Nucleic Acids Research, 2026, **54**, gkag091
https://doi.org/10.1093/nar/gkag091
Methods

Improved reconstruction of transcripts and coding sequences from RNA-seq data

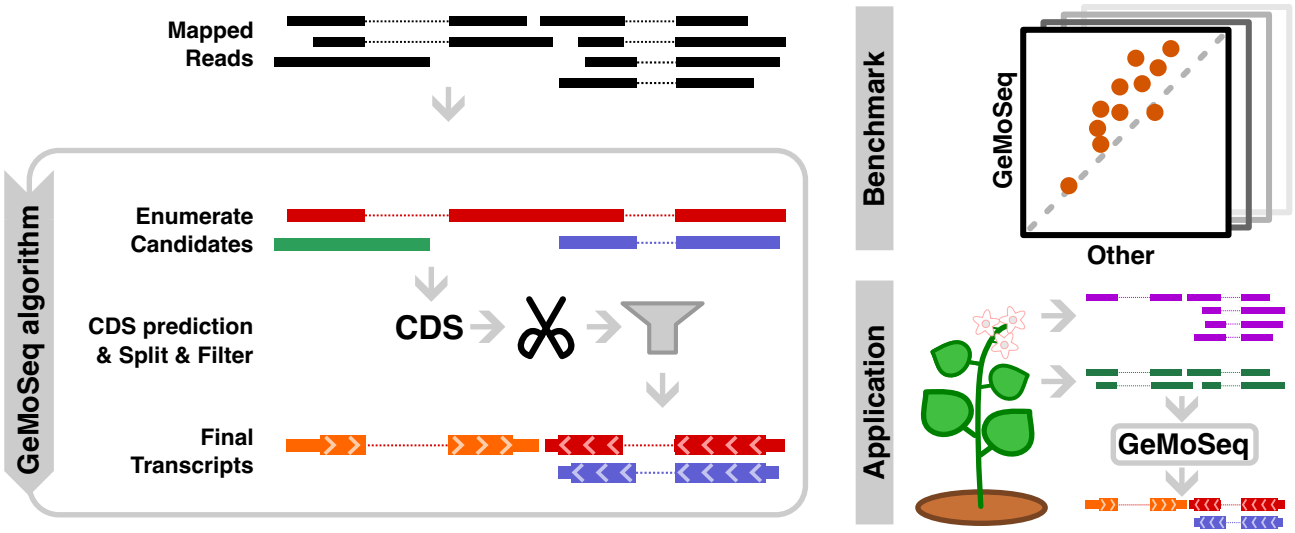
Jan Grau[1,*], **Deborah Weise**[1], **Marika Panster**[2], **Martin H. Schattat**[2], **Jens Keilwagen**[3]

[1]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Von-Seckendorff-Platz 1, 06120 Halle, Saxony Anhalt, Germany
[2]Institute of Biology, Martin Luther University Halle-Wittenberg, Weinbergweg 10, 06120 Halle, Saxony Anhalt, Germany
[3]Institute for Biosafety in Plant Biotechnology, Julius Kühn-Institut, Erwin-Baur-Str. 27, 06484 Quedlinburg, Saxony Anhalt, Germany
*To whom correspondence should be addressed. Email: grau@informatik.uni-halle.de

Abstract

Annotation of genes and transcripts is a key prerequisite for understanding the information that is encoded in newly sequenced genomes. One source of information suited for this purpose is RNA-seq data mapped to the respective genome sequence. RNA-seq-based approaches for transcript reconstruction generate transcript models from these data by combining regions of contiguous coverage (exons) and split read mappings (introns). Understanding phenotypes as a consequence of proteins encoded in a genome further requires the annotation of coding sequences within transcript models. We present GeMoSeq, a novel approach for transcript reconstruction from RNA-seq data that combines combinatorial enumeration of candidate transcripts with heuristics for splitting candidate transcripts into regions of contiguous coverage and subsequent likelihood-based quantification. Prediction of coding sequences is an integral part of the GeMoSeq algorithm. We benchmark GeMoSeq against previous approaches using a large collection of public RNA-seq data for seven species. For the majority of species, we observe an improved prediction performance of GeMoSeq, especially on the level of coding sequences and for species with dense genomes. We combine GeMoSeq with the homology-based approach GeMoMa to re-annotate two recently sequenced genomes of *Nicotiana benthamiana* lab strains, which illustrates the main purpose of GeMoSeq: the initial annotation of newly sequenced genomes with protein-coding genes.

Graphical abstract



Introduction

Gene prediction is one of the longest-standing problems in bioinformatics, and pioneering approaches date back to the 1990s [1, 2]. Accurate predictions of genes, transcripts and transcript variants are essential for virtually any downstream tasks in genomics and transcriptomics, including the construction of pangenomes [3], phylogenomics [4], the detection of orthologs [5] or the analysis of differentially expressed genes or transcripts [6]. Since many phenotypes are expressed via the proteins encoded by (protein-coding) genes [7], the annotation of coding sequences (CDSs) within transcripts is an important additional requirement for functional genomics [8].

Received: April 9, 2025. **Revised:** January 13, 2026. **Accepted:** January 20, 2026

Methods for computational gene prediction can be roughly categorized into *ab initio*, homology-based, and expression-based approaches. *Ab initio* gene prediction uses mathematical or statistical models to represent typical sequence signals of, for instance, coding exons or splice sites to predict transcript models in newly sequenced genomes based on the genomic sequence alone. The most popular instance of a (mostly) *ab initio* approach today is Augustus [9–11], while the recently developed Tiberius builds upon specialized neural network layers and has been reported to further improve prediction accuracy [12].

In homology-based gene prediction, protein sequences or transcript models from well-annotated reference species are mapped to the newly sequenced genome based on sequence homology on the level of (translated) protein sequence. Splicing in eukaryotes can be considered by performing spliced alignments in approaches like miniprot [13] or Spaln [14], or by including the conservation of intron positions as a feature, like in GeneMapper [15] or GeMoMa [16].

Finally, expression-based approaches start from experimental evidence of expressed transcripts, which initially have been cDNA or EST sequences [17], but today are mostly next- or third-generation RNA sequencing data. Hence, we will refer to such methods as RNA-seq-based in the following. One pioneering approach for gene prediction, also termed transcript reconstruction in this context, from RNA-seq data has been Cufflinks [18]. More recent alternatives include Scallop [19] and StringTie [20], where the latter is currently the most popular choice according to citations. Finally, IsoQuant [21] has been specifically designed for third-generation RNA-seq data,

Mixtures and combinations of these prototypical approaches have been implemented in methods and pipelines like MAKER2 [22] or BRAKER3 [23], and many tools allow for the inclusion of hints from at least one additional source of evidence. For instance, Augustus and GeMoMa both make use of additional RNA-seq-based evidence [24–26].

Each of the prototypical approaches has specific limitations. Purely *ab initio* gene prediction is prone to false positive exon predictions and errors in general [23, 27]. RNA-seq-based approaches may only predict genes and transcript variants that are sufficiently expressed and, hence, depend on the diversity of the sequencing libraries considered. By design, homology-based predictions are limited to those genes/proteins that are present in the reference and may, hence, miss orphan, i.e., species-specific genes.

The latter limitation also applies to our homology-based prediction method, GeMoMa. Hence, we aimed at developing an RNA-seq-based prediction method that may complement the GeMoMa predictions to also yield predictions for genes that are not present in the reference species. Since the homology-based prediction of GeMoMa is intrinsically limited to protein-coding genes, this approach should also focus on protein-coding genes and include an internal prediction of coding sequences (CDSs) within transcript models. While non-coding RNAs play important roles in an organism, protein-coding genes are still a common starting point for elucidating the functions encoded in a genome [8]. In addition, the joint prediction using a homology-based and an RNA-seq-based approach allows for identifying high-confidence transcript models with double evidence.

This novel approach, termed GeMoSeq, differs from previous approaches in two central aspects. First, we generate candidate transcripts combinatorially from the exons and introns with RNA-seq evidence, which are subsequently filtered by likelihood-based quantification. Second, we apply several heuristics that split overlapping candidate transcripts while taking CDS predictions into account. We compare the prediction performance of GeMoSeq with the previous approaches Cufflinks, Scallop and StringTie in benchmark studies considering diverse RNA-seq libraries for seven species, namely the three plant species *Arabidopsis thaliana*, *Oryza sativa* and *Solanum lycopersicum*, the three animal species *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus*, and the yeast *Saccharomyces cerevisiae*. We further illustrate how the combination of homology-based (GeMoMa) and RNA-seq-based (GeMoSeq) predictions may be used to balance sensitivity and precision. Finally, we apply the combined pipeline of homology-based and RNA-seq-based (GeMoSeq) prediction to re-annotate two recently *Nicotiana benthamiana* genomes.

Materials and methods

GeMoSeq algorithm

The GeMoSeq algorithms start from a set of reads mapped to the respective reference genome. Mapped reads are then used to build a base-resolution *read graph*, which is further processed into a *splicing graph*, which roughly represents exons and introns. Based on the splicing graph, candidate transcripts are enumerated, which are further tested for possible splits, quantified, and filtered to yield the final prediction. In Fig. 1, we illustrate the general workflow of the algorithm, while each of its steps is described in detail in the following.

Defining genomic regions

The input of the GeMoSeq algorithm are mapped reads in (coordinate sorted) BAM format together with the strand orientation of the sequenced library. Reads are filtered for mapping quality (default: 40). To reduce computational complexity, we first partition the genomic sequence into *regions* with contiguous coverage by mapped reads. In this step, we count a genomic position covered, if either a read base has been mapped to this position or if it is contained in the split region (as indicated by "N" stretches in the cigar string) of a mapped read. Small gaps in the coverage profile (default: 50 bp) may be closed by adding *dummy* reads, which are disregarded in later quantification steps. In the case of a strand-specific library, regions are defined for each strand separately. Regions with fewer than 10 reads in total are discarded. If the coverage on one strand is 50-fold larger than the coverage of the other strand and the Jaccard coefficient of the sets of covered positions in the two regions exceeds 0.5, both regions are merged, retaining the strand orientation of the regions with the larger coverage.

