

RESEARCH

Open Access



Galba: genome annotation with miniprot and AUGUSTUS

Tomáš Brůna¹, Heng Li^{2,3}, Joseph Guhlin⁴, Daniel Honsel⁵, Steffen Herbold⁶, Mario Stanke⁷, Natalia Nenasheva⁷, Matthis Ebel⁷, Lars Gabriel⁷ and Katharina J. Hoff^{7*}

*Correspondence:

katharina.hoff@uni-greifswald.de

¹ U.S. Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

² Department of Data Sciences, Dana-Farber Cancer Institute, Boston 02215, MA, USA

³ Department of Biomedical Informatics, Harvard Medical School, Boston 02215, MA, USA

⁴ Genomics Aotearoa and Laboratory for Evolution and Development, Department of Biochemistry, University of Otago, Dunedin 9016, New Zealand

⁵ Institute of Computer Science, University of Göttingen, 37077 Göttingen, Germany

⁶ Faculty for Computer Science and Mathematics, University of Passau, 94032 Passau, Germany

⁷ Institute of Mathematics and Computer Science, and Center for Functional Genomics of Microbes, University of Greifswald, 17489 Greifswald, Germany

Abstract

Background: The Earth Biogenome Project has rapidly increased the number of available eukaryotic genomes, but most released genomes continue to lack annotation of protein-coding genes. In addition, no transcriptome data is available for some genomes.

Results: Various gene annotation tools have been developed but each has its limitations. Here, we introduce GALBA, a fully automated pipeline that utilizes miniprot, a rapid protein-to-genome aligner, in combination with AUGUSTUS to predict genes with high accuracy. Accuracy results indicate that GALBA is particularly strong in the annotation of large vertebrate genomes. We also present use cases in insects, vertebrates, and a land plant. GALBA is fully open source and available as a docker image for easy execution with Singularity in high-performance computing environments.

Conclusions: Our pipeline addresses the critical need for accurate gene annotation in newly sequenced genomes, and we believe that GALBA will greatly facilitate genome annotation for diverse organisms.

Keywords: Gene prediction, Protein coding gene, Miniprot, AUGUSTUS

Background

The Earth Biogenome Project (EBP) aims at sequencing and annotating all eukaryotic life on Earth within ten years [1]. It has brought about an explosion of genomic data: for instance, the Wellcome Sanger Institute alone currently aims at sequencing and assembling 60 genomes per day. This provides an unprecedented opportunity to study the diversity of life on Earth. Generating genome assemblies is now easier than ever thanks to cheaper sequencing, e.g. with Nanopore technology (for review of technology see [2]). However, while the number of available genomes continues to rapidly increase, the annotation of protein-coding genes remains a bottleneck in the analysis of these data [3]. This is, for instance, obvious from screening through Data Note Genome Announcements at Wellcome Open Research [4], or from counting genomes and their annotations



at NCBI Genomes, where on April 3rd 2023, only 23% of 28,754 species are listed with the annotation of at least one annotated Coding Sequence (CDS) [5].

Genome annotation remains a bottleneck because it is currently not a straightforward approach. Large centers, such as Ensembl at EBI or the NCBI, are facing computational and human resources bottlenecks to apply their in-house annotation pipelines to all incoming genomes, while small and less experienced teams simply might not know where to start because not all annotation pipelines work equally well in all genomes.

BRAKER3 [6], a pipeline that combines the gene prediction tools GeneMark-ETP [7] and AUGUSTUS [8, 9] for fully automated structural genome annotation with short read transcriptome data (RNA-Seq) and a large database of proteins (such as an OrthoDB clade partition [10]) was recently demonstrated to have high accuracy for the particular input scenario of genome file, RNA-Seq short read data, and a protein database. However, despite the EBP encouraging the sequencing of transcriptomes alongside genomes [3], it can be difficult to obtain RNA-Seq data for some organisms for logistical or financial reasons, or an initial genome annotation can be desired before a transcriptome is sequenced. Also, some genes may not be expressed in tissues being sequenced and thus do not have RNA-Seq support. Conservation species often need to be annotated for gene-level genetic load estimation, frequently lacking RNA-Seq data. In invasomics, annotation of protein coding genes is of particular importance for exploratory gene drive studies, and generating probes for expression and localization studies. For both, high-quality rapid annotation is essential to move towards downstream analyses.

In the lack of transcriptome evidence, it is a common procedure to annotate novel genomes by leveraging spliced alignment information of proteins from related species to the target genome. Since the resulting alignments usually only cover a fraction of all existing genes in a genome and do not cover untranslated regions (UTRs), protein alignments are commonly combined with gene prediction tools that employ statistical models (e.g. AUGUSTUS, SNAP [11], and variants of GeneMark [12–14]) to identify the other fraction of genes as good as possible. MAKER [15–17] was an early pipeline that automated this for the gene prediction step (though it lacks automated training of gene predictors). FunAnnotate [18] was originally designed to train gene finders using RNA-Seq data but also provides a workaround for protein input on fungi. It has since also been applied to other eukaryotic genomes (a random example: [19]). In contrast to these algorithms, which usually use evidence from one or a low number of donor proteomes, BRAKER2 [20] is a pipeline that leverages a large database of proteins with GeneMark-EP [13] and AUGUSTUS to predict protein-coding genes. BRAKER2 fully automates the training of GeneMark-EP and AUGUSTUS in novel genomes. BRAKER2 was previously demonstrated to have higher accuracy than MAKER [20].

In order to allow for the alignment of a large number of protein sequences in a reasonable time, GeneMark-EP first runs self-training GeneMark-ES [12, 14] to generate genomic seeds. Subsequently, DIAMOND [21] quickly returns hits of proteins against those initial candidate protein-coding sequences found in the genome, and Spaln [22, 23] is applied to run accurate spliced-alignment of the best matching protein sequences against the genomic seeds. BRAKER2 executes one iteration of this process to expand the genomic seed space by AUGUSTUS predictions. This complex sub-pipeline is called ProtHint and was introduced to make the alignment of a large database of

proteins against the genome for evidence generation computationally feasible on desktop machines. BRAKER2 generally achieves high accuracy in small and medium-sized genomes. In large genomes (e.g., the genome of a chicken or mouse), self-training GeneMark-ES performs poorly during seed generation, leading to lower prediction accuracy of BRAKER2.

With the appearance of miniprot [24], a very fast and accurate tool for spliced-aligning proteins to genome sequences, the question arose whether it is necessary to run a complicated pipeline such as ProtHint in order to generate evidence and training genes to annotate novel genomes with protein evidence with high accuracy. Moreover, miniprot has no problems processing average vertebrate-sized genomes and therefore promises to overcome the main shortcoming of BRAKER2 in terms of accuracy in large genomes.

With regard to the EBP, we expect the appearance of a large number of genomes for which suitable reference proteomes for running BRAKER2 will not be fully available. BRAKER2 requires a large protein database input; it usually fails to run with reference proteins of only one species because its components, ProtHint and GeneMark-EP, rely heavily on evidence derived from multiple alignments (requiring ≥ 4 supporting alignments to classify a hint as high-confidence). This hinders BRAKER2's ability to annotate genomes of poorly sequenced clades where only one reference relative is often available.

In order to address these open questions and challenges, we designed GALBA. GALBA is a fully automated pipeline that takes protein sequences of one or many species and a genome sequence as input, aligns the proteins to the genome with miniprot, trains AUGUSTUS, and then predicts genes with AUGUSTUS using the protein evidence. In this manuscript, we describe the GALBA pipeline and evaluate its accuracy in 14 genomes with existing reference annotation. Further, we present three use cases of *de novo* genome annotation in insects, vertebrates, and one land plant. We also evaluate the effect of merging GALBA and BRAKER2 gene sets with TSEBRA [25], the transcript selector for BRAKER.

Our pipeline is fully open source, containerized, and addresses the critical need for accurate gene annotation in large newly sequenced genomes. We believe that GALBA will greatly facilitate genome annotation for diverse organisms and is thus a valuable resource for the scientific community.

Results

We first briefly describe the GALBA pipeline and the effect of several features on gene prediction accuracy. Subsequently, we present accuracy results of the final software in 14 species. Further, we present three different use cases for GALBA.

GALBA pipeline

GALBA is a pipeline that connects three main components to predict protein coding genes: Firstly, we employ miniprot [24] to splice-align input protein sequences to the genome, and then use miniprothint [26] to score the resulting alignments and categorize the evidence into low- and high-confidence classes. We utilize the high-confidence alignment-derived genes with the highest alignment score per locus to train the gene prediction tool AUGUSTUS [8, 9]. Subsequently, we run AUGUSTUS with the Python package Pygustus to predict genes using the protein evidence in multithreading mode.

