# Supplementary Materials<br>GALBA: Genome Annotation with Miniprot and AUGUSTUS

Tomáš Brůna[2], Heng Li[3], Joseph Guhlin[4], Daniel Honsel[5], Steffen Herbold[6], Mario Stanke[1], Natalia Nenasheva[1], Matthis Ebel[1], Lars Gabriel[1], and Katharina J. Hoff*[1]

[1]Institute of Mathematics and Computer Science & Center for Functional Genomics of Microbes, University of Greifswald, 17489 Greifswald, Germany
[2]US Department of Energy Joint Genome Institute, Berkeley, CA 94720, USA
[3]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215, USA & Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[4]Genomics Aotearoa and Laboratory for Evolution and Development, Department of Biochemistry, University of Otago, PO Box 56, 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

July 28, 2023

## Contents

**S1 Supplementary Figures** **S1**

**S2 Supplementary Tables** **S6**

**S3 Supplementary Methods** **S13**
S3.1 Assembly Quality Estimation . . . . . S13
S3.2 Annotation Parameter Computation . . . . . S13
S3.3 Running FunAnnotate . . . . . S13
S3.4 Running GALBA . . . . . S14
S3.5 Running BRAKER2 . . . . . S14
S3.6 Running TSEBRA . . . . . S14
S3.7 Data Preparation for Estimating the Number of Genomes at NCBI that May Benefit from GALBA . . . . . S14

**S4 Supplementary Results** **S15**
S4.1 Accuracy Improvements during GALBA Development . . . . . S15

## S1 Supplementary Figures

*corresponding author: katharina.hoff@uni-greifswald.de

```
##ATN GAGGCC---CGCTCACCgtactgactgatgccatcggtatcgattcggagctagcttagtcaagCACAAGCGCTATAGCCTAC
##ATA E..A..-..R..S..P.                                               .T..$$R..Y..!A..Y..
##AAS |  |           |                                                 |    |  +   |  |
##AQA E  A  F  H  -  P                                                 T  E R  W   A  Y
```

Figure S1: Custom alignment format produced by miniprot executed with option `--aln`. Here, `ATN` stands for target nucleotides, `ATA` for translated target codons, `AAS` for amino acid alignment quality, and `AQA` for query protein amino acids. "$" and "!" represent frameshifts. If an intron is longer than 200bp, only 100+100bp are shown while an integer in the middle may indicate the total intron length, e.g.: `...gtcatgcta∼500∼tacgatgactag....`

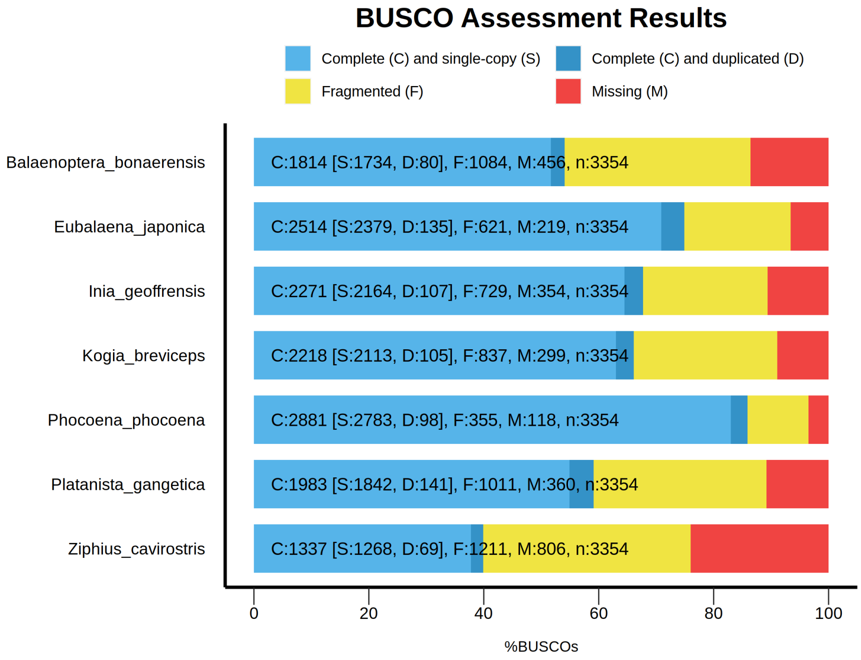

Figure S2: BUSCO scores (obtained with vertebrata_odb10) in whale and dolphin genome assemblies.

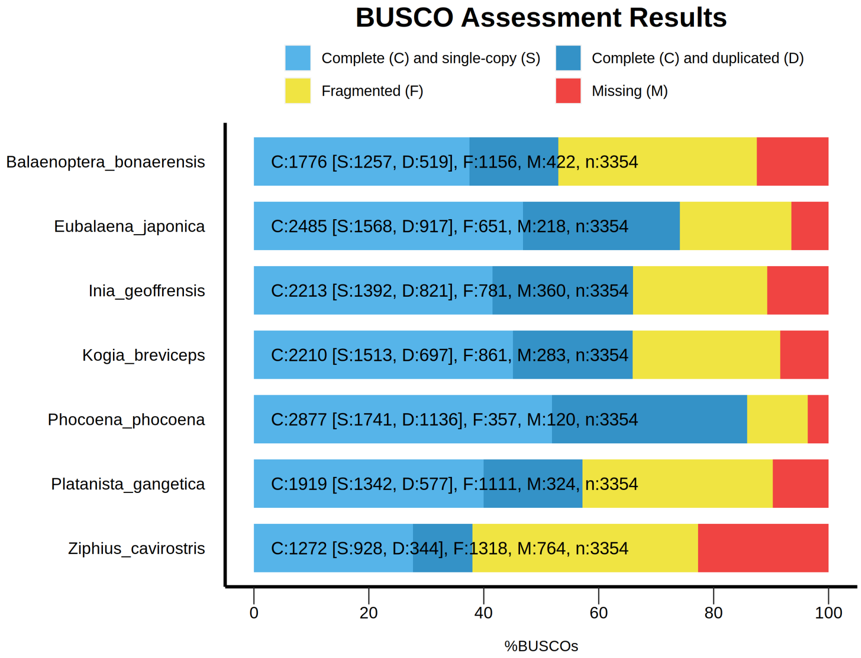

Figure S3: BUSCO scores (obtained with vertebrata_odb10) of proteins predicted with GALBA in whale and dolphin genomes.

