FLAG: Find, Label, Annotate Genomes, a fully automated tool for genome gene structural and functional annotation of highly fragmented non-model species

William Troy[1], Joana Damas[1], Alexander J. Titus[2], and Brandi L. Cantarel[1]

Abstract:
Recent advances in long-read sequencing technologies and the efforts of projects aimed at increasing the universe sequenced reference genomes, such as the Earth Biogenome project have led to a growth in the number of whole genome sequencing projects for non-model organisms. Still, the vast majority (6,880) of the 28,727 unique publicly available eukaryotic genomes lack gene structure annotations. Annotating genes in a genome of interest is a multistep process, and while there are many tools available for each step of the annotation, there are few end-to-end workflows that (i) run on multiple computing environments, (ii) automatically run without initial data training and (iii) perform accurately with fragmented genomes. Here we present "Find, Label, Annotate Genomes" (FLAG), a fully automated genome annotation workflow that does not require species-specific extrinsic evidence to create highly complete and accurate gene annotations of even highly fragmented genomes. Using FLAG, various eukaryotic genome assemblies, including 1 plant, 5 living animals, and 1 extinct animal were annotated. In all living species, FLAG annotations provided an average of 18% increase in complete BUSCO scores when compared to BRAKER2. Furthermore, FLAG has more accurately predicted the number of protein-coding genes, with an average error rate 15x lower than BRAKER2.

Introduction:
Because of the highly repetitive and complex nature of eukaryotic genomes, accurately finding genes is an active area of research. The last five years has seen great advances in sequencing technologies that led to the availability of more accurate and complete genome assemblies. Eukaryotic genome annotation can involve gathering transcript sequencing data for the species of interest, comparison of genes from closely related species, and iterating genome gene structure annotation following training gene models based on "high confidence" gene models (1–6). End-to-end workflows such as BRAKER2 (3, 6) and MAKER (4, 7) can take a diversity of supporting evidence to train the gene-finding tools such as AUGUSTUS or GENEMARK and use those models to produce gene predictions. BRAKER2 is the genome annotation workflow used by Ensembl and has been shown to be more accurate than previously published eukaryotic annotation workflows (3). These methods can require large computational resources (> 24 h training times), tend to be less accurate for highly fragmented genomes, and require some computational expertise to install and run. While the NCBI does annotate genome assemblies as a part of its public service, it only has the resources to annotate less than 200 genomes per year, with ⅓ of those consisting of reannotations. Furthermore, it is not possible in all cases to generate transcriptome data for the various tissues of all organisms for which we can collect

genome data, especially for rare, endangered, or extinct species. Genome annotation quality is highly dependent on genome assembly accuracy and completeness (3, 8). Accuracy is decreased when (i) sequences are missing; (ii) assemblies are highly fragmented, which can fragment genes on multiple contigs, and (iii) contigs are misassembled. Typical crustacean and tree genomes that are of low quality or are highly gapped have over 100k extra protein-coding genes when compared to complete annotations of related species (9–11).

We developed "Find, Label, Annotate Genomes" (FLAG) as an end-to-end automated workflow to annotate eukaryotic genomes. While FLAG can accept species-specific extrinsic evidence and homology-based training data, it can create highly complete and accurate annotations when this data is unavailable. Unlike other eukaryotic annotation workflows, where the user must select a gene prediction tool, such as Gnomon, Augustus, GeneMark-ES, or SNAP (3, 4, 6, 12), FLAG allows the user to combine results from homology-based alignments and multiple gene finding methods including Liftoff(13), Augustus (1), and Helixer(8). We combine the results of all gene predictors using a novel combine and filter algorithm. Genes are functionally annotated and labeled with a gene symbol when possible using Entap (14). The final output of FLAG is (i) a filtered and labeled gene annotation file, in GTF format, and (ii) annotation statistics generated using AGAT (15) and BUSCO (16, 17). BUSCO scores can generally be extrapolated to indicate gene prediction accuracy, with a higher BUSCO score indicating a higher number of highly conserved genes within a specific lineage being annotated. When an annotation is complete, the BUSCO scores (complete, fragmented, and missing percentages) for gene annotations should resemble those scores for the assembly.