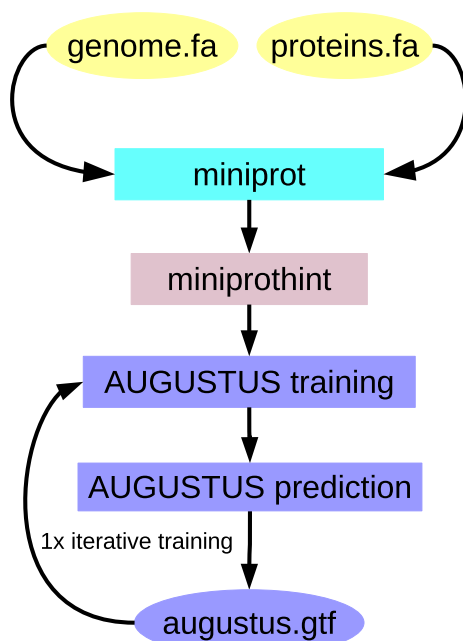


Fig. 1 The GALBA pipeline. Miniprot performs rapid spliced alignment of proteins against the genome. Subsequently, miniprothint (2) scores and classifies these alignments. Training genes for AUGUSTUS are generated from the best high quality miniprot alignment per locus (1). After training, AUGUSTUS predicts genes using the alignment evidence generated by miniprothint. AUGUSTUS parameters are refined by one iteration of training (3). The numbering of steps in the figure caption corresponds to the order in which steps were introduced into GALBA during development, see Additional file 1: Results section S4.1

After the first round of prediction, we select genes with 100% evidence support according to AUGUSTUS for a second round of training, while all other predicted genes are used to delineate flanking intergenic regions for the training of parameters for non-coding sequences. Then, we obtain the final set of predicted genes by AUGUSTUS (see Fig. 1). The idea of GALBA is that training AUGUSTUS on the basis of miniprot alignments will enable AUGUSTUS (with hints) to obtain a gene set that is more accurate and more complete than the miniprot alignments on their own. We show that GALBA works as expected in terms of accuracy with respect to reference annotations on the example of 14 species in Additional file 1: Table S10. This is also reflected by the drastically increasing complete BUSCOs when moving from training gene set to AUGUSTUS gene set within GALBA (see Additional file 1: Table S12).

GALBA was implemented in Perl, building on the existing codebase of BRAKER [27].

Effect of mutation rate from reference to target

GALBA is designed to be used with reference proteomes of (possibly several) closely related species. It is predictable that spliced protein to genome alignment with miniprot works better the lower the mutation rate from donor to target is. We provide results of GALBA runs with single-species reference protein inputs in *D. melanogaster* next to a phylogenetic tree that indicates mutation rates to provide users a reference for how similar a donor species should be to achieve good results with GALBA (see Fig. 2).

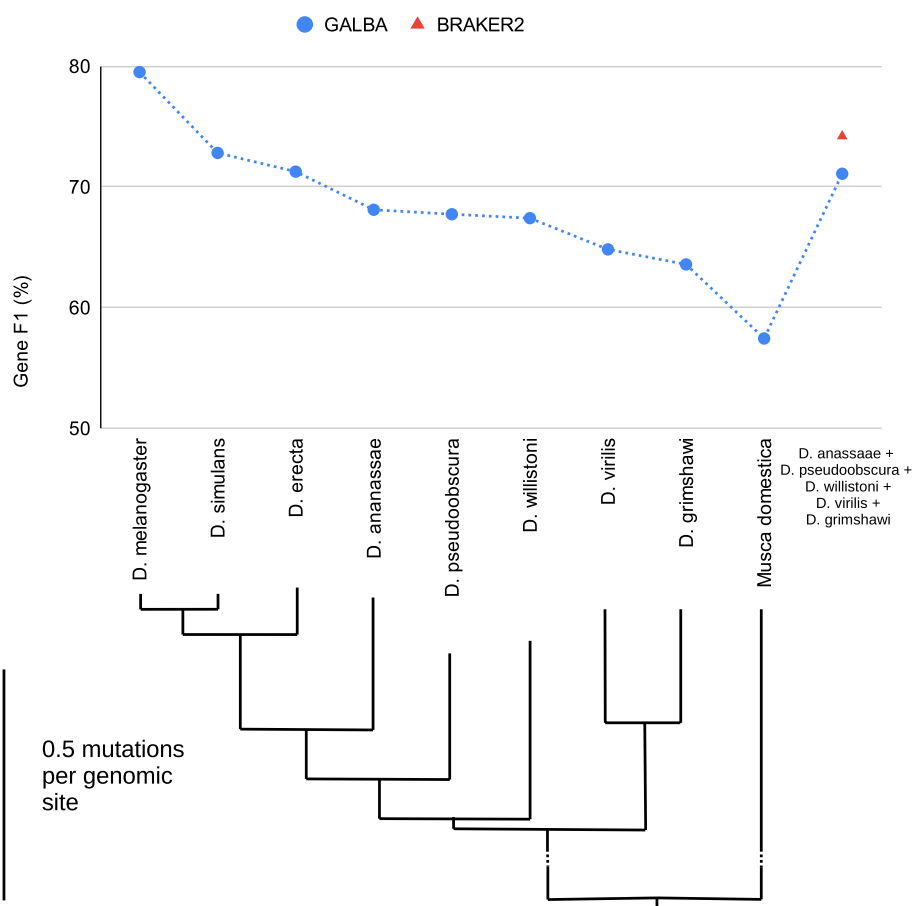


Fig. 2 Gene prediction of GALBA provided with either a proteome of a single reference species (corresponding to phylogenetic tree from [57]), or executed with a combination of the species listed on the right. BRAKER2 can only be executed with a certain level of redundancy in the protein reference set, and results are therefore only provided for the combined protein input set

When executed using all annotated proteins of the target species itself, GALBA achieves a gene F1 of 79.5% (F1-scores are in this manuscript defined as $\frac{2 \cdot \text{Sensitivity} \cdot \text{Specificity}}{\text{Sensitivity} + \text{Specificity}}$). When moving to *D. ananassae*, the accuracy drops by ~7.5% points. Gene F1 does not drop below 63.6% when moving away to *D. grimshawi*, and even with *Musca domestica* input, GALBA maintains an accuracy of 57%. Interestingly, accuracy is restored to 71% when using a combined input of five protein donors. This last experiment can in fact also be performed with BRAKER2, which scores 3% points higher accuracy compared to GALBA.

Accuracy in genomes with reference annotation

We provide accuracy results measured in genomes of 14 species by comparison to existing annotations (see Figs. 3 and 4 for sensitivity and specificity on gene level, and Table 1 for F1-scores for gene, transcript, and exon levels). The annotations of the small model organisms *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster* have undergone extensive curation [28], and thus we believe that benchmarking on these data sets gives a realistic estimate of the true accuracy of gene prediction

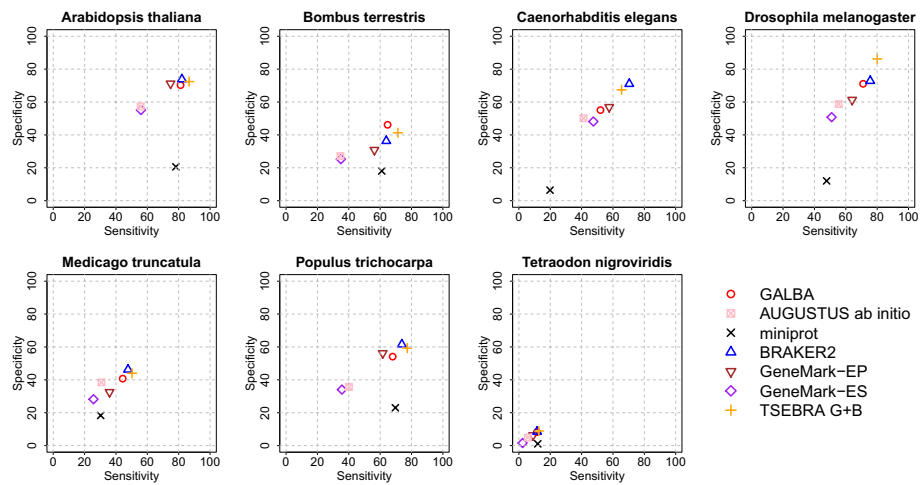


Fig. 3 Sensitivity and Specificity on gene level in 7 genomes smaller than 500 Mb. We show accuracy of miniprot raw alignments, AUGUSTUS ab initio trained on filtered miniprot alignments, GALBA (AUGUSTUS with hints by miniprot), BRAKER2, GeneMark-EP, GeneMark-ES, and a combination of GALBA and TSEBRA (labelled as TSEBRA G+B)