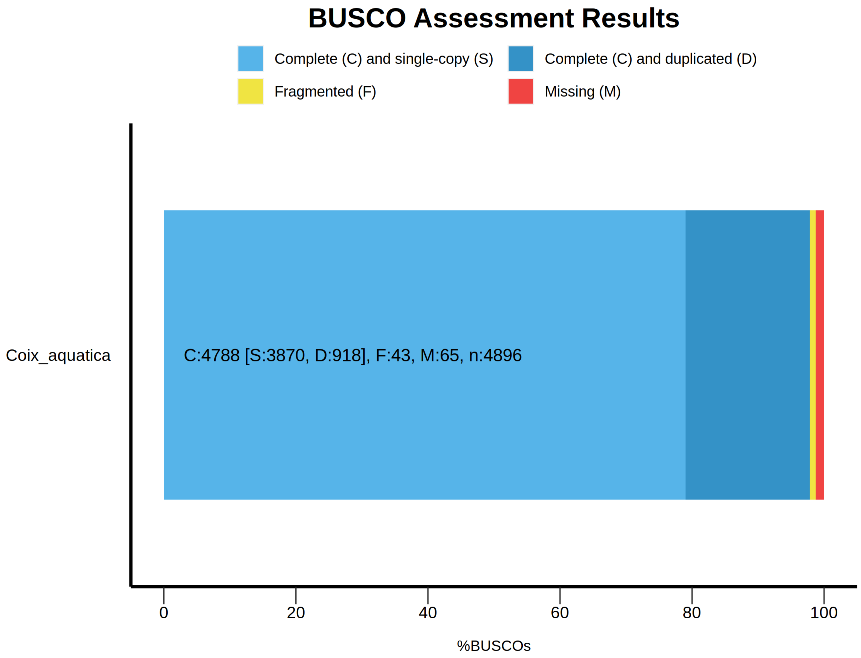

Figure S4: BUSCO scores (obtained with poales_odb10) of proteins predicted with GALBA in *Coix aquatica*.

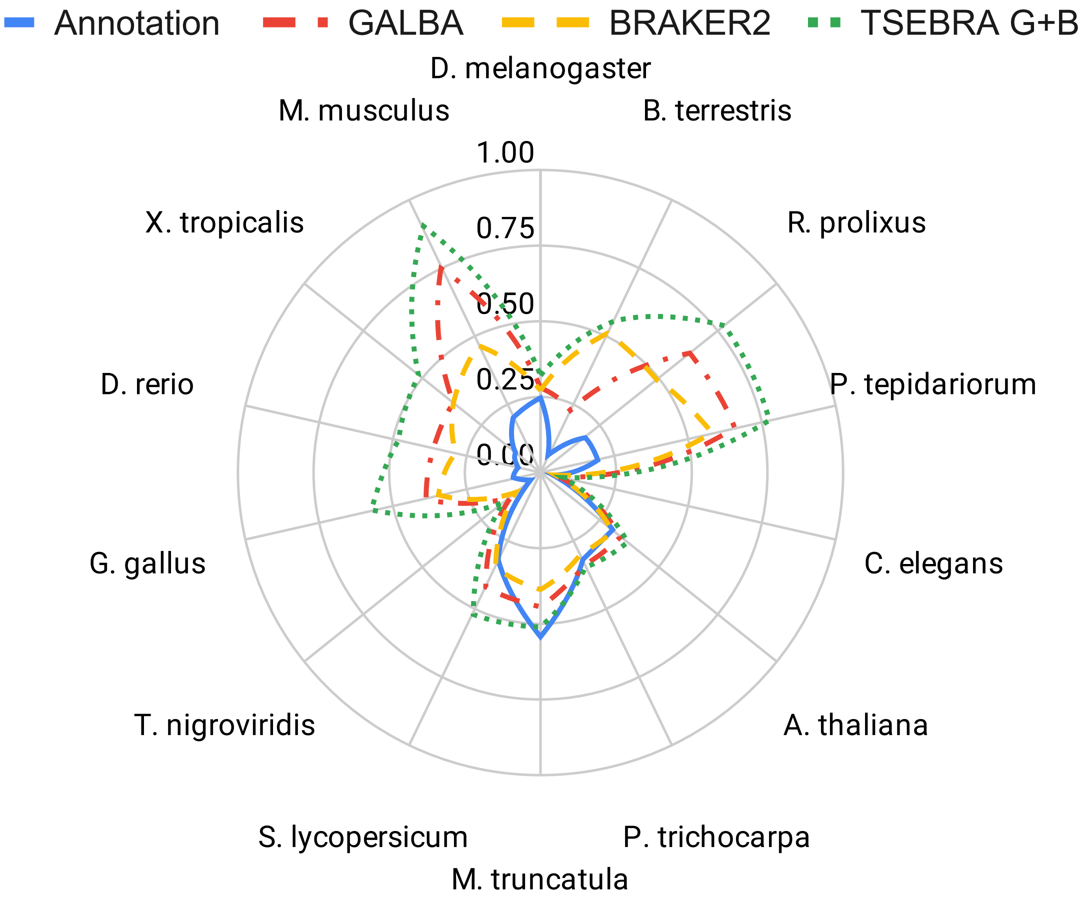

Figure S5: Network plot of mono-exonic to multi-exonic gene ratios. The species are in a clockwise fashion from the top insects with increasing genome size, followed by *C. elegans*, the only metazoan in our experiments, followed by plants with increasing genome size, followed by vertebrates with increasing genome size. We show this ratio for the reference annotation, GALBA, BRAKER2, and combination of GALBA and BRAKER2 with TSEBRA.

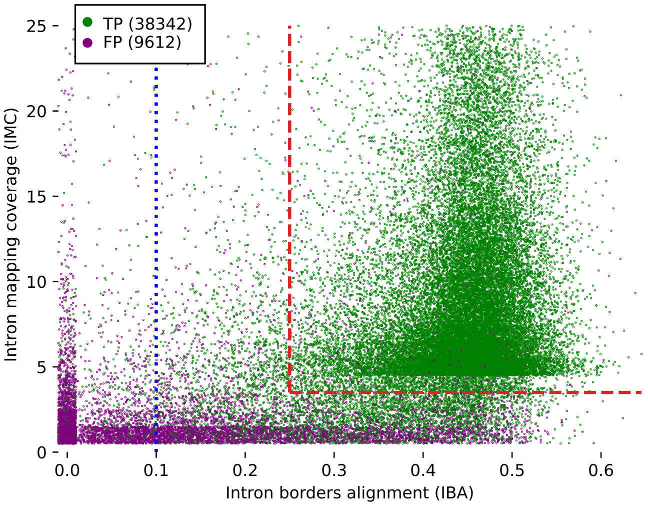

Figure S6: Introns predicted by miniprot, characterized by miniprothint-derived IMC and IBA scores. The predictions originate from running miniprot on *D. melanogaster* with reference proteomes of five other *Drosophila* species (see Figure 2 in the main manuscript for the list of reference species). A small random offset was added to each item to reduce the amount of overlapping data points. Miniprothint discards all introns with IBA < 0.1 (the blue dotted line). This step improved the prediction Specificity from 80.0% to 89.8% at the cost of a Sensitivity decrease from 80.3% to 78.8%. Miniprothint also defines a set of high-confidence hints characterized by IBA >= 0.25 and IMC >= 4 (the red dashed lines). This further improved the Specificity to 98.5% while reducing the Sensitivity to 68.9%.

