Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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 *Genome Research* Open Access option.

Creative Commons License This manuscript is Open Access.This article, published in *Genome Research*, 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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

Alejandro Paniagua[1,2*], Cristina Agustín-García[1,*], Francisco J. Pardo-Palacios[1], Thomas Brown[3,4], Maite De Maria[5], Nancy D. Denslow[5], Camila J. Mazzoni[3,4], Ana Conesa[1]

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

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

[3]Department of Evolutionary Genetics, Leibniz Institute for Zoo and Wildlife Research Berlin, Germany

[4] Berlin Center for Genomics in Biodiversity Research, Germany

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

contact: ana.conesa@csic.es

* equally contributing

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

Abstract

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

Introduction

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

Results

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

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

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

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

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

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

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

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

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

Evidence-driven annotation of the Florida manatee genome

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

To further complete the annotation, we incorporated the BUSCO genes identified in the manatee genome with the PacBio transcriptome and used these two sources of evidence in combination during the gene prediction step. This strategy resulted in a BUSCO completeness of 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 and 39,977 transcripts. This transcriptome was used for downstream analysis.

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

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

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

Methods

Sample acquisition, nucleic acid extraction, and sequencing

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

Obtaining transcripts from RNA sequencing data

Assembly of WTC11 short read data

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

Reconstruction of WTC11 and Florida manatee long-read transcriptomes

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

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

Transcriptome filtering

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

Benchmarking of gene prediction methods on WTC11

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

Obtaining training sets from long-read transcripts

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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.

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

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

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

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

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

Annotation of the Florida manatee genome using long-reads

AUGUSTUS *ab initio* predictions on the manatee genome

The AUGUSTUS HMM was trained using BUSCO genes identified within the LRGASP Florida manatee genome (Pardo-Palacios et al. 2024b). The training gene set consisted of 5000 genes, with a flanking region length of 1000 nucleotides. Files were converted from GTF to GenBank format with the appropriate flanking region using the AUGUSTUS script gff2gbSmallDNA.pl.

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

Evidence-driven annotation of the manatee genome

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

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

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

Comparison of the Florida manatee long-reads based genome annotation with the NCBI *Trichechus manatus latirostris* Annotation Release 102

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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**

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

Competing interests

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

Acknowledgments

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

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

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

References

Blaxter ML. 2022. Sequence locally, think globally: The Darwin Tree of Life Project. *Proc Natl Acad Sci U S A* **119**: e2115642118. doi:10.1073/pnas.2115642118

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience* **8**: 1–13. doi:10.1093/gigascience/giz100

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* **10**: 1–9. doi:10.1186/1471-2105-10-421

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome Annotation and Curation Using MAKER and MAKER-P. *Curr Protoc Bioinformatics* **48**: 4.11.1-4.11.39. doi:10.1002/0471250953.bi0411s48

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: i884–i890. doi:10.1093/bioinformatics/bty560

Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, Love MI, Göke J. 2023. Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nature Methods 2023 20:8* **20**: 1187–1195. doi:10.1038/s41592-023-01908-w

Cocquet J, Chong A, Zhang G, Veitia RA. 2006. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88**: 127–131. doi:10.1016/j.ygeno.2005.12.013

Cook DE, Valle-Inclan JE, Pajoro A, Rovenich H, Thomma BPHJ, Faino L. 2019. Long-Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing. *Plant Physiol* **179**: 38–54. doi:10.1104/pp.18.00848

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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 Il.
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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

Scoring of Eukaryotic, Prokaryotic, and Viral Genomes ed. J. Kelley. *Mol Biol Evol* **38**: 4647–4654. doi:10.1093/molbev/msab199

Nip KM, Hafezqorani S, Gagalova KK, Chiu R, Yang C, Warren RL, Birol I. 2023. Reference-free assembly of long-read transcriptome sequencing data with RNA-Bloom2. *Nature Communications 2023 14:1* **14**: 1–14. doi:10.1038/s41467-023-38553-y

Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín R, Estevan-Morió E, Liu T, Nanni A, McIntyre L, et al. 2024a. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nature Methods 2024 21:5* **21**: 793–797. doi:10.1038/s41592-024-02229-2

Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, Loveland JE, De María M, Adams MS, Balderrama-Gutierrez G, et al. 2024b. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat Methods* **21**: 1349–1363. doi:10.1038/s41592-024-02298-3.