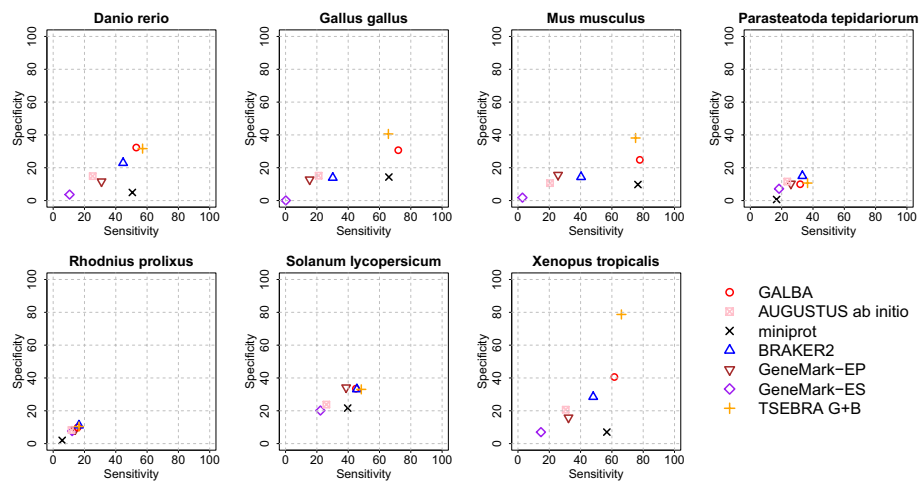


Fig. 4 Sensitivity and Specificity on gene level in 7 genomes larger than 500 Mb. We show accuracy of miniprot raw alignments, AUGUSTUS ab initio trained on filtered miniprot alignments, GALBA (AUGUSTUS with hints by miniprot), BRAKER2, GeneMark-EP, GeneMark-ES, and a combination of GALBA and TSEBRA (labelled as TSEBRA G+B)

pipelines. Annotations of the other species are much less reliable. Therefore, we report gene prediction sensitivity measured on two more reliable subsets created by selecting transcripts that (1) are complete and have all introns supported by RNA-Seq mapping (Additional file 1: Table S3); (2) have identical gene structures in two distinct reference annotations (Additional file 1: Table S4).

We decided to show GALBA and BRAKER2 results with identical multi-species protein input side-by-side. Since users of BRAKER2 may be familiar with the Transcript Selector for BRAKER (TSEBRA) for combining several gene sets, we also provide TSEBRA results for which the GALBA and BRAKER2 outputs including their evidence were combined, enforcing the predictions by GALBA to avoid a drop of all transcripts without

Table 1 F1-scores of gene predictions for the genomes of 14 different species

	<i>Arabidopsis thaliana</i>		<i>Bombus terrestris</i>		<i>Caenorhabditis elegans</i>		<i>Danio rerio</i>		<i>Drosophila melanogaster</i>			
	Gene	Transcript	Exon	Gene	Transcript	Exon	Gene	Transcript	Exon	Gene	Transcript	Exon
GALBA	75.32	60.09	84.82	53.89	45.19	82.82	53.51	42.28	80.99	71.07	55.05	82.74
BRAKER2	78.20	62.09	85.14	46.32	38.99	79.15	70.71	56.71	88.01	30.32	23.87	73.02
TSEBRA G+B	78.92	61.16	84.98	52.30	43.25	81.62	66.44	49.09	83.81	40.73	29.17	76.77
GALBA ⁵	71.15	57.16	84.16	49.57	41.65	81.80	47.16	38.31	78.40	32.10	25.43	75.58
	<i>Medicago truncatula</i>		<i>Parasteatoda tepidariorum</i>		<i>Populus trichocarpa</i>		<i>Rhodnius prolixus</i>		<i>Tetraodon nigroviridis</i>			
	Gene	Transcript	Exon	Gene	Transcript	Exon	Gene	Transcript	Exon	Gene	Transcript	Exon
GALBA	42.44	40.90	73.57	15.17	13.17	56.26	60.26	46.39	77.75	11.75	11.16	53.64
BRAKER2	46.94	46.94	74.95	20.67	18.40	63.50	67.14	56.02	82.27	13.25	12.77	54.62
TSEBRA G+B	46.93	42.35	74.01	16.51	13.63	55.51	67.09	48.65	78.18	12.75	11.36	53.03
GALBA ⁵	43.32	42.45	74.81	15.19	13.70	59.07	53.44	46.28	78.86	11.29	11.05	53.53
	<i>Gallus gallus</i>		<i>Mus musculus</i>		<i>Solanum lycopersicum</i>		<i>Xenopus tropicalis</i>		Average			
	Gene	Transcript	Exon	Gene	Transcript	Exon	Gene	Transcript	Exon	Gene	Transcript	Exon
GALBA	43.03	35.07	69.29	37.62	31.45	62.75	38.37	36.46	71.55	48.93	39.23	83.77
BRAKER2	23.92	16.29	46.50	27.80	26.96	57.39	38.36	35.91	69.33	35.76	27.84	77.91
TSEBRA G+B	50.17	35.34	83.75	50.58	31.88	79.05	39.26	35.22	70.50	49.15	37.59	82.80
GALBA ⁵	40.59	34.76	70.10	30.05	27.23	61.72	38.54	37.24	72.71	39.83	32.87	81.34

We show a direct comparison of GALBA, BRAKER2, and a combination of GALBA with BRAKER2 by TSEBRA (TSEBRA G+B) with the same input data. In addition, we provide GALBA⁵ results with one reference gene set only (labeled with ⁵ in Additional file 1: Table S1)

support by evidence. In large vertebrate genomes, GALBA shows a large improvement in accuracy compared to BRAKER2 (between 10 and 30% points in the gene F1-score). In small and medium-sized genomes, BRAKER2 is usually superior to GALBA. In *A. thaliana*, *D. melanogaster*, *M. truncatula*, *P. tepidariorum*, *R. prolixus*, and *T. nigroviridis*, BRAKER2 is $\geq 5\%$ more accurate on the gene level than GALBA. GALBA shows particularly poor accuracy in *C. elegans* (17% points less than BRAKER2) and *P. trichocarpa* (7% points less than BRAKER2). In *B. terrestris* and *S. lycopersicum*, GALBA performs marginally better than BRAKER2.

This general impression also holds when looking at the subset of multi-exon genes that are supported by RNA-Seq from VARUS sampling (see Additional file 1: Table S3), and when inspecting Sensitivity in the subset of genes that are supported by more than one annotation provider (see Additional file 1: Table S4). In large vertebrate genomes, GALBA here achieves astonishing exon F1-scores of $> 90\%$, and gene F1-scores $> 70\%$, outperforming BRAKER2 by up to 42% points on the gene level.

Since BRAKER2 was originally designed to run with a large database of proteins instead of a hand-picked proteome of few closely related species, we show BRAKER2 results with OrthoDB v11 partitions for different taxonomic phyla (Arthropoda, Metazoa, Vertebrates, Viridiplantae), excluding proteins of the target species, and adding the hand-picked proteomes of close relatives by concatenation. This input does not change accuracy results much (see Additional file 1: Table S7). To the best of our knowledge, BRAKER2 is the most suitable pipeline for annotation scenarios where closer relatives have not been sequenced and annotated, yet. Therefore, we also provide BRAKER2 results with OrthoDB partitions, excluding proteins of species that are in the same taxonomic order as the target species.¹ In *M. truncatula*, *P. tepidariorum*, *P. trichocarpa*, and *T. nigroviridis*, BRAKER2 is even more accurate than GALBA using the remotely related protein set (see Additional file 1: Table S7).

It is an interesting question whether combining the GALBA and BRAKER2 gene sets (with the same protein input) with TSEBRA provides increased or restored accuracy. In general, TSEBRA tends to increase the ratio of mono-exonic to multi-exonic genes (see Table 2 and Additional file 1: Figure S5). In species where both GALBA and BRAKER2 shows initial comparable accuracy, TSEBRA application usually increases the accuracy by a few percentage points. However, if the GALBA gene prediction accuracy is particularly poor (e.g., in the case of *C. elegans*), then TSEBRA does not fully restore accuracy to the better gene finder (here BRAKER2). For large vertebrate genomes, the TSEBRA approach consistently yields very good results (despite increasing the amount of single-exon genes), although the effect varies between about 1% point on gene level in *D. rerio* and 13% points in *M. musculus*.