# S2 Supplementary Tables

Table S1: Donor proteins used for annotating each species genome with GALBA, FunAnnotate, and BRAKER2. Note: The proteins for whales and dolphins were applied to all whale and dolphin species with GALBA. *) Proteins were not used in the combined set but only for single protein set input experiments. $^s$) Proteins were used to demonstrate GALBA accuracy with reference proteins from this species, alone (GALBA$^s$ in Table 1).

| Species | Reference Protein File |
|---|---|
| Arabidopsis thaliana | |
| Arabidopsis lyrata subsp. lyrata$^s$ | GCF_000004255.2_v.1.0_protein.faa.gz |
| Arabidopsis thaliana x Arabidopsis arenosa | GCA_019202795.1_ASM1920279v1_protein.faa.gz |
| Camelina sativa | GCF_000633955.1_Cs_protein.faa.gz |
| Arabidopsis suecica | GCA_019202805.1_ASM1920280v1_protein.faa.gz |
| Capsella rubella | GCF_000375325.1_Caprub1_0_protein.faa.gz |
| Bombus terrestris | |
| Bombus vancouverensis nearcticus | GCF_011952275.1_Bvanc_JDL1245_protein.faa.gz |
| Bombus huntii | GCF_024542735.1_iyBomHunt1.1_protein.faa.gz |
| Bombus affinis | GCF_024516045.1_iyBomAffi1.2_protein.faa.gz |
| Bombus pyrosoma | GCF_014825855.1_ASM1482585v1_protein.faa.gz |
| Bombus vosnesenskii | GCF_011952255.1_Bvos_JDL3184-5_v1.1_protein.faa.gz |
| Bombus bifarius | GCF_011952205.1_Bbif_JDL3187_protein.faa.gz |
| Bombus impatiens$^s$ | GCF_000188095.3_BIMP_2.2_protein.faa.gz |
| Caenorhabditis elegans | |
| Caenorhabditis auriculariae | GCA_904845305.1_CAUJ_protein.faa.gz |
| Caenorhabditis bovis | GCA_902829315.1_CBOVIS_v1.1_protein.faa.gz |
| Caenorhabditis brenneri | GCA_000143925.2_C_brenneri-6.0.1b_protein.faa.gz |
| Caenorhabditis briggsae$^s$ | GCF_000004555.2_CB4_protein.faa.gz |
| Caenorhabditis remanei | GCF_000149515.1_ASM14951v1_protein.faa.gz |
| Danio rerio | |
| Cyprinus carpio | GCF_018340385.1_ASM1834038v1_protein.faa.gz |
| Carassius auratus | GCF_003368295.1_ASM336829v1_protein.faa.gz |
| Puntigrus tetrazona | GCF_018831695.1_ASM1883169v1_protein.faa.gz |
| Sinocyclocheilus rhinocerous | GCF_001515625.1_SAMN03320098_v1.1_protein.faa.gz |
| Sinocyclocheilus anshuiensis | GCF_001515605.1_SAMN03320099.WGS_v1.1_protein.faa.gz |
| Onychostoma macrolepis$^s$ | GCA_012432095.1_ASM1243209v1_protein.faa.gz |
| Carassius gibelio | GCF_023724105.1_carGib1.2-hapl.c_protein.faa.gz |
| Pimephales promelas | GCF_016745375.1_EPA_FHM_2.0_protein.faa.gz |
| Labeo rohita | GCF_022985175.1_IGBB_LRoh.1.0_protein.faa.gz |
| Megalobrama amblycephala | GCF_018812025.1_ASM1881202v1_protein.faa.gz |
| Sinocyclocheilus grahami | GCF_001515645.1_SAMN03320097.WGS_v1.1_protein.faa.gz |
| Ctenopharyngodon idella | GCF_019924925.1_HZGC01_protein.faa.gz |
| Drosophila melanogaster | |
| Drosophila ananassae$^s$ | GCF_017639315.1_ASM1763931v2_protein.faa.gz |
| Drosophila erectra* | GCF_003286155.1_DereRS2_protein.faa.gz |
| Drosophila grimshawi | GCF_018153295.1_ASM1815329v1_protein.faa.gz |
| Drosophila pseudoobscura | GCF_009870125.1_UCI_Dpse_MV25_protein.faa.gz |
| Drosophila simulans* | GCF_016746395.2_Prin_Dsim_3.1_protein.faa.gz |
| Drosophila virilis | GCF_003285735.1_DvirRS2_protein.faa.gz |
| Drosophila willistoni | GCF_018902025.1_UCI_dwil_1.1_protein.faa.gz |
| Musca domestica* | GCF_000371365.1_Musca_domestica-2.0.2_protein.faa.gz |
| Gallus gallus | |
| Lagopus muta | GCF_023343835.1_bLagMut1_primary_protein.faa.gz |
| Tympanuchus pallidicinctus | GCF_026119805.1_pur_lepc_1.0_protein.faa.gz |
| Lagopus leucura | GCF_019238085.1_USGS_WTPT01_protein.faa.gz |
| Centrocercus urophasianus | GCF_019232065.1_USGS_Curo_1.0_protein.faa.gz |
| Centrocercus urophasianus | GCF_019232065.1_USGS_Curo_1.0_protein.faa.gz |
| Coturnix japonica$^s$ | GCF_001577835.2_Coturnix_japonica_2.1_protein.faa.gz |
| Meleagris gallopavo | GCF_000146605.3_Turkey_5.1_protein.faa.gz |
| Medicago truncatula | |
| Trifolium pratense$^s$ | GCF_020283565.1_ARS_RC_1.1_protein.faa.gz |
| Pisum sativum | GCF_024323335.1_CAAS_Psat_ZW6_1.0_protein.faa.gz |
| Cicer arietinum | GCF_000331145.1_ASM33114v1_protein.faa.gz |
| Mus musculus | |
| Arvicanthis niloticus | GCF_011762505.1_mArvNil1.pat.X_protein.faa.gz |
| Grammomys surdaster | GCF_004785775.1_NIH_TR_1.0_protein.faa.gz |
| Mastomys coucha | GCF_008632895.1_UCSF_Mcou_1_protein.faa.gz |
| Mus pahari | GCF_900095145.1_PAHARI_EIJ_v1.1_protein.faa.gz |