Peng S, Dahlgren AR, Hales EN, Barber AM, Kalbfleisch T, Petersen JL, Bellone RR, Mackowski M, Cappelli K, Capomaccio S, et al. 2022. Long-read RNA Sequencing Improves the Annotation of the Equine Transcriptome. *bioRxiv*. doi:10.1101/2022.06.07.495038

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology 2015 33:3* **33**: 290–295. doi:10.1038/nbt.3122

Prjibelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, Tilgner HU. 2023. Accurate isoform discovery with IsoQuant using long reads. *Nature Biotechnology 2023 41:7* **41**: 915–918. doi:10.1038/s41587-022-01565-y

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. 2021. Towards complete and error-free genome

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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.

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**. doi:10.1093/bioinformatics/btg1080

Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. 2020. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature Communications 2020 11:1* **11**: 1–12. doi:10.1038/s41467-020-15171-6

Tang S, Lomsadze A, Borodovsky M. 2015. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res* **43**: e78–e78. doi:10.1093/nar/gkv227

Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, Kariyawasam H, Du MRM, Schuster J, Wang C, et al. 2021. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biol* **22**: 1–24. doi:10.1186/s13059-021-02525-6.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology 2010 28:5* **28**: 511–515. doi:10.1038/nbt.1621

Veeneman BA, Shukla S, Dhanasekaran SM, Chinnaiyan AM, Nesvizhskii AI. 2016. Two-pass alignment improves novel splice junction quantification. *Bioinformatics* **32**: 43–49. doi:10.1093/bioinformatics/btv642

Volden R, Schimke KD, Byrne A, Dubocanin D, Adams M, Vollmers C. 2023. Identifying and quantifying isoforms from accurate full-length transcriptome sequencing reads with Mandalorion. *Genome Biol* **24**: 1–15. doi:10.1186/s13059-023-02999-6

Wang J, Li ST, Zhang Y, Zheng HK, Xu Z, Ye J, Yu J, Wong GKS. 2003. Vertebrate gene predictions and the problem of large genes. *Nat Rev Genet* **4**: 741–749. doi:10.1038/nrg1160

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahmanian S, Forner S, Matheos D, Zeng W, Williams B, Trout D, et al. 2020. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv*. doi:10.1101/672931

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 plant reference genome and improved gene annotation using long-read and paired-end sequencing data. *Scientific Data 2019 6:1* **6**: 1–9. doi:10.1038/s41597-019-0127-1

Xu QS, Liang YZ. 2001. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems* **56**: 1–11. doi:10.1016/S0169-7439(00)00122-2

Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics 2012 13:5* **13**: 329–342. doi:10.1038/nrg3174

Zhang R, Kuo R, Coulter M, Calixto CPG, Entizne JC, Guo W, Marquez Y, Milne L, Riegler S, Matsui A, et al. 2022. A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-seq analysis. *Genome Biol* **23**: 1–37. doi:10.1186/s13059-022-02711-0

Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

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

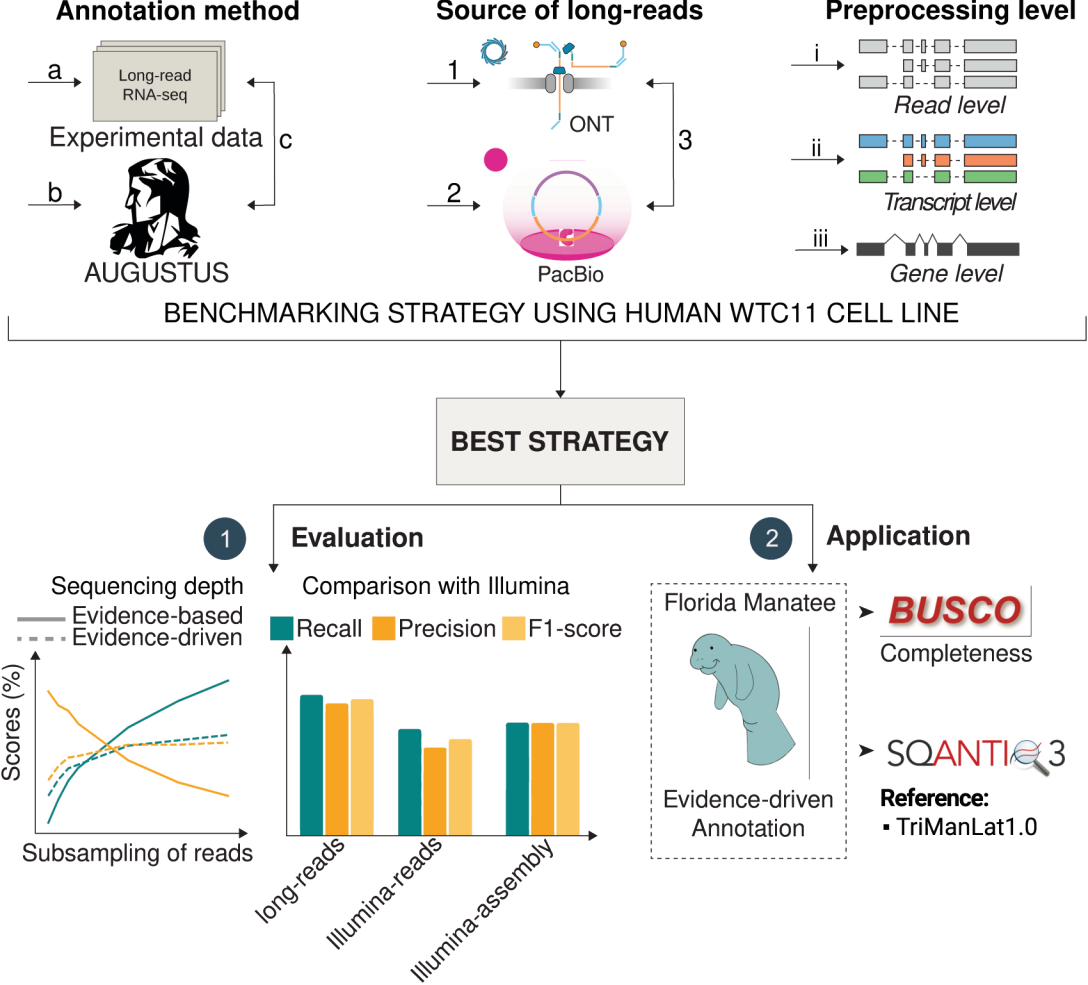
Figure 2. Assessment of the incorporation of long-read data to evidence-driven annotation. (A) Distribution of SQANTI3 categories at the read unique splice-junctions combination, transcript, and collapsed transcript level for PacBio data processed using Iso-Seq3 and ONT and ONT+PacBio data processed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

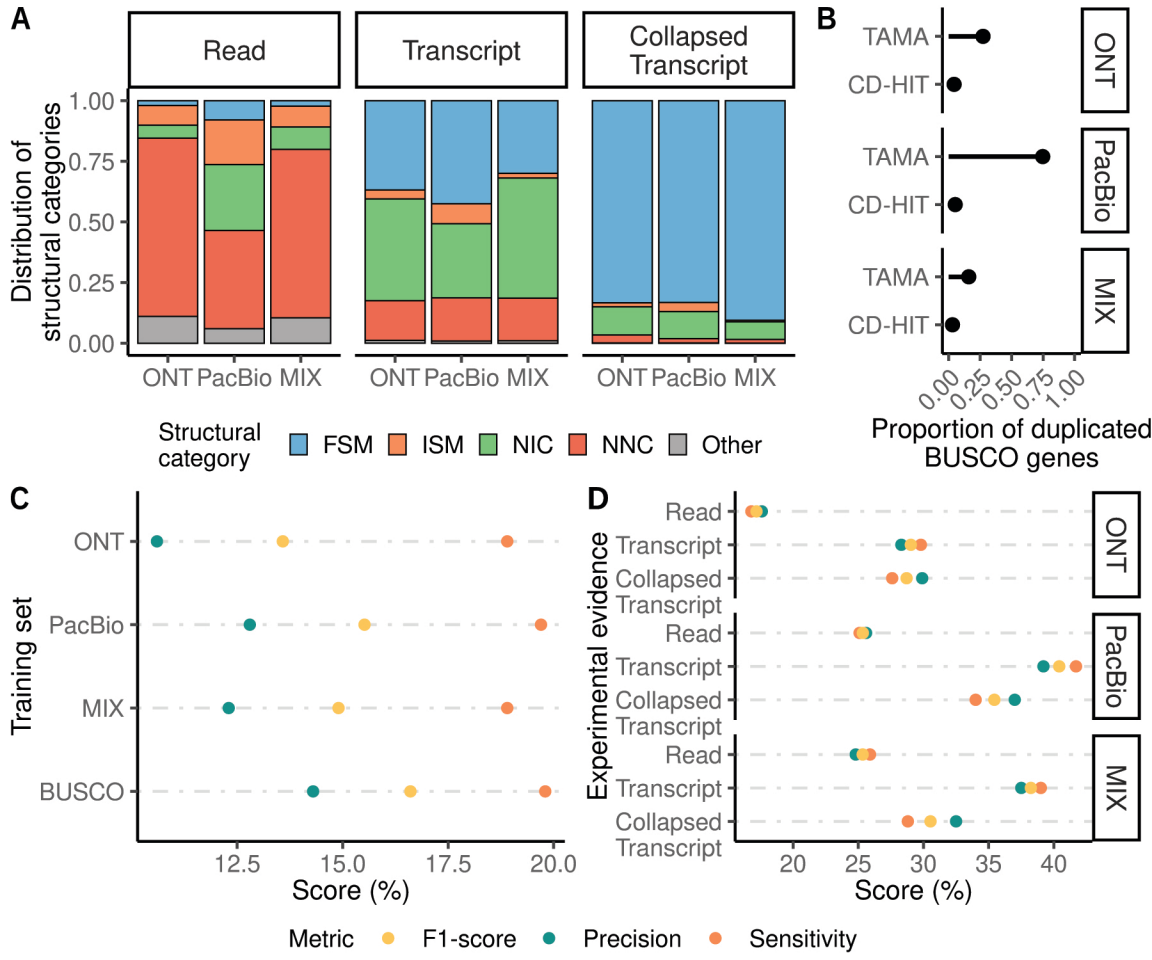
Downloaded from genome.cshlp.org on May 30, 2026 . Published by Cold Spring Harbor Laboratory Press