We annotated six eukaryotic genomes using FLAG, which consisted of 1 tree, 2 insects, 1 fish, 1 bird, and 1 mammal. We compared FLAG and BRAKER2 annotations. Our comparison used similar metrics as GenomeQC (18), including BUSCO scores, total number of protein-coding genes, single-exon protein-coding genes, and mean gene length. FLAG can be downloaded and installed locally or run with a transcript, protein, and related species annotation on the Form Bio platform using the default run parameters (Supplementary Figure 1).

Materials and Methods:

Input Data Types

Genome annotation requires two main types of input, the target genome sequence and external evidence. For external evidence, FLAG accepts the assembly and gene annotations from closely related organisms, RNASeq or transcript data from the species of interest or closely related species and protein sequences. Gene annotation should be in gene transfer format (GTF) or GFF3 format, which can be downloaded from publicly available genome annotation sources such as Ensembl, Gencode, the EBI or the NCBI. Genome assemblies and annotations from closely related species should be of good quality and completeness, and from species diverged within 50 million years of evolution or with 97% or higher genome sequence identity. While it is best to supply transcript or protein data from the organism of interest, if that data is not available, one

can input transcript or protein sequences of a closely related species. When no species-specific information is available, users can use publicly available data from RefSeq (19), UniProt (20), and UniRef50 (21). Users of the Form Bio platform will be able to select clade-specific databases such as mammalian, avian, sauropsids, insect, fish, etc.

Annotation Methods

Based on user selection and input data, FLAG can run several different tools to determine genes (i) that are homologous to some sequencing data from the target or a closely related species, (ii) share excess similarity to proteins in a protein reference database and (iii) are predicted using ab-initio gene finding methods (Figure 1). FLAG has three main sub-workflows, (i) the RNA workflow, which can create a transcript assembly and align transcripts to the target genome, (ii) a protein workflow, which identifies putative genes by homology with known protein sequences, and (iii) a gene finding workflow that uses evidence from the RNA and protein workflows to inform on ab-initio gene predictions, synthesize all evidence into gene models, and annotate genes with orthologous genes name (Figure 2).

The RNA and protein workflows produce alignments to the target genome that can be used as evidence to train gene finders and can be used as external evidence when determining gene confidence. The RNA workflow creates alignments of transcripts to the target genome assembly. If RNASeq data is provided, the workflow uses Trinity (22) to assemble raw sequencing reads into transcripts. Transcripts from the same species are then aligned to the genome using Pasa (23), which consists of pblat for its alignments (24) and Splign (25). Transcripts from closely related species or from a database are aligned to the target genome using blast (26). The protein workflow aligns either user-provided or publicly available protein sequences to the target genome using miniprot (27), genomethreader (28), exonerate (29), and/or prosplign.

The gene finding workflow produces the intron/exon structure of genes, including the prediction of alternative transcripts. This workflow has four ab-initio gene finders implemented: Liftoff (13), Augustus (1), Transdecoder (30), and Helixer. If RNA and/or protein data is provided, Augustus can be trained with genome alignments from the RNA and protein workflows (8). Otherwise, Augustus can be trained using prediction from other tools including Helixer or Liftoff. Augustus can also be run using a pre-trained model from a closely related organism. Transdecoder can also use genome alignments from the RNA and protein workflows as input data to inform on its ab-initio gene structural predictions; however, while it runs fast in our testing, its accuracy was the lowest out of all the implemented gene predictors. Helixer uses pre-trained models for vertebrates, invertebrates, land plants, and fungi. Users can choose to run one or several tools in parallel. Models from these gene finders are combined using Evidence Modeler v2.0.0 (31). The gene structures from Evidence Modeler and all gene finders are then separately filtered in our novel filtering algorithm. We start by separating the annotations into two categories: 1) highly conserved and complete using BUSCO and 2) supported by multiple

evidence types or gene predictors. Next, we filter them down to only be the most complete in each category. We also provide a further filtered annotation with gFacs (32), that filters more possibly false positives, including genes with internal stop codons.

Results:

We compared annotations generated by FLAG to public BRAKER2 annotations for 6 unrelated non-model species. In solving the problem of annotating genes in highly fragmented genome assemblies and genome assemblies to which external evidence is not available, we selected 6 unrelated non-model species with public BRAKER2 annotations. Of them, 3/6 have genome assemblies containing 32,000+ contigs each (Table 1) and were annotated purely based on proteins and transcripts from other species, regardless of if species-specific information was used in their BRAKER2 annotations. These species include *Erynnis tages* (Dingy skipper Butterfly) (33), *Iphiclides podalirius* (Scarce Swallowtail Butterfly) (34), *Juglans sigillata* (Iron Walnut Tree), *Macquaria ambigua* (Golden Perch) (35), *Zosterops japonicus* (Japanese white-eye) (36), and *Giraffa camelopardalis antiquorum* (Kordofan Giraffe) (37). Additionally, we also annotated an extinct species, *Anomalopteryx didiformis* (Little-bush Moa) (38), whose genome assembly has over 1.7 million contigs. Given its least ideal conditions, no protein or RNA data, a very fragmented genome and degraded DNA, there is no available gene annotation of the Little-bush moa for assessing FLAG accuracy.