| | |
|---|---|
| Apodemus sylvaticus | GCF_947179515.1_mApoSyl1.1_protein.faa.gz |
| Mus caroli$^s$ | GCF_900094665.1_CAROLI_EIJ_v1.1_protein.faa.gz |
| Rattus rattus | GCF_011064425.1_Rrattus_CSIRO_v1_protein.faa.gz |
| Rattus norvegicus | GCF_015227675.2_mRatBN7.2_protein.faa.gz |
| Homo sapiens | GCF_000001405.40_GRCh38.p14_protein.faa.gz |
| Parasteatoda tepidariorum | |
| Trichonephila inaurata | GCA_019973955.1_Tnin_1.0_protein.faa.gz |
| Caerostris extrusa | GCA_021605095.1_Cext_1.0_protein.faa.gz |
| Caerostris darwini | GCA_021605075.1_Cdar_1.0_protein.faa.gz |
| Oedothorax gibbosus | GCA_019343175.1_Ogib_1.0_protein.faa.gz |
| Trichonephila clavata | GCA_019973975.1_Tnct_1.0_protein.faa.gz |
| Trichonephila clavipes | GCA_019973935.1_Tncv_1.0_protein.faa.gz |
| Araneus ventricosus$^s$ | GCA_013235015.1_Ave_3.0_protein.faa.gz |
| Nephila pilipes | GCA_019974015.1_Npil_1.0_protein.faa.gz |
| Rhodnius prolixus | |
| Nesidiocoris tenuis | GCA_902806785.1_CYROTEf_10X_genome_protein.faa.gz |
| Cimex lectularius$^s$ | GCF_000648675.2_Clec_2.1_protein.faa.gz |
| Halyomorpha halys | GCF_000696795.2_Hhal_2.0_protein.faa.gz |
| Nezara viridula | GCA_928085145.1_PGI_NEZAVIv3_protein.faa.gz |
| Populus trichocarpa | |
| Populus tomentosa | GCA_018804465.1_PTv2_protein.faa.gz |
| Populus euphratica | GCF_000495115.1_PopEup_1.0_protein.faa.gz |
| Populus alba | GCF_005239225.1_ASM523922v1_protein.faa.gz |
| Populus deltoides$^s$ | GCA_015852605.2_ASM1585260v2_protein.faa.gz |
| Solanum lycopersicum | |
| Solanum stenotomum | GCF_019186545.1_ASM1918654v1_protein.faa.gz |
| Solanum tuberosum | GCF_000226075.1_SolTub_3.0_protein.faa.gz |
| Solanum verrucosum | GCF_900185275.1_falcon-dt-bn_protein.faa.gz |
| Solanum pennellii$^s$ | GCF_001406875.1_SPENNV200_protein.faa.gz |
| Tetraodon nigroviridis | |
| Micropterus salmoides | GCF_014851395.1_ASM1485139v1_protein.faa.gz |
| Gasterosteus aculeatus aculeatus | GCF_016920845.1_GAculeatus_UGA_version5_protein.faa.gz |
| Sebastes umbrosus | GCF_015220745.1_fSebUmb1.pri_protein.faa.gz |
| Etheostoma cragini | GCF_013103735.1_CSU_Ecrag_1.0_protein.faa.gz |
| Gymnodraco acuticeps | GCF_902827175.1_fGymAcu1.1_protein.faa.gz |
| Pseudochaenichthys georgianus | GCF_902827115.1_fPseGeo1.1_protein.faa.gz |
| Dissostichus mawsoni | GCA_011823955.1_KU_Dm_1.0_protein.faa.gz |
| Cyclopterus lumpus | GCF_009769545.1_fCycLum1.pri_protein.faa.gz |
| Notolabrus celidotus | GCF_009762535.1_fNotCel1.pri_protein.faa.gz |
| Etheostoma spectabile | GCF_008692095.1_UIUC_Espe_1.0_protein.faa.gz |
| Anarrhichthys ocellatus | GCF_004355925.1_GSC_Weel_1.0_protein.faa.gz |
| Cottoperca gobio | GCF_900634415.1_fCotGob3.1_protein.faa.gz |
| Takifugu rubripes$^s$ | GCF_901000725.2_fTakRub1.2_protein.faa.gz |
| Xenopus tropicalis | |
| Xenopus laevis$^s$ | GCF_001663975.1_Xenopus_laevis_v2_protein.faa.gz |
| Hymenochirus boettgeri | GCA_019447015.1_UCB_Hboe_1.0_protein.faa.gz |
| Eleutherodactylus coqui | GCA_019857665.1_UCB_Ecoq_1.0_protein.faa.gz |
| Engystomops pustulosus | GCA_019512145.1_UCB_Epus_1.0_protein.faa.gz |
| Bufo bufo | GCF_905171765.1_aBufBuf1.1_protein.faa.gz |
| Spea bombifrons | GCF_027358695.1_aSpeBom1.2.pri_protein.faa.gz |
| Rana temporaria | GCF_905171775.1_aRanTem1.1_protein.faa.gz |
| Bufo gargarizans | GCF_014858855.1_ASM1485885v1_protein.faa.gz |
| Bombina bombina | GCF_027579735.1_aBomBom1.pri_protein.faa.gz |
| Wales and dolphins | |
| Lipotes vexillife | GCF_000442215.2_Lipotes_vexillifer_v1.1_protein.faa.gz |
| Delphinapterus leucas | GCF_002288925.2_ASM228892v3_protein.faa.gz |
| Monodon monoceros | GCF_005190385.1_NGI_Narwhal_1_protein.faa.gz |
| Tursiops truncatus | GCF_011762595.1_mTurTru1.mat.Y_protein.faa.gz |
| Neophocaena asiaeorientalis | GCF_003031525.2_Neophocaena_asiaeorientalis_V1.1_protein.faa.gz |
| Phocoena sinus | GCF_008692025.1_mPhoSin1.pri_protein.faa.gz |
| Lagenorhynchus obliquidens | GCF_003676395.1_ASM367639v1_protein.faa.gz |
| Pontoporia blainvillei | GCA_011754075.1_ASM1175407v1_protein.faa.gz |
| Globicephala melas | GCF_006547405.1_ASM654740v1_protein.faa.gz |
| Orcinus orca | GCF_937001465.1_mOrcOrc1.1_protein.faa.gz |
| Physeter catodon | GCF_002837175.2_ASM283717v2_protein.faa.gz |
| Coix aquatica | |
| Zea mays | GCF_902167145.1_Zm-B73-REFERENCE-NAM-5.0_protein.faa.gz |

| | |
|---|---|
| Sorghum bicolor | GCF_000003195.3_Sorghum_bicolor_NCBIv3_protein.faa.gz |
| Miscanthus lutarioriparius | GCA_904845875.1_Mlu_assembly_protein.faa.gz |
| Panicum hallii | GCF_002211085.1_PHallii_v3.1_protein.faa.gz |

| | miniprot raw | | | | miniprothint all | | | | miniprothint HC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TP | FP | Sn | Sp | TP | FP | Sn | Sp | TP | FP | Sn | Sp |
| five close relatives | 38,342 | 9,612 | 80.3 | 80.0 | 37,639 | 4,230 | 78.8 | 89.9 | 32,896 | 511 | 68.9 | 98.5 |
| ODB order excluded | 29,640 | 390,978 | 62.1 | 7.1 | 25,427 | 82,094 | 53.3 | 23.7 | 18,315 | 1,878 | 38.4 | 90.7 |

Table S2: Comparison of intron predictions by spliced alignment using a protein set of closely related species (see Table S1), and the OrthoDB v.11 (ODB) Arthopoda partition (proteins from species of the same order excluded) on *D. melanogaster*. The reference annotation has 47,739 introns. The values in the table—True Positives (TP), False Positives (FP), Sensitivity (Sn), Specificity (Sp)—are shown for the raw miniprot result, all miniprothint predictions, and high-confidence (HC) miniprothint predictions (see Figure 4 for details).

| | Gene Sensitivity | | Exon Sensitivity | |
|---|---|---|---|---|
| | GALBA | BRAKER2 | GALBA | BRAKER2 |
| *A. thaliana* | 86.76 | **91.22** | 89.97 | **91.02** |
| *B. terrestris* | **78.31** | 76.03 | **89.74** | 86.98 |
| *C. elegans* | 59.56 | **77.03** | 80.59 | **88.30** |
| *D. melanogaster* | 72.43 | **77.73** | 81.43 | **82.16** |
| *M. truncatula* | 62.40 | **69.19** | 88.56 | **91.63** |
| *P. tepidariorum* | 44.92 | **45.26** | 81.19 | **82.02** |
| *P. trichocarpa* | 75.80 | **83.51** | 90.59 | **92.41** |
| *R. prolixus* | 42.25 | **47.90** | 77.37 | **81.48** |
| *S. lycopersicum* | 75.88 | **77.17** | 94.02 | **94.55** |
| *T. nigroviridis* | **71.12** | **71.12** | **91.91** | 90.61 |
| *X. tropicalis* | **72.21** | 54.95 | **91.45** | 83.97 |

Table S3: Feature prediction Sensitivity in a subset of annotated multi-exon genes that have support by spliced RNA-Seq to genome alignments in all introns.

