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

* 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.

Abstract

GeneMark-ETP is a computational tool developed to find genes in eukaryotic genomes in consistence with genomic-, transcriptomic- and protein-derived evidence. Earlier developed GeneMark-ET and GeneMark-EP+ addressed more narrow tasks of data integration, working either with fragments of transcripts (short RNA reads) or with homologous protein sequences.

Both the transcript- and protein-derived evidence have uneven distribution across a genome. Therefore, GeneMark-ETP finds the genomic loci where extrinsic data helps to identify genes with "high confidence" and then proceeds with the analysis of the regions between the high-confidence genes. If the number of high-confidence genes is sufficiently large, the GHMM model training is done in a single-iteration. Otherwise, several iterations of self-training are necessary prior to making the final prediction of the whole gene complement. Since the difficulty of the gene prediction task ramps up significantly in large plant and animal genomes, the focus of the new development was on large genomes.

The GeneMark-ETP performance was favorably compared with the ones of GeneMark-ET, GeneMark-EP+, BRAKER1 and BRAKER2, the methods using a single type of extrinsic evidence. Comparison was also made with TSEBRA, a tool constructing an optimal combination of gene predictions made by BRAKER1 and BRAKER2, thus utilizing both transcript- and protein-derived evidence.

Introduction

Computational methods of gene identification in novel eukaryotic genomes use a combination of intrinsic and extrinsic evidence. While the intrinsic evidence, the genome-specific patterns of nucleotide ordering, are sufficient for accurate gene prediction in genomes of fungi and protists (Lomsadze et al. 2005; Ter-Hovhannisyan et al. 2008) the extrinsic evidence plays a critical role in accurate gene prediction in large eukaryotic genomes (Mudge and Harrow 2016). Strictly extrinsic evidence-based approaches, using either a protein space, such as *exonerate* (Slater and Birney 2005), GenomeThreader (Gremme et al. 2005), or ProSplign (Kiryutin et al. 2007), or the transcript space, such as StringTie (Pertea et al. 2015; Kovaka et al. 2019), PsiCLASS (Song et al. 2019), and Cufflinks (Trapnell et al. 2010), are limited to finding subsets of the whole gene complement. When the protein-based evidence is used, this would be a subset of genes whose protein products show detectable similarity to cross-species orthologs; for the transcripts-based evidence, this would be a subset of genes with sufficiently high expression. Recently developed GeMoMa (Keilwagen et al. 2018) combined RNA-Seq based with protein-based predictions and delivered quite accurate results for genomes of the species whose close relatives have sequenced and well-annotated genomes.

In absence of extrinsic evidence, a gene finding algorithm fully relies on models combining intrinsic (genomic sequence derived) features such as k-mer frequency patterns, splice site motifs, intron/exon length distributions, etc. Several efficient *ab initio* algorithms were developed at a time when the volume of extrinsic evidence was rather small (e.g., Genie (Kulp et al. 1996), GENSCAN (Burge and Karlin 1997), GeneID (Parra et al. 2000), SNAP (Korf 2004), AUGUSTUS (Stanke and Waack 2003), GeneMark-ES (Lomsadze et al. 2005)). The *ab initio* methods were observed to be less accurate for the large eukaryotic genomes that carry long non-coding regions, leaving an ample space for false positive predictions (Guigo et al. 2006; Coghlan et al. 2008; Goodswen et al. 2012; Scalzitti et al. 2020). More recently developed gene finders have relied on intrinsic evidence as well as extrinsic evidence. AUGUSTUS (Stanke et al. 2008), GeneMark-ET (Lomsadze et al. 2014) and BRAKER1 (Hoff et al. 2016) are examples of tools integrating genomic and transcript data. On the other hand, AUGUSTUS-PPX (Keller et al. 2011), GeneMark-EP+ (Bruna et al. 2020) and BRAKER2 (Bruna et al. 2021) integrate genomic features with evidence obtained from mapped to genome cross-species proteins.

The majority of the existing tools integrating all the three sources of evidence work as *combiners,* e.g., FINDER (Banerjee et al. 2021), LoReAn (Cook et al. 2019), GAAP (Kong et al. 2019), IPred (Zickmann and Renard 2015) Evigan (Liu et al. 2008), EVidenceModeler (Haas et al. 2008), JIGSAW (Allen and Salzberg 2005), Combiner (Allen et al. 2004), or GAZE (Howe et al. 2002). A combiner first generates independent sets of genome-wide gene predictions: *ab initio*-, transcriptomic- and protein mapping- based. Next, at a post-processing step, these sets are combined into a final set of predictions.

An alternative approach would integrate the three sources of evidence upon the prediction of each gene. For a self-training algorithm—one working without an expert-defined training set—the integration would be included into the cycles of iterative model training and gene prediction.

Here we introduce GeneMark-ETP, a new computational tool integrating genomic, transcriptomic, and protein information throughout *all* the stages of the algorithm's training and gene prediction. This integration is facilitated upon gene prediction in long transcripts assembled from RNA-Seq reads and supported by verification of the consistency of protein and transcript information. The estimation of parameters of the statistical models (Generalized Hidden Markov Models, GHMM) used in GeneMark-ETP is done by unsupervised training. Protein based evidence, producing hints to locations of introns and exons in genomic DNA, is generated by using homologous proteins of any evolutionary distance, including remote homologs. Accurate accounting for DNA sequence repeats plays a significant role as well.

Tests of the GeneMark-ETP performance were done on both compact and large, GC-homogeneous and GC-heterogeneous eukaryotic genomes. The results were compared with performances of GeneMark-ET, GeneMark-EP+ as well as their virtual combination. The performance of GeneMark-ETP was also compared with the performances of the pipelines BRAKER1 and BRAKER2 as well as with TSEBRA (Gabriel et al. 2021), a recently developed tool combining the BRAKER1 and BRAKER2 predictions.

Results

Assessment of the GeneMark-ETP prediction accuracy

The gene prediction accuracy of GeneMark-ETP (Fig. 1) was assessed for seven genomes representing diverse genomic organizations and taxonomic clades: *Arabidopsis thaliana*, *Caenorhabditis elegans, Drosophila melanogaster, Solanum lycopersicum*, *Danio rerio*, *Gallus gallus* and *Mus musculus*. For the three shorter (compact) genomes *A. thaliana*, *C. elegans, D. melanogaster* we accepted genome annotation as the ground truth. For the four large genomes, *S. lycopersicum*, *D. rerio*, *G. gallus* and *M. musculus* estimations of the gene prediction sensitivity (Sn) were computed for genes present in both NCBI and Ensemble annotations (see Methods) while the gene prediction specificity (Sp) was computed in comparison with the union of the NCBI and Ensemble annotations. We have observed a significant increase in accuracy in comparison with both GeneMark-ET and GeneMark-EP+. Moreover, the improvement was reached also in comparison with both BRAKER1 and BRAKER2 (Figs. 2, 3, S1; Tables S1, S2). The most notable improvements occurred in large genomes, especially the GC-heterogeneous ones. For the groups of compact, large homogeneous, and large heterogeneous genomes, the GeneMark-ETP gene level F1 values increased on average over GeneMark-EP+ by 14.1%, 33.6%, and 55.3%, respectively (Table S1), while the exon level F1 values for the same three groups of genomes increased on average by 5.2%, 15.4%, and 43.2%, respectively (Table S1).