If the length of the resulting region exceeds a user-defined threshold (default: 750 kbp), the region is split into sub-regions at the positions with the lowest coverage. This is repeated until the length of each sub-region is below the defined threshold. The final region or sub-regions, respectively, are used to build the read graph in the following step.

Generating and processing the read graph

The read graph is generated separately for each genomic region. In the read graph, each genomic position within the region is represented by a node. Two nodes are connected by an

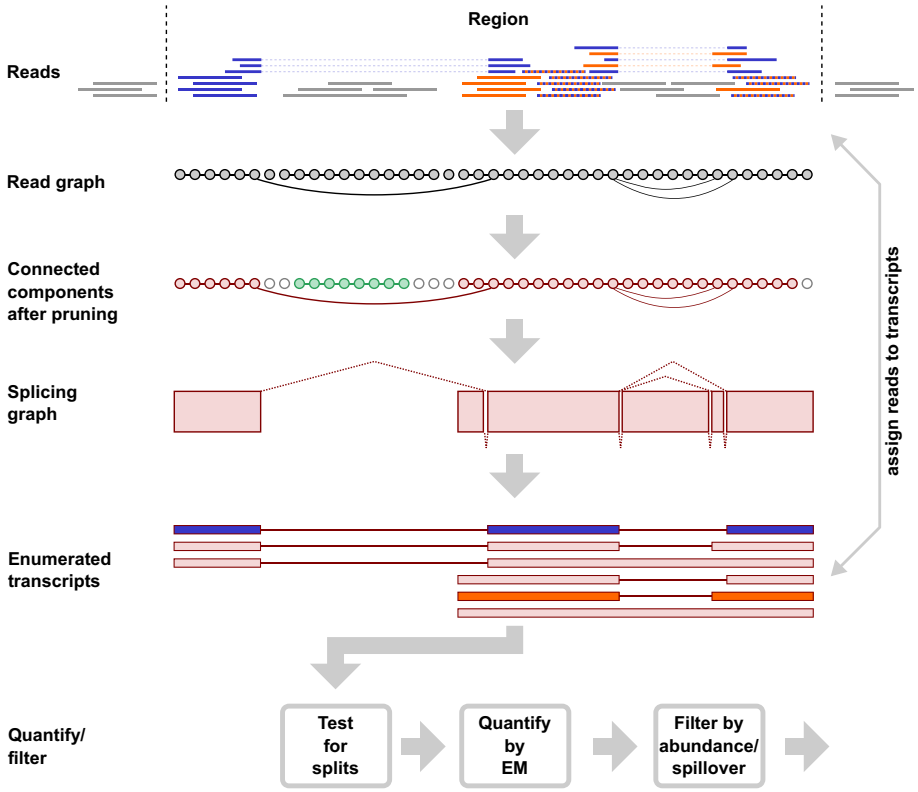


Figure 1. Schematic representation of the GeMoSeq algorithm. First, a nucleotide-resolution read graph is constructed from the mapped reads, which is decomposed into connected components and converted to a splicing graph. Based on the splicing graph, candidate transcript models are enumerated combinatorially and tested for possible splits. Reads are assigned to compatibility classes of transcript models as indicated for the two highlighted examples of the enumerated transcripts and corresponding highlighting of the original reads, where striped reads are compatible with both example transcripts. Transcript abundance is quantified by an EM algorithm on the compatibility classes of reads, and finally filtered and reported.

edge if two adjacent bases of a read have been mapped to the respective position. This may apply to nodes of the read graph that represent directly adjacent genomic positions if these are contained in a read stretch that has been mapped contiguously, but also to nodes representing more distant genomic positions in case of split reads (cf. Fig. 1, Read graph). In the latter case, edges are only considered if they connect nodes that are more distant than a user-defined threshold (default: 10 bp) as these may later be considered as putative introns of transcript models. For each node and each edge, we further record the set of supporting reads.

Edges are subsequently pruned by absolute and relative coverage according to user-defined thresholds. After pruning, the read graph is decomposed into connected components to reduce computational complexity in the following steps.

Generating the splicing graph

Each connected component of the read graph is further processed into a splicing graph. In the splicing graph, stretches of connected nodes without alternative edges that represent contiguous genomic positions are merged to *contiguous regions* (cf. Fig. 1, Splicing graph, red boxes). In case of alternative edges from one node to multiple alternative nodes of the read graph, the respective contiguous regions are connected by edges. This may include edges between contiguous regions directly adjacent in genomic coordinates, but also to non-neighboring or more distant contiguous regions. The latter edges will later be considered as putative introns. Both contiguous regions and edges in the splicing graph inherit supporting reads from the respective nodes and edges of the previous read graph.

Enumerating candidate transcripts

Based on the splicing graph, we combinatorially enumerate possible transcript models. Conceptually, we start from each of the contiguous regions that are connected to other contiguous regions only at their right end, i.e. that do not have incoming edges, reading the splicing graph from left to right. From one contiguous region, we follow outgoing (right) edges to the next contiguous region one by one, where each alternative edge taken results in the generation of an additional, nascent transcript model. This depth-first procedure is repeated, combinatorially generating nascent transcript models in each step, until we reach a contiguous region without an outgoing (right) edge. In the enumerated transcript models, directly adjacent contiguous regions are merged into exons, while edges between these are considered as introns. All transcript models generated in this manner are reported as putative transcripts for the following steps. In some cases, this procedure may result in a combinatorial explosion of transcript models. Hence, if the number of enumerated transcripts exceeds 1000, the procedure is stopped before completion, and only transcripts enumerated up to this point are considered. To reduce the loss of putatively relevant transcript models, outgoing edges are considered in the order of their coverage, and further coverage-based heuristics are applied to decide if an edge is followed in this case.

Splitting candidate transcripts

If two neighboring true transcripts are located close to each other on the genome, it may frequently happen that the coverage profiles of both transcripts overlap. With the previously described procedure, such true transcripts may share a contiguous region in the splicing graph (as these are defined by contiguous coverage without alternative edges) and, subsequently, are represented by a joint transcript model. To counteract this issue, we implemented multiple splitting heuristics that are applied to each of the enumerated transcript models.

These split heuristics include rules for the preservation of coding sequences (CDS). To this end, we apply a simple "longest ORF" CDS prediction, where the spliced sequence according to a transcript model is conceptually translated in all three reading frames, and the longest stretch between a start codon and the next in-frame stop codon is considered as CDS. The same CDS prediction is also applied when reporting final transcript models in GeMoSeq.

A first split heuristic is based on the coverage profile within the candidate transcript model. Here, we search for pronounced "dips" in the coverage profile, i.e. local coverage minima (Supplementary Fig. S1), or a pronounced coverage imbalance, where one part of the transcript model has substantially less coverage than another part (Supplementary Fig. S2). Specifically, we compute the average coverage in sliding windows of 50 bp (small) and 1000 bp (large) and search for small windows with less than 20% of the coverage of the less-covered, neighboring large window. In addition, we search for positions, where the coverage in the preceding large window is at least 4-fold greater/less than in the subsequent large window. Positions fulfilling either of these criteria are tested for a possible split. A split is performed only if the original transcript model has at least one intron and the two split candidate transcripts still each contain a proper CDS prediction with a user-defined minimum number of amino acids (default: 70). In case of a split, the two resulting new transcript models may be further split repeating the same procedure.

A second split heuristic is applied only in case of non-strand-specific libraries, where overlapping transcripts may occur in head-to-head or tail-to-tail configuration. Here, we consider the orientation of splice sites within the enumerated transcript (Supplementary Fig. S3). The orientation of splice sites may be determined from the canonical donor (GT/GC) and acceptor (AG) di-nucleotides at the first and last positions of putative introns. If we identify an exon in the transcript model that separates a contiguous stretch of introns indicating one orientation from a contiguous stretch of introns indicating the opposite orientation, this exon is considered for a split following the same logic as for the first heuristic but without the requirement of a CDS prediction in both split transcript models. Since we cannot determine an accurate split position from splice sites alone, the exon at the border between the two regions with differing orientation is included in both split transcript models.

A third heuristic is considered only if neither of the two previous heuristics had an effect. In that case, the CDS prediction of the original transcript model is taken into account, and we test if the remainder of the transcript (i.e. 5" UTR or 3" UTR according to the original CDS prediction) may harbor another CDS of sufficient length. In addition, we require that both resulting split transcript models contain at least two introns. The latter heuristic shall reduce the number of false-positive splits in long UTRs with declining coverage. If these criteria are met, the original transcript is split into two transcript models.

Assigning reads to transcripts

For subsequent quantification, reads need to be assigned to candidate transcript models. Specifically, we assign reads to *compatibility classes* based on their concordance to one or multiple transcript models, which is inspired by the definition of equivalence classes in kallisto [28]. A read is compatible with a transcript model if all of its mapped bases fall into the transcript"s exons, and all split regions in the mapping match its introns. By this definition, a read may be compatible with multiple candidate transcript models as illustrated in Fig. 1.

We define the read r and candidate transcript model t

$$\delta_{t,r} = \begin{cases} 1 \text{ if } r \text{ compatible with } t \\ 0 \text{ otherwise.} \end{cases}$$

Since contiguous regions and edges in the splicing graph and, hence, exons and introns of the transcript model store lists of supporting reads, the assignment can be solved efficiently by set operations (union, set difference) on those (sorted) lists.

Quantifying transcript abundance

Quantification of candidate transcript abundance is performed using a likelihood-based approach, again inspired by kallisto [28]. Let R denote the set of reads that have been assigned to any of the candidate transcripts that originated from the current connected component of the read graph. Let T denote the set of those candidate transcripts. Let denote w_t the *a priori* weight of a candidate transcript t, and v_r the weight of a read r, chosen according to the heuristics described in Supplementary Methods.

The likelihood of the set of reads given the abundance vector $\underline{\alpha}$ of the candidate transcripts is proportional to

$$P(R|\underline{\alpha}) \propto \prod_{r \in R} \left(\sum_{t \in T} \delta_{t,r} \cdot w_t \cdot \alpha_t \right)^{v_r}. \quad (1)$$

Transcript abundances α_t are determined by maximizing the likelihood (equation 1) with respect to the vector $\underline{\alpha}$ using the expectation maximization (EM) algorithm.