To test the typical conditions for non-model organisms, we did not use species-specific evidence, but instead used protein and RNA alignments from sequences derived from closely related organisms. Genomes and annotations used for all liftoffs were from the comparison species; in cases where there are 2 comparison species, only 1 was selected for Liftoff. We found that using miniprot reduced computational time (430x compared to exonerate and 655x compared to genomethreader) without decreasing in accuracy when aligning *D. melanogaster* proteins obtained from the NCBI to the *D. melanogaster* genome.

When using BUSCO to align genes to an ortholog database, compared to BRAKER2 on our test annotations, FLAG, on average, has 18% more complete genes, 6.68% less fragmented genes, and 12.68% fewer missing genes. Similarly, in the case of A. didiformis, compared to Augustus, FLAG has 66.3% more complete genes, 13.8% less fragmented genes, and 52.5% fewer missing genes (Figure 3). Using AGAT to calculate the total number of genes, we found when comparing BRAKER2 and FLAG to closely related species, annotated by the NCBI, FLAG has a 15x lower gene count error rate (Table 2). These related species include *D. plexippus* (GCF_009731565.1), *C. elaphus* (GCF_910594005.1), *M. salmoides* (GCF_014851395.1), *H. rustica* (GCF_015227805.1), *J. microcarpa x J. regia* (GCF_004785595.1), *J. regia* (GCF_001411555.2), *D. novaehollandiae* (GCF_003342905.1), and *T. guttatus* (GCF_000705375.1).

In many cases, using multiple gene finding methods improved accuracy and robustness as it increased sensitivity without sacrificing specificity. For example, when we use the *D. plexippus*

genome as input evidence in Liftoff to create annotations for *I. podalirius* and *E. tages* genomes, it only predicted 5000 gene annotations of ~13k expected, since *D. plexippus* mapped quite poorly to the genome assemblies being annotated. Adding Helixer and Augustus predicted annotations for nearly all of the expected genes (Supplementary Table 1).
Using default parameters, depending on data size FLAG runs from 3 hours to 4 days for genomes that are between 800 Mb and 3.3GB. Although this is highly variable and heavily reliant on how much supporting evidence is provided and which de-novo genome annotation tool is used . Using Helixer and Liftoff for gene prediction greatly improves runtime and will produce very similar results to Augustus alone (Supplementary Figure 2)

Discussion:
Here, we present Find, Label, Annotate Genomes (FLAG), a highly parallelized and accurate genome annotation tool for non-model organisms, which can be run as a point-and-click tool using the Form Bio platform. The tool is very flexible to customization as users are able to provide as little or as much supporting evidence such as protein and RNA sequencing data from the target or closely related species. Using modern gene finding tools, such as Helixer and Liftoff, decreases computation time without a decrease in accuracy. We provide a plethora of run options and pre-curated databases for users, reducing the overall annotation times by facilitating setup and quicker computational turnaround times, and requiring less curation of the produced gene models. We show that FLAG default parameters provide excellent results for most Eukaryotic genomes, allowing for users with little to no experience to rapidly produce accurate gene annotations for their species of interest.

In the testing of FLAG we annotate various already published non-model highly fragmented genome assemblies and show dramatically better results than comparable tools, even in cases where genomes were very fragmented and with varying degrees of sequencing error. Our annotated proteins BUSCO scores closely resemble those derived from running BUSCO on the target genome assembly, while maintaining accurate total protein-coding gene counts, single exon gene counts, and mean gene lengths. In addition, FLAG also annotates tRNAs, performs a functional annotation of the annotated genes and combines structural and functional annotations to label genes in a fully automated fashion.

Software Availability Statement:
FLAG is available and open source at: https://github.com/formbio/FLAG

Conflict of interest:
The authors of this manuscript, identified as BLC, JD, and WT, are employees of Form Bio Inc and hold stocks in Colossal Biosciences Inc. Additionally, AJT, an employee of Colossal Biosciences Inc, holds stocks in Form Bio Inc.

