Supplementary Materials

GeneMark-ETP: Automatic Gene Finding in Eukaryotic Genomes in Consistence with Extrinsic Data

Tomas Bruna [1], Alexandre Lomsadze [2] and Mark Borodovsky [1,2,3,]

1 School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA

2 Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

3 School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Supplementary Methods

S1. Additional details on the processing of the GeneMarkS-T predictions

S1.1 Prediction of 5' complete and 5' partial genes.

To improve accuracy of gene predictions by GeneMarkS-T we used the following inequality:

$$(b-a)-(a-1)+1000*\ln\frac{AAI_{partial}}{AAI_{complete}}>0 \qquad \text{(S1)}$$

Here (a) and (b) are positions of the starts of the local alignments within the target protein when aligned against the longer and shorter protein queries, recalculated to longer query. $AAI_{partial}$ and $AAI_{complete}$ are, respectively, the percentages of amino acid identities in the alignments of the longer and shorter proteins to the target protein (see Fig. S9 and S10).

Inequality (S1) is used to discriminate between two possibilities: a correctly predicted incomplete protein vs a true complete protein incorrectly extended into incomplete one. If the extension is not supported by a protein alignment, then the complete protein option is selected. To characterize the alignment of a protein query and a protein target we selected the following features:

I) The length of the query sequence upstream to the start of the local alignment of the initial long query and target (fragment "a-1" in Fig. S9).
II) The difference between starts of the two alignments with the target protein ("b-a" in Fig. S10).
III) Ratio of the AAI of an alignment within the range "a-c" to the AAI of alignment within the range "b-c"

A large length “a-1” would indicate possible translation of a part of the 5’ UTR region of the query gene. A small length “b-a” (Fig. S9) is interpreted as alignment of a fragment of translation of 5’ UTR of a query gene. Increase of the AAI ratio favors the 5’ partial candidate.

The first two features are measured in numbers of AA and the third one is dimensionless. We scale the third feature by using logarithm with a factor 1,000, i.e. 1,000*log(). The larger value of the ratio would correspond to more conserved sequence alignment in the range “b-a”.

S1.2 Removal of the 3’ partial predictions

The 3' partial predictions were rarely observed; they usually originated from a gene prediction error rather than from a 3’ partial assembly. This frequency pattern could be expected since RNA-Seq libraries, prepared with the poly-A tail enrichment of mRNA transcripts, should predominantly carry transcripts complete at 3' ends (Zhao et al. 2014). This consideration justifies the removal of all the 3' partial genes from the list of candidates for high-confidence genes.

S1.3. Adjustment of gene predictions being shorter than the longest ORF

Most eukaryotic genes are translated from the start codon closest to the transcript 5' end (Kozak 1999). Still, the translation can be initiated at one of the downstream starts; e.g., when the most upstream start has a weak translation initiation signal (the Kozak pattern (Kozak 1987)). GeneMarkS-T accounts for the possibility of non-5'-most translation start codons by predicting the translation start based on the strength of its Kozak pattern (derived in species-specific self-training). However, because the Kozak pattern is relatively weak, the GeneMarkS-T non-5'-most start codon predictions exhibit a higher false-positive rate than the 5'-most start codon predictions (the bias in reference annotations towards the 5'-most translation initiation sites cannot be excluded as well). To improve gene start prediction, GeneMark-ETP uses extrinsic evidence. If the translation of a predicted gene is extended to the 5’-most start codon, and this translation is supported by external protein evidence, GeneMark-ETP extends the predicted gene with non-5'-most start to the longest open reading frame. Obviously, the remaining non-5'-most start gene predictions, that may appear among the high-confidence genes, become more reliable than the genes predicted by GeneMarkS-T in the original set.

S1.4 Complete genes with full protein support

A gene predicted by GeneMarkS-T is said to have full protein support if there is a protein in a database whose significant BLASTp alignment to the predicted protein satisfies condition (S2)

$$(|Q_{start} - T_{start}| \leq 5) \wedge (|(Q_{len} - Q_{end}) - (T_{len} - T_{end})| \leq 20) \qquad \text{(S2)}$$

The terms used in (S2) are determined by features of an alignment of the query (predicted protein) and target, a protein found by the DIAMOND similarity search (Fig. S11). Here, Q_{start}, Q_{end} , respectively, are the positions of the start and end of the alignment within the query protein; T_{start}, T_{end} , respectively, are the positions of the start and end of the alignment within the target protein; Q_{len}, T_{len} , respectively, are the lengths of the query protein and the aligned target.

Experiments with multiple sequence alignment (MSA) of orthologous proteins demonstrated that internal section of MSA is usually the most conserved, while the protein N-proximal region is less conserved and the least conserved region in MSA is usually C-proximal region. Therefore, condition (S2) allows misalignments both at the start of the query to target alignment and at the end of the alignment, even to a larger degree. A query protein whose alignment to at least one target out of the 25 best satisfies condition (S2) is classified as fully supported by the target.

S1.5 Assessment of accuracy of the transcript classification rules

To assess the accuracy of classification made with help of Inequality (1) and Condition (2), we used the following approach. First, we prepared test sets of complete and partial genes. The ground-truth labels were determined by comparisons with reference annotations; the set of training data contained GeneMarkS-T gene predictions in transcripts from each genome (Table 3). Next, we selected features for computations of scores used in Inequality (1) and Condition (2). To prove that (1) and (2) produce efficient classification, we used two approaches. We trained random forest and logistic regression classifiers (with Python's scikit-learn machine learning library) — using all alignment features offered by DIAMOND's tabular output (Buchfink et al. 2015) — to classify predictions as complete/partial by using (1), or true/false by using (2). We observed that use of Inequality (1) and Condition (2) for classification of GeneMarkS-T predictions in the test set (not overlapping with the training set) produced more accurate results than ones generated with application of general-purpose random forest or logistic regression models.

S2. Analysis of the ProtHint support for additional candidates for high-confidence genes

GeneMarkS-T gene predictions not fully supported by proteins could be classified as high-confidence genes (Section 2.2.3). Such predictions should satisfy several conditions, one of which is no contradiction to the ProtHint hints. To give more details, let consider the prediction process step-by-step. First, a gene predicted in a transcript by GeneMarkS-T is mapped to genomic DNA. Next, ProtHint uses the gene mapped to DNA as the gene seed, and ProtHint hints are generated as described in GeneMark-EP+ (Bruna et al. 2020). At this point, elements of the transcript-mapped exon-intron structure (introns, start and stop codons) are compared to the ProtHint hints. The conflict exists if (i) at least one of ProtHint's introns overlaps an exon, or (ii) a ProtHint defined stop codon overlaps an exon or intron, or (iii) a ProtHint start codon overlaps an exon or intron (except the start-to-start overlap).

S3. MAKER2 vs GeneMark-ETP experiments

Three model organisms representing three different types of genome organization were selected for MAKER2 and GeneMark-ETP experiments:

- *Drosophila melanogaster* – small GC homogeneous genome.
- *Danio rerio* – large GC homogenous genome
- *Mus musculus* – large GC heterogeneous genome

The same input information was provided to MAKER2 and GeneMark-ETP.

Repeat coordinates predicted by RepeatMasker software were reformatted to MAKER2 supported GFF format as:

```
rmasker_out2maker_gff.pl < genome.fasta.out > repeatmasker.gff
```

Transcripts assembled in GeneMark-ETP runs from RNA-Seq by HISAT2/StringTie2 were provided as transcriptome input to MAKER2.