Filtering transcripts

Based on the transcript abundances α_t obtained in the previous step, candidate transcripts are filtered by absolute and relative abundance. Specifically, candidate transcripts are ordered according to their abundance, and candidate transcripts are added to the output set until:

(1) The absolute abundance falls below a user-defined threshold (default: 20),
(2) The (relative) abundance falls below a user-defined threshold (default: 0.05),
(3) The already added transcripts explain a user-defined fraction of the total abundance (default: 0.9), or
(4) The fraction $\frac{\alpha_t}{\alpha_{t'}}$ of abundance relative to the next candidate transcript t' in sorted order is larger than a user-defined threshold (default: 20).

In case of strand-specific libraries, a small fraction of reads may also occur on the opposite strand of the original transcripts, and such "spillover," single-exon transcripts are filtered by a heuristic based on local coverage. Multi-exon

candidate transcripts are also filtered if their strand orientation according to their donor and acceptor splice sites does not match the strand orientation according to the library, and single-exon candidate transcripts are filtered if the theoretical abundance of their CDS (i.e. abundance multiplied by fraction of CDS relative to total length) falls below the threshold on absolute abundance (criterion (a) above).

Handling of long-read data

For handling long-read (e.g., Pacific Biosciences or Oxford Nanopore Technologies) mapping data, a few adaptations of the described algorithm have been implemented in GeMoSeq, since current long reads may contain mismatches and insertions/deletions relative to the reference genome at a higher frequency than short (e.g., Illumina) reads, and typically span multiple introns.

First, the requirement of at least 10 reads per region is lifted. Second, edges are pruned from the read graph if their number of supporting reads is less than a fifth of the number of supporting reads of an alternative edge that shares one node with the edge in question while the pair of second nodes has a distance of at most 5. Together with ignoring short deletions (according to the CIGAR string) directly adjacent to introns, this shall prevent the read-centric enumeration of spurious candidate transcripts due to sequencing artifacts in the next step. Third, the combinatorial enumeration of transcripts is replaced by a read-centric algorithm. A typical long read spans multiple exons and introns, and ideally even covers a transcript end-to-end. For this reason, we derive the respective combination of contiguous regions and introns of the splicing graph directly from each long read, which are considered candidate transcripts and are technically stored as bit vectors. We then filter these candidate transcripts for redundancies, which include perfectly identical bit vectors, but also fully contained transcripts with low abundance. Due to the extended handling of introns in the second adaptation, small sequencing artifacts at the borders of introns will not affect the resulting bit vectors and will be considered redundant in this step. The resulting candidate transcripts are then subjected to the same quantification approach as those derived by combinatorial enumeration for short reads. Here, the weighting of reads in the quantification process is adapted as described in Supplementary Methods.

Due to the differing procedures for enumerating candidate transcripts, a fundamental step of the GeMoSeq algorithm, GeMoSeq does not offer a "hybrid" mode for jointly predicting from short-read and long-read data. For this purpose, we recommend performing independent runs of GeMoSeq in the respective modes for short-read and long-read data. The resulting predictions can afterwards be merged using the *GeMoMa Annotation Filter* tool of GeMoMa as described below, and specific weights can be assigned to the individual predictions.

Tuning heuristics

Heuristics have been derived and tuned by systematic assessment of prediction performance and manual inspection of critical cases (e.g., missing transcripts, fusion transcripts) based on the following libraries (ENA/SRA accessions): *A. thaliana*: ERR1886195, ERR2245559, and DRR09053; *D. melanogaster*: SRR3659025; *M. musculus*: ERR2983488; *O. sativa*: SRR1785721. In addition, long-read (PacBio SMRT) data for *A. thaliana* (SRR12350726) have been used for testing the application of GeMoSeq to long reads. None of these datasets have been used to evaluate prediction performance in the benchmark studies.

Benchmark data

GeMoSeq has been benchmarked on short-read (Illumina) data for *A. thaliana*, *O. sativa*, *S. lycopersicum*, *C. elegans*, *D. melanogaster*, *M. musculus*, yeast (*S. cerevisiae*), and, in addition, *H. sapiens*. Genome sequences and annotations of all species have been obtained from the sources listed in Supplementary Table S1.

For each species, 40 paired-end libraries were selected at random from those present in the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena/browser/home) that have been sequenced using Illumina HiSeq or NovaSeq instruments. Due to the focus on protein-coding genes, libraries have been stratified for library selection: 85% of the libraries were sequenced after poly-A enrichment (field `library_selection` cDNA, PolyA or Oligo-dT), while 15% where sequenced after rRNA depletion (Inverse rRNA (selection)). In addition, we aimed at a considerable fraction of strand-specific and non-strand-specific libraries, which is not represented by a standardized field in ENA. Hence, strand-specificity of libraries was determined by custom scripts after mapping the reads to the respective reference genome as detailed below. For *S. cerevisiae*, the initially sampled libraries did not contain a sufficient number of non-strand-specific libraries. For this reason, the number of sampled libraries was increased to 60, yielding 9 non-strand-specific libraries (cf. Supplementary Fig. S4). A list of all sampled libraries (ENA accessions) is provided as Supplementary Table S3, where a few libraries failed processing for different reasons (*D. melanogaster*: 2; *O. sativa*: 1; *S. cerevisiae*: 3).

Benchmarking procedure

Sampled libraries were mapped to the respective reference genome (Supplementary Table S1) using STAR 2.7.10b with parameters `--alignIntronMax` and `--alignMatesGapMax` set to 100,000 for *M. musculus*, *H. sapiens* and *D. melanogaster*, and to 20,000 for the remaining species to avoid spurious mappings with very long splits that sometimes occur in STAR mappings. Further non-standard parameters were `--runThreadN 4 --outSAMtype BAM SortedByCoordinate`. Finally, we set `--outSAMstrandField intronMotif` to allow for gene prediction using cufflinks from non-strand-specific libraries.

After mapping, the strand specificity of each library was determined using a custom script based on featureCounts [29] using the respective reference annotation (Supplementary Table S1). If the number of reads assigned by featureCounts using one strand orientation was at least 2-fold the number for the other strand, the library was reported as `FR_FIRST_STRAND` (following cufflinks nomenclature) or `FR_SECOND_STRAND`, respectively. If the ratio between both parameters was less than 2, the library was reported as `FR_UNSTRANDED`.

For benchmarking purposes, GeMoSeq, StringTie v2.2.1, Cufflinks v2.2.1, and Scallop v0.10.5 were applied to each of the mapped libraries separately.

GeMoSeq was run with default parameters except for parameter `s` set to the key for the respective strand specificity

of the library. StringTie [20, 30, 31] was run with default parameters except for `-p 6`, and the parameters for strand specificity were set to `--rf` in case of `FR_FIRST_STRAND`, `--fr` for `FR_SECOND_STRAND` in case of strand-specific libraries. Cufflinks [18] was run with default parameters except for `-p 6`, and `--library-type` set to the key for the respective strand specificity of the library. Scallop [19] was run with default parameters except for `--library_type` set to `first` in case of `FR_FIRST_STRAND` and `second` for `FR_SECOND_STRAND` in case of strand-specific libraries.

The output of GeMoSeq, StringTie, Cufflinks, and Scallop was compared against the respective reference annotation (Supplementary Table S1) using gffcompare v0.12.6 [32]. The gffcompare output was analyzed primarily on the "Transcript" level and is intended to provide a measure of prediction accuracy on the level of transcript models, irrespective of CDS annotations.

On the level of CDS, the output of all four tools was further compared against the reference using the "Analyzer" tool of GeMoMa [16, 25, 26] with default parameters. Since StringTie, Cufflinks, and Scallop do not perform CDS prediction internally, we followed two alternative strategies: first, we made the simplistic CDS prediction of GeMoSeq applicable to external GFF/GTF files in a separate tool, and applied this tool to the output of StringTie, Cufflinks, and Scallop. Second, we applied TransDecoder v5.7.0 to the output of these three tools (cf. Supplementary Methods), since the simplistic prediction built into GeMoSeq might unnecessarily reduce the prediction performance of the alternative tools. We did not apply TransDecoder to GeMoSeq output, since the CDS predictions of TransDecoder might contradict those that have been used by GeMoSeq-internal heuristics when splitting candidate transcripts.

Benchmarking long-read data

In addition, we benchmarked GeMoSeq on long-read (PacBio, ONT) data. GeMoSeq and StringTie provide a dedicated long-read mode, while Scallop-LR [33] requires PacBio-specific header files and has not been updated for five years. Hence, we limited the comparison to GeMoSeq and StringTie in this case.

Benchmark data have been obtained from ENA for the same seven species as for the Illumina short-read data. Datasets were sampled from those available at ENA with at least 100,000 reads to yield 5 datasets per instrument of the PacBio Sequel and Sequel II, and the ONT GridION, MinION, and PromethION instruments. If fewer than five datasets were available for an instrument, further datasets were selected from the remaining at random. For *O. sativa*, this resulted in 15, and for *S. lycopersicum,* in 13 datasets, while for the remaining species the desired number of 25 datasets could be obtained. A list of all sampled libraries (ENA accessions) is provided as Supplementary Table S4.

Long reads were mapped to the respective genome sequence using minimap2 2.17-r941 [34] with parameters `-x splice` and `-G` set to 100,000 for *M. musculus* and *D. melanogaster*, and to 20,000 for the remaining species.

Strand specificity of the libraries was determined in analogy to the Illumina short-read data, but with an additional flag `-L` for the featureCounts call.

StringTie was run on the long-read datasets with additional parameter `-L` and GeMoSeq was run with additional parameters `lr=true mrpg=5 mrpt=5`, where the first indicates the long-read mode of GeMoSeq, while the other two parameter lower the number of supporting reads that are required to report a gene or transcript.

The remainder of the benchmarking procedure (gffcompare, CDS prediction, Analyzer) was completed in analogy to previous short-read data.

Homology-based prediction using GeMoMa

Since GeMoSeq shall complement the homology-based predictions of GeMoMa, we additionally used GeMoMa to predict transcript models for all seven species, considering existing transcript models from three closely related species as reference. Conceptually, this scenario is not typical for a homology-based prediction, because the seven well-annotated species considered in this manuscript would rather be used as a reference when annotating newly sequenced genomes, and may have also been used to annotate those considered as a reference here.

