



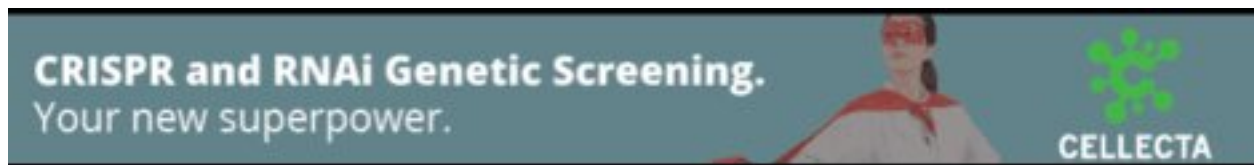
Evaluation of strategies for evidence-driven genome annotation using long-read RNA-seq

Alejandro Paniagua, Cristina Agustin-García, Francisco J Pardo-Palacios, et al.

Genome Res. published online December 23, 2024

Access the most recent version at doi:[10.1101/gr.279864.124](https://doi.org/10.1101/gr.279864.124)

P<P	Published online December 23, 2024 in advance of the print journal.
Accepted Manuscript	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
Open Access	Freely available online through the <i>Genome Research</i> Open Access option.
Creative Commons License	This manuscript is Open Access. This article, published in <i>Genome Research</i> , is available under a Creative Commons License (Attribution 4.0 International license), as described at http://creativecommons.org/licenses/by/4.0/ .
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here .



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

1 Evaluation of strategies for evidence-driven genome
2 annotation using long-read RNA-seq

3

4 Alejandro Paniagua^{1,2*}, Cristina Agustín-García^{1,*}, Francisco J. Pardo-Palacios¹, Thomas
5 Brown^{3,4}, Maite De Maria⁵, Nancy D. Denslow⁵, Camila J. Mazzoni^{3,4}, Ana Conesa¹

6 ¹Institute for Integrative Systems Biology, Spanish National Research Council, Paterna, Spain

7 ²Department of Computer Science, Universitat de València, Valencia, 46100, Spain

8 ³Department of Evolutionary Genetics, Leibniz Institute for Zoo and Wildlife Research Berlin,
9 Germany

10 ⁴Berlin Center for Genomics in Biodiversity Research, Germany

11 ⁵ Department of Physiological Sciences and Center for Environmental and Human Toxicology,
12 University of Florida, Gainesville, FL 32611, USA

13 contact: ana.conesa@csic.es

14 * equally contributing

15

16

17

18

19

20

21

22 **Abstract**

23 While the production of a draft genome has become more accessible due to long-read
24 sequencing, the annotation of these new genomes has not been developed at the same pace.
25 Long-read RNA sequencing (lrRNA-seq) offers a promising solution for enhancing gene
26 annotation. In this study, we explore how sequencing platforms, Oxford Nanopore R9.4.1
27 chemistry or PacBio Sequel II CCS, and data processing methods influence evidence-driven
28 genome annotation using long reads. Incorporating PacBio transcripts into our annotation
29 pipeline significantly outperformed traditional methods, such as *ab initio* predictions and short-
30 read-based annotations. We applied this strategy to a non-model species, the Florida Manatee,
31 and compared our results to existing short-read-based annotation. At the *loci* level, both
32 annotations were highly concordant, with 90% agreement. However, at the transcript level, the
33 agreement was only 35%. We identified 4,906 novel *loci*, represented by 5,707 isoforms, with
34 64% of these isoforms matching known sequences in other mammalian species. Overall, our
35 findings underscore the importance of using high-quality curated transcript models in
36 combination with *ab initio* methods for effective genome annotation.

37 **Introduction**

38 The development of long-read, single-molecule sequencing technologies such as Pacific
39 Biosciences and Oxford Nanopore with the ability to produce ultra-long reads, have radically
40 transformed our capacity for obtaining high-quality, even telomere-to-telomere, genome drafts
41 and have boosted the establishment of several large-scale initiatives to sequence the genomes of
42 all species on Earth. Projects such as Darwin Tree of Life (Blaxter 2022), the Vertebrate
43 Genome Project (Rhie et al. 2021), and the Earth BioGenome Project (EBP) (Lewin et al. 2018)
44 have undertaken the ambitious goal of sequencing the planetary biodiversity by establishing
45 new protocols and pipelines for DNA extraction, sequencing, and genome assembly suitable to
46 any non-model or poorly characterized species.

47 As the catalogue of newly sequenced genomes increases, there is a concomitant necessity for
48 genome annotation, which has not evolved at the same speed and still represents a bottleneck in
49 our efforts to define the gene pools of living organisms (Yandell and Ence 2012). Some of the
50 reasons behind this are the difficulty in annotating large fragmented "draft" genomes, the
51 occurrence of annotation mistakes due to errors and contaminations in draft assemblies
52 (Salzberg 2019), the lack of reference data for many non-model species, and the intrinsic
53 uncertainty in the discovery of functional coding and non-coding elements in newly assembled
54 genomes (Yandell and Ence 2012).

55 Currently, genome annotation approaches can be broadly divided into three main types:
56 evidence-based methods, *ab initio*, and evidence-driven gene predictions (Yandell and Ence
57 2012). Evidence-based annotation or evidence alignment methods use experimental data, such
58 as RNA-seq, protein sequences, or expressed sequence tags (EST) from the organism of interest
59 or related species, which are mapped to the genome to identify genes (Yandell and Ence 2012;
60 Jung et al. 2020). For this approach, RNA-seq has been shown to overperform other forms of
61 evidence, since this type of data is able to capture splice sites, exon, and alternative spliced exon
62 boundaries (Yandell and Ence 2012). However, evidence-based methods have several
63 downsides, one of them being the amount of data available and tissues from which samples can
64 be obtained (Liang et al. 2009; Scalzitti et al. 2020), which is oftentimes limited for non-model
65 organisms, resulting in incomplete annotations. Another drawback of these methods is the
66 propagation of incorrect annotations when using information from related species (Scalzitti et
67 al. 2020). Recently, long-read transcriptome sequencing (lrRNA-seq) has started to be used for
68 this purpose (Zhang et al. 2022; Peng et al. 2022). lrRNA-seq has the potential to capture full-
69 length transcripts and reveal the complexity of the transcriptomes. However, lrRNA-seq data
70 also contain sequencing and library preparation errors and benchmarking studies have
71 demonstrated limitations of these data to deliver accurate transcript models (Pardo-Palacios et
72 al. 2024b).

73 *Ab initio* approaches use mathematical models, like Support Vector Machines (SVMs) or hidden
74 Markov models (HMMs), to predict genes from the genomic sequence (Yandell and Ence 2012;
75 Scalzitti et al. 2020). In this strategy, the models combine signal sensors (e.g. splice sites or
76 polyadenylation signals) and content sensors (e.g. nucleotide composition or length of exons
77 and introns) to make the predictions (Huang et al. 2016; Scalzitti et al. 2020). The main
78 advantage of this approach is that no experimental evidence is needed (although it can be used,
79 see next section) to detect genes and their genic structures (Yandell and Ence 2012), allowing
80 the discovery of unidentified genes (Scalzitti et al. 2020). However, *ab initio* prediction
81 programs tend to be error-prone, detecting non-coding nucleotides in exons (Scalzitti et al.
82 2020) and have low accuracy when predicting gene structures (Yandell and Ence 2012).
83 Another limitation is that the majority of predictor tools tend to identify genetic coding
84 sequences (CDS) rather than untranslated regions (UTRs) or alternative isoforms, rendering
85 incomplete annotations. In addition, *ab initio* methods need a trained model, which ideally
86 should be organism-specific (Yandell and Ence 2012), and is not always available (Yandell and
87 Ence 2012; Scalzitti et al. 2020).

88 Evidence-driven genome annotation is a powerful alternative for improving the detection of
89 genes in genomes. In this case, experimental evidence such as gene or protein expression is used
90 as support during the gene predictions by the *ab initio* programs, in order to increase their
91 precision and overcome the sensitivity limitations of the evidence-alone methods. Several tools,
92 such as AUGUSTUS (Stanke and Waack 2003; Stanke et al. 2006) and SNAP (Korf 2004), and
93 popular annotation pipelines, like BRAKER (Stanke et al. 2008; Hoff et al. 2019) and MAKER
94 (Campbell et al. 2014) have successfully implemented this strategy. Moreover, although the
95 majority of genomes have been and are still being annotated using short-read RNA-seq as
96 evidence, lrRNA-seq is increasingly being generated to serve this purpose. Potentially, long-
97 reads used as transcriptional data for evidence-driven approaches could overcome some of the
98 limitations of the short-reads to faithfully inform complete gene structures. However, no
99 extensive studies have been carried out to evaluate the best strategy for using lrRNA-seq in the

100 evidence-driven approach, leading to a disconnect between the latest sequencing technologies
101 and the genome annotation pipelines (Cook et al. 2019).

102 In this work, we investigate various alternatives for evidence-driven genome annotation using
103 lrRNA-seq data. Specifically, we evaluate different sequencing technologies (PacBio Sequel II
104 and Nanopore R9.41. chemistry) and read pre-processing levels (from raw-reads to
105 reconstructed transcripts and gene models). For benchmarking purposes, we used the well-
106 annotated human genome and long-read cDNA reads of the human cell line WTC11 generated
107 by the Long-read RNA-seq Genome Annotation Assessment Project (LRGASP) (Pardo-
108 Palacios et al. 2024b). We exemplify the utility of the best-performing approach on the
109 annotation of the non-model species, *Trichechus manatus latirostris* (Florida manatee).

110 **Results**

111 Evidence-driven genome annotation adds extrinsic evidence to *ab initio* algorithms to improve
112 gene prediction. This approach requires careful consideration of both the modelling and
113 prediction stages at which the evidence is incorporated, in addition to the type of supporting
114 data utilized. We evaluated the utilization of lrRNA-seq data at both stages, the type of
115 sequencing technology employed, and the level of data pre-processing. To provide a sound
116 benchmarking scenario, cDNA long-read data from the human WTC11 cell line was used and
117 results were compared to the GENCODE human annotation (Supplemental Fig. S1A,B). After
118 determining the best procedure for training and prediction, we assessed the volume of
119 sequencing data required to achieve optimal outcomes and compared this strategy to evidence-
120 driven gene predictions supported by Illumina short-reads (Supplemental Fig. S1C). Lastly, we
121 applied the long-read evidence driven approach to the annotation of the Florida manatee
122 genome, recently sequenced by the LRGASP consortium (Pardo-Palacios et al. 2024), using
123 blood and brain lrRNA-seq data as experimental evidence sources. We then compared the long-
124 read supported genome annotation to the existing NCBI *Trichechus manatus latirostris*
125 Annotation Release 102 annotation obtained using short reads (Fig. 1).