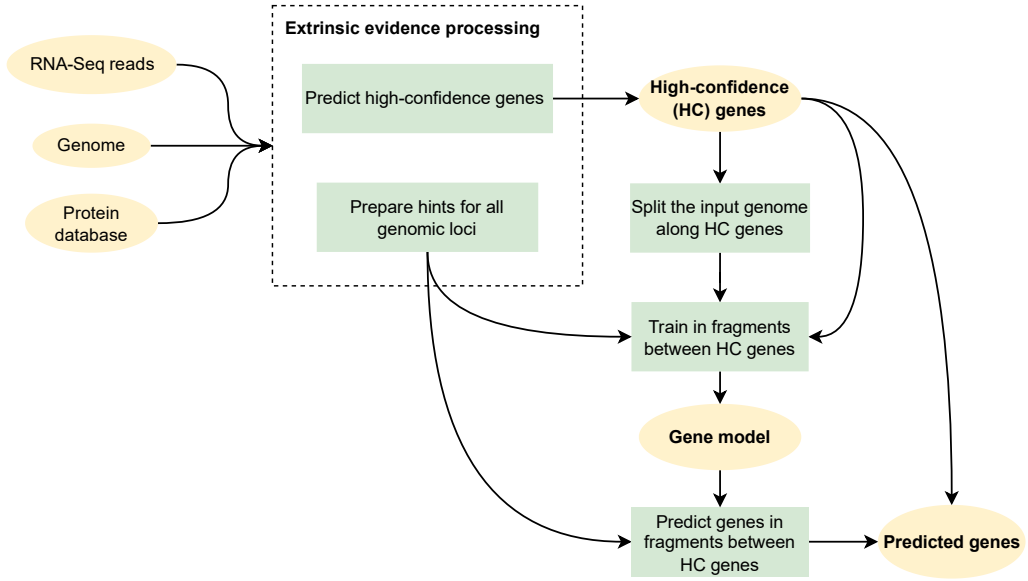

Figure 1. High-level diagram of the GeneMark-ETP algorithm

In comparison with TSEBRA (Figs. 2, 3, S1; Table S2), the average gene level F1 values increased, respectively, by 2.2%, 8.3%, and 39.5% (Table S2) while the average exon F1 values by 0.6%, 1.0%, and 19.1% (Table S2).

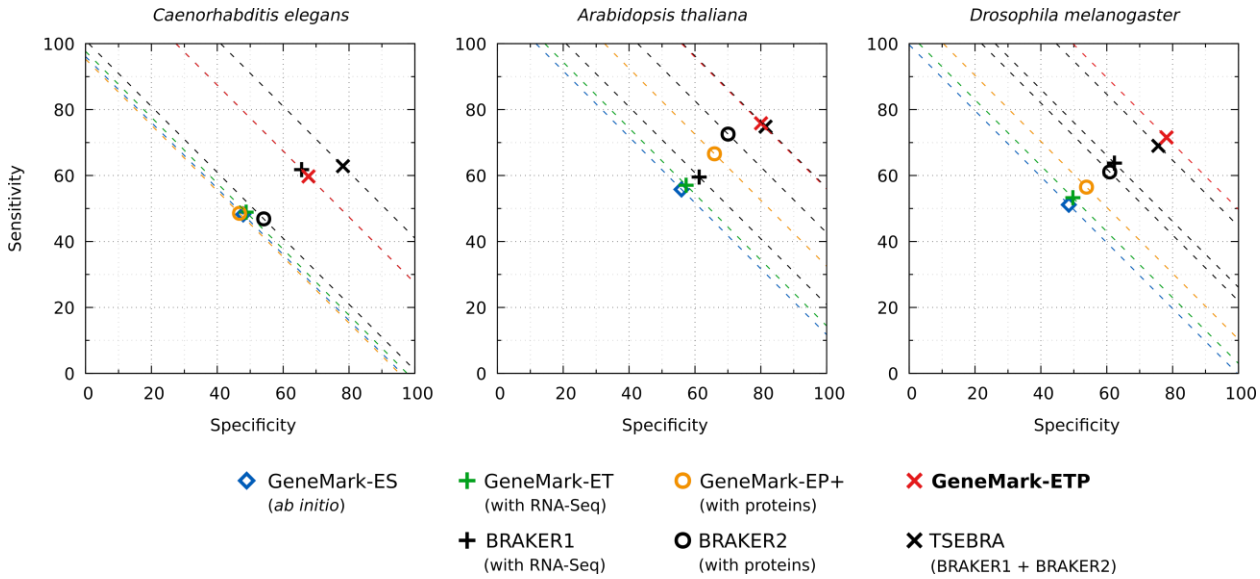

Figure 2. Gene level Sensitivity (Sn) and Specificity (Sp) of GeneMark-ETP. Sn = Tp/(Tp+Fn) and Sp = Tp/(Tp+Fp) where Tp, Fp and Fn are the numbers of true positive, false positive and false negative gene predictions, respectively. The dashed lines correspond to constant levels of (Sn+Sp)/2. The species-specific protein databases used for derivation of protein-based evidence did not include proteins originated from the species from the same taxonomic order.

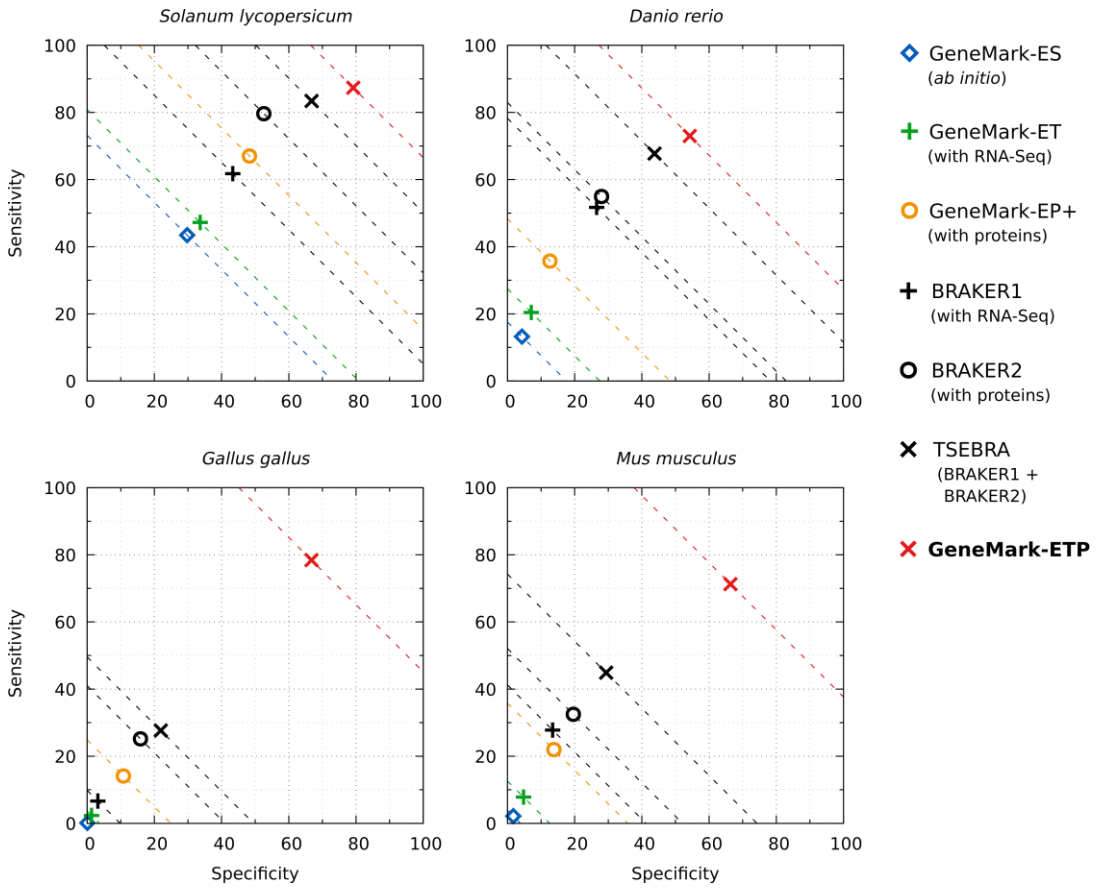

Figure 3. Gene level Sn and Sp of GeneMark-ETP for the four larger genomes. All other specifications are the same as in Fig. 1.

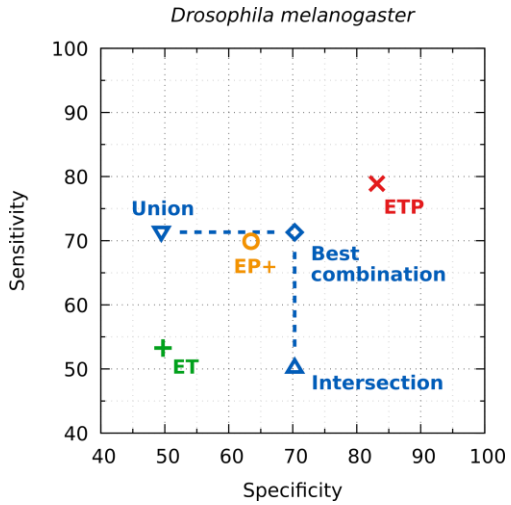

Figure 4*:* Gene-level Sn and Sp of the artificial combinations of GeneMark-ET and GeneMark-EP+ gene predictions made in genome of *D. melanogaster* are shown along with the Sn and Sp of GeneMark-ETP. Proteins of the same species were excluded from the reference protein database.