The species and genome versions used as reference for predicting transcript models in the genomes of *A. thaliana*, *O. sativa*, *S. lycopersicum*, *C. elegans*, *D. melanogaster*, *M. musculus*, and *S. cerevisiae* based on sequence homology are listed in Supplementary Table S5.

We use the *Extract RNA Evidence* tool of GeMoMa to extract intron information from the respective RNA-seq libraries, specifying the strand specificity determined previously, and merge these from the individual libraries using the *Denoise Introns* tool of GeMoMa. This intron information may guide GeMoMa in the specific selection of splice sites during the homology-based prediction of transcript models. For predictions using GeMoMa, we specify the respective reference annotations and the files with the intron information as input to the *GeMoMa pipeline* tool. Predictions for individual species are stored for further processing in the next step.

Merging predictions from multiple libraries

First, we merge the individual predictions of GeMoSeq used for the benchmark and we merge the individual predictions of GeMoMa based on the three reference species per target species using the *GeMoMa Annotation Filter* tool of GeMoMa (cf. Supplementary Methods).

Second, the resulting merged predictions of GeMoSeq (RNA-seq-based) and GeMoMa (homology-based) are merged using the auxiliary *Merge* tool of GeMoSeq. Here, we consider the modes "intersect," which only reports transcript models with matching CDSs in the GeMoSeq and GeMoMa predictions, "union," which reports the redundancy-filtered union of the GeMoSeq and GeMoMa predictions, and "intermediate," which reports all transcript models of the "intersect" mode and additional, quality-filtered predictions from the "union" model (cf. Supplementary Methods).

N. benthamiana data

We use the combination of RNA-seq-based predictions of GeMoSeq and homology-based predictions of GeMoMa to re-annotate the recently published "NbLab360" [35] and "NbKLAB" [36] strains of *Nicotiana benthamiana* (cf. Supplementary Table S7). To this end, we again sampled public RNA-seq libraries of *N. benthamiana* from the European Nucleotide Archive as described in Supplementary Methods.

A list of all sampled 319 libraries (ENA accessions) is provided as Supplementary Table S6.

Homology-based predictions were based on the reference species *A. thaliana*, *S. lycopersicum*, and *Nicotiana tabacum*, and based on previous *N. benthamiana* genome versions 1.0.1 and 2.6.1. For *A. thaliana* and *S. lycopersicum*, we used the genome versions listed in Supplementary Table S1, while the genome versions of *N. benthamiana* and *N. tabacum* are listed in Supplementary Table S7.

Predicting transcripts in *N. benthamiana*

For predicting transcript models in the NbLab350 and NbKLAB genomes of *N. benthamiana*, we essentially follow the same procedure as previously described for the benchmark studies. Briefly, we map the RNA-seq reads to the respective genome sequence and run GeMoSeq for each of the *N. benthamiana* RNA-seq datasets separately and merge these individual predictions using the *GeMoMa Annotation Filter* tool. We also run GeMoMa for the three reference species and the two previous *N. benthamiana* genome versions and merge these as well. Finally, we merge the merged predictions of GeMoSeq and GeMoMa using the auxiliary "merge" tool of GeMoSeq. In addition to the "intersect," "union" and "intermediate" modes, we run the "merge" tool in "annotate" mode, which results in the same set of output predictions as the union tool, but annotates transcript predictions as "high" confidence if they had been reported in the "intersect" mode, as "medium" confidence if they had been reported in the "intermediate" mode, and as "low" confidence otherwise.

Experimental validation of transcript predictions in *N. benthamiana*

Gene selection

Transcripts for validation were chosen based on RNA-seq read coverage (> 500 reads in sequencing run SRR25703580) to ensure sufficient expression in mature leaf tissue.

Primer design

Primer pairs were designed to amplify cDNA fragments spanning all predicted and corrected splice sites for each selected transcript (Supplementary Table S8) and were obtained from BioLegio (Nijmegen, Netherlands).

cDNA

Total RNA was extracted from wild-type *N. benthamiana* leaf tissue using the NucleoSpin RNA Plant kit (Macherey-Nagel, Düren, Germany) according to the manufacturer's instructions. First-strand cDNA was synthesized from total RNA using the SuperScript™ VILO™ cDNA Synthesis Kit (Invitrogen, Carlsbad, CA, USA) with oligo(dT) primers following the manufacturer's protocol.

PCR

PCR amplifications were performed using the synthesized cDNA as template with Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific, Waltham, MA, USA) on different dilutions of the cDNA synthesis product. To facilitate simultaneous processing on a single thermal cycler, amplicons were grouped by expected product size into three categories: small (< 800 bp), medium (800 – 2000 bp), and large (> 2000 bp).

PCR direct sequencing

PCR products were visualized by electrophoresis on 1% agarose gels, and products of expected size were excised and purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel, Düren, Germany). Purified products were directly sequenced by Sanger sequencing (Microsynth AG, Balgach, Switzerland) using the respective PCR primers. For transcripts yielding low product concentrations, a second round of PCR was performed using purified first-round products as a template to generate sufficient material for sequencing. For transcript niben101_Niben101Scf00127g04004.1_R0, initial PCR using primers G123 and G124 with 1:100 diluted cDNA yielded low product concentrations, and a nested PCR approach was employed using internal primers G168 and G171 with the first-round PCR product as template. Direct sequencing of the nested PCR product revealed a strong signal overlap, suggesting multiple transcript variants.

PCR product subcloning

To resolve individual sequences, the nested PCR product was cloned into the pJet1.2/blunt cloning vector (Thermo Fisher Scientific, Waltham, MA, USA) and transformed into *E. coli* TOP10 competent cells according to the manufacturer"s instructions. Plasmid DNA from six individual colonies was isolated and sequenced using pJet1.2 forward and reverse sequencing primers as well as one internal sequencing primer needed to cover the whole cDNA fragment. Sequencing results were analyzed using Geneious Prime 2019 (Biomatters Ltd., Auckland, New Zealand) and aligned to predicted transcript models to verify exact matches of exon-intron boundaries and coding sequences.

Results and discussion

Benchmark using short-read data

GeMoSeq has been benchmarked on short-read (Illumina) data for three plant species (*A. thaliana*, *O. sativa* and *S. lycopersicum*), three animal species (*C. elegans*, *D. melanogaster*, and *M. musculus*), and one yeast (*S. cerevisiae*). These species have been specifically selected for their mature genome sequences and high-quality genome annotations, since these genome annotations serve as a "gold standard" in the evaluation of prediction performance.

Evaluation of the level of transcripts

After applying Cufflinks, Scallop, StringTie, and GeMoSeq to each of the RNA-seq datasets used for benchmarking (cf. Benchmark data), we evaluated their prediction performance on the level of transcripts using gffcompare [32]. On the level of transcripts, gffcompare reports sensitivity and precision, which have been used to compute the respective F1-measure as shown in Fig. 2A. In Supplementary Fig. S12, we additionally consider the pairwise differences in F1 measure of GeMoSeq versus all three alternative tools together with a statistical assessment (two-sample Wilcoxon test) of their significance.

In general, F1 values well below the theoretical maximum of 1.0 are to be expected, since not all genes will be expressed in a single experiment. For the majority of species, we observe that Cufflinks yields lower F1 values than the remaining tools. Cufflinks has been pioneering for the prediction of transcript models from RNA-seq data, but its development has stalled

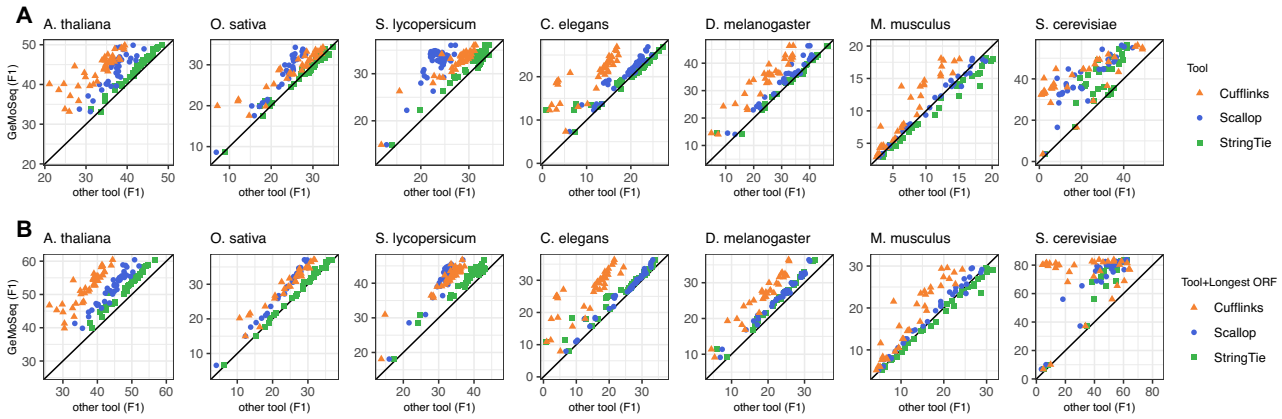


Figure 2. Benchmark results of GeMoSeq compared with Cufflinks, Scallop, and StringTie (**A**) based on the evaluation using gffcompare on the level of transcripts and (**B**) using GeMoMa Analyzer on the level of CDS. Each panel displays the F1 measure based on the sensitivity and precision of GeMoSeq (ordinate) compared with the remaining tools (abscissa, color coded). Points above the main diagonal indicate an improved performance of GeMoSeq, while points below the diagonal indicate an improved performance of the respective alternative tool.

over the last few years, and it has been *de facto* replaced by StringTie. The prediction performance of GeMoSeq and StringTie is widely comparable, with some species-specific differences. For *A. thaliana* and *S. lycopersicum*, GeMoSeq yields slightly but consistently and significantly better F1 values than StringTie. For *O. sativa*, *C. elegans*, and *D. melanogaster*, the prediction performance of GeMoSeq and StringTie is similar, and differences in prediction performance are not significant. For *M. musculus*, StringTie typically achieves higher F1 values than GeMoSe,q and this difference in prediction performance is statistically significant. Scallop typically performs worse than StringTie (and GeMoSeq). For *M. musculus*, the F1 measures achieved by Scallop and GeMoSeq are widely similar. For *S. cerevisiae*, we observe the largest absolute difference in prediction performance between GeMoSeq and all three alternative tools, and the improvement achieved by GeMoSeq is highly significant.