Tables:

Table 1. Assembly statistics on the genomes used for FLAG annotation tests.

Genome	***E. tages***	***I. podalirius***	***J. sigillata***	***M. ambigua***	***Z. japonicus***	***G. camelopardalis antiquorum***	***A. didiformis***
Size (Mb)	329	430	585	661	1,005	2,963	1,190
Scaffolds	41	260	6,413	7,165	29,442	10,445	1,909
Scaffold N50	11,903,105	15,090,856	238,362	252,501	9,685,185	27,504,022	3,350,051
Scaffold L50	13	13	699	730	29	25	111
Contigs	111	446	54,794	7,352	73,843	32,009	1,734,944
Contig N50	8,006,790	5,183,580	46,167	243852	38,076	228,443	823
Contig L50	16	26	3,256	757	7,569	3,139	321,557

Table 2. Gene annotation statistics of BRAKER2 and FLAG against respective NCBI comparison species being what is deemed as expected truth, shaded in gray. For A. didiformis the closest related species was T. guttatus, however, the assembly for A. didiformis was built off of D. novaehollandiae and as such the current A. didiformis assembly should resemble a mix between the two. Statistics were calculated using AGAT v1.0.0.

Genome	**Annotation Method**	**# Protein Coding Genes**	**# Single Exon Genes**	**Mean Gene Length**
E. tages	BRAKER2	17,826	-	-
E. tages	FLAG	12,929	1,784	10,948
I. podalirius	BRAKER2	15,237	2,618	6,721
I. podalirius	FLAG	12,477	1,669	12,092
D. plexippus (*E. tages* and *I.*	NCBI	13,105	1,015	10,501

podalirius comparison)				
G. camelopardalis antiquorum	BRAKER2	59,082	25,710	9,869
G. camelopardalis antiquorum	FLAG	20,532	2,258	45,055
C. elaphus (*G. camelopardalis antiquorum* comparison)	NCBI	22,941	3,251	49,621
M. ambigua	BRAKER2	36,108	5,000	6,296
M. ambigua	FLAG	25,367	2,172	12,216
M. salmoides (*M. ambigua* comparison)	NCBI	27,179	1,088	19,634
Z. japonicus	BRAKER2	71,966	6,090	4,120
Z. japonicus	FLAG	18,012	2,807	23,102
H. rustica (*Z. japonicus* comparison)	NCBI	15,516	580	35,255
J. sigillata	BRAKER2	89,512	10,653	2,993
J. sigillata	FLAG	34,806	8,982	4,026
J. microcarpa x J. regia (*J. sigillata* comparison)	NCBI	27,962	4,238	5,586
J. regia (*J. sigillata* comparison)	NCBI	30,709	5,255	5,309
A. didiformis	Augustus (3 runs)	9,797-13,763	106-122	24,369-35,722
A. didiformis	FLAG	14,329	756	30,972
D. novaehollandiae (*A. didiformis* comparison)	NCBI	15,554	552	37,263
T. guttatus (*A. didiformis* comparison)	NCBI	15,723	820	22,119

Figure Legends:

Figure 1. Overall workflow diagram of FLAG, not including the input genome as it is an input to all steps shown here. Some steps can be run in parallel and/or threaded to maximize CPU and RAM usage.

Figure 2. Sub-workflows for the major analysis including a) the RNA workflow, b) the protein workflow, and c) the gene prediction workflow.

Figure 3. BUSCO scores of BRAKER2, Augustus, and FLAG annotations compared against their respective genome assembly BUSCO scores. Lineages used include lepidoptera for *I. podalirius* and *E. tages*, eudicots for *J. sigillata*, actinopterygii for *M. ambigua*, sauropsida for *Z. japonicus* and *A. didiformis*, and mammalia for *G. camelopardalis antiquorum*. All BUSCO scores were calculated with BUSCO v5.3.2 and ortho DB v10 (39).