Proteins from the following species in OrthoDB were used as input to MAKER2 and GeneMark-ETP:

For *Drosophila melanogaster* 274,283 proteins from:

Drosophila ananassae
Drosophila biarmipes
Drosophila bipectinate
Drosophila busckii
Drosophila elegans
Drosophila erecta
Drosophila eugracilis
Drosophila ficusphila
Drosophila grimshawi
Drosophila hydei
Drosophila mojavensis
Drosophila obscura
Drosophila pseudoobscura
Drosophila rhopaloa
Drosophila serrata
Drosophila takahashii
Drosophila virilis
Drosophila willistoni
Drosophila yakuba

For *Danio rerio* 181,842 proteins from:

Cyprinus carpio
Sinocyclocheilus anshuiensis

Sinocyclocheilus 5ahari
Sinocyclocheilus rhinocerous

For *Mus musculus* 207,553 proteins from:

Cavia porcellus
Cricetulus griseus
Fukomys damarensis
Ictidomys tridecemlineatus
Marmota marmota marmota
Mesocricetus auratus
Mus caroli
Mus 5ahari
Octodon degus
Rattus norvegicus

MAKER2 was executed with three gene finders: AUGUSTUS, GeneMark.hmm and SNAP.
The following model file were used by gene finders:

For prediction in *Drosophila melanogaster*:

AUGUSTUS – "fly" from AUGUSTUS distribution.
GeneMark.hmm – model created by GeneMark-ETP.
SNAP – "D.melanogaster.hmm" from SNAP distribution.

For prediction in *Danio rerio*:

AUGUSTUS – "zebrafish" from AUGUSTUS distribution.
GeneMark.hmm – model created by GeneMark-ETP.
SNAP – model trained according to instructions from SNAP distribution. The training set matches the test set used for evaluation of MAKER2 performance. All the other training steps were done using scripts from SNAP distribution.

For prediction in *Mus musculus*:

AUGUSTUS – "human" from AUGUSTUS distribution.
GeneMark.hmm – model created by GeneMark-ETP on mouse genome for medium GC bin.
SNAP – "mam46.hmm" mammalian model for medium GC bin from SNAP distribution.

MAKER2 was executed with the following setting in the MAKER2 configuration file:

```
genome=genome.fasta
est=transcriptome.fasta
protein=proteindb.fasta
model_org=   #empty
rm_gff=repeatmasker.gff
snaphmm=snap.model
gmhmm=genemark.mod
```

```
augustus_species=model_name
est2genome=1
protein2genome=1
alt_splice=1
always_complete=1
keep_preds=1 for D. melanogaster
keep_preds=0 for D. rerio and M. musculus
split_hit=20000
max_dna_len=1000000
```

MAKER2 was executed on Azure cloud LINUX node with 96 cores in MPI mode.

Accuracy of MAKER2 and GeneMark-ETP runs was estimated as it is described in the main text of this paper. Accuracy on exon, transcript and gene levels is shown in Tables S11.

Supplementary Figures

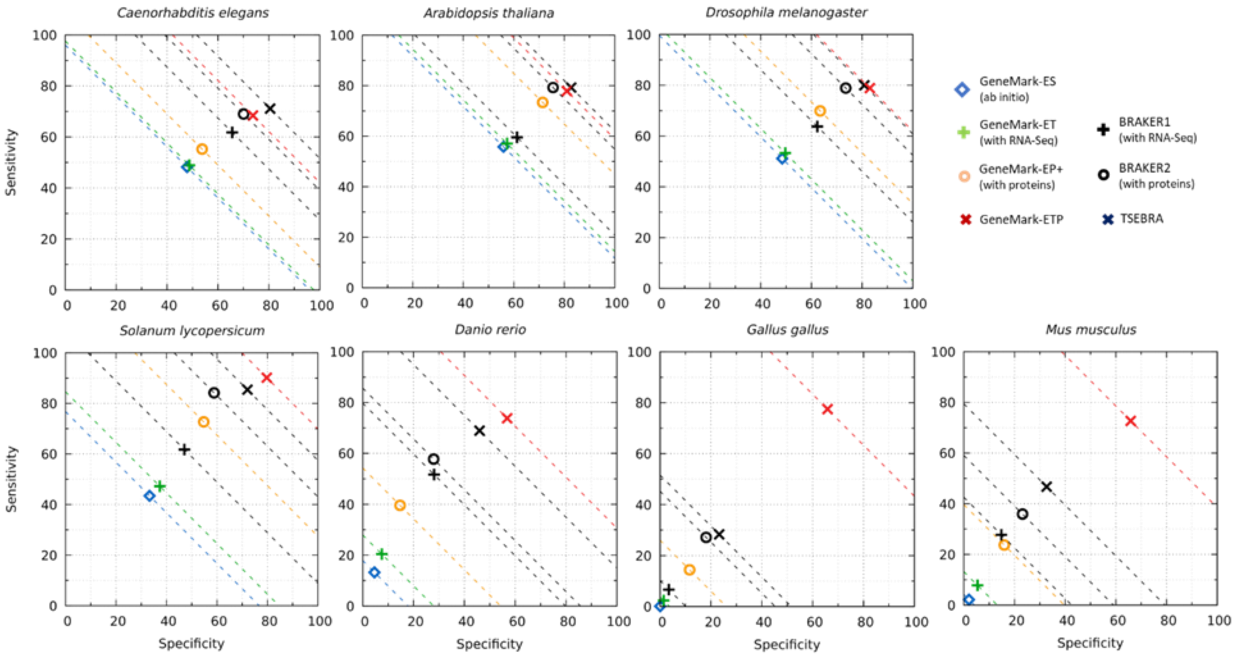

Figure S1. Gene level accuracy of the seven gene prediction tools (see legends to Figs. 1, 2). Compared to the figures in the main text, where we used smaller size reference protein databases for each species (all proteins of the same taxonomic order were excluded from the corresponding IP_0 databases), here we used larger size databases (proteins from the same species excluded from the corresponding IP_0 databases).

