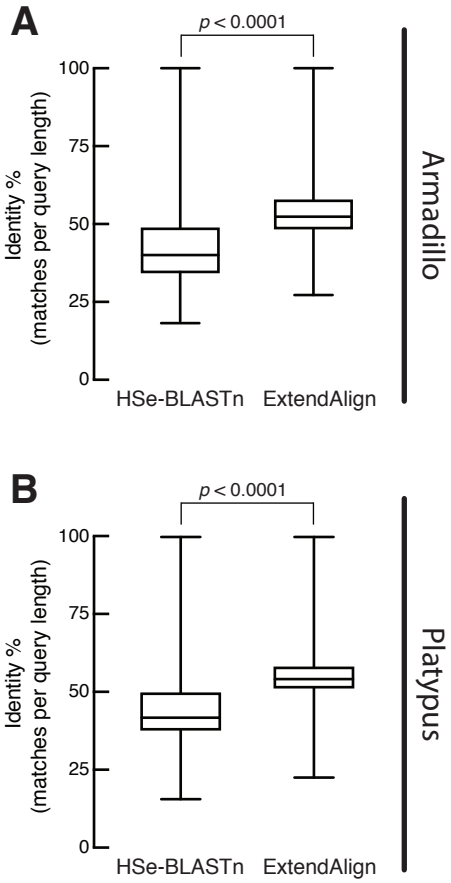
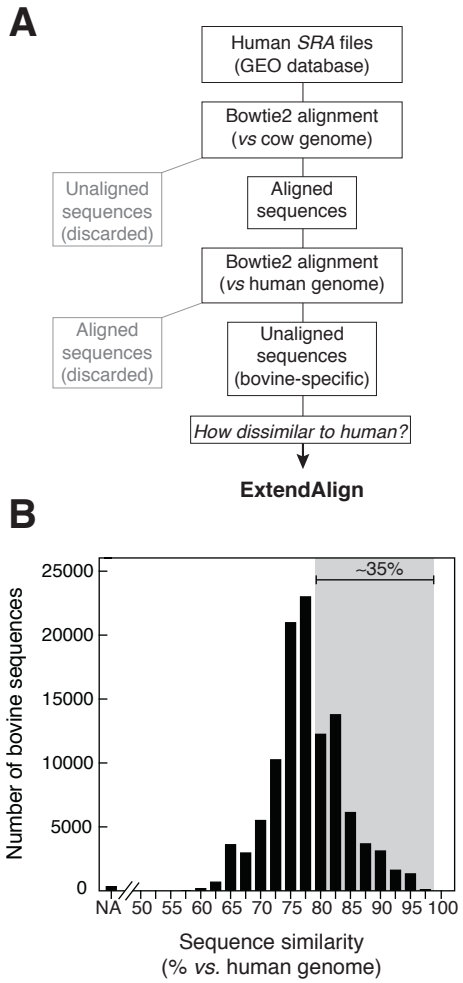
Flores-Torres *et al.*, Figure 3



Flores-Torres *et al.*, Figure 4



Flores-Torres *et al.*, Figure 5



Flores-Torres *et al.*, Figure 6