| | Gene Sensitivity | | Exon Sensitivity | |
|---|---|---|---|---|
| | GALBA | BRAKER2 | GALBA | BRAKER2 |
| *D. rerio* | **70.16** | 58.78 | **93.49** | 89.4 |
| *G. gallus* | **72.00** | 30.16 | **94.08** | 37.61 |
| *M. musculus* | **77.85** | 40.31 | **95.18** | 61.38 |

Table S4: Feature prediction Sensitivity in a subset of reliably annotated genes. A gene is regarded as reliable if a minimum of two annotation sets contain this exact gene structure.

| Tool | Version (or commit) |
|---|---|
| GALBA | 1.0.6 |
| Python | 3.8 |
| miniprot | 0.9-r224-dirty |
| augustus | 3.5.0 |
| miniprothint | a38f300 |
| miniprot-boundary-scorer | 37493bc |
| braker.pl | 3.0.0 |
| TSEBRA | b0d6c4f |
| GeneMark-EP/ETP | ede6bc5 |
| BUSCO | 5.4.2 |
| FunAnnotate | v1.8.14 |
| Exonerate | v2.4.0 |
| DIAMOND | v2.0.15 |
| EvidenceModeler | 1.1.1 |
| GeneMark (FunAnnotate) | v4.71_lic |
| tbl2asn | 25.8 |
| bedtools | v2.30.0 |
| augustus (FunAnnotate) | 3.3.2 |
| tRNAscan-SE | 2.0.9 |
| minimap2 | 2.24-r1122 |
| RepeatModeler | 2.0.4 |
| RepeatMasker | 4.1.4 |
| NCBI/RMBLAST | 2.13.0+ |
| TRF | 4.09 |
| RECON | 1.08 |
| RepeatScout | 1.0.5 |
| GenomeTools | 1.6.0 |
| LTR_Retriever | v2.9.0 |
| Ninja | 0.97 |
| MAFFT | 7.471 |
| CD-HIT | 4.8.1 |
| Singularity | 3.10.0-dirty |

Table S5: Software versions.

| Species | BUSCO seed species | BUSCO DB |
|---|---|---|
| *Arabidopsis thaliana* | cacao | embryophyta |
| *Bombus terrestris* | fly | arthropoda |
| *Caenorhabditis elegans* | trichinella | metazoa |
| *Dano rerio* | human | vertebrata |
| *Drosophila melanogaster* | nasonia | arthropoda |
| *Gallus gallus* | human | tetrapoda |
| *Medicago truncatula* | cacao | embryophyta |
| *Mus musculus* | chicken | tetrapoda |
| *Parasteatoda tepdariorum* | fly | arthropoda |
| *Populus trichocarpa* | cacao | embryophyta |
| *Rhodnius prolixus* | fly | arthropoda |
| *Solanum lycopersicum* | cacao | embryophyta |
| *Tetraodon nigroviridis* | human | vertebrata |
| *Xenopus tropicalis* | human | tetrapoda |

Table S6: Seed species and BUSCO DB used for BUSCO with FunAnnotate. Parameters were selected in such a way that the species that the AUGUSTUS parameters were trained on is not part of the same order as the target species. We use this scenario to simulate what will happen when annotating representatives of novel clades.

| | *Arabidopsis thaliana* | | | *Bombus terrestris* | | | *Caenorhabditis elegans* | | | *Danio rerio* | | | *Drosophila melanogaster* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon |
| BRAKER2 ODB+ | 76.95 | 61.46 | 85.11 | 47.41 | 39.23 | 79.58 | 69.31 | 56.27 | 87.90 | 29.89 | 23.59 | 72.86 | 76.80 | 58.68 | 83.88 |
| BRAKER2 ODB$^o$ | 71.17 | 56.33 | 83.97 | 37.32 | 29.42 | 75.49 | 51.30 | 41.62 | 80.48 | 27.20 | 21.82 | 72.15 | 60.61 | 46.03 | 76.66 |
| FunAnnotate | 77.26 | 61.81 | 87.03 | 35.51 | 29.04 | 71.66 | 45.53 | 37.39 | 77.84 | 8.95 | 7.40 | 47.04 | 58.24 | 44.68 | 74.41 |
| | *Medicago truncatula* | | | *Parasteatoda tepidariorum* | | | *Populus trichocarpa* | | | *Rhodnius prolixus* | | | *Tetraodon nigroviridis* | | |
| | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon |
| BRAKER2 ODB+ | 46.93 | 45.06 | 74.82 | 21.48 | 19.08 | 64.09 | 66.13 | 57.20 | 82.95 | 13.35 | 12.83 | 54.88 | 9.39 | 8.24 | 58.56 |
| BRAKER2 ODB$^o$ | 44.80 | 43.52 | 74.76 | 19.33 | 17.36 | 62.60 | 63.65 | 55.09 | 82.61 | 12.77 | 12.41 | 54.38 | 9.21 | 8.20 | 58.47 |
| FunAnnotate | 33.33 | 33.33 | 67.89 | 13.71 | 12.48 | 55.20 | 50.11 | 44.38 | 75.94 | 6.89 | 6.89 | 29.51 | 4.42 | 4.11 | 36.91 |
| | *Gallus gallus* | | | *Mus musculus* | | | *Solanum lycopersicum* | | | *Xenopus tropicalis* | | | **Average** | | |
| | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon |
| BRAKER2 ODB+ | 23.11 | 15.60 | 45.57 | 27.20 | 16.90 | 57.27 | 38.45 | 36.12 | 69.41 | 36.48 | 28.23 | 78.21 | 41.63 | 34.18 | 71.08 |
| BRAKER2 ODB$^o$ | 20.14 | 18.53 | 42.83 | 27.01 | 26.41 | 66.09 | 37.50 | 36.29 | 71.29 | 31.18 | 23.97 | 75.91 | 36.66 | 31.22 | 69.78 |
| FunAnnotate | 15.4 | 10.05 | 44.21 | NA | NA | NA | 31.94 | 31.94 | 66.28 | NA | NA | NA | NA | NA | NA |

Table S7: F1-scores of gene predictions from BRAKER2 executed with OrthoDB v11 partitions (species excluded) and proteins of closely related species (BRAKER2 ODB+), and BRAKER2 results with OrthoDB v11 partitions where proteins from the same order as the target species have been excluded (BRAKER2 ODB$^o$), and results of FunAnnotate. FunAnnotate went out of memory for *M. musculus* and *X. tropicalis* on our HPC nodes that had 189 GB RAM.