For both compact and large genomes, the accuracy of GeneMark-ETP was observed to be significantly higher than the accuracy of the virtual combinations of the sets of gene predictions made by GeneMark-ET and GeneMark-EP+ separately, the union, the intersection or the 'best' combination (Figs. 4, S2, Methods Section 5.2).

Refinement of the gene predictions in assembled transcripts

To produce a set of high-confidence genes we relied on gene predictions made by GeneMarkS-T (Tang et al. 2015) made in assembled transcripts. Some of these predictions were refined if the alignments of the predicted proteins with the homologous proteins found in the protein database indicated a possible better match with a modified protein (Methods Section 2.2.2). We assessed the accuracy of these refined predictions (Table 1). We found that for each of the seven species, and species-specific protein databases of the smaller and the larger size (Tables 1, S3), the gene-level specificity increased, on average, by 25 percentage points, reaching values higher than 90%.

Table 1. The gene-level Sn and Sp values of *all* the GeneMarkS-T gene predictions in the assembled transcripts and for those GeneMarkS-T predictions that were selected as high-confidence (HC) genes. The Sn and Sp values in the second column are shown for these HC genes. From the corresponding PD_0 protein databases employed in these analyses proteins from the species of the same taxonomic order as the species in question were excluded.

Species	-	GeneMarkS-T	Final HC
C. elegans	Sn	**47.6**	35.8
	Sp	63.8	**88.4**
A. thaliana	Sn	51.7	**57.0**
	Sp	80.0	**97.3**
D. melanogaster	Sn	**60.5**	55.2
	Sp	82.0	**94.7**
S. lycopersicum	Sn	68.0	**75.1**
	Sp	74.6	**92.8**
D. rerio	Sn	60.5	**67.3**
	Sp	57.2	**84.6**
G. gallus	Sn	49.8	**74.7**
	Sp	43.3	**85.6**
M. musculus	Sn	50.0	**63.7**
	Sp	59.5	**90.4**

When the smaller protein databases were used (protein of the same taxonomic order excluded from the species-specific reference database - PD_0, see Materials) we observed a noticeable increase in gene prediction sensitivity for five of the seven tested genomes (Table S3). For the larger databases, (with only the proteins from *the same* species excluded from PD_0, the increase was observed for all seven genomes (Tables 1, S3). This increase was largely due to the introduction of the refinement of the gene prediction in transcripts. For example, in the case of *D. rerio,* 2,753 out of 22,979 genes predicted in transcripts were initially classified as 5' partial by GeneMarkS-T (Table 2). Comparison with annotation revealed 1,384 truly 5' partial predictions

and 1,369 those that contained true complete gene inside. The refinement shortened 1,159 of the 1,369 wrong predictions (reaching 85% sensitivity in this set). At the same time, this refinement incorrectly shortened 122 genes from the 1,384 true *partial* genes (9% error rate). The results of this type of analysis for all seven genomes are shown in Table S4.

Table 2. Confusion matrix for the complete/5' partial classification of the genes predicted in the *D. rerio* transcripts. Proteins of the species from the same taxonomic order as *D. rerio* were removed from the reference protein database PD_0.

	Complete genes	Partial genes
Predicted complete	1,159	122
Predicted partial	210	1,262

Analysis of the balance of extrinsic and intrinsic evidence

For each of the seven genomes we divided the whole complements of predicted genes into four categories by the type of extrinsic support: *fully extrinsic*: all elements of the exon-intron structure were supported by significant (high scoring) extrinsic evidence; *partially extrinsic*: some elements of the exon-intron structure were determined due to significant extrinsic evidence while other were predicted *ab initio*; *ab initio anchored* which meant that the whole gene was predicted *ab initio,* while a match to a low scoring extrinsic evidence for some gene elements was detected *a posteriori*; *ab initio unsupported*: none of the gene elements predicted *ab initio* were supported by any extrinsic evidence even *a posteriori*.

We observed that the reliability of gene predictions could be reduced significantly upon the decrease of the level of extrinsic support. Particularly, in the three largest genomes, the predictions in the *ab initio unsupported* category had gene level Sp values below 1% (Table 3) and exon level Sp below 3% (Table S5). When we removed such gene predictions from the reported lists of predicted genes we found that in these four genomes, the gene-level Sp increased, on average, by 19.6% while Sn decreased by 0.4% (Table S6). For the three smaller genomes, such a pruning would increase on average the gene level Sp by 3.4% with decrease of Sn by 2.5%.

Table 3. Distribution of predicted genes among the four categories of extrinsic support along with average Sp values (gene level) for each category. Descriptions of the smaller and larger species-specific protein databases are given in Materials.

Species	Types of evidence for a predicted gene	Smaller protein DB		Larger protein DB	
		# of genes	Specificity, %	# of genes	Specificity, %
C. elegans	Fully extrinsic	7,690	88.9	10,833	91.7
	Partially extrinsic	4,832	56.4	5,473	52.9
	Ab initio anchored	3,912	53.4	1,433	42.1
	Ab initio unsupported	1,192	24.5	658	17.6
A. thaliana	Fully extrinsic	16,513	97.2	18,155	97.5
	Partially extrinsic	4,855	63.8	5,838	55.2
	Ab initio anchored	1,787	49.9	1,350	29.8
	Ab initio unsupported	2,857	28.0	1,027	8.2
D. melanogaster	Fully extrinsic	8,076	95.1	10,055	96.4
	Partially extrinsic	2,339	49.5	2,763	44.5
	Ab initio anchored	1,023	57.9	160	45.0
	Ab initio unsupported	1,342	42.1	343	15.5
S. lycopersicum	Fully extrinsic	17,656	92.7	18,438	92.0
	Partially extrinsic	4,823	42.9	5,395	39.1
	Ab initio anchored	1,431	29.0	1,345	19.4
	Ab initio unsupported	4,079	7.2	3,082	3.3
D. rerio	Fully extrinsic	15,766	85.9	16,273	86.0
	Partially extrinsic	11,036	15.4	11,763	14.7
	Ab initio anchored	1,860	10.8	1,628	5.8
	Ab initio unsupported	11,257	0.7	10,908	0.3
G. gallus	Fully extrinsic	11,893	85.7	11,593	86.4
	Partially extrinsic	4,464	20.6	4,935	21.4
	Ab initio anchored	589	7.6	628	6.7
	Ab initio unsupported	9,186	0.5	9,043	0.4
M. musculus	Fully extrinsic	13,554	91.9	13,988	92.1
	Partially extrinsic	6,285	20.1	6,564	18.8
	Ab initio anchored	1,070	7.5	1,218	4.8
	Ab initio unsupported	18,463	0.6	17,750	0.1

Discussion

The purpose of GeneMark-ETP was to generate gene predictions in a eukaryotic genome in consistence with the genomic sequence patterns, protein-coding region determinants elucidated from transcripts, as well as homologous proteins footprints. Solving this task required training of the two GHMM models – one for the assembled transcripts and one for genomic DNA, mapping of the genes predicted in transcripts to genome and finding a set of proteins homologous to a not yet fully predicted gene. All these steps have led to integration of the three layers of information for each genomic locus.

One of the principal differences with the earlier developed tools, GeneMark-ES, GeneMark-ET and GeneMark-EP+ was in the method of training of the genomic GHMM model. In all the just

cited tools the gene prediction process started with the heuristic model with parameters determined based on functions approximating dependence of the K-mers frequencies on genome GC content. In GeneMark-ETP we start the process of genomic GHMM training with a model derived from a set of the HC gene loci identified by integration of genomic, transcriptomic and protein evidence. In experiments with well-studied genomes the numbers of HC genes were frequently so large that thus derived initial GHHM model would not change in the further training iterations, thus they became unnecessary.

Here we would like to discuss the steps of the algorithm that distinguish GeneMark-ETP from the earlier tools as well as to comment on the possible sources of errors and the comparisons with other computational tools.

Identification of a set of genes predicted with high confidence

The accuracy of gene prediction in full transcripts is known to be critically affected by assembly errors, as well as by the presence of mRNA transcripts missing their natural 5' or 3' ends. Accurate assembly of complete transcripts from short reads has presented a challenge for quite a while (Steijger et al. 2013). New tools, such as StringTie2 (Kovaka et al. 2019), have demonstrated a significantly improved performance.

