GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes

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

[*] To whom correspondence should be addressed. Tel: +1 404 894 8432; Email: borodovsky@gatech.edu

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Supplemental Material

Supplemental Methods

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

S1.1 Corrections of the 5' end gene predictions

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 CDSs, those having start codons within transcripts. On the other hand, the 5' incomplete CDSs predicted by GeneMarkS-T with the start codons residing near the first nucleotide of a transcript carry more frequent errors that should be corrected. We need to discriminate between a correctly predicted 5' incomplete CDS and an incorrect 5' incomplete CDS with a true complete CDS residing inside.

Incomplete CDSs 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 CDS, *the 5' partial CDS is predicted*.

Otherwise, the CDS starting with the internal ATG is selected as the *predicted complete CDS*. 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, then the transcript is removed from consideration.

The quantitative meaning of the *better support* provided by the protein alignment data is formalized by 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 (see Supplemental Fig. S10 and S11). $AAI_{partial}$ and $AAI_{complete}$ are, respectively, the percentages of *amino acid identities* in the alignments of the longer and shorter query proteins with 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. S10 and S11).

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

Notably, "a-1" is the length of the unaligned N proximal part of the long query.
A large "a-1" is likely to indicate the presence of a translated 5' UTR region situated upstream of 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 for 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 in the *AAI* ratio favors the 5' partial candidate. The *AAI* ratio is scaled using a logarithm with a factor of 1,000, i.e., 1,000*log(…).

S1.2 Removal of the 3' partial CDS predictions

The 3' partial predictions were rarely observed. This low frequency could be expected since RNA-Seq libraries used in our experiments, prepared with the poly-A tail enrichment of mRNA transcripts, should predominantly carry 3' end complete transcripts (Zhao et al. 2014). Therefore, all the 3' partial genes were removed from the list of candidates for high-confidence CDSs.

S1.3. Extensions of GeneMarkS-T CDS 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 predictions of CDSs with non-5'-most start codons carry a higher false-positive rate than the predictions of CDSs with 5'-most start codons. Therefore, we use the following rule. If a CDS

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 predicted CDS is considered a candidate for an HC CDS along with the one with non-5'-most start.

S1.4 Complete genes with uniform protein support

In the described above similarity searches we have dealt with local pairwise alignments. Still, being interested in the accurate prediction of all protein-coding exons, we are concerned about *uniform* protein support showing evolutionary conservation over the whole protein-coding region. We say that a *uniform protein support* exists for a predicted *complete* CDS if there is a significant BLASTp alignment (with E-value better than 10^{-3}) of the translation of the predicted CDS *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. S9).

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. Predicted CDS is called a complete CDS with uniform protein support if a translated query 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 CDSs are included in the set of high-confidence CDSs.

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 CDSs. 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 CDSs (Condition S1) as well as selecting CDSs with uniform protein support (Condition S2) with the 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 CDS 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 the use of conditions S1 and S2 produced

more accurate results than the results generated by the 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 predicted CDSs not uniformly supported by proteins (and not satisfying Condition S2) could still be included in the set of HC CDSs. 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 CDS predicted by GeneMarkS-T is mapped to genomic DNA. Second, the translation of the initially predicted CDS and its genomic locus is used by ProtHint as the protein and CDS seeds to generate hints for the next round of CDS 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 CDSs

An additional round of selection is made to subject CDSs that satisfy Condition S2 to a more stringent restriction. Let $I^g_{complete}$ be a set of complete alternative CDSs of protein-coding gene g and $I^g_{partial}$ is a set of its alternative partial CDSs. 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 CDS isoforms for a gene g, denoted as $\mathrm{s}(g_{complete})$. A score of a CDS isoform $\mathrm{s}(i)$ selected as complete HC CDS isoform must satisfy the inequality:

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

Among the partial alternative CDSs 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 CDS 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 CDSs of gene g are retained.

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

The numbers of predicted alternative CDSs are smaller than the numbers of annotated alternative CDSs (Table 3), because we predict alternative CDSs only for the HC genes, a subset of all genes. Moreover, the CDS isoforms of the HC genes must have full protein support (Condition S3) which further limits the number of predicted CDS isoforms.

