Building better genome annotations across the tree of life

Running title: Genome annotation across the tree of life

Adam H. Freedman[1*], Timothy B. Sackton[1]

[1] Faculty of Arts and Sciences, Informatics Group, Harvard University, 52 Oxford Street, Cambridge, MA 02138

* Corresponding author: adamfreedman@fas.harvard.edu

ABSTRACT

Recent technological advances in long read DNA sequencing accompanied by dramatic reduction in costs have made the production of genome assemblies financially achievable and computationally feasible, such that genome assembly no longer represents the major hurdle to evolutionary analysis for most non-model organisms. Now, the more difficult challenge is to properly annotate a draft genome assembly once it has been constructed. The primary challenge to annotations is how to select from the myriad gene prediction tools that are currently available, determine what kinds of data are necessary to generate high quality annotations, and evaluate the quality of the annotation. To determine which methods perform the best and determine whether the inclusion of RNA-seq data is necessary to obtain a high-quality annotation, we generated annotations with 10 different methods for 21 different species spanning vertebrates, plants, and insects. We found that the RNA-seq assembler Stringtie and the annotation transfer method TOGA were consistently top performers across a variety of metrics including BUSCO recovery, CDS length, and false positive rate, with the exception that TOGA performed less in plants with larger genomes. RNA-seq alignment rate was best with RNA-seq assemblers. HMM-based methods such as BRAKER, MAKER, and multi-genome AUGUSTUS mostly underperformed relative to Stringtie and TOGA. In general, inclusion of RNA-seq data will lead to substantial improvements to genome annotations, and there may be cases where complementarity among methods may motivate combining annotations from multiple sources.

30 **INTRODUCTION**

31 The reporting in 2001 of the first draft of the human genome sequence (Lander et al. 2001;
32 Venter et al. 2001) ushered in a new era of genome-scale analysis, with a concomitant, rapid
33 increase in the development of bioinformatics tools and resources to interrogate genomes for
34 evolutionary patterns and features of biomedical interest. But even as genomes became
35 available for other model organisms such as mouse (*Mus musculus*) (Mouse Genome
36 Sequencing Consortium et al. 2002) and rhesus macaque (*Macaca mulatta*) (Rhesus Macaque
37 Genome Sequencing and Analysis Consortium et al. 2007)—and had been previously published
38 for smaller genomes such as *Drosophila melanogaster* (Adams et al. 2000)—the prohibitive cost
39 of generating genome assemblies meant that research groups working on non-model organisms
40 continued to operate in the genomic dark. Absent genome assemblies and annotations, such
41 groups were forced to embark on time-consuming efforts to sequence small sets of conserved
42 genes with Sanger sequencing using primers designed with other genomes, target anonymous
43 loci such as AFLPs or *de novo* assembled RAD-seq reads. These methods imposed an
44 analytical glass ceiling on the types of inferences that could be made and prevented the framing
45 of research findings in a genomic context. While the advent of RNA-seq inched non-model
46 organism research closer to understanding patterns at functional loci, *de novo* assembled
47 transcriptomes presented novel analytical challenges and potential distortions of evolutionary
48 patterns relative to what would be obtained with access to a genome assembly (Freedman et al.
49 2021).

50 Recent technological advances in long read DNA sequencing such as Pacific
51 Biosciences HiFi (Wenger et al. 2019) and Oxford Nanopore (Jain et al. 2018), accompanied by
52 dramatic reduction in costs have made the production of genome assemblies financially
53 achievable and computationally feasible, such that genome assembly no longer represents the
54 major hurdle to evolutionary analysis for most non-model organisms. Now, the more difficult

55 challenge is to properly annotate a draft genome assembly once it has been constructed. The
56 challenge is not so much the difficulty or computational resources required to run any one
57 genome annotation tool, but how to a) select from the myriad gene prediction tools that are
58 currently available, b) determine what kinds of data are necessary to generate high quality
59 annotations, and c) evaluate the quality of the predicted transcript and gene models.