Since GALBA may also be executed with a single reference proteome, we provide results of such experiments, using the closest relative from our selection of protein donor species. Using a single protein donor instead of a set of several with GALBA usually leads to a decrease in accuracy (on average 4% points gene F1). This effect can be less

¹ For this, we used the orthodb-clades pipeline [29] to generate the protein sets.

Table 2 Ratios of mono-exonic to multi-exonic genes per species

Species	Annotation	GALBA	BRAKER2	TSEBRA G+B
<i>A. thaliana</i>	0.30	0.34	0.31	0.37
<i>B. terrestris</i>	0.06	0.23	0.51	0.56
<i>C. elegans</i>	0.04	0.07	0.05	0.08
<i>D. melanogaster</i>	0.25	0.28	0.27	0.32
<i>D. rerio</i>	0.08	0.36	0.29	0.48
<i>G. gallus</i>	0.09	0.39	0.35	0.57
<i>M. musculus</i>	0.20	0.75	0.47	0.90
<i>M. truncatula</i>	0.54	0.44	0.39	0.51
<i>P. tepidariorum</i>	0.19	0.66	0.58	0.78
<i>P. trichocarpa</i>	0.32	0.34	0.30	0.35
<i>R. prolixus</i>	0.19	0.63	0.49	0.78
<i>S. lycopersicum</i>	0.32	0.42	0.35	0.52
<i>T. nigroviridis</i>	0.04	0.13	0.10	0.16
<i>X. tropicalis</i>	0.11	0.37	0.37	0.52

We show this ratio for the reference annotation, GALBA, BRAKER2, and combination of GALBA and BRAKER2 with TSEBRA

strongly observed in species where GALBA performs comparably poorly (e.g., *R. prolixus* or *P. tepidariorum*).

We also report results of FunAnnotate (see Additional file 1: Table S7) with the same protein and genome input as GALBA and BRAKER2, but these results are not directly comparable since this pipeline requires specification of a *seed species* for training AUGUSTUS, and of a BUSCO [30] lineage, and accuracy results may heavily depend on the selection of these (here used seed species and BUSCO lineages are listed in Additional file 1: Table S6). FunAnnotate was competitive with GALBA (and BRAKER2) only in the case of predicting genes in *A. thaliana*.

Use case examples

GALBA is widely applicable to eukaryotic genomes of different sizes and assembly quality. In the following, we present three use cases.

Insect genomes

We compare annotation results for four Hymenoptera species across three pipelines: GALBA, BRAKER2, and FunAnnotate. For this, we select three high-quality wasp genomes from [31], *Vespula vulgaris*, *V. germanica*, and *V. pensylvanica*, previously annotated using FunAnnotate with multiple rounds of annotation polishing, and one additional wasp generated with short-read assembly, [32] *Polistes dominula* (see Table 6). Input proteome to all three consisted of UniProt Swiss-Prot [33] release 2023_01, combined with published proteomes from RefSeq [34] release 104 of *Apis mellifera* HA v3.1 [35] and *Polistes canadensis* [36].

Compared to the other pipelines, GALBA consistently predicts the most genes. BUSCO scores are comparable with BRAKER2 and higher than FunAnnotate (see Table 3). GeneValidator [37], which scores individual proteins, serves as a larger metric for analyzing genome annotation results and scores individual protein predictions. GALBA predicts more higher-quality proteins, however the lower quartile for GALBA

Table 3 Summary across four Hymenopteran insect genomes and *de novo* annotation pipelines

Species	Method	#Genes	#Transcripts	#Good Predictions	#Bad Predictions	Score Quartiles	BUSCO C (%)	Δ BUSCO C
<i>Vespula vulgaris</i>	GALBA	14,087	16,766	5,393	11,373	0, 67, 90	95.8	-0.9
	BRAKER2	12,338	13,808	4,974	8,834	45, 67, 90	95.8	-0.9
	Funannotate	12,200	12,200	2,970	9,230	0, 45, 67	82.7	12.2
<i>Vespula pensylvanica</i>	GALBA	14,071	16,897	5,767	11,130	0, 67, 90	98.0	-1.8
	BRAKER2	12,891	14,327	5,134	9,193	45, 67, 90	97.4	-1.2
	Funannotate	12,580	12,580	3,146	9,434	0, 45, 90	85.6	10.6
<i>Vespula germanica</i>	GALBA	14,413	17,070	5,354	11,716	0, 64, 90	94.8	-1.2
	BRAKER2	12,956	14,409	4,919	9,490	45, 67, 90	94.6	-1
	Funannotate	10,267	10,267	3,177	7,090	45, 67, 90	84.7	8.9
<i>Polistes dominula</i>	GALBA	15,590	18,505	5,645	12,860	0, 64, 90	96.4	-0.7
	BRAKER2	15,322	17,075	5,145	11,930	22, 64, 90	96.2	-0.5
	Funannotate	9,637	9,637	2,061	7,576	0, 45, 67	65.6	30.1

Number of good and bad predictions, as well as score quartiles, as summarized by GeneValidator. BUSCO completeness according to the hymenopteran lineage hymenoptera_odb10. (Δ BUSCO C, defined as the difference of BUSCO C on genome level - BUSCO C in the predicted gene set)

is always 0, while for BRAKER2 the average lower quartile is 39.3. Taken together, this shows GALBA predicts a larger number of both high-quality and low-quality proteins. Both pipelines outperform FunAnnotate in every metric. However, FunAnnotate was designed for use with RNA-Seq data (on fungi), so this is likely to be expected.

Vertebrate genomes

Three years ago, the Zoonomia consortium presented a large whole-genome alignment of various vertebrates [38]. Many of the genomes in this alignment have not been annotated for protein-coding genes until today. Most of the unannotated assemblies in the alignment were produced by short-read genome sequencing and are thus fragmented and incomplete, and for many species (reflected by a low N50, a very large number of scaffolds, and BUSCO completeness far below 100%), there is no transcriptome data available in the Sequencing Read Archive [39]. We *de novo* annotated all whale and dolphin assemblies from that alignment that lack RNA-Seq evidence (see Table 6). The selected reference protein sets are listed in Additional file 1: Table S1.

We were able to apply multi-threaded GALBA to these genomes without any problems. GALBA predicted between 53k and 78k genes in these assemblies. The ratio of mono- to multi-exonic genes suggests an overprediction of single-exon genes. It should be noted that AUGUSTUS is capable of predicting incomplete genes that span sequence borders, and that the high single-exon count is not caused by genome fragmentation alone. Removing all incomplete genes from the prediction does not substantially

Table 4 Summary of protein-coding gene structures predicted in the previously unannotated whale and dolphin genomes of Zoonomia [38], and in *Coix aquatica*

Species	#Genes	#Transcripts	Mono:Mult	Max exons	#Incomplete	BUSCO C (%)	ΔBUSCO C
<i>Balaenoptera bonaerensis</i>	78,621	85,752	1.18	117	19,085	53.0	1.1
<i>Eubalaena japonica</i>	65,123	75,137	1.02	124	10,478	74.1	0.8
<i>Inia geoffrensis</i>	53,435	63,147	0.86	117	8,405	66.0	1.7
<i>Kogia breviceps</i>	72,288	81,084	1.21	160	15,792	65.9	0.2
<i>Phocoena phocoena</i>	56,156	68,654	0.93	158	6,365	85.8	0.1
<i>Platanista gangetica</i>	72,926	80,263	1.13	67	16,080	57.2	1.9
<i>Ziphius cavirostris</i>	75,609	81,048	1.41	77	29,926	38.0	1.9
<i>Coix aquatica</i>	93,399	98,979	1.07	80	102	97.8	0

Number of genes (#Genes), number of transcripts (#Transcripts), number of incompletely predicted transcripts where start- and/or stop-codon are lacking (#Incomplete), Mono:Mult ratio (considering only the first of each possible alternative splicing isoforms of genes with multiple isoforms), the maximum number of exons in a single gene, BUSCO completeness according to vertebrata_odb10, the difference to BUSCO completeness on genome level (ΔBUSCO C, defined as the difference of BUSCO C on genome level - BUSCO C in the predicted gene set)

decrease the mono:mult ratio (data not shown). BUSCO-completeness of predicted genes is comparable to the BUSCO-completeness of the corresponding genomic assemblies (see Table 4 and Additional file 1: Figures S3 and S2). OMArk [40], a tool that provides an estimate on annotation quality for a much larger set of conserved genes than BUSCO, also indicates a high level of completeness in these genomes (see Additional file 1: Table S8). However, the number of unexpected duplicate HOGs is large for these annotations. The consistency report of OMArk shows that the predicted genes are to a large extent possibly incomplete/fragmented (which is here likely caused by the genome assembly quality).