S4. Computing the species-specific repeat penalty parameter

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

We have the set of the HC CDSs, 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 CDSs in the genomic sequences containing the HC CDSs for which we compute the gene level F1 value (Supplemental Fig. S12-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 CDSs is made based on genome annotation. We also observed that the value *q* was robust with respect to the size of the HC CDSs 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. S12-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. S12-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 having 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 available to MAKER2.

Repeat coordinates predicted by RepeatMasker in the MAKER2 supported GFF format:

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

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

As a protein database input for both MAKER2 and GeneMark-ETP we used the OrthoDB proteins as follows:

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

For *Mus musculus*:

AUGUSTUS – the "human" model from the AUGUSTUS distribution.
GeneMark.hmm – the 'medium GC' model created by GeneMark-ETP for the *Mus musculus* genome.

SNAP – the “mam46.hmm” mammalian model for the medium GC bin from SNAP distribution.

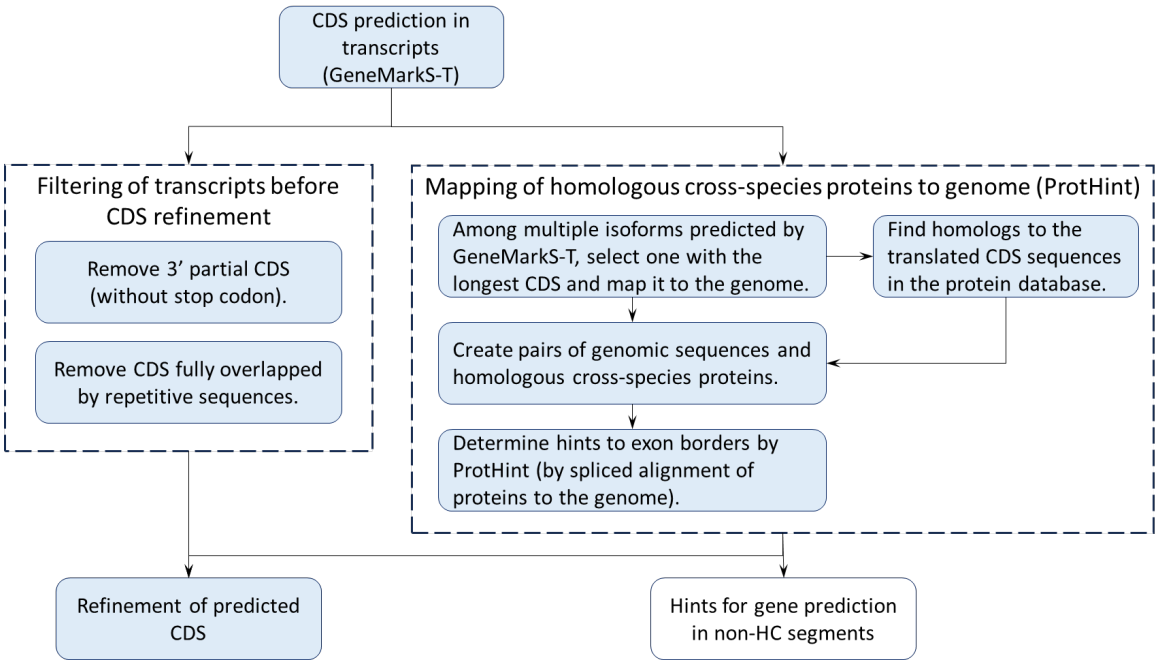
MAKER2 was executed with the following settings in the MAKER2 configuration file:

