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

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

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

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

Abstract

New large-scale genomic initiatives, such as the Earth BioGenome Project, require efficient methods for eukaryotic genome annotation. A new automatic tool, GeneMark-ETP, presented here, finds genes by integration of genomic-, transcriptomic- and protein-derived evidence. The algorithm was developed with a focus on large plant and animal genomes. 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 rounds of gene prediction and parameter re-estimation. Upon reaching convergence, GeneMark-ETP makes the final predictions of the whole complement of genes. The GeneMark-ETP performance was expectably better than the performance of GeneMark-ET or GeneMark-EP+, the gene finders using a single type of extrinsic evidence, either short RNA-seq reads or mapped to genome homologous proteins. Subsequently, for comparisons with the tools utilizing both transcript- and protein-derived extrinsic evidence, we have chosen MAKER2 and a more recent tool, TSEBRA, combining BRAKER1 and BRAKER2. The results demonstrated that GeneMark-ETP delivered state-of-the-art gene prediction accuracy with the margin of outperforming existing approaches increasing for larger and more complex eukaryotic genomes.

Introduction

New initiatives aiming for massive sequencing of genomes of eukaryotic species, e.g., the Earth BioGenome Project (Lewin et al. 2022) require accurate and efficient tools of genome annotation. The development of such tools remains an active area of research. One of the long-standing challenges is to achieve optimal integration of the intrinsic, *ab initio*, evidence of protein-encoding by a nucleotide sequence with the extrinsic evidence of gene presence derived from mapped to genome RNA or protein sequences. Gene finding methods developed prior to the advent of NGS and the transition to massive genomic sequencing have been heavily focused on using intrinsic evidence. *Ab initio* gene finders 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)). Among these methods we should mention GeneMark-ES, which reached full automation by implementation of iterative unsupervised training (Lomsadze et al. 2005; Ter-Hovhannisyan et al. 2008). The accuracy achieved by this *ab initio* method in fungal and protist genomes is difficult to improve, even by the addition of all the available extrinsic evidence. However, with an increase in genome length and decrease in gene density typical for plant and animal genomes, the accuracy of pure *ab initio* methods deteriorates. The addition of extrinsic evidence plays critical role for large genomes (Guigo et al. 2006; Coghlan et al. 2008; Goodswen et al. 2012; Scalzitti et al. 2020). Splice alignments of cross-species proteins were used in several tools, e.g. in exonerate (Slater and Birney 2005), GenomeThreader (Gremme et al. 2005), and ProSplign (Kiryutin et al. 2007). The goal of such methods was to find genes whose protein products were homologous to known proteins. On the other hand, RNA sequences carry yet another type of extrinsic evidence. The tools such as Cufflinks (Trapnell et al. 2010), StringTie (Pertea et al. 2015; Kovaka et al. 2019), or PsiCLASS (Song et al. 2019) were developed to map to genome short RNA-seq reads and, thus, help identify exon borders of genes with detectable levels of expression. Nevertheless, we should state that an approach based *solely* on extrinsic evidence has its limitations; in each genome it could reliably identify only a subset of the whole gene complement.

The goal of improving the integration of all three types of evidence continues to motivate the development of new gene finding methods for more than two decades, 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), and FINDER (Banerjee et al. 2021).

Recent gene finding tools combining intrinsic and extrinsic evidence, such as BRAKER1 (Hoff et al. 2016) and BRAKER2 (Bruna et al. 2021), have attempted to excel in automation along with accuracy. Automatic BRAKER1 integrates genomic and transcript data in a pipeline containing

GeneMark-ET (Lomsadze et al. 2014) and AUGUSTUS (Stanke et al. 2008). Automatic BRAKER2 integrates genomic and protein data in a pipeline, including AUGUSTUS and GeneMark-EP+ (Bruna et al. 2020). Both BRAKER1 and BRAKER2 have been combined into a single pipeline TSEBRA (Gabriel et al. 2021).

The gene finding method described here, GeneMark-ETP, integrates genomic, transcriptomic, and protein information in several stages (Fig. 1). First, the protein-coding regions are predicted in transcripts assembled from RNA-seq reads by an earlier developed *ab initio* method, GeneMarkS-T (Tang et al. 2015). The predicted intronless CDS sequences are validated or modified if the protein level evidence suggests so. The set of thus predicted CDS sequences is mapped to the genome and defines gene models with high specificity, the high-confidence or HC genes. The set of HC genes is large enough to be used as an initial training set for the generalized hidden Markov model (GHMM), which is used and updated in further iterations by GeneMark-ETP. Prediction of the genes situated in genomic segments between the HC genes is done by combining a stream of external hints into the Viterbi algorithm for GHHM. The hints to exon borders of a gene model in a given locus are generated by spliced alignments of the homologous cross-species proteins found in a protein database with the addition of information from short RNA reads mapped to the same locus. The rounds of gene prediction and model parameters estimation continue until convergence.

For benchmarking of the new method, we have selected seven eukaryotic genomes, both GC-homogeneous and GC-heterogeneous: *Arabidopsis thaliana*, *Caenorhabditis elegans, Drosophila melanogaster, Solanum lycopersicum*, *Danio rerio*, *Gallus gallus*, and *Mus musculus*. Along with GeneMark-ETP we tested other gene prediction tools. Particularly, we made comparisons with GeneMark-ET, GeneMark-EP+ and their 'virtual combination', with the pipelines BRAKER1, BRAKER2, MAKER2 as well as with TSEBRA combining BRAKER1 and BRAKER2. The tests demonstrated the state-of-the-art performance of GeneMark-ETP, whose margin of improvement over other tools was increasing with the increase in the genome length.

Results

Evidence integration in the initial steps of the algorithm

The gene prediction accuracy of GeneMark-ETP and other gene prediction tools was characterized by Sensitivity (or Recall), Sn = Tp/(Tp+Fn), Specificity (or Precision), Sp = Tp/(Tp+Fp), where Tp, Fp and Fn are the numbers of true positive, false positive and false negative gene predictions, respectively, and the F1 score, F1 = 2(Sn)(Sp)/(Sn+Sp).

The initial step of the data analysis in the GeneMark-ETP pipeline (Fig. 1) was assembling short RNA-seq reads into transcripts by StringTie2 (Kovaka et al. 2019). Next, the assembled transcripts were used as input for iterative unsupervised training of GeneMarkS-T (Tang et al. 2015) done simultaneously with the rounds of *ab initio* gene predictions in the same transcripts. Running GeneMarkS-T was the first step in the GeneMarkS-TP module generating a set of high confidence gene predictions, the HC genes, based on transcript and protein evidence available for the corresponding genomic loci (see Methods).