126 **Long-Read Sequencing Data for Curated Gene Set Generation and HMM** 127 **Training**

128 We first evaluated the utilization of long-reads as evidence for model training using the human
129 WTC11 cell line data from LRGASP. Training the HMM model requires the use of a set of
130 high-quality, non-redundant gene models, which in this case should be obtained from the long
131 reads. Due to the potential noisiness of raw lrrRNA-seq data, we envisioned three levels of data
132 preprocessing and assessed their suitability as reliable training sets. The first level represents a
133 minimal read preprocessing scenario aiming at removing read redundancy by collapsing them
134 by their junction pattern into *unique junction chain (UJC)* sequences, without further read
135 correction or filtering. The second preprocessing level used the transcript models reconstructed
136 with a suitable lrrRNA-seq analysis pipeline (Iso-Seq3 for PacBio data, and FLAIR for
137 Nanopore and PacBio+Nanopore datasets) coupled to basic filtering to provide a set of reliable
138 error-corrected transcripts that included alternative isoforms. The third preprocessing level
139 collapsed transcripts belonging to the same gene in one transcript model per gene to reduce
140 sequence redundancy. See Supplemental Table S1 for a detailed description of the number of
141 elements at each step. The characteristics of the data resulting from these three preprocessing
142 levels were inspected by running SQANTI3 and BUSCO analyses.

143 Using SQANTI3, we found that most UJC had at least one splice-site that was not present in the
144 reference, i.e. were Novel-Not-in-Catalog (NNC) transcripts, whether we considered Nanopore,
145 PacBio reads, or the combination of both (Fig. 2A). In contrast, the transcript reconstruction
146 pipelines generated transcriptomes with a significant reduction in the proportion of NNC at the
147 transcript level (Fig. 2A). One key aspect of our pipeline was the removal of monoexons, as
148 between 62.5% and 82.1% of the obtained monoexons did not match any known transcript and
149 were overly present in the two transcriptomes containing Oxford Nanopore Technologies
150 (ONT) data (Supplemental Fig. S2). Other filtering criteria included in our pipeline were the
151 presence of coverage of splice-junctions by short-reads and the number of long-reads associated
152 with each transcript. Moreover, we kept only those transcripts with at least one BLAST hit, a

153 query coverage over 85% and E-value lower than 1×10^{-50} . This, combined with previous
154 filtering steps, reduced the proportion of NNC transcripts to below 10% for the PacBio
155 transcriptome generated using Iso-Seq3 and the two FLAIR transcriptomes using either ONT
156 data or a combination of ONT and PacBio reads (MIX). At this preprocessing level, most
157 transcripts were either contained in the reference (Full-Splice-Match, FSM), or displayed a
158 novel combination of annotated donor and acceptor sites (Novel-In-Catalog, NIC), indicating an
159 improvement in the reliability of the transcript models.

160 The third preprocessing level aimed at reducing gene redundancy, which was still high,
161 especially for the PacBio transcriptome. Non-redundancy is important to avoid overfitting
162 during training (Hoff and Stanke 2019). To reduce the redundancy, we collapsed transcripts
163 using TAMA and then applied a second clustering step at the protein level using CD-HIT. We
164 evaluated the efficiency of this last preprocessing step using the BUSCO dataset eutheria_odb10
165 including 11,366 genes. For the Iso-Seq transcriptome, the fraction of duplicated BUSCO genes
166 was reduced from an initial value of 0.75 at the transcript level to 0.05 after clustering using
167 CD-HIT. We obtained similar results for the ONT and MIX transcriptomes, with a final
168 proportion of duplicated BUSCO genes of 0.04 and 0.03, respectively (Fig. 2B), suggesting a
169 successful collapse of transcripts into gene models without compromising the discovery of true
170 BUSCOs (Supplemental Fig. S3).

171 Next, we used the three sets of collapsed non-redundant transcripts, derived from the PacBio,
172 ONT, and MIX long-reads, to train the *ab initio* model and assessed their performance. For
173 completeness, we included a fourth training set consisting of the BUSCO genes identified in the
174 human genome, as this is a common strategy for *ab initio* gene prediction (Simão et al. 2015).
175 Generally, to evaluate the performance of the gene prediction models, the gene set is divided
176 into training and test sets. The training set is used to train the model, which is then evaluated by
177 predicting genes in the test set and comparing predictions to the true gene annotations.
178 However, it has been shown that this evaluation method overestimates the performance of the
179 models (Guigó et al. 2000). To address this problem, rather than using a long-read-derived test

180 set, we reannotated Chromosome 19 of the human genome. The predictions obtained with the
181 trained models were compared to the 1420 protein-coding genes in the reference annotation of
182 Chromosome 19. Performance was evaluated as sensitivity, precision, and F1 score at the
183 nucleotide (nt), exon, and gene levels. We considered predicted genes as true positives (TP)
184 only if they matched the CDS of a gene in the reference annotation from start to end positions,
185 with all splice sites correctly annotated. To identify the best set of parameters for each type of
186 data, we tested different values for the flanking region used to learn non-coding region patterns
187 and the number of genes included in the training set. We found that, in general, shorter flanking
188 regions (around 1000 bp) and a training set size of 4000-5000 genes provided the best
189 performance results, beyond which no further improvement was observed. Additionally, some
190 data-type specific behaviors were noted (Supplemental Fig. S4). This pattern was observed at
191 the nt, exon, and gene levels and was especially notable for the BUSCO training set.

192 We then evaluated the performance of the best model obtained with each data type. The F1-
193 score at the gene level ranged from 0.13 to 0.16 (Fig. 2C), with models obtained by BUSCO and
194 PacBio data scoring the highest. Sensitivity was very similar between the four models, while the
195 precision for the BUSCO training set was slightly better (Fig. 2C). Next, we characterized the
196 exon number and length of the predictions. We denoted as true positive (TP) genes those with
197 all their junctions and ends matching a gene in the reference, partial true positive (PTP), those
198 with partial overlap with a reference gene, false genes (FG), those gene predictions that did not
199 overlap with any gene in the reference, and missed genes, the genes that were completely
200 missed by the gene prediction algorithm. The challenges associated with *ab initio* gene
201 prediction tools are well-documented, as these tools tend to overestimate the presence of genes
202 (Wang et al. 2003; Dragan et al. 2016; Scalzitti et al. 2020). We found that the number and
203 length of exons in these four types of predicted genes were consistent among the different
204 training sets. TP genes had generally fewer and longer exons than other groups (Supplemental
205 Fig. S5A, B), revealing that these gene structures are easy to predict. FN genes, on the contrary,
206 while displaying a similar number of exons, had shorter exons (Supplemental Fig. S5A, B). This

207 finding agrees with research by Scalzitti et al. (2020) which demonstrated that AUGUSTUS
208 accuracy decreases for genes with short exons. Finally, FGs typically had fewer but longer
209 exons compared to all predictions (Supplemental Fig. S5A, B), with BUSCO training resulting
210 in the lowest (635) and ONT in the highest (1,082) number of false genes. Based on these
211 results we concluded that, at least for our experimental settings, long-read-based transcript
212 models did not provide an advantage over employing BUSCO genes for AUGUSTUS *ab initio*
213 gene prediction and selected this last approach for our subsequent analyses.

214 **PacBio transcript models achieve higher performance when used as experimental** 215 **evidence in the gene prediction step**

216 Once established as the best approach for model training, we evaluated the utilization of lrrRNA-
217 seq data at the gene prediction step in the evidence-driven strategy implemented by
218 AUGUSTUS. We reannotated Chromosome 19 of the human genome HMM model trained with
219 BUSCO genes and provided as experimental evidence each of our three types of long-read
220 preprocessed data: a) FLNC PacBio reads and raw ONT long-reads independently and in
221 combination; b) filtered protein-coding transcripts obtained with transcript reconstruction
222 algorithms, and c) collapsed transcripts derived from those transcriptomes. Note that
223 AUGUSTUS developers recommend using PacBio long-reads with minimum processing for
224 evidence-driven annotations (Hoff and Stanke 2019).

225 We found that the utilization of reads as experimental evidence, particularly from the ONT
226 platform, resulted in the lowest precision and sensitivity (Fig. 2D). We hypothesize that this
227 outcome was due to the presence of reads with unannotated splice sites or aligning to non-
228 coding regions, which lead to false gene predictions, supporting the notion that pre-processing
229 of raw reads improves evidence-driven annotation. Employing transcript models as external
230 evidence provided superior performance than the non-redundant, collapsed transcripts set. In
231 particular, the best performance was achieved using the transcriptome inferred from PacBio data
232 as experimental evidence. This transcriptome recovered 959 protein-coding genes encoded in

233 Chromosome 19, whereas the collapsed version of this transcriptome recovered only 587 genes.
234 The combined PacBio and ONT transcriptome achieved a slightly inferior F1-score compared
235 with the results obtained with the PacBio transcriptome, whereas the ONT transcriptome
236 resulted in the lowest F1-score (Fig. 2D).

237 These results show that the genome annotation substantially improves when long-reads are
238 processed into transcripts models or collapsed transcripts and provided as experimental
239 evidence to *ab initio* algorithms for prediction, improving both sensitivity and precision.

240 **Evidence-driven lrrNA-seq genome annotation reaches an optimal trade-off**
241 **between sensitivity and precision with increasing sequencing depth.**

242 Once we established that processing long reads into transcript models was the best choice for
243 leveraging lrrNA-seq data in genome annotation, we asked to which extent the amount of
244 sequencing throughput impacts annotation. We sampled various proportions of the total PacBio
245 reads and conducted both evidence-based and evidence-driven annotation procedures, using the
246 annotation of human Chromosome 19 as ground truth. We evaluated sensitivity and precision as
247 in previous sections but additionally studied the proportion of missed *loci*. A *locus* was
248 considered missed if no annotation features overlapped it. This metric demonstrates how the
249 presence of imperfect but potentially meaningful predictions changes with sequencing depth in
250 the two annotation approaches.

251 While the evidence-based model achieved higher sensitivity, the precision dropped drastically
252 when more data was incorporated (Fig. 3A). On the contrary, the evidence-driven approach
253 resulted in both metrics reaching a plateau when approximately half of the data was included in
254 the predictions, representing 3,501,946 FLNC reads. Despite the sensitivity being improved
255 when more reads were used, this increase was not as pronounced as the values obtained with the
256 evidence-based approach. However, the precision obtained with the evidence-driven annotation
257 method increased as more data was added at the prediction step, compared with the precision
258 values of the evidence-based approach. In this sense, the evidence-driven strategy reached a

259 trade-off between sensitivity and precision, limiting the incorporation of low-quality predictions
260 and redundancy in the annotation. In addition, the evidence-driven approach was able to capture
261 more *loci* even when small amounts of data were provided as extrinsic evidence (Fig. 3B). This
262 contrasted with the evidence-based strategy, where despite increasing recall as more reads were
263 used, the number of *loci* captured was still lower than the evidence-driven annotation method,
264 especially at the lowest sequencing depth. We concluded that the evidence-driven approach
265 overcomes one of the main limitations of evidence-based methods, which is the amount of
266 available data, while providing a better balance between precision and recall.