References:

1. Stanke,M., Keller,O., Gunduz,I., Hayes,A., Waack,S. and Morgenstern,B. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.*, **34**, W435–W439.
2. Stanke,M. and Morgenstern,B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.*, **33**, W465–W467.
3. Brůna,T., Hoff,K.J., Lomsadze,A., Stanke,M. and Borodovsky,M. (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinforma.*, **3**, lqaa108.
4. Holt,C. and Yandell,M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, **12**, 491.
5. Ejigu,G.F. and Jung,J. (2020) Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. *Biology*, **9**, 295.
6. Hoff,K.J., Lange,S., Lomsadze,A., Borodovsky,M. and Stanke,M. (2016) BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, **32**, 767–769.
7. Cantarel,B.L., Korf,I., Robb,S.M., Parra,G., Ross,E., Moore,B., Holt,C., Alvarado,A.S. and Yandell,M. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res*, **18**, 188–196.
8. Stiehler,F., Steinborn,M., Scholz,S., Dey,D., Weber,A.P.M. and Denton,A.K. (2020) Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics*, **36**, 5291–5298.
9. Salzberg,S.L. (2019) Next-generation genome annotation: we still struggle to get it right. *Genome Biol.*, **20**, 92.
10. Trouern-Trend,A.J., Falk,T., Zaman,S., Caballero,M., Neale,D.B., Langley,C.H., Dandekar,A.M., Stevens,K.A. and Wegrzyn,J.L. (2020) Comparative genomics of six Juglans species reveals disease-associated gene family contractions. *Plant J. Cell Mol. Biol.*, **102**, 410–423.
11. Veldsman,W.P., Ma,K.Y., Hui,J.H.L., Chan,T.F., Baeza,J.A., Qin,J. and Chu,K.H. (2021) Comparative genomics of the coconut crab and other decapod crustaceans: exploring the

molecular basis of terrestrial adaptation. *BMC Genomics*, **22**, 313.
12. Tatusova,T., DiCuccio,M., Badretdin,A., Chetvernin,V., Nawrocki,E.P., Zaslavsky,L., Lomsadze,A., Pruitt,K.D., Borodovsky,M. and Ostell,J. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.*, **44**, 6614–6624.
13. Shumate,A. and Salzberg,S.L. (2021) Liftoff: accurate mapping of gene annotations. *Bioinformatics*, **37**, 1639–1643.
14. Hart,A.J., Ginzburg,S., Xu,M. (Sam), Fisher,C.R., Rahmatpour,N., Mitton,J.B., Paul,R. and Wegrzyn,J.L. (2020) EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Mol. Ecol. Resour.*, **20**, 591–604.
15. Dainat,J., Hereñú,D., Davis,E., Crouch,K., LucileSol, Agostinho,N., pascal-git, Zollman,Z. and tayyrov (2023) NBISweden/AGAT: AGAT-v1.1.0. 10.5281/zenodo.7950165.
16. Simão,F.A., Waterhouse,R.M., Ioannidis,P., Kriventseva,E.V. and Zdobnov,E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
17. Manni,M., Berkeley,M.R., Seppey,M., Simao,F.A. and Zdobnov,E.M. (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. 10.48550/arXiv.2106.11799.
18. Manchanda,N., Portwood,J.L., Woodhouse,M.R., Seetharam,A.S., Lawrence-Dill,C.J., Andorf,C.M. and Hufford,M.B. (2020) GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. *BMC Genomics*, **21**, 193.
19. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation | Nucleic Acids Research | Oxford Academic.
20. UniProt: the Universal Protein Knowledgebase in 2023 | Nucleic Acids Research | Oxford Academic.
21. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B., Wu,C.H., and the UniProt Consortium (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.
22. Haas,B.J., Papanicolaou,A., Yassour,M., Grabherr,M., Blood,P.D., Bowden,J., Couger,M.B., Eccles,D., Li,B., Lieber,M., *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, **8**, 1494–1512.
23. Haas,B.J., Delcher,A.L., Mount,S.M., Wortman,J.R., Smith,R.K., Hannick,L.I., Maiti,R., Ronning,C.M., Rusch,D.B., Town,C.D., *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, **31**, 5654–5666.
24. Wang,M. and Kong,L. (2019) pblat: a multithread blat algorithm speeding up aligning sequences to genomes. *BMC Bioinformatics*, **20**, 28.
25. Kapustin,Y., Souvorov,A., Tatusova,T. and Lipman,D. (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct*, **3**, 20.
26. McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, **32**, W20–W25.
27. Li,H. (2023) Protein-to-genome alignment with miniprot. *Bioinformatics*, **39**, btad014.
28. Gremme,G., Brendel,V., Sparks,M.E. and Kurtz,S. (2005) Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.*, **47**, 965–978.
29. Slater,G.S.C. and Birney,E. (2005) Automated generation of heuristics for biological