Gene prediction in an assembled transcript was done by a specialized tool, GeneMarkS-T. Predicted proteins were searched against a protein database. The proteins found in the similarity searches could fully support the predicted gene, thus making the prediction more confident. Predicted 5' partial genes were further refined (see Methods). The resulting set of genes was named *high-confidence* (HC) *genes*. In our test, the HC genes had on average significantly better match to the 'true' genes than the set of genes derived from the initial GeneMarkS-T gene predictions (Tables 1, S3). Thus, the set of HC genes was identified by using genomic, transcriptomic, and protein data in concert.

Identification of a set of genes predicted with the least confidence

In all the seven genomes, we saw that genome-specific proportions of genes predicted with full and partial extrinsic support went down with the increase in the genome size (Table 3). For example, the percentage of genes predicted with extrinsic support diminished, from 81.5% for *D. melanogaster* to 49.2% for *M. musculus* (the numbers a given for the case of using larger reference databases).

At the same time, the increase in genome size was accompanied by the increase in the proportion of the genes predicted *ab initio* (Table 3). For instance, the percentage of genes predicted *ab initio* were 3.8% and 18.5% for *D. melanogaster* and 48.0% and 50.8% for *M. musculus* for smaller and the larger protein databases, respectfully.

Importantly, the fraction of false positives among the *ab initio* predictions grew even faster with the genome size increase and led to a significant drop in Specificity (Tables 3, S6). The fast growth in the rate of false positive predictions could be caused by a combination of several factors; an

increase in the average length of intron and intergenic regions; increased frequency of pseudogenes; increase in the size of populations of transposable elements (repeats), etc. Notably, in the large genomes, the percentage of false positive predictions was in the range observed in our experiments with gene prediction in simulated non-coding regions (data not shown).

Analysis of the results showed that the gene level specificity of *ab initio* gene predictions could drop significantly, reaching below 10% for *ab initio unsupported* predictions (the fourth category in Table 3). We observed that such genes comprised more than 10% of the total number of predictions in genomes larger than 300 Mbp (Figs. S3, S4). At the same time, in all the genomes under consideration, the fraction of predictions supported extrinsically (fully or partially) was above 50%. Therefore, we came up with the following empirical rule. For genomes larger than 300 Mbp where a 10% threshold on the fraction of *ab initio* predictions was exceeded, it was reasonable to eliminate gene predictions that fell into the category *ab initio unsupported*. For such genomes, all the genes remaining in the final list of predictions would have at least a few elements of the exon-intron structure supported by extrinsic evidence which was either used in the prediction or detected *a posteriori* (the first three categories in Table 3).

Transition to the GC-content-specific models

For gene prediction in *GC-heterogeneous* genomes and transcripts, we used several GC-content-specific sets of the GHMM parameters. This diversification led to the improvement in gene prediction accuracy. The resulting performance was certainly better than the ones of GeneMark-ET or GeneMark-EP+, the tools that used a single model designed for an average genomic GC composition (Fig. 2).

The "GC-heterogeneous" mode could be used for any genome. However, if a genome is rather a true *GC-homogeneous* one, the use of this mode would increase runtime and slightly decrease the accuracy, due to splitting the overall training set into smaller subsets. Therefore, the degree of GC-heterogeneity is assessed at a pre-processing step, and, the "GC-heterogeneous" mode is used if needed.

Processing of repetitive elements

Transposable elements (TEs), particularly families of retrotransposons with thousands of copies of very similar TE sequences, occupy substantial portions of eukaryotic genomes. Errors in gene prediction may happen in presence of repetitive elements with composition similar to protein-coding genes (Yandell and Ence 2012; Torresen et al. 2019). Getting information on the repeat locations, e.g., predicted *de novo*, helps to reduce the errors. However, some of the predicted repeat sequences may overlap with the protein-coding genes of the host (Bayer et al. 2018).

One could *hard-mask* all repeats longer than a chosen threshold T (Lomsadze et al. 2014; Bruna et al. 2020). Such an approach carries disadvantages: (i) repeats shorter than T would not be masked and (ii) protein-coding exons overlapped by the masked repeats could be difficult to find.

To deal with this issue, the authors of AUGUSTUS have introduced a repeat length-dependent penalty function used in the GHMM Viterbi algorithm (Stanke et al. 2008). A single parameter of this function, *q*, had the same value for all species. We have shown that in GeneMark-ETP use of a species-specific parameter *q* produced even better results (Fig. 4, see also Fig. S5, Table S7). The species-specific *q* value was determined after identification of the HC genes (see Methods). We have observed that the suggested approach was robust with respect to the size of the sample of the HC genes (data not shown).

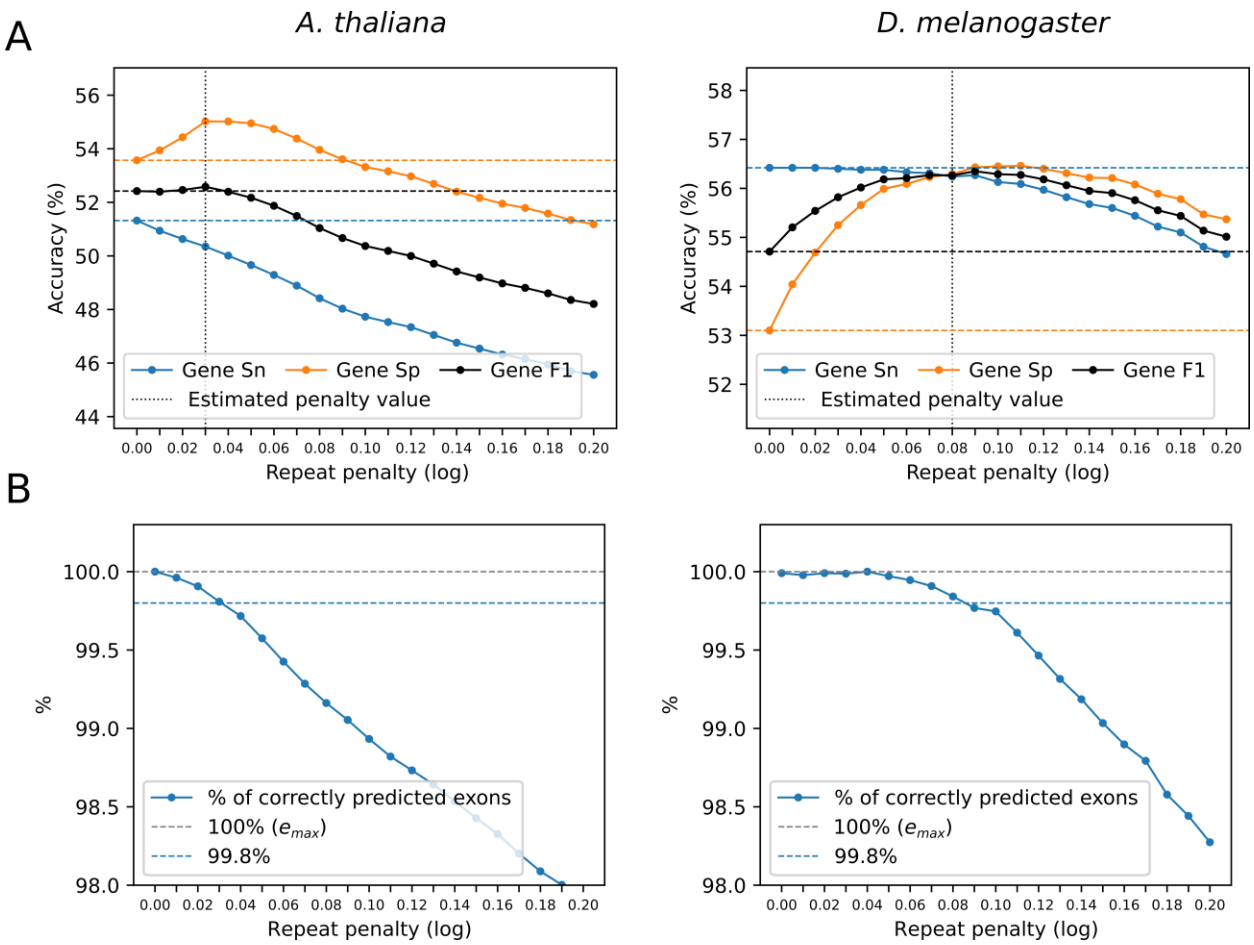

Figure 4. A. Dependence of the accuracy measures on the repeat penalty parameter *q* observed for genomes of *A. thaliana* and *D. melanogaster*. **B.** Dependence of % of correctly predicted exons of the HC genes (Sn) on the repeat penalty parameter *q* for the same genomes as in A (see Methods).

