GeneMark-ETP: Automatic Gene Finding in Eukaryotic Genomes in Consistency 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

New large scale initiatives, such as the Earth BioGenome Project, require efficient automatic tools for eukaryotic genome annotation. A new automatic tool, GeneMark-ETP, presented here, finds genes by integration of genomic-, transcriptomic- and protein-derived evidence. GeneMark-ETP first identifies genomic loci where extrinsic data is sufficient for gene prediction with 'high confidence' and then proceeds with finding the remaining genes across the whole genome. The initial set of parameters of the statistical model is estimated on the training set made from the high confidence genes. Subsequently, the model parameters are iteratively updated in the cycles of gene prediction and parameter re-estimation. Upon reaching convergence GeneMark-ETP makes the final prediction of the whole complement of genes. The algorithm development was made with a focus on large plant and animal genomes. GeneMark-ETP performance was compared favorably with the ones of the gene finders using a single type of extrinsic evidence delivered by either short RNA reads (GeneMark-ET), or by mapped to genome homologous proteins (GeneMark-EP+). These outcomes could be expected. Moreover, comparisons were made with the pipelines utilizing both transcript- and protein-derived extrinsic evidence. For these experiments we have chosen TSEBRA, combining BRAKER1 and BRAKER2, as well as MAKER2. The results demonstrated that GeneMark-ETP delivered state-of-the-art gene prediction accuracy with a large margin of improvement in large eukaryotic genomes.

Introduction

Initiatives aiming to massive sequencing of genomes of eukaryotic species, e.g., the Earth BioGenome Project (Lewin et al. 2022) require accurate automatic tools of genome annotation. Therefore, the development of computational methods of gene prediction continues to be an important area of research. One of the long-standing issues is how to integrate the intrinsic, ab initio, evidence of protein-encoding by DNA sequence with the extrinsic evidence for a gene presence, the one derived from mapped to genome transcript or protein sequences. *Ab initio* gene finding algorithms could rely on k-mer frequency patterns, models of splice site and start/stop site motifs, intron/exon length distributions, etc., all embedded in an HMM type model (e.g., Genie (Kulp et al. 1996), GENSCAN (Burge and Karlin 1997), GeneID (Parra et al. 2000), AUGUSTUS (Stanke and Waack 2003), SNAP (Korf 2004), GeneMark-ES (Lomsadze et al. 2005)). Automatic *ab initio* gene finders can deliver sufficiently accurate predictions of protein-coding genes in fungal and protist genomes (Lomsadze et al. 2005; Ter-Hovhannisyan et al. 2008). Addition of the extrinsic evidence, however, was necessary to improve the accuracy of gene prediction in large eukaryotic genomes having low gene density (Guigo et al. 2006; Coghlan et al. 2008; Goodswen et al. 2012; Scalzitti et al. 2020). It was recognized that the approaches based *solely* on extrinsic evidence could only identify subsets of the whole gene complement in a genome. Particularly, cross-species protein splice alignments, employed in *exonerate* (Slater and Birney 2005), GenomeThreader (Gremme et al. 2005), or ProSplign (Kiryutin et al. 2007), helped identify genes whose protein products were homologous to database proteins. On the other hand, finding a group of genes having detectable levels of expression could be facilitated by mapping to genome of either short RNA-Seq reads or whole assembled transcripts, e.g. with the help of Cufflinks (Trapnell et al. 2010), StringTie (Pertea et al. 2015; Kovaka et al. 2019), or PsiCLASS (Song et al. 2019).

Among other recent tools we should mention BRAKER1 (Hoff et al. 2016) and BRAKER2 (Bruna et al. 2021). BRAKER1 includes GeneMark-ET (Lomsadze et al. 2014) and AUGUSTUS (Stanke et al. 2008) and works to integrate genomic and transcript data. BRAKER2 includes AUGUSTUS (Stanke et al. 2008) and GeneMark-EP+ (Bruna et al. 2020) and delivers gene predictions by integration of genomic and protein data.

Still, the larger task is to automatically and efficiently integrate all the three streams of evidence. Several tools were developed to address this problem, e.g., GAZE (Howe et al. 2002), Combiner (Allen et al. 2004), JIGSAW (Allen and Salzberg 2005), Evigan (Liu et al. 2008), EVidenceModeler (Haas et al. 2008), MAKER2 (Holt and Yandell 2011), IPred (Zickmann and Renard 2015), GeMoMa (Keilwagen et al. 2018), LoReAn (Cook et al. 2019), GAAP (Kong et al. 2019), or FINDER (Banerjee et al. 2021).

The majority of the above mentioned tools work as *combiners.* First, a combiner generates sets of genome-wide gene predictions: *ab initio*-, transcriptomic-, and mapped proteins-based. Next, at a post-processing step, these sets are combined into a final set of predictions.

Alternatively, the integration of the three types of evidence is done upon prediction of each gene. GeneMark-ETP, integrates genomic, transcriptomic, and protein information throughout *all* the stages of the algorithm's training and gene prediction (Fig. 1). Gene prediction is made not only in genomic DNA, but, also, in long transcripts assembled from RNA-Seq reads. Protein-coding regions predicted in transcripts are verified, and, if necessary, modified as a result of comparison of the initially predicted proteins with homologous cross-species proteins. In the absence of transcript evidence for a gene situated in genomic DNA, hints to the locations of introns and exons are generated by mapping of homologous proteins to the genome locus. Accounting for repetitive sequence and transposable elements plays a significant role as well. The estimation of parameters of the statistical models used in GeneMark-ETP (generalized hidden Markov models, GHMM) is made by iterative training.

Tests of the GeneMark-ETP performance were done on both GC-homogeneous and GC-heterogeneous eukaryotic genomes. Comparisons were made with GeneMark-ET, GeneMark-EP+ as well as with their virtual combinations. Further comparisons were made with the pipelines BRAKER1 (Hoff et al. 2016), BRAKER2 (Bruna et al. 2021), MAKER2 (Holt and Yandell 2011) as well as with TSEBRA (Gabriel et al. 2021), a tool combining the BRAKER1 and BRAKER2 gene predictions. The tests demonstrated state-of-the-art performance GeneMark-ETP with the margin of improvement over other tools increasing with the genome length.