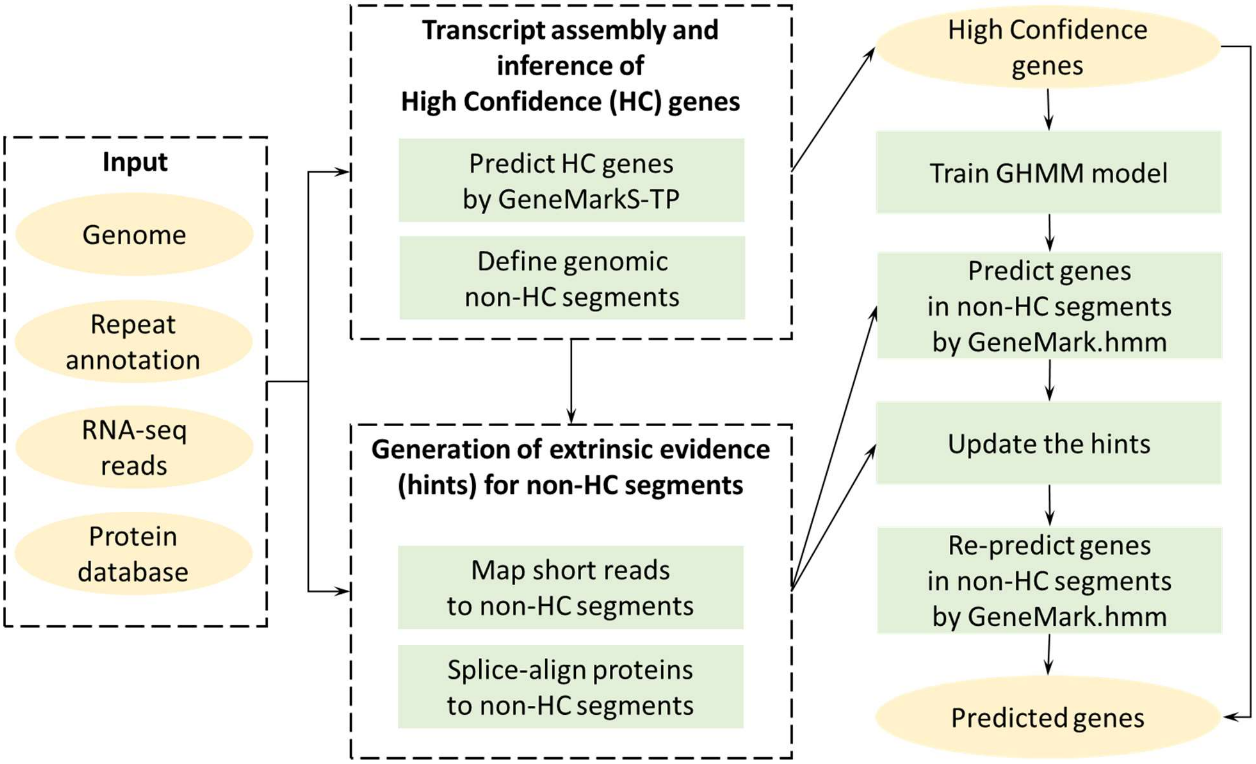
Input
Genome
Repeat annotation
RNA-seq reads
Protein database
Transcript assembly and inference of High Confidence (HC) genes
Predict HC genes by GeneMarkS-TP
Define genomic non-HC segments
Generation of extrinsic evidence (hints) for non-HC segments
Map short reads to non-HC segments
Splice-align proteins to non-HC segments
High Confidence genes
Train GHMM model
Predict genes in non-HC segments by GeneMark.hmm
Update the hints
Re-predict genes in non-HC segments by GeneMark.hmm
Predicted genes

Figure 1. Conceptual diagram of the GeneMark-ETP processing of genomic, RNA-seq, and protein data. Due to limited space, many important details are omitted (Supplemental Figs. S1A-S1G).

Both the numbers of GeneMarkS-T gene predictions in transcripts and the numbers of predicted HC genes were in correlation with the numbers of annotated genes (Fig. 2 and Table 1). The sets of the HC genes predicted in each genome were large enough to make between 1/3 to 2/3 of the final set of the predicted genes.

The intronless CDS predicted by GeneMarkS-T in transcripts were translated, and the amino acid sequences were aligned with homologous cross-species proteins. The alignments guided the corrections made by GeneMarkS-TP in the predicted CDS regions. The magnitude of improvement in Sn and Sp depended on the size of database proteins used by GeneMarkS-TP (Table 1). Thus corrected CDS sequences were splice-aligned to genome to delineate exon-intron structures, the high-confidence (HC) genes named so due to the high Specificity. With respect to the mapped to the genome initial prediction made by GeneMarkS-T, the Specificity values of the

HC gene predictions increased, on average, by 25 percentage points, reaching close to or even higher than 90% (Table 1).

The numbers of predicted HC genes determined with support of the larger 'Species excluded' databases did not differ significantly from the numbers of HC genes predicted with the use of the smaller 'Order excluded' databases (Table 1 and Fig. 2). The Specificity (Sp) values did not change significantly upon transition from the smaller to the larger database (except *D. melanogaster*).

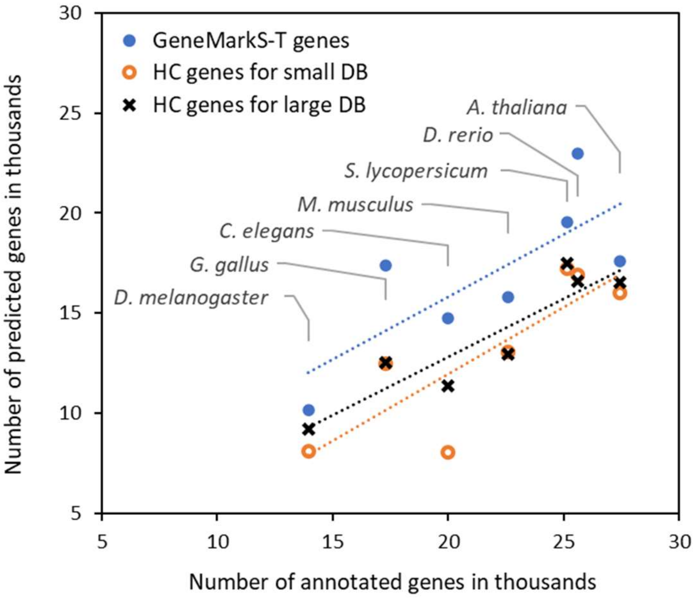

Figure 2. Sizes of gene sets predicted at preliminary stages of running GeneMark-ETP (i) genes predicted in assembled transcripts by GeneMarkS-T (blue dots), (ii) HC genes predicted by GeneMarkS-TP with small protein database (orange circles) and with the large protein database (black crosses).