Robustness of GeneMark-ETP

If in the first step of the GeneMark-ETP run (Fig.1), more than a certain number of HC genes were identified (wih 4,000 as the default), then a single round of the GHMM training and gene prediction was sufficient. Otherwise, an iterative training of the GHMM model was conducted ending with generation of the final set of gene predictions (Fig. S6).

This rule was motivated by the observation that if in the first step more than 4,000 HC genes were identified, the additional iterative training, with effective increase of the training set size, did not improve gene prediction accuracy. This outcome was likely due to reaching the training set size at which the parameters of the 5-order three-periodic Markov chain, computed from frequencies of phased nucleotide 6-mers, approached the stationary values. For a genome with mid-GC nucleotide composition, such a training set size is 4 Mbp, that, assuming an average size of coding region to be 1000nt, corresponds to 4,000 genes. For low and high GC genomes, reaching stationarity of the parameter estimates could occur for even smaller training sets.

As could be expected, the GeneMark-ETP accuracy was less affected by the change in size of the protein database in comparison with GeneMark-EP+ that was using protein data only. For example, for *D. melanogaster*, when a larger protein database (proteins of the same species excluded from PD_0) was changed to a smaller database (proteins of the same order excluded from PD_0), the gene level F1 of GeneMark-ETP decreased by 6.3% (Table S1) while the F1 of GeneMark-EP+ decreased by 11.4%. Certainly, the use of the HC genes derived from GeneMarkS-T predictions that did not have full-length protein support did contribute into this effect (Methods Section 2.2.3).

While the increase in the volume of proteins from the closely related species should, generally, lead to increase in gene prediction accuracy, the accuracy of GeneMark-ETP (similarly to GeneMark-EP+) did not critically depend on presence of such proteins in the reference database.

Comparison with other computational tools

GeneMark-ET, GeneMark-EP+, and their virtual combination

GeneMark-ETP performed better than either GeneMark-ET or GeneMark-EP+ in all the tests (see Results). Since both GeneMark-ET or GeneMark-EP+, use only a single source of extrinsic evidence, this result should have been expected.

The *virtual tool* considered here made an artificial combination of the sets of genes predicted separately by GeneMark-ET and GeneMark-EP+ (Method Section 5.2). The largest sensitivity of such tool could be achieved by the *union* of the two sets while the largest specificity could be achieved if the *intersection* of the two sets is used. Implementation of the best-balanced combination would require either a removal of false positives from the *union* set, or an addition of true positives to the *intersection* set. When a gene finder is running on a novel genome information on true and false positives is not immediately available. Nevertheless, even if this ideal correction would be made, the accuracy of the best virtual tool still fell below the accuracy of GeneMark-ETP (Figs 3, S2).

BRAKER1, BRAKER2 and TSEBRA

Earlier developed pipelines—BRAKER1 (Hoff et al. 2016), combining AUGUSTUS and GeneMark-ET, using transcripts as a source of extrinsic evidence, and BRAKER2 (Bruna et al. 2021), combining AUGUSTUS and GeneMark-EP+ supported by cross-species protein data are

frequently used tools. WE have shown that GeneMark-ETP gene prediction accuracy was higher than either BRAKER1 or BRAKER2, especially for large genomes (Figs. 1, 2). Again, this result could be expected due to the use of twice as many types of extrinsic information in GeneMark-ETP. The recently developed TSEBRA applies well designed rules to select a subset of all predictions made by either BRAKER1 or BRAKER2 and, thus, achieves higher accuracy than any of the BRAKERs (Gabriel et al. 2021). It was shown that TSEBRA performed better than EVidenceModeler, one of the best combiners, as well.

In our tests, it was demonstrated that GeneMark-ETP achieved higher accuracy than TSEBRA in large genomes (Fig. 2), particularly in the GC-heterogeneous ones (*G. gallus*, *D. rerio*) where BRAKER1 and BRAKER2 use single statistical models tuned up for "average GC" in each genome. Nevertheless, GeneMark-ETP reached higher than TSEBRA prediction accuracy in the GC-homogeneous genomes of *S. lycopersicum* and *D. rerio*. Therefore, there should be yet another factor beyond the training of the GC diversified the statistical models. Such additional source of accuracy improvement is, arguably, use of extrinsic hints that integrate both assembled transcript and protein information. The accuracy advantage of GeneMark-ETP was much smaller in the group of compact genomes (Fig. 1), with TSEBRA achieving higher accuracy than GeneMark-ETP in the case of *C. elegans*.

All over, the new tool integrated transcriptomic and protein evidence of presence of protein-coding function in genomic sequences into hints used consistently at *all* stages of the algorithm training and gene prediction. We argue that such an approach is more effective than combining the predictions made with a particular single type of extrinsic evidence in a "post-processing" step.

Materials

For the assessment of the GeneMark-ETP gene prediction accuracy, we selected seven genomes from diverse eukaryotic clades (Tables 4, S8). The group included relatively short GC-homogeneous genomes of the well-studied model organisms: *A. thaliana*, *C. elegans,* and *D. melanogaster*. The group also included larger genomes, both GC-homogenous (*S. lycopersicum*, *D. rerio*) and GC-heterogeneous (*G. gallus*, *M. musculus*). In all the genomic datasets, contigs with no chromosome or organelle assignment were excluded from the analysis.

To generate the reference sets of proteins used as a source of extrinsic evidence we used the OrthoDB v10.1 protein database (Kriventseva et al. 2019); for more details see (Bruna et al. 2020; Bruna et al. 2021). For each of the seven species, we built an initial species-specific protein database (PD_0) containing all proteins from the largest clade considered for the given species (Table S9). Also, for each given species, we created two smaller reference databases by removing from PD_0 either all proteins of this species itself and its strains, or proteins from all the species that belonged to the same taxonomic order. These, the larger and the smaller databases, were devised to mimic practical scenarios when a species in question would have either a larger or a smaller set of proteins from close relatives present in the reference database. All over, the

numbers of proteins in the databases used in the computations ranged from 2.6 to 8.3 million (Table S9).

Transcript datasets, such as the sets of Illumina paired reads, were selected from the NCBI SRA database. The read length varied between 75 to 151 nt; the total volume of RNA-Seq collections varied from 9 Gb for *D. melanogaster* to 83 Gb for *M. musculus* (Table S10).

Table 4. Genomes used for the assessment of the GeneMark-ETP gene prediction accuracy. For the larger genomes, the numbers in parentheses characterize selected subsets of genes presumed to be more reliably annotated. To compute the numbers of introns per gene we used averages among annotated alternative transcripts.

Species	Genome length (Mb)	Reference annotation statistics					
		# of protein-coding genes		# of transcripts		# of introns per gene	
C. elegans (roundworm)	100	19,969		28,544		4.8	
A. thaliana (thale cress)	119	27,445		40,828		4.0	
D. melanogaster (fruit fly)	138	13,951		22,395		2.8	
S. lycopersicum (tomato)	807	25,158	(15,138)	31,911	(15,150)	4.4	(4.3)
D. rerio (zebrafish)	1,345	25,611	(17,894)	42,934	(19,978)	8.4	(8.4)
G. gallus (chicken)	1,050	17,279	(10,736)	38,534	(12,733)	9.0	(9.2)
M. musculus (mouse)	2,723	22,405	(16,531)	58,318	(20,708)	6.0	(8.6)

Methods

Outline of the GeneMark-ETP algorithmic steps

In GeneMark-ES, -ET, -EP, iterative unsupervised training was used to estimate the parameters of the GHHM models (Lomsadze et al. 2005; Lomsadze et al. 2014; Bruna et al. 2020). The iterative cycles of model training and gene prediction resulted in getting a final set of model parameters employed in the prediction of the final set of genes. Since GeneMark-ETP relies on a larger set of extrinsic data, the training procedure was significantly modified.