267 These results collectively highlight the superiority of evidence-driven annotation approaches
268 over purely experimental methods in achieving a good balance between sensitivity and precision
269 when working with limited lrrNA-seq data.

270 **The evidence-driven approach outperforms the rest of genome annotation** 271 **approaches when long-read technologies are used**

272 We finally asked if evidence-driven genome annotation using long reads outperforms traditional
273 approaches such as *ab initio* methods or annotation based on short-reads. For this, we annotated
274 the human Chromosome 19 using Illumina data in both evidence-driven and evidence-based
275 strategies. Moreover, we evaluated the utilization of either Illumina raw reads, as recommended
276 by AUGUSTUS developers, or a short-read assembled transcriptome as the source of
277 experimental evidence. Performance was compared to the PacBio transcriptome evidence-based
278 and evidence-driven annotation and with the AUGUSTUS *ab initio* approach.

279 As expected, the lowest precision and sensitivity were obtained with the *ab initio* method (Fig.
280 3C), possibly due to the identification of fragmented and false genes, as frequently described
281 (Scalzitti et al. 2020). Notably, when considering evidence-based approaches, employing an
282 Illumina assembled transcriptome resulted in higher precision than using PacBio transcripts,
283 while the sensitivity achieved with the long-read transcript models was higher compared with
284 Illumina. The F1-score of both evidence-based approaches was very similar, 35.1% and 35.9%

285 for the PacBio and Illumina transcriptomes, respectively (Fig. 3D). In the case of the evidence-
286 driven methods, the lowest precision and sensitivity values were obtained when raw Illumina
287 reads were used as the source of evidence. This again underscores the limitations of raw reads
288 as evidence source for evidence-driven genome annotation as shown for the long reads. Despite
289 the improved F1-score when Illumina transcript models were provided for evidence-driven
290 annotation, the best performance of this strategy was achieved by employing the PacBio
291 transcriptome. Furthermore, the combination of the evidence-driven method and long-read
292 technologies outperformed the rest of the approaches achieving a balance between sensitivity
293 and precision leading to the highest F1-score (Fig. 3E). We also attempted to incorporate short-
294 read data to retain transcript models with coverage of all their splice-junctions; however, the
295 combination of short and long read data did not yield improved results (Supplemental Table
296 S2).

297 Altogether, these results show that the combination of long-read-derived transcript models with
298 the evidence-driven annotation method provides superior results than other strategies,
299 establishing new guidelines for genome annotation

300 **Evidence-driven annotation of the Florida manatee genome**

301 Given that the PacBio transcriptome evidence-driven strategy achieved the highest sensitivity
302 and precision for gene prediction in the human genome, we used this approach for the
303 annotation of the Florida manatee genome draft obtained by the LRGASP consortium (Pardo-
304 Palacios et al. 2024b). We compared the results with other annotation methods by analyzing
305 BUSCO completeness. This analysis measures the proportion of genes that are complete (C),
306 either as single-copy (S) or duplicated (D), as well as fragmented (F) and missing (M) genes,
307 out of the total 11,366 genes in the Eutheria_odb10 dataset. The available PacBio lrrNA-seq
308 dataset consisted of 5,434,382 FLNC blood and 2,709,782 FLNC brain reads. Using the Iso-Seq
309 pipeline, we generated 461,040 transcript models. SQANTI3 filtering was applied to remove
310 transcripts with non-canonical splice junctions, intrapriming artifacts, and RT-switching

311 (Cocquet et al., 2006) while retaining only multi-exon transcripts with coding sequences (CDS)
312 longer than 300 nucleotides. Although a high-quality reference gene annotation was
313 unavailable, we showed in the well-annotated WTC11 dataset that this feature-based filtering
314 effectively reduced the proportion of NNC and intergenic transcripts. Following this stringent
315 filtering, we identified 54,491 high-confidence transcript models. This final BUSCO
316 completeness was C:50.1%[S:14.0%, D:36.1%], F:3.1%, M:46.8%, n:11366 (Fig. 4A).

317 Following our benchmarking guidelines, the AUGUSTUS HMM model was trained with the
318 BUSCO genes identified in the LRGASP Florida manatee genome. This model predicted
319 15,735 genes on the manatee genome, with a final BUSCO completeness of C:30.7%[S:30.4%,
320 D:0.3%], F:12.3%, M:57.0%, n:11366. After adding the PacBio transcriptome during the
321 prediction step for an evidence-driven strategy, 35,662 transcripts and 20,782 genes were
322 predicted representing BUSCO completeness of C:61.3%[S:44.4%, D:16.9%], F:10.3%,
323 M:28.4%, n:11366 (Fig. 4A). Compared to the evidence-based method, the evidence-driven
324 approach achieved a higher BUSCO completeness with almost 20,000 fewer transcripts. These
325 results are in agreement with the results obtained during the human benchmark, where the
326 evidenced-driven method controlled the number of missed *loci* obtained while balancing
327 precision and sensitivity.

328 To further complete the annotation, we incorporated the BUSCO genes identified in the manatee
329 genome with the PacBio transcriptome and used these two sources of evidence in combination
330 during the gene prediction step. This strategy resulted in a BUSCO completeness of
331 C:87.9%[S:56.4%,D:31.5%],F:4.4%,M:7.7%, n:11366 and contained a total of 21,082 genes
332 and 39,977 transcripts. This transcriptome was used for downstream analysis.

333 **Comparison of the long-read based (LRB) manatee genome annotation with the** 334 **NCBI annotation**

335 We compared our long-read-based (LRB) annotation to the available NCBI *Trichechus manatus*
336 *latirostris* Annotation Release 102, obtained using short-reads and the NCBI Eukaryotic

337 Genome Annotation Pipeline, to evaluate any gain or loss of information. We first determined
338 the number of isoforms per gene in each of the annotations. Although the NCBI and LRB
339 annotation had a similar proportion of genes that only encode one isoform (65.73% and 66.42%,
340 respectively), we noticed that the number of genes coding multiple isoforms was slightly
341 different. The public annotation contained a higher number of genes with two or three isoforms
342 (22.62%) compared to the LRB one (19.26%). In contrast, the proportion of predicted genes at
343 the LRB annotation with 4 or more isoforms was 14.31%; while in the case of the NCBI
344 annotation, this percentage was 11.64% (Fig. 4B).

345 To further understand these differences, we then used the SQANTI3 framework to classify the
346 LRB transcript models when compared against the NCBI annotation as reference annotation,
347 and the NCBI transcripts compared against the LRB models as reference. In this comparison,
348 identical annotations are identified as Full-Splice-Match (FSM) in both directions, and
349 differences are revealed as other SQANTI3 transcript categories (Supplemental Fig. S6A). The
350 reciprocal comparisons returned 35.11% of transcripts as FSM when assessing NCBI annotation
351 against the LRB reference, while 24.11% were FSM when LRB models were compared to the
352 NCBI annotation. In addition, 85.7% of the transcripts in the LRB transcriptome and 90.2% of
353 the transcripts in the NCBI annotation could be assigned to *loci* found in both annotations. In
354 these shared *loci*, transcripts classified as Incomplete-Splice-Match represent new isoforms with
355 alternative TSS or TTS, while Novel-in-Catalog and Novel-Not-in-Catalog indicate transcripts
356 with different splicing patterns (Supplemental Fig. S6A). These results revealed that, despite an
357 overall good agreement between the two annotations in the identification of expressed *loci*, gene
358 and transcript models were overly different, as only 25-35% of these were perfect matches of
359 each other, suggesting the source and/or strategy for using experimental evidence strongly
360 impacts annotation results.

361 Next, we focused on the isoforms assigned to novel *loci* in each annotation. Using cuffcompare,
362 we identified 4,906 novel *loci* exclusively present in the LRB annotation and 2,225 *loci* unique
363 to the NCBI annotation. To better characterize these *loci*, we evaluated the support for the

364 isoforms found in these *loci* using known sequences from other mammalian species. Of the
365 5,707 isoforms of the LRB annotation associated with novel *loci*, 3671 had at least one BLAST
366 hit against the UniProt Mammalia curated protein database with an E-value lower than 1×10^{-3} .
367 Of the total 2036 transcripts without a BLAST hit, 693 (34.04%) were supported by
368 experimental evidence, while for the transcripts with a BLAST hit, 2481 (67.58%) were
369 supported (Supplemental Fig. S6B). In comparison, for the NCBI annotation, 3,137 out of 3,212
370 isoforms had at least one BLAST hit. These results indicate the capacity of long-read methods
371 to support the annotation of novel genes, while also highlighting the known risk of false
372 discoveries associated with *ab initio* methods (Scalzitti *et al.* 2020).

373 Next, we characterized events of gene fragmentation, where genes are fragmented in one
374 annotation compared to the other. An example is shown in Supplemental Figure S6C. In this
375 case, we found that the public annotation had 1610 isoforms overlapping multiple genes of the
376 new annotation, while the LRB annotation only has 496 putative fusion isoforms, suggesting a
377 higher level of fragmentation in the LRB annotation. We hypothesised that this potential
378 fragmentation may result from the *ab initio* predictions in absence of experimental evidence.
379 However, 1,117 out of the total 1,610 NCBI fusion isoforms overlapped multiple
380 experimentally supported LRB genes. These experimentally supported fragmented LRB genes
381 exhibited a similar median length and number of exons as the other supported LRB isoforms
382 (Supplemental Fig. S6D,E). Since the distribution of lengths and number of exons of these LRB
383 fragmented genes was similar to the non-fragmented genes, we evaluated if the fragments
384 originated from the same gene. We used BLAST to analyze fragmented genes of the LRB
385 annotation overlapped by one NCBI isoform to determine if these supported genes were part of
386 the same human gene. Of the 1087 NCBI fusion isoforms that overlapped supported LRB genes
387 with at least one BLAST hit, 1041 had LRB genes matching the same BLAST hit. This result
388 suggests that the LRB annotation included supported, but fragmented genes.

389 To understand if this fragmentation was already present in the experimental evidence provided
390 for the predictions or caused during the gene prediction step, we used SQANTI3 to evaluate the

391 experimental evidence of the fragmented supported genes of the LRB transcriptome. We
392 detected 100 isoforms from the experimental data that overlapped multiple genes of the LRB
393 annotation, corresponding to 54 different *loci*. Of these 54 *loci*, 17 lacked experimental evidence
394 to support any of the fragmented genes, while in the other 37 cases, at least one of the
395 fragmented genes was supported by the experimental data (Supplemental Fig. S7A, B). These
396 results suggest that while gene fragmentation can occur during gene prediction, the primary
397 cause is the presence of fragmented experimental evidence. Since we did not observe evidence
398 of gene fragmentation when benchmarking our evidence-based approach, we asked if
399 differences in the quality of the manatee lrRNA-seq data with respect to the human WTC11
400 PacBio reads could explain the different behaviour. We found a substantial shift towards shorter
401 reads in the manatee blood and brain transcriptome data when compared to the WTC11 cell line,
402 possibly reflecting a poorer library preparation quality for the manatee field sample
403 (Supplemental Fig. S8A). Accordingly, we found that fragmented genes corresponded to
404 transcript models with CDSs lengths above 1.5kb, which were depleted in the manatee dataset
405 (Supplemental Fig. S8B). We concluded that biases in transcript capture by long-read methods
406 may compromise the capacity for providing supporting evidence in evidence-driven annotation
407 methods.