| Species | Single | Duplicated | Duplicated, Expected | Duplicated, Unexpected | Missing | Consistent | Consistent, partial hits | Consistent, fragmented hits | Inconsistent | Inconsistent, partial hits | Inconsistent fragmented | Contaminants | Unknown |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Balaenoptera bonaerensis* | 54.48 | 44.28 | 43.47 | 0.81 | 1.23 | 71.10 | 11.13 | 41.53 | 28.52 | 5.27 | 21.69 | 0 | 0.39 |
| *Eubalanea japonica* | 67.59 | 31.03 | 30.25 | 0.77 | 1.39 | 67.91 | 9.08 | 32.42 | 31.78 | 4.26 | 23.26 | 0 | 0.31 |
| *Inia geoffrensis* | 69.99 | 27.92 | 27.34 | 0.58 | 2.08 | 69.99 | 9.16 | 31.67 | 29.75 | 4.10 | 21.56 | 0 | 0.26 |
| *Kogia previceps* | 63.27 | 35.13 | 34.66 | 0.48 | 1.59 | 67.60 | 10.14 | 35.59 | 32.19 | 4.87 | 23.90 | 0 | 0.22 |
| *Phocoena phocoena* | 76.97 | 21.48 | 20.77 | 0.70 | 1.55 | 67.44 | 8.11 | 28.47 | 32.21 | 4.71 | 23.76 | 0 | 0.35 |
| *Platanista gangetica* | 54.51 | 44.30 | 43.62 | 0.67 | 1.19 | 70.54 | 11.59 | 37.75 | 29.11 | 4.51 | 21.66 | 0 | 0.34 |
| *Ziphius cavirostris* | 48.30 | 49.67 | 49.12 | 0.57 | 2.02 | 74.62 | 13.19 | 44.83 | 25.18 | 4.13 | 18.42 | 0 | 0.20 |
| *Coix aquatica* | 85.25 | 10.98 | 8.59 | 2.39 | 3.67 | 48.01 | 6.81 | 9.88 | 49.27 | 7.61 | 32.42 | 0 | 2.72 |

Table S8: OMArk results (in percent) in genomes that were *de novo* annotated with GALBA. The number of conserved HOGs for whales and dolphins is 13,050, the number of conserved HOGs for *Coix aquatica* is 20,501.

| OrthoDB partition | Size (#sequences) | Test species |
|---|---|---|
| arthropoda_odb11 | 4,307,558 | *Bombus terrestris*, *Drosophila melanogaster*, *Parasteatoda tepidariorum*, *Rhodnius prolixus* |
| metazoa_odb11 | 15,257,394 | *Caenorhabditis elegans* |
| vertebrata_odb11 | 9,805,833 | *Danio rerio*, *Gallus gallus*, *Tetraodon nigroviridis*, *Mus musculus* |
| viridiplantae_odb11 | 5,310,477 | *Arabidopsis thaliana*, *Medicago truncatula*, *Populus trichocarpa*, *Solanum lycopersicum* |

Table S9: Overview of the OrthoDB partitions and the test species for which they were used. For results in Table 4, each test species, species belonging to the same taxonomic order were excluded from the databases for each experiment. We used the orthodb-clades pipeline to generate the protein sets. For results in Table S7, only the target species were excluded, and this ODB partition was subsequently combined with the close relatives input from Table S1 by concatenation prior to execution of BRAKER2.

Table S10: F1-scores of miniprot, filtered miniprot alignments used as training genes for AUGUSTUS, AUGUSTUS *ab initio* predictions trained on filtered miniprot alignments, and of the final output of GALBA, which are AUGUSTUS predictions with hints from miniprothint scored miniprot alignments. Note that AUGUSTUS *ab initio* is not an intermediate output of GALBA, these results were only computed for the curious reader. GALBA supports producing this additional output with the command line flag `--AUGUSTUS_ab_initio`.

| | *Arabidopsis thaliana* | | | *Bombus terrestris* | | | *Caenorhabditis elegans* | | | *Danio rerio* | | | *Drosophila melanogaster* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon |
| miniprot | 32.61 | 33.01 | 66.56 | 27.86 | 33.43 | 75.82 | 9.69 | 9.13 | 46.37 | 9.00 | 9.32 | 54.32 | 19.23 | 21.03 | 55.85 |
| Training genes | 66.45 | 53.66 | 81.81 | 45.24 | 34.87 | 79.51 | 15.11 | 12.27 | 53.43 | 27.08 | 20.61 | 75.45 | 37.81 | 28.90 | 66.14 |
| AUGUSTUS *ab initio* | 56.51 | 45.31 | 79.58 | 30.45 | 24.01 | 73.95 | 45.24 | 36.89 | 79.89 | 18.73 | 15.10 | 69.99 | 57.03 | 43.63 | 74.82 |
| GALBA | 75.32 | 60.09 | 84.82 | 53.89 | 45.19 | 82.82 | 53.51 | 42.28 | 80.99 | 40.16 | 30.07 | 77.53 | 71.07 | 55.05 | 82.74 |

| | *Medicago truncatula* | | | *Parasteatoda tepidariorum* | | | *Populus trichocarpa* | | | *Rhodnius prolixus* | | | *Tetraodon nigroviridis* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon |
| miniprot | 22.82 | 22.82 | 61.80 | 1.25 | 1.30 | 20.78 | 34.38 | 34.36 | 66.16 | 2.96 | 2.96 | 39.50 | 1.98 | 1.95 | 39.04 |
| Training genes | 29.71 | 29.71 | 69.59 | 4.22 | 4.03 | 47.04 | 36.63 | 32.05 | 71.80 | 4.58 | 4.58 | 46.77 | 6.58 | 6.06 | 55.51 |
| AUGUSTUS *ab initio* | 31.62 | 31.62 | 68.60 | 15.54 | 14.28 | 64.30 | 37.79 | 33.34 | 72.77 | 9.69 | 9.68 | 52.97 | 5.26 | 4.85 | 54.55 |
| 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 | 9.52 | 7.70 | 58.57 |

| | *Gallus gallus* | | | *Mus musculus* | | | *Solanum lycopersicum* | | | *Xenopus tropicalis* | | | **Average** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon |
| miniprot | 23.55 | 28.27 | 77.56 | 17.33 | 19.05 | 68.50 | 28.01 | 28.01 | 67.15 | 12.49 | 14.75 | 62.07 | 17.37 | 18.53 | 57.25 |
| Training genes | 40.65 | 25.24 | 83.09 | 50.61 | 27.97 | 80.89 | 37.71 | 37.71 | 74.38 | 41.04 | 30.63 | 81.99 | 31.67 | 24.88 | 69.10 |
| AUGUSTUS *ab initio* | 17.65 | 11.67 | 72.81 | 14.00 | 9.05 | 59.92 | 24.88 | 24.88 | 65.52 | 24.65 | 19.12 | 76.64 | 27.79 | 23.10 | 69.02 |
| 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 | 42.93 | 35.23 | 72.58 |

Table S11: F1-scores of GeneMark-ES, GeneMark-EP, and BRAKER2. Genes predicted by GeneMark-ES and GeneMark-EP are here intermediate products of BRAKER2 executed with protein sets listed in Table S1.