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

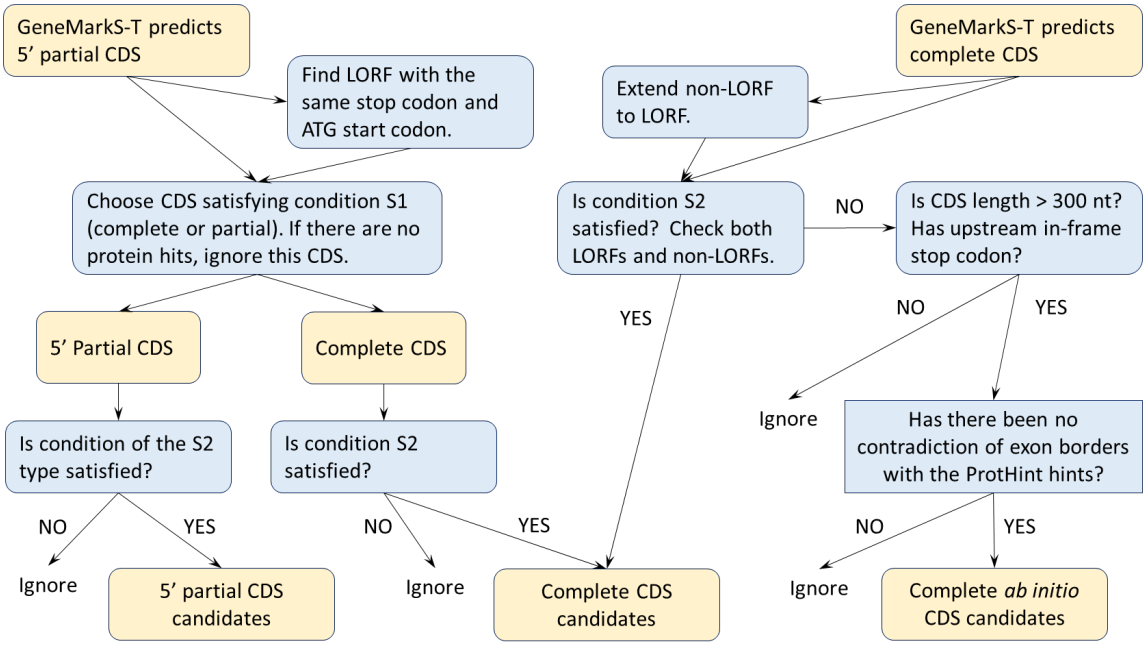
A LINUX node with 96 cores was used to execute MAKER2 in MPI mode at the Azure cloud.

The gene prediction accuracy of MAKER2 and GeneMark-ETP (Supplemental Table S5) was estimated as described in the main text (see Methods).

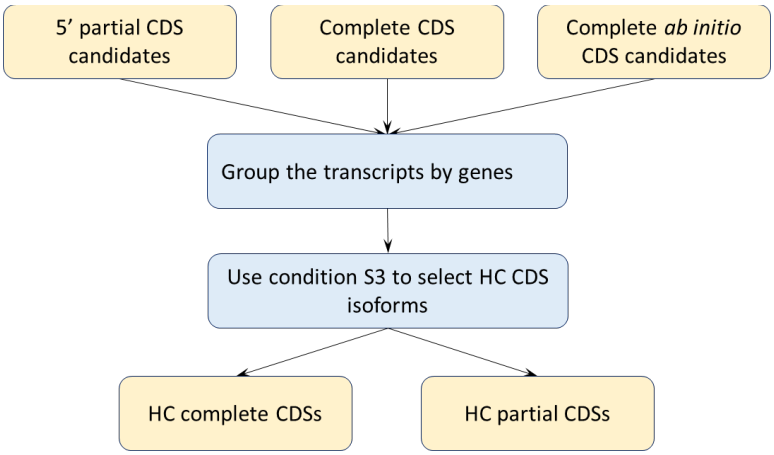
Supplemental Figures



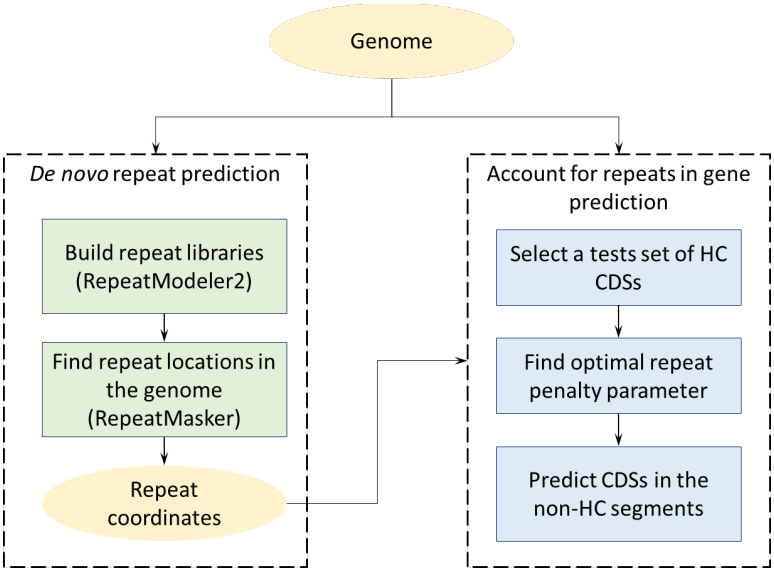
Supplemental Figure S1. Schematics showing the transcript processing steps in GeneMarkS-TP (see Fig. 2).