408 Finally, we evaluated if LRB annotation predicted proteins could potentially improve the
409 protein sequences of the public annotation. We selected 2,166 isoforms from the NCBI
410 annotation that specifically matched the splice-junctions of an isoform in the LRB annotation,
411 either fully (FSM) or partially (ISM), and exhibited a truncation of at least 50 base pairs at one
412 end. Among the 1,597 matched LRB isoforms, 1,002 (66.4%) had a BLAST hit in the UniProt
413 curated mammalian database, with differences of fewer than three amino acids at the protein
414 ends. We further aligned the sequences of two of these truncated proteins, their matching LRB
415 proteins, and their orthologs in two closely related manatee species. Figure 4C shows the
416 alignment data for B cell lymphoma 3 protein and the S100P-binding protein isoform X4. In
417 both cases, the protein sequence in the NCBI annotation lacked an N-terminal fragment present

418 in the LRB version and the two manatee-related species. These results confirm the potential of
419 the LRB annotation to improve the current gene annotations.

420 **Discussion**

421 Long-read sequencing technologies, concerted with global efforts to sequence all Earth's
422 organisms, herald a new era of possibilities and requirements for genome annotation.
423 Traditionally, genome annotation has relied on *ab initio* algorithms alone or combined with
424 short-read sequencing and proteomics data to improve gene predictions. However, with the
425 increased availability and throughput of lrRNA-seq, there is a notable shift towards using
426 transcriptome data generated from these platforms to assist genome annotations in non-model
427 species, as exemplified in several recent studies involving the tea plant (Xia et al. 2019) and the
428 ant *Harpegnathos saltator* (Shields et al. 2021).

429 Long-read sequencing offers the advantage of producing full-length transcripts, potentially
430 providing a more complete representation of gene models than is obtained with short reads.
431 However, challenges such as sequencing errors and library artifacts can compromise the
432 reliability of long-read data. Initiatives like the Long-read RNA-seq Genome Annotation
433 Assessment Project (LRGASP) have demonstrated that *de novo* transcript reconstruction using
434 solely long-read data still faces significant hurdles (Pardo-Palacios et al. 2024b).

435 In response to these challenges, we hypothesized that long reads could be optimally utilized in
436 evidence-driven strategies, where experimental evidence is integrated with *ab initio* genome
437 prediction algorithms to enhance their efficacy. Aware of the limitations of long-read
438 transcriptome sequencing, our study aimed to assess how long-read RNA-seq (lrRNA-seq) data
439 can be best used to support genome annotation efforts. Our results indicate that processing
440 lrRNA-seq data into transcript models followed by SQANTI3 curation, rather than using raw
441 reads, and providing this evidence information at the prediction step is the most effective
442 strategy, possibly because the reconstructed transcript models represent more accurate transcript
443 structures than the raw reads. Additionally, we observed that PacBio Sequel II sequencing data

444 yielded better results than ONT v9.4.1 chemistry data or a combination of both. This result is in
445 agreement with conclusions of the LRGASP project (Pardo-Palacios et al. 2024b), that revealed
446 that the longer read distribution and sequence accuracy of the long reads were more important
447 for accurate transcript model prediction than the number of reads. Our results suggest that this is
448 also true when using long-read transcript models for evidence-driven annotation. Moreover, our
449 comparative analysis of results with human and manatee lrRNA-seq data suggests that narrow
450 read length distributions can compromise the accurate support of longer gene models, resulting
451 in gene fragmentation. These results stress the importance of ensuring high-quality full-length
452 transcript models to support genome annotation pipelines.

453 While our study provides important insights regarding the utilization of long-read sequencing
454 technologies in genome annotation, we acknowledge several limitations in our work. First, we
455 provided AUGUSTUS with relatively simple lrRNA-seq datasets (one cell line in human, two
456 tissues for the manatee) to evaluate genome prediction accuracy. Genome annotation efforts
457 often utilize multiple tissue types to capture the broadest range of expressed genes, and our
458 study did not extensively assess this aspect. We acknowledge the need for future studies to
459 explore the impact of tissue diversity on annotation accuracy. However, obtaining diverse tissue
460 samples can be challenging for certain non-model species, as this was the case for the manatee,
461 where access to multiple tissue samples was extremely difficult. Our results suggest that only
462 moderate amounts of long-read data are necessary to enhance *ab initio* prediction accuracy and
463 sensitivity, and that the combination of the *ab initio* with the available lrRNA-seq evidence is an
464 effective way to overcome the problems associated with the limited data.

465 Additionally, in this work, we only evaluated FLAIR and Iso-Seq tools for transcript model
466 construction, as these are widely used tools, and both can operate without reference annotations.
467 However, other methods exist for long-read transcript reconstruction (Bushmanova et al. 2019;
468 Sahlin and Medvedev 2020; Wyman et al. 2020; Tian et al. 2021; Chen et al. 2023; Lienhard et
469 al. 2023; Nip et al. 2023; Prjibelski et al. 2023; Volden et al. 2023). As the LRGASP assessment
470 indicated that *de novo* reconstruction of transcript models from lrRNA-seq data still poses

471 challenges, a follow-up study should evaluate the performance of other algorithms for evidence-
472 driven annotation.

473 The LRGASP consortium demonstrated the impact of the tools and sequencing technologies in
474 the identification of different transcripts. In our case, the differences observed between the LRB
475 manatee annotation and the NCBI annotation are likely influenced by not only the distinct types
476 of RNA-seq data used but also by the differences in the annotation pipelines employed.

477 Finally, while we do not claim that the genome annotation strategy outlined in this study
478 represents the best possible pipeline for lrRNA-seq-based genome annotation, the major
479 takeaway of this work, namely, the utilization of high-quality curated transcript models rather
480 than raw long reads in conjunction with *ab initio* methods as most effective approach for
481 genome annotation offers a valuable perspective for ongoing global efforts to develop quality
482 annotation pipelines using long-read sequencing to annotate the planet's biodiversity.

483 **Methods**

484 **Sample acquisition, nucleic acid extraction, and sequencing**

485 Total RNA was extracted from the brain tissue of an adult West Indian manatee (*Trichechus*
486 *manatus*), sourced from a captive individual at Tierpark Berlin. The extracted RNA displayed a
487 RIN of 7.8. A PacBio IsoSeq library was prepared from this RNA sample and sequenced using
488 a single SMRT cell.

489 **Obtaining transcripts from RNA sequencing data**

490 Assembly of WTC11 short read data

491 Quality control and adapter trimming of the reads was performed using fastp v0.23.2 (Chen et
492 al. 2018). This involved the detection and removal of adaptors from the reads, ensuring data
493 integrity and accuracy. Following pre-processing, reads were aligned to the reference genome
494 using STAR v2.7.10a (Dobin et al. 2013). To enhance splice junction quantification, the
495 alignment process utilized the --twopassMode option (Veeneman et al. 2016).

496 To generate a short-read assembled transcriptome, the reads were aligned using TopHat2 (Kim
497 et al. 2013) with the options `--no-discordant`, `--no-mixed`, and `-r 400`, to accommodate a
498 fragment length of 600 bp with an end length of 100 bp. Subsequently, the mapped reads of
499 each sample were individually assembled using StringTie2.2.1 (Pertea et al. 2015) with default
500 parameters. Finally, a non-redundant set of transcripts was generated by executing StringTie in
501 merge mode, utilizing default parameters. Only stranded transcripts were kept.

502 **Reconstruction of WTC11 and Florida manatee long-read transcriptomes**

503 The Iso-Seq pipeline was used to generate transcript models from PacBio long-read data. This
504 pipeline includes preceding subreads into circular consensus sequencing (CCS) reads (`ccs`
505 `v6.0.0`). Full-length non-concatemer (FLNC) reads were obtained from the CCS reads after
506 removing the sequencing primers (`lima` with `--isoseq` and `--peek-guess` options) and the poly(A)
507 tails (`isoseq refine` with `--require-polya` option). After that, the FLNC reads of the samples were
508 clustered (`isoseq cluster` with `--verbose --use-qvs` options). The high-quality clustered reads
509 were then mapped to the genome using `minimap2 v2.17` (Li 2018). Finally, the mapped reads
510 were collapsed into unique isoforms (`isoseq collapse` with `--do-not-collapse-extra-5exons`
511 option). Only those transcript models supported by at least 2 reads were kept.

512 To obtain the WTC11 ONT transcriptome and the ONT+PacBio (MIX) transcriptome we used
513 FLAIR v1.5.1 (Tang et al. 2020). Long reads were aligned to the human genome using the
514 FLAIR align module with default options. Splice junctions were corrected using FLAIR correct
515 module, incorporating the splice-junctions generated by STAR during short-read alignment.
516 Splice junctions with at least three supporting short reads were kept and used to correct the
517 long-read splice junctions. The corrected long reads were then collapsed into transcript models
518 using the FLAIR collapse module with the options `-s 2 --stringent --check_splice --filternosubset`
519 to reduce the number of redundant isoforms per gene.

520 **Transcriptome filtering**

521 SQANTI3 v4.2 (Pardo-Palacios et al. 2024a) was used to filter reconstructed transcriptomes and
522 eliminate artifacts. We incorporated Illumina short-reads as orthogonal data with the `-c`
523 parameter. Filtering consisted of the removal of monoexons, transcripts with non-sense
524 mediated decay signals, transcripts with non-canonical splice-junctions, and transcripts with
525 evidence of intrapriming. We retain transcripts with coding potential identified by
526 GeneMarkS-T (Tang et al. 2015) and an ORF of at least 300 nucleotides. For the long-read
527 defined transcriptomes, only transcripts with at least two associated long-reads were retained.
528 BUSCO v5.4.7 (Manni et al. 2021) was run in transcriptome mode using the Eutheria_odb10
529 dataset, including 11,366 genes, on these transcriptomes. The BUSCO completeness results
530 were expressed as **C:89.0%[S:85.8%, D:3.2%], F:6.9%, M:4.1%, n:11,366**, where C
531 represents the percentage of complete genes found in the input data, S indicates the percentage
532 of these genes found as single copies, D represents the percentage of duplicated genes, F stands
533 for fragmented genes, M refers to missing genes, and n denotes the total number of genes in the
534 BUSCO dataset used.

535 **Benchmarking of gene prediction methods on WTC11**

536 The software AUGUSTUS v3.1 (Hoff and Stanke 2019) was used for *ab initio*, evidence-based,
537 and evidence-driven gene prediction strategies. Evidence was obtained from the long-read or
538 short-read data at three levels of processing: raw-reads without preprocessing (raw level),
539 reconstructed or assembled transcript models (transcript level), and collapsed transcripts into
540 unique *loci* (gene level).