In order to focus on results, we have to skip a detailed description of the algorithm (see Methods and the illustrated graphics shown in Supplemental Figs. S1A-S1G). At the step preceding the generation of the final set of gene predictions we have a large set of gene predictions obtained and corrected at all the intermediate steps. This set of predicted genes could be divided into: (i) fully extrinsic predictions having all the exon borders supported by a significant (high scoring) extrinsic evidence; (ii) partially extrinsic predictions, with significant extrinsic evidence for some exon borders; (iii) gene predictions with detected extrinsic match, the genes predicted *ab initio* and having an a posteriori detected match of some exon borders to an extrinsic evidence; (iv) *ab initio* gene predictions with no extrinsic match, for which an a posteriori extrinsic support is not detected for any of the gene borders. The HC genes belong to the first two categories.

One could see that for all the species, the Specificity values decrease significantly upon a decrease in the level of extrinsic support (Table 2), thus indicating an increase in false positive rates.

Table 1. Summary statistics characterizing the sets of the GeneMarkS-T gene predictions in assembled transcripts as well as the sets of the high-confidence gene predictions (HC genes). Two versions of reference protein databases were used for each species: the database called 'Species excluded', containing all the proteins from an OrthoDB segment but proteins from the same species, as well as the smaller database called 'Order excluded' containing all the proteins from the same OrthoDB segment but proteins from the same taxonomic order (see Materials). Additional data is provided in Supplemental Table S1.

Species	# of annotated genes	# of genes predicted by GeneMarkS-T	Sn/Sp of GeneMarkS-T predicted genes	# of HC genes (Order excluded DB)	Sn/Sp of HC genes (Order excluded DB)	# of HC genes (Species excluded DB)	Sn/Sp of HC genes (Species excluded DB)
C. elegans	19,969	14,746	46.8/63.4	8,062	35.7/88.4	11,399	51.7/90.6
A. thaliana	27,445	17,589	51.2/79.9	16,008	55.0/94.7	16,551	58.8/97.6
D. melanogaster	13,951	10,163	59.6/81.8	8,109	59.6/81.8	9,223	63.7/96.3
S. lycopersicum	25,158	19,526	67.8/77.8	17,231	74.9/95.2	17,489	75.8/95.1
D. rerio	25,611	22,992	59.6/59.9	16,918	67.0/88.5	16,573	66.9/90.4
G. gallus	17,279	17,381	49.6/47.0	12,473	74.4/89.1	12,564	74.0/88.4
M. musculus	22,611	15,819	49.6/63.2	13,057	63.5/93.2	12,965	63.9/94.5

In the largest genomes of *D. rerio*, *G. gallus,* and *M. musculus*, the gene predictions that had no minimal match in extrinsic data had, on average, Sp values below 1.5% at the gene level and below 3% at the exon level (Supplemental Table S2). In general, the growth of the false positive rate in large genomes is expected due to the increase of the average length of non-coding regions as well as the increase of numbers of repetitive sequences carrying no host genes. For the four largest genomes, removal of the predictions with no extrinsic match from the GeneMark-ETP output led to an increase in the gene-level Sp, on average, by 21% with a simultaneous decrease in Sn by 0.3% (Supplemental Table S3). For the three compact genomes, *A. thaliana*, *C. elegans*, *D. melanogaster*, such a change in the output led to the increase, on average, the gene level Sp by 3.7% and a decrease of Sn by 1.7%. This observation justifies the removal of the intermediate gene predictions with no extrinsic match from the GeneMark-ETP output for sufficiently long eukaryotic genomes (longer than 300 Mb, the default threshold).

Assessment of the gene prediction accuracy

The task of improving gene finding accuracy is challenging in compact genomes where the changes in Sn and Sp are observed to be incremental from the tools developed earlier to those developed later (Fig. 3). However, the double-digit changes of the Sn and Sp values occurred in the large genomes, both GC-homogenous *S. lycopersicum*, *D. rerio*, and, especially, GC-heterogeneous *G. gallus* and *M. musculus* (Fig. 4, Supplemental Fig. S2; Supplemental Tables S4, S5).

Table 2. Distribution of the intermediate sets of predicted genes among the four categories characterized by the degree of connection to extrinsic data. The average Sp values (gene level) are given for the genes of each category. Descriptions of the species-specific protein databases (the smaller one – 'Order excluded' and the larger one – 'Species excluded') are given in Materials.

Species	Gene predictions	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
	With extrinsic match	4,020	54.7	1,548	45.2
	With no extrinsic match	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
	With extrinsic match	1,794	50.2	1,360	30.1
	With no extrinsic match	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
	With extrinsic match	1,043	57.1	165	44.9
	With no extrinsic match	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
	With extrinsic match	1,577	38.4	1,484	29.7
	With no extrinsic match	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
	With extrinsic match	1,973	11.4	1,663	7.3
	With no extrinsic match	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
	With extrinsic match	527	8.9	579	7.1
	With no extrinsic match	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
	With extrinsic match	957	10.1	1,155	7.3
	With no extrinsic match	20,711	1.2	19,666	0.5

The increases in gene level F1 achieved by GeneMark-ETP over GeneMark-ET (utilizing RNA-seq reads to augment the training process) were, on average, 19.6, 47.8, and 66.3 points in the three groups of compact, large homogeneous, and large heterogeneous genomes, respectively. The increases of F1 over GeneMark-EP+ (using cross-species proteins to support both training and prediction) were, respectively, 14.2, 33.9, and 55.7 points (Supplemental Table S4). GeneMark-

ETP also had shown improvements in F1 scores over BRAKER1 and BRAKER2, the pipelines where GeneMark-ET and GeneMark-EP+, respectively, play key roles. In comparison with BRAKER1, the positive changes in F1, on average, 9.7, 29.2, and 58.5 points, in the same three groups of genomes. In comparison with BRAKER2, the F1 values were increased by 11.2, 21.9, and 47.6 points.