Plant genome

We chose the genome of the plant *Coix aquatica* [41] (see Table 6) to demonstrate the ability of GALBA to *de novo* annotate large chromosome-scaffolded genomes (see Table 6). This species is one of many that currently lack an annotation of protein-coding genes at NCBI Genomes (even though the publication [41] describes an annotation approach and statistics on predicted protein coding genes), and there is no RNA-Seq data of this species available at the Sequence Read Archive (even though [41] report having used RNA-Seq data for annotation). In practice, a *Coix aquatica* focused scientist would request the gene set from the authors of [41], but here, we took it as a *de novo* annotation example. Four reference proteomes used with GALBA are listed in Additional file 1: Table S1.

GALBA predicted 93k genes with a mono- to multi-exonic gene ratio of 1.07 in *Coix aquatica*. This is an overprediction compared to the number of 39,629 genes reported by [41]. However, the BUSCO sensitivity in the GALBA gene set is with ~98% very high and comparable to BUSCO completeness of the assembly (see Additional file 1: Figure S4). OMArk also attests to a high degree of HOG completeness. Compared to the whale

and dolphin gene predictions, the predictions in this plant genome show a much lower degree of fragmentation (see Additional file 1: Table S8). About half of the predicted proteins are placed as inconsistent, and most of these are identified by fragmented hits. GALBA here provided a quick and simple means to obtain a gene set.

Runtime

We report wallclock time passed when running GALBA on *D. melanogaster* using proteins of *D. ananassae*, *D. pseudoobscura*, *D. willistoni*, *D. virilis*, and *D. grimshawi* on an HPC node with Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz using 48 threads. A complete GALBA run took 3:24 h. A full BRAKER2 run on the same node took 3:03 h. The most time-consuming step of GALBA (and BRAKER2) is often the metaparameter optimization for AUGUSTUS. This step can optionally be disabled (`--skipOptimize`), leading to slightly lower prediction accuracy in most cases. Without this optimization step, a GALBA run with the same input data took 0:44 h.

As a second example, we report wallclock time of 8:52 h for *de novo* annotation of the *Coix aquatica* genome on an HPC node with Intel(R) Xeon(R) Gold 6240 CPU @ 2.60GHz using 72 threads (including metaparameter optimization). On the same data set and architecture, BRAKER2 required 11:11 h.

Discussion

Obtained accuracy results of GALBA are far from perfect when compared to reference annotations. However, GALBA provides substantially higher accuracy than BRAKER2 in the genomes of large vertebrates because GeneMark-ES within BRAKER2 performs poorly in such genomes when generating seed regions for spliced-alignment of proteins to the genome. We estimate, that to date, ~1k unannotated genomes without transcriptome data could benefit from structural annotation with GALBA (see Additional file 1: Methods S3.7).

In smaller genomes, BRAKER2 remains superior because with the GeneMark-ES seeding process, it is able to produce protein to genome alignments with a higher specificity than miniprot (compare Fig. 5 and Additional file 1: Table S11).

Further, we demonstrate that GALBA can process highly fragmented as well as large genomes in multi-threading mode, mainly attributed to the usage of Pygustus. We expect the Pygustus approach to be adopted in BRAKER to improve stability.

Implementing pipelines that leverage protein-to-genome alignment for training and running gene finders is not straightforward. In this work, we once more demonstrate that alignment scoring is crucial for achieving high gene prediction accuracy when protein evidence is used as the sole extrinsic evidence source.

While neither GALBA nor BRAKER2 can compete with pipelines that integrate RNA-Seq as an additional source of evidence, such as BRAKER3, GALBA is a valuable addition to closing the annotation gap for already deposited genomes and for future genomes generated within the EBP for which RNA-Seq data is not available.

Combining multiple gene sets commonly yields higher accuracy than using a single gene set of a single gene predictor. However, the authors caution users that combining gene sets from different sources may not always lead to improved accuracy, and users of genome annotation pipelines should proceed with caution. Recommended estimates for

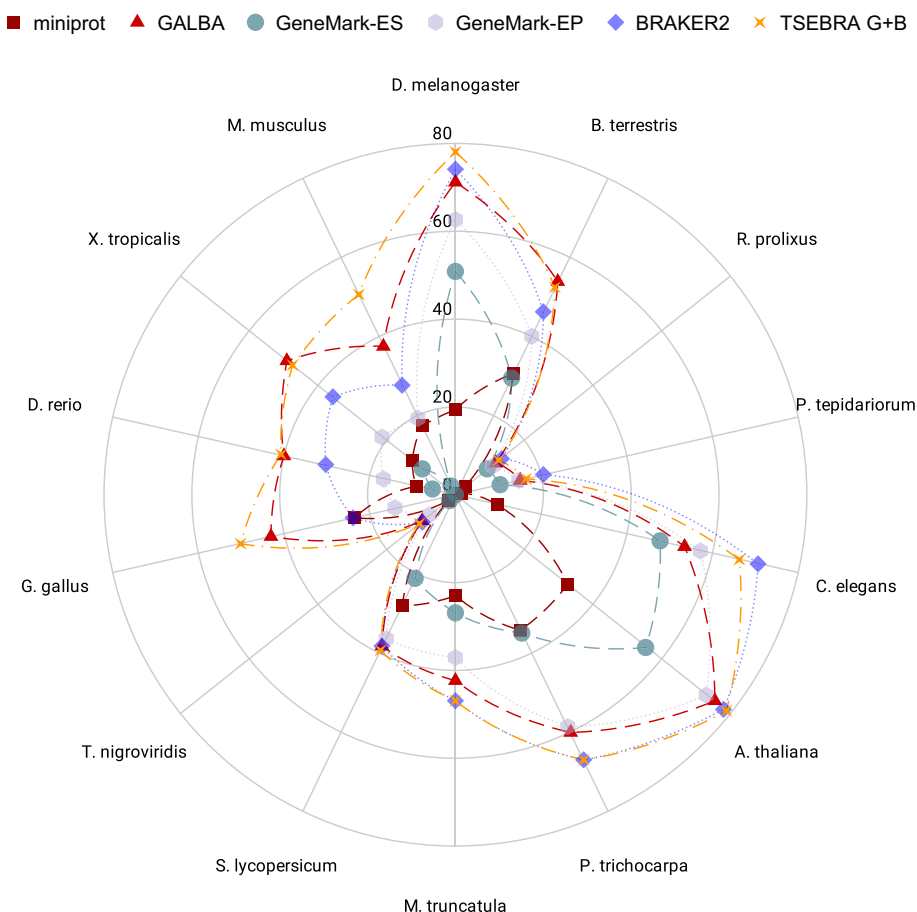


Fig. 5 Network plot of gene F1 accuracy for (clockwise starting from the top, increasing genome sizes) insects, metazoa, plants, and vertebrates. We show accuracy of GALBA and its intermediate product miniprot, and of BRAKER2 and its intermediate GeneMark-ES and GeneMark-EP gene sets. Accuracy of the combiner TSEBRA combining the final gene sets of both GALBA and BRAKER2 is also shown as TSEBRA G+B

gene set quality are BUSCO Sensitivity, the number of predicted genes, and the mono-to-multi-exon gene ratio.

Both GALBA and BRAKER2 tend to heavily overpredict single-exon genes, most likely a result of incorrectly splitting genes. For plants, a desired mono- to multi-exonic gene ratio of 0.2 was recently postulated by [42]. This particular ratio certainly does not hold for non-plant species, and also the reference annotations of plants used in this manuscript often deviated from that recommendation. Nevertheless, GALBA, BRAKER2, and TSEBRA output may benefit from downstream mono-exonic gene filtering. The EBP would benefit from future developments to address the split gene problem in pipelines for fully automated annotation of protein-coding genes.

