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

# current address: U.S. Department of Energy, Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Supplemental Methods

S1. GeneMarkS-TP: predicting genes in RNA transcripts with protein database support.

S1.1 Corrections of the 5' end gene predictions

First, the CDS prediction in assembled transcripts is done by GeneMarkS-T (Tang et al, 2015). We have observed that GeneMarkS-T made very few errors when predicting 5' complete genes, those with start codons within transcripts. On the other hand, the 5' incomplete genes predicted by GeneMarkS-T with the CDS start residing near the first nucleotide of a transcript, carry more frequent errors and should be corrected. We need to discriminate between a correctly predicted 5' incomplete CDS vs an incorrect 5' incomplete CDS with a true complete CDS residing inside.

Incomplete genes predicted by GeneMarkS-T in transcripts serve as queries in searches for homologous proteins (targets) in a reference protein database (e.g. by DIAMOND (Buchfink et al. 2015)). If among the similarity search hits (targets) exists at least one target that i/ is common for both queries and ii/ shows *better support* for the 5' partial gene, *the 5' partial gene is predicted*. Otherwise, the CDS starting with the internal ATG is selected as the *predicted complete gene*. If the sets of protein targets in the two searches (those with 25 best scores, the default setting) do not overlap, the 5' partial CDS is selected. If both similarity searches do not produce targets, the transcript is removed from consideration.

The quantitative meaning of the *better support* follows from analysis of alignment data in the form of the following condition:

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

Here a and b are the starting positions of the local alignments within the target protein for the longer and shorter protein queries respectively (Supplemental Fig. S3 and S4). $AAI_{partial}$ and $AAI_{complete}$ are, respectively, the percentages of *amino acid identities* in the alignments of the longer and shorter query proteins to the target protein. $AAI_{partial}$ is defined within the range "a-c", $AAI_{complete}$ is defined within the range "b-c", where c is the common end position of the two local pairwise alignments (Supplemental Figs. S3 and S4).

If condition (S1) is fulfilled, the longer query is selected, the 5' partial gene.
If condition (S1) is not fulfilled, the shorter query, a *complete* gene is selected.

Notably, "a-1" is the length of unaligned N proximal part of the long query.
A large "a-1" is likely to indicate a presence of translated 5' UTR region situated upstream to a complete gene. A small "b-a" indicates that an extension of the complete gene candidate does not extend the zone of two proteins similarity, again a support of the complete gene prediction.

The larger value of the *AAI* ratio the more conservation exists between query and target protein subsequences in the range "b-a". Therefore, the increase of the *AAI* ratio favors the 5' partial candidate. The *AAI* ratio is scaled by using logarithm with a factor 1,000, i.e. 1,000*log(…).

S1.2 Removal of the 3' partial gene predictions

The 3' partial predictions were rarely observed. This frequency pattern could be expected since RNA-Seq libraries used in our experiments, prepared with the poly-A tail enrichment of mRNA transcripts, should predominantly carry transcripts complete at 3' end (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. Extensions of GeneMarkS-T gene predictions to the longest ORFs

Most eukaryotic genes are translated from the ATG start codon closest to the transcript 5' end (Kozak 1999). Still, the translation can be initiated at one of the downstream ATG starts; e.g., when the most upstream start has a weak translation initiation signal known as the Kozak pattern (Kozak 1987). GeneMarkS-T computes Kozak pattern score (with respect to the model with parameters derived in species-specific self-training) to account for the possibility of non-5'-most translation start codons. However, the Kozak pattern is relatively weak. We have observed that the gene predictions with non-5'-most start codons carry a higher false-positive rate than the predictions with 5'-most start codons. Therefore, GeneMark-ETP uses the following rule. If a gene predicted in a transcript could be extended to the 5'-most start codon, and the translation of this extension is supported by alignment to a target protein, the extended version of the predicted gene is considered as a candidate for an HC gene along with the one with non-5'-most start.

S1.4 Complete genes with uniform protein support

In the considered above similarity searches we have dealt with local pairwise alignments. Still, being interested in accurate prediction of all protein-coding exons, we are concerned about a *uniform* protein support showing evolutionary conservation over the whole protein-coding region. We say that a *uniform protein support* exists for a predicted *complete* gene if there is a significant BLASTp alignment (with E-value better than 10^{-3}) of the translation of the predicted gene *Q* to a protein in a database *T* and the following condition is satisfied:

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

Here, Q_{start}, Q_{end}, (T_{start}, T_{end}) are, respectively, the positions of the start and end of the alignment within the query protein (within the target protein); Q_{len}, T_{len} are the lengths of the query and target proteins, respectively (Supplemental Fig. S5).

Experiments with multiple sequence alignments (MSA) of orthologous proteins demonstrated that internal sections of MSA were usually most conserved, while the N-proximal regions of the proteins were less conserved and the least conserved regions in MSA were usually C-proximal regions. Therefore, testing for conservation of the N- and C- proximal regions provided sufficient evidence of evolutionary conservation across the pair of proteins. Condition S2 allows some misalignment at the alignment start and even to a larger degree at the alignment end. A gene prediction is called a complete gene with uniform protein support if a query complete protein has an alignment to at least one target (out of the best scored 25, the default setting) that satisfies condition S2. All such predicted genes are included in the set of high-confidence gene predictions.

S1.5 Tests of conditions S1 and S2

To assess the degree of improvement in the quality of gene sets selected with conditions S1 and S2, we used the following approach. We have prepared test sets of transcript sequences with complete and partial genes. The ground-truth labels were determined from reference annotations. GeneMarkS-T was run on these sequences. Next, for each transcript, the alignments of the longer and shorter queries with the target proteins were made and the features used in conditions S1 and S2 were selected. We assessed the efficiency of the empirical rules for selecting partial and complete genes (condition S1) as well as selecting genes with uniform protein support (conditions S2) with efficiency of two other possible 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) i/ to classify gene predictions as complete or partial (compared to the use of condition S1), ii/ to claim uniform protein support (compared to the use of condition S2). The training sets for the two ML methods did not overlap with the test set. We observed that use of conditions S1 and S2 produced more accurate results than the results generated by application of general-purpose random forest or logistic regression models (data not shown).

S2. ProtHint filter for high-confidence gene candidates (in the *ab initio* category)

Some GeneMarkS-T gene predictions not uniformly supported by proteins (and not satisfying Condision S2) still could be included in the set of HC genes. Such predictions should satisfy several conditions (see Main text), one of which is no contradiction to the ProtHint hints. To detect such a conflict, we proceed as follows. First, a gene predicted by GeneMarkS-T is mapped to a particular locus of genomic DNA. Second, the translation of the initially predicted gene and its genomic locus is used by ProtHint as the protein and gene seeds to generate hints for the next round of gene prediction in the same locus (Bruna et al. 2020). Next, the borders of the thus determined exons are compared to the ProtHint hints. We say that the contradiction exists if (i) at least one of ProtHint's introns overlaps a mapped exon, or (ii) a ProtHint defined stop codon overlaps an exon or intron of the mapped gene, or (iii) a ProtHint start codon overlaps an exon or intron of the mapped gene (except the start-to-start overlap).