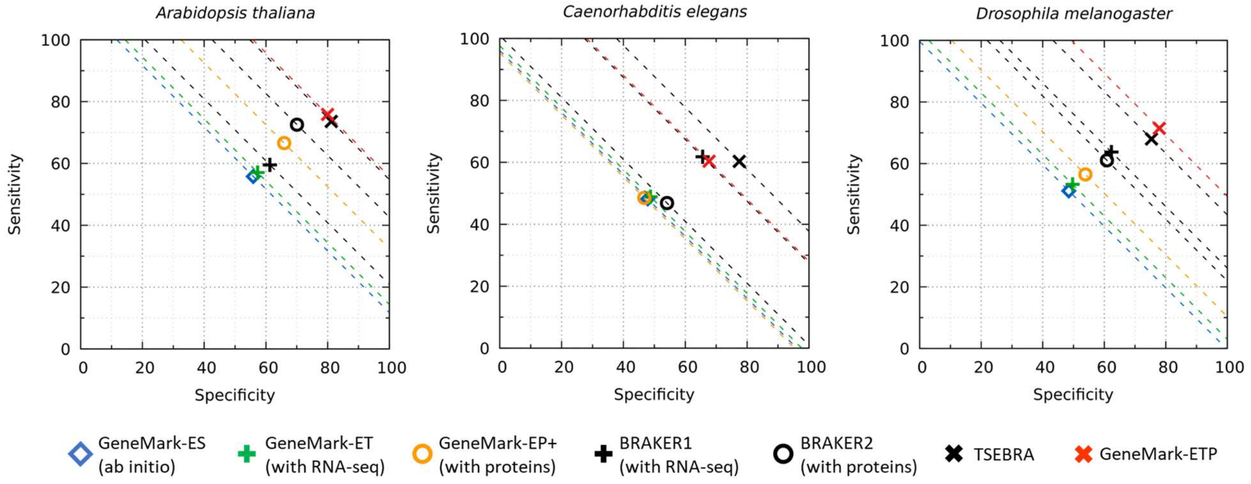

Figure 3. Gene level Sensitivity and Specificity of GeneMark-ETP and six other gene prediction tools for the three compact genomes. The dashed lines correspond to constant levels of (Sn+Sp)/2. The 'Order excluded' protein databases were used. The true positives were defined with respect to the complete set of genes annotated in each genome.

The observed changes quantify the advantage of the tool integrating two sources of extrinsic evidence over the tools and pipelines using just a single source. More challenging comparisons come next.

First, we considered virtual combinations of the sets of gene predictions made separately by GeneMark-ET and GeneMark-EP+ (see Methods). The union of the two sets would increase Sn with a decrease of Sp, while the sets intersection would have larger Sp along with decrease of Sn (Fig. 5, the case of *D. melanogaster*). The 'ideal' combination of the two sets could be made by removing false positives from the union set or adding true positives made by either method to the intersection (see Methods). When the union set is changed by taking away the incorrect predictions, the point for $Union$ in Fig. 5 moves horizontally to the right. If one could add to the intersection set only correct predictions made by one of the tools but not the other, the point for $Intersection$ in Fig. 5 moves move up vertically. The crossing of the two lines characterizes the accuracy of the best virtual combination of the two sets to predicted genes (Fig. 5).

In terms of F1 score, the difference between GeneMark-ETP and the 'ideal' combination were 10.1, 18.0, and 56.6 points for genomes of *D. melanogaster*, *S. lycopersicum*, and *G. gallus*, respectively.

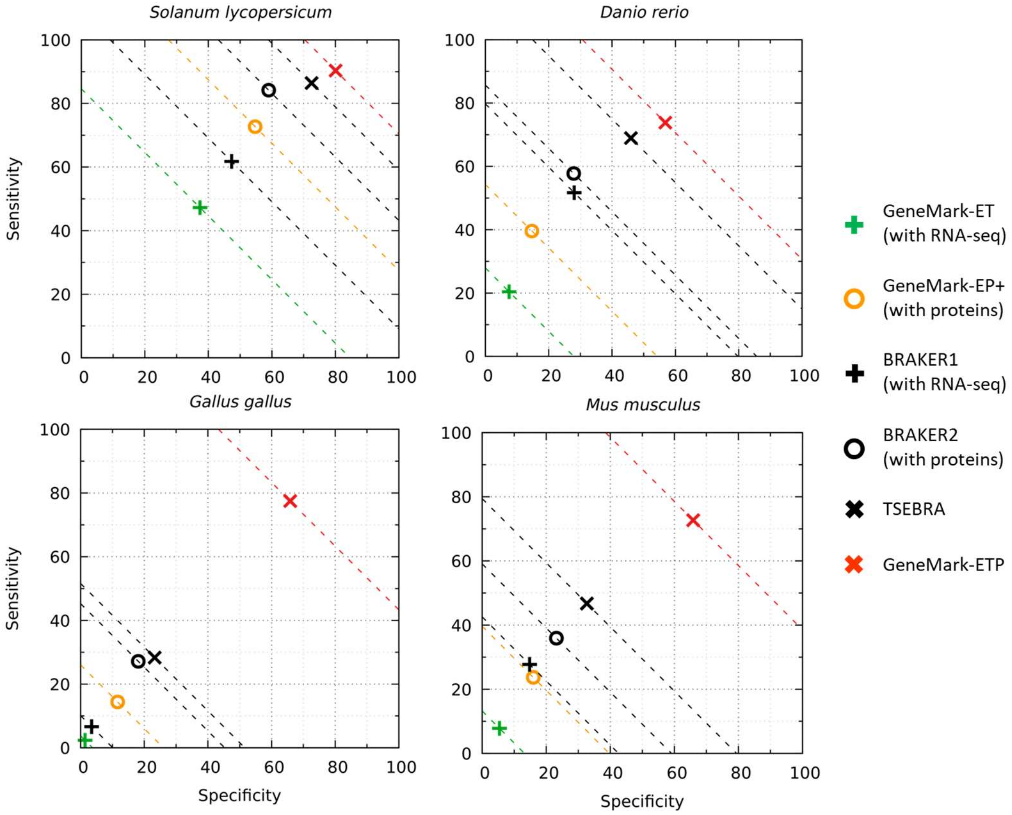

Figure 4. Gene level Sn and Sp of gene predictions made by GeneMark-ETP and five other gene prediction tools in the four large genomes. The Sn and Sp values were defined with respect to selected subsets of genes from genome annotation (see text). In these genomes (with length > 300 Mb), GeneMark-ETP output did not include genes with no match to extrinsic support (see text). The 'Order excluded' protein databases were used.

Finally, the comparisons were made with the gene finding methods TSEBRA (Gabriel et al. 2021) and MAKER2 (Holt and Yandell 2011). TSEBRA combines predictions made independently by BRAKER1 (supported by RNA-seq evidence) and BRAKER2 (supported by evidence inferred from homologous cross-species proteins) by filtering out less supported predictions from the union of the sets of predicted genes generated by the two pipelines. In comparison with TSEBRA the gene level F1 values of GeneMark-ETP were comparable in the group of the compact genomes while, on average, were better by 8.2 and 39.0 points, respectively, for the large GC-homogeneous genomes, and the large GC-inhomogeneous genomes, (Figs. 3, 4; Supplemental Table S5).