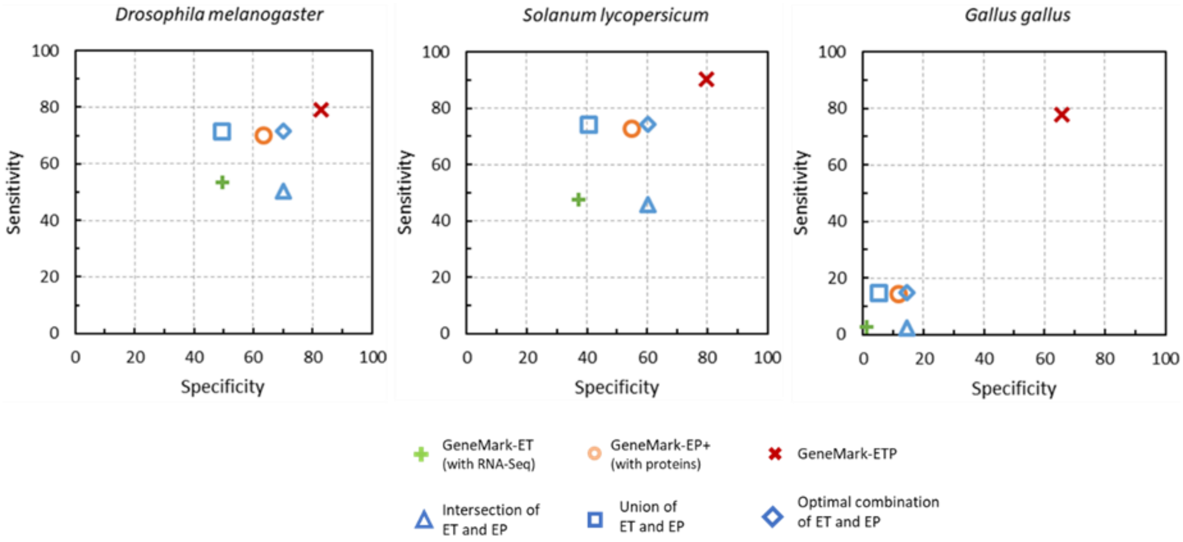

Figure S2. Gene-level accuracy of the virtual optimal combinations of GeneMark-ET and GeneMark-EP+ (Fig. 3) along with the prediction accuracy of GeneMark-ETP. The results for *D. melanogaster* were obtained for a large reference database (only proteins of the same species were excluded from the corresponding IP_0 database); for the other two genomes, proteins of the same taxonomic order were removed from the corresponding IP_0 database.

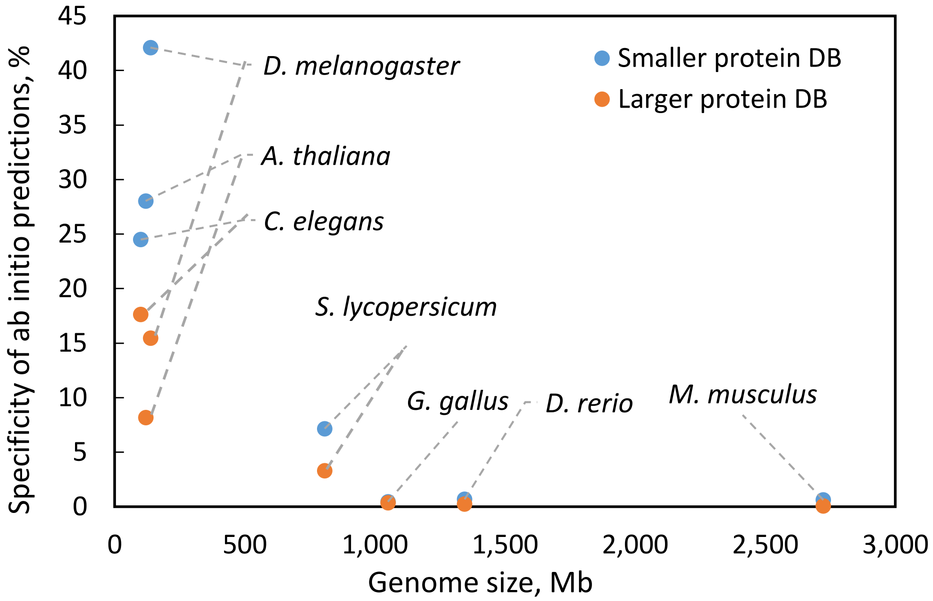

Figure S3. *Ab initio* gene predictions were divided into two categories (see Table 2). Those that could be at least partly supported by available extrinsic evidence in an *a posteriori* analysis and those that could not receive support by any extrinsic evidence at any stage of the analysis (unverified). The figure shows the dependence of the Specificity of unverified *ab initio* gene predictions on the size of the genome, as observed in our experiments.

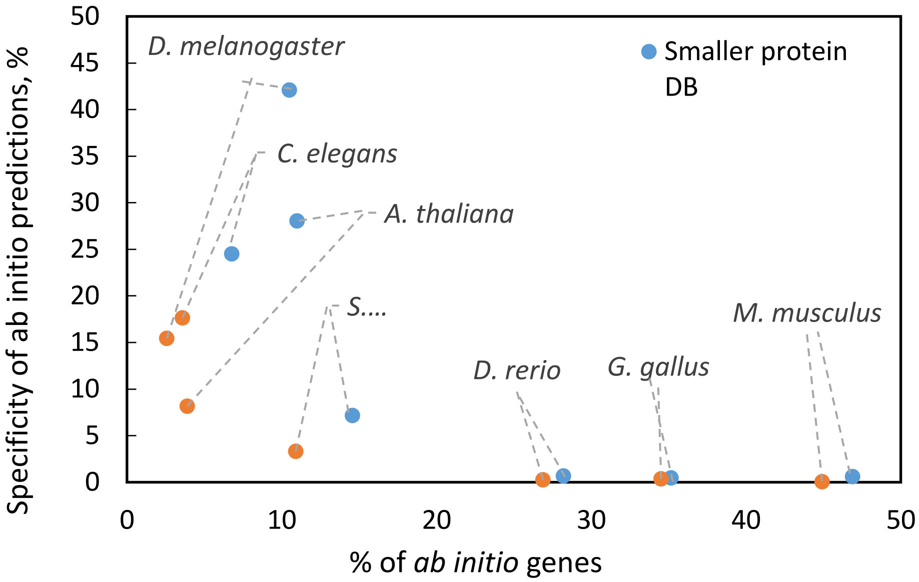

Figure S4. The Y-axis shows the same variable as the one described in Fig. S7. The X-axis shows a fraction (%) of *ab initio* predicted unverified genes among the whole set of genes predicted in each genome.

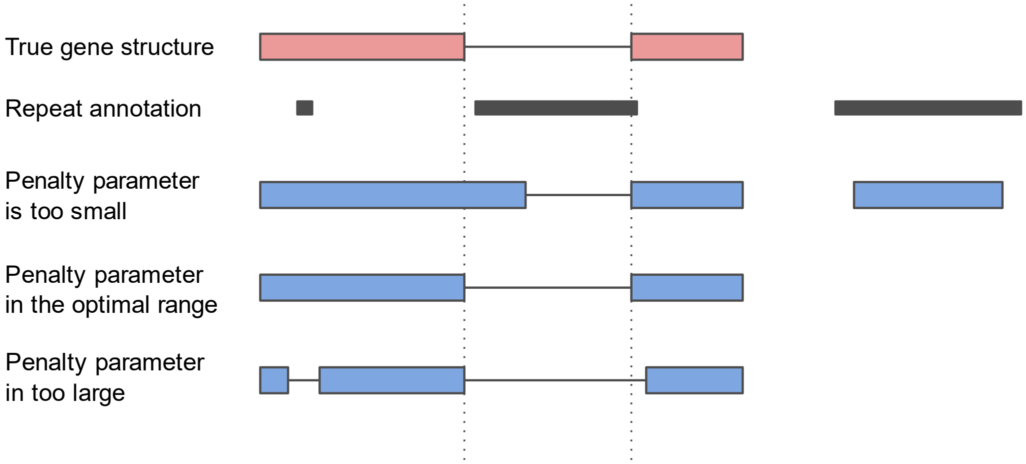

Figure S5. Possible aberrations in gene prediction caused by inadequate selection of the repeat penalty parameter.