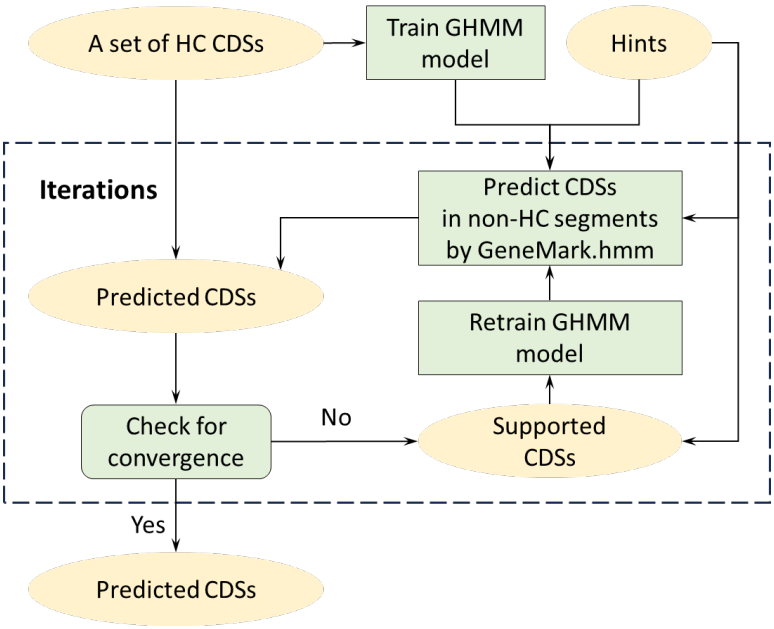
Supplemental Figure S2. Schematics of the generation of HC CDS candidates in GeneMarkS-TP (the refinement block in Fig. 2).



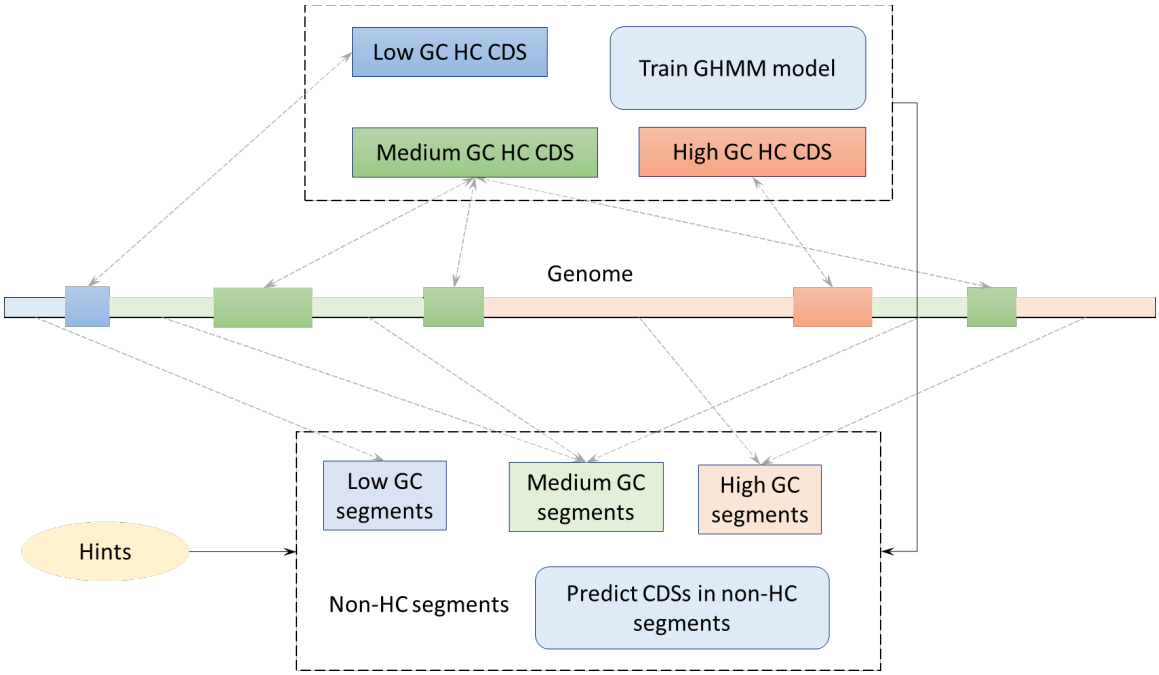
Supplemental Figure S3. Schematics of the selection of HC CDSs in GeneMarkS-TP (see Fig.2).