Results

Assessment of the GeneMark-ETP prediction accuracy and comparison with other gene finding methods

To assess the gene prediction accuracy of GeneMark-ETP (see Fig. 1) we used 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 annotations as the ground truth. For the four larger genomes, *S. lycopersicum*, *D. rerio*, *G. gallus*, and *M. musculus* conservative estimates of the gene prediction sensitivity (recall) were computed on the sets of genes present in both NCBI and Ensembl annotations (see Methods) while the specificity (precision) values were computed for the unions of the NCBI and Ensemble annotations.

We have observed a significant increase in gene prediction accuracy of GeneMark-ETP in comparison with both GeneMark-ET and GeneMark-EP+. Moreover, an improvement was also reached in comparison with both BRAKER1 and BRAKER2.

The most notable changes of Sn and Sp values occurred in the large genomes, both the GC-homogenous *S. lycopersicum*, *D. rerio*, and, especially, the GC-heterogeneous ones *G. gallus* and *M. musculus* (Figs. 2, 3, S1; Tables S1, S2). As a single parameter characterizing gene prediction accuracy, we used the gene level F1 value (F1=TP/(TP+(FN+FP)/2)), ranging from zero to one, or expressed as percentage points. For the three groups of compact, large homogeneous, and large

heterogeneous genomes the absolute increase in F1 reached by GeneMark-ETP over GeneMark-ET was, on average, 19.6%, 47.8%, and 66.3% respectively, while the increase of F1 over GeneMark-EP+ was, respectively, 14.2%, 33.9%, and 55.7% (Table S1). In comparison with BRAKER1 (using RNA-seq reads as extrinsic evidence) the positive changes in the gene level F1 were 9.7%, 29.2%, and 58.5%, on average, in each of the three groups, respectively. In comparison with BRAKER2 (using mapped homologous proteins as extrinsic evidence) the increases in F1 in the three groups were, respectively, 11.2%, 21.9%, and 47.6%. These outcomes should have been expected, since each of the tools taken for comparison was using a single source of extrinsic evidence, either the evidence of expression - the native RNA reads or evidence of evolutionary conservation - homologous cross-species proteins.

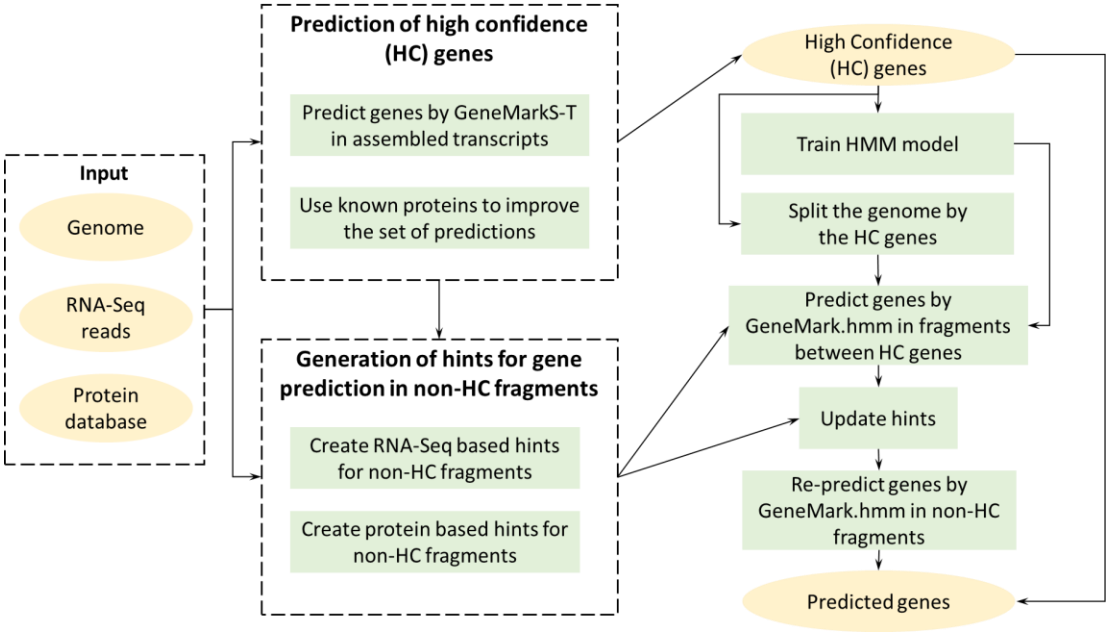

Figure 1. High-level diagram of GeneMark-ETP

Other comparisons were made with the tools using two types of extrinsic evidence.

First, we considered virtual combinations of the sets of gene predictions made separately by GeneMark-ET and GeneMark-EP+. These could be the union, the intersection, and the 'ideal' combination of the two sets of predicted genes made by removing false positives from the union set (see Methods and Fig S2). For both compact and large genomes, the accuracy of GeneMark-ETP was observed to be higher than the accuracy of the three virtual combinations. Particularly, for *D. melanogaster S. lycopersicum* and *G. galus*, the changes in F1 in comparison with the 'ideal' combination were 10.1%, 18.0%, and 56.6% respectively.

In comparison with TSEBRA creating a combination of the two sets of predictions made by BRAKER1 and BRAKER2 by filtering out less supported predictions from the union of the two sets of predicted genes, the gene level F1 values were improved by GeneMark-ETP for the three groups of genomes by 1.0%, 8.2%, and 39.0% (Figs. 2, 3; Table S2), respectively.