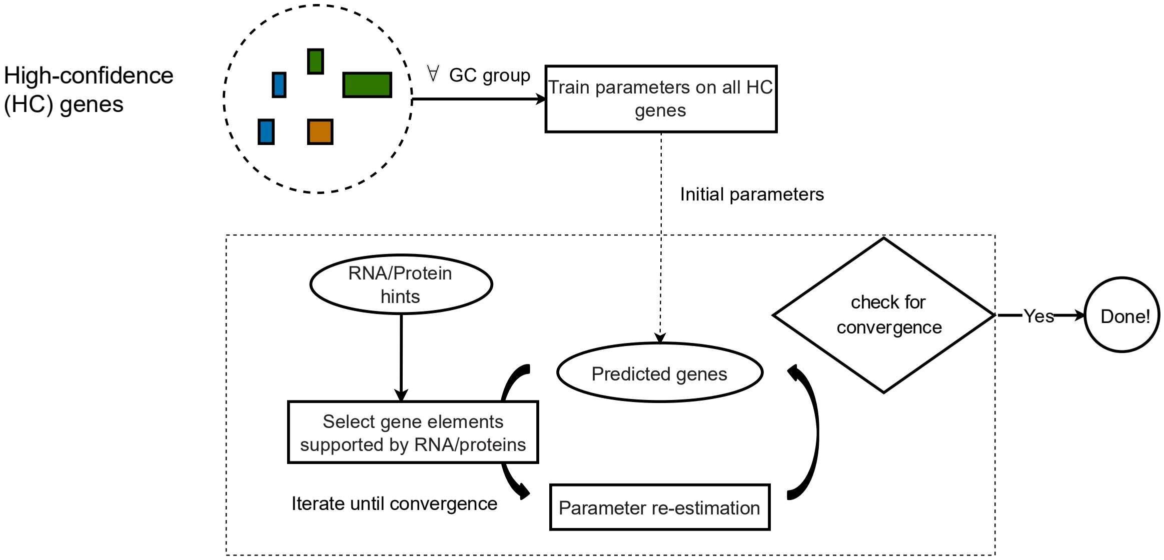
High-confidence (HC) genes
∀ GC group
Train parameters on all HC genes
Initial parameters
RNA/Protein hints
check for convergence
Yes
Done!
Predicted genes
Select gene elements supported by RNA/proteins
Iterate until convergence
Parameter re-estimation

Figure S6. Workflow of the GHMM model training procedure for the GeneMark.hmm algorithm in GeneMark-ETP.

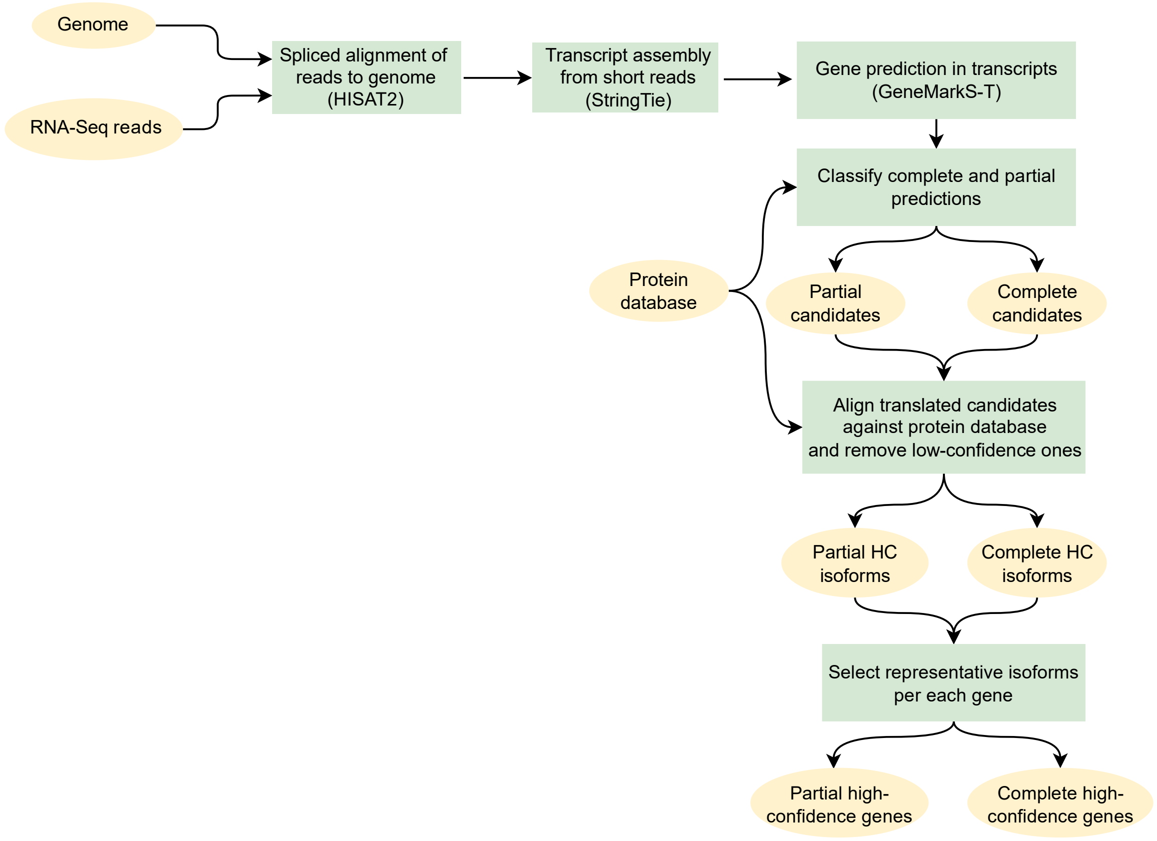
Genome
RNA-Seq reads
Spliced alignment of reads to genome (HISAT2)
Transcript assembly from short reads (StringTie)
Gene prediction in transcripts (GeneMarkS-T)
Classify complete and partial predictions
Protein database
Partial candidates
Complete candidates
Align translated candidates against protein database and remove low-confidence ones
Partial HC isoforms
Complete HC isoforms
Select representative isoforms per each gene
Partial high-confidence genes
Complete high-confidence genes

Figure S7. High-level schematics of the procedure of the identification of high-confidence (HC) genes and selection of representative HC isoforms.

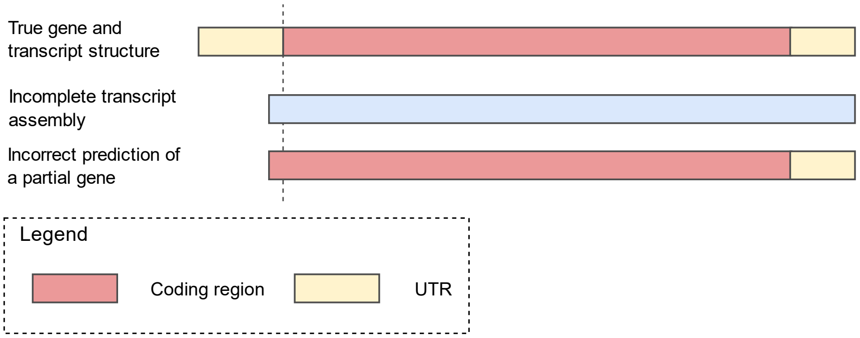
True gene and transcript structure
Incomplete transcript assembly
Incorrect prediction of a partial gene
Legend
Coding region
UTR

Figure S8. An example of an incorrect prediction of a partial gene. The assembly contains a complete coding region and a part of true 5’ UTR. The coding region predicted in transcript was incorrectly extended to the 5’ end of the transcript due to the shortened 5’ UTR.