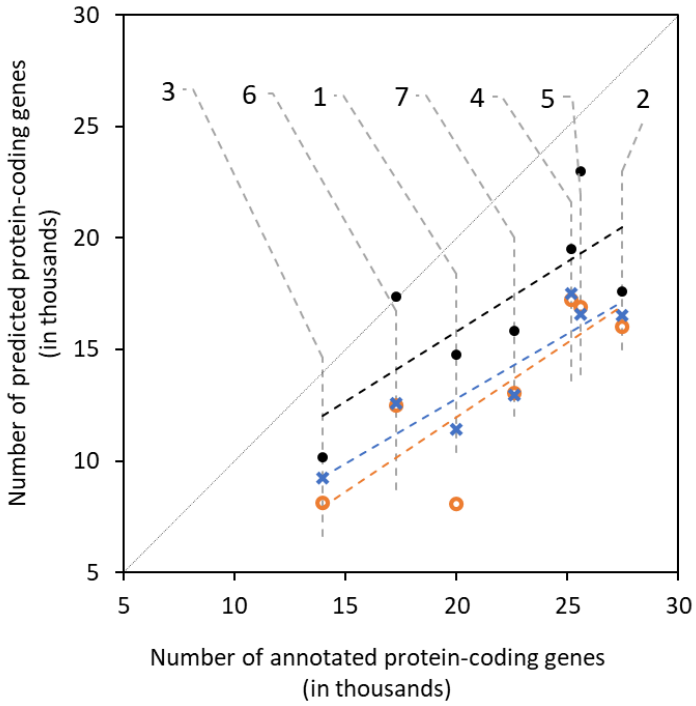
Supplemental Figure S4. Schematics of the repetitive sequence identification and processing. *De novo* repeats prediction module (shown on the left) is not a part of GeneMark-ETP (see Fig.1).