541 Obtaining training sets from long-read transcripts

542 To obtain a reliable and non-redundant set of long-read transcripts to use as training sets for
543 AUGUSTUS, we further filtered the 3 transcriptomes based on the SQANTI3 output. Only
544 transcript models with all their splice-junctions supported by short-reads were included in the
545 training set. For the ONT+Pacbio transcriptome, we only kept the 25% most expressed

546 transcripts. Moreover, protein sequences predicted by GeneMarkS-T in the filtered WTC11
547 transcript models were searched in the manually reviewed mammalian proteins available at
548 UniProt (<https://www.uniprot.org>) using BLAST+ v2.12.0 (Camacho et al. 2009) with an E-
549 value cutoff of 1×10^{-50} . Only proteins with at least one BLAST hit and a query coverage over
550 85% were kept for the training set.

551 To minimize the number of isoforms per gene, TAMA Collapse (Kuo et al. 2020) was applied
552 to the three filtered WTC11 transcriptomes, utilizing options `-x no_cap -m 100 -z 100`.
553 Subsequently, a second clustering, based on the sequence of the predicted proteins in the
554 transcripts, was conducted using CD-HIT v4.8.1 (Fu et al. 2012). Sequences with an identity
555 higher than 80% were clustered with the option `-c 0.8`. The gene coding the longest protein was
556 kept as the representative of the cluster. Finally, BUSCO genes in Eutheria_odb10 dataset were
557 identified in the final training sets after clustering at the protein level.

558 AUGUSTUS HMM training and ab initio gene prediction with WTC11 data

559 The AUGUSTUS Hidden Markov Model (HMM) was trained using non-redundant transcripts
560 sourced from the Iso-Seq WTC11 PacBio transcriptome, the FLAIR WTC11 ONT
561 transcriptome, or a combination of FLAIR PacBio and ONT WTC11 transcriptomes (MIX).
562 Additionally, BUSCO genes in Eutheria_odb10 dataset identified in the human genome were
563 utilized for HMM training.

564 Various parameters were explored during training, including the number of genes in the training
565 set and the length of the flanking region. Flanking region lengths of 1000 nts, 2500 nts, 5000
566 nts, 7500 nts, and 10000 nts were tested. Training gene sets of sizes ranging from 200 to 5000
567 were generated from the three non-redundant training sets derived from the WTC11 long-reads
568 transcriptomes.

569 The AUGUSTUS script `gff2gbSmallDNA.pl` was employed to convert files from GTF to
570 GenBank format, incorporating the appropriate flanking region. Evaluation of HMMs was

571 conducted using Monte-Carlo cross-validation (Xu and Liang 2001). Each training gene set was
572 randomly generated three times using different seeds, resulting in a total of 117 training sets.

573 Gene prediction with experimental evidence on the human genome

574 AUGUSTUS was run using the --sofmasking=on option and different sources of experimental
575 evidence to predict genes on Chromosome 19 of the human genome. The experimental evidence
576 given to AUGUSTUS was: Illumina short-reads, ONT and PacBio long-reads individually and
577 in combination, the Iso-Seq3 WTC11 PacBio transcriptome, the FLAIR WTC11 ONT
578 transcriptome, the FLAIR PacBio and ONT WTC11 transcriptome, the Illumina transcriptome
579 and the 3 sets of non-redundant collapsed transcripts generated from each of the three long-
580 reads transcriptomes.

581 For the four transcriptomes, only multiexon transcripts identified as protein-coding and with no
582 features of bad quality, i.e. signals of non-sense mediated decay, signals of intrapriming, non-
583 canonical splice-junctions and a coding sequence (CDS) shorter than 300 nt, were given to
584 AUGUSTUS.

585 The default AUGUSTUS configuration file for long-reads evidence was used as input for the
586 raw long-read data. For the Illumina-assembled transcriptome and the long-reads
587 transcriptomes, only the CDSs identified by GeneMarkS-T were converted to AUGUSTUS
588 hints. For the Illumina short-reads aligned with STAR, the default AUGUSTUS configuration
589 file for short-reads evidence was used. Ultimately, these sets of predicted genes were compared
590 to the reference human annotation of Chromosome 19 using Cuffcompare v2.2.1 (Trapnell et al.
591 2010).

592 WTC11 genome annotation evaluation

593 The performance of gene prediction strategies was evaluated by comparison to the annotation of
594 the human Chromosome 19 using Cuffcompare v2.2.1 with the -e 0 -d 0 options. The reference
595 annotation was constructed from the GENCODE annotation selecting the coding region of the
596 transcript with the longest open reading frame (ORF) for each gene on Chromosome 19.

597 A predicted gene was considered a true positive (TP) when all exonic and genic nucleotides
598 matched the corresponding features in the reference. Nucleotides, exons, or genes not matching
599 any feature in the reference were considered false positives (FP), while features in the reference
600 not perfectly matched by gene predictions were categorized as false negatives (FN).

601 Cuffcompare provided the sensitivity (Sn) and precision (Pr) values at the nucleotide, exon, and
602 gene levels. The F1-score was calculated as the harmonic mean of these two values.

$$603 \quad Sn = \frac{TP}{TP+FN}; Pr = \frac{TP}{TP+FP}; F1\text{-score} = 2 \times \frac{Sn \times Pr}{Sn+Pr}$$

604 Assessment of sequencing depth in evidence-based and evidence-driven annotation

605 To evaluate the effect of sequencing depth on genome annotation we performed evidence-driven
606 and evidence-based annotation using an increasing proportion of the total WTC11 PacBio
607 FLNC reads. Subsampling of the FLNC reads from the WTC11 cell line was conducted using
608 SAMtools v1.16 (Danecek et al. 2021). The fraction of sampled reads was 0.1, 0.15, 0.2, 0.25,
609 0.5, 0.75, and the full dataset. Following subsampling, the FLNC reads were processed using
610 the Iso-Seq v4.0.0 pipeline. The same strategy described in the previous sections was used to
611 filter the transcriptomes. CDSs identified using GeneMarkS-T were included in AUGUSTUS
612 v3.1 as experimental evidence. Finally, the seven sets of predicted genes, alongside the
613 experimental data, were compared to the reference human annotation of the Chr19 using
614 Cuffcompare v2.2.1.

615 **Annotation of the Florida manatee genome using long-reads**

616 AUGUSTUS *ab initio* predictions on the manatee genome

617 The AUGUSTUS HMM was trained using BUSCO genes identified within the LRGASP
618 Florida manatee genome (Pardo-Palacios et al. 2024b). The training gene set consisted of 5000
619 genes, with a flanking region length of 1000 nucleotides. Files were converted from GTF to
620 GenBank format with the appropriate flanking region using the AUGUSTUS script
621 `gff2gbSmallIDNA.pl`.

622 Subsequently, the trained model was utilized to predict genes on the manatee genome. For the
623 *ab initio* predictions on the manatee genome, BUSCO was employed in protein mode, utilizing
624 the Eutheria_odb10 dataset.

625 Evidence-driven annotation of the manatee genome

626 AUGUSTUS v3.1 was executed with the --softmasking=on option, utilizing the CDSs identified
627 by GeneMarkS-T within the filtered PacBio transcripts as extrinsic evidence. To predict several
628 isoforms per gene, we included the --alternatives-from-evidence=on option.

629 Subsequently, the protein sequences of these predictions were obtained with the AUGUSTUS
630 getAnnoFasta.pl script. For assessing the BUSCO completeness, BUSCO was run in protein
631 mode, employing the Eutheria_odb10 and the protein sequences of the predicted genes obtained
632 using the created transcriptome.

633 Finally, BUSCO genes identified in the genome were concatenated with the long-read evidence
634 and provided to AUGUSTUS as hints for evidence-driven annotation. Protein sequences of the
635 predictions were obtained to assess the BUSCO completeness of the annotation.

636 Comparison of the Florida manatee long-reads based genome annotation with the NCBI

637 *Trichechus manatus latirostris* Annotation Release 102

638 To compare the NCBI short-read-based genome annotation of the Florida manatee with our
639 long-read-based annotation, we first mapped the CDS sequences from NCBI to the LRGASP
640 Florida manatee genome using minimap2 with the options -ax splice -uf -MD. After obtaining
641 the primary sorted alignment using SAMtools v1.16, we generated the GFF file using
642 spliced_bam2gff v1.2. Both annotations were compared using Cuffcompare v2.2.1 with options
643 -e 0, -d 0, and -G. To validate the unique genes of each annotation, we extracted their protein
644 sequences and used BLAST v2.13 to search these sequences on a custom-made database with
645 representatives of all the mammalian proteins, using an E-value cutoff of 1-e3.

646 To study the structural classification of isoforms in each annotation, SQANTI3 v5.2 was run
647 using the liftOver of the NCBI annotation (only the protein-coding sequences) as input and the
648 complete LRB annotation as the reference annotation; and vice versa. We identified those
649 isoforms classified by SQANTI3 as fusion and further characterized those cases in which NCBI
650 isoforms overlapped multiple supported genes of the LRB annotation, i.e. the genes determined
651 using both the long-read experimental data or BUSCO genes as support for AUGUSTUS during
652 the prediction step. After filtering the supported predicted genes from the LRB annotation, we
653 divided these genes into two groups based on whether they were part of a fusion isoform in the
654 NCBI annotation or not. Then, the number of exons and length of these genes present in the
655 LRB annotation were studied.

656 To determine if these supported LRB genes composing the NCBI fusion isoforms were
657 fragments of the same protein, BLAST v2.13 was run using their protein sequences and
658 utilizing a custom-made database with all the human proteins. The options used for running
659 BLAST were an E-value cutoff of 1-e3 and -num_alignments 10. The BLAST custom-made
660 database was created with all the human-curated proteins available at UniProt. After that, a
661 custom-made Python script was developed to test if the BLAST hits recovered were the same
662 for the supported LRB genes overlapped by the same NCBI fusion isoform. To address whether
663 the fragmentation was caused by the experimental evidence provided for the predictions or
664 during the gene prediction step, SQANTI3 was run using the experimental data as input and the
665 supported LRB genes overlapped by NCBI fusion isoforms as reference. In case of
666 fragmentation during the gene prediction step, the experimental evidence would be classified as
667 fusion, overlapping multiple genes.

668 Comparison with other species

669 To demonstrate the potential of our LRB annotation to improve the NCBI *Trichechus manatus*
670 *latirostris* Annotation Release 102, we selected transcripts with matching splice junctions in
671 both annotations but with truncated CDS in the NCBI annotation. We extracted the
672 corresponding protein sequences and searched them using BLASTP v2.15 on the NCBI

673 Reference Proteins (refseq_protein) database. The hits of two different species were selected
674 based on these filters: lowest E-value and both highest query cover and percent of identity. The
675 four sequences were aligned using Clustal Omega (Sievers and Higgins 2021) from the EMBL-
676 EBI web page (Madeira et al. 2022).