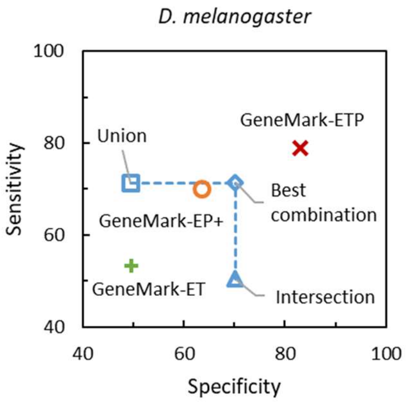

Figure 5. Gene-level Sn and Sp of gene predictions made in the *D. melanogaster* genome by different methods. The true positives were defined with respect to the complete annotation of *D. melanogaster* genome. The 'Species excluded' database was used in GeneMark-EP+ and GeneMark-ETP.

The frequently used MAKER2 pipeline runs AUGUSTUS, SNAP, and GeneMark-ES. As extrinsic evidence, MAKER2 uses transcript and protein data. Runs of MAKER2 and GeneMark-ETP were done for *D. melanogaster*, *D. rerio,* and *M. musculus* (see Methods for details). For all three species the GeneMark-ETP performance was better than the one of MAKER2 (Supplemental Table S6). The gene level F1 improvement was observed for all three genomes as follows: 23.3, 21.7, and 27.8 points for *D. melanogaster*, *D. rerio,* and *M. musculus*, respectively.

Discussion

The GeneMark-ETP algorithm is efficiently using previously developed constructs and approaches such as anchored exon borders (GeneMark-ET), external hints generated from multiple protein spliced alignments (GeneMark-EP+), unsupervised training of the GHMM model (GeneMark-ES). However, a distinct feature of GeneMark-ETP is the generation of a set of high confidence gene predictions by the GeneMarkS-TP module. These predictions do not change in the subsequent steps of the analysis when the genes in genomic segments between HC genes are predicted.

The initial set of the HC gene candidates was generated by GeneMarkS-T. We demonstrated that after further modifications and selections, the set of the generated HC gene predictions had much higher Specificity than the initial set of the candidates (Fig. 2, Table 1). The number of HC gene predictions made for the 'Species excluded' databases did not differ significantly (except *D. melanogaster*) from the numbers of HC genes predicted with the 'Order excluded' databases (Table 1 and Fig. 2). Also, the Specificity (Sp) values did not change significantly upon transition to the larger database (except *D. melanogaster*). This observation indicated that more distant proteins carry sufficient information to convert initial HC gene candidates into a set of high confidence genes, the sets with close to 90% Specificity. Therefore, the estimates of the accuracy for the sets of HC genes generated with smaller databases, 'Order excluded', will be adequate estimates for novel genomes produced for the species having limited, if any, number of closely related species with sequenced and annotated genomes.

The patterns of change in Sensitivity depended on the structure of the protein database used by GeneMarkS-TP. For instance, when the 'Order excluded' databases were used, the Sensitivity of the HC gene predictions in the *C. elegans,* or *D. melanogaster* genomes decreased in comparison with the Sensitivity of the initial candidates. However, with the larger 'Species excluded' database, the set of the HC genes had higher Sensitivity values than the initial GeneMarkS-T predictions for all seven species (Table 1).

It was interesting to analyze the patterns of dependence of the gene prediction accuracy on the level of extrinsic support (see Methods). In all seven genomes, we observed a large difference in Specificity between gene predictions having full extrinsic support and those with partial extrinsic support (Table 2). The magnitude of the difference was 30-40% in the compact genomes as well as in the tomato genome while in the larger genomes *D. rerio, G. gallus* and *M. musculus* the difference was more than 50%. Notably, the Specificity of gene predictions with the second and the third categories of the external support level decreased as the size of the reference database increased. This change is explained by a move of the many gene predictions made in these two categories with a smaller protein database to the highest category, with full extrinsic support, when the larger database is used.

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

As 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 constructed an 'optimal' combination of the sets of genes predicted separately by GeneMark-ET and GeneMark-EP+ (see Methods).

The comparison on *D. melanogaster* genome shows the GeneMark-ETP is increasing both Sensitivity and Specificity roughly by 10% above the values reached by the best virtual combination (Fig. 5). One of the sources of the performance improvement available to GeneMark-ETP, but not to the other two tools, is linking the intron hints into groups bound to the same gene as a result of processing of gene predictions made in assembled transcripts.

Direct comparisons with the performance of GeneMark-ET or GeneMark-EP+ were made for the seven genomes (Figs. 3, 4, Supplemental Table S4). We comment here on the performance improvements that GeneMark-ETP made above the best of the two tools, measuring the improvement in the F1 points at the gene level. For compact genomes, F1 improved by 11-19 points, and for large genomes – by 26-60 points, with the maximum on the *G. gallus* genome.

The higher accuracy among GeneMark-ET or GeneMark-EP+ (running with the 'Order excluded' database) was most frequently observed for GeneMark-EP+, which has a mechanism for enforcing protein derived external hints into the gene predictions, while GeneMark-ET used the RNA-seq derived hints only in training. If the 'Species excluded' database is used (Supplemental Table S4), then the gap in F1 on gene level between GeneMark-ETP and GeneMark-EP+ would decrease by up to 4 points even though the absolute value of F1 reached by GeneMark-ETP increased. These absolute values of F1 for GeneMark-ETP were standing between 64.2 (*D. rerio*) and 84.6 (*S. lycopersicum*).

BRAKER1, BRAKER2 and TSEBRA

Comparisons of the GeneMark-ETP performance with ones of the pipelines BRAKER1 and BRAKER2 carry clear parallels with the comparisons made with GeneMark-ET or GeneMark-EP+. Each of the two pipelines uses a single type of external evidence, either RNA-seq reads or proteins. While BRAKER1 outperforms GeneMark-ET and BRAKER-2 outperforms GeneMark-EP+, the crucial components of the respective pipelines, GeneMark-ETP shows even better performance, especially for the four large genomes (Figs. 3, 4, Supplemental Table S5). This result is not surprising since GeneMark-ETP uses both types of extrinsic information.

On the other hand, the TSEBRA pipeline is a strong competitor. It selects a subset of all predictions made by either BRAKER1 or BRAKER2 by the rules that increase Specificity without compromising Sensitivity (Gabriel et al. 2021). TSEBRA 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. Interestingly, for *A. thaliana* and *C. elegance,* TSEBRA shows slightly better performance than GeneMark-ETP, with improvement at the gene level F1 by 1.5 and 4 points. TSEBRA F1 measure for *D. melanogaster* is lower than the one for GeneMark-ETP by 0.5 points. Further comparisons show higher GeneMark-ETP F1 performance by 6.4, 9.0, 45.6, and 29.4 points for the larger genomes: *S. lycopersicum*, *D. rerio*, *G. gallus,* and *M. musculus*, respectively (Supplemental Table S5). The double digits of improvement in F1 are observed for GC inhomogeneous genomes. For these genomes, BRAKER1 and BRAKER2 used single statistical models tuned up for genome-specific "average GC". An additional factor facilitating the improvement of gene prediction accuracy of GeneMark-ETP is the construction of hints produced in the concerted processing of the transcripts and homologous cross-species proteins.