S3. Alternative HC isoforms

An additional round of selection is made to filter out possible false positives among HC isoforms. Here we consider the HC isoforms that satisfy Condition S2. Let $I_{complete}^{g}$ be a set of complete isoforms of gene g and $I_{partial}^{g}$ is a set of its partial isoforms. Each isoform i is assigned a score $\mathrm{s}(i)$ -- the *bitscore* of its best hit to a protein in the protein database. We compute the maximum score of all the complete isoforms for a gene g, denoted as $\mathrm{s}(g_{complete})$. A score of an isoform $\mathrm{s}(i)$ selected as complete HC isoform must satisfy the inequality:

$$\mathrm{s}(i) \geq 0.8 \times \mathrm{s}(g_{complete}) \quad (i \in I_{complete}^{g}) \qquad \text{(S3)}$$

Among the partial alternative isoforms of gene g, we determine the maximum score $\mathrm{s}(g_{partial})$. If $\mathrm{s}(g_{partial})$ is larger than $\mathrm{s}(g_{complete})$, the partial isoform with this largest score is selected as the partial HC isoform. In this case, all the complete HC isoforms are removed. Otherwise, if $\mathrm{s}(g_{partial})$, is lower than $\mathrm{s}(g_{complete})$, then only complete HC isoforms of gene g are retained.

If all alternative HC candidates were defined *ab initio*, then the one with the longest protein-coding region is selected as the predicted HC gene.

S4. Computing the species-specific repeat penalty parameter

For each genome, after identification of the HC genes and the first iteration of the GHMM model training, we estimate species-specific parameter q.

We have the set of the HC genes, the first version of the full GHMM model, and the coordinates of the repeats identified in genomic DNA. GeneMark.hmm is run several times with different *q* values to predict genes in the genomic sequences containing the HC genes for which we compute the gene level F1 value (Supplemental Fig. S7-A). The value *q* delivering the F1 maximum was chosen as the species-specific repeat penalty. We have shown that this value is close to *q* found when the test set of genes is made based on genome annotation. We also observed that the value *q* was robust with respect to the size of the HC genes set (data not shown).

Moreover, we have found that the use of the exon level Sn led to more robust estimation of q in comparison with use of the gene level F1 (data not shown). Practically, we first find the q' value maximizing the number of correctly predicted exons in the set of HC genes, e_{max} (Supplemental Fig. S7-B). Then, the value $q*$ at which $0.998 \times e_{max}$ exons are correctly predicted (marked for *A. thaliana* and *D. melanogaster* in panel A of Supplemental Fig. S7-A) is selected as q. To reduce the runtime of the repeat penalty parameter estimation we use simulated annealing (Kirkpatrick et al. 1983).

S5. Data sets used in computational experiments with MAKER2

Three model organisms representing three different types of genome organization were selected:

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

The following information was provided to MAKER2.

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 6ahari
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 6ahari
Octodon degus
Rattus norvegicus