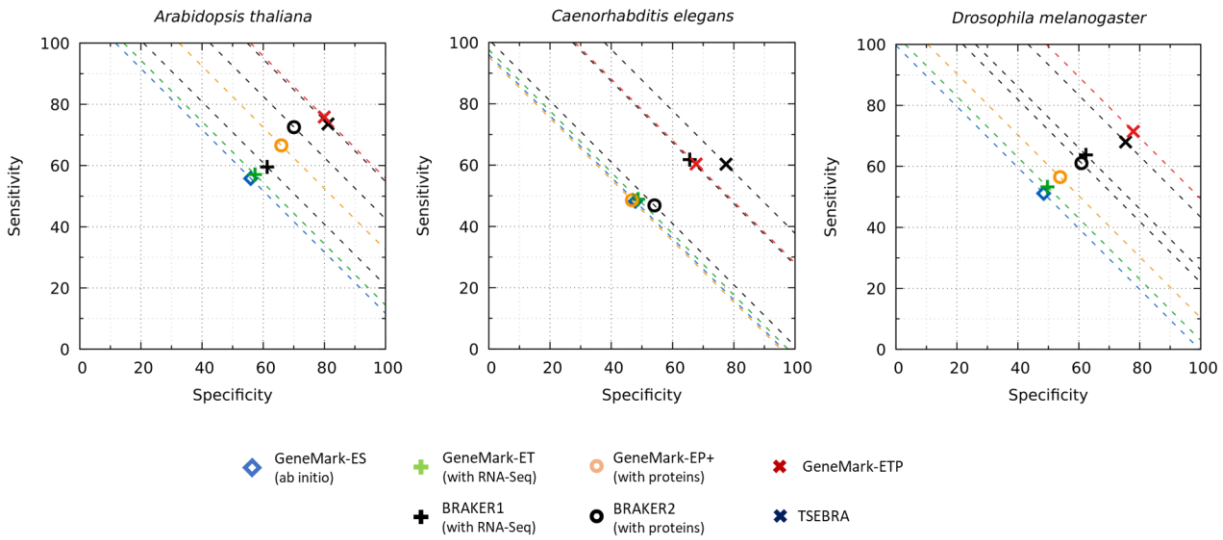

Figure 2. Gene level Sensitivity (Sn) and Specificity (Sp) of GeneMark-ETP for the three compact genomes. The dashed lines correspond to constant levels of (Sn+Sp)/2. The species-specific protein databases used for derivation of protein-based evidence did not include proteins from the species of the same taxonomic order. 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.

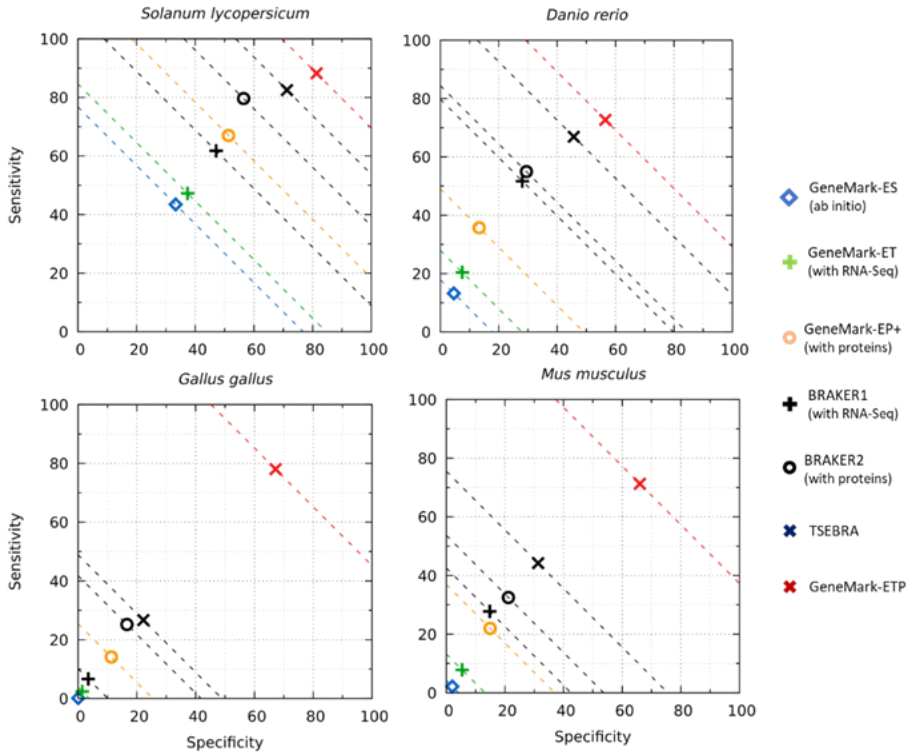

Figure 3. Gene level Sn and Sp of GeneMark-ETP for the four large genomes. All other specifications are the same as in Fig. 2.

One more genome annotation tool, MAKER2, uses a combination of the three gene finders, AUGUSTUS, SNAP and GeneMark.hmm. As extrinsic evidence MAKER2 uses transcript and protein data. The comparisons with MAKER2 were done for *D. melanogaster, D. rerio* and *M. musculus*. For all the three species the GeneMark-ETP accuracy was higher than the one of MAKER2 (Table S3). Particularly, for *D. melanogaster, D. rerio* and *M. musculus,* the GeneMark-ETP gene level F1 values were higher by 23.3%, 21.7%, and 27.8%, respectively.

Results of initial integration of available extrinsic evidence: transcripts and proteins

Gene predictions in assembled transcripts were made by the *ab initio* gene finder GeneMarkS-T. This automatic tool creates its own models by unsupervised training on a set of transcripts (Tang et al. 2015). GeneMarkS-T gene predictions were verified with respect to available protein data and modified if necessary. Particularly, if an alignment of the protein product of a predicted gene to a protein in a database showed a better support to an alternative gene start within the transcript, the initial gene prediction was altered (Methods Section 2). In our computational experiments, we demonstrated that for a reference protein database compiled for a given species, whether of the smaller or of the larger size, this correction increased the gene-level specificity, on average, by 25 percentage points, reaching values close to or even higher than 90% (Tables 1, S4). Thus, refined genes were dubbed *high-confidence genes*.

It should be underlined that the set of the initial GeneMarkS-T gene predictions and the set of the HC genes may overlap but are not subsets of each other.

Table 1. The gene-level Sn and Sp values determined for i/ the initial GeneMarkS-T gene predictions in the assembled transcripts and for ii/ the high-confidence (HC) genes obtained by modification of the initially predicted genes due to the results of comparison to homologous proteins. Reference protein database for each species did not contain proteins of the species from the same *taxonomic order* (see Materials). For a full set of results see Tables S4 and S5.

Species/ Accuracy		Initial GeneMarkS-T predictions	The set of HC genes
C. elegans	Sn	**46.8**	35.7
	Sp	63.4	**88.4**
A. thaliana	Sn	51.2	**56.7**
	Sp	79.9	**97.3**
D. melanogaster	Sn	**59.6**	55.0
	Sp	81.8	**94.7**
S. lycopersicum	Sn	67.8	**74.9**
	Sp	77.8	**95.2**
D. rerio	Sn	59.6	**67.0**
	Sp	59.9	**88.5**
G. gallus	Sn	49.6	**74.4**
	Sp	47.0	**89.1**
M. musculus	Sn	49.6	**63.5**
	Sp	63.2	**93.2**

Relative contribution of extrinsic and intrinsic evidence to gene prediction

We divided the whole set of genes predicted in a given genome into the following categories: *fully extrinsic*: all elements of the exon-intron structure were supported by significant (high scoring) extrinsic evidence; *partially extrinsic*: with significant extrinsic evidence for some gene structure elements while other had been predicted *ab initio*; *ab initio anchored* 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*: a gene was predicted *ab initio* while there was no even *a posteriori* support for any of the gene structure elements.