| | *Arabidopsis thaliana* | | | *Bombus terrestris* | | | *Caenorhabditis elegans* | | | *Danio rerio* | | | *Drosophila melanogaster* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon |
| GeneMark-ES | 55.54 | 44.72 | 78.60 | 29.32 | 23.34 | 71.11 | 47.82 | 39.61 | 82.22 | 5.43 | 4.64 | 46.88 | 50.88 | 39.20 | 71.21 |
| GeneMark-EP | 72.98 | 58.96 | 83.73 | 39.88 | 32.80 | 73.55 | 57.25 | 47.52 | 84.68 | 16.91 | 14.35 | 59.76 | 62.67 | 48.52 | 77.73 |
| 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 | 74.19 | 57.18 | 82.95 |

| | *Medicago truncatula* | | | *Parasteatoda tepidariorum* | | | *Populus trichocarpa* | | | *Rhodnius prolixus* | | | *Tetraodon nigroviridis* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon |
| GeneMark-ES | 26.89 | 26.89 | 64.60 | 10.34 | 9.62 | 50.51 | 34.83 | 30.59 | 70.30 | 9.38 | 9.38 | 49.49 | 1.87 | 1.76 | 40.28 |
| GeneMark-EP | 26.89 | 26.89 | 64.60 | 14.73 | 13.69 | 56.75 | 58.80 | 51.79 | 80.15 | 10.74 | 10.74 | 50.73 | 7.41 | 6.91 | 55.18 |
| 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 | 9.80 | 8.34 | 58.57 |

| | *Gallus gallus* | | | *Mus musculus* | | | *Solanum lycopersicum* | | | *Xenopus tropicalis* | | | **Average** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon | Gene | Transcript | Exon |
| GeneMark-ES | 0.11 | 0.07 | 0.26 | 2.21 | 1.37 | 20.55 | 21.17 | 21.17 | 60.33 | 9.47 | 7.70 | 60.63 | 21.80 | 18.57 | 54.78 |
| GeneMark-EP | 13.99 | 8.97 | 38.81 | 19.42 | 12.08 | 54.81 | 36.30 | 36.30 | 69.68 | 21.26 | 17.23 | 69.70 | 33.53 | 28.36 | 66.10 |
| 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 | 42.05 | 35.42 | 70.41 |

Table S12: BUSCO scores of training genes (best miniprot alignment per locus) and the AUGUSTUS predictions with hints, which are the final product of GALBA. The following BUSCO data sets were used per species: *A. thaliana* - brassicales_odb10; *B. terrestris* - hymenoptera_odb10; *C. elegans* - nematoda_odb10; *D. rerio* - actinopterygii_odb10; *D. melanogaster* - diptera_odb10; *M. truncatula* - embryophyta_odb10; *P. tepidariorum* - arachnida_odb10; *P. trichocarpa* - eudicots_odb10; *R. prolixus* - insecta_odb10; *T. nigroviridis* - actinopterygii_odb10; *G. gallus* - aves_odb10; *M. musculus* - euarchontoglires_odb10; *S. lycopersicum* - solanales_odb10; *T. nigroviridis* - tetrapoda_odb10.

| Species | Gene Set | Complete | Fragmented | Missing |
|---|---|---:|---:|---:|
| *A. thaliana* | training genes | 70.4 | 3.5 | 26.1 |
| | AUGUSTUS with hints | 99.4 | 0.4 | 0.2 |
| *B. terrestris* | training genes | 55.7 | 11.6 | 32.7 |
| | AUGUSTUS with hints | 95.4 | 2.5 | 2.1 |
| *C. elegans* | training genes | 62.0 | 5.5 | 32.5 |
| | AUGUSTUS with hints | 98.9 | 0.7 | 0.4 |
| *D. rerio* | training genes | 56.1 | 7.4 | 36.5 |
| | AUGUSTUS with hints | 95.7 | 1.5 | 2.8 |
| *D. melanogaster* | training genes | 74.6 | 8.4 | 17 |
| | AUGUSTUS with hints | 98.3 | 0.5 | 1.2 |
| *G. gallus* | training genes | 54.5 | 10.3 | 35.2 |
| | AUGUSTUS with hints | 97.1 | 0.8 | 2.1 |
| *M. truncatula* | training genes | 64.9 | 6.7 | 28.4 |
| | AUGUSTUS with hints | 98.7 | 0.7 | 0.6 |
| *M. musculus* | training genes | 55.1 | 7.1 | 37.8 |
| | AUGUSTUS with hints | 97.8 | 0.8 | 1.4 |
| *P. tepidariorum* | training genes | 68.7 | 5.0 | 26.3 |
| | AUGUSTUS with hints | 94.7 | 2.5 | 2.8 |
| *P. trichocarpa* | training genes | 62.5 | 7.8 | 29.7 |
| | AUGUSTUS with hints | 98.4 | 0.3 | 1.3 |
| *R. prolixus* | training genes | 64.8 | 16.4 | 18.8 |
| | AUGUSTUS with hints | 95.4 | 2.3 | 2.3 |
| *S. lycopersicum* | training genes | 67.0 | 2.7 | 30.3 |
| | AUGUSTUS with hints | 98.1 | 0.7 | 1.2 |
| *T. nigroviridis* | training genes | 42.7 | 9.4 | 47.9 |
| | AUGUSTUS with hints | 84.7 | 6.7 | 8.6 |
| *X. tropicalis* | training genes | 55.4 | 8.4 | 36.2 |
| | AUGUSTUS with hints | 95.9 | 0.8 | 3.3 |

# S3 Supplementary Methods

## S3.1 Assembly Quality Estimation

We used seqstats from https://github.com/clwgg/seqstats to compute genome sizes, (scaffold) N50, and the total number of sequences.

## S3.2 Annotation Parameter Computation

In order to count genes and alternative transcripts thereof, we renamed the genes and transcripts in reference annotations with the script rename_gtf.py from https://github.com/Gaius-Augustus/TSEBRA as follows:

```
rename_gtf.py --gtf annot.gtf --out annot_tsebra.out
```

Subsequently, we extracted the last gene id as number of genes, and computed the number of transcripts:

```
cat annot_tsebra.gtf | perl -ne ’ \
     if(m/transcript_id \"([^"]+)\"/){print $1."\n";}’| sort -u | wc -l
```

The ratio of mono-exonic to multi-exonic genes was computed with analyze_exons.py from https://github.com/Gaius-Augustus/GALBA:

```
analyze_exons.py -f file.gtf
```

In case of RNA-Seq supported ’reliable’ genes, the number was computed with complete_supported_subset_table.sh from https://github.com/gatech-genemark/BRAKER2-exp:

```
complete_supported_subset_table.sh prediction.gtf annot.gtf completeTranscripts.gtf \
    pseudo.gff3 varus.gff
```

## S3.3 Running FunAnnotate

FunAnnotate was executed from a singularity container as follows:

```
# only once, to get the singularity container
singularity pull docker://nextgenusfs/funannotate

export GENEMARK_PATH=/path/to/GeneMark-ES-ET-EP_v4.71_lic/gmes_funannotate

species="name of species"
buscoSeedSpecies="name of seed species"
buscodb="name of busco db"
genomepath="/path/to/genome.fasta.masked"
protpath="/path/to/proteins.fa"

# calculateGenomeSizeFromFasta.pl adds up the length of all sequences in a fasta
genomeSize=$(perl ~/calculateGenomeSizeFromFasta.pl $genomepath)
maxIntronLen_f=$(echo "3.6 * sqrt($genomeSize)" | bc -l)
maxIntronLen=$(printf "%.0f" "$maxIntronLen_f")

mkdir -p fun tmp
singularity run funannotate_latest.sif funannotate predict \
    --input $genomepath --out fun --species $species \
    --busco_seed_species $buscoSeedSpecies --busco_db $buscodb \
    --organism other --protein_evidence $protpath \
    --max_intronlen $maxIntronLen --cpus 72 --tmpdir tmp --no-progress \
    --repeats2evm
```

For accuracy evaluation, the gff3 output of FunAnnotate was converted from gff3 to gtf format using `gff3_to_gtf.pl` from GeneMark-ET, and with `compute_accuracies.sh` from BRAKER:

```
gff3_to_gtf.pl funannotate.gff3 funannotate.gtf
compute_accuracies.sh annot.gtf pseudo.gff3 funannotate.gtf gene trans cds
```

FunAnnotate sometimes modifies sequence names in the output, automatically. We had to revert these sequence name changes to match the reference annotation. This was in particular the case for *Medicago truncatula*:

```
cat funannotate.gtf | perl -pe 's/Mrun/Mtrun/' > funannotate.f.gtf
mv funannotate.f.gtf funannotate.gtf
```

## S3.4 Running GALBA

GALBA was executed as follows:

```
galba.pl --genome=genome.fa --prot_seq=proteins.fa --threads 72
```

The number of threads varied between runs, depending on HPC node availability.