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

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

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

For *Mus musculus*:
AUGUSTUS – the "human" model from the AUGUSTUS distribution.
GeneMark.hmm – the medium GC model created by GeneMark-ETP on the mouse genome.
SNAP – the "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
```

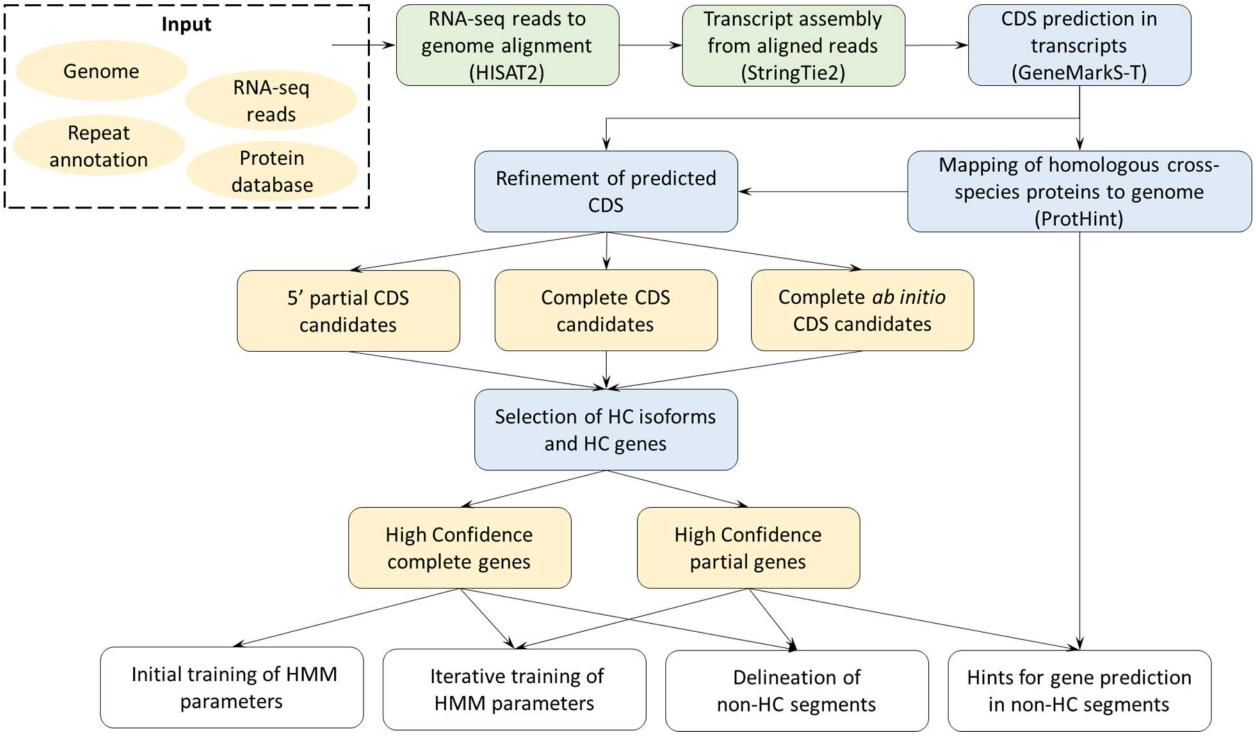
```
est=transcriptome.fasta
protein=proteindb.fasta
model_org=   #empty
rm_gff=repeatmasker.gff
snaphmm=snap.model
gmhmm=genemark.mod
augustus_species=model_name
est2genome=1
protein2genome=1
alt_splice=1
always_complete=1
keep_preds=1 for D. melanogaster
keep_preds=0 for D. rerio and M. musculus
split_hit=20000
max_dna_len=1000000
```

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

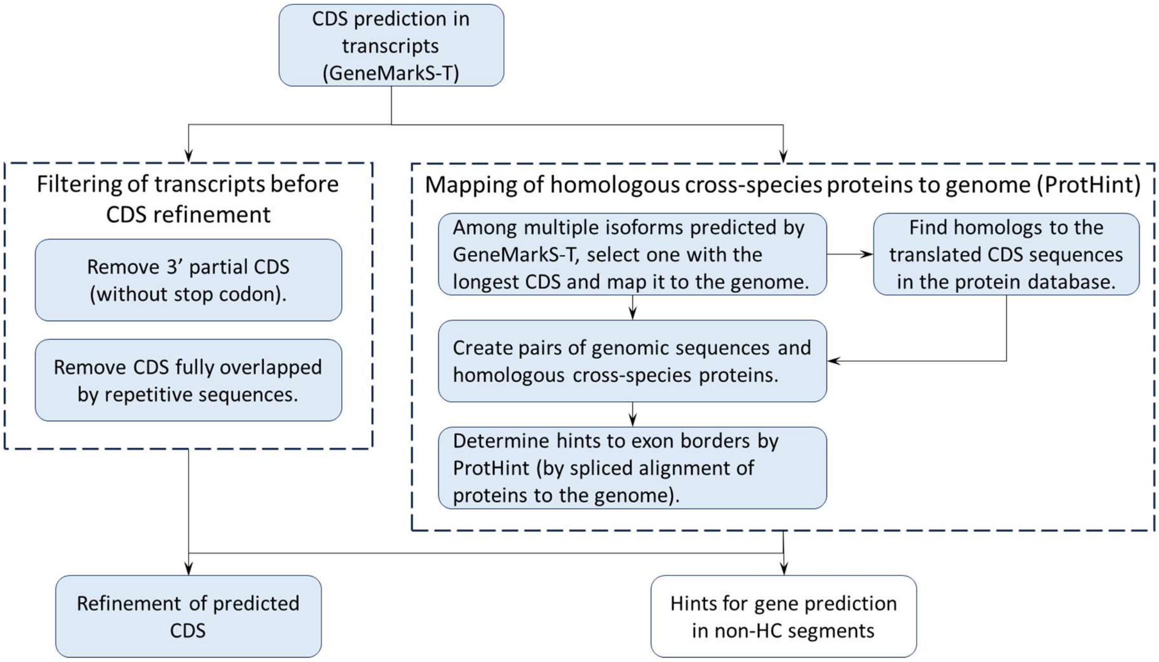
The gene prediction accuracy of MAKER2 and GeneMark-ETP (shown in Supplemental Table S6) was estimated as described in the main text.

Supplemental Figures

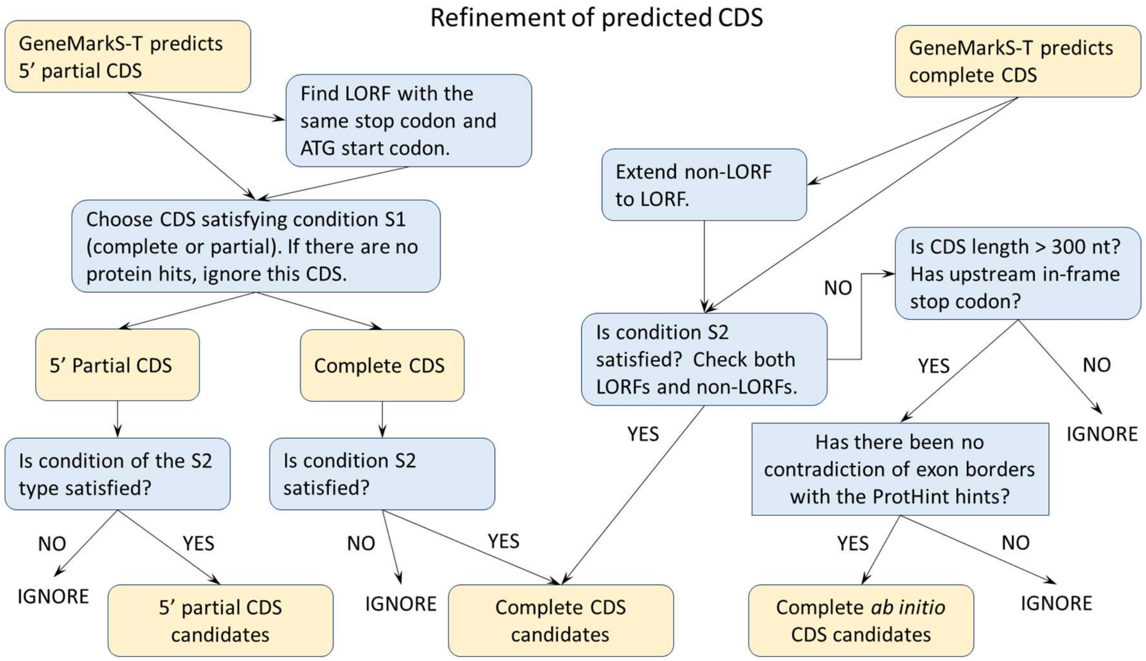
Supplemental Figures 1Sa-1Sg extend the flowchart of GeneMark-ETP shown in Fig. 1.


Input
Genome
RNA-seq reads
Repeat annotation
Protein database
RNA-seq reads to genome alignment (HISAT2)
Transcript assembly from aligned reads (StringTie2)
CDS prediction in transcripts (GeneMarkS-T)
Refinement of predicted CDS
Mapping of homologous cross-species proteins to genome (ProtHint)
5' partial CDS candidates
Complete CDS candidates
Complete *ab initio* CDS candidates
Selection of HC isoforms and HC genes
High Confidence complete genes
High Confidence partial genes
Initial training of HMM parameters
Iterative training of HMM parameters
Delineation of non-HC segments
Hints for gene prediction in non-HC segments

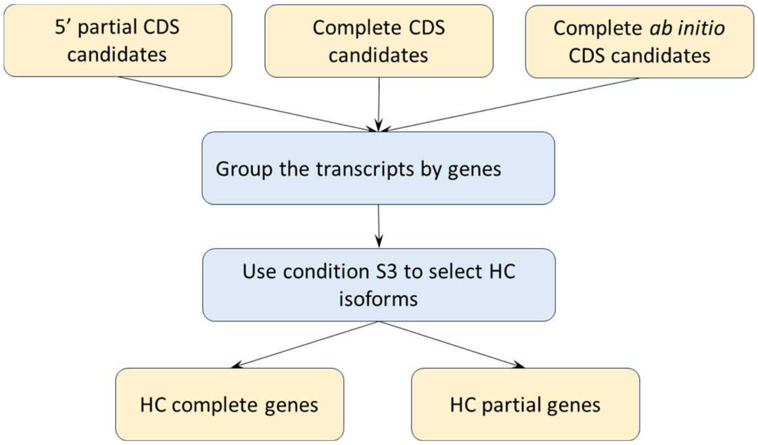
Supplemental Figure S1A. A high-level diagram illustrates the high-confidence (HC) gene identification procedure. Additional details are shown in Supplemental Figs. S1B-S1D.


CDS prediction in transcripts (GeneMarkS-T)
Filtering of transcripts before CDS refinement
Remove 3' partial CDS (without stop codon).
Remove CDS fully overlapped by repetitive sequences.
Mapping of homologous cross-species proteins to genome (ProtHint)
Among multiple isoforms predicted by GeneMarkS-T, select one with the longest CDS and map it to the genome.
Find homologs to the translated CDS sequences in the protein database.
Create pairs of genomic sequences and homologous cross-species proteins.
Determine hints to exon borders by ProtHint (by spliced alignment of proteins to the genome).
Refinement of predicted CDS
Hints for gene prediction in non-HC segments

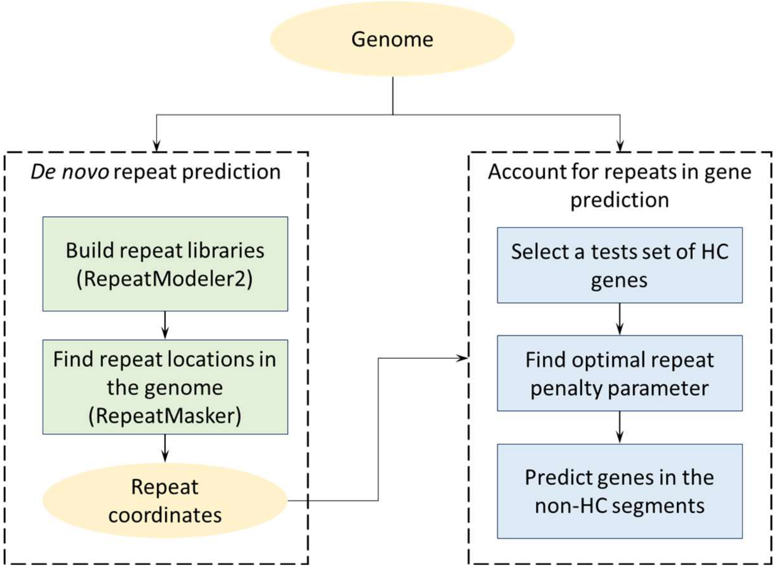
Supplemental Figure S1B. Details of the transcript processing.


Refinement of predicted CDS
GeneMarkS-T predicts 5' partial CDS
Find LORF with the same stop codon and ATG start codon.
Choose CDS satisfying condition S1 (complete or partial). If there are no protein hits, ignore this CDS.
5' Partial CDS
Complete CDS
Is condition of the S2 type satisfied?
NO
YES
IGNORE
5' partial CDS candidates
Is condition S2 satisfied?
NO
YES
IGNORE
Complete CDS candidates
GeneMarkS-T predicts complete CDS
Extend non-LORF to LORF.