Another difference with the previous developments is that GeneMark-ETP predicts genes both in genomic DNA as well as in assembled transcripts. Gene prediction in transcripts is done by a self-training tool GeneMarkS-T with GHMM (generalized hidden Markov model) designed for sequences with intron-less genes (Tang et al. 2015). On the other hand, gene prediction in eukaryotic DNA sequences requires GHMM with an exon-intron model (Lomsadze et al. 2005).

At the start of a genome analysis, GeneMark-ETP generates a set of *high-confidence* (HC) genes (Fig. S7). GeneMarkS-T plays a central role in this step. Next, if the number of the HC genes is large enough to make a reliable training set, the parameters of the 'eukaryotic' GHMM are estimated, and the Viterbi algorithm is used to predict genes in the regions between the HC genes. If the set of HC genes is not large enough, the initial parameters of the GHMM model are

further refined by self-training in the genomic regions situated between HC genes. The use of the transcript and protein evidence continues in all the steps (Fig. 5).

Selection of a set of genes predicted with high confidence

1 Gene prediction in assembled transcripts

Reads of each RNA-Seq library used in the input for GeneMark-ETP are splice-aligned to the genome by HISAT2 (Kim et al. 2019) and assembled into a set of transcripts by StringTie2 (Kovaka et al. 2019). All the sets of transcripts are merged into the final non-redundant transcriptome where the low-abundance transcripts are filtered out by a procedure described in the StringTie2 paper (Kovaka et al. 2019). Next, GeneMarkS-T (Tang et al. 2015) predicts genes in the assembled transcripts. These genes are refined based on the results of the protein similarity search as described below.

Refinement of the GeneMarkS-T predictions based on protein information

The GeneMarkS-T predicted gene is considered to be 5' partial if the predicted protein-coding region effectively starts from the first nucleotide of a transcript. An incorrectly predicted 5' partial gene could have a true gene inside, with true start located further downstream in the transcript sequence (Fig. S8). The first downstream ATG from the 5' end of the transcript could be either a start of the coding region (making the longest ORF for the in-frame stop located downstream) or it could be situated inside the longer coding region disrupted in the transcript assembly. To determine if an ATG is a true start codon or an internal codon, the two alternative coding regions are translated into proteins and used as queries in similarity searches by DIAMOND (Buchfink et al. 2015) against a reference protein database. If among the targets of both sets of alignments exists at least one that is i/ common for both queries and ii/ shows better support for the prediction starting at the start of the transcript, the longer sequence is selected as the predicted 5' partial gene. Otherwise, the shorter sequence with the ATG start is selected as the predicted complete gene. Note, that if the sets of targets (we choose 25 bests from each DIAMOND search) do not overlap, the 5' partial prediction is selected.

We needed to assess to which one of the competing shorter and longer protein queries the same target protein provides a stronger support. The details of this assessment is described in Section 2.4 of Supplementary Materials. The 5' partial gene prediction (longer protein query) is chosen if inequality (S1) is fulfilled for at least one common target (Fig. S9), otherwise, the shorter protein query (complete gene) is selected. Another type of refinement— when a predicted complete gene is changed into a 5' partial gene —is addressed by similar approach: using alignments of the common target to two possible queries and employing inequality (S1). We have observed that the current version of GeneMarkS-T makes very few errors of this type.

2 High-confidence genes

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) described in Section 2.4 of Supplementary Materials. To find a target satisfying condition (S2), we examine 25 top-scoring alignments of the query to the target proteins. If such a target exists, the query—a 5' complete gene with full protein support—is classified as a *high-confidence* gene.

A 5' complete gene predicted in a transcript may not make the "longest ORF" with respect to the predicted 3'end of the gene, though it was observed that the vast majority of annotated eukaryotic genes do make the longest ORFs. To correct possible underprediction, both the original prediction and its extension to the "longest ORF" are checked by condition (S2) and, if fulfilled, one of them or even both are classified as HC (alternative) isoforms.

5'-partial genes with full protein support

A 5' partial gene (see Fig. S10) can be classified as a high-confidence gene if the C-terminal of its protein translation is supported by at least one protein alignment. If the best-scoring protein alignment does not cover the 5' partial protein from the start (see Fig. S11 where $Q_{start} \neq 1$), the 5' partial gene is shortened to the first in-frame ATG codon covered by the protein alignment.

Any gene predicted as 3' partial (unambiguously defined by the lack of a stop codon) is not considered as a candidate for an HC gene.

Genes predicted ab initio

Complete GeneMarkS-T gene predictions that either have no significant BLASTp hits in the protein database or do not satisfy condition (S2), for an alignment of the predicted protein and any of its best targets in the protein database, still could make high-confidence genes. To be qualified as such, all of the following conditions have to be satisfied: (i) a length of protein-coding region is longer than $299\ nt$, (ii) an in-frame stop codon triplet is present in the 5' UTR, (iii) the GeneMarkS-T log-odds score is ≥ 50 and (iv) the gene structure mapped to genomic DNA does not create any conflict with ProtHint hints (see Section 2 of Supplementary Methods for more details). A single HC isoform (the one with the longest protein-coding region) is selected per locus, where several isoforms are predicted based on multiple transcript assemblies.

High-confidence alternative isoforms

Alternative isoforms of the same gene may belong to the set of HC genes. Selection of HC alternative isoforms is done as follows.

Let $I^g_{complete}$ be a set of all complete isoforms of gene g and $I^g_{partial}$ is a set of all its partial isoforms. Each isoform i is assigned a score $\mathrm{s}(i)$ -- the *bitscore* of its best hit to a protein in the reference protein database.

We compute the maximum $\mathrm{s}(i)$ score of all the complete isoforms for each gene g (Eq. 1).

$$\mathrm{s}(g_{complete}) = \max_{i \in I^g_{complete}} \mathrm{s}(i) \tag{1}$$

The score of an isoform selected as HC complete isoform must satisfy inequality (2).

$$\mathrm{s}(i) \geq 0.8 \times \mathrm{s}(g_{complete}) \quad (i \in I^g_{complete}) \tag{2}$$

For the partial alternative isoforms, we have

$$\mathrm{s}(g_{partial}) = \max_{i \in I^g_{partial}} \mathrm{s}(i) \tag{3}$$

If this score is larger than $\mathrm{s}(g_{complete})$, the partial transcript with this largest score is selected as the partial HC isoform. Moreover, all the complete HC isoforms are removed in this case. Otherwise, if $\mathrm{s}(g_{partial})$, is lower than $\mathrm{s}(g_{complete})$, then only the complete isoforms are retained.

The GHHM model training

Single step model training

A set of predicted HC genes is used for the initial and often final GHHM parameter estimation. First, the set of HC genes is checked for possible redundancy. In the loci with several complete HC isoforms the isoform with the longest protein-coding region is selected. Next, we determine the GC content distribution of the selected HC genes and if more than 80% of them are contained in a 10% wide GC content interval, the genome is characterized as GC homogeneous, else as GC heterogeneous.

In the *GC homogeneous case*, the loci of all the selected HC genes are used for the estimation of parameters of a GHHM model (Fig. S6). If the set of genes is large enough (more than 4,000), the GHMM model parameter estimation is done by training on the set of the HC loci, the sequences containing these HC genes with 2,000 margins. Otherwise, an iterative *extended training* of the GHMM parameters is done similar to the approach described earlier (Lomsadze et al. 2014; Bruna et al. 2020; Lomsadze et al., 2005).

In the *GC heterogeneous* case, the sequence set of *HC loci* is split into the three GC bins: low, medium, and high. The borders of the medium GC bin with a fixed width (9% by default) are selected within the GC one dimensional range to include the largest possible number of the HC loci. Setting up these boundaries automatically determines the borders of the low and medium GC bins. The sets of the HC loci contained in each GC bin are used for the training of the three GC-specific GHHM models. The training in each bin is finished in a single step if more than 4,000

HC genes is identified in total. Otherwise, the several iterations of *extended* training are done in each bin.

Notably, gene prediction in transcripts by GeneMarkS-T is made with a set of the GC-specific statistical models derived as described in Tang et al, 2015.

Extended GHMM model training

The logic of extended model training is similar but not identical to iterative training used in GeneMark-ET and GeneMark-EP+ (Lomsadze et al. 2014; Bruna et al. 2020).