Considering sensitivity on the level of transcripts as reported by gffcompare (Supplementary Fig. S5), we again observe lower sensitivity values for Cufflinks, while GeMoSeq, StringTie and Scallop achieve similar levels of sensitivity for the majority of species. For *M. musculus*, we again observe an advantage of StringTie, while for *S. cerevisiae* GeMoSeq yields substantially higher sensitivity values than any of the other tools.

Regarding precision (Supplementary Fig. S6), we find a clear advantage of GeMoSeq compared with the remaining three approaches for *A. thaliana*, *O. sativa*, *S. lycopersicum* and *D. melanogaster*, while for some *M. musculus* datasets, StringTie achieves the highest precision, and for *S. cerevisiae* either StringTie or Scallop achieve improved precision values compared with GeMoSeq for a subset of datasets.

In summary, we find an advantage of GeMoSeq over StringTie, Scallop, and Cufflinks, especially for the plant species and yeast. For the animal species except *M. musculus*, the performance of StringTie and GeMoSeq is on the same level, while Cufflinks and Scallop perform worse. For *M. musculus*, we observe an advantage when using StringTie instead of GeMoSeq. The lower prediction performance of GeMoSeq compared with StringTie for *M. musculus* may raise concerns regarding the utility of GeMoSeq for human datasets. Hence, we additionally tested all tools for a collection of human datasets (cf. Supplementary Table S3). We find (Supplementary Fig. S15A/D) that results for *H. sapiens* are largely comparable to those obtained for *M. musculus*. However, due to its consistent performance, StringTie is part of many genome annotation pipelines and, hence, genome annotations may be biased towards StringTie in some cases. Improved levels of the F1 measure when using GeMoSeq are mostly due to improved precision of its predictions, while sensitivity reaches similar levels as StringTie and Scallop for the majority of species.

Evaluation of the level of coding sequences

Since GeMoSeq has been primarily developed to complement the homology-based predictions of GeMoMa and for this reason has a focus on protein-coding genes, we further compare the performance of GeMoSeq and the three alternative approaches on the level of CDS (coding sequence) prediction. GeMoSeq makes simplistic CDS predictions (longest ORF) internally, while the other three tools predict exons but no CDSs. Hence, we apply the CDS prediction of GeMoSeq to the predictions of these tools for the purpose of benchmarking. We analyze prediction performance by the *Analyzer* tool of GeMoMa, as gffcompare does not allow for a CDS-centric evaluation. The results of this evaluation are presented in Fig. 2B for the F1 measure, where we find an improvement of GeMoSeq compared with the other approaches for *A. thaliana*, *S. lycopersicum*, *D. melanogaster* and *S. cerevisiae* and, although less pronounced, for *O. sativa* and *C. elegans*. The improvement in prediction performance of GeMoSeq compared with the previous tools is highly significant for all of these species (Supplementary Fig. S13). For *M. musculus* (and *H. sapiens*, Supplementary Fig. S15B/E), StringTie still yields slightly but significantly better F1 values than GeMoSeq. Evaluation on the levels of sensitivity (Supplementary Fig. S7) and precision (Supplementary Fig. S8) supports the previous observation that the improved performance of GeMoSeq is widely due to improved levels of precision, while sensitivity values are more similar to those of the other tools, especially StringTie.

Since the GeMoSeq-internal CDS prediction based on the longest ORFs within transcripts might be too simplistic and the alternative tools might profit from a more sophisticated CDS prediction, we also tested a CDS prediction using

TransDecoder for these tools. For GeMoSeq, we abstain from a CDS prediction using TransDecoder since its results might contradict the GeMoSeq-internal prediction, which, however, is also being used in the split heuristics of GeMoSeq. We present the results of the evaluation based on the TransDecoder CDS prediction in Supplementary Figs S9, S10, S11, S14, and S15 C/F. Surprisingly, we find that the prediction performance of all tools drops considerably when using TransDecoder instead of the simplistic CDS prediction of GeMoSeq. Here, GeMoSeq yields a substantially improved prediction performance considering the F1 measure, sensitivity and precision, even for *M. musculus*. The reasons for this observation may be manifold. First, we may have used TransDecoder in an incorrect manner. However, we followed the instructions of the TransDecoder manual (cf. Supplementary Methods) and did not spot any obvious flaws. Second, the training data for TransDecoder, as extracted based on predicted transcript models may not have been sufficient to accurately train the TransDecoder model. Third, the CDS annotations of the (high-quality) reference genomes may be biased towards the longest ORFs, since a similar heuristic might have been used when generating these "gold standard" annotations in the first place. Based on the data available, we cannot clarify which of these options (or additional explanations) is correct. However, we consider the third explanation the most likely since, historically, the longest ORF heuristic has been applied widely.

Assessing the influence of split heuristics

The improvement in prediction performance of GeMoSeq is especially pronounced for *S. cerevisiae* with its densely packed genome, of which 72% are transcribed in pre-mRNAs of protein coding genes with intergenic distances of, on average, only a few hundred base pairs (cf. Supplementary Table S2). For this reason, one source of the improvement might be the implementation of several heuristics for splitting candidate transcripts that appear as merged on the level of read mappings. Hence, we additionally evaluate the prediction performance of a GeMoSeq variant with all split heuristics turned off. We compare GeMoSeq with and without split heuristics based on the F1 measure on the level of CDSs (Supplementary Fig. S16). Indeed, we find that the improvement due to the split heuristics is especially pronounced for *S. cerevisiae*. Considerably improved F1 values can also be observed for *A. thaliana*, *C. elegans* and *D. melanogaster*. The genomes of these species can also be considered dense based on median intergenic distances below 1 kbp, large portions of genomic sequence covered by primary transcripts and a high gene density (Supplementary Table S2). For the remaining species, we find a similar performance of both GeMoSeq variants. Hence, we may conclude that the implementation of split heuristics has an important contribution to the prediction performance of GeMoSeq, especially for dense genomes. Since these heuristics are a unique feature of GeMoSeq, it also indicates that split heuristics are partly responsible for the improvement in prediction performance compared with Cufflinks, Scallop and StringTie. However, since we also observed an improvement for *O. sativa* and *S. lycopersicum* in the previous benchmark, other aspects of the GeMoSeq algorithm (e.g., combinatorial enumeration of candidate transcripts, EM-based quantification) contribute as well, but are hard to benchmark individually, because they are deeply integrated into the complete GeMoSeq algorithm.

Assessing determinants of GeMoSeq's performance

In the benchmark studies, we also observe that the range of performance values achieved by GeMoSeq and the alternative tools is library-specific. Hence, we sought for characteristics of sequencing libraries that might be indicative of GeMoSeq's performance. The libraries considered for benchmarking cover a wide range of sequencing depths. However, sequencing depth alone is only moderately correlated to prediction performance (Supplementary Fig. S17). Instead, we found that the number of transcripts (reference annotation) covered by more than 10 reads (Supplementary Fig. S18) and the number of transcripts with a coverage above 5 (Supplementary Fig. S19) show a better correlation with the F1 measure determined for these libraries. So, as might be expected, prediction performance depends on the number of true transcripts that show expression in the sequencing library used as input of GeMoSeq. This cannot be achieved by a single specific library, e.g. specific for a tissue, developmental stage or stress, which, however, might be important for capturing splicing variants. For this reason, we recommend running GeMoSeq for diverse, specific libraries and join prediction results afterwards as exemplified for *N. benthamiana* below.

Artificial reads simulated based on the respective reference genomes and annotations might provide a more controlled environment for evaluating further determinants of GeMoSeq's performance and have also been considered in the original StringTie publication [20]. Following the example of StringTie, we use Flux Simulator [37] to generate artificial reads for the seven genomes considered (cf. Supplementary Methods). However, in the evaluation of prediction performance, we find inconsistent results on the level of transcripts (gffcompare) and CDS (GeMoMa Analyzer) likely due to technical reasons (Supplementary Fig. S20), which render results based on simulated reads less informative when focusing on protein-coding genes.

Runtime and memory requirements

In addition to prediction performance, we also record the total runtime and peak memory consumption of GeMoSeq, Cufflinks, Scallop and StringTie on the benchmark datasets as measured on a compute cluster with nodes equipped with Intel Xeon E5-2680v3 processors and 128 GB of RAM. We observe (Supplementary Fig. S21) that GeMoSeq requires a total runtime typically shorter than Cufflinks but longer than Scallop or StringTie. The median runtime of GeMoSeq is 14 min and 90% of the datasets are processed within 46 min. On a more recent machine (Macbook Air M3, 24 GB of RAM), absolute runtimes for these datasets are reduced to approx. 2 min (median) and 10 min (90% percentile), respectively. For human datasets (Supplementary Fig. S15G), the general pattern is similar with a median runtime of 44 min for GeMoSeq on the compute cluster.

The implementation of the GeMoSeq algorithm is multithreaded. But using a higher number of threads, the (unparallelized) I/O performance becomes the limiting factor. As an alternative, instances of GeMoSeq can be run in parallel for individual chromosomes, and their predictions can be simply concatenated afterwards. To support this option, GeMoSeq allows for a naming schema of genes and transcripts that includes the chromosome/sequence name.

Regarding memory, GeMoSeq has a maximum peak memory consumption that is comparable to Scallop (55 GB vs.

52 GB), but the median memory consumption is substantially higher than for any of the other tools. One reason for the increased memory consumption is the combinatorial enumeration of candidate transcripts with associated read information. However, the median peak memory consumption of GeMoSeq is approx. 16 GB and 90% of the datasets required at most 25 GB, which makes GeMoSeq executable on current standard compute servers. For human datasets (Supplementary Fig. S15H), memory consumption is comparable to the *M. musculus* datasets. If required, memory consumption of GeMoSeq can be reduced by adjusting the parameters for the maximum length of a region before it is split and the maximum region coverage before reads are downsampled.