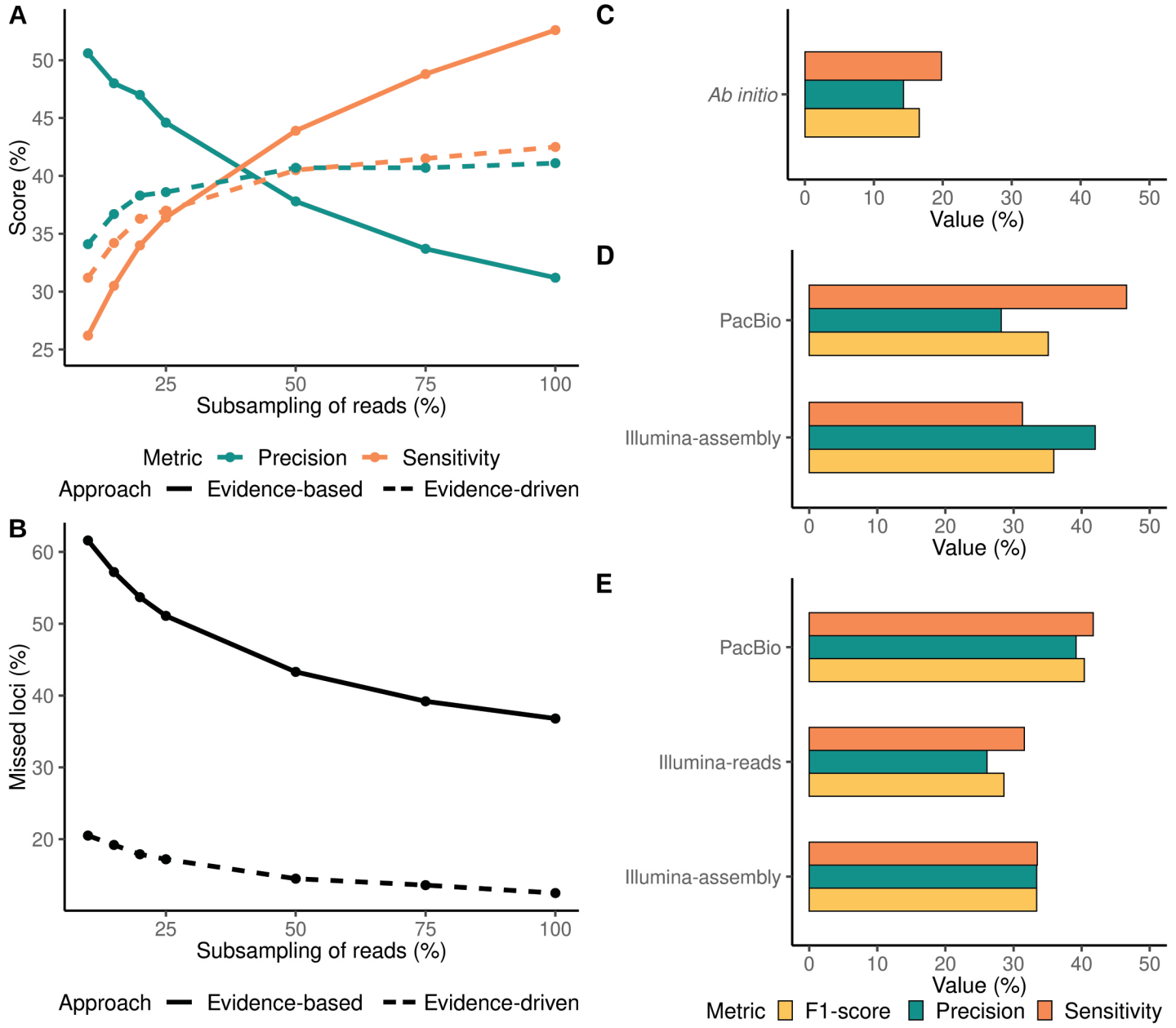
914 predictions by the three models trained with long-read data and the model trained using BUSCO genes.
915 (D) Selection of the different long-reads-based extrinsic evidence sources used for the evidence-driven
916 gene predictions. Reads of PacBio, ONT, and a MIX of both (Read); transcript models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