60 Currently available genome annotation tools approach the genome annotation problem
61 in very different ways. Early computational tools for annotation used Hidden Markov Models
62 (HMMs) to scan genomes for sequences representing protein-coding intervals, with
63 AUGUSTUS (Stanke and Waack 2003) being the most widely used example. Recent
64 implementations of this approach, such as BRAKER1 (Hoff et al. 2016) and BRAKER2 (Brůna
65 et al. 2021) wrap optimized implementations of AUGUSTUS, using protein and RNA-seq
66 evidence, respectively—and with the latest release, both—to train HMMs. Transcript
67 assemblers such as Stringtie (Pertea et al. 2015) implement a graph-based framework to
68 directly assemble transcripts from splice-aware alignments of RNA-seq reads to the genome.
69 Tools such as Comparative AUGUSTUS (CGP) (Nachtweide and Stanke 2019) and TOGA
70 (Kirilenko et al. 2023) use whole genome alignments to transfer annotation evidence between
71 genomes, with the former involving multi-way transfer of HMM-based gene predictions, and the
72 latter lifting over annotations from a high quality reference annotation in an exon-aware fashion.

73 A large part of difficulty in determining what strategy will work for “my genome” is
74 annotation methods have, for the most part, been benchmarked and optimized with genomes
75 from a very small slice of the tree of life: small genomes such as *D. melanogaster*,
76 *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, or with emphasis on vertebrates (Pertea et
77 al. 2015; Hoff et al. 2016; Cantarel et al. 2008; Shao and Kingsford 2017) —no surprise given
78 the implications for discoveries relevant to human health and disease. A related problem is that
79 genome annotation will in the not too far future and, by necessity, need to be conducted by the

group that has assembled a genome, rather than relying on Ensembl or NCBI to implement their automated pipelines. While the number of genomes annotated per year by NCBI has remained flat over the last few years, the number of published genome assemblies continues to increase at an unprecedented rate. In summary, publicly accessible generators of and repositories for annotations cannot keep up, such that the wait times between genome submission date and annotation completion will increase to the point of being impractical.

Motivated by the practical challenges facing research groups seeking to annotate new genome assemblies, here we evaluate the contents of genome annotations produced by ten different methods across a broad swath of the tree of life. These methods sample the current state of the art for different approaches to the annotation problem. Our taxonomic sampling includes two vertebrate clades (birds and mammals), two insect clades (*drosophila* spp., and heliconiine butterflies), and two plant clades (rosids and monocots), totaling 21 species. While "genome annotation" is often treated as an omnibus term that includes both the prediction of the genomic positions of genes and constituent isoforms, and the assigning of gene symbols and functions, we focus on the first of these two components. Our goals are to determine a) which methods are consistently top performers with respect to various sensitivity and specificity metrics b) the contents of individual annotations with respect to gene model fragmentation and fusion, c) whether the inclusion of RNA-seq data is essential for producing a high-quality annotation, d) whether species and taxonomic group affect annotation method performance, and e) whether there is complementarity among methods, such that integration of > 1 method might lead to detectable improvements in sensitivity.

RESULTS

We evaluated ten different methods for genome annotation. At the core of five of these—MAKER (Cantarel et al. 2008) with both protein and RNA-seq evidence; BRAKER1; BRAKER2;

105 CGP using protein evidence; and CGP with RNA-seq evidence—are HMM-based *ab initio*
106 predictions by AUGUSTUS, with MAKER and BRAKER including predictions from additional *ab*
107 *initio* tools. CGP enables evidence-based prediction within a species, as well as annotation
108 transfer across species in the in the whole-genome alignment. TOGA performs an exon-aware
109 liftover of annotations from a related genome that, ideally, has a complete, high quality genome
110 annotation. Four methods assembly transcripts directly from RNA-seq alignments to a genome
111 with one of two aligners: Stringtie with either HISAT2 (Kim et al. 2019) or STAR (Dobin et al.
112 2013) as aligner, and Scallop (Shao and Kingsford 2017) with either of these aligners. Many of
113 our performance metrics use NCBI annotations as benchmarks. Derived from multiple lines of
114 evidence, outputs of the NCBI annotation pipeline are good approximations to complete
115 annotations. Futhermore, the species we annotated are either well-established model
116 organisms or others with a long history of study and substantial genomic resources, such that
117 their NCBI annotations are of the highest quality.
118
119 **BUSCO recovery**
120 The ability of a gene prediction method to recover genes known to be conserved across a wide
121 array of taxa is a good approximation for its ability to recover at least some sub-sequence of any
122 gene, including those showing less conservation. BUSCO (Simão et al. 2015) scores –
123 measuring the proportion of BUSCO targets in a search that have matches to query
124 transcripts—varied considerably across methods and among species and broader taxonomic
125 groups. Methods built on HMMs (BRAKER and CGP) consistently produced BUSCO scores for
126 dipterans, heliconiine butterflies, and rosid plants that were higher than for other methods, or
127 tied with TOGA (Fig. 1). In contrast, RNA-seq assemblers consistently outperform HMM-based
128 methods in mammals and monocots, and one of the three bird species (Fig. 1). $BRAKER_{RNA\text{-}seq}$
129 recovered more BUSCOs than $BRAKER_{protein}$ in 16 of 20 species (Fig. 1), and in cases where