Benchmark using long-read data

Although databases of raw sequencing data (NCBI SRA, ENA) are currently dominated by short-read, specifically Illumina, data, the amount of long-read RNA-seq data as generated by Pacific Biosciences (PacBio) or Oxford Nanopore Technologies (ONT) instruments is increasing and these will likely become the preferred technologies for the annotation of transcript models in the future. Since long RNA reads are conceptually capable of representing transcripts and transcript variants end-to-end, they are suited to improve the accuracy and completeness of genome annotations. However, the specific error profile of long-read data poses new challenges for algorithms predicting transcript models. For this reason, GeMoSeq provides a specific long-read mode that adapts some algorithmic details (cf. section *Handling of long-read data*). StringTie has a long-read mode as well, while Cufflinks has not been designed for long-read data and a version of Scallop (Scallop-LR [33]) specifically tailored to long-read data requires PacBio-specific header files and has not been updated for five years. However, with IsoQuant [21], a tool specifically tailored to long-read data is available. Hence, we restrict the benchmark based on long-read data to the comparison of IsoQuant, and StringTie and GeMoSeq in their long-read modes. To this end, we sampled long-read data for the same seven species also considered for the short-read data and stratify these for (PacBio and ONT) instruments (cf. section Benchmarking long-read data).

On the level of transcript models evaluated using gffcompare (Supplementary Fig. S22), we find an improved prediction performance of GeMoSeq relative to StringTie for *A. thaliana*, *S. lycopersicum* and *C. elegans*, while for *O. sativa*, *D. melanogaster*, *M. musculus* and *S. cerevisiae*, either GeMoSeq or StringTie yields better performance for a subset of datasets, which is in some cases related to specific instrument models and/or strand specificity. In the comparison to IsoQuant GeMoSeq yields an improvement for *A. thaliana*, *C. elegans*, *M. musculus* and *S. cerevisiae*, while results are inconclusive for the remaining species (Supplementary Fig. S23). This picture is widely preserved on the level of CDS using the longest-ORF heuristic for StringTie (Supplementary Fig. S24) and IsoQuant (Supplementary Fig. S25) predictions. Using TransDecoder for CDS prediction on the transcript models of StringTie and IsoQuant again leads to a substantial drop in the F1 measure (Supplementary Figs S26 and S27).

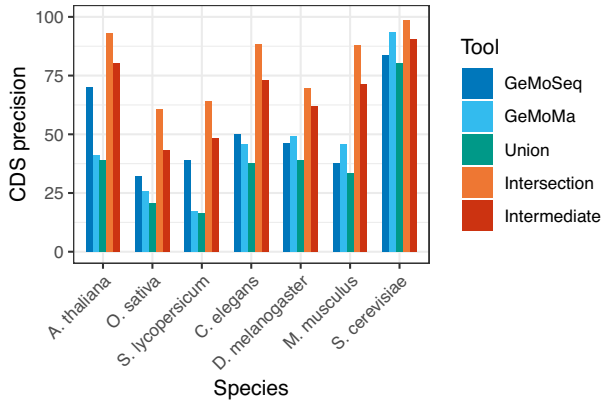


Figure 3. Precision on the level of CDS for GeMoSeq (RNA-seq-based) and GeMoMa (homology-based) predictions, and for combinations of the prediction sets.

Combining RNA-seq-based and homology-based predictions

While the previous benchmarks demonstrated the accuracy of the RNA-seq-based prediction of transcripts and CDSs using GeMoSeq, the original intention for developing GeMoSeq has been to complement the homology-based predictions of GeMoMa, especially with regard to highly species-specific genes and transcripts. Hence, we evaluate the improvements of annotation completeness and accuracy combining GeMoMa and GeMoSeq predictions in the following. To this end, we additionally perform homology-based transcript prediction using GeMoMa for the seven species considered in the benchmark studies (cf. Materials and methods, Reference species listed in Supplementary Table S5). We first merge all RNA-seq-based GeMoSeq predictions based on the datasets used for the benchmark, and we merge all homology-based GeMoMa predictions across the reference species considered. Afterwards, the individually merged GeMoSeq and GeMoMa predictions are further combined by determining the intersection and union of the sets of predictions of both tools per species. We might expect that the intersection of independent sources of evidence (homology and RNA-seq) yields high precision, while the union of both sources would achieve high sensitivity. We additionally consider an "intermediate" variant that supplements the intersection set with further high confidence predictions of the individual tools (cf. Materials and methods).

First, we consider sensitivity on the level of CDSs of the individual prediction sets of GeMoSeq and GeMoMa, and the different combinations of these. In Supplementary Fig. S28, we indeed observe that the union of the predictions of both tools yields the best sensitivity. However, the improvement over the individual predictions of GeMoSeq and GeMoMa is lower than we expected, and the intersection of both sets results in only mildly reduced sensitivity. This indicates that the overlap between the RNA-seq-based predictions of GeMoSeq and the homology-based predictions of GeMoMa is substantial, and may provide a set of high-confidence transcript models. As expected, we find the largest precision values for the intersection of GeMoSeq and GeMoMa predictions, while precision drops substantially for the union of both (Fig. 3). The intermediate variant of the combination yields a considerably improved precision compared with the individual predictions except for *S. cerevisiae*, for which we observe

precision values above 75% in all cases. Notably, we observe precision values above 85% for the intersection for four of the species (*A. thaliana*, *C. elegans*, *M. musculus*, *S. cerevisiae*), while precision is substantially lower for the remaining species. Since the intersection of GeMoSeq and GeMoMa predictions provides evidence of the respective transcript models from two independent sources (RNA-seq and homology), we assume the predictions in the intersection to be likely correct. Hence, this finding may indicate that the current reference annotation of those species (*O. sativa*, *S. lycopersicum*, *D. melanogaster*) is less complete than for the other species. Turning to the F1 measure combining sensitivity and precision (Supplementary Fig. S29), we find that – as intended – the intermediate variant for combing GeMoSeq and GeMoMa predictions yields the best F1 values for four of the species (*A. thaliana*, *C. elegans*, *D. melanogaster*, *M. musculus*). For the remaining species, F1 values are comparable to the intersection case, that, however, achieved substantially lower sensitivity, which might be relevant for practical applications.

Applying GeMoSeq to *N. benthamiana* LAB genomes

Nicotiana benthamiana is used as a model plant in many experimental labs worldwide, especially since it is conveniently infiltrated with *Agrobacterium tumefaciens* for transient expression of genes. Recently, two genomes and corresponding annotation of "lab" strains have been published [35, 36]. At first sight, the transcript annotations of these genomes appear to be rather incomplete with mostly one or only few transcript variants per gene. In addition, both transcript annotations use transcript IDs that are unrelated to the previously used *N. benthamiana* v1.0.1 and v2.6.1 genomes, which complicates the comparisons with previous publications. Hence, we set out to apply the combination of RNA-seq-based GeMoSeq predictions and homology-based GeMoMa predictions to both lab genomes. For GeMoSeq, we compile a collection of 319 *N. benthamiana* RNA-seq datasets from the European Nucleotide Archive (cf. Materials and methods, Supplementary Table S6). For GeMoMa, we consider *A. thaliana*, *S. lycopersicum* and *N. tabacum* as reference species and additionally include the previous v.1.0.1 and v.2.6.1 genome versions to establish the link between this new annotation and the previous ones. We merge the GeMoSeq and GeMoMa predictions as described for the benchmark studies, and assign confidence levels "high," "medium," and "low" as described in Materials and methods. In the merge, transcript IDs of the v1.0.1 and v2.6.1 are available either in the ID field or as entries of the "alternative" tag of the GFF format if these could be transferred to the respective lab genome.

In Fig. 4A, we show an upset plot (inclusive intersection size) of the NbLab360 v103 annotation and the annotation generated using GeMoSeq and GeMoMa on different levels of confidence. The NbLab360 annotation contains 45,796 (protein coding) transcripts, while the high-confidence intersection of GeMoSeq and GeMoMa contains 35,684 transcripts, 21,193 of which are shared (i.e., identical on the level of CDSs) between both annotations. Also considering medium-confidence GeMoSeq/GeMoMa predictions, the number of transcripts increases to 65,892 belonging to 47,307 genes, which corresponds to 1.39 transcript per gene on average. A number of 26,874 of these transcripts is shared with the NbLab360 annotation. Also including low-confidence transcripts, the number of transcripts increases dramatically to 231,219 transcripts. Hence, we recommend to consider low-confidence transcripts only when aiming for maximum sensitivity.

We further scrutinize the 14,491 transcripts that are present in the high-confidence GeMoSeq/GeMoMa set but not in the NbLab360 annotation (Fig. 4B). We find that 4262 of these transcripts do not overlap an existing NbLab360 annotation and may, hence, be considered transcripts of missing genes. If a GeMoSeq/GeMoMa transcript does not match any NbLab360 transcript, but another GeMoSeq/GeMoMa transcript of the same gene perfectly matches an NbLab360 transcript, these GeMoSeq/GeMoMa transcripts may be considered missing alternative transcripts. We find 3,494 of such missing alternatives, which more frequently contain additional exons (2,180 cases) than lack exons (206 cases) present in the matching variant. GeMoSeq/GeMoMa transcripts that overlap but do not perfectly match any NbLab360 transcript may differ in multiple aspects, where alternative transcript start or end positions are the most frequent (6,286 out of 6,735 differing transcripts).

For the transcripts from the "missing gene" category, we aim at obtaining an overview of the functions that are encoded by these genes. To this end, we extract the protein sequences of all high-confidence transcripts and use PANNZER2 [38] to assign functional annotation and GO terms. Among the transcripts from the "missing gene" category, the terms (biological product) "regulation of DNA-templated transcription" (161 transcripts), "defense response" (116), "translation" (103), "protein ubiquitination" (82), "chromatin remodeling" (80), "response to auxin" (67) and "methylation" (55) are the most frequent. Terms with significant enrichment (Fisher"s exact test, adjusted P-value < 0.05) are "killing of cells of another organism" (9 transcripts; log-odds ratio 3.6; $p = 1.3 \cdot 10^{-2}$), "cell population proliferation" (11; 3.5; $p = 1.4 \cdot 10^{-3}$), "hydrotropism" (10; 3.0; $p = 4.3 \cdot 10^{-2}$), "photosynthesis, light harvesting" (17; 2.2; $p = 9.0 \cdot 10^{-3}$), and "response to auxin" (67; 1.9; $p = 1.8 \cdot 10^{-12}$). Together, this indicates that the genes that are only present in our re-annotation of the NbLab genome fulfill important, central functions for *N. benthamiana* plants.