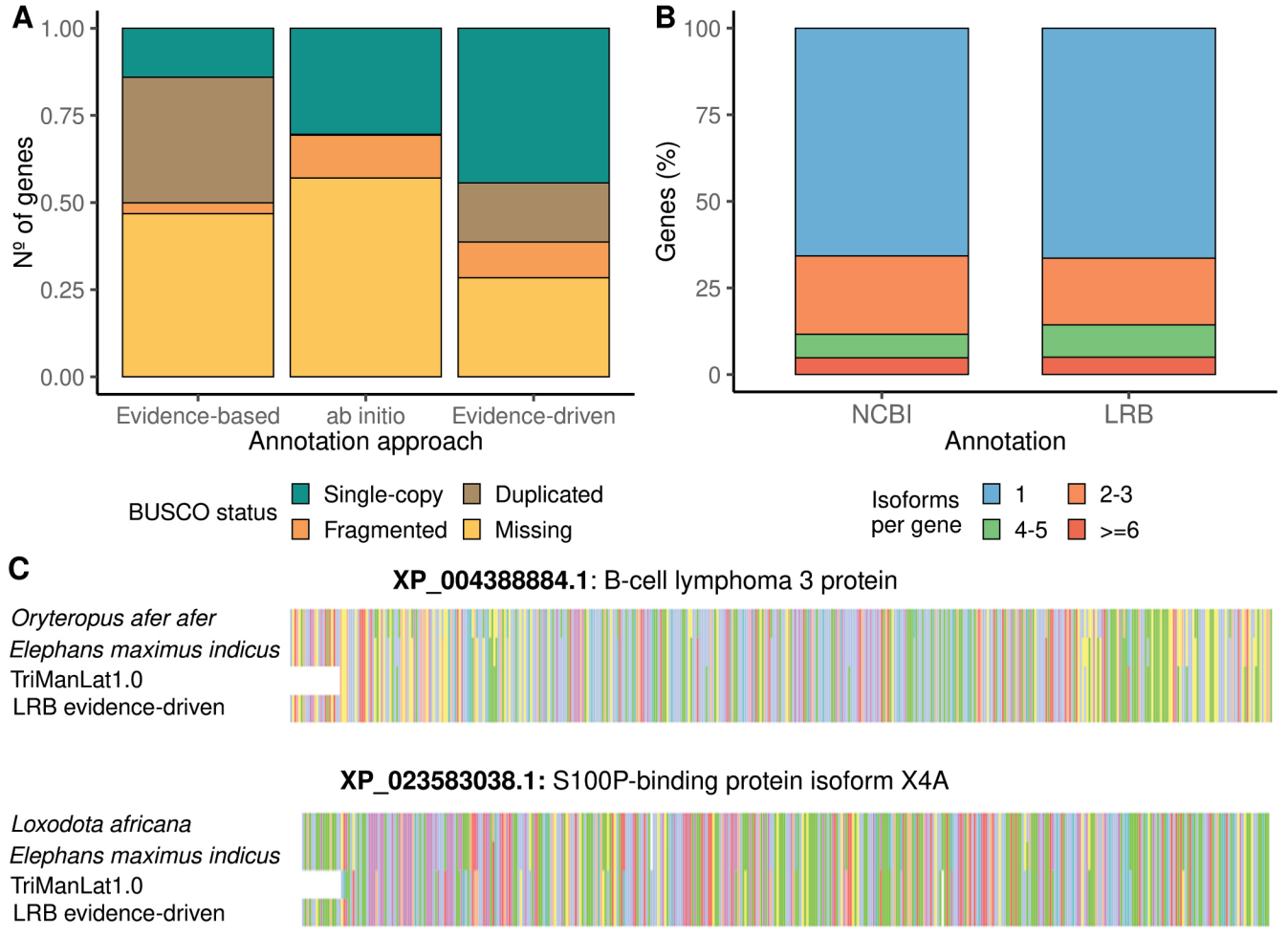
919 **Figure 3.** Performance analysis of gene prediction as a function of the number of reads. Sensitivity,
920 precision (A) and number of missed *loci* (B) were obtained with different sample sizes of WTC11 cell
921 line PacBio FLNC reads using evidence-based and evidence-driven approaches. Performance 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
a
Long-read RNA-seq
Experimental data
c
b
AUGUSTUS
Source of long-reads
1
ONT
3
2
PacBio
Preprocessing level
i
Read level
ii
Transcript level
iii
Gene level
BENCHMARKING STRATEGY USING HUMAN WTC11 CELL LINE
BEST STRATEGY
1
Evaluation
Sequencing depth
Evidence-based
Evidence-driven
Scores (%)
Subsampling of reads
Comparison with Illumina
Recall
Precision
F1-score
long-reads
Illumina-reads
Illumina-assembly
2
Application
Florida Manatee
Evidence-driven Annotation
BUSCO
Completeness
SQANTI3
Reference:
▪ TriManLat1.0