sequence comparison. *BMC Bioinformatics*, **6**, 31.
30. Brian Haas, Bob Zimmerman, Michael R Crusoe, Charles Plessy and Matt MacManes TransDecoder.
31. Haas,B.J., Salzberg,S.L., Zhu,W., Pertea,M., Allen,J.E., Orvis,J., White,O., Buell,C.R. and Wortman,J.R. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.*, **9**, R7.
32. Caballero,M. and Wegrzyn,J. (2019) gFACs: Gene Filtering, Analysis, and Conversion to Unify Genome Annotations Across Alignment and Gene Prediction Frameworks. *Genomics Proteomics Bioinformatics*, **17**, 305–310.
33. Darwin Tree of Life.
34. Mackintosh,A., Laetsch,D.R., Baril,T., Ebdon,S., Jay,P., Vila,R., Hayward,A. and Lohse,K. (2022) The genome sequence of the scarce swallowtail, Iphiclides podalirius. *G3 GenesGenomesGenetics*, **12**, jkac193.
35. Pavlova,A., Harrisson,K.A., Turakulov,R., Lee,Y.P., Ingram,B.A., Gilligan,D., Sunnucks,P. and Gan,H.M. (2022) Labile sex chromosomes in the Australian freshwater fish family Percichthyidae. *Mol. Ecol. Resour.*, **22**, 1639–1655.
36. Venkatraman,M., Fleischer,R.C. and Tsuchiya,M.T.N. (2021) Comparative Analysis of Annotation Pipelines Using the First Japanese White-Eye (Zosterops japonicus) Genome. *Genome Biol. Evol.*, **13**, evab063.
37. Coimbra,R.T.F., Winter,S., Kumar,V., Koepfli,K.-P., Gooley,R.M., Dobrynin,P., Fennessy,J. and Janke,A. (2021) Whole-genome analysis of giraffe supports four distinct species. *Curr. Biol.*, **31**, 2929-2938.e5.
38. Cloutier,A., Sackton,T., Grayson,P., Edwards,S. and Baker,A. (2019) First nuclear genome assembly of an extinct moa species, the little bush moa (Anomalopteryx didiformis).
39. Kriventseva,E.V., Kuznetsov,D., Tegenfeldt,F., Manni,M., Dias,R., Simão,F.A. and Zdobnov,E.M. (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.*, **47**, D807–D811.