Table 2. Distribution of predicted genes among the four categories along with average Sp values (gene level) for each category. Descriptions of the species-specific protein databases (smaller and larger) 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	7,676	88.9	10,778	91.6
	Partially extrinsic	4,804	56.4	5,417	54.4
	Ab initio anchored	4,020	54.7	1,548	45.2
	Ab initio unsupported	1,298	24.9	778	18.0
A. thaliana	Fully extrinsic	16,445	97.2	18,083	97.5
	Partially extrinsic	4,825	64.4	5,807	55.7
	Ab initio anchored	1,794	50.2	1,360	30.1
	Ab initio unsupported	2,964	27.9	1,128	9.4
D. melanogaster	Fully extrinsic	8,059	95.1	9,952	96.8
	Partially extrinsic	2,328	49.3	2,751	44.9
	Ab initio anchored	1,043	57.1	165	44.9
	Ab initio unsupported	1,369	41.6	377	15.9
S. lycopersicum	Fully extrinsic	17,639	95.2	18,420	95.0
	Partially extrinsic	5,174	47.3	5,813	44.3
	Ab initio anchored	1,577	38.4	1,484	29.7
	Ab initio unsupported	4,714	14.8	3,703	9.2
D. rerio	Fully extrinsic	15,691	89.8	15,501	92.6
	Partially extrinsic	10,905	16.6	11,769	16.6
	Ab initio anchored	1,973	11.4	1,663	7.3
	Ab initio unsupported	12,534	0.8	11,879	0.3
G. gallus	Fully extrinsic	11,856	89.3	11,547	89.9
	Partially extrinsic	4,857	19.6	5,337	20.1
	Ab initio anchored	527	8.9	579	7.1
	Ab initio unsupported	11,332	0.4	11,352	0.3
M. musculus	Fully extrinsic	13,556	94.6	13,769	96.2
	Partially extrinsic	7,376	20.6	7,606	19.6
	Ab initio anchored	957	10.1	1,155	7.3
	Ab initio unsupported	20,711	1.2	19,666	0.5

We observed that the reliability of gene predictions was 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 1.5% (Table 2) and exon level Sp

below 3% (Table S6). When, for these four genomes, we removed such predictions from the reported lists of predicted genes, we found that the gene-level Sp increased, on average, by 21% while Sn decreased by 0.3% (Table S7). For the three compact genomes, such a filtering did increase, on average, the gene level Sp by 3.7% while Sn was lowered by 1.7%.

Discussion

GeneMark-ETP as a gene finder was designed to integrate intrinsic, genomic, and two types of extrinsic evidence, RNA and proteins. Reaching this aim required solving multiple tasks. One of them was iterative training of parameters of the two GHMM models that were used i/ in the algorithm of gene prediction in assembled transcripts or ii/ in the algorithm for gene prediction in genomic DNA. The method of training of the first GHMM did follow the path described for GeneMarkS-T (Tang et al. 2015). However, the method of training of the GHMM model for the gene prediction across the whole genome was different from the methods described for GeneMark-ES, GeneMark-ET and GeneMark-EP+. In those tools the initial values of parameters of the GHMM model were defined by the functions approximating dependence of the k-mer frequencies on genome GC content (Lomsadze et al. 2005; Lomsadze et al. 2014; Bruna et al. 2020). In GeneMark-ETP the initial values of the GHMM parameters were trained on the sequences of the loci containing the HC genes. Then, GeneMark-ETP iterated over gene prediction and parameter re-estimation steps until convergence is reached (Fig S3). It was observed that if more than 4,000 HC genes were found in the initial step of the HC gene identification, then the model derived from the set of HC genes would not change significantly in the further iterations. Such an outcome was due to reaching stationary values of the parameter estimates with respect to the training set size.

An important feature of the GHMM training implemented in GeneMark-ES, GeneMark-ET and GeneMark-EP+, was step by step unfreezing of the subsets of the GHMM model parameters. For instance, the transition probabilities between hidden states, i.e., intron, exon etc., as well as distributions of durations of hidden states, were fixed during the initial iterations while the values of emission probabilities, derived from the k-mer frequencies, 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 where all the GHMM parameters were estimated at once. We attribute this streamlining of the training process to having more accurate initial parameters of GHMM derived from the sequences of the HC loci.

In what follows we discuss the algorithmic steps that make GeneMark-ETP different from the other tools mentioned above.

Identification of a set of genes predicted with high confidence.

The task of transcript assembly from short RNA-Seq reads has presented a well-known challenge (Steijger et al. 2013). The assembly errors affected downstream analysis, particularly gene

prediction. However, recently developed short read assembly tools, such as StringTie2 (Kovaka et al. 2019), demonstrated a significantly improved performance.

Within GeneMark-ETP, gene prediction in assembled transcripts was done by GeneMarkS-T (Tang et al. 2015). Predicted protein products were searched for similarity against a protein database. The full-length high scoring hits identified the uniformly supported genes, named *high-confidence* (HC) *genes.* In our tests, the HC genes had, on average, a significantly better match to the 'true' (annotated) genes than the genes in the initial set of GeneMarkS-T predictions (Tables 1, S4). Many initially predicted 5' partial genes were modified based on the high scoring protein hits (see Methods). However, some of the 5' end partial predictions were included in the set of HC genes as well.

Genes predicted with lower confidence.

We observed that with increase in the genome length the fractions of genes predicted with extrinsic support, full or partial, diminished, e.g., from 96% for *D. melanogaster* to 51% for *M. musculus* (Table 2, larger reference databases). At the same time, the fractions of the genes predicted *ab initio* did grow from 4% for *D. melanogaster to* 49% for *M. musculus*.

Importantly, with the genome length increase the *fraction of false positives* among the genes predicted *ab initio* grew even faster than the fraction of the *ab initio* predicted genes. This trend produced a significant drop in Specificity (Tables 2, S6). The growth in the rate of false positive predictions could be expected due to an increase in the average length of intron and intergenic regions, an increase in the size of populations of transposable elements (repeats), an increased frequency of pseudogenes, etc.