Is condition S2 satisfied? Check both LORFs and non-LORFs.
YES
NO
Is CDS length > 300 nt? Has upstream in-frame stop codon?
YES
NO
IGNORE
Has there been no contradiction of exon borders with the ProtHint hints?
YES
NO
Complete *ab initio* CDS candidates
IGNORE

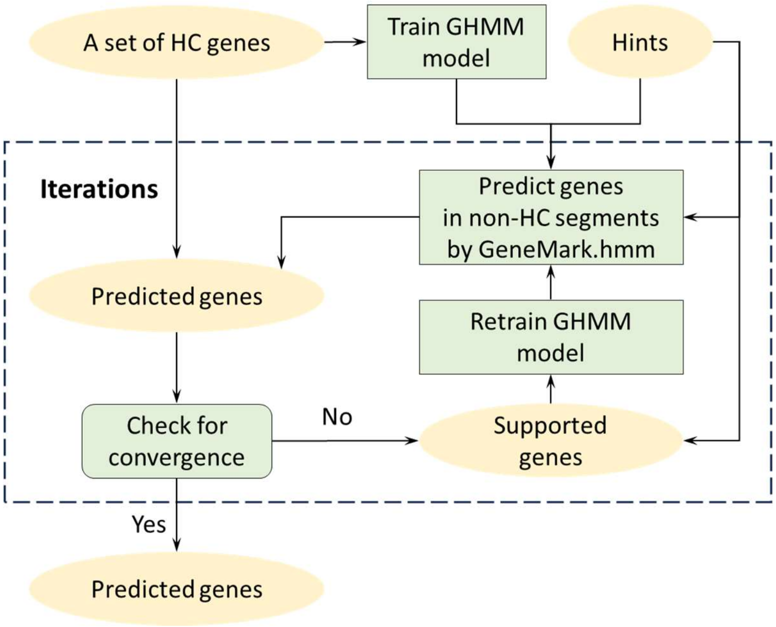
Supplemental Figure S1C. HC gene candidate generation procedure (refinement block in Supplemental Fig. S1A).


5' partial CDS candidates
Complete CDS candidates
Complete *ab initio* CDS candidates
Group the transcripts by genes
Use condition S3 to select HC isoforms
HC complete genes
HC partial genes

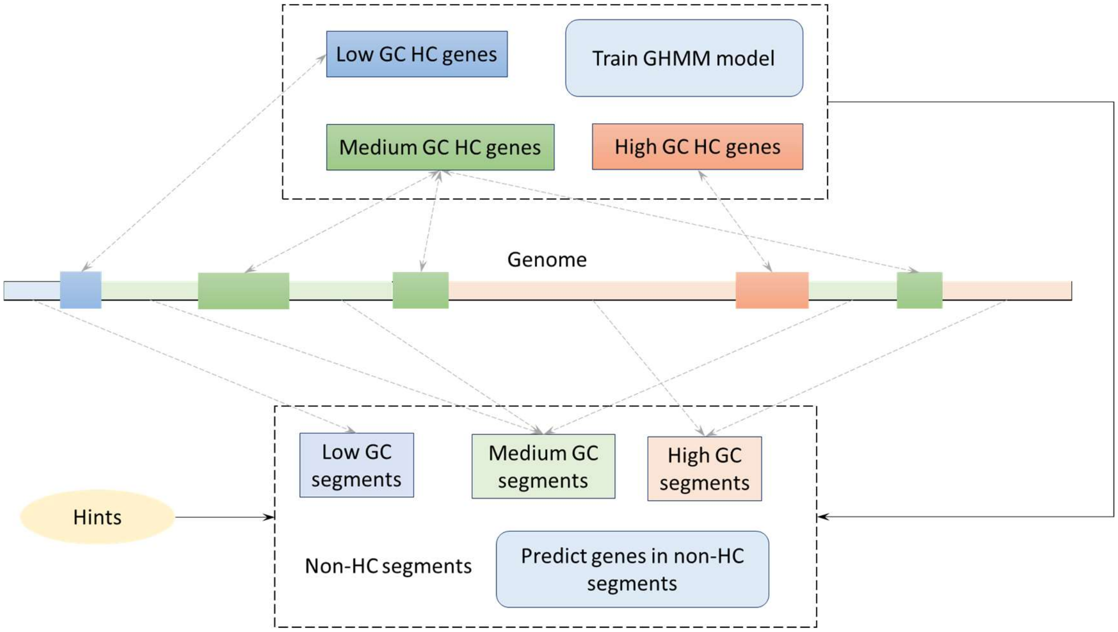
Supplemental Figure S1D. Details on the selection of HC genes and isoforms.


Genome
De novo repeat prediction
Build repeat libraries (RepeatModeler2)
Find repeat locations in the genome (RepeatMasker)
Repeat coordinates
Account for repeats in gene prediction
Select a tests set of HC genes
Find optimal repeat penalty parameter
Predict genes in the non-HC segments

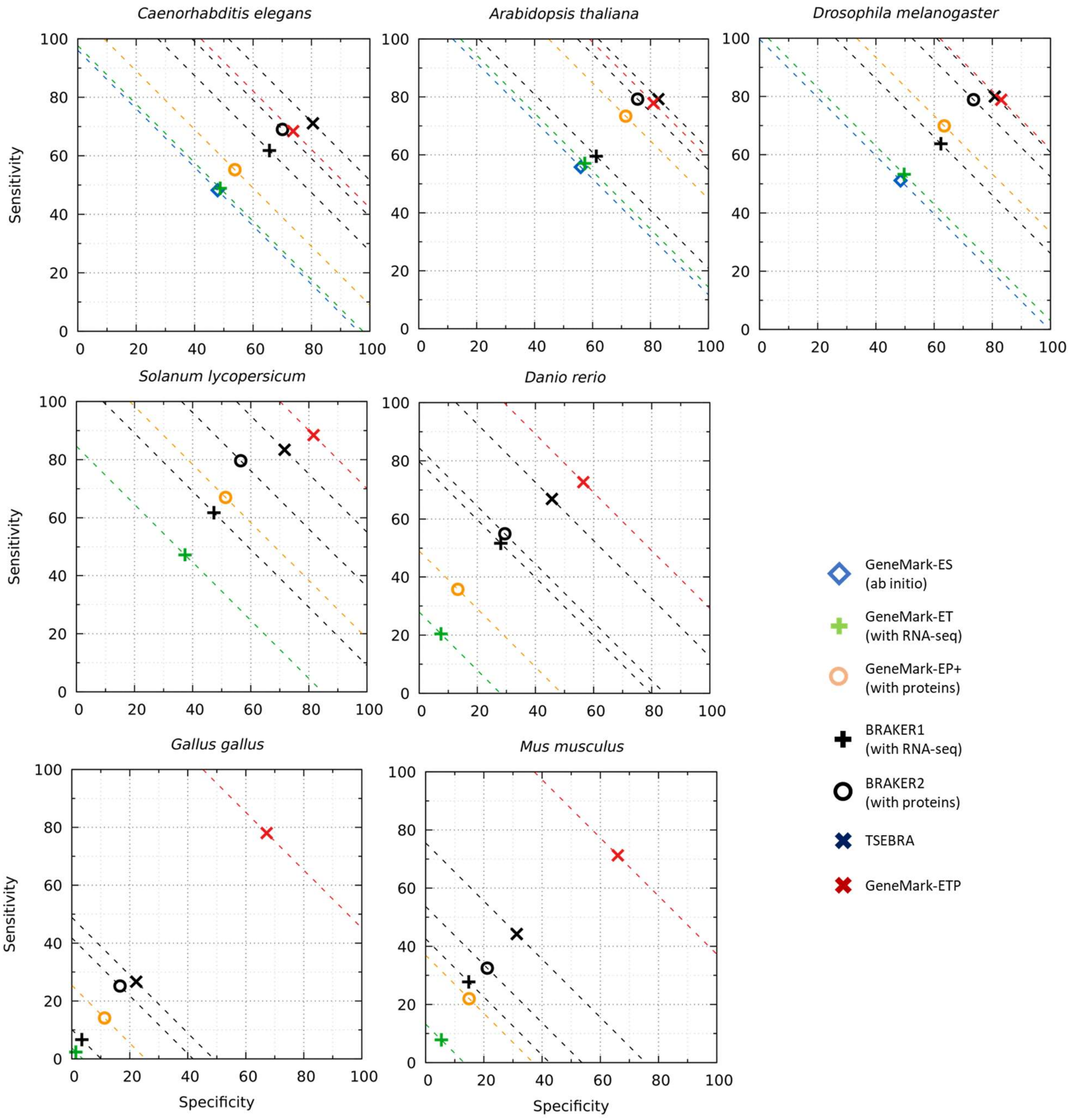
Supplemental Figure S1E. Details on the repetitive sequence identification and processing. *De novo* repeat prediction block is not included in GeneMark-ETP.



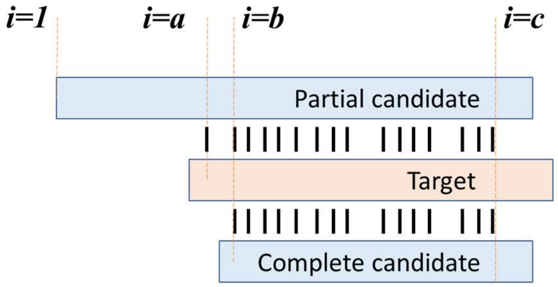
Supplemental Figure S1F. Workflow of the GHMM model training procedure for the GeneMark.hmm algorithm in GeneMark-ETP.



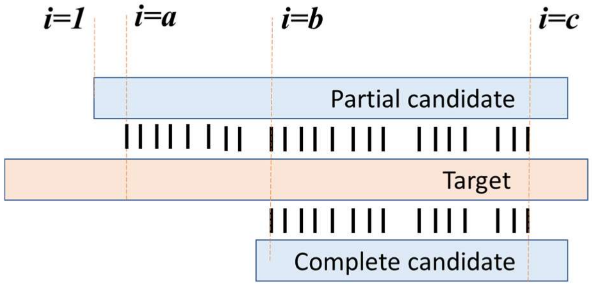
Supplemental Figure S1G. Schematics of the identification and use of the non-HC segments in training and gene prediction.



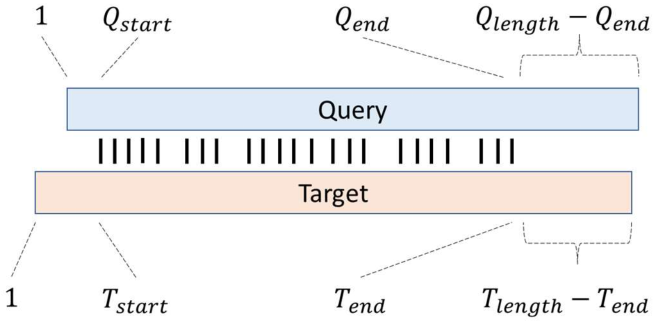
Supplemental Figure S2. Gene level accuracy of the seven gene prediction tools (see legends to Figs. 3-4). Compared to the figures in the main text, where we used smaller-size reference protein databases for each species (all proteins of the same taxonomic order were excluded from the corresponding IP_0 databases), here we used larger-size databases (proteins from the same species excluded from the corresponding IP_0 databases).



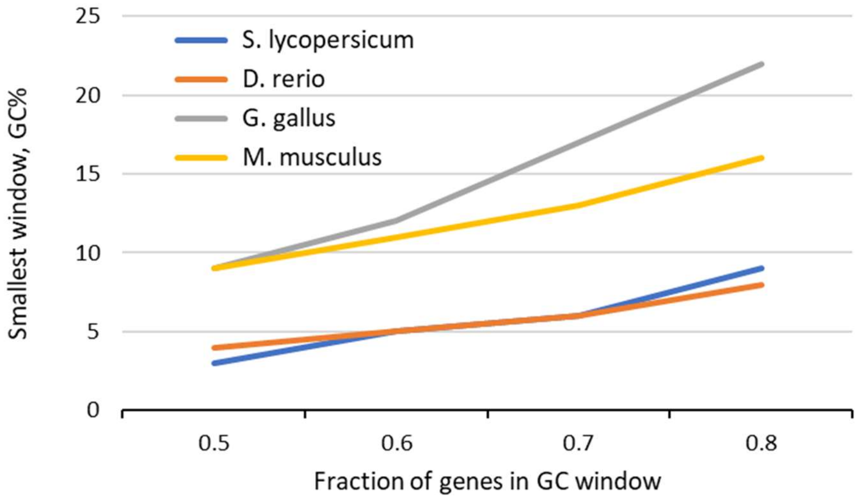
Supplemental Figure S3. The GeneMarkS-T gene prediction to be classified as *a complete gene*. Condition S1 is not fulfilled. Here *a, b* and *c* are defined as in Supplemental Fig. S5.



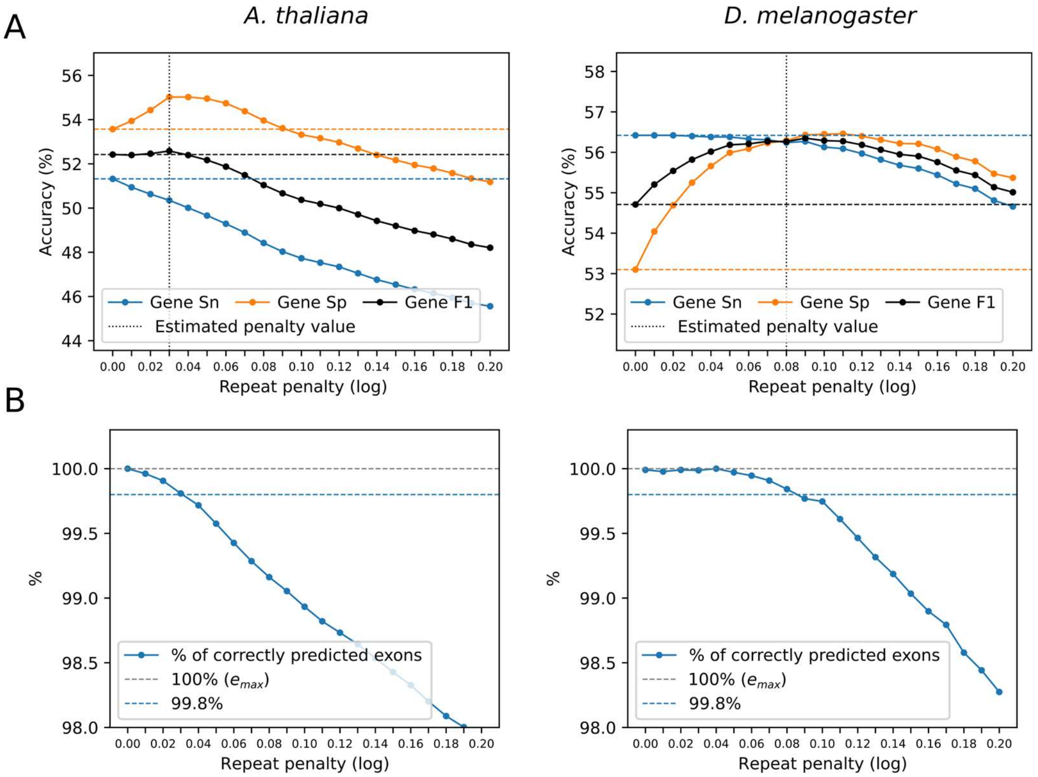
Supplemental Figure S4. The GeneMarkS-T gene prediction to be classified as a 5' *partial gene*. Condition S1 is fulfilled. Here *a* and *b* are positions of the starts of the local alignments of respective longer and shorter protein queries, while *c* is the end position of the local pairwise alignments.



Supplemental Figure S5. The features used in condition S2.



Supplemental Figure S6. Analysis of the genome GC content inhomogeneity. The graphs show the size of the narrowest GC% window that would contain a given amount of the annotated genes (fraction of a whole gene complement in a given genome)



Supplemental Figure S7. A. Dependence of the gene level Sn, Sp and F1 values (determined for the full sets of HC genes) on the repeat penalty parameter *q* (natural log) for genomes of *A. thaliana* and *D. melanogaster*. **B.** Dependence of fraction (%) of correctly predicted exons of *the HC genes* (Sn) on the repeat penalty parameter *q* for the same genomes as in A. (Suppl. Materials)