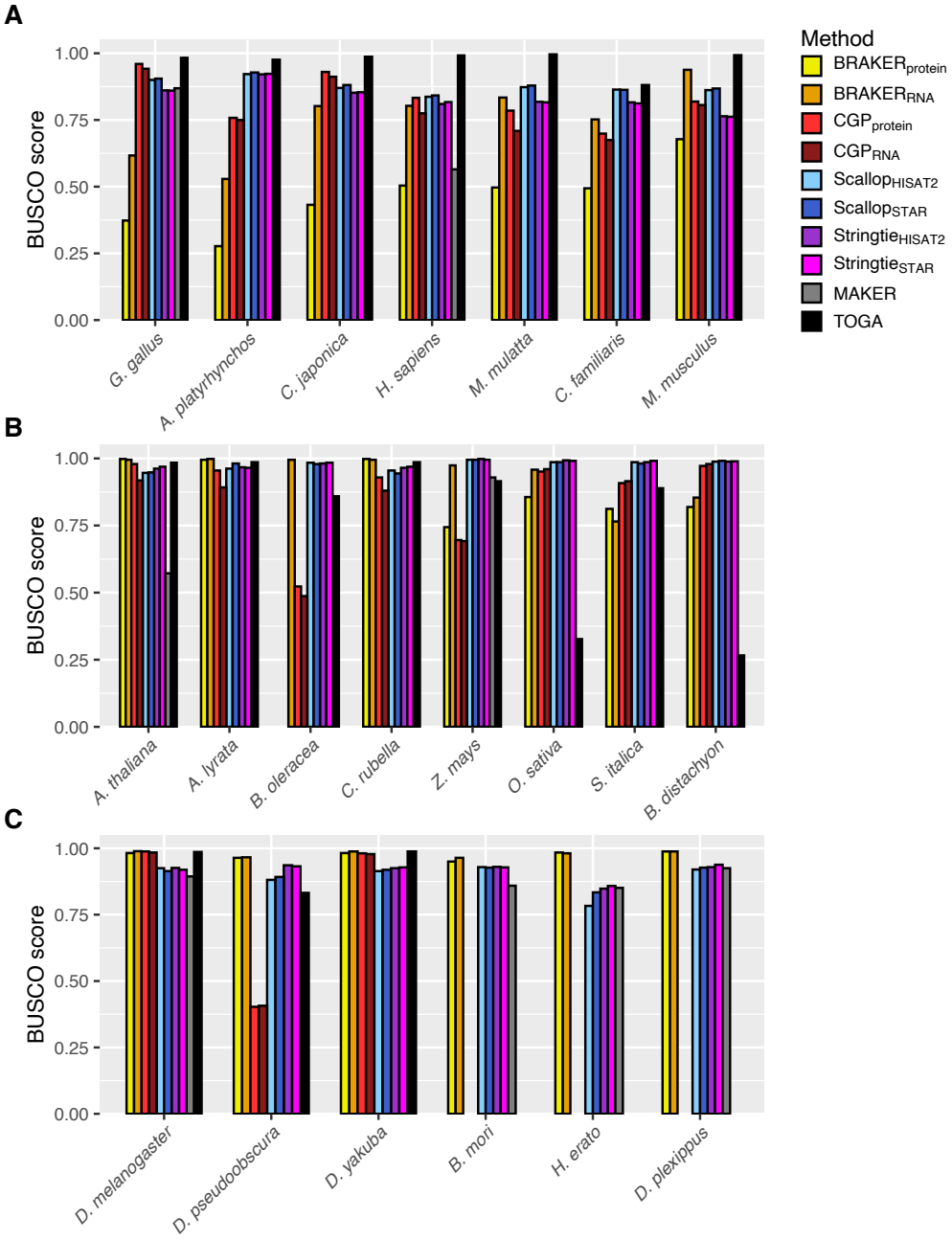


Figure 1. BUSCO scores by annotation method for (A) vertebrates, (B) plants,, and (C) insects.

130

131 BRAKER$_{protein}$ recovered more the difference in recovery is typically small. BUSCO scores were

132 very similar between CGP$_{protein}$ and CGP$_{RNA\text{-}seq}$, with CGP$_{protein}$ recovering slightly more BUSCOs

133 in 14 of 16 species (Fig. 1). CGP failed to produce more than a handful of transcript predictions
134 for heliconiines, and $BRAKER_{protein}$ consistently failed with *B. oleracea*, hence why these results
135 are not included. For *D. pseudoobscura* and *B. oleracea*, CGP BUSCO recovery was poor, with
136 BUSCO scores of approximately 50% or less. TOGA consistently produced the highest BUSCO
137 scores in birds and mammals, and for most species in other groups had scores comparable to
138 those produced by the top-performing method (Fig. 1). TOGA BUSCO scores relative to other
139 methods was lower in plants, particularly in two of four monocots (Fig. 1). Somewhat
140 surprisingly, MAKER, a putative full annotation workflow that leverages both protein and RNA-
141 seq evidence, consistently lagged in performance behind other standalone methods. These
142 patterns are not influenced by variation in the presence of BUSCOs in the underlying genome
143 assemblies, which might produce taxonomic effects (correlation between BUSCO score and
144 annotation BUSCO score/genome BUSCO score, Pearson's $\rho = 0.998$, $p = 2.2 \times 10^{-16}$).
145 While there is considerable overlap between the BUSCOs that are recovered between
146 classes of methods, there are species for which methods that use RNA-seq will recover
147 BUSCOs that methods that use protein evidence cannot (Supplemental Fig. S1). This is
148 particularly the case for species with larger genomes, such as birds, mammals, and *Z. mays*,