Since no established "gold standard" for *N. benthamiana* transcripts has been established, the quality of transcript models cannot be evaluated in the same manner as for the benchmark species. Hence, we validate example predictions from the "missing gene," "missing alternative," and "differing transcript" categories experimentally and assess the global completeness of the different annotation variants using BUSCO [39, 40] on the level of protein sequences.

Experimental validation

To experimentally validate the quality of our computational predictions, we selected 20 gene models from different prediction categories for PCR-based validation on cDNA from wild-type *N. benthamiana* leaf tissue (Table 1). These selected genes represent the annotation categories "missing gene," "missing alternative," and "differing transcript" compared to the NbLab360 annotation. We successfully obtained PCR products for 16 of the 20 tested transcripts (80% PCR success rate). For 14 genes yielding PCR products, we obtained high-quality Sanger sequencing results through

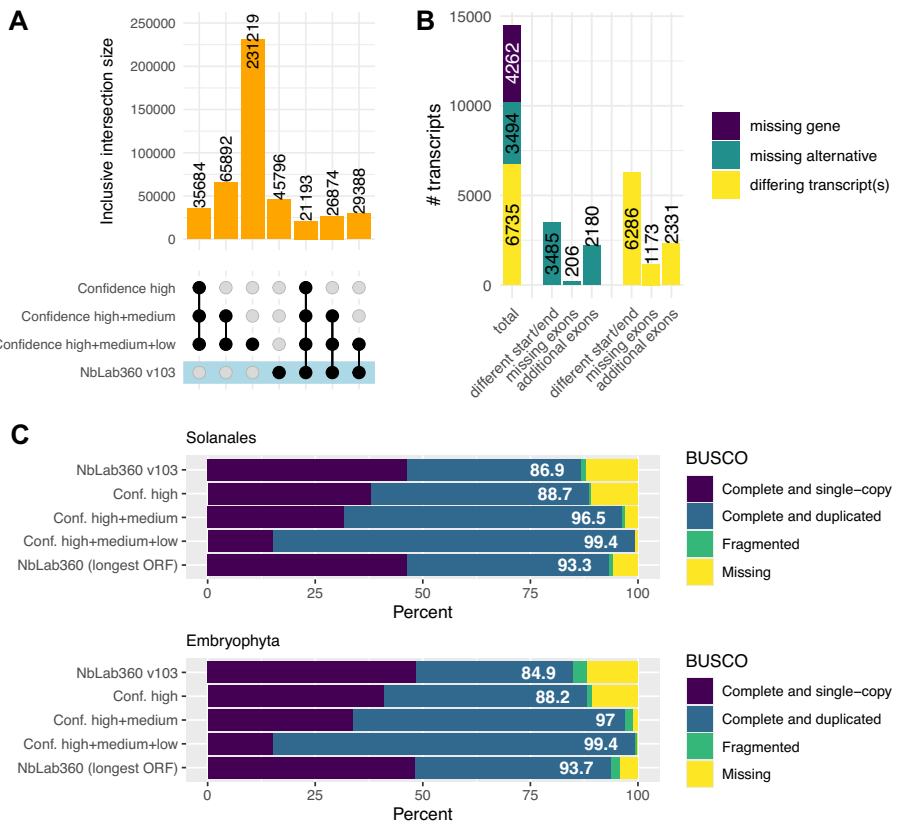


Figure 4. Comparison of the original NbLab360 annotation to the merged prediction of GeMoSeq and GeMoMa, considering transcripts of different levels of confidence. (**A**) Intersection sizes (inclusive, i.e., size of intersection shown irrespective of the occurrence in other sets/intersections) of the CDS contained in each of the annotation files. (**B**) Breakdown of transcripts that are present in the "Confidence high" set of GeMoSeq/GeMoMa but are missing from the NbLab360 annotation. (**C**) BUSCO completeness of the annotation files (extracted proteins) for the solanales and embryophyta datasets.

Table 1. Experimental validation of predicted transcripts in *N. benthamiana*. Twenty predicted transcripts from different categories (cf. Fig. 4B) were tested by PCR amplification of cDNA from wild-type leaf tissue followed by Sanger sequencing. Transcript identifiers have been assigned according to the corresponding homology-based prediction of GeMoMa. PP (PCR Product): successful amplification of expected PCR product; SM (Sequence Match): sequencing result matches computational prediction. ∗) Two distinct isoforms detected; one matched prediction; Categories indicate the relationship between the tested prediction and the NbLab360 v103 annotation: "missing gene" indicates genes absent from the original annotation; "missing alternative" indicates alternative transcript variants not present in the original annotation; "differing transcripts" indicates structural differences in exon-intron structure compared to annotated transcripts

Predicted Transcript	PP	SM	Category	Supp. Fig.
niben101_Niben101Scf04174g05001.1_R0	No	-	missing gene	S33
niben101_Niben101Scf05710g00007.1_R0	Yes	Yes	missing gene	S34
niben261_Niben261Chr11g0918006.1_R0	Yes	Yes	missing gene	S35
niben261_Niben261Chr12g0701036.1_R0	No	-	missing gene	S36
nitab_Nitab4.5_0000928g0110.1_R0	Yes	Yes	missing gene	S37
nitab_Nitab4.5_0002157g0030.1_R1	Yes	Yes	missing gene	S38
nitab_Nitab4.5_0005016g0030.1_R1	No	-	missing gene	S39
nitab_Nitab4.5_0009033g0020.1_R1	Yes	Yes	missing gene	S40
solyc_mRNA:Solyc01g110020.4.1_R0	Yes	Yes	missing gene	S41
niben101_Niben101Scf02053g01004.1_R0	Yes	Yes	missing alternative: acceptor shift	S42
niben101_Niben101Scf00784g00007.1_R0	Yes	Yes	missing alternative: +1 intron	S43
niben261_Niben261Chr06g1105011.1_R0	Yes	Yes	missing alternative: +1 intron	S44
nitab_Nitab4.5_0005550g0060.1_R0	Yes	Yes	missing alternative: +1 intron	S45
solyc_mRNA:Solyc11g005330.2.1_R2	Yes	-	missing alternative: +1 intron	S46
niben101_Niben101Scf13411g05021.1_R0	Yes	Yes	differing transcripts: multiple introns	S47
solyc_mRNA:Solyc06g073460.4.1_R0	No	-	differing transcripts: multiple introns	S48
niben101_Niben101Scf00127g04004.1_R0	Yes	Yes∗	differing transcripts: larger differences	S49
niben101_Niben101Scf04196g00003.1_R0	Yes	Yes	differing transcripts: acceptor shift	S40
nitab_Nitab4.5_0002370g0020.1_R0	Yes	Yes	differing transcripts: 1 exon different	S51
niben261_Niben261Chr11g1317009.1_R0	Yes	Yes	differing transcripts: +1 intron	S52

direct sequencing of purified PCR products. All 14 sequences confirmed the GeMoSeq/GeMoMa predicted models, showing exact matches of exon-intron boundaries and coding sequences (see Supplementary Figs S33--S52).

Four transcripts failed to produce detectable PCR products in initial amplification attempts (Table 1). However, RNA-seq read mappings for all tested transcripts, including these four, indicated sufficient expression levels (Supplementary Figs S33--S52), suggesting that the lack of PCR products is likely due to technical rather than biological factors. Since we obtained sufficient PCR-based validation results across all prediction categories represented in our test set (Table 1), we did not perform extensive optimization of PCR conditions (e.g., annealing temperature, primer design variations, or alternative polymerase systems) for these initial failures. Hence, the absence of PCR products cannot be interpreted as evidence against the existence of these transcripts, and further optimization efforts with alternative primer pairs or adjusted PCR conditions might yield amplification products for these transcripts.

Two PCR-positive transcripts did not yield reliable direct sequencing results. For the first transcript (Solyc11g005330.2), we observed a weak band signal accompanied by smearing on the agarose gel, indicating low specificity of the amplification reaction. Purification and further amplification did not improve PCR product quality and only insufficient product concentration for reliable sequencing were obtained. For the second transcript (niben101_Niben101Scf00127g04004.1_R0), direct Sanger sequencing of purified PCR products yielded superimposed sequencing signals, suggesting the presence of multiple DNA molecules in the PCR reaction. To resolve this ambiguity, we cloned the PCR fragments into the pJet1.2 vector and sequenced individual colonies. Sequencing of six colonies revealed two distinct populations of cDNA fragments differing in the extent of the first intron, indicating two distinct transcript isoforms (Supplementary Fig. S49). Four colonies (67%) contained a splice variant differing from our primary prediction, while two colonies (33%) exactly matched the predicted sequence. Examination of RNA-seq read mappings for this locus confirmed the presence of both splice variants, with split reads supporting both the predicted and alternative acceptor sites of the first intron, consistent with the coexistence of both isoforms in the transcriptome (Supplementary Fig. S53).

In total, we obtained experimental confirmation for 15 out of 16 genes with interpretable sequencing results, resulting in a 94% confirmation rate. All confirmed sequences showed exact matches of exon-intron boundaries and coding sequences to GeMoSeq/GeMoMa predictions. Although this validation represents a limited sample size, the high confirmation rate provides strong experimental support for the accuracy of the combined RNA-seq/homology-based gene prediction approach. The precise matching of splice sites and coding sequences at the nucleotide level demonstrates that the combined GeMoSeq/GeMoMa predictions are not only structurally correct but also immediately suitable for functional studies and protein-level analyses.

BUSCO completeness

For BUSCO completeness, we consider only CDS annotations with proper start and stop codons and without premature stops. In Fig. 4C, we present the BUSCO completeness of the original NbLab360 v103 annotation and GeMoSeq/GeMoMa-derived annotations on different levels of confidence using the "Solanales" and "Embroyphyta" BUSCO sets. We observe that even the high-confidence predictions of GeMoSeq/GeMoMa achieve a better BUSCO completeness (single-copy and duplicated) than the original NbLab360 annotation, while completeness further increases when also considering medium and low-confidence transcripts. In previous benchmark studies, we noted that the longest-ORF heuristic of GeMoSeq typically results in better prediction performance than using TransDecoder. Hence, we additionally apply this heuristic to the NbLab360 v103 annotation, which may also repair previously missing start and stop codons. Indeed, we find that this increases BUSCO completeness, although completeness is still lower than for the combination of high-confidence and medium-confidence GeMoSeq/GeMoMa predictions.