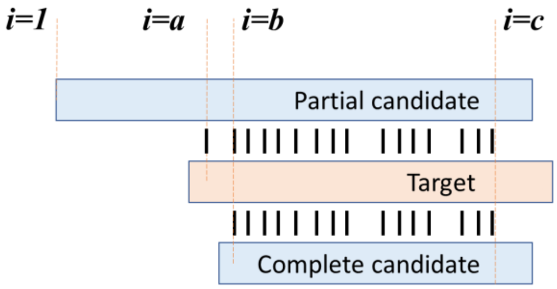

Figure S9. The GeneMarkS-T gene prediction could be classified as complete gene. (a) and (b) are positions of the starts of the local alignments of respective longer and shorter protein queries. (c) is the end position of the local alignments.

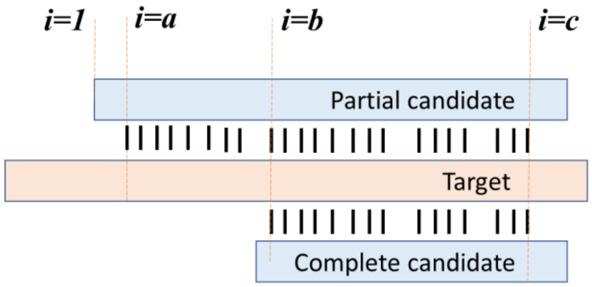

Figure S10. The GeneMarkS-T gene prediction could be classified as a 5' partial gene. (a) and (b) are positions of the starts of the local alignments of respective longer and shorter protein queries. (c) is the end position of the local alignments.

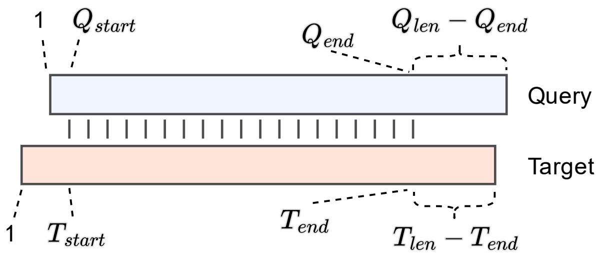

Figure S11. The alignment features used to select complete high-confidence genes based on protein support.

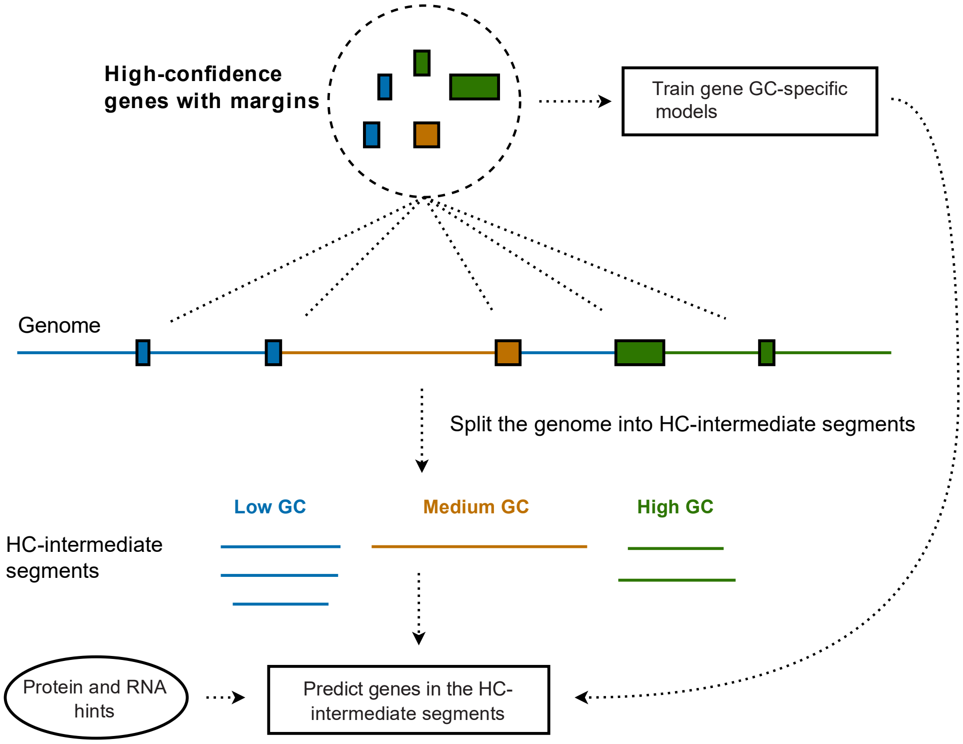
High-confidence genes with margins
Train gene GC-specific models
Genome
Split the genome into HC-intermediate segments
Low GC
Medium GC
High GC
HC-intermediate segments
Protein and RNA hints
Predict genes in the HC-intermediate segments

Figure S12. Schematics of the identification and use of the HC-intermediate regions in training and gene prediction

Supplementary Tables

Table S1. Gene- and exon-level prediction accuracy of the *ab initio* GeneMark-ES, the RNA-Seq-based GeneMark-ET, the protein-based GeneMark-EP+, and GeneMark-ETP. The accuracy estimates are shown for the smaller (order excluded) and for the larger (species excluded) protein databases (see Materials).

		ES	ET	Smaller protein DB		Larger protein DB	
				EP+	ETP	EP+	ETP
C. elegans	Gene Sn	48.2	48.9	48.5	60.4	55.2	68.4
	Gene Sp	47.9	48.8	46.8	67.7	53.8	73.8
	Gene F1	48.0	48.8	47.6	63.8	54.5	71.0
	Exon Sn	81.8	81.7	81.1	82.9	83.3	85.9
	Exon Sp	83.1	83.7	82.0	90.1	84.9	91.4
	Exon F1	82.5	82.7	81.5	86.4	84.1	88.6
A. thaliana	Gene Sn	55.8	57.1	66.6	75.8	73.4	77.9
	Gene Sp	55.9	57.3	65.9	80.0	71.5	81.0
	Gene F1	55.9	57.2	66.3	77.8	72.4	79.4
	Exon Sn	76.9	77.1	79.8	82.3	81.5	82.9
	Exon Sp	80.8	82.1	84.9	90.9	86.3	91.0
	Exon F1	78.8	79.5	82.3	86.4	83.8	86.8
D. melanogaster	Gene Sn	51.2	53.3	56.5	71.5	69.9	78.9
	Gene Sp	48.5	49.7	53.9	77.9	63.5	83.1
	Gene F1	49.8	51.4	55.1	74.6	66.5	80.9
	Exon Sn	67.8	68.6	70.2	76.4	76.5	80.7
	Exon Sp	72.8	74.2	77.3	89.7	81.1	91.4
	Exon F1	70.2	71.3	73.6	82.5	78.8	85.7
S. lycopersicum	Gene Sn	43.4	47.2	67.0	88.2	72.7	90.2
	Gene Sp	33.3	37.4	51.3	81.4	54.8	79.8
	Gene F1	37.7	41.7	58.1	84.7	62.5	84.6
	Exon Sn	82.6	83.5	90.5	96.7	92.1	97.2
	Exon Sp	69.4	74.2	80.0	92.6	80.7	91.6
	Exon F1	75.5	78.6	84.9	94.6	86.0	94.3
D. rerio	Gene Sn	13.2	20.4	35.7	72.7	39.6	73.8
	Gene Sp	4.6	7.5	13.3	56.5	14.7	56.8
	Gene F1	6.9	11.0	19.4	63.6	21.4	64.2
	Exon Sn	75.3	79.1	84.9	93.6	86.2	94.0
	Exon Sp	40.8	50.3	55.9	85.1	56.5	85.1
	Exon F1	52.9	61.5	67.4	89.2	68.2	89.3
G. gallus	Gene Sn	0.1	2.4	14.1	78.0	14.4	77.5
	Gene Sp	0.1	1.4	11.3	67.2	11.6	65.9
	Gene F1	0.1	1.8	12.6	72.2	12.9	71.2
	Exon Sn	0.3	15.1	28.7	95.4	29.0	95.4
	Exon Sp	0.2	27.0	53.4	90.7	53.8	90.3
	Exon F1	0.2	19.3	37.3	93.0	37.7	92.8
M. musculus	Gene Sn	2.2	7.8	22.0	71.3	23.7	72.7
	Gene Sp	2.0	5.4	15.0	66.0	16.0	65.9
	Gene F1	2.1	6.4	17.8	68.6	19.1	69.1
	Exon Sn	25.4	49.7	57.3	91.2	58.1	91.7
	Exon Sp	25.4	50.9	64.2	90.7	64.8	90.7
	Exon F1	25.4	50.3	60.6	91.0	61.3	91.2