Figures

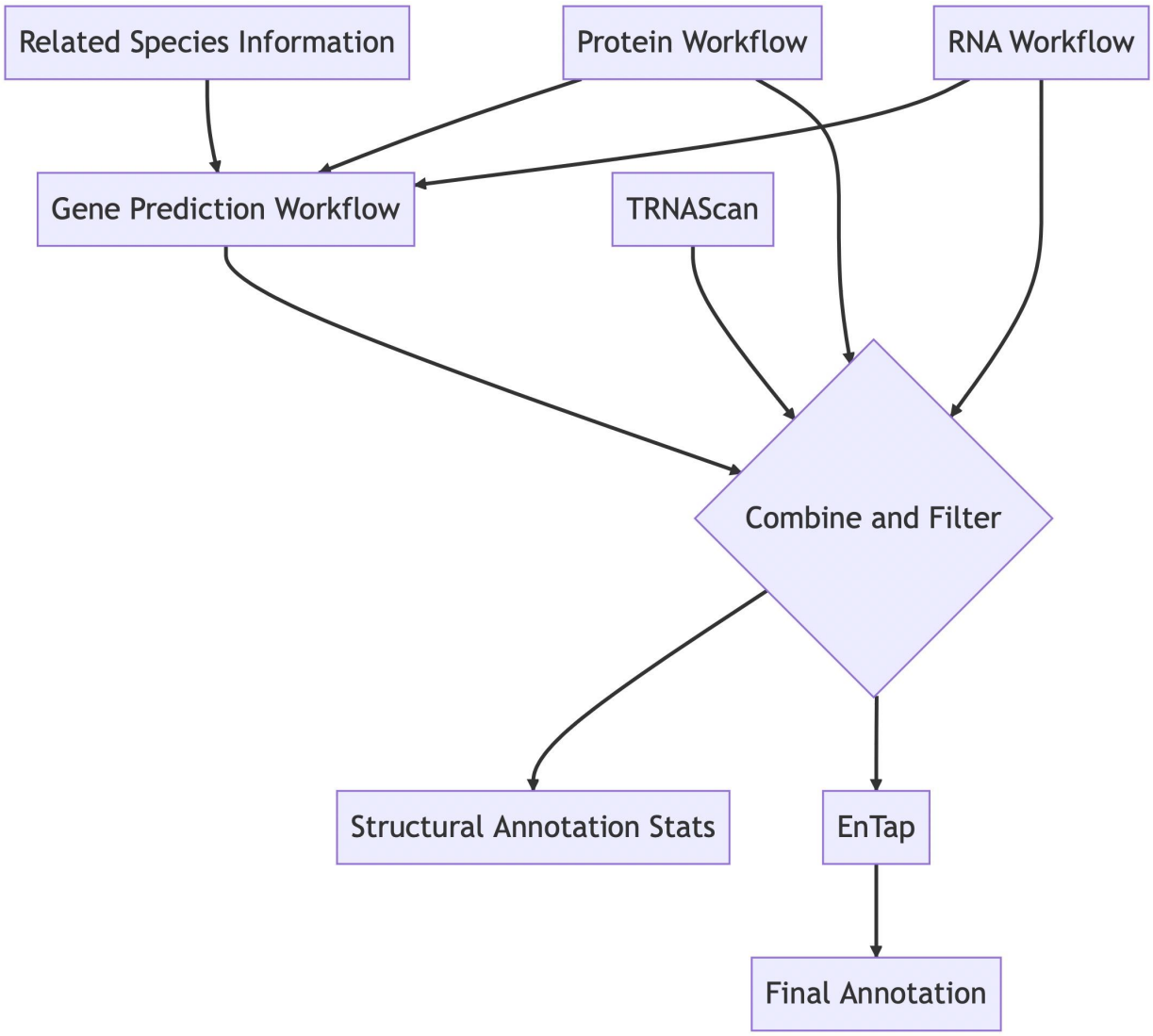

Figure 1. Overall workflow diagram of FLAG, not including the input genome as it is an input to all steps shown here. Some steps can be run in parallel and/or threaded to maximize CPU and RAM usage.

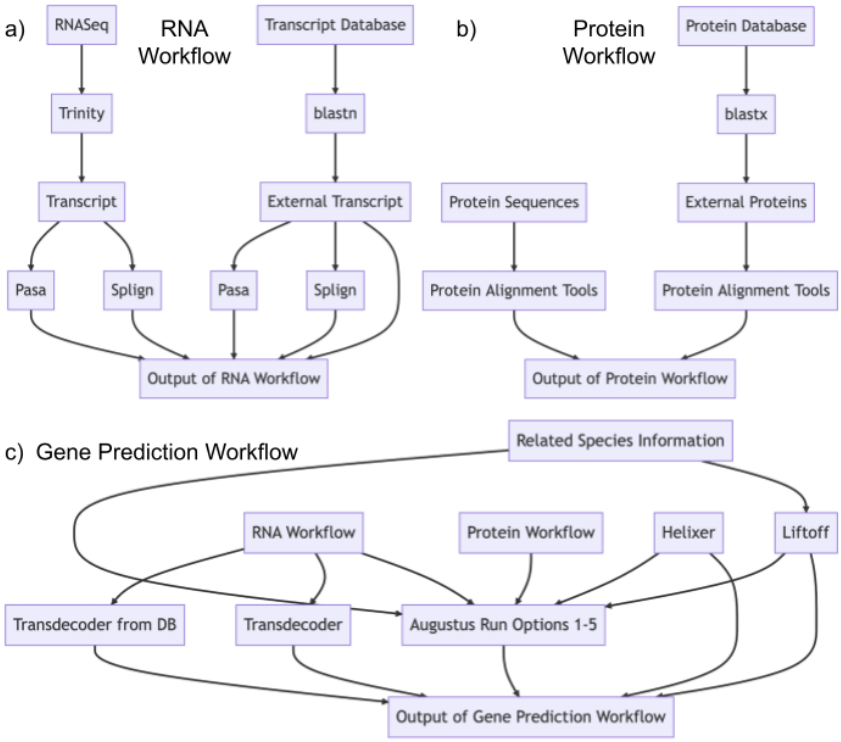

Figure 2. Sub-workflows for the major analysis including a) the RNA workflow, b) the protein workflow, and c) the gene prediction workflow.