The specificity of *unsupported ab initio* predictions was reaching down below 10% (Table 2, the fourth category). In genomes larger than 300 Mbp, such genes comprised more than 10% of the total number of predictions. In the final lists of gene predictions made in genomes larger than 300 Mbp, we eliminated *unsupported ab initio* predictions. As a result, the genes on the list would have at least one gene structure element supported by an extrinsic evidence which was either used in the prediction or detected to exist *a posteriori*.

Using the GC-content-specific models

To make more accurate gene predictions in *GC-heterogeneous* genomes, we switched from a single statistical GHMM model to multiple GC-content-specific models. In general, GeneMark-ETP could be run in the "GC-heterogeneous" mode on any genome. However, for a true *GC-homogeneous* genome, this choice is likely to increase runtime and decrease 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 GeneMark-ETP is run in the "GC-heterogeneous" mode if necessary.

Processing of repetitive elements

Transposable elements or repetitive sequences that, most frequently, do not carry the 'host' genes or even exons could be masked to reduce the search space for gene prediction and improve the accuracy (Lomsadze et al. 2014; Bruna et al. 2020). While being quite reasonable, such an approach has disadvantages: (i) repeats shorter than a chosen threshold T would remain unmasked and (ii) protein-coding exons overlapped by the masked regions would be difficult to find.

To address this issue, the authors of AUGUSTUS have introduced a penalty function used in the Viterbi algorithm (Stanke et al. 2008). The function decreased the likelihood of prediction of protein coding regions inside the masked regions. A single parameter of this function, *q*, had the same value for all species. We have shown that the use of a *species-specific* parameter *q* would improve the algorithm performance (see Methods). The value *q* for a given genome was determined (trained) on the set of the HC genes used as the development set (Table S8, see Methods). The value *q* found by the application of this approach was shown to be robust with respect to the size of the set of the HC genes (data not shown).

Comparison with other gene finders

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

As it could be expected, in all the tests GeneMark-ETP performed better than either GeneMark-ET or GeneMark-EP+, the tools using only a single source of extrinsic evidence (see Results).

To raise the comparison bar, we considered a virtual combination of the sets of genes predicted separately by GeneMark-ET and GeneMark-EP+ (see the Method Section). The largest sensitivity of such a combination could be achieved with the *union* of the two sets of predicted genes while the largest specificity could be delivered by using the *intersection* of the two sets of predicted genes. Yet, the best overall accuracy could be achieved by either removal of false positives from the *union* of the two sets, or addition of true positives to the *intersection* of the two sets of gene predictions. These modifications cannot be implemented when a gene finder is running on a novel genome since information on true and false positives is not immediately available. Nevertheless, these ideal modifications could be made for the genomes used for the tests, as, arguably, a reliable annotation of these genomes was known. We observed that even the accuracy of the ideal combination of the gene predictions of the two complementary tools still fell below the accuracy of GeneMark-ETP (Fig. S2).

BRAKER1, BRAKER2, TSEBRA, and MAKER2

Earlier developed automatic gene finding pipeline BRAKER1 (Hoff et al. 2016), combining AUGUSTUS and GeneMark-ET, used experimentally determined transcripts as a source of extrinsic evidence. The pipeline BRAKER2 (Bruna et al. 2021), combining AUGUSTUS and GeneMark-EP+, relied on finding and mapping homologous cross-species proteins to the genome of interest. We have shown that the gene prediction accuracy of GeneMark-ETP was higher than

either BRAKER1 or BRAKER2, especially for large genomes (Figs. 1, 2). This result could be expected since GeneMark-ETP could use both types of extrinsic information.

On the other hand, the TSEBRA pipeline (Gabriel et al. 2021) selects a subset of all predictions made by either BRAKER1 or BRAKER2 and, thus, achieves higher accuracy than any of the two BRAKERs. It was shown that TSEBRA performed better than EVidenceModeler (Haas et al. 2008), one of the best combiners, as well.

We demonstrated that while GeneMark-ETP has an accuracy comparable with TSEBRA for compact genomes (Fig. 2), it achieved higher accuracy than TSEBRA in large genomes (Fig. 3), particularly in the GC-heterogeneous ones (*G. gallus*, *M. musculus*). For these genomes, BRAKER1 and BRAKER2 used single statistical models tuned up for genome-specific "average GC". Nonetheless, GeneMark-ETP reached higher than TSEBRA prediction accuracy in the GC-homogeneous genomes as well, i.e., for *S. lycopersicum* and *D. rerio*. An additional source of accuracy improvement in GeneMark-ETP was, arguably, the use of hints generated by concerted analysis of the assembled transcripts and homologous proteins.

In addition, we demonstrated that the GeneMark-ETP gene prediction accuracy was higher than the one of MAKER2 in the tests on the three genomes of model species (Table S3). All the gene finders employed in MAKER2 were run with parameters that, arguably, corresponded to the best-case scenario of training (see Comparison with MAKER2 in the Results).

Observed differences in performance could be caused by the difference in i/ the selection of the front-end gene finders, ii/ in methods of generating the transcriptome and protein-based hints, as well as in iii/ the methods of the evidence integration into gene predictions.

Also, it should be noted that for a GC-heterogeneous genome of *M. musculus* GeneMark-ETP used the models with the GC specific parameters. In MAKER2, the GC specific parameters were used in AUGUSTUS but not in SNAP or GeneMark.hmm.

The new tool, GeneMark-ETP, integrates transcriptomic and homologous protein extrinsic evidence into local hints that are used consistently at *all* the stages of the algorithm training and gene prediction. We have demonstrated in the tests that such an approach is more effective than the combiner approach, when the sets of gene predictions made with a single type of extrinsic evidence are combined in a "post-processing" step.

Materials

To assess the gene prediction accuracy, we selected genomes of the seven eukaryotic species (Tables 4, S9). These were *A. thaliana*, *C. elegans,* and *D. melanogaster*, the well-studied model organisms with GC-homogeneous and compact in size genomes. There were also *S. lycopersicum*, and *D. rerio* with large GC-homogenous genomes, as well as *G. gallus* and *M. musculus* having large GC-heterogeneous genomes. In all the genomic datasets, sequences of organelles and contigs without chromosome assignment were excluded from the analysis.

Table 4. Genomic, transcriptomic and protein data used for assessment of gene prediction accuracy. For the four species with large genomes, the numbers in parentheses characterize subsets of genes presumed to be more reliably annotated (see text). To compute the numbers of introns per gene we used averages among annotated alternative transcripts.