Table S2. Comparison of gene- and exon-level prediction accuracy between RNA-Seq-based BRAKER1, protein-based BRAKER2, TSEBRA (a tool generating the combination of BRAKER1 and BRAKER2 results), and GeneMark-ETP. The accuracy estimates are shown for the smaller (order excluded) and for the larger (species excluded) protein databases (see Materials).

		BRAKER1	Smaller protein DB			Larger protein DB		
			BRAKER2	TSEBRA	ETP	BRAKER2	TSEBRA	ETP
C. elegans	Gene Sn	61.8	46.8	60.3	60.4	69.0	71.1	68.4
	Gene Sp	65.6	54.1	77.5	67.7	70.1	80.5	73.8
	Gene F1	63.6	50.2	67.8	63.8	69.6	75.5	71.0
	Exon Sn	85.0	74.0	76.6	82.9	84.8	83.9	85.9
	Exon Sp	88.5	87.8	93.4	90.1	91.5	93.8	91.4
	Exon F1	86.7	80.3	84.2	86.4	88.0	88.6	88.6
A. thaliana	Gene Sn	59.6	72.6	73.6	75.8	79.2	79.3	77.9
	Gene Sp	61.3	70.1	81.2	80.0	75.6	82.8	81.0
	Gene F1	60.4	71.3	77.2	77.8	77.4	81.0	79.4
	Exon Sn	78.3	81.0	79.6	82.3	83.1	82.7	82.9
	Exon Sp	82.5	88.4	93.7	90.9	88.2	93.2	91.0
	Exon F1	80.4	84.5	86.1	86.4	85.6	87.6	86.8
D. melanogaster	Gene Sn	63.8	61.1	68.0	71.5	78.9	80.0	78.9
	Gene Sp	62.3	60.9	75.4	77.9	73.6	80.9	83.1
	Gene F1	63.0	61.0	71.5	74.6	76.1	80.4	80.9
	Exon Sn	77.0	71.4	72.1	76.4	80.1	79.8	80.7
	Exon Sp	80.9	83.4	89.9	89.7	88.5	92.2	91.4
	Exon F1	78.9	76.9	80.0	82.5	84.1	85.6	85.7
S. lycopersicum	Gene Sn	61.8	79.6	82.5	88.2	84.2	85.4	90.2
	Gene Sp	47.1	56.5	71.3	81.4	58.9	72.1	79.8
	Gene F1	53.5	66.1	76.5	84.7	69.3	78.2	84.6
	Exon Sn	90.7	94.2	94.9	96.7	95.4	96.1	97.2
	Exon Sp	75.5	82.8	90.3	92.6	82.3	90.2	91.6
	Exon F1	82.4	88.1	92.5	94.6	88.4	93.0	94.3
D. rerio	Gene Sn	51.7	55.0	66.9	72.7	57.8	69.0	73.8
	Gene Sp	28.1	29.5	45.7	56.5	27.9	46.0	56.8
	Gene F1	36.4	38.4	54.3	63.6	37.6	55.2	64.2
	Exon Sn	91.1	88.0	89.4	93.6	89.4	90.1	94.0
	Exon Sp	75.4	78.9	87.2	85.1	76.2	86.8	85.1
	Exon F1	82.5	83.2	88.3	89.2	82.2	88.4	89.3
G. gallus	Gene Sn	6.6	25.2	26.7	78.0	27.2	28.3	77.5
	Gene Sp	3.5	16.6	22.2	67.2	18.1	23.3	65.9
	Gene F1	4.6	20.0	24.2	72.2	21.7	25.6	71.2
	Exon Sn	66.1	35.0	59.8	95.4	35.3	60.0	95.4
	Exon Sp	48.1	59.2	74.4	90.7	60.6	74.4	90.3
	Exon F1	55.7	44.0	66.3	93.0	44.6	66.4	92.8
M. musculus	Gene Sn	27.8	32.5	44.2	71.3	35.9	46.7	72.7
	Gene Sp	14.8	21.2	31.3	66.0	23.2	32.7	65.9
	Gene F1	19.3	25.7	36.7	68.6	28.2	38.5	69.1
	Exon Sn	83.9	57.6	77.4	91.2	59.3	78.1	91.7
	Exon Sp	67.5	71.6	83.3	90.7	72.7	83.5	90.7
	Exon F1	74.8	63.8	80.2	91.0	65.3	80.7	91.2

Table S3. A gene-level accuracy evaluation of initial GeneMarkS-T predictions and the refined ones which are used to identify the high-confidence genes. The accuracy is shown separately for complete and partial predictions as well as for both sets together (Combined). The first three columns (Raw GeneMarkS-T) show the accuracy of unprocessed GeneMarkS-T predictions in all assembled transcripts. The remaining columns (HC genes) show the accuracy of the processed, high-confidence gene sets. The accuracy of HC genes is shown for the smaller (order excluded) and for the larger (species excluded) protein databases (see Materials).