Supplemental Tables

Supplemental Table S1 GeneMarkS-TP processing of transcript protein and genomic data. Transformation of the initially predicted genes in assembled transcripts (Section A) into a set of HC genes (Section B).

A

Species	# of annotated genes	# of RNA-seq paired reads (M)	RNA-seq library size (Gb)	# of genes predicted by GeneMarkS-T	Ratio of # of GeneMarkS-T to # of annotated genes	Sn/Sp for GeneMarkS-T predictions
C. elegans	19,969	132.5	21.5	14,746	0.74	46.8 / 63.4
A. thaliana	27,445	63.9	14.1	17,589	0.64	51.2 / 79.9
D. melanogaster	13,951	58.5	9.0	10,163	0.73	59.6 / 81.8
S. lycopersicum	25,158	130.7	30.5	19,526	0.78	67.8 / 77.8
D. rerio	25,611	75.7	17.2	22,992	0.90	59.6 / 59.9
G. gallus	17,279	95.3	25.3	17,381	1.01	49.6 / 47.0
M. musculus	22,611	411.1	83.0	15,819	0.70	49.6 / 63.2

B

Species	# of proteins in the Order excluded DB	# of HC genes found with the Order excluded DB	Ratio of # of HC genes to # of annotated genes	Sn/Sp for predicted HC genes
C. elegans	8,168,321	8,062	0.40	35.7 / 88.4
A. thaliana	3,160,482	16,008	0.58	55.0 / 94.7
D. melanogaster	1,785,203	8,109	0.58	59.6 / 81.8
S. lycopersicum	3,116,328	17,231	0.68	74.9 / 95.2
D. rerio	4,791,893	16,918	0.66	67.0 / 88.5
G. gallus	4,933,362	12,473	0.72	74.4 / 89.1
M. musculus	1,835,426	13,057	0.58	63.5 / 93.2

Species	# of proteins in the Species excluded DB	# of HC genes found with the Species excluded DB	Ratio of # of HC genes to # of annotated genes	Sn/Sp for predicted HC genes
. elegans	8,245,445	11,399	0.57	51.7 / 90.6
A. thaliana	3,483,291	16,551	0.60	58.8 / 97.6
D. melanogaster	2,588,444	9,223	0.66	63.7 / 96.3
S. lycopersicum	3,456,742	17,489	0.70	75.8 / 95.1
D. rerio	4,973,735	16,573	0.65	66.9 / 90.4
G. gallus	4,984,020	12,564	0.73	74.0 / 88.4
M. musculus	2,228,727	12,965	0.57	63.9 / 94.5

Supplemental Table S2. Distribution of the predicted exons among four categories of extrinsic support. Average Specificity values (exon level) are given for each category. Descriptions of the smaller and larger protein databases compiled for each species are given in Methods.

Species	Intermediate set of gene predictions	Smaller protein DB		Larger protein DB	
		# of exons	Specificity, %	# of exons	Specificity, %
C. elegans	Fully extrinsic	53,534	97.2	74,548	97.3
	Partially extrinsic	38,696	88.4	37,472	86.2
	With extrinsic match	21,962	83.8	7,279	74.3
	With no extrinsic match	4,769	54.5	2,286	37.2
A. thaliana	Fully extrinsic	102,615	98.8	108,633	98.8
	Partially extrinsic	25,406	85.2	26,650	77.4
	With extrinsic match	6,538	63.5	4,759	37.6
	With no extrinsic match	7,384	24.7	2,829	11.5
D. melanogaster	Fully extrinsic	35,300	97.7	42,821	97.7
	Partially extrinsic	12,443	82.2	11,455	76.9
	With extrinsic match	3,175	76.3	329	52.9
	With no extrinsic match	2,766	36.3	1,084	9.4
S. lycopersicum	Fully extrinsic	108,024	98.4	110,645	98.3
	Partially extrinsic	25,610	75.6	26,784	72.0
	With extrinsic match	5,507	59.5	4,893	47.3
	With no extrinsic match	11,112	17.0	8,799	12.5
D. rerio	Fully extrinsic	156,781	97.6	156,506	98.1
	Partially extrinsic	102,256	70.6	105,941	69.8
	With extrinsic match	9,398	34.4	7,360	27.4
	With no extrinsic match	43,023	2.5	40,983	1.9