A
Read
Transcript
Collapsed Transcript
Distribution of structural categories
1.00
0.75
0.50
0.25
0.00
ONT PacBio MIX
ONT PacBio MIX
ONT PacBio MIX
Structural category
FSM
ISM
NIC
NNC
Other
B
TAMA
CD-HIT
ONT
TAMA
CD-HIT
PacBio
TAMA
CD-HIT
MIX
0.00
0.25
0.50
0.75
1.00
Proportion of duplicated BUSCO genes
C
Training set
ONT
PacBio
MIX
BUSCO
12.5
15.0
17.5
20.0
Score (%)
D
Experimental evidence
Read
Transcript
Collapsed Transcript
ONT
PacBio
MIX
20
25
30
35
40
Score (%)
Metric
F1-score
Precision
Sensitivity


A
Score (%)
50
45
40
35
30
25
25
50
75
100
Subsampling of reads (%)
Metric
Precision
Sensitivity
Approach
Evidence-based
Evidence-driven
B
Missed loci (%)
60
50
40
30
20
25
50
75
100
Subsampling of reads (%)
Approach
Evidence-based
Evidence-driven
C
Ab initio
0
10
20
30
40
50
Value (%)
D
PacBio
Illumina-assembly
0
10
20
30
40
50
Value (%)
E
PacBio
Illumina-reads
Illumina-assembly
0
10
20
30
40
50
Value (%)
Metric
F1-score
Precision
Sensitivity


A
1.00
0.75
0.50
0.25
0.00
Nº of genes
Evidence-based
ab initio
Evidence-driven
Annotation approach
B
100
75
50
25
0
Genes (%)
NCBI
LRB
Annotation
BUSCO status
Single-copy
Duplicated
Fragmented
Missing
Isoforms per gene
1
2-3
4-5
>=6
C
XP_004388884.1: B-cell lymphoma 3 protein
Oryteropus afer afer
Elephans maximus indicus
TriManLat1.0
LRB evidence-driven
XP_023583038.1: S100P-binding protein isoform X4A
Loxodota africana
Elephans maximus indicus
TriManLat1.0
LRB evidence-driven