We also observe that the number of duplicated BUSCOs for GeMoSeq/GeMoMa predictions is substantially larger than for the original NbLab360 v103 annotation, which contains one transcript variant per gene. Hence, we aggregate the BUSCO results on the level of genes using the "BUSCORecomputer" module of GeMoMa. We find (Supplementary Fig. S31 and S32) that this indeed decreases the number of duplicated BUSCOs for the GeMoSeq/GeMoMa predictions to similar levels as for the NbLab360 v103 annotation, except for the low-confidence set.

We apply the same procedure to the NbKLAB genome and corresponding annotation (Supplementary Fig. S30) with similar results, but a considerably larger number of missing genes (6,700) and slightly lower BUSCO completeness of the NbKLAB annotation. One notable difference is that applying the longest-ORF heuristic to the NbKLAB annotation does not result in different BUSCO completeness, likely because a similar heuristic has already been applied for CDS prediction initially and the number of missing start/stop codons is already substantially lower for the original NbKLAB annotation than for NbLab360.

Conclusion

In benchmark studies, we found that GeMoSeq yields an improved prediction performance compared with the previous approaches Cufflinks, Scallop and StringTie for six out of seven species considered. This improvement is especially pronounced for the two plant species *A. thaliana* and *S. lycopersicum* and for yeast (*S. cerevisiae*), which have rather dense genomes. We further demonstrated how the combination of RNA-seq-based predictions of GeMoSeq with homology-based predictions using GeMoMa may be used to balance sensitivity and precision in the annotation of transcript models. Applying the combination of GeMoSeq and GeMoMa to recently sequenced genomes of two *N. benthamiana* lab strains, we illustrated the utility of this strategy to yield high-quality annotations of transcript models and coding sequences for newly sequenced genomes. We provide the GeMoSeq tool and the generated annotations of *N. benthamiana* lab strains to the scientific community.

Acknowledgements

We thank Lilya Kopertekh for valuable discussions. We thank Theresa Staps for extracting *N. benthamiana* RNA. We ac-

knowledge the financial support of the Open Access Publication Fund of the Martin Luther University Halle-Wittenberg (https://www.uni-halle.de).

Author contributions: J.G. (Conceptualization [equal] Investigation [equal] Methodology [lead] Software [lead] Visualization [equal] Writing – original draft [lead] Writing – review & editing [equal]), D.W. (Investigation [supporting] Methodology [supporting] Software [supporting] Writing – review & editing [equal]), M.P. (Investigation [supporting] Validation [equal] Writing – review & editing [equal]), M.H.S. (Conceptualization [supporting] Investigation [supporting] Validation [equal] Visualization [supporting] Writing – review & editing [equal]), J.K. (Conceptualization [equal] Investigation [equal] Methodology [supporting] Software [supporting] Visualization [equal] Writing – review & editing [equal])

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

Institutional funding of Martin Luther University Halle-Wittenberg and Julius Kühn-Institute. Funding to pay the Open Access publication charges for this article was provided by the Open Access Publication Fund of the Martin Luther University Halle-Wittenberg (https://www.uni-halle.de).

Data availability

The source code of GeMoSeq is available from GitHub at https://github.com/Jstacs/Jstacs/tree/master/projects/gemoseq. The code version used when generating the results of this manuscript is available from Zenodo at https://doi.org/10.5281/zenodo.17716915. A binary version of GeMoSeq is available from https://www.jstacs.de/index.php/GeMoSeq. The annotation files for the *N. benthamiana* lab strains are available from Zenodo at https://doi.org/10.5281/zenodo.14901380.

References

1. Guigó R, Knudsen S, Drake N *et al.* Prediction of gene structure. *J Mol Biol* 1992;226:141–57. https://doi.org/10.1016/0022-2836(92)90130-C
2. Snyder EE, Stormo GD. Identification of protein coding regions in genomic DNA. *J Mol Biol* 1995;248:1–18. https://doi.org/10.1006/jmbi.1995.0198
3. Golicz AA, Batley J, Edwards D. Towards plant pangenomics. *Plant Biotechnol J* 2016;14:1099–105. https://doi.org/10.1111/pbi.12499
4. Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. *Nat Rev Genet* 2020;21:428–44. https://doi.org/10.1038/s41576-020-0233-0
5. Langschied F, Bordin N, Cosentino S *et al.* Quest for orthologs in the era of biodiversity genomics. *Genome Biol Evol* 2024;16:evae224. https://doi.org/10.1093/gbe/evae224
6. Van den Berge K, Hembach KM, Soneson C *et al.* RNA sequencing data: Hitchhiker"s guide to expression analysis. *Annu Rev Biomed Data Sci* 2019;2:139–73. https://doi.org/10.1146/annurev-biodatasci-072018-021255
7. Guigó R. Genome annotation: from human genetics to biodiversity genomics. *Cell Genom* 2023;3:100375.
8. Mudge JM, Harrow J. The state of play in higher eukaryote gene annotation. *Nat Rev Genet* 2016;17:758–72. https://doi.org/10.1038/nrg.2016.119
9. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 2003;19:ii215–25. https://doi.org/10.1093/bioinformatics/btg1080
10. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 2005;33:W465–67. https://doi.org/10.1093/nar/gki458
11. Keller O, Kollmar M, Stanke M *et al.* A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* 2011;27:757–63. https://doi.org/10.1093/bioinformatics/btr010
12. Gabriel L, Becker F, Hoff KJ *et al.* Tiberius: end-to-end deep learning with an HMM for gene prediction. *Bioinformatics* 2024;40:btae685. https://doi.org/10.1093/bioinformatics/btae685
13. Li H. Protein-to-genome alignment with miniprot. *Bioinformatics* 2023;39:btad014. https://doi.org/10.1093/bioinformatics/btad014
14. Gotoh O. Spaln3: improvement in speed and accuracy of genome mapping and spliced alignment of protein query sequences. *Bioinformatics* 2024;40:btae517. https://doi.org/10.1093/bioinformatics/btae517
15. Chatterji S, Pachter L. Reference based annotation with GeneMapper. *Genome Biol* 2006;7:R29. https://doi.org/10.1186/gb-2006-7-4-r29
16. Keilwagen J, Wenk M, Erickson JL *et al.* Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res* 2016;44:e89. https://doi.org/10.1093/nar/gkw092
17. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. *Genome Res* 2004;14:988–95. https://doi.org/10.1101/gr.1865504
18. Trapnell C, Williams BA, Pertea G *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;28:511–5. https://doi.org/10.1038/nbt.1621
19. Shao M, Kingsford C. Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat Biotechnol* 2017;35:1167–9. https://doi.org/10.1038/nbt.4020
20. Pertea M, Pertea GM, Antonescu CM *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;33:290. https://doi.org/10.1038/nbt.3122
21. Prjibelski AD, Mikheenko A, Joglekar A *et al.* Accurate isoform discovery with IsoQuant using long reads. *Nat Biotechnol* 2023;41:915–8. https://doi.org/10.1038/s41587-022-01565-y
22. Holt C, Yandell M. Maker2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 2011;12:491. https://doi.org/10.1186/1471-2105-12-491
23. Gabriel L, Brůna T, Katharina J *et al.* BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP AUGUSTUS and TSEBRA. *Genome Res* 2024;34:769–77. https://doi.org/10.1101/gr.278090.123
24. Stanke M, Schöffmann O, Morgenstern B *et al.* Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 2006;7:62. https://doi.org/10.1186/1471-2105-7-62
25. Keilwagen J, Hartung F, Paulini M *et al.* Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics* 2018;19:189. https://doi.org/10.1186/s12859-018-2203-5
26. Keilwagen J, Hartung F, Grau J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol Biol* 2019;1962:161–77.

27. Liang C, Mao L, Ware D *et al.* Evidence-based gene predictions in plant genomes. *Genome Res* 2009;19:1912–23. https://doi.org/10.1101/gr.088997.108
28. Bray NL Pimentel H, Melsted P *et al.* Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;34:525–7. https://doi.org/10.1038/nbt.3519
29. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2013;30:923–30. https://doi.org/10.1093/bioinformatics/btt656
30. Kovaka S, Zimin AV, Pertea GM *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* 2019;20:278. https://doi.org/10.1186/s13059-019-1910-1
31. Shumate A, Wong B, Pertea G *et al.* Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLOS Comput Biol* 2022;18:1–18. https://doi.org/10.1371/journal.pcbi.1009730
32. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. *F1000Research* 2020;9:304. https://doi.org/10.12688/f1000research.23297.1
33. Tung LH, Shao M, Kingsford C. Quantifying the benefit offered by transcript assembly with Scallop-LR on single-molecule long reads. *Genome Biol* 2019;20:287. https://doi.org/10.1186/s13059-019-1883-0
34. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–100. https://doi.org/10.1093/bioinformatics/bty191
35. Ranawaka B, An J, Lorenc MT *et al.* A multi-omic Nicotiana benthamiana resource for fundamental research and biotechnology. *Nat Plants* 2023;9:1558–71. https://doi.org/10.1038/s41477-023-01489-8
36. Ko SR, Lee S, Koo H *et al.* High-quality chromosome-level genome assembly of Nicotiana benthamiana. *Sci Data* 2024;11:386. https://doi.org/10.1038/s41597-024-03232-0
37. Griebel T, Zacher B, Ribeca P *et al.* Modelling and simulating generic RNA-seq experiments with the Flux Simulator. *Nucleic Acids Res* 2012;40:10073–83. https://doi.org/10.1093/nar/gks666
38. Törönen P, Medlar A, Holm L. Pannzer2: a rapid functional annotation web server. *Nucleic Acids Res* 2018;46:W84–8. https://doi.org/10.1093/nar/gky350
39. Manni M, Berkeley MR, Seppey M *et al.* BUSCO: assessing genomic data quality and beyond. *Curr Protoc* 2021;1:e323. https://doi.org/10.1002/cpz1.323
40. Manni M, Berkeley MR, Seppey M *et al.* BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* 2021;38:4647–54. https://doi.org/10.1093/molbev/msab199

Received: April 9, 2025. **Revised:** January 13, 2026. **Accepted:** January 20, 2026