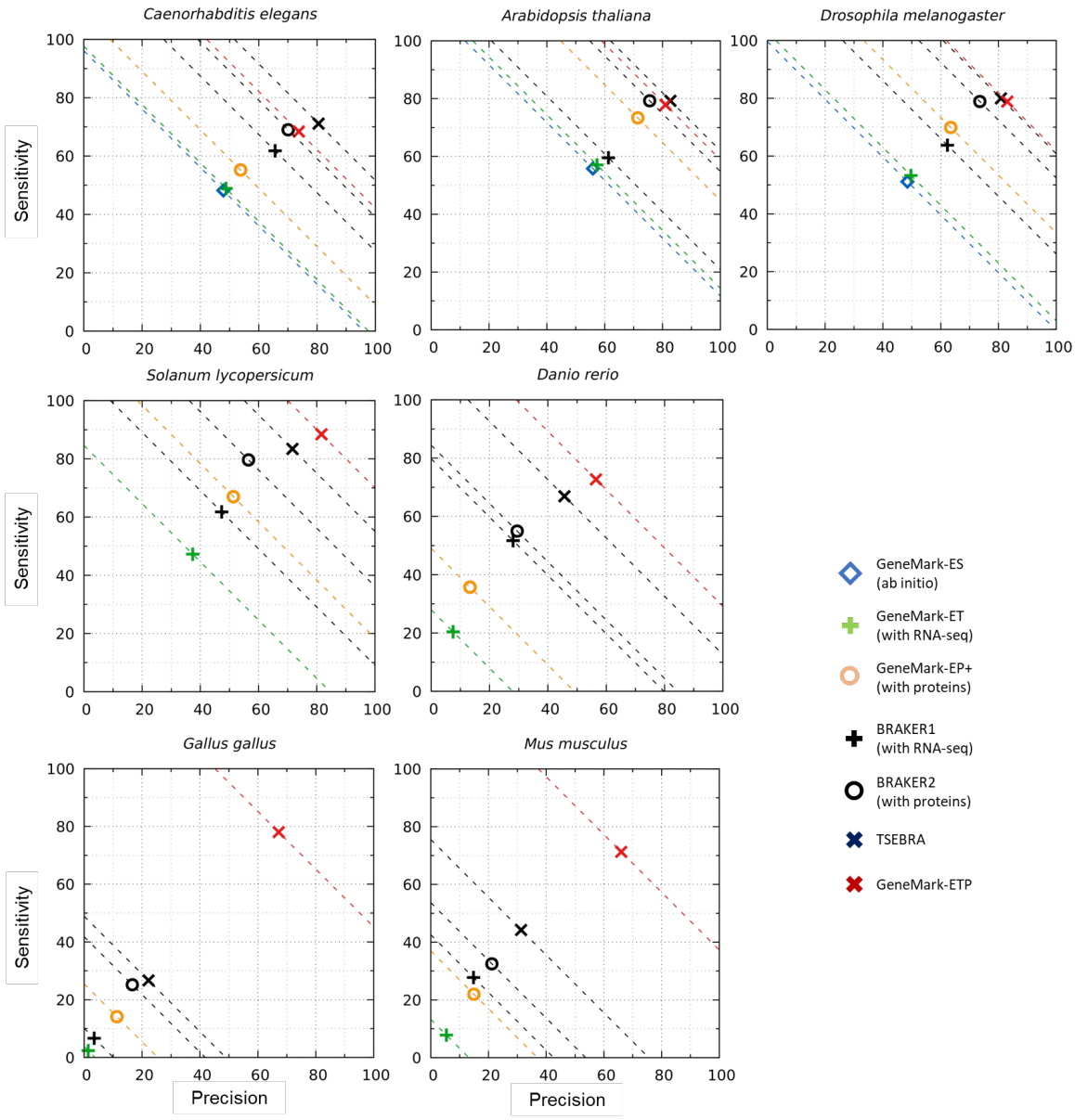
Supplemental Figure S5. Workflow of the training of the GHMM model used in GeneMark.hmm (see Fig.1).



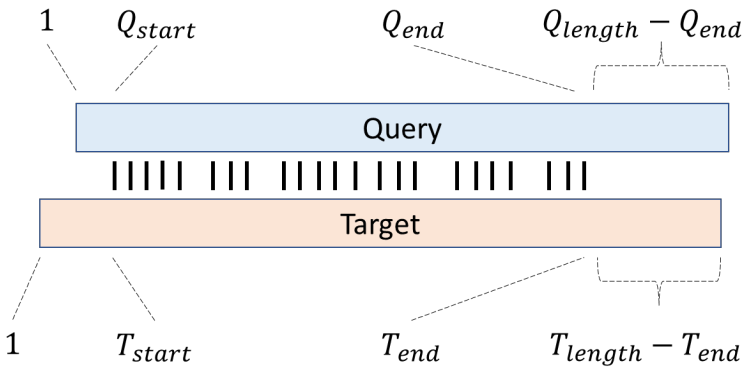
Supplemental Figure S6. Schematics of the identification and the use of the non-HC segments in the GHMM training and CDS prediction (see Fig.1).