677 **Data sets**

678 Transcriptomics data used in this study was obtained from the LRGASP project and
679 downloaded from the ENCODE database (<https://www.encodeproject.org/>), with the following
680 accession: ENCSR673UKZ for the human WTC11 Illumina RNA-seq (3 samples, 143,171,620
681 paired-end reads), ENCSR513JKI for the manatee Illumina RNA-seq (9 samples, 513,272,173
682 paired-end reads), ENCSR539ZXJ for WTC11 Nanopore raw reads (3 samples, 30,664,338
683 reads), ENCSR507JOF for WTC11 PacBio raw subreads (3 samples, 6,943,271 FLNC reads),
684 and ENCSR272BQI (3 Sequel II samples) and ENCSR583MLP (1 Sequel I sample) for
685 manatee peripheral blood mononuclear cells (5,434,382 FLNC reads). Additionally, this study
686 generated a total of 2,709,782 FLNC reads from the brain sample of a single manatee
687 individual.

688 The human reference genome used for lrRNA-seq and RNA-seq data processing and structural
689 annotation was downloaded from the GENCODE database (<https://www.gencodegenes.org/>)
690 (GRCh38.p13 release 38 reference primary assembly). The corresponding reference annotation
691 was downloaded from GENCODE and filtered to contain only protein-coding genes. The
692 LRGASP *Trichechus manatus latirostris* genome, assembled using Nanopore and Illumina
693 reads, was downloaded from GenBank (<https://www.ncbi.nlm.nih.gov/>) (GCA_030013775).
694 The short-read based TriManLat1.0 genome assembly (GCF_000243295.1) and the NCBI
695 *Trichechus manatus latirostris* Annotation Release 102 were downloaded from RefSeq
696 (<https://www.ncbi.nlm.nih.gov/refseq/>).

697 **Data access**

698 The transcriptomics data generated in this study have been submitted to the ENA database
699 (<https://www.ebi.ac.uk/ena/browser/>) under accession number PRJEB76292. The custom
700 scripts, final Florida manatee annotation, and AUGUSTUS hints generated in this study can be
701 found on GitHub (https://github.com/alexpan00/lr_evidence_driven) and as Supplemental Code.

702 **Competing interests**

703 A.C. has received in-kind funding from Pacific Biosciences for library preparation and
704 sequencing. A.C. collaborates with Oxford Nanopore in the Marie Skłodowska-Curie Actions
705 Doctoral Network project LongTREC.

706 **Acknowledgments**

707 This research was supported by the Ministry of Science, Innovation and Universities of Spain
708 (FPU21/01597 and PID2020-119537RB-I00), the National Institutes of Health
709 (1R21HG011280-01) and the Comunitat Valenciana Grant ACIF/2018/290

710 We thank Dr. Walsh and Homosassa Spring State Park for supporting this research by providing
711 manatee blood samples. Any use of trade, firm, or product names is for descriptive purposes
712 only and does not imply endorsement by the United States Government. We thank Sylke
713 Winkler and Gene Myers for generating the manatee Iso-Seq data. We thank Margaret E.
714 Hunter for her valuable contributions to the revision of this manuscript. The computations were
715 performed on the HPC cluster Garnatxa at the Institute for Integrative Systems Biology
716 (I2SysBio), I2SysBio is a joint research institute of the University of Valencia (UV) and the
717 Spanish National Research Council (CSIC).

718 *Author contribution:* A.C conceptualized and supervised the study. T.B. and C.J.M. provided
719 the Florida manatee brain samples. A.P. and C.A.G analysed the data. F.J.P.P contributed in the
720 interpretation of the results. A.P. and C.A.G wrote the initial manuscript with the contribution of
721 all of the authors.

722 **References**

- 723 Blaxter ML. 2022. Sequence locally, think globally: The Darwin Tree of Life Project. *Proc Natl*
724 *Acad Sci U S A* **119**: e2115642118. doi:10.1073/pnas.2115642118
- 725 Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019. maSPAdes: a de novo
726 transcriptome assembler and its application to RNA-Seq data. *Gigascience* **8**: 1–13.
727 doi:10.1093/gigascience/giz100
- 728 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
729 BLAST+: Architecture and applications. *BMC Bioinformatics* **10**: 1–9. doi:10.1186/1471-
730 2105-10-421
- 731 Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome Annotation and Curation Using
732 MAKER and MAKER-P. *Curr Protoc Bioinformatics* **48**: 4.11.1-4.11.39.
733 doi:10.1002/0471250953.bi0411s48
- 734 Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor.
735 *Bioinformatics* **34**: i884–i890. doi:10.1093/bioinformatics/bty560
- 736 Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, Love MI, Göke J. 2023. Context-aware
737 transcript quantification from long-read RNA-seq data with Bambu. *Nature Methods*
738 *2023 20:8* **20**: 1187–1195. doi:10.1038/s41592-023-01908-w
- 739 Cocquet J, Chong A, Zhang G, Veitia RA. 2006. Reverse transcriptase template switching and
740 false alternative transcripts. *Genomics* **88**: 127–131. doi:10.1016/j.ygeno.2005.12.013
- 741 Cook DE, Valle-Inclan JE, Pajoro A, Rovenich H, Thomma BPHJ, Faino L. 2019. Long-Read
742 Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA
743 Sequencing. *Plant Physiol* **179**: 38–54. doi:10.1104/pp.18.00848

- 744 Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T,
745 McCarthy SA, Davies RM. 2021. Twelve years of SAMtools and BCFtools. *Gigascience*
746 **10**: 1–4. doi:10.1093/gigascience/giab008.
- 747 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M,
748 Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–
749 21. doi:10.1093/bioinformatics/bts635
- 750 Dragan MA, Moghul I, Priyam A, Bustos C, Wurm Y. 2016. GeneValidator: identify problems
751 with protein-coding gene predictions. *Bioinformatics* **32**: 1559.
752 doi:10.1093/bioinformatics/btw015
- 753 Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation
754 sequencing data. *Bioinformatics* **28**: 3150–3152. doi:10.1093/bioinformatics/bts565
- 755 Guigó R, Agarwal P, Abril JF, Burset M, Fickett JW. 2000. An Assessment of Gene Prediction
756 Accuracy in Large DNA Sequences. *Genome Res* **10**: 1631. doi:10.1101/gr.122800
- 757 Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-Genome Annotation with
758 BRAKER. *Methods Mol Biol* **1962**: 65. doi:10.1007/978-1-4939-9173-0_5
- 759 Hoff KJ, Stanke M. 2019. Predicting Genes in Single Genomes with AUGUSTUS. *Curr Protoc*
760 *Bioinformatics* **65**: e57. doi:10.1002/cpbi.57
- 761 Huang Y, Chen SY, Deng F. 2016. Well-characterized sequence features of eukaryote genomes
762 and implications for ab initio gene prediction. *Comput Struct Biotechnol J* **14**: 298.
763 doi:10.1016/j.csbj.2016.07.002
- 764 Jung H, Ventura T, Sook Chung J, Kim WJ, Nam BH, Kong HJ, Kim YO, Jeon MS, Eyun S II.
765 2020. Twelve quick steps for genome assembly and annotation in the classroom. *PLoS*
766 *Comput Biol* **16**: e1008325. doi:10.1371/journal.pcbi.1008325

- 767 Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: Accurate
768 alignment of transcriptomes in the presence of insertions, deletions and gene fusions.
769 *Genome Biol* **14**: 1–13. doi:10.1186/gb-2013-14-4-r36
- 770 Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**: 1–9. doi:10.1186/1471-
771 2105-5-59
- 772 Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, Archibald AL, Burt DW. 2020. Illuminating
773 the dark side of the human transcriptome with long read transcript sequencing. *BMC*
774 *Genomics* **21**: 1–22. doi:10.1186/s12864-020-07123-7
- 775 Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R,
776 Edwards S V., Forest F, Gilbert MTP, et al. 2018. Earth BioGenome Project: Sequencing
777 life for the future of life. *Proc Natl Acad Sci U S A* **115**: 4325–4333.
778 doi:10.1073/pnas.1720115115
- 779 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–
780 3100. doi:10.1093/bioinformatics/bty191
- 781 Liang C, Mao L, Ware D, Stein L. 2009. Evidence-based gene predictions in plant genomes.
782 *Genome Res* **19**: 1912. doi:10.1101/gr.088997.108
- 783 Lienhard M, Van Den Beucken T, Timmermann B, Hochradel M, Börno S, Caiment F, Vingron
784 M, Herwig R. 2023. IsoTools: a flexible workflow for long-read transcriptome
785 sequencing analysis. *Bioinformatics* **39**. doi:/10.1093/bioinformatics/btad364
- 786 Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, Madhusoodanan N, Kolesnikov
787 A, Lopez R. 2022. Search and sequence analysis tools services from EMBL-EBI in 2022.
788 *Nucleic Acids Res* **50**: gkac240–gkac240. doi:10.1093/nar/gkac240
- 789 Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel
790 and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for

- 791 Scoring of Eukaryotic, Prokaryotic, and Viral Genomes ed. J. Kelley. *Mol Biol Evol* **38**:
792 4647–4654. doi:10.1093/molbev/msab199
- 793 Nip KM, Hafezqorani S, Gagalova KK, Chiu R, Yang C, Warren RL, Birol I. 2023. Reference-
794 free assembly of long-read transcriptome sequencing data with RNA-Bloom2. *Nature*
795 *Communications* 2023 14:1 **14**: 1–14. doi:10.1038/s41467-023-38553-y
- 796 Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín R,
797 Estevan-Morió E, Liu T, Nanni A, McIntyre L, et al. 2024a. SQANTI3: curation of long-
798 read transcriptomes for accurate identification of known and novel isoforms. *Nature*
799 *Methods* 2024 21:5 **21**: 793–797. doi:10.1038/s41592-024-02229-2
- 800 Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, Loveland JE,
801 De María M, Adams MS, Balderrama-Gutierrez G, et al. 2024b. Systematic assessment of
802 long-read RNA-seq methods for transcript identification and quantification. *Nat Methods*
803 **21**: 1349–1363. doi:10.1038/s41592-024-02298-3.
- 804 Peng S, Dahlgren AR, Hales EN, Barber AM, Kalbfleisch T, Petersen JL, Bellone RR,
805 Mackowski M, Cappelli K, Capomaccio S, et al. 2022. Long-read RNA Sequencing
806 Improves the Annotation of the Equine Transcriptome. *bioRxiv*.
807 doi:10.1101/2022.06.07.495038
- 808 Perteu M, Perteu GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie
809 enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature*
810 *Biotechnology* 2015 33:3 **33**: 290–295. doi:10.1038/nbt.3122
- 811 Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, Tilgner HU.
812 2023. Accurate isoform discovery with IsoQuant using long reads. *Nature Biotechnology*
813 2023 41:7 **41**: 915–918. doi:10.1038/s41587-022-01565-y
- 814 Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W,
815 Fungtammasan A, Kim J, et al. 2021. Towards complete and error-free genome