## S3.5 Running BRAKER2

BRAKER2 was executed with singularity as follows:

```
singularity exec braker3.sif braker.pl --genome=genome.fa --prot_seq=proteins.fa --threads 72
```

The number of threads varied between runs, depending on HPC node availability.

## S3.6 Running TSEBRA

TSEBRA was executed as follows:

```
tsebra.py -g braker.gtf --keep_gtf galba.gtf \
    -e braker_hintsfile.gff,galba_hintsfile.gff -c default.cfg -o tsebra.gtf
```

## S3.7 Data Preparation for Estimating the Number of Genomes at NCBI that May Benefit from GALBA

We downloaded the list of all NCBI-available eukaryotic genomes on 27 July$^{th}$ 2023, excluding those that are flagged as atypical. This results in a list of 32,512 genomes.

To overlay this information with RNA-Seq availability in SRA, we extracted the first two words of the Organism name from that table to enlarge the chance to find data for this type of species in SRA. (There are a large number of entries in the genome data set that are strains or varieties or just highly similar species. SRA often contains only RNA-Seq for one or few of these, but those data may potentially be used to annotate all strains/varieties).

```
cut -f 3 eukaryotes_no_atypical.tsv  | perl -pe 's/(\S+\s+\S+).*/$1/;' | \
   sort -u > genome_species.lst
```

This resulted in 11,947 unique species name stems that can be used to query SRA for RNA-Seq data. Since the download of a full list of all eukaryotic RNA-Seq SRA accessions turned out to be difficult (the download is routinely truncated at 10,428 entries, but there are millions of runs), we modified the VARUS Perl script RunListRetriever.pl to download the number of available RNA-Seq data sets for each species. This returned 5,468 species for which a minimum of one RNA-Seq data set is available. The modified script is available at `https://github.com/Gaius-Augustus/VARUS/blob/galba/RunListRetriever/FindData.pl` and was executed as follows:

```
FindData.pl --file genome_species.lst > genome_rnaseq.lst
```

We manually confirmed the lack of RNA-Seq data for ten randomly selected species from the output to gain confidence that our approach is reliable.

We merged these results with the original genome list running a python script with the following content:

```
import pandas as pd

# Read the 'rna' file and create a dictionary with V1 as keys and V2 as values
rna_dict = {}
with open('genome_rnaseq.lst', 'r') as rna_file:
for line in rna_file:
v1, v2 = line.strip().split('\t')
rna_dict[v1] = v2

# Read the 'd' file into a pandas DataFrame
d_df = pd.read_csv('eukaryotes_no_atypical.tsv', sep='\t')

# Update the DataFrame based on the dictionary lookup
d_df['Modified_Organism_Name'] = d_df['Organism Name'].apply(lambda x: ' '.join(x.split()[:2]))
d_df['SRA entries'] = d_df['Modified_Organism_Name'].map(rna_dict)
d_df['RNA-Seq exists'] = d_df['SRA entries'].apply(lambda x: 0 if int(x) == 0 else 1)

# Write the updated DataFrame to a new file or update the original file
d_df.to_csv('eukaryotes_no_atypical_rnaseq.tsv', sep='\t', index=False)
```

The resulting file was used to count the number of genomes at NCBI that do or do not have RNA-Seq evidence in SRA.

It is important to note that this approach does not guarantee 100% correct results due to the implementation for automatically querying SRA that we chose with truncated species name stems, and due to miss-labeling of data in SRA (e.g. transcriptome libraries are sometimes labeled as genomic, etc.). It is an approximation that is likely close to reality.

# S4 Supplementary Results

## S4.1 Accuracy Improvements during GALBA Development

When we started with the GALBA development, we simply ran miniprot, used the alignments as training genes for AUGUSTUS (without any processing), and then predicted genes with AUGUSTUS using the alignment evidence. We call this the baseline version of GALBA (see Figure S7). In that early version, the selection of training genes depended on an arbitrary order of similar genes in a DIAMOND [1] output (DIAMOND is used by both BRAKER and GALBA to remove bias resulting from redundancy in training genes). We added a step that selects the highest-scoring alignment per locus as the initial training genes. This improved the gene F1 accuracy by ∼2 percentage points (assessed on *D. melanogaster* with reference proteomes of five other *Drosophila* species).

We developed miniprothint, an alignment scorer that improves the quality of protein alignment evidence used for gene prediction. It scores alignments based on three criteria: entire exon alignment, intron border alignment, and intron mapping coverage. The evidence is separated into high- and low-confidence categories, with high-confidence evidence used for training gene selection and enforced during gene prediction. The integration of miniprothint led to a further increase in gene F1 by ∼5 percentage points. In Figure 2 (main manuscript), we demonstrate the effect of using IBA and IMC to select high-confidence evidence from miniprot alignments. In Supplementary Table S2, we also report the accuracy of intron prediction with a large reference proteome of remote proteins as input.

Using a protein to genome alignment to generate training genes for a gene finder typically identifies only a subset of existing gene structures in the genome. To generate putative training genes, AUGUSTUS excises

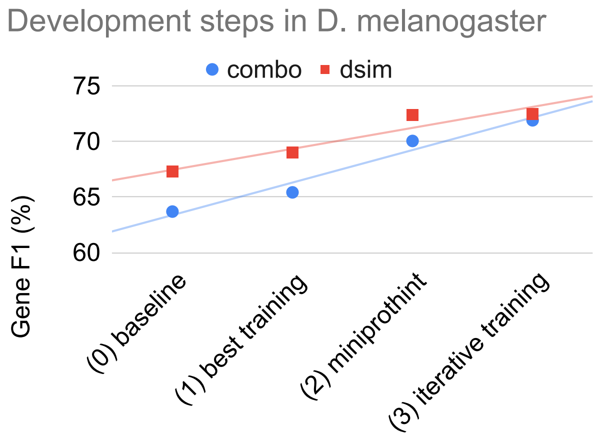

Figure S7: Gene prediction F1-scores of GALBA across development steps using two different reference proteomes: dsim = *D. simulans*, combo = *D. ananassae*, *D. grimshawi*, *D. pseudoobscura*, *D. virilis*, and *D. willistoni*.

training genes (on the basis of e.g. protein alignments) with flanking regions, which may contain parts of neighboring genes and lead to sub-optimal training results. We implemented a training approach in GALBA, which includes filtering predicted genes with 100% evidence support and training with filtered flanking regions to improve the accuracy of gene prediction, referred to as iterative training. This provided additional ∼2 percentage points accuracy increase on the gene F1 level.

The observed effects can also be measured on a single species reference proteome (with slightly different absolute numbers), as exemplarily shown by using the proteins of the very close relative *D. simulans*, only (see Figure S7).

# References

[1] Buchfink, B., Xie, C., Huson, D.H.: Fast and sensitive protein alignment using DIAMOND. Nature Methods **12**(1), 59–60 (2015)