Species	Genome length (Mb)	Reference annotation					
		# 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)

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 protein database (PD_0) containing the proteins from the largest clade selected for the given species (Table S10). Next, for each species, we created two reference databases by removing from PD_0 either i/ all proteins of this very *species* and its strains, or ii/ proteins of all the species from the same taxonomic *order*. These, *the larger* and *the smaller* databases, were devised to reflect practical scenarios when a species in question would appear on either a *larger* or a *smaller* evolutionary distance from the species present in the reference database. Overall, the numbers of proteins in the reference databases used in this study ranged from 2.6 to 8.3 million (Table S10).

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 S11).

Methods

Outline of the GeneMark-ETP workflow

In the earlier developed automatic gene finders, GeneMark-ES, -ET, -EP+, estimation of the parameters of the GHHM models was done by iterative unsupervised training (Lomsadze et al. 2005; Lomsadze et al. 2014; Bruna et al. 2020). At the end of iterations, a final set of parameters was used to make the final gene predictions. In the case of GeneMark-ETP the training procedure had significant differences.

In GeneMark-ETP we predict genes in both genomic DNA and assembled transcripts. Gene prediction in transcripts uses GHMM designed for sequences with intron-less genes while gene prediction in genomic DNA requires GHMM with an exon-intron model (Lomsadze et al. 2005). Genes are predicted in transcripts by an earlier developed self-trained tool GeneMarkS-T (Tang et al. 2015), and the predictions are further refined by GeneMarkS-TP (see Suppl. Methods). At

this point, GeneMark-ETP generates a set of *high-confidence* (HC) genes (Fig. S4). Next, the initial parameters of the 'eukaryotic' GHMM are estimated. The initial parameters of the GHMM model are further refined by iterative training in the genomic regions situated between HC genes. The final set of parameters defined upon convergence of the training is used to predict genes in the regions between the HC genes. The use of the transcript and protein extrinsic evidence continues in all the prediction steps (Fig. 1).

Selection of a set of high confidence genes.

Initial gene prediction in assembled transcripts

Short reads of each RNA-Seq library 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). The low-abundance transcripts are filtered out and the remaining transcripts are merged by StringTie2 into the non-redundant transcriptome (Kovaka et al. 2019). Since StringTie2 assembles transcripts from RNA-Seq reads *anchored* to genome (as a result of the read mapping to genome), the information on intron positions remains in the system and is carried out to determine the exon/intron structure of a gene in terms of genomic co-ordinates.

The GeneMarkS-T intron-less gene predictions in assembled transcripts appeared to have very few errors when the predicted genes were complete genes. On the other hand, the 5' partial genes predicted by GeneMarkS-T carried more frequent errors. The errors could be corrected with the aid of the data on homologous proteins.

Verification and modification of 5' partial gene predictions (GeneMarkS-TP).

Some GeneMarkS-T predictions could be corrected based on protein information. Incomplete, 5' partial genes start from the first nucleotide of a transcript. A 5' partial gene is erroneously predicted if a true complete gene is residing inside this 5' partial gene. To verify the 5' partial gene prediction two alternative protein translations are used as queries in against a reference protein database by DIAMOND (Buchfink et al. 2015). One query corresponds to the full length 5' partial gene and another one to a gene which ATG start is the closest to the 5' end (in the same reading frame as the first one). As a result of the similarity searches there may exist a protein target common for both queries searches (with E-value 10^{-3}). In this case we check if a stronger evolutionary support is present for the 5' partial prediction (condition S1 in Suppl. Methods). If the condition S1 is fulfilled then the 5' partial gene is predicted, otherwise, a complete gene starting with the internal ATG is selected as a modified prediction. The 5' partial gene is predicted also if the sets of protein targets in the two searches (those with 25 best scores) do not overlap. When both similarity searches do not produce targets, the transcript is removed from consideration. The details of the algorithm of the gene prediction verification and correction (that we call GeneMarkS-TP) are described in Section 1.1 of Suppl. Methods.

As a result of this verification and modification procedure, we arrive to a set of genes G predicted in transcripts that contains complete genes predicted by GeneMarkS-T, verified 5' partial genes

predicted by GeneMarkS-T and the set of modified initial predictions consisting of verified complete genes.

Compilation of a set of high-confidence genes

Complete and partial HC genes with uniform protein support

Genes in the whole set G could be divided into complete and partial genes. We select from this set a subset of genes having protein support along the whole gene.

A complete gene predicted by GeneMarkS-TP is said to have *uniform protein support* if alignment of its protein to a protein in a database satisfies condition S2 (Fig. S5, see Suppl. Methods). The search for the target protein is done by DIAMOND (in blastp mode) with the gene translation as query. Further on, 25 top-scoring hits are examined for fulfillment of condition S2. If at least one target protein makes an alignment with the query satisfying S2, then this gene is classified as a *complete high-confidence* gene.

A complete gene predicted in a transcript may not make the "longest ORF" with respect to its 5' end. However, it was observed that most of the time eukaryotic genes do make the longest ORFs. To correct possible under prediction, the gene extension to the "longest ORF" is checked for fulfillment of condition S2 and, if fulfilled, both variants are kept as candidates for alternative complete HC isoforms.

A predicted 5' partial gene (Fig. S6, S7) has a uniform protein support if the C-terminal of the protein translation is supported by at least one protein alignment and a condition of the same kind as condition S2 is fulfilled (see Suppl. Methods). Such genes are selected to be *partial HC genes.*

A gene predicted as 3' partial (lacking a stop codon) cannot be selected as an HC gene.

HC genes predicted ab initio on a transcript level.

Complete GeneMarkS-T gene predictions without protein support along the whole gene (condition S2 not fulfilled) still could be classified as HC genes. These predictions have to satisfy the following conditions: (i) the length of predicted CDS is longer than $299\ nt$, (ii) transcript 5' UTR contains in-frame stop codon triplet, (iii) the GeneMarkS-T log-odds score is larger than zero, and (iv) the gene structure mapped to genomic DNA does not create a conflict with a ProtHint hint (see Section S2 of Suppl. Methods).

Alternative HC isoforms

Alternative isoforms may be present in the set of HC genes. Selection of alternative HC 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 protein database. We compute the maximum score of all the complete isoforms

for the gene g, denoted as $s(g_{complete})$. The score of an isoform $s(i)$ selected as complete HC isoform must satisfy the inequality:

$$s(i) \geq 0.8 \times s(g_{complete}) \quad (i \in I^{g}_{complete})$$

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

The GHHM model training

Single step model training

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

For the *GC homogeneous* genome, the loci of all the selected HC genes (the sequences containing the HC genes with 1,000nt margins) are used for estimation of the GHHM parameters (Fig. S8).