		Raw GeneMarkS-T			HC genes					
					Smaller protein DB			Larger protein DB		
		Complete	Partial	Combined	Complete	Partial	Combined	Complete	Partial	Combined
C. elegans	Sn	42.9	3.9	46.8	33.6	2.1	35.7	47.7	4.0	51.7
	Sp	82.0	18.2	63.4	88.8	81.5	88.4	91.5	80.7	90.6
A. thaliana	Sn	49.8	1.4	51.2	55.6	1.1	56.7	57.3	1.6	58.8
	Sp	89.1	17.0	79.9	97.4	92.3	97.3	97.8	90.8	97.6
D. melanogaster	Sn	56.4	3.2	59.6	53.3	1.8	55.0	60.6	3.1	63.7
	Sp	87.5	38.1	81.8	95.0	85.3	94.7	96.9	85.0	96.3
S. lycopersicum	Sn	66.3	1.4	67.8	73.7	1.3	74.9	74.2	1.5	75.6
	Sp	84.1	26.6	77.8	95.4	87.2	95.2	95.4	84.8	95.1
D. rerio	Sn	55.3	4.3	59.6	62.8	4.2	67.0	62.4	4.5	66.9
	Sp	68.4	32.8	59.9	89.7	78.9	88.5	92.8	75.3	90.4
G. gallus	Sn	43.9	5.7	49.6	67.9	6.5	74.4	66.3	7.7	74.0
	Sp	64.0	23.0	47.0	89.5	86.1	89.1	90.0	80.3	88.4
M. musculus	Sn	48.4	1.2	49.6	60.8	2.7	63.5	60.5	3.4	63.9
	Sp	80.4	9.6	63.2	95.1	68.0	93.2	96.7	69.8	94.5

Table S4. Accuracy of the transcript classification as complete/partial (described in the main text). The transcripts used in this evaluation were i/ classified as partial by GeneMarkS-T, ii/ had a correctly predicted stop codon, and iii/ contained no assembly errors. The names of rows and columns are the same as in the confusion matrix shown in Table 2, see Results. Sensitivity represents the percentage of complete transcripts that were classified as such. The error rate is defined as the percentage of partial transcripts incorrectly classified as complete. The results are shown for the smaller (order excluded) and for the larger (species excluded) protein databases.

`		Smaller protein DB			Larger protein DB		
		Complete	Partial	Accuracy	Complete	Partial	Accuracy
C. elegans	Predicted complete	1488	127		1982	78	
	Predicted partial	273	207		393	471	
	Sensitivity (complete)			84.5			83.5
	Error rate (partial)			38.0			14.2
A. thaliana	Predicted complete	1476	55		1442	22	
	Predicted partial	107	203		165	249	
	Sensitivity (complete)			93.2			89.7
	Error rate (partial)			21.3			8.1
D. melanogaster	Predicted complete	273	76		299	9	
	Predicted partial	48	254		130	388	
	Sensitivity (complete)			85.1			69.7
	Error rate (partial)			23.0			2.3
S. lycopersicum	Predicted complete	897	81		868	63	
	Predicted partial	81	322		119	358	
	Sensitivity (complete)			91.7			87.9
	Error rate (partial)			20.1			15.0
D. rerio	Predicted complete	1152	107		1052	69	
	Predicted partial	249	1242		364	1318	
	Sensitivity (complete)			82.2			74.3
	Error rate (partial)			7.9			5.0
G. gallus	Predicted complete	3232	197		2972	114	
	Predicted partial	449	849		715	937	
	Sensitivity (complete)			87.8			80.6
	Error rate (partial)			18.8			10.9
M. musculus	Predicted complete	2026	16		1879	8	
	Predicted partial	497	205		642	216	
	Sensitivity (complete)			80.3			74,53
	Error rate (partial)			7.2			3.6

Table S5. Distribution of the predicted exons among four categories along with average Specificity values (exon level) for each category. The categories differ by the strength of extrinsic evidence for predicted genes (see text). Descriptions of the species-specific smaller and larger protein databases are given in Methods.

Species	Types of support	Smaller protein DB		Larger protein DB	
		# of exons	Specificity, %	# of exons	Specificity, %
C. elegans	Fully extrinsic	53,534	97.17	74,548	97.28
	Partially extrinsic	38,696	88.43	37,472	86.24
	Ab initio anchored	21,962	83.77	7,279	74.31
	Ab initio unsupported	4,769	54.54	2,286	37.18
A. thaliana	Fully extrinsic	102,615	98.84	108,633	98.79
	Partially extrinsic	25,406	85.15	26,650	77.37
	Ab initio anchored	6,538	63.51	4,759	37.57
	Ab initio unsupported	7,384	24.74	2,829	11.52
D. melanogaster	Fully extrinsic	35,300	97.72	42,821	97.67
	Partially extrinsic	12,443	82.17	11,455	76.88
	Ab initio anchored	3,175	76.28	329	52.89
	Ab initio unsupported	2,766	36.26	1,084	9.41
S. lycopersicum	Fully extrinsic	108,024	98.37	110,645	98.29
	Partially extrinsic	25,610	75.58	26,784	71.95
	Ab initio anchored	5,507	59.52	4,893	47.29
	Ab initio unsupported	11,112	17.02	8,799	12.48
D. rerio	Fully extrinsic	156,781	97.59	156,506	98.12
	Partially extrinsic	102,256	70.55	105,941	69.77
	Ab initio anchored	9,398	34.35	7,360	27.35
	Ab initio unsupported	43,023	2.51	40,983	1.85
G. gallus	Fully extrinsic	129,144	98.16	126,410	98.24
	Partially extrinsic	50,046	75.08	53,784	75.18
	Ab initio anchored	2,968	31.2	3,008	25.37
	Ab initio unsupported	33,168	0.71	33,111	0.58
M. musculus	Fully extrinsic	141,520	99.1	143,186	99.29
	Partially extrinsic	55,236	72.95	55,394	72.5
	Ab initio anchored	5,202	49.81	5,063	43.12
	Ab initio unsupported	61,229	2.08	58,337	1.15

Table S6. Gene- and exon-level prediction accuracy of GeneMark-ETP with and without filtering of pure *ab initio* predictions. The superior F1 accuracy is highlighted in bold. The unverified *ab initio* predictions were removed from the GeneMark-ETP outputs for genomes larger than 300 Mbp in length (the bottom four genomes). For each genome, the results are shown for the smaller (order excluded) and for the larger (species excluded) protein databases.

		Smaller protein DB		Larger protein DB	
		All predictions	*Ab initio* removed	All predictions	*Ab initio* removed
C. elegans	Gene Sn	60.4	58.7	68.4	67.7
	Gene Sp	67.7	71.1	73.8	76.2
	Gene F1	63.8	**64.3**	71.0	**71.7**
	Exon Sn	82.9	80.9	85.9	85.3
	Exon Sp	90.1	91.6	91.4	92.4
	Exon F1	**86.4**	86.0	88.6	**88.7**
A. thaliana	Gene Sn	75.8	72.8	77.9	77.5
	Gene Sp	80.0	86.7	81.0	84.2
	Gene F1	77.8	**79.1**	79.4	**80.7**
	Exon Sn	82.3	81.1	82.9	82.7
	Exon Sp	90.9	94.6	91.0	92.6
	Exon F1	86.4	**87.3**	86.8	**87.4**
D. melanogaster	Gene Sn	71.5	67.4	78.9	78.4
	Gene Sp	77.9	82.3	83.1	85.0
	Gene F1	**74.6**	74.1	80.9	**81.6**
	Exon Sn	76.4	74.8	80.7	80.6
	Exon Sp	89.7	92.6	91.4	93.0
	Exon F1	**82.5**	82.7	85.7	**86.4**
S. lycopersicum	Gene Sn	89.5	88.2	90.6	90.2
	Gene Sp	70.6	81.4	70.9	79.8
	Gene F1	78.9	**84.7**	79.5	**84.6**
	Exon Sn	97.1	96.7	97.4	97.2
	Exon Sp	87.1	92.6	87.0	91.6
	Exon F1	91.8	**94.6**	91.9	**94.3**
D. rerio	Gene Sn	72.9	72.7	73.8	73.8
	Gene Sp	39.4	56.5	40.3	56.8
	Gene F1	51.2	**63.6**	52.2	**64.2**
	Exon Sn	93.9	93.6	94.2	94.0
	Exon Sp	73.7	85.1	74.1	85.1
	Exon F1	82.5	**89.2**	82.9	**89.3**
G. gallus	Gene Sn	78.1	78.0	77.5	77.5
	Gene Sp	40.7	67.2	40.0	65.9
	Gene F1	53.5	**72.2**	52.8	**71.2**
	Exon Sn	95.5	95.4	95.4	95.4
	Exon Sp	76.9	90.7	76.5	90.3
	Exon F1	85.2	**93.0**	85.0	**92.8**
M. musculus	Gene Sn	71.7	71.3	72.8	72.7
	Gene Sp	34.5	66.0	35.3	65.9
	Gene F1	46.5	**68.6**	47.6	**69.1**
	Exon Sn	91.6	91.2	92.0	91.7
	Exon Sp	70.1	90.7	70.7	90.7
	Exon F1	79.4	**91.0**	79.9	**91.2**