MAKER2

We should note the uniform, double-digit improvement of the F1 metrics in the tests made for comparison of GeneMark-ETP and MAKER2 (Supplemental Table S6). The three genomes of model species *D. melanogaster, D. rerio*, and *M. musculus* represented compact, large GC homogeneous and GC inhomogeneous groups. The improvements in gene level F1 were 23.3, 21.7, and 27.8 points, respectively. 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). We used limited volumes of protein (see Suppl. Materials Section S5) as the runtime of MAKER2 with a larger protein database, such as 'Order excluded' would be prohibitive. Notably, a comparison of the results of GeneMark-ETP in this set of experiments with the runs of GeneMark-ETP where the larger databases, 'Order excluded' were used for the same three species, shows that the 10x increase in the database size brings about an F1 improvement of about 5.0 points for *D. melanogaster* and *M. musculus,* does not make a noticeable change in F1 for *D. rerio.*

Difference with the previously developed GeneMark tools

GeneMark-ETP was designed to integrate in an optimal way the intrinsic and 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 the GHMM model for 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 iterates over gene prediction and parameter re-estimation steps until convergence is reached (Supplemental Fig. S1F). 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 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 the ability to streamline the training process to having more accurate initial parameters of GHMM derived from the sequences of the HC loci.

BRAKER3

Since 2015 we have closely collaborated with a group of Mario Stanke on development of the gene prediction pipelines known as BRAKERs (Hoff et al. 2016; Bruna et al. 2021; Gabriel et al. 2023). The reason to initiate the collaboration was that the gene finding tools created by our groups have had complementary strengths. The research behind the GeneMark-ES, -ET, -EP algorithms was focused on machine learning methods of self-training. The comprehensive AUGUSTUS algorithm needs extensive training sets on which to perform supervised discriminative training. A combination of the two approaches may not follow a linear, sequential path that could be associated with the architectures of BRAKER1 and BRAKER2. In TSEBRA, being a union of BRAKER1 and BRAKER2, the two sets of gene predictions made by each pipeline are combined by yet another algorithm to increase Sensitivity without compromising Specificity. A similar approach is implemented in BRAKER3, being a union of GeneMark-ETP and AUGUSTUS. The current paper describing GeneMark-ETP is important for understanding of the forthcoming publication on BRAKER3 focused on multi-layer integration of GeneMark-ETP with AUGUSTUS.

Methods

Data Sets

For computational experiments with GeneMark-ETP, we selected genomes of the seven eukaryotic species (Table 3, Supplemental Table S7). Among them were three well-studied genomes of model organisms *A. thaliana*, *C. elegans,* and *D. melanogaster,* having GC-homogeneous and compact in size genomes. The larger genomes of *S. lycopersicum*, and *D. rerio* were GC-homogenous, while the other large genomes of *G. gallus* and *M. musculus* were GC-heterogeneous (see Methods). In all cases, sequences of organelles, as well as contigs without chromosome assignment, were excluded.

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 each of the seven species, we built an initial protein database (PD_0) containing proteins from the organisms present in the Kingdom or the Phylum or the Class segment of OrthoDB, where the given species belongs (Supplemental Table S8). 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, the one called ‘Species excluded’, or ii/ proteins of all the species from the same taxonomic *order*, the database called ‘Order excluded’ (see also Bruna et al. 2020). These, *the larger* and *the smaller* databases, were supposed to simulate practical scenarios when a species of interest would appear on either a *larger* or a *smaller* evolutionary distance from the species present in the reference database. Overall, in our study, the numbers of proteins in such species-specific databases ranged from 2.6 to 8.3 million (Supplemental Table S8).

Table 3. Genomes and gene annotations used as references for the assessment of gene prediction accuracy. The numbers in parentheses provided for the four large genomes characterize sets of genes and transcripts in the intersection of NCBI and Ensembl annotations (see Methods). The numbers of introns per gene were computed from averages for each gene among annotated alternative transcripts. Alternative transcripts that differ only by UTR regions are not considered.

Species	Genome length (Mb)	Reference annotation statistics					
		# coding genes		# coding transcripts		introns per gene	
C. elegans (roundworm)	100	19,969		28,544		4.8	
A. thaliana (thale cress)	119	27,445		40,827		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	25,610	(17,893)	42,929	(19,975)	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)

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* (Supplemental Table S9).

Algorithm Overview

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, the final set of parameters was used for the final round of gene predictions. The model training and gene prediction procedure implemented in GeneMark-ETP is distinctly different (Fig. 1).

GeneMarkS-TP: generation of high confidence (HC) gene predictions

Initial gene prediction in assembled transcripts

Besides complete genomes of the species we have considered, a substantial volume of extrinsic evidence is available in the libraries of RNA-seq reads and protein databases. This wealth of information allows us to develop a new component of the algorithm: gene prediction in assembled transcripts with cross-species protein support.

For a given species, the short reads from the selected RNA-seq libraries are splice-aligned to the genome by HISAT2 (Kim et al. 2019) and assembled into transcripts by StringTie2 (Kovaka et al. 2019). After filtering out the low-abundance transcripts, the remaining transcripts are merged by StringTie2 into a non-redundant set (Supplemental Table S10).

A complete transcript should contain a protein-coding region (CDS) along with 5’ and 3’ untranslated regions (UTRs). Predictions of the borders of CDS sequences in the transcripts are made by GeneMarkS-T, a self-training tool (Tang et al. 2015). Converting the intronless CDS sequences into sequences of exons in the genomic DNA is a standard task. To efficiently solve it we use data on RNA-seq reads *anchored* to genome that have been generated by StringTie2 in the process of the transcripts assembly. Given the CDS sequence, the genomic co-ordinates of the full set of exon borders are immediately inferred along with the full set of introns mapped by short reads. The statistics and the values of Sn and Sp metrics for the initially predicted genes are shown in Supplemental Table S1A of Suppl. Materials.

Correcting gene predictions

The GeneMarkS-T predictions of *complete* CDS sequences in the transcripts appear to be quite accurate. On the other hand, predictions of *5’ incomplete* CDS sequences (5’ partial genes) could be less precise. The predicted 5’ partial gene are verified and corrected if the data on homologous proteins are available.

A 5’ partial CDS is supposed to start from the first nucleotide of a transcript. However, a true complete CDS may reside inside this 5’ partial gene. To discriminate between the two possibilities, we proceed as follows. Translations of the predicted 5’ partial CDS as well as of the longest internal ORF situated in the same reading frame are used as queries in a similarity search by DIAMOND (Buchfink et al. 2015). The protein target common for both searches (E-value<10^{-3}) is aligned to both queries. The strengths of evolutionary conservation along the pairwise alignments are analyzed (condition S1 in Suppl. Methods). The 5’ partial CDS is confirmed if condition S1 is fulfilled, otherwise, the 5’ partial CDS is replaced by a shorter complete CDS (see Section S1.1 of Suppl. Methods, Supplemental Tables S11, S12).