GeMoMa is a different approach towards an accurate mapping of annotated protein-coding genes from one species to the genome of another [43–45]. GeMoMa does not work with protein sequence input in FASTA format but requires a *gff3* or *gtf* file with the annotation of a related species. It was previously shown that GeMoMa has higher base Sensitivity in the human genome using the zebrafish annotation as the donor, while miniprot has higher base Sensitivity in the fruit fly when using the mosquito annotation

as input. It is to be expected that a pipeline such as GALBA will yield more accurate results using GeMoMa instead of miniprot if GeMoMa achieves higher accuracy with a given input scenario. We have previously demonstrated that combining GeMoMa with BRAKER [46] and TSEBRA can be beneficial for annotating plant and insect genomes [47–49]. Particularly for larger genomes, it is worth replacing BRAKER2 with GALBA in such workflows in the future.

Recently, Helixer demonstrated the potential of modern machine learning for genome annotation [50]. Accuracy is not competitive, yet, possibly because these methods do not currently allow for the integration of extrinsic evidence. However, we believe that once an improved and more accurate gene finder on the basis of modern machine learning technology has been trained, it will be of great advantage not only in terms of accuracy, but also in terms of reduced runtime compared to methods such as GALBA.

We intend to expand GALBA in the future. For example, we might incorporate Helixer for faster trimming of the flanking regions of training genes for AUGUSTUS. Also, there is room for improvement in the hints generation given that the protein donors for GALBA might not always be closely related (see Additional file 1: Table S2).

There is a substantial gap in data processing between producing a GALBA (or BRAKER2) output and submission of the annotation to e.g. NCBI Genomes. This gap is already addressed in FunAnnotate, and also to some extent in MOSGA, a web service that executes BRAKER [51]. We expect the definition of a new standard for third-party genome annotation tagging in the foreseeable future. We will then adapt GALBA to produce an annotation that matches this novel standard in order to facilitate genome annotation tagging.

Conclusions

GALBA is an easy-to-use pipeline for the annotation of protein coding genes. It has competitive accuracy, in particular, it is superior to the BRAKER2 pipeline in the annotation of large vertebrate genomes.

Methods

Sequences for accuracy estimation

For estimating prediction accuracy of gene prediction tools, genomes with an already existing annotation are required. Here, we resort to using the genomes and annotations of 14 species (see Table 5), collected from two previous publications. Data of *Arabidopsis thaliana*, *Bombus terrestris*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Rhodnius prolixus*, *Parasteatoda tepidariorum*, *Populus trichocarpa*, *Medicago truncatula*, *Solanum lycopersicum*, and *Xenopus tropicalis* prepared as described in [20],² annotation supporting RNA-Seq evidence described at [53]. In addition, we used the following genomes and annotations from [7]³: *Danio rerio*, *Gallus gallus*, and *Mus musculus*. For each species, *reliable* transcripts were identified, either by definition if at least two annotation providers report a transcript identically, or if all introns of a transcript have support by a spliced alignment from RNA-Seq evidence sampled with VARUS [55]

² genomes, repeat masking and annotation processing documented at [52].

³ Described at [54].

Table 5 Summary of genomes and annotations used for accuracy evaluation

Species	Size (Mbp)	#Genes	#Transcripts	Mono:Mult	#ReliableTx
<i>Arabidopsis thaliana</i>	119	27,445	48,149	0.30	17,800 ^b
<i>Bombus terrestris</i>	249	10,581	22,091	0.06	7481 ^b
<i>Caenorhabditis elegans</i>	100	20,172	33,624	0.0	15,819 ^b
<i>Dano rerio</i>	1345	25,611	42,934	0.08	19,978 ^a
<i>Drosophila melanogaster</i>	138	13,930	30,561	0.25	10,321 ^b
<i>Gallus gallus</i>	1050	17,279	38,534	0.09	12,733 ^a
<i>Medicago truncatula</i>	420	44,464	44,464	0.54	20,059 ^b
<i>Mus musculus</i>	2723	22,405	58,318	0.20	20,708 ^a
<i>Parasteatoda tepdariorum</i>	1445	18,602	27,516	0.19	7926 ^b
<i>Populus trichocarpa</i>	389	34,488	52,085	0.35	22,203 ^b
<i>Rhodnius prolixus</i>	706%MCE- PASTEBIN%	15,061	15,075	0.19	3340 ^b
<i>Solanum lycopersicum</i>	773	33,562	33,562	0.32	13,803 ^b
<i>Tetraodon nigroviridis</i>	359	19,589	23,105	0.04	2112 ^b
<i>Xenopus tropicalis</i>	1449	21,821	45,081	0.11	14,683 ^b

Data extracted from Table 4 in [7] and computed from raw data of [7, 20]. Note that #ReliableTx (for reliable transcripts) has two different meanings: ^aTranscripts that are annotated identically by at least two reference annotation providers, ^bTranscripts that have support in all introns by RNA-Seq evidence

Table 6 Genomes *de novo* annotated with GALBA using reference protein sets listed in Additional file 1: Table S1 as use cases that demonstrate the applicability of GALBA

Species	Assembly	Size (Gbp)	nSeqs	N50 (nt)	BUSCO C (%)	RM (%)
<i>Vespula vulgaris</i>	GCA_014466185.1	0.18	35	8,304,510	94.9	19.5
<i>Vespula germanica</i>	GCA_014466195.1	0.18	133	8,396,154	93.6	19.9
<i>Vespula pensylvanica</i>	GCA_014466175.1	0.18	225	8,532,720	96.2	19.4
<i>Polistes dominula</i>	GCA_001465965.1	0.21	1,483	1,625,592	95.7	48.1
<i>Balaenoptera bonaerensis</i>	GCA_000978805.1	2.23	421,444	20,082	54.1	34.0
<i>Eubalaena japonica</i>	GCA_004363455.1	2.69	1,353,963	39,813	74.9	43.3
<i>Inia geoffrensis</i>	GCA_004363515.1	2.60	1,213,610	26,707	67.7	43.8
<i>Kogia breviceps</i>	GCA_004363705.1	2.76	1,252,072	28,812	66.1	41.3
<i>Phocoena phocoena</i>	GCA_004363495.1	2.70	1,331,158	115,969	85.9	44.7
<i>Platanista gangetica</i>	GCA_004363435.1	2.67	1,098,790	23,933	59.1	44.7
<i>Ziphius cavirostris</i>	GCA_004364475.1	3.15	3,758,276	3,608	39.9	45.1
<i>Coix aquatica</i>	GCA_009725075.1	1.62	2,012	148,397,812	97.8	83.3

nSeqs number of sequences in the assembly, BUSCO C percentage of BUSCOs detected as complete, RM percentage of repeatmasked nucleotides in assembly

As protein input, we manually selected the reference protein sets listed in Additional file 1: Table S1 from NCBI Genomes. These include close relatives of the target species. In short, we used NCBI Taxonomy [56] to identify species that are closely related to the target species and that have a protein sequence set originating from nuclear genome annotation. In order to enable a direct comparison with BRAKER2 (which cannot be executed with a protein set from only one reference species), we ensured to pick a minimum of three protein sets for annotating each species.

Since GALBA is a pipeline that may also be executed with only one reference proteome, we also present accuracy with such single-species protein sets. In general, we

selected the closest relative, with the exception of experiments in *Drosophila melanogaster*, where we excluded *D. simulans* and *D. erecta* from the combined protein set, and from selection as single species reference because they have less than 0.2 expected mutations per genomic site and are thus extremely similar to the target species (see Fig. 2).

Successful generation of high-quality protein to genome alignments depends on the phylogenetic distance between donor and target species. We demonstrate this by evaluating GALBA in single-reference-mode on *D. melanogaster*, using protein donor species arranged on a phylogenetic tree from [57].

Software

All software versions used to generate results in this manuscript are listed in Additional file 1: Table S5.

Miniprot extensions

Miniprot was modified to output detailed residue alignment in a compact custom format to facilitate alignment parsing for scoring with miniprothint. An example of this format is shown in Additional file 1: Figure S1. Further, a new option `-I` was introduced that automatically sets the maximal size of introns to $3.6 \cdot \sqrt{\text{genomeSize}}$. On the *Drosophila-Anopheles* benchmark dataset used in the miniprot paper [24], the new feature doubles the alignment speed and reduces the number of spurious introns by 16.3% at the cost of missing 0.5% of introns that are longer than the threshold.