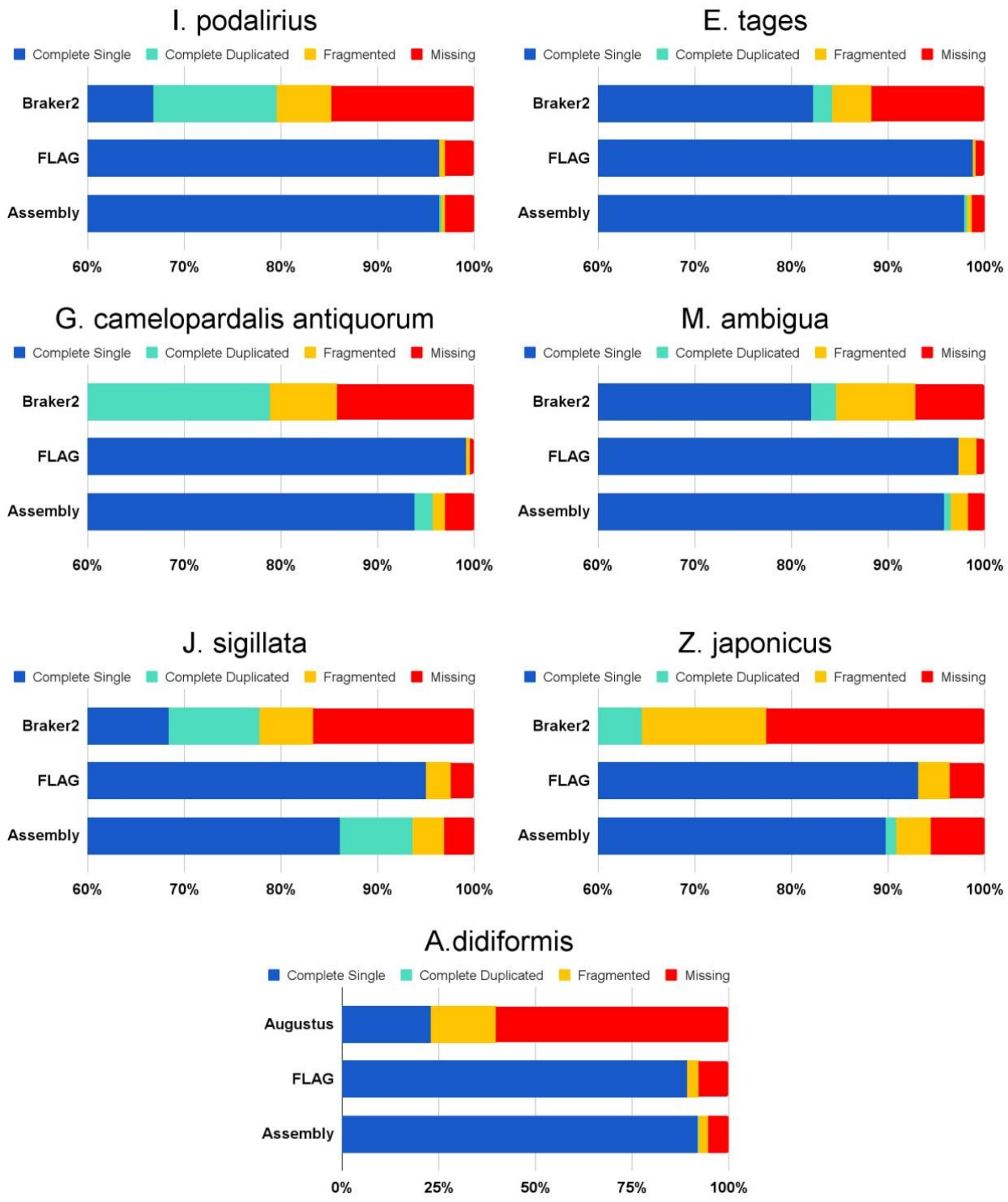

Figure 3. BUSCO scores of BRAKER2, Augustus, and FLAG annotations compared against their respective genome assembly BUSCO scores. Lineages used include lepidoptera for *I. podalirius* and *E. tages*, eudicots for *J. sigillata*, actinopterygii for *M. ambigua*, sauropsida for *Z. japonicus* and *A. didiformis*, and mammalia for *G. camelopardalis antiquorum*. All BUSCO scores were calculated with BUSCO v5.3.2 and ortho DB v10 (39).