For the *GC heterogeneous* genome, the set of *HC loci* sequences is split into three GC bins: low GC, mid GC, and high GC. The left border of the mid GC bin with a fixed width (9% by default) is chosen to include into the GC bin the largest possible number of the HC loci. Setting up the mid GC interval immediately determines the low and high GC bins. The three sets of the HC loci are used to train the three GC-specific GHHM models.

Note that by design GeneMarkS-T predicts genes in transcripts by a set of the GC-specific statistical models (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).

For *GC homogeneous GC*, at the initialization of iterations, the GHMM model parameters are derived from the sequences of the HC loci. The new round of gene prediction is then made only in the genomic sequences situated between HC genes, *the HC-intermediate regions*. Description of the steps that follow is given below in the section 'Gene prediction in HC-intermediate regions'.

For *GC heterogeneous* genomes the extended GHHM training works as follows. First, GeneMark-ETP calculates the GC content of each HC-intermediate region and assigns the regions to the corresponding GC bins. The parameters of the initial GC specific GHHM model are trained on the

sequences of the HC loci of a corresponding bin. Subsequently, a GC specific model is used for gene prediction in the HC-intermediate regions of the same GC bin. From this point on, the extended training on the HC-intermediate regions in each GC bin is made in the same way as in the GC homogeneous case. Thus, the GC-specific models are used in the iterative training and gene prediction up to the final gene prediction step.

Accounting for repeats

Repeat identification.

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 be caused by repetitive elements with composition similar to protein-coding regions (Yandell and Ence 2012; Torresen et al. 2019). Identification of the repetitive sequence could be done independently from gene finding. This step, *de novo* repeat finding, helps to reduce the gene prediction errors. Still, some of the predicted repeats may overlap with the protein-coding genes (Bayer et al. 2018). We generated species-specific repeat libraries *de novo* using RepeatModeler2 (Flynn et al. 2020). Repetitive sequences—interspersed and tandem repeats—were then identified and soft-masked by RepeatMasker (www.repeatmasker.org).

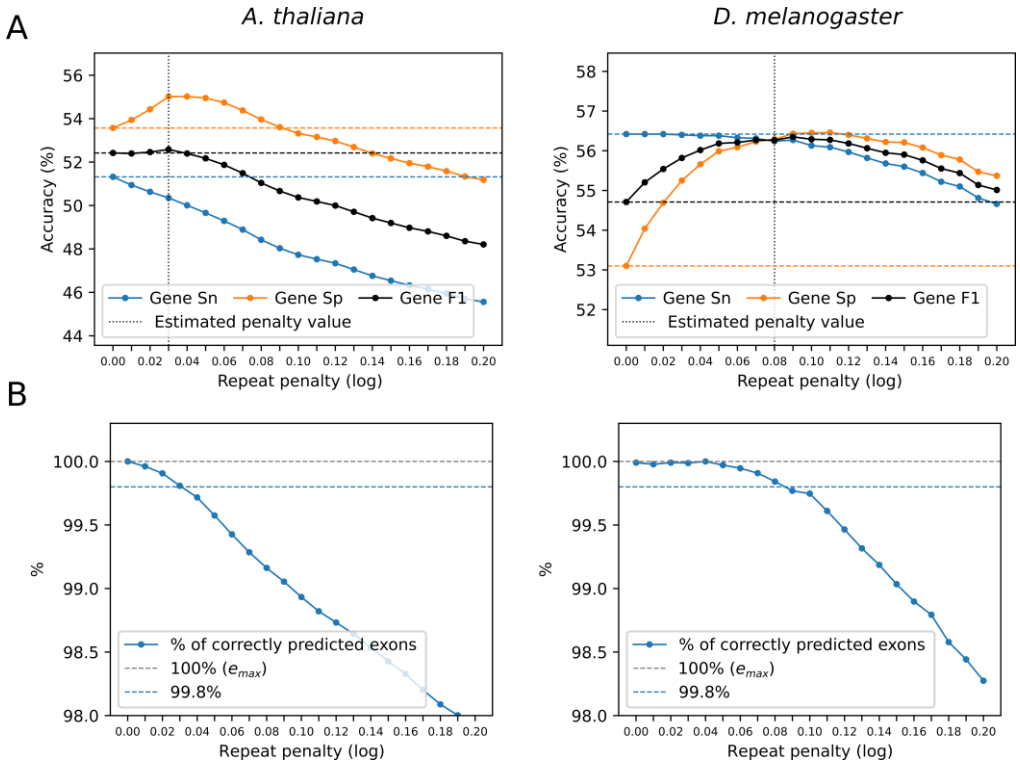

Figure 4. A. Dependence of the gene Sn, Sp and F1 values on the repeat penalty parameter *q* for genomes of *A. thaliana* and *D. melanogaster*. **B.** Dependence of fraction (%) of correctly predicted exons of *the HC genes* (Sn) on the repeat penalty parameter *q* for the same genomes as in A.

Selection of the species-specific repeat penalty

To account for an overlap of a protein-coding region with a repetitive sequence, GeneMark-ETP introduces a state for the overlap and computes the probability of a sequence (of length *n*) that would appear in such an overlap:

$$P(seq|coding\ state\ overlapping\ repeat) = \frac{P(seq|coding\ state)}{q^n} \quad (1)$$