Selecting HC genes: genes with uniform protein similarity support

The genes with uniform protein support are selected from the set of genes predicted in assembled transcripts. A *complete CDS* is said to have *uniform protein support* if a pairwise alignment of the predicted protein to some known protein satisfies condition S2 (see Suppl. Methods). A complete CDS having uniform protein support, when mapped to genome defines a *complete high-confidence (HC) gene*.

A predicted *5’ partial CDS* is said to have a uniform protein support if condition S2 is fulfilled for a pairwise alignment of the C-terminal of the 5’ partial protein translation with some database protein (Supplemental Figs. S3, S4). Such a 5’ partial CDS with uniform protein support defines a *partial HC gene* in genomic DNA.

One more step of analysis is made with the predicted complete CDS sequences. They are checked if an extension to the “longest ORF” is a possibility. If such a longer ORF exists and its protein product has a uniform protein support, satisfies condition S2 (Supplemental Fig. S5), both CDS variants are designated as the candidates for alternative *complete HC gene isoforms* (see below).

Selecting HC genes: genes without uniform protein support.

If the condition S2 is not satisfied, a predicted *complete* CDS is checked if: (i) the length of the CDS is longer than $299\ nt$, (ii) the 5' UTR contains in-frame stop codon triplet, and (iii) the exons mapped to genomic DNA do not create a conflict with the ProtHint hints (see Section 2 of Suppl. Methods). If conditions (i)-(iii) are satisfied, then the CDS mapping to genome defines an HC gene.

Notably, if a CDS is lacking a stop codon it cannot give raise to an HC gene.

Selecting of HC genes: alternative isoforms

The initially determined set of complete HC genes may contain predicted alternative isoforms. In this case, an additional round of selection is made (see Section S3 of Suppl. Methods).

The whole set of the HC genes makes the output of GeneMarkS-TP. The statistics and the values of Sn and Sp metrics for the sets of the HC genes are shown in Supplemental Table S1B of Suppl. Materials.

The GHMM model training

Single step GHMM model training

The predicted HC genes are used for training of the GHMM parameters. In the presence of HC genes with alternative isoforms only the longest CDS sequence is retained.

The set of thus defined HC genes may have a significant GC content spread. The GC content distribution is built at this stage. If more than 70% of the selected HC genes could be contained in a 9% wide GC content interval, the genome is characterized as GC homogeneous, otherwise as GC heterogeneous (Supplemental Fig. S6).

In the *GC homogeneous* case, the sequences of the HC genes extended by 1,000nt margins (making a set of the *HC loci*) are used for estimation of the GHMM parameters. In the *GC heterogeneous* case, the *HC loci* sequences are split into three GC bins: low GC, mid GC, and high GC. The borders of the mid GC bin (with the default 9% width; selection of the 9% value is illustrated in Supplemental Fig. S6) are defined by the interval position at which the interval would contain the largest number of the HC loci. Setting up the mid GC interval immediately determines the low and high GC intervals (bins). Then the three sets of the HC loci corresponding to the selected intervals are used to train the three GC-specific GHMM models.

Note that the original GeneMarkS-T was developed under assumption that a transcriptome could be GC inhomogeneous. GeneMarkS-T predicts genes in transcripts by a set of the GC-specific 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). At the initialization of iterations for *GC homogeneous* genomes, the GHMM model parameters are derived from the sequences of the HC loci. The following round of gene prediction is then made only in the genomic sequences situated between HC genes, *the non-HC segments*. The subsequent steps are described below in the section 'Gene prediction in non-HC segments'.

The extended GHMM training for *GC heterogeneous* genomes works as follows. The parameters of the initial GC specific GHMM models are trained on the HC loci sequences of the corresponding bins. Subsequently, a GC specific model is used for gene prediction in the non-HC segments of the same GC bin. From this point on, the extended training for the non-HC segments in each GC bin is done separately and similarly to the GC homogeneous case. The GC-specific models are used in the rounds of iterative parameters estimation and gene prediction that occur until convergence (Supplemental Fig. S1F). If the number of the HC genes in each bin is large (more than 4,000), then, based on our experience, a single iteration is sufficient for convergence.

In general, GeneMark-ETP could be run in the 'GC-heterogeneous' mode on any genome. However, for a 'true' GC-homogeneous genome, this choice will increase runtime and sometimes even decrease gene prediction accuracy due to splitting the overall training set into smaller subsets. Therefore, the degree of GC-heterogeneity is assessed prior to GHMM parameter estimation.

Gene prediction in non-HC segments

At the first step, GeneMark.hmm is using the trained GHMM models to create initial gene predictions in the non-HC segments (Supplemental Fig. S1F, S1G). The gene predictions are used in ProtHint to generate protein-based hints as in GeneMark-EP+ (Bruna et al. 2020). An additional set of hints comes from RNA-seq reads mapped to genome by HISAT2 (Kim et al. 2019). All over, there are the following categories of hints: (i) RNA-seq and ProtHint-derived hints that agree with each other; (ii) high score ProtHint hints to intron borders; (iii) RNA-seq-based hints to intron borders that may or may not coincide with the intron borders predicted *ab initio*; (iv) *partial HC genes* that could be extended into the non-HC segments. Note that partial HC genes can appear only at the border with non-HC segments and technically we consider partial HC genes to belong to non-HC segments.

The partial hints of the first three categories point to separate elements of a multi-exon gene. Hints of the fourth category represent 'chains' of introns that should belong to the same gene. The requirement of corroboration of *ab initio* predictions with the RNA-seq based hints of category (iii) allows to filter out the 'false positive' intron hints mapped from expressed non-coding RNA. The whole set of hints is now ready for enforcement in a run of GeneMark.hmm.

The hints are *enforced* in the non-HC segments in the cycles of gene prediction. 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. Frequently, GeneMark-ETP converged in the second iteration. The set of genes predicted in the non-HC segments along with the set of the HC genes constitute the *final set of genes* predicted by GeneMark-ETP.