- 816 assemblies of all vertebrate species. *Nature* 2021 592:7856 **592**: 737–746.
817 doi:10.1038/s41586-021-03451-0
- 818 Sahlin K, Medvedev P. 2020. De Novo Clustering of Long-Read Transcriptome Data Using a
819 Greedy, Quality Value-Based Algorithm. *Journal of Computational Biology* **27**: 472–484.
820 doi: 10.1089/cmb.2019.0299
- 821 Salzberg SL. 2019. Next-generation genome annotation: We still struggle to get it right.
822 *Genome Biol* **20**: 1–3. doi: 10.1186/s13059-019-1715-2
- 823 Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson JD. 2020. A benchmark study of
824 ab initio gene prediction methods in diverse eukaryotic organisms. *BMC Genomics* **21**: 1–
825 20. doi:10.1186/s12864-020-6707-9
- 826 Shields EJ, Sorida M, Sheng L, Sieriebriennikov B, Ding L, Bonasio R. 2021. Genome
827 annotation with long RNA reads reveals new patterns of gene expression and improves
828 single-cell analyses in an ant brain. *BMC Biol* **19**: 1–19. doi:10.1186/s12915-021-01188-
829 w
- 830 Sievers F, Higgins DG. 2021. The Clustal Omega Multiple Alignment Package. *Methods Mol*
831 *Biol* **2231**: 3–16. doi:10.1007/978-1-0716-1036-7_1
- 832 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO:
833 assessing genome assembly and annotation completeness with single-copy orthologs.
834 *Bioinformatics* **31**: 3210–3212. doi:10.1093/bioinformatics/btv351
- 835 Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped
836 cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**: 637–644.
837 doi:10.1093/bioinformatics/btn013
- 838 Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a
839 generalized hidden Markov model that uses hints from external sources. *BMC*
840 *Bioinformatics* **7**: 1–11. doi:10.1186/1471-2105-7-62.

- 841 Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron
842 submodel. *Bioinformatics* **19 Suppl 2**. doi:10.1093/bioinformatics/btg1080
- 843 Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. 2020.
844 Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic
845 leukemia reveals downregulation of retained introns. *Nature Communications* 2020 **11:1**
846 **11**: 1–12. doi:10.1038/s41467-020-15171-6
- 847 Tang S, Lomsadze A, Borodovsky M. 2015. Identification of protein coding regions in RNA
848 transcripts. *Nucleic Acids Res* **43**: e78–e78. doi:10.1093/nar/gkv227
- 849 Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, Kariyawasam H, Du
850 MRM, Schuster J, Wang C, et al. 2021. Comprehensive characterization of single-cell
851 full-length isoforms in human and mouse with long-read sequencing. *Genome Biol* **22**: 1–
852 24. doi:10.1186/s13059-021-02525-6.
- 853 Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold
854 BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals
855 unannotated transcripts and isoform switching during cell differentiation. *Nature*
856 *Biotechnology* 2010 **28:5** **28**: 511–515. doi:10.1038/nbt.1621
- 857 Veeneman BA, Shukla S, Dhanasekaran SM, Chinnaiyan AM, Nesvizhskii AI. 2016. Two-pass
858 alignment improves novel splice junction quantification. *Bioinformatics* **32**: 43–49.
859 doi:10.1093/bioinformatics/btv642
- 860 Volden R, Schimke KD, Byrne A, Dubocanin D, Adams M, Vollmers C. 2023. Identifying and
861 quantifying isoforms from accurate full-length transcriptome sequencing reads with
862 Mandalorion. *Genome Biol* **24**: 1–15. doi:10.1186/s13059-023-02999-6
- 863 Wang J, Li ST, Zhang Y, Zheng HK, Xu Z, Ye J, Yu J, Wong GKS. 2003. Vertebrate gene
864 predictions and the problem of large genes. *Nat Rev Genet* **4**: 741–749.
865 doi:10.1038/nrg1160

- 866 Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahmanian S, Forner S, Matheos D,
867 Zeng W, Williams B, Trout D, et al. 2020. A technology-agnostic long-read analysis
868 pipeline for transcriptome discovery and quantification. *bioRxiv*. doi:10.1101/672931
- 869 Xia E, Li F, Tong W, Yang H, Wang S, Zhao J, Liu C, Gao L, Tai Y, She G, et al. 2019. The tea
870 plant reference genome and improved gene annotation using long-read and paired-end
871 sequencing data. *Scientific Data 2019 6:1* **6**: 1–9. doi:10.1038/s41597-019-0127-1
- 872 Xu QS, Liang YZ. 2001. Monte Carlo cross validation. *Chemometrics and Intelligent*
873 *Laboratory Systems* **56**: 1–11. doi:10.1016/S0169-7439(00)00122-2
- 874 Yandell M, Ence D. 2012. A beginner’s guide to eukaryotic genome annotation. *Nature Reviews*
875 *Genetics 2012 13:5* **13**: 329–342. doi:10.1038/nrg3174
- 876 Zhang R, Kuo R, Coulter M, Calixto CPG, Entizne JC, Guo W, Marquez Y, Milne L, Riegler S,
877 Matsui A, et al. 2022. A high-resolution single-molecule sequencing-based Arabidopsis
878 transcriptome using novel methods of Iso-seq analysis. *Genome Biol* **23**: 1–37.
879 doi:10.1186/s13059-022-02711-0
- 880
- 881
- 882
- 883
- 884
- 885
- 886
- 887
- 888

889

890

891

892

893

894

895

896

897

898

899 **Figure 1.** Overview of the evidence-driven, *ab initio* and evidence-based annotation using long-read
900 technologies. We compared three genome annotation strategies, evidence-based annotation using lrrRNA-
901 seq data (a), *ab initio* gene prediction using AUGUSTUS (b) and evidence-driven annotation combining
902 both (c). We tested Nanopore (1) and PacBio (2) data independently and in combination (3) with
903 processing levels ranging from raw reads (i) to transcripts (ii) and collapsing to the gene level (iii). For
904 the best strategy defined in our benchmark, we evaluated the effect of the sequencing depth and compared
905 to evidence-driven annotation using Illumina short-reads data. Finally, we applied our strategy to a non-
906 model species, the Florida manatee, applying BUSCO and SQANTI3 to evaluate the completeness of our
907 annotation and comparing to the available Florida manatee annotation.

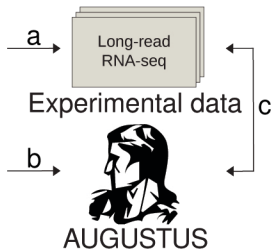
908 **Figure 2.** Assessment of the incorporation of long-read data to evidence-driven annotation. (A)
909 Distribution of SQANTI3 categories at the read unique splice-junctions combination, transcript, and
910 collapsed transcript level for PacBio data processed using Iso-Seq3 and ONT and ONT+PacBio data
911 processed using FLAIR. Full-Splice-Match (FSM), Incomplete-Splice-Match (ISM), Novel-in-catalog
912 (NIC), Novel-Not-in-Catalog (NNC). (B) Redundancy of the generated transcriptome and the final
913 collapsed transcripts set based on the proportion of duplicated BUSCO genes. (C) Evaluation of gene

914 predictions by the three models trained with long-read data and the model trained using BUSCO genes.
915 (D) Selection of the different long-reads-based extrinsic evidence sources used for the evidence-driven
916 gene predictions. Reads of PacBio, ONT, and a MIX of both (Read); transcript models of PacBio, ONT,
917 and a MIX of both (Transcript) and collapsed transcripts identified in those transcriptomes (Collapsed
918 Transcript) were used.

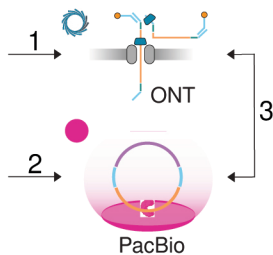
919 **Figure 3.** Performance analysis of gene prediction as a function of the number of reads. Sensitivity,
920 precision (A) and number of missed *loci* (B) were obtained with different sample sizes of WTC11 cell
921 line PacBio FLNC reads using evidence-based and evidence-driven approaches. Performance of the
922 different genome annotation approaches with Illumina short-reads and PacBio long-reads technologies at
923 the gene level. (C) *Ab initio* predictions. (D), evidence-based models using PacBio and Illumina
924 assembled transcriptomes. (E), evidence-driven approach with PacBio, Illumina reads or Illumina
925 assembled transcriptomes as the source of evidence for the prediction step.

926 **Figure 4.** LRB annotation of the Florida manatee. (A) Assessing the BUSCO completeness of the Florida
927 manatee's genome annotation with the different approaches. (B) Number of isoforms per gene of NCBI
928 annotation (left) and the LRB annotation (right). (C) Multiple sequence alignments of proteins of two
929 manatee-related species, the sequence identified in the new annotation, and the manatee sequence available at
930 the NCBI. At the top, B cell lymphoma 3 protein is shown. At the bottom, S100P-binding protein isoform X4
931 is shown.