For each genome, GeneMark-ETP estimates species-specific parameter q minimizing errors in gene predictions. The estimation of q is made after the first iteration of the model training. At this point there is a set of the HC genes, a version of the full GHMM model, and the coordinates of the repeats identified in genomic DNA. Next, the Viterbi algorithm is run in an *ab initio* mode with several different *q* values 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 prediction accuracy of the HC genes. The value *q* for which F1 would reach maximum (Fig. 4A) turns out to be a good approximation of the value *q* found if the test set would be made with precise knowledge of genome annotation (data not shown). Moreover, the use of the exon level Sn in such computations led to more robust estimation of *q* in comparison with use of the gene level F1 (data not shown). Technically, we first find the q' value maximizing the number of correctly predicted exons in the set of HC genes, e_{max} (Fig. 4B). Then, we use the value *q** at which $0.998 \times e_{max}$ exons are correctly predicted (marked in panel A of Fig. 4). To reduce the runtime of the repeat penalty parameter estimation we use 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. S8). 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 RNA-seq based hints to introns mapped to genome by HISAT2 (Kim et al. 2019). These hints provide transcript-based evidence for the HC intermediate regions. The whole set of hints is then ready for enforcement in a run of GeneMark.hmm. All over, 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; 4/ partial HC genes. Note that hints of categories 1-3 point to separate elements of a gene (the RNA-Seq mapped introns or the ProtHint introns). Hints of category 4 represent 'chains' that should belong to the same gene. The requirement of the overlap that appears in description of hints of category 3 is introduced to filter out introns mapped from expressed non-coding RNA, e.g., lncRNA. These hints are *enforced* in HC-intermediate regions in the cycles of gene prediction and model training. The iterations stop when the identity of the training sets in two consecutive iterations reaches 99%. The number of

iterations observed in our experiments was rarely above three, while, frequently, GeneMark-ETP converged in the second iteration. 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.

In addition, we expect that if the high-quality long reads or their assemblies are available, GeneMarkS-T and GeneMarkS-TP refinement routine could be run on the long reads to predict and modify intron-less genes. Thus predicted genes would be mapped to the genome, e.g. with Minimap2 (Li 2018), to create HC genes candidates. Thus, this type of data could be used by GeneMark-ETP.

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 unsupported *ab initio* predictions (Table 2). The evidence-supported genes must have at least one element of the gene structure supported externally. The *support* here implies either presence of hints used in the process of gene prediction or a match to an extrinsic evidence detected in an *a posteriori* analysis (see Discussion Section 'Genes predicted with lower confidence'. We have observed that in large genomes the fractions of correct predictions among *unsupported ab initio* predictions were sharply decreasing with the increase in genome size (Sp values in Table 2). For each genome, the lists of genes predicted by GeneMark-ETP contain either the *full set* of predicted genes or the set of genes remaining after filtering out *unsupported ab initio* predictions from the full set. The reported here gene prediction accuracy for large genomes (longer than 300 Mbp) was computed for the reduced outputs.

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 were 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, i.e., in the NCBI and the Ensembl records (Table S9). On the other hand, the values of gene prediction specificity were defined by comparison with the union of genes annotated by either NCBI or Ensembl.

Statistics of the sets of genes used for accuracy assessment are given in Table 2. In all the tests, regions of annotated pseudogenes were excluded from consideration.

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

We compared GeneMark-ETP with 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 straightforward ways: by making either union U or

intersection I. The intersection I contains only genes from both sets that have identical gene structures. The set U contains the most comprehensive set of predicted genes, while the set I, arguably, contains the set of 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 the incorrect predictions, the point for *U* in Fig. 5 will move horizontally. If one can add to the set I only correct predictions the point for *I* in Fig. 5 will move up vertically. The crossing of the two lines at the point (Sn, Sp) characterizes the accuracy of the best virtual combiner approach.

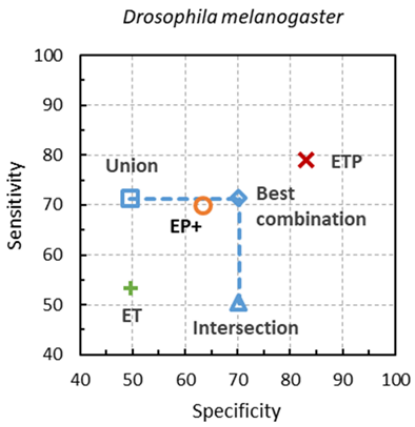

Figure 5. Gene-level Sn and Sp of the artificial combinations of GeneMark-ET and GeneMark-EP+ gene predictions made in the *D. melanogaster* genome are shown along with Sn and Sp of GeneMark-ETP. The proteins of the *D. melanogaster* species were excluded from the reference protein database PD_0.

Running BRAKER1, BRAKER2, TSEBRA, and MAKER2

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

Execution of MAKER2 has some degree of freedom as the rules of training of AUGUSTUS, SNAP and GeneMark.hmm are not specified. Therefore, we wanted to generate such gene prediction accuracy figures for MAKER2 that would be recognized as upper bounds among possible options for training. Thus, the models used for AUGUSTUS and SNAP were either the models provided by the code developers (available with the respective software distribution) or generated by supervised training on the genes given in the GenBank annotation. Both MAKER2 and GeneMark-ETP use the GeneMark.hmm gene finder. MAKER2 uses the version of GeneMark.hmm self-trained by GeneMark-ES, thus, having no use of extrinsic evidence in model training. Assuming that more precise models would be obtained with addition of extrinsic evidence into training, we

prepared the GeneMark.hmm models trained on a set of high confidence genes (determined by GeneMark-ETP). To get the accuracy figures, the gene predictions made by MAKER2 were processed in the same way as the ones made by GeneMark-ETP or other gene finders on the genomes of *D. melanogaster, D. rerio* and *M. musculus*.

MAKER2 was executed in the default mode with gene finders AUGUSTUS, SNAP, and GeneMark.hmm. AUGUSTUS and SNAP were run with models provided in the corresponding distributions, except that in the genome of *D. rerio* the SNAP predictions were made with the model trained on genes annotated by Ensembl. GeneMark.hmm was run with models trained on the high confidence genes obtained by GeneMark-ETP. The repeat coordinates, RNA-seq and protein data sets were the same for MAKER2 and GeneMark-ETP (see details in the Supplementary materials).

Summary

A new eukaryotic gene prediction software tool, GeneMark-ETP was shown to generate more accurate gene predictions in comparison with the earlier developed tools. The improvement over other tools was shown to be more pronounced for larger genomes, both GC-homogenous and GC-heterogeneous. All over, we demonstrated the advantage of the simultaneous integration of several sources of extrinsic evidence into gene prediction over a post-processing-style integration which combines 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. GeneMark-ETP was included in the recently developed pipeline BRAKER3 (Gabriel et al. 2023).

Funding

National Institutes of Health [GM128145, 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, Bruna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, Stanke M. 2023. BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *bioRxiv* doi:10.1101/2023.06.10.544449: 2023.2006.2010.544449.

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.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**: 491.

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.
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.
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.
Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Balint M, Barker KB, Baumgartner B, Belov K, Bertorelle G et al. 2022. The Earth BioGenome Project 2020: Starting the clock. *Proc Natl Acad Sci U S A* **119**.
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.
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.