G. gallus	Fully extrinsic	129,144	98.2	126,410	98.2
	Partially extrinsic	50,046	75.1	53,784	75.2
	With extrinsic match	2,968	31.2	3,008	25.4
	With no extrinsic match	33,168	0.7	33,111	0.6
M. musculus	Fully extrinsic	141,520	99.1	143,186	99.3
	Partially extrinsic	55,236	73.0	55,394	72.5
	With extrinsic match	5,202	49.8	5,063	43.1
	With no extrinsic match	61,229	2.1	58,337	1.2

Supplemental Table S3. Gene and exon level prediction accuracy of GeneMark-ETP with and without filtering of gene predictions with no extrinsic match. The larger F1 values are shown in bold. The gene predictions with no extrinsic match were removed from the GeneMark-ETP outputs for genomes longer than 300 Mbp (the bottom four genomes). For each species, the results are shown for the smaller (order excluded) and larger (species excluded) databases of reference proteins.

		Smaller protein DB		Larger protein DB	
		All intermediate predictions	Output	All intermediate predictions	Output
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**

Supplemental Table S4. 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 Data Sets section).

		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		47.2	67.0	**88.2**	72.7	**90.2**
	Gene Sp		37.4	51.3	**81.4**	54.8	**79.8**
	Gene F1		41.7	58.1	**84.7**	62.5	**84.6**
	Exon Sn		83.5	90.5	**96.7**	92.1	**97.2**
	Exon Sp		74.2	80.0	**92.6**	80.7	**91.6**
	Exon F1		78.6	84.9	**94.6**	86.0	**94.3**
D. rerio	Gene Sn		20.4	35.7	**72.7**	39.6	**73.8**
	Gene Sp		7.5	13.3	**56.5**	14.7	**56.8**
	Gene F1		11.0	19.4	**63.6**	21.4	**64.2**
	Exon Sn		79.1	84.9	**93.6**	86.2	**94.0**
	Exon Sp		50.3	55.9	**85.1**	56.5	**85.1**
	Exon F1		61.5	67.4	**89.2**	68.2	**89.3**
G. gallus	Gene Sn		2.4	14.1	**78.0**	14.4	**77.5**
	Gene Sp		1.4	11.3	**67.2**	11.6	**65.9**
	Gene F1		1.8	12.6	**72.2**	12.9	**71.2**
	Exon Sn		15.1	28.7	**95.4**	29.0	**95.4**
	Exon Sp		27.0	53.4	**90.7**	53.8	**90.3**
	Exon F1		19.3	37.3	**93.0**	37.7	**92.8**
M. musculus	Gene Sn		7.8	22.0	**71.3**	23.7	**72.7**
	Gene Sp		5.4	15.0	**66.0**	16.0	**65.9**
	Gene F1		6.4	17.8	**68.6**	19.1	**69.1**
	Exon Sn		49.7	57.3	**91.2**	58.1	**91.7**
	Exon Sp		50.9	64.2	**90.7**	64.8	**90.7**
	Exon F1		50.3	60.6	**91.0**	61.3	**91.2**

Supplemental Table S5. Comparison of gene- and exon-level prediction accuracy between RNA-seq-based BRAKER1, protein-based BRAKER2, TSEBRA (a tool combining BRAKER1 and BRAKER2), and GeneMark-ETP. Note that the low accuracy in genomes of *G. gallus* and *M. musculus* observed for BRAKER1, BRAKER2, and TSEBRA could be explained in part by the use of a single statistical model for the genome-wide gene prediction rather than the use of the local GC-specific models as in GeneMark-ETP.

		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**

Supplemental Table S6. Performance of MAKER2 and GeneMark-ETP on the genomes of the three model species. The results shown for MAKER2 are supposed to give upper bounds with respect to the freedom of choosing the methods of MAKER2 training. The protein databases are described in Section S3 of Suppl. Materials.

		D. melanogaster		*D. rerio*		*M. musculus*	
		MAKER2	GeneMark-ETP	MAKER2	GeneMark-ETP	MAKER2	GeneMark-ETP
exon	Sn	75.2	**80.7**	83.3	**93.9**	79.2	**91.7**