149 the monocot in our sample with a genome > 2Gb, and that is six to 10-fold larger than the
150 genomes of other monocots in our study. In some cases, 10-15% of BUSCOs are recovered by
151 methods leveraging RNA-seq but not by methods relying on protein evidence. However, there
152 are very few BUSCOs recovered by methods leveraging RNA-seq that TOGA does not also
153 recover; the taxonomic exception is monocots, for which TOGA can perform poorly, likely due to
154 known issues with whole genome alignment for plants.
155
156
157

158 **Annotation composition: number and length of CDS**

159 The constituent CDS predictions that underlie BUSCO recovery rates varied widely among
160 methods. Across diverse taxa, TOGA consistently produced CDS whose length distributions
161 closely approximated those generated by NCBI, and with few exceptions Stringtie did as well
162 (Supplemental Fig. S2). CGP annotations contained larger proportions of short predictions than
163 other methods, with weaker trends towards shorter CDS observed in BRAKER. Scallop length
164 distributions were often shifted towards shorter CDS to a similar degree as the HMM-based
165 method with the greatest proportion of short CDS transcripts.

166 Annotations with tendencies towards shorter and far larger numbers of CDS relative to
167 NCBI are likely indicative of fragmented transcript models. We observed this tendency for
168 several species, particularly those with larger genomes (Fig. 2; Supplemental Fig. S3). Notable
169 extreme outliers were the 7-fold and 4-fold larger CDS counts produced by CGP for macaque
170 and human, respectively (Fig. 2A). The best approximations to median CDS length and number
171 of NCBI annotations were typically produced by TOGA or a Stringtie assembly (Fig. 2;
172 Supplemental Fig. S3).