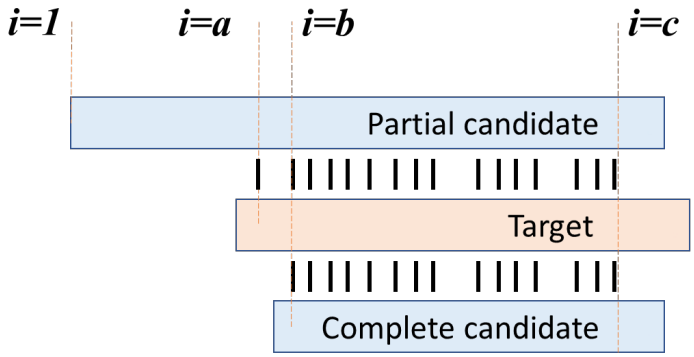
Supplemental Figure S7. Numbers of protein-coding genes predicted at initial stages of running GeneMark-ETP (i) genes predicted in assembled transcripts by GeneMarkS-T (black dots), (ii) HC genes predicted by GeneMarkS-TP with the 'Order excluded' protein database (orange circles) and with the 'Species excluded' database (blue crosses). The number of genes annotated in each genome is taken from the RefSeq annotation (Supplemental Table S7). The numerical designations of the species are as follows: 1 - *C. elegans*, 2 - *A. thaliana*, 3 - *D. melanogaster*, 4 - *S. lycopersicum*, 5 - *D. rerio*, 6 - *G. gallus*, 7 - *M. musculus*.



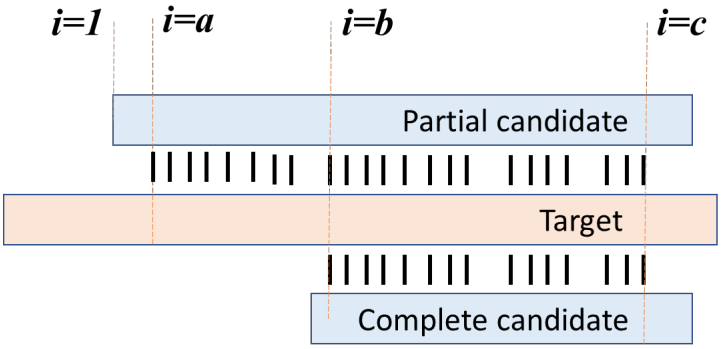
Supplemental Figure S8. Gene level accuracy of the seven gene prediction tools (see legends to Figs. 3-4). Compared to the figures in the main text, where we used the 'Order excluded' protein databases for each species, here we used the larger 'Species excluded' databases.



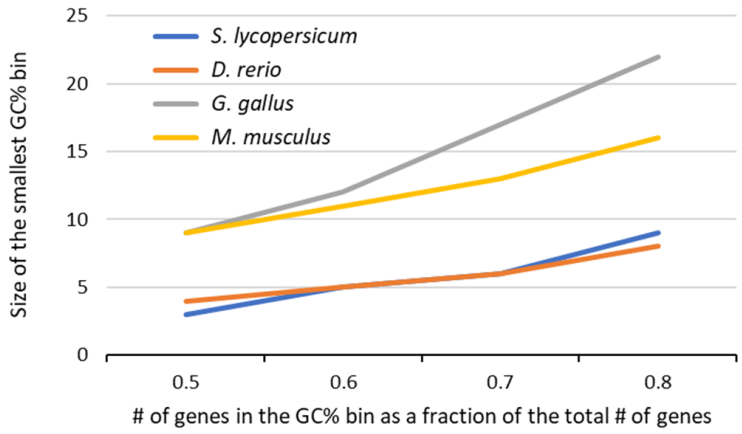
Supplemental Figure S9. Depictions of the alignment features used in Condition S2.