	Sp	74.0	**91.4**	79.2	**84.9**	77.4	**87.9**
	F1	74.6	**85.7**	81.2	**89.2**	78.3	**89.8**
gene	Sn	60.2	**79.0**	47.7	**73.5**	41.6	**73.1**
	Sp	55.3	**83.0**	37.6	**56.2**	34.8	**59.7**
	F1	57.7	**81.0**	42.0	**63.7**	37.9	**65.7**

Supplemental Table S7. Genomic sequences and annotations used in the tests. A date in parenthesis is the date of the last update prior to the data use.

Species	Assembly version	Main annotation	Supplementary annotation*
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**

*Supplementary annotation was used to produce the 'set of reliable genes' by comparison of two annotations and selecting identically annotated genes. **The subset for *M. musculus* genes identical between Ensemble and NCBI was selected by choosing a subset of the GENCODE transcripts with the following attributes: *CCDS* (Agreement with the 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).

Supplemental Table S8: Selection of proteins from OrthoDB v10.1. Numbers in bold black font show the number of species in the largest OrthoDB segment, IP$_{0,}$ considered for a given species. Numbers in bold blue font show the number of species excluded from IP$_0$ in the 'Order excluded' segment of OrthoDB. The 'Species excluded' segment of OrthoDB constitutes proteins in IP$_0$ but those from the species of interest itself.

Species	# of species in the OrthoDB clade						Name of the OrthoDB segment	# of proteins in the OrthoDB segment (M)
	Genus	Family	Order	Class	Phylum	Kingdom		
C. elegans	3	3	**5**	6	7	**448**	Metazoa	8.3
A. thaliana	2	8	**10**	-	100	**117**	Plantae	3.5
D. melanogaster	20	20	**56**	148	**170**	-	Arthropoda	2.6
S. lycopersicum	2	10	**11**	-	100	**117**	Plantae	3.5
D. rerio	1	5	**5**	50	**246**	-	Chordata	5.0
G. gallus	1	3	**4**	62	**246**	-	Chordata	5.0
M. musculus	3	5	**20**	**111**	-	-	Mammalia	2.3

Supplemental Table S9. RNA-seq libraries used for computational experiments.

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

Supplemental Table S10. Numbers of GeneMark-ETP predicted genes and transcripts including alternative transcripts.

Species	GeneMark-ETP with order excluded DB			Reference annotation statistics		
	# coding genes	# coding transcripts	introns per transcript	# coding genes	# coding transcripts	introns per gene
C. elegans	18,820	19,806	5.4	19,969	28,544	4.8
A. thaliana	26,449	27,708	4.2	27,445	40,827	4.0
D. melanogaster	12,850	14,138	2.9	13,951	22,395	2.8
S. lycopersicum	24,420	26,341	4.3	25,158	31,911	4.4
D. rerio	28,608	31,961	7.5	25,610	42,929	8.4
G. gallus	17,275	21,433	7.7	17,279	38,534	9.0
M. musculus	23,956	27,686	6.7	22,405	58,318	6.0

Supplemental Table S11. Analysis of the correctness of the results of *re-classification* of the GeneMarkS-T predicted genes as complete/partial conducted by comparison with annotation. The genes used in this analysis were i/ predicted partial (incomplete) by GeneMarkS-T, ii/ had correctly predicted stop codon, and iii/ were verified for absence of assembly errors. The results are shown for the smaller protein database (Order excluded) and for the larger one (Species excluded) (see Data Sets in Main text).

		Smaller protein DB			Larger protein DB		