At the initialization of iterations for genomes with *homogeneous GC content,* the GHHM model parameters are derived from the sequences of the HC loci contrary to the use of the cruder heuristic model in GeneMark-ET and GeneMark-EP+. The gene prediction is then made only in the genomic sequences situated between HC genes, *the HC-intermediate regions*. These predictions, serving as ProtHint gene seeds, are translated and used as queries for a protein database search initiating the full run of ProtHint (Bruna et al. 2020). High scoring hints are enforced in the Viterbi algorithm upon processing of the HC-intermediate regions. In all the iterations but the last one, the newly predicted genes serve as seed genes for the next run of ProtHint. It was observed in our tests on the six genome that convergence to a certain set of genes in training as well as to a certain set of predicted genes characterized by particular (not changing) values of Sn and Sp could be reached as early as after two and frequently after three iterations. After the final iteration, the set of HC genes was merged with the set of genes predicted in the HC-intermediate regions. Overall, we have to say that the number of iterations is species specific and for the species that have rather small sets of available transcripts and limited number of proteins with high similarity score, the number of iterations could be larger than three.

An important trait of the GHMM training process implemented in GeneMark-ET and GeneMark-EP+ algorithms, was step by step unfreezing of the subsets of the GHMM model parameters. For instance, the Markov chain transition probabilities and durations of functional regions, i.e., intron, exon etc., were fixed during the initial iterations while the values of emission probabilities were free to change. In the later iterations all the parameters were made free. Such gradual unfreezing of the parameters was shown to be unnecessary for GeneMark-ETP. From the second iteration on, all the GHMM parameters are estimated simultaneously. We attribute this streamlining of the training process to availability of the more accurate initial parameters of GHMM derived from the sequences of HC loci.

In *GC heterogeneous* genomes the extended GHHM training worked as follows. First, GeneMark-ETP calculated the GC content of each HC-intermediate region and assigned the regions to the corresponding GC bins. The parameters of the initial GC specific GHHM model were trained on the thus selected sets of sequences of HC loci. Subsequently, a GC specific model was used for gene prediction in the HC-intermediate regions of a corresponding GC bin. From this point on, the extended training on the HC-intermediate regions from a particular GC bin was essentially

made in the same way as described above for the GC homogeneous case. The iterative training and the final gene prediction step were executed using the GC-specific models updated from iteration to iteration.

Accounting for repeats

Repeat identification and masking

To identify repetitive sequences, we used RepeatModeler2 (Flynn et al. 2020) and RepeatMasker (www.repeatmasker.org). Repeat libraries were generated *de novo* using RepeatModeler2. Repeat sequences—interspersed and tandem repeats—were then identified and soft-masked by RepeatMasker.

Selection of the species-specific repeat penalty parameter

To account for an overlap of a protein-coding region with a repetitive sequence, the GeneMark-ETP algorithm changes the probability (likelihood) of a sequence in such an overlap by formula (4) with penalty parameter q (n is the length of an overlap):

$$P(seq|coding\ state\ overlapping\ repeat) = \frac{P(seq|coding\ state)}{q^n} \qquad (4)$$

GeneMark-ETP estimates species-specific parameter q for each genome. The goal is to find the value minimizing disruptions to correct gene predictions. The q estimation step is made after the first iteration of the GeneMark-ETP model training; this, we have full GHMM model, a set of the HC genes, and the coordinates of the repeats identified in genomic DNA prior to the first iteration. The Viterbi algorithm is then run several times (with different *q* values) in an *ab initio* mode to predict genes in the soft-masked genomic sequences containing the HC genes (Fig. 4). In each run we compute gene level F1 value of the gene prediction accuracy determined on the test set made from the HC genes. Then we identify the value *q* for which F1 would reach maximum (Fig. 4A). Thus determined best *q* value was good approximation of the one determined on a test set derived with "full" knowledge of genome annotation (Table S7). Moreover, we found that selecting the *q* value by using the exon level Sn was a more robust method in comparison of using the gene level F1. Technically, we would compute the best q by maximizing the number of correctly predicted exons in the HC genes, e_{max}. Such q value would be larger than or equal to 1 (Fig. 4B). We found that value $q*$ at which $0.998 \times e_{max}$ exons were correctly predicted was a good estimate of the best value q (marked in panel A of Fig. 4). To reduce the running time of the search for the best repeat penalty parameter, we used simulated annealing (Kirkpatrick et al. 1983).

Gene prediction in the HC intermediate regions

Integration of intrinsic and extrinsic evidence

The models trained on the set of the HC loci are used in GeneMark.hmm to create initial gene predictions in the HC-intermediate segments (Fig. S12). These gene predictions can be refined by incorporation of the protein and transcript evidence. This task is solved as follows. The initial gene predictions are used in ProtHint to generate protein-based hints (Bruna et al., 2020). On the other hand, we have a set of genes predicted by GeneMarkS-T in transcripts that were mapped to genome by HISAT2. The mapping that falls into the HC intermediate regions constitute transcript based evidence for the HC intermediate regions. The whole set of hints is then ready for enforcement in a run of the Viterbi algorithm for GHMM (Fig. 6). To reiterate, we have the following categories of hints: 1/ RNA-Seq and ProtHint-derived hints that agree with each other; 2/ solely high score ProtHint hints; 3/ solely RNA-Seq-based intron hints, if they overlap but do not coincide with the *ab initio* predicted introns; the requirement of the overlap filters out introns mapped from expressed lncRNA; 4/ exons of partial HC genes; partial HC genes are determined at the stage of HC gene identification (Methods Section *2.2.2)*. Note that category 1-3 may not necessarily point to elements of the same gene (the RNA-Seq mapped introns or the ProtHint introns). Hints of category 4 should belong to the same gene.

The genes predicted in the HC-intermediate segments along with the full set of the HC genes constitute the *final set of genes* predicted by GeneMark-ETP.

While we did not make experiments with long RNA reads, we could argue that if the high-quality long reads or their assemblies are available, GeneMarkS-T could be run on the long reads to predict intron-less genes that in turn would be mapped to the genome, e.g. with Minimap2 (Li 2018). Thus, the GeneMark-ETP run could be implemented with this type of data.

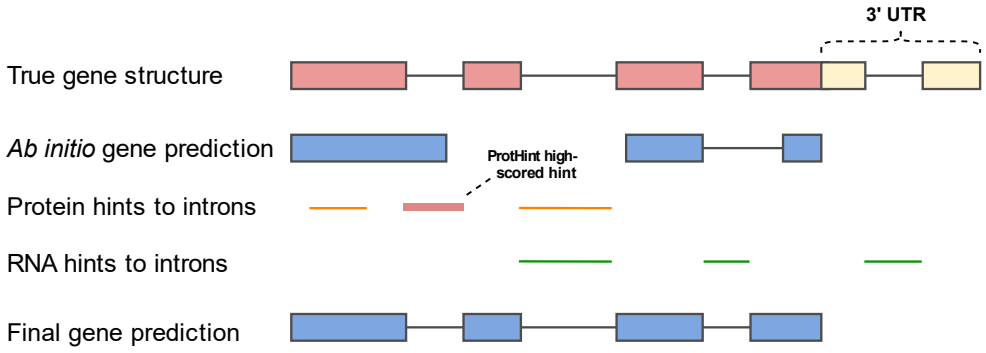

Figure 6. Integration of extrinsic evidence into the GeneMark-ETP gene predictions in HC-intermediate segments. The figure shows that a low score extrinsic evidence not corroborated by other extrinsic evidence or by *ab initio* gene prediction is ignored. The low score evidence is shown by thin lines.

Filtering out *unsupported ab initio* predictions

The genes predicted in the HC-intermediate segments could be split into two non-overlapping sets: evidence-supported predictions and pure *ab initio* predictions (see Discussion and Table 3). The evidence-supported genes must have at least one element of the gene structure supported externally. We have observed that in larger genomes the fraction of correct predictions among *unsupported ab initio* predictions were sharply decreasing with the genome size (see Sp values in Table 3). Therefore, for genomes larger than 300 Mbp we offer two types of outputs, one with the full set of gene predictions and the other with *unsupported ab initio* predictions removed. The results of the accuracy tests in the larger genomes are given for this reduced output.