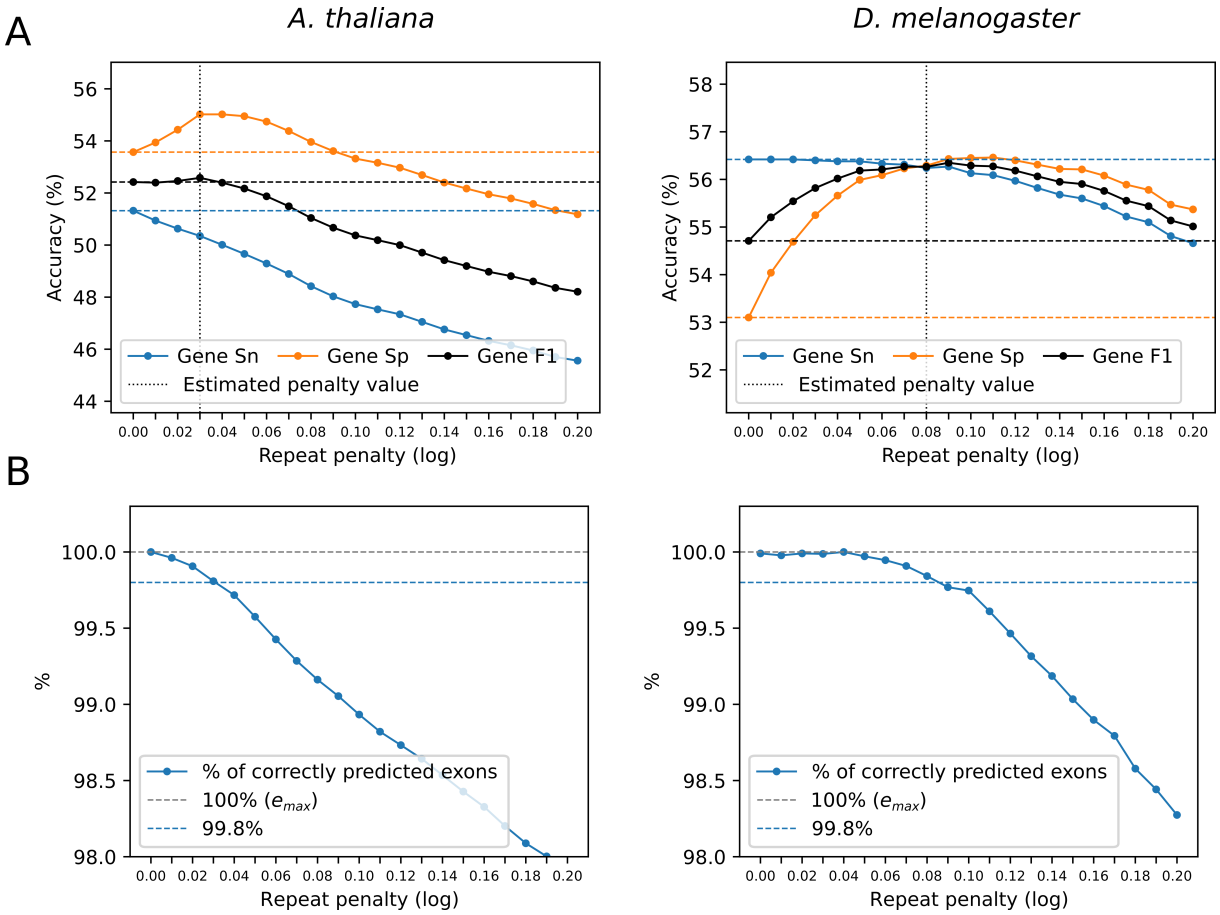
Supplemental Figure S10. Illustration for the case when Condition S1 is not fulfilled, and the GeneMarkS-T prediction is classified as *complete CDS*. Here *a* and *b* are positions of the starts of the local alignments of respective longer and shorter protein queries, while *c* is the end position of the local pairwise alignments.



Supplemental Figure S11. Illustration for the case when Condition S1 is fulfilled, and the GeneMarkS-T gene prediction is classified as a 5' *partial CDS*. Here *a, b* and *c* are defined as in Supplemental Fig. S9.



Supplemental Figure S12. Results of analysis of the genome GC content inhomogeneity. For each genome, the graphs show the sizes of the narrowest GC% bin in the genome-specific GC content distribution (Y axis) that would contain the number of genes corresponding to a fixed fraction of the total number of annotated genes (X axis). It can be seen from the graph that the *G. gallus* genome is the most GC heterogeneous, followed by the *M. musculus* genome. The remaining two genomes are GC homogeneous: 80% of the whole gene complement can be placed into the GC bin with 10% width.



Supplemental Figure S13. A. Dependence of the gene level Sn, Pr, and F1 values (determined for the full sets of HC CDSs) 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 CDSs (Sn) on the repeat penalty parameter *q* for the same genomes as in A. (See Section S4 of Suppl. Materials)

References

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

Kirkpatrick S, Gelatt CD, Vecchi MP. 1983. Optimization by Simulated Annealing. *Science* **220**: 671-680.

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.

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

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.