173

174 **False positives: intergenic predictions**

175 For all but *A. thaliana*, the false positive rate (FPR) at which predicted genes fell entirely within
176 intergenic regions relative to the respective NCBI annotations was lowest for TOGA; for *A.*
177 *thaliana*, Stringtie (with STAR alignments) had the lowest FPR, and that for TOGA was nearly
178 identical. In general, when TOGA did not have the lowest FPR, an RNA-seq assembler did (Fig.
179 3). For all species, FPR for the best performing method was ≤ 10% and for many species those
180 predictions occurred < 5% of the time (Fig. 3). Regardless of the species and evidence type,
181 FPR for CGP was much larger, often exceeding 50%, with predictions using RNA-seq evidence
182 being less prone to FPR than those relying on protein evidence (Fig. 3). While BRAKER FPR

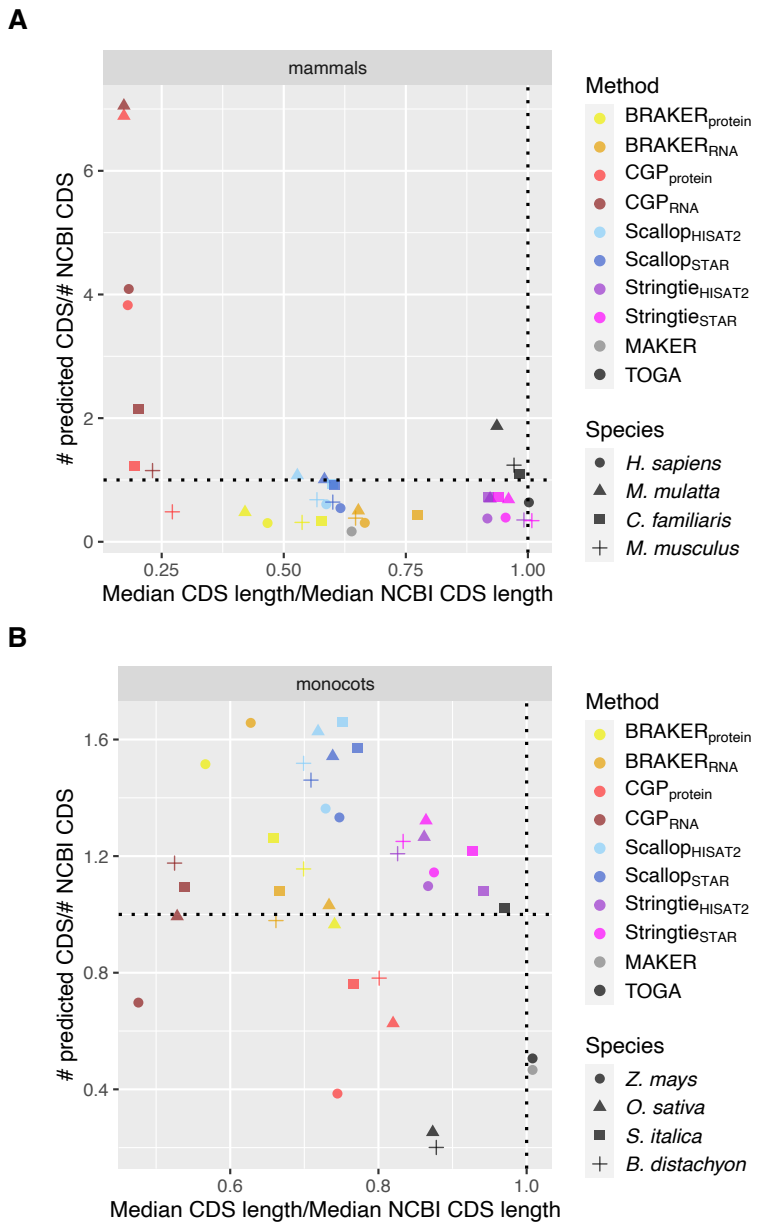


Figure 2. Joint distributions of number of predicted CDS (normalized by number of NCBI predictions) over median predicted CDS length (normalized by median NCBI CDS length) for (A) mammals, (B) monocots. Dotted lines indicated equivalence to NCBI annotation, such that methods that are closest to the intersection of those lines best approximate CDS length and number of NCBI annotations. For dipterans, birds and rosids, see Figure S3.