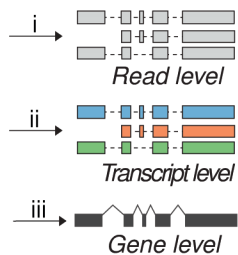
Annotation method



Source of long-reads



Preprocessing level

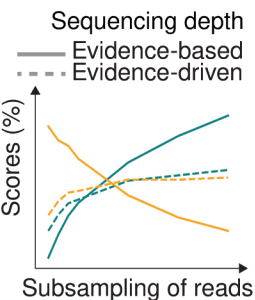


BENCHMARKING STRATEGY USING HUMAN WTC11 CELL LINE

BEST STRATEGY

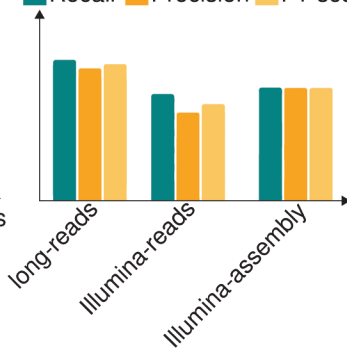
1

Evaluation



Comparison with Illumina

■ Recall ■ Precision ■ F1-score



2

Application

Florida Manatee

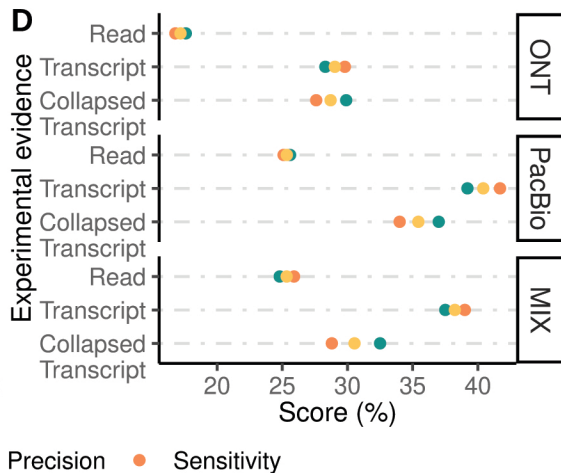
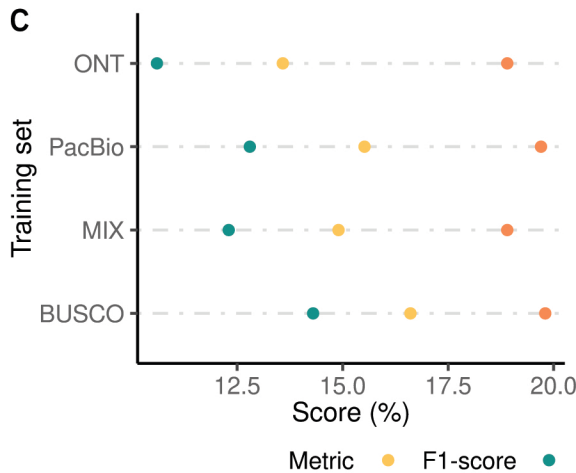
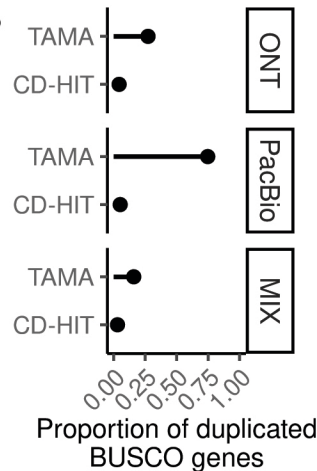
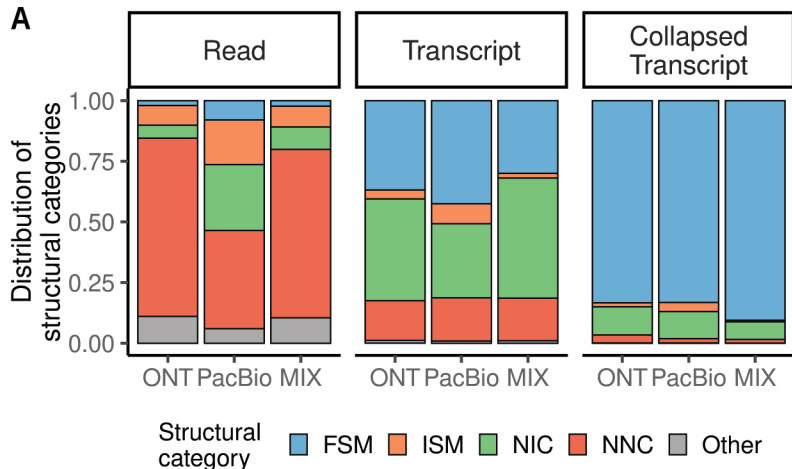


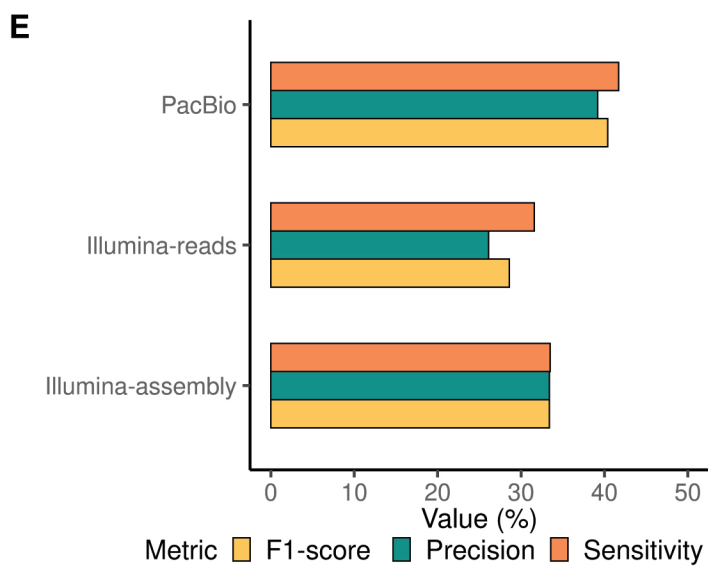
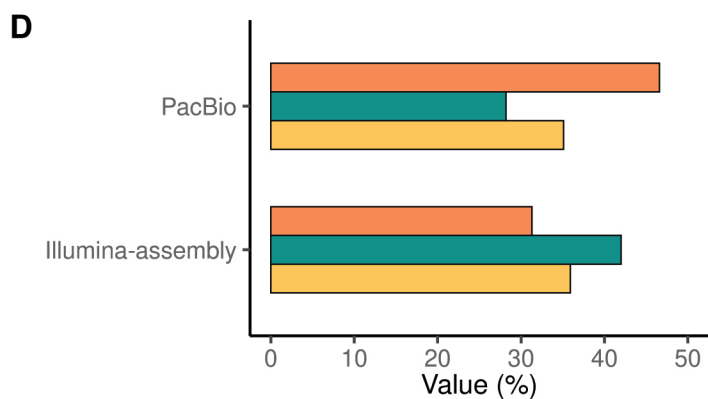
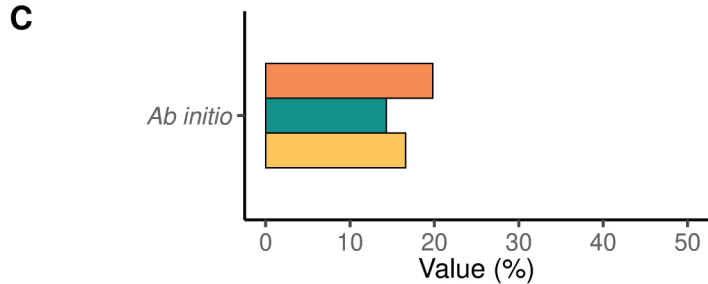
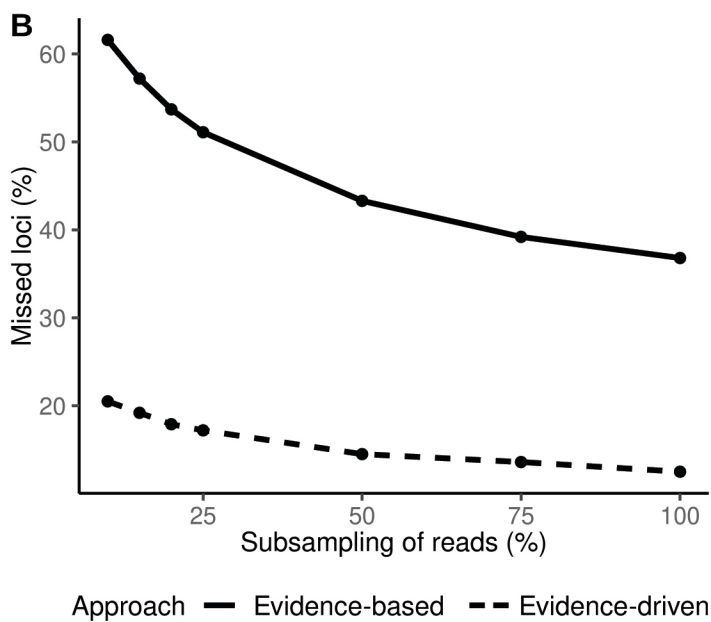
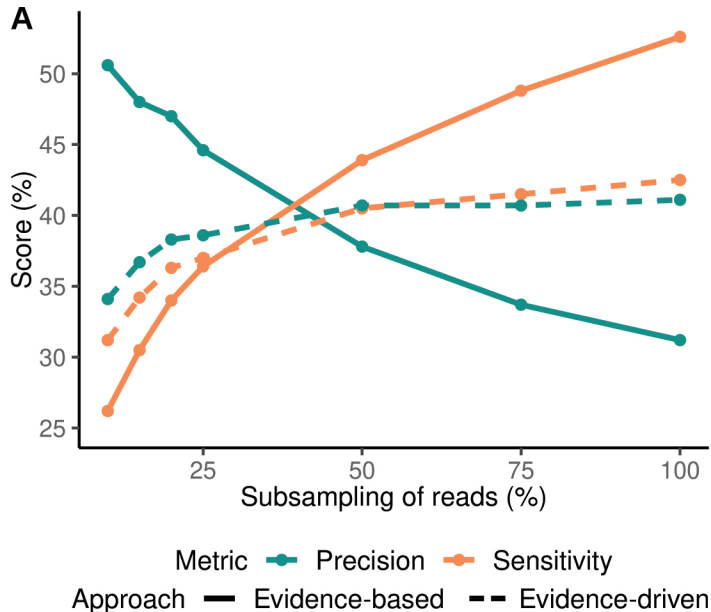
Evidence-driven Annotation

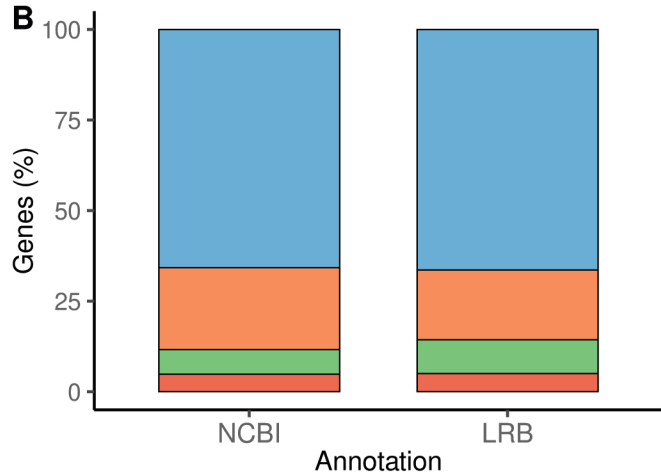
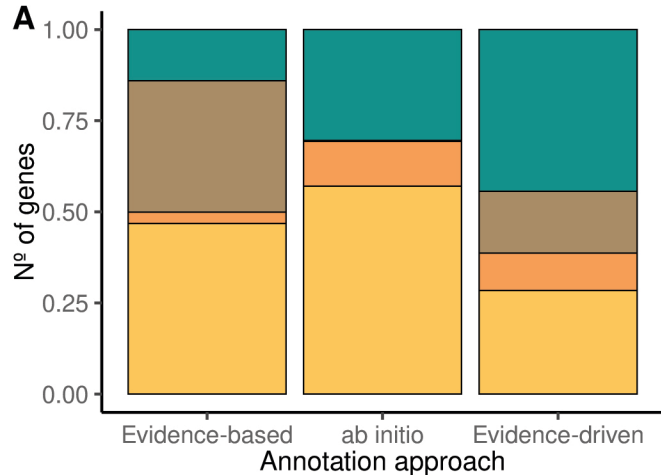
➤ **BUSCO**
Completeness

➤ **SOANTI**₃

Reference:
• TriManLat1.0







C

Oryteropus afer afer
Elephants maximus indicus
 TriManLat1.0
 LRB evidence-driven

XP_004388884.1: B-cell lymphoma 3 protein



XP_023583038.1: S100P-binding protein isoform X4A

Loxodota africana
Elephants maximus indicus
 TriManLat1.0
 LRB evidence-driven