Miniprothint

During early development of GALBA, it became clear that miniprot (like any spliced aligner) may produce spurious alignments if the reference proteins originate from distantly related species (compare Additional file 1: Table S2). Furthermore, conflicting alignments of homologous proteins from multiple donor species negatively impacted the quality of the AUGUSTUS training gene set. To solve these problems, we wrote an alignment scorer—here called miniprothint—that scores all predicted introns by computing the intron border alignment (IBA) and the intron mapping coverage (IMC) scores. Briefly, the IBA score characterizes the conservation of exons adjacent to the scored intron, with larger weights given to parts close to the donor and acceptor splice sites. The IMC score counts how many times a given intron was exactly mapped by spliced alignments of distinct target proteins. See [58], pages 20 and 21, for a precise definition of both scores.

Based on these scores, miniprothint discards the least reliable evidence and separates the remaining evidence into two classes: high- and low-confidence (see Additional file 1: Figure S6 for more details). High-confidence evidence is used to select training gene candidates for AUGUSTUS and is enforced during gene prediction with AUGUSTUS. Low-confidence evidence is supplied to AUGUSTUS in the form of prediction hints. In comparison to the scoring defined in [58], miniprothint adds penalties for in-frame stop codons and frameshifts (common in the alignments of remote homologs) and significantly improves the computational speed of alignment scoring. The speed improvements

are, in part, achieved by taking advantage of minimap2's compact alignment format (see Additional file 1: Figure S1).

Iterative training

When generating putative training genes for AUGUSTUS from any kind of extrinsic evidence, typically, only some of the actually existing gene structures will be identified in the genome. Otherwise, one would not need to train a gene finder to find the others. In the case of AUGUSTUS, training genes are excised from the genome with flanking and hopefully truly intergenic regions. There is a certain risk that a flanking region will, in fact, carry parts of neighboring genes. Using such "contaminated" intergenic regions can lead to sub-optimal training results. Therefore, we implemented the training of AUGUSTUS in GALBA as follows (e.g., suggested in [9]):

- 1 `et` training on the original training genes derived from evidence with possibly contaminated flanking regions
- 2 prediction of genes with the evidence by AUGUSTUS after initial training
- 3 selection of predicted genes with 100% evidence support, other genes are only eliminated from flanking regions
- 4 `et` training with training genes with filtered flanking regions that are free of predicted genes
- 5 `optimize_augustus.pl` for metaparameter optimization

Multithreading AUGUSTUS

AUGUSTUS is not multithreaded and the gene prediction and metaparameter optimization steps can have a relatively long running time. To address this issue, the BRAKER pipelines split the genome into individual sequence files and execute AUGUSTUS using the Perl module `ParallelForkManager`. However, this approach can strain the file system when dealing with highly fragmented genomes, as a large number of files need to be generated.

To overcome this limitation, we developed `Pygustus`, a Python wrapper for AUGUSTUS that supports parallel execution. This allows for multithreading of AUGUSTUS prediction on genomes of any size and fragmentation level. Large chromosomes are split into overlapping chunks that are not too large for fast parallel execution. The overlaps are introduced to prevent the truncation of genes. Conversely, many short sequences are joined into temporary FASTA files of which there are not too many to strain the file system. `Pygustus` automatically and invisibly to the user decides what sequences to split or join, and assemblies are allowed to have simultaneously very many (small) sequences and (few) very large sequences. The annotation is then done in parallel and the redundancies in annotations from overlapping runs are removed.

In GALBA, we use `Pygustus` to multithread AUGUSTUS predictions, thereby enabling efficient genome annotation without compromising the file system. This approach can be particularly useful for researchers dealing with large and complex genomes, where computational efficiency is critical.

Repeat masking

The genomes of 14 species used for accuracy assessment were previously masked for repeats in [13] and [7]. In short, species-specific repeat libraries were generated with RepeatModeler2 [59]. Subsequently, the genomes were masked with RepeatMasker [60] using those libraries. For vertebrate genomes, an additional step of masking with TandemRepeatsFinder [61] was performed.⁴

The same approach was adopted for each whale and dolphin genome (including the TandemRepeatsFinder step). The additional TandemRepeatsFinder step was not applied to the insects and the plant in Table 6. For *Polistes dominula*, we used repeat masking as provided by NCBI Genomes. Genomes of *Vespula* species were masked with RepeatModeler and RepeatMasker as described in [31].

Accuracy evaluation

For selected genomes, we used the existing reference annotation to assess Sensitivity⁵ and Specificity⁶ of predictions by GALBA, BRAKER2, FunAnnotate, and TSEBRA on gene, transcript and exon level. For this purpose, we used the script `compute_accuracies.sh` that is a part of the BRAKER code. To summarize Sensitivity and Specificity, we computed the F1-score as

$$\frac{2 \cdot \text{Sensitivity} \cdot \text{Specificity}}{\text{Sensitivity} + \text{Specificity}}$$

Prediction quality estimation

For estimating the quality of gene prediction in previously unannotated genomes, we provide BUSCO Sensitivity of both genomes and predicted proteomes [30], and OMAR results [40]. For BUSCO assessment of use case insect assembly and proteome completeness, we used *hymenoptera_odb10*. In dolphins and whales, we used the *vertebrate_odb10* lineage. For *Coix aquatica*, we used the *poales_odb10*. Further, we report basic metrics such as the number of predicted genes, the number of transcripts, the recently suggested mono-exonic to multi-exonic gene ratio [42], and the maximum number of exons per gene across all predicted genes.

To provide a more fine-grained view on the insect annotation use case, we use GeneValidator [37], which scores the predicted proteins to a reference set by length, coverage, conserved regions, and identifies putative merges. Each predicted protein receives an individual score, with 90 being considered a good prediction, and a score of 0 indicating a very poor prediction, or a lack of BLAST hits to the reference proteome to estimate potential lengths and conserved regions. In this instance, we use our input proteome for the prediction tools (Swiss-Prot and RefSeq of *A. mellifera* and *P. canadensis*) consisting of 611,968 proteins.

⁴ See [53].

⁵ $\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$

⁶ $\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$

Assembly statistics

We used seqstats and BUSCO to report basic assembly metrics (see Additional file 1: Methods).

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-023-05449-z>.

Additional file 1. Supplementary Material.

Acknowledgements

We thank Stefan Kemnitz from the University of Greifswald Computing Center for support in designing the software container. We thank Felix Becker for help with publishing python packages to PyPI.

Author Contributions

T.B. developed miniprot boundary scorer and miniprothint; H.L. modified miniprot; N.N. evaluated intron accuracy on data sets that gave rise to the development of miniprothint; D.H. implemented Pygustus; M.E. ran FunAnnotate and participated in experimental design; J.G. contributed use case; S.H. and M.S. supervised Pygustus development; L.G. provided BRAKER2 ODB results; T.B., H.L., and K.J.H. conceptualized the pipeline; K.J.H. and T.B. implemented the pipeline; all authors wrote the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. The position of L.G. is funded by the US National Institute of Health Grant GM128145 to M.S. The Ph.D. project of N.N. is partially funded by German Research Foundation Grant 277249973 to K.J.H. The positions of N.N. and M.E. are partially funded by *Project Data Competency* granted to K.J.H. and M.S. by the government of Mecklenburg-Vorpommern. H.L. is supported by US National Institute of Health Grant R01HG010040. D.H. was funded by German Research Foundation Grant 391397397 to S.H. and M.S. Funding bodies did not play any role in the design of the study or collection, analysis, or interpretation of data or in writing the manuscript.

Availability of data and materials

The datasets analysed during the current study have previously been made available. We are summarizing data sources in the GALBA-data repository at <https://github.com/KatharinaHoff/GALBA-data>. The corresponding author of this manuscript can be contacted if somebody wants to request the data from this study.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 18 April 2023 Accepted: 21 August 2023

Published online: 31 August 2023

References

1. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, et al. Earth BioGenome project: sequencing life for the future of life. *Proc Natl Acad Sci.* 2018;115(17):4325–33.
2. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol.* 2021;39(11):1348–65.
3. Lawniczak MK, Durbin R, Flicek P, Lindblad-Toh K, Wei X, Archibald JM, Baker WJ, Belov K, Blaxter ML, Marques Bonet T, et al. Standards recommendations for the Earth BioGenome Project. *Proc Natl Acad Sci.* 2022;119(4):2115639118.
4. Hope H, Willis S, Markie M, Elliott L. Wellcome Open Research. <https://wellcomeopenresearch.org/browse/articles> Accessed 10 April 2023. 2023.
5. for Biotechnology Information NC. NCBI Genomes. <https://www.ncbi.nlm.nih.gov/genome/browse/#/eukaryotes/> Accessed 10 April 2023. 2023.
6. Gabriel L, Hoff KJ, Bruna T, Lomsadze A, Borodovsky M, Stanke M. The BRAKER3 genome annotation pipeline. *Plant and Animal Genomes Conference.* 2023;30.
7. Bruna T, Lomsadze A, Borodovsky M. GeneMark-ETP: automatic gene finding in eukaryotic genomes in consistency with extrinsic data. *bioRxiv.* 2023. <https://doi.org/10.1101/2023.01.13.524024>.
8. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics.* 2008;24(5):637–44.
9. Hoff KJ, Stanke M. Predicting genes in single genomes with AUGUSTUS. *Curr Protoc Bioinform.* 2019;65(1):57.

10. Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, Zdobnov EM. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* 2023;51(D1):445–51.
11. Korf I. Gene finding in novel genomes. *BMC Bioinform.* 2004;5(1):1–9.
12. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* 2008;18(12):1979–90.
13. Bruna T, Lomsadze A, Borodovsky M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform.* 2020;2(2):026.
14. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* 2005;33(20):6494–506.
15. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 2008;18(1):188–96.
16. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.* 2011;12(1):1–14.
17. Campbell MS, Holt C, Moore B, Yandell M. Genome annotation and curation using MAKER and MAKER-P. *Curr Protoc Bioinform.* 2014;48(1):4–11.
18. FunAnnotate. 2023. <https://github.com/nextgenusfs/funannotate> Accessed Accessed 10 April 2023.
19. Ranz JM, González PM, Clifton BD, Nazario-Yepiz NO, Hernández-Cervantes PL, Palma-Martínez MJ, Valdivia DI, Jiménez-Kaufman A, Lu MM, Markow TA, et al. A de novo transcriptional atlas in *Danaus plexippus* reveals variability in dosage compensation across tissues. *Commun Biol.* 2021;4(1):791.
20. Bruna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform.* 2021;3(1):108.
21. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* 2015;12(1):59–60.
22. Gotoh O. Direct mapping and alignment of protein sequences onto genomic sequence. *Bioinformatics.* 2008;24(21):2438–44.
23. Iwata H, Gotoh O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Res.* 2012;40(20):161–161.
24. Li H. Protein-to-genome alignment with minipro. *Bioinformatics.* 2023;39(1):014.
25. Gabriel L, Hoff KJ, Bruna T, Borodovsky M, Stanke M. TSEBRA: transcript selector for BRAKER. *BMC Bioinform.* 2021;22(1):1–12.
26. Bruna T. miniprohint. <https://github.com/tomasbruna/miniprohint.git> Accessed Accessed 10 April 2023. 2023.
27. Gabriel L, Bruna T, Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. BRAKER. <https://github.com/Gaius-Augustus/BRAKER>. Accessed 10 April 2023. 2023.
28. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet.* 2012;13(5):329–42.
29. Bruna T. OrthoDB-clades. <https://github.com/tomasbruna/orthodb-clades> Accessed Accessed 10 April 2023. 2023.
30. Manni M, Berkeley MR, Seppey M, Zdobnov EM. BUSCO: assessing genomic data quality and beyond. *Curr Protoc.* 2021;1(12):323.
31. Harrop TW, Guhlin J, McLaughlin GM, Permina E, Stockwell P, Gilligan J, Le Lec MF, Gruber MA, Quinn O, Lovegrove M, et al. High-quality assemblies for three invasive social wasps from the *Vespa* genus. *G3: Genes Genom Genet.* 2020;10(10):3479–88.
32. Standage DS, Berens AJ, Glastad KM, Severin AJ, Brendel VP, Toth AL. Genome, transcriptome and methylome sequencing of a primitively eusocial wasp reveal a greatly reduced dna methylation system in a social insect. *Mol Ecol.* 2016;25(8):1769–84.
33. Uniprot. The universal protein knowledgebase in 2023. *Nucleic Acids Res.* 2023;51(D1):523–31.
34. O'Leary NA, Wright MW, Brister JR, Ciufu S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016;44(D1):733–45.
35. Wallberg A, Bunikis I, Pettersson OV, Mosbech M-B, Childers AK, Evans JD, Mikheyev AS, Robertson HM, Robinson GE, Webster MT. A hybrid de novo genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. *BMC Genom.* 2019;20:1–19.
36. Patalano S, Vlasova A, Wyatt C, Ewels P, Camara F, Ferreira PG, Asher CL, Jurkowski TP, Segonds-Pichon A, Bachman M, et al. Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies. *Proc Natl Acad Sci.* 2015;112(45):13970–5.
37. Drăgan M-A, Moghul I, Priyam A, Bustos C, Wurm Y. GeneValidator: identify problems with protein-coding gene predictions. *Bioinformatics.* 2016;32(10):1559–61.
38. A comparative genomics multitool for scientific discovery and conservation. *Nature.* 2020;587(7833):240–245.
39. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The sequence read archive: a decade more of explosive growth. *Nucleic Acids Res.* 2022;50(D1):387–90.
40. Nevers Y, Rossier V, Train C, Altenhoff AM, Dessimoz C, Glover N. Multifaceted quality assessment of gene repertoire annotation with OMArk. *bioRxiv.* 2022;2022–11.
41. Guo C, Wang Y, Yang A, He J, Xiao C, Lv S, Han F, Yuan Y, Dong X, et al. The coix genome provides insights into panicoidae evolution and papery hull domestication. *Mol Plant.* 2020;13(2):309–20.
42. Vuruputoor VS, Monyak D, Fetter KC, Webster C, Bhattarai A, Shrestha B, Zaman S, Bennett J, McEvoy SL, Caballero M, et al. Welcome to the big leaves: best practices for improving genome annotation in non-model plant genomes. *bioRxiv.* 2022. <https://doi.org/10.1101/2022.10.03.510643>
43. Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* 2016;44(9):89–89.
44. Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinform.* 2018;19:1–12.
45. Keilwagen J, Hartung F, Grau J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Gene prediction: Methods Protoc.* 2019;161–177.

46. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*. 2016;32(5):767–9.
47. Erbbi M, Keilwagen J, Hoff KJ, Steffen R, Altmüller J, Oettler J, Schrader L. Transposable elements and introgression introduce genetic variation in the invasive ant *Cardiocondyla obscurior*. *Mol Ecol*. 2021;30(23):6211–28.
48. Wöhner TW, Emeriewen OF, Wittenberg AH, Schneiders H, Vrijenhoek I, Halász J, Hrotkó K, Hoff KJ, Gabriel L, Lempe J, et al. The draft chromosome-level genome assembly of tetraploid ground cherry (*Prunus fruticosa* Pall.) from long reads. *Genomics*. 2021;113(6):4173–83.
49. Woehner TW, Emeriewen OF, Wittenberg AH, Nijbroek K, Wang RP, Blom E-J, Keilwagen J, Berner T, Hoff KJ, Gabriel L, et al. The structure of the tetraploid sour cherry 'Schattenmorelle' (*Prunus cerasus* L.) genome reveals insights into its segmental allopolyploid nature. *bioRxiv*. 2023–03;2023.
50. Stiehler F, Steinborn M, Scholz S, Dey D, Weber AP, Denton AK. Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics*. 2020;36(22–23):5291–8.
51. Martin R, Hackl T, Hattab G, Fischer MG, Heider D. Mosga: modular open-source genome annotator. *Bioinformatics*. 2020;36(22–23):5514–5.
52. Bruna T. EukSpecies-BRAKER2. <https://github.com/gatech-genemark/EukSpecies-BRAKER2>. Accessed 10 April 2023. 2023.
53. Bruna T. BRAKER2-exp. <https://github.com/gatech-genemark/BRAKER2-exp>. Accessed 10 April 2023. 2023.
54. Bruna T. GeneMark-ETP-exp. <https://github.com/gatech-genemark/GeneMark-ETP-exp>. Accessed 10 April 2023. 2023.
55. Stanke M, Bruhn W, Becker F, Hoff KJ. VARUS: sampling complementary RNA reads from the sequence read archive. *BMC Bioinform*. 2019;20:1–7.
56. Schoch CL, Ciuffo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*. 2020;2020:baaa062.
57. König S, Romoth L, Stanke M. Comparative genome annotation. *Comp Genom Methods Protoc* 2018;189–212.
58. Bruna T. Unsupervised algorithms for automated gene prediction in novel eukaryotic genomes. Ph.D thesis, Georgia Institute of Technology. 2022.
59. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci*. 2020;117(17):9451–7.
60. Chen N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinform*. 2004;5(1):4–10.
61. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27(2):573–80.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