		Total incomplete predictions	True Complete	True Partial	Total incomplete	True Complete	True Partial
		2,095			2,924		
C. elegans	Re-classified complete		1,488	127		1,982	78
	Re-classified partial		273	207		393	471
		1,841			1,878		
A. thaliana	Re-classified complete		1,476	55		1,442	22
	Re-classified partial		107	203		165	249
		651			826		
D. melanogaster	Re-classified complete		273	76		299	9
	Re-classified partial		48	254		130	388
		1,381			1,408		
S. lycopersicum	Re-classified complete		897	81		868	63
	Re-classified partial		81	322		119	358
		2,750			2,803		
D. rerio	Re-classified complete		1,152	107		1,052	69
	Re-classified partial		249	1,242		364	1,318
		4,727			4,738		
G. gallus	Re-classified complete		3,232	197		2,972	114
	Re-classified partial		449	849		715	937
		2,744			2,745		
M. musculus	Re-classified complete		2,026	16		1,879	8
	Re-classified partial		497	205		642	216

Supplemental Table S12. The values of Sn and Sp are determined for i/ a set of initial GeneMarkS-T predictions – complete and partial and ii/ a set of high-confidence (HC) genes, obtained with the use of Condition S1. The Sn and Sp are shown separately for the complete and partial GeneMarkS-T predictions as well as for complete and partial HC genes. The true positive prediction of a partial gene is called if the partial prediction coincides with a part of a gene in the reference annotation. The Sn and Sp of the HC genes are shown for the smaller protein database (Order excluded) and for the larger one (Species excluded) (see Data Sets). There is a significant increase in the Sp values of the partial genes that occurred in transition from the GeneMarkS-T to the HC genes. Note that for majority of the species the use of a larger protein database does not improve the Sp of the partial genes.

		GeneMarkS-T predictions		HC genes (processed by GeneMarkS-TP)			
				Smaller protein DB		Larger protein DB	
		Complete	Partial	Complete	Partial	Complete	Partial
C. elegans	Sn	42.9	3.9	33.6	2.1	47.7	4.0
	Sp	82.0	18.2	88.8	81.5	91.5	80.7
A. thaliana	Sn	49.8	1.4	55.6	1.1	57.3	1.6
	Sp	89.1	17.0	97.4	92.3	97.8	90.8
D. melanogaster	Sn	56.4	3.2	53.3	1.8	60.6	3.1
	Sp	87.5	38.1	95.0	85.3	96.9	85.0
S. lycopersicum	Sn	66.3	1.4	73.7	1.3	74.2	1.5
	Sp	84.1	26.6	95.4	87.2	95.4	84.8
D. rerio	Sn	55.3	4.3	62.8	4.2	62.4	4.5
	Sp	68.4	32.8	89.7	78.9	92.8	75.3
G. gallus	Sn	43.9	5.7	67.9	6.5	66.3	7.7
	Sp	64.0	23.0	89.5	86.1	90.0	80.3
M. musculus	Sn	48.4	1.2	60.8	2.7	60.5	3.4
	Sp	80.4	9.6	95.1	68.0	96.7	69.8

Supplemental Table S13. The values of the genome specific masking penalty parameter *q* (in natural logarithms). For GC-heterogeneous genomes, the optimal *q* value was estimated for each GC bin. Descriptions of the smaller (Order excluded) and larger (Species excluded) protein databases are given in the Data Sets section.

	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

References

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

Kozak M. 1987. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* **15**: 8125-8148.

Kozak M. 1999. Initiation of translation in prokaryotes and eukaryotes. *Gene* **234**: 187-208.

Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM. 2014. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15**: 419.