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 be caused by the presence of 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. We generate species-specific repeat libraries *de novo* using RepeatModeler2 (Flynn et al. 2020). Repetitive sequences could then be identified in genomic sequence by RepeatMasker (www.repeatmasker.org). Some of the predicted repeats may overlap with protein-coding genes (Bayer et al. 2018). To conduct gene finding in genomic sequence with 'soft masked' repeats, the authors of AUGUSTUS have introduced a constant bonus function used in the Viterbi algorithm (Stanke et al. 2008). The function increased the likelihood of prediction of non-coding regions inside the sequences designated as repetitive. Here, we have introduced a penalty function to decrease the likelihood of prediction of protein-coding region inside the repeat region (1). A single parameter of this function, q, is defined (trained) by GeneMark-ETP for each genome. Technically, we introduce a state for an overlap of a repeat and CDS and compute the probability of a sequence (of length n) to appear in such an overlap:

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

The species-specific parameter q is estimated after compiling the set of HC genes and the first round of the GHMM model training (see Suppl. Methods, Supplemental Fig. S7, Supplemental Table S13).

The accuracy assessment

Selection of 'gold standard' gene sets and computation of accuracy measures

To assess the gene prediction accuracy, we had to define the 'gold standards' for all the seven genomes. Since annotations of well-studied genomes of *A. thaliana*, *C. elegans*, and *D. melanogaster* have been updated multiple times, we used these complete annotations in computation of the accuracy measures. Arguably, the annotations of the four larger genomes are less perfect. Therefore, for these genomes, *S. lycopersicum*, *D. rerio*, *G. gallus*, and *M. musculus,* we computed the Sensitivity values using a set of genes with identical annotations in the NCBI and the Ensembl records (Table 3). On the other hand, the Specificity was computed by using the union of gene annotations made by NCBI and Ensembl. In all the tests, regions of annotated pseudogenes were excluded from consideration.

As a single parameter characterizing gene prediction accuracy, we used the harmonic mean of Sensitivity and Specificity, F1 = 2(Sn)(Sp)/(Sn+Sp) or F1 = TP/(TP+(FN+FP)/2); for convenience the F1 score is multiplied by 100. Strictly speaking the later formula for F1 is not equivalent to the former one if Specificity and Sensitivity are computed using different sets of positives examples (annotated genes), which is the case for the four larger genomes. Therefore, to avoid ambiguities we use the former F1 formula in all the computations.

The Sn and Sp values were computed at the exon and the gene levels. Predicted exon had to match exactly an annotated exon to be counted as a true positive. A predicted gene was counted as true positive if at least one of its predicted alternative isoforms matched an alternative isoform of an annotated gene.

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

Earlier developed gene finders, GeneMark-ET or GeneMark-EP+, were designed to use a single source of extrinsic support, either RNA-seq reads or protein database. They were run to get reference points on what accuracy could be achieved with the single source of evidence. For a fair comparison with GeneMark-ETP, we considered a virtual combination of the sets of genes predicted separately by GeneMark-ET and GeneMark-EP+. The largest Sensitivity of such a combination could be achieved when we consider a *union* of the two sets of predicted genes while the largest Specificity could be seen when the *intersection* of the two sets of predicted genes is considered. Yet, the best overall accuracy (Sn+Sp)/2 could be achieved by either removal of false positives from the *union* of the single tool made gene predictions, or addition of true positives to the *intersection* of the two sets of predictions (see Results). These changes cannot be made when a gene finder is running on a novel genome since information on true and false positives is not immediately available. Nevertheless, such modifications could be made for gene predictions in genomes with known annotations.

Running BRAKER1, BRAKER2, TSEBRA, and MAKER2

Earlier developed automatic gene finding pipeline BRAKER1 combines AUGUSTUS and GeneMark-ET (Hoff et al. 2016). The pipeline BRAKER2 combines AUGUSTUS and GeneMark-EP+ (Bruna et al. 2021). 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 selects a subset of all the gene predictions made separately by BRAKER1 and BRAKER2, and, thus, predicts genes supported by both RNA-seq and proteins. 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 in the MAKER2 publication (Holt and Yandell 2011). Hence, we wanted to present the accuracy figures that would, arguably, correspond to the upper limit achieved by the optimal training option for MAKER2. The models for AUGUSTUS and SNAP were either the models provided by the code developers (available with the respective software distribution) or the models generated by supervised training on the genes annotated by NCBI (RefSeq). As an exception, the SNAP training was done on the Ensembl annotated *D. rerio* genome. Both MAKER2 and GeneMark-ETP use the GeneMark.hmm gene finder. MAKER2 uses models for GeneMark.hmm self-trained by GeneMark-ES using no extrinsic evidence. We assume that more precise models are obtained if extrinsic evidence is added into the training. Therefore, we trained the GeneMark.hmm models on a set of high confidence genes (determined by GeneMark-ETP). To get the accuracy figures, the gene predictions made for the genomes of *D. melanogaster, D. rerio* and *M. musculus* by MAKER2 were processed in the same way as the ones made by GeneMark-ETP or other gene finders. The repeat coordinates, RNA-seq and protein data sets were the same for MAKER2 and GeneMark-ETP (see Suppl. Materials). Notably, we compiled special reduced size protein databases for running the experiments with MAKER2, since the runtime of MAKER2 sharply increases with the increase in volume of protein data. The minimal size of the 'Order Excluded' database used in the experiments described for the seven genomes was 2.5 mln proteins. The maximum size database that was used in the experiments for comparison of GeneMark-ETP with MAKER2 was about 300,000 proteins (see Section S5 of the Suppl. Materials). 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, by design, the GC specific parameters were used in AUGUSTUS but not in SNAP or GeneMark.hmm.

Runtime of GeneMark-ETP

The runtime of GeneMark-ETP depends linearly on the genome size and is comparable to the one observed for GeneMark-EP+. The GeneMark-ETP runtime dependence on the size of protein databases is also linear with much smaller coefficient of proportionality. The size of RNA-seq libraries is critical for the HISAT2 runtime but not for the rest of the GeneMark-ETP operations with RNA data processing. To give examples of absolute runtime values, on a machine with 64 CPU cores, GeneMark-ETP runtimes on genomes of *D. melanogaster*, *D. rerio*, and *M. musculus* were 1.0, 4.5, and 6.5 hours, respectively. The 'Order excluded' protein databases and RNA reads sets described in Table 1 were used in these experiments reflected in Figs. 3 and 4.

Supplemental Materials

URL to be determined by Genome Research (A file with Supplemental Materials 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. GeneMark-ETP was included in the recently developed pipeline BRAKER3 (Gabriel et al. 2023).

Authors Contribution

T.B, A.L. and M.B designed the project; T.B and A.L. designed and wrote the code, prepared all the initial data, and conducted computational experiments; T.B, A.L. and M.B wrote and edited the manuscript.

Competing Interests Statement

GeneMark-ETP and its part GeneMark.hmm, are distributed under the Creative Commons license. Licensing GeneMark.hmm by commercial companies may create a conflict of interest for AL and MB. TB declares having no competing interests.

Funding

National Institutes of Health [GM128145, in part]. Funding for the open access charge: Georgia Tech.

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.

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.

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**.

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.

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.