Table S7. The values of the masking penalty parameter estimated by GeneMark-ETP for each of the tested genomes (natural logarithms). For GC-heterogeneous genomes, the optimal masking penalty parameter was estimated for each of the GC bins. For each species, the results are shown for the smaller and larger protein databases (see caption to Table S6).

	Smaller protein DB			Larger protein DB		
C. elegans	0.06			0.05		
A. thaliana	0.03			0.03		
D. melanogaster	0.08			0.08		
S. lycopersicum	0.04			0.04		
D. rerio	0.08			0.09		
GC	Low	Medium	High	Low	Medium	High
G. gallus	0.15	0.17	0.12	0.14	0.16	0.11
M. musculus	0.13	0.14	0.14	0.14	0.14	0.14

Table S8: Data sources used in the tests. A date in parenthesis shows the date of the last update. *The reliable subset for *M. musculus* was selected by choosing a subset of GENCODE transcripts with the following attributes: *CCDS* (Agreement with RefSeq annotation), *transcript_support_level=1* (All splice junctions of the transcript were supported by at least one non-suspect mRNA), and *basic* (prioritizes full-length protein-coding transcripts over partial or non-protein-coding transcripts within the same gene).

Species	Assembly version	Main annotation	Supplementary annotation used to prepare the reliable subset
C. elegans	GCF_000002985.6	Wormbase WS284 (Feb 2022)	-
A. thaliana	GCF_000001735.4	Araport11 (Mar 2021)	-
D. melanogaster	GCF_000001215.4	FlyBase r6.44 (Feb 2022)	-
S. lycopersicum	GCF_000188115.4	NCBI annot. Release 103 (Jun 2019)	ITAG3.2 (Jun 2017)
D. rerio	GCF_000002035.6	NCBI annot. Release 106 (Oct 2019)	Ensembl GRCz11.105 (Oct 2021)
G. gallus	GCF_000002315.6	NCBI annot. Release 104 (Mar 2020)	Ensembl GRCg6a.105 (Oct 2021)
M. musculus	GCF_000001635.27	GENCODE M28 (Dec 2021)	RefSeq*

Table S9: Composition of the clades of OrthoDB v10.1 used by GeneMark-ETP. The bold case black numbers show the largest numbers of species that could be used to support gene predictions for a given species (left column). These numbers correspond to the species-specific IP_0 databases (see Materials). The numbers of species removed from the species-specific IP_0 databases to make corresponding larger and smaller protein databases are shown in blue.

Species	# of species in the OrthoDB clade						Name of the largest OrthoDB segment	# of proteins in the OrthoDB segment
	Genus	Family	Order	Class	Phylum	Kingdom		
C. elegans	3	3	5	6	7	**448**	Metazoa	8,266,016
A. thaliana	2	8	10	-	100	**117**	Plantae	3,510,742
D. melanogaster	20	20	56	148	**170**	-	Arthropoda	2,601,995
S. lycopersicum	2	10	11	-	100	**117**	Plantae	3,510,742
D. rerio	1	5	5	50	**246**	-	Chordata	5,003,104
G. gallus	1	3	4	62	246	-	Chordata	5,003,104
M. musculus	3	5	20	111	246	-	Chordata	5,003,104

Table S10: RNA-Seq libraries used for the experiments with GeneMark-ETP.

Species	RNA-Seq library ID	Number of paired reads (M)	Read length (nt)	Library size (Gb)
C. elegans	SRR065717	29.1	76	4.4
	SRR065719	73.3	76	11.1
	SRR473298	19.9	100	4.0
	SRR2054452	10.2	100	2.0
	Total	132.5		21.5
A. thaliana	SRR934391	20.0	101	4.0
	SRR5588566	24.7	125	6.2
	SRR7169927	19.2	101	3.9
	Total	63.9		14.1
D. melanogaster	SRR023505	8.4	76	1.3
	SRR023546	8.9	76	1.4
	SRR023608	11.9	76	1.8
	SRR026433	22.1	76	3.4
	SRR027108	7.2	76	1.1
	Total	58.5		9.0
S. lycopersicum	SRR2002284	56.2	73	8.2
	SRR7959012	25.4	149	7.6
	SRR7959019	27.9	149	8.3
	SRR14055940	21.2	150	6.4
	Total	130.7		30.5
D. rerio	SRR9735169	28.2	75	4.2
	SRR10004226	21.6	150	6.5
	SRR10040127	25.9	126	6.5
	Total	75.7		17.2
G. gallus	ERR2812450	44.9	150	13.5
	SRR3971633	24.0	100	4.8
	SRR6337028	10.0	100	2.0
	SRR11038071	16.4	151	5.0
	Total	95.3		25.3
M. musculus	SRR567480	155.7	101	31.5
	SRR567482	161.1	101	32.5
	SRR567497	94.3	101	19.0
	Total	411.1		83.0

Table S11: Performance of MAKER2 and GeneMark-ETP gene prediction algorithms on three model species.

Drosophila melanogaster				
		MAKER	GeneMark-ETP	diff
exon	Sn	75.2	**80.7**	5.6
	Sp	74.0	**91.4**	17.5
	F1	74.6	**85.7**	11.2
gene	Sn	60.2	**79.0**	18.8
	Sp	55.3	**83.0**	27.7
	F1	57.7	**81.0**	23.3
transcript	Sn	38.3	**54.8**	16.5
	Sp	51.7	**79.4**	27.7
	F1	44.0	**64.8**	20.8
Danio rerio				
		MAKER	GeneMark-ETP	diff
exon	Sn	83.3	**93.9**	10.7
	Sp	79.2	**84.9**	5.7
	F1	81.2	**89.2**	8.0
gene	Sn	47.7	**73.5**	25.9
	Sp	37.6	**56.2**	18.6
	F1	42.0	**63.7**	21.7
transcript	Sn	42.8	**68.5**	25.7
	Sp	35.1	**55.8**	20.7
	F1	38.5	**61.5**	22.9
Mus musculus				
		MAKER	GeneMark-ETP	diff
exon	Sn	79.2	**91.7**	12.6
	Sp	77.4	**87.9**	10.5
	F1	78.3	**89.8**	11.5
gene	Sn	41.6	**73.1**	31.5
	Sp	34.8	**59.7**	24.9
	F1	37.9	**65.7**	27.8
transcript	Sn	33.4	**61.9**	28.5
	Sp	32.2	**61.9**	29.7
	F1	32.8	**61.9**	29.1

References

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59-60.