The accuracy assessment of GeneMark-ETP

Selection of gene sets with reliable annotation

Since annotations of well-studied genomes of *A. thaliana*, *C. elegans*, and *D. melanogaster* have been updated multiple times, we considered these complete annotations as "gold standards", against which the gene prediction accuracy parameters was determined. Arguably, the reference annotations of the other four genomes have been less trustworthy. Therefore, to assess the sensitivity parameters we selected genes with identical annotations in two different sources, e.g., in the NCBI and the Ensemble records (Table S8). On the other hand, the values of prediction specificity for these genomes were defined by comparison with the union of genes annotated by either NCBI or Ensemble or by both.

Description of the sets of genes used for accuracy assessment is given in Table 3. In all the tests, regions of annotated pseudogenes were excluded from consideration.

A virtual combination of GeneMark-ET and GeneMark-EP+ predictions

We compared the accuracy of GeneMark-ETP with the accuracy delivered by a "virtual" tool which output was made of a combination of genes predicted by GeneMark-ET and GeneMark-EP+. Predictions made by GeneMark-ET and GeneMark-EP+ could be combined in two simple ways: by making either union U or intersection I. The intersection contained only genes with identical gene structures. The set U presents the most comprehensive set of predicted genes, while the set I, arguably, presents the most reliable predictions. The sensitivity of U genes is designated as Sn and the specificity of I genes is designated as Sp. Now, if one can reduce set U by taking away only the incorrect predictions, the point in Fig. 3 will move horizontally. If one can add to the set I only correct predictions the point in Fig. 3 will move up vertically. The crossing of the two lines at the point (Sn, Sp) characterizes the accuracy of the virtual tool, implementing the best version of the virtual combiner approach.

Running BRAKER1, BRAKER2, and TSEBRA

To make comparisons with the transcript-supported BRAKER1 (Hoff et al. 2016) and protein-supported BRAKER2 (Bruna et al. 2021) we have run BRAKER1 and BRAKER2, respectively, with

the same RNA-Seq libraries and protein databases, as the ones used in experiments with GeneMark-ETP. Also, we ran TSEBRA (Gabriel et al. 2021) that generated a set of genes supported by both RNA-Seq and proteins. TSEBRA selects a subset of the union of gene predictions made by BRAKER1 and BRAKER2. TSEBRA was shown to achieve higher accuracy than (i) either BRAKER1 or BRAKER2 running alone, as well as (ii) EVidenceModeler (Haas et al. 2008), one of the frequently used combiner tools.

Summary

A new eukaryotic gene prediction software tool, GeneMark-ETP was shown to generate better—and in the case of large genomes significantly more accurate—eukaryotic gene predictions in comparison with the earlier developed tools. The algorithm constructs a genomic parse into coding and non-coding regions supported by the combined evidence extracted from genomic, transcriptomic, and protein sequences. Integration of the intrinsic and extrinsic data is consistently implemented through the major steps of the algorithm: the GHMM models training and gene prediction. The margin of the prediction accuracy improvement does grow with the increase of the genome complexity from relatively compact genomes to large, GC-heterogeneous genomes. All over, we believe that we managed to demonstrate the advantage of the simultaneous integration of several sources of evidence into gene prediction over a post-processing-style integration combining several separate streams of gene predictions, each with its own type of extrinsic evidence.

Supplementary materials

URL to be determined (a file is submitted along with the main text)

Availability

GeneMark-ETP is available on GitHub at https://github.com/gatech-genemark/GeneMark-ETP.git and http://topaz.gatech.edu/GeneMark/license_download.cgi. All scripts and data used to generate figures and tables in this manuscript are available at https://github.com/gatech-genemark/GeneMark-ETP-exp. The runtime of GeneMark-ETP depends linearly on the genome size and is comparable to the one of GeneMark-EP+. For example, on a machine with 64 CPU cores, GeneMark-ETP runs on genomes of *D. melanogaster*, *D. rerio*, and *M. musculus* for 1.0, 4.5, and 6.5 hours, respectively.

Funding

National Institutes of Health [GM128145 to M.B., in part]. Funding for the open access charge: National Institutes of Health [GM128145].

Conflict of interest statement. None declared.

References

Allen JE, Pertea M, Salzberg SL. 2004. Computational gene prediction using multiple sources of evidence. *Genome Research* **14**: 142-148.

Allen JE, Salzberg SL. 2005. JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics* **21**: 3596-3603.

Banerjee S, Bhandary P, Woodhouse M, Sen TZ, Wise RP, Andorf CM. 2021. FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences. *BMC Bioinformatics* **22**: 205.

Bayer PE, Edwards D, Batley J. 2018. Bias in resistance gene prediction due to repeat masking. *Nat Plants* **4**: 762-765.

Bruna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform* **3**: lqaa108.

Bruna T, Lomsadze A, Borodovsky M. 2020. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform* **2**: lqaa026.

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

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**: 78-94.

Coghlan A, Fiedler TJ, McKay SJ, Flicek P, Harris TW, Blasiar D, n GC, Stein LD. 2008. nGASP--the nematode genome annotation assessment project. *BMC Bioinformatics* **9**: 549.

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

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* **117**: 9451-9457.

Gabriel L, Hoff KJ, Bruna T, Borodovsky M, Stanke M. 2021. TSEBRA: transcript selector for BRAKER. *Bmc Bioinformatics* **22**.

Goodswen SJ, Kennedy PJ, Ellis JT. 2012. Evaluating high-throughput ab initio gene finders to discover proteins encoded in eukaryotic pathogen genomes missed by laboratory techniques. *PLoS One* **7**: e50609.

Gremme G, Brendel V, Sparks ME, Kurtz S. 2005. Engineering a software tool for gene structure prediction in higher organisms. *Inform Software Tech* **47**: 965-978.

Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E et al. 2006. EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol* **7 Suppl 1**: S2 1-31.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**: R7.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**: 767-769.

Howe KL, Chothia T, Durbin R. 2002. GAZE: A generic framework for the integration of gene-prediction data by dynamic programming. *Genome Research* **12**: 1418-1427.

Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. 2018. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics* **19**: 189.

Keller O, Kollmar M, Stanke M, Waack S. 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**: 757-763.
Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907-915.
Kirkpatrick S, Gelatt CD, Vecchi MP. 1983. Optimization by Simulated Annealing. *Science* **220**: 671-680.
Kiryutin B, Souvorov A, Tatusova T. 2007. Prosplign: protein to genomic alignment tool. In *11th Annual International Conference in Research in Computational Molecular Biology*, San Francisco, USA.
Kong J, Huh S, Won JI, Yoon J, Kim B, Kim K. 2019. GAAP: A Genome Assembly + Annotation Pipeline. *Biomed Res Int* **2019**: 4767354.
Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**: 59.
Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**: 278.
Kozak M. 1999. Initiation of translation in prokaryotes and eukaryotes. *Gene* **234**: 187-208.
Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simao FA, Zdobnov EM. 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research* **47**: D807-D811.
Kulp D, Haussler D, Reese MG, Eeckman FH. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. *Proc Int Conf Intell Syst Mol Biol* **4**: 134-142.
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.
Liu Q, Mackey AJ, Roos DS, Pereira FC. 2008. Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction. *Bioinformatics* **24**: 597-605.
Lomsadze A, Burns PD, Borodovsky M. 2014. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* **42**: e119.
Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* **33**: 6494-6506.
Mudge JM, Harrow J. 2016. The state of play in higher eukaryote gene annotation. *Nat Rev Genet* **17**: 758-772.
Parra G, Blanco E, Guigo R. 2000. GeneID in Drosophila. *Genome Res* **10**: 511-515.
Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290-295.
Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson JD. 2020. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC Genomics* **21**: 293.
Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31.
Song L, Sabunciyan S, Yang G, Florea L. 2019. A multi-sample approach increases the accuracy of transcript assembly. *Nat Commun* **10**: 5000.
Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**: 637-644.
Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**: ii215-225.
Steijger T, Abril JF, Engstrom PG, Kokocinski F, Consortium R, Hubbard TJ, Guigo R, Harrow J, Bertone P. 2013. Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* **10**: 1177-1184.
Tang S, Lomsadze A, Borodovsky M. 2015. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res* **43**: e78.
Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res* **18**: 1979-1990.

Torresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, Gruca A, Grynberg M, Kajava AV, Promponas VJ et al. 2019. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res* **47**: 10994-11006.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511-515.

Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13**: 329-342.

Zickmann F, Renard BY. 2015. IPred - integrating ab initio and evidence based gene predictions to improve prediction accuracy. *BMC Genomics* **16**: 134.