was typically less than that of CGP, regardless of the evidence type used, FPR was higher than the best performing RNA-seq assembler in 16 of 18 species. The disparity between the low rates for RNA-seq assemblers compared to other methods was most evident in rosids,

187 monocots, and mammals (Fig. 3). Nevertheless, the observed FPR suggest that, even for the
188 best performing methods, hundreds or even thousands of gene predictions will fall outside of the
189 genomic intervals for known real coding sequence. This raises the question of whether these
190 sequences are false negatives in the NCBI annotation, or whether they are, in fact, false
191 positives. We take a conservative approach and assume that most of these predictions are false
192 positives, asking whether there are transcript features that might be predictive of such putative
193 false positives so they can be removed. The same patterns held at the transcript level (gene vs.
194 transcript intergenic prediction rate, Pearson's $\sigma = 0.994$, $p < 2.2 \times 10^{-16}$).

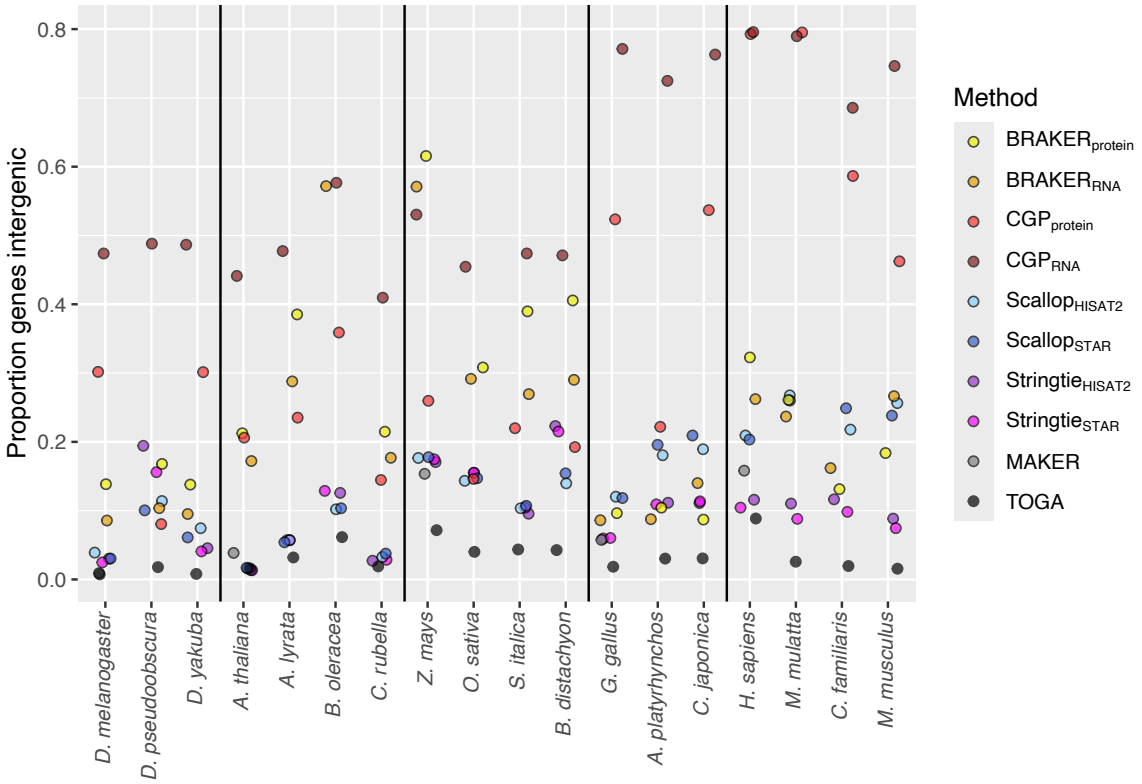


Figure 3. Proportion of predicted protein-coding genes that fall entirely outside the intervals for NCBI protein-coding genes, organized by species and method.

195

196 To better understand whether features of predicted transcripts distinguish those that at
197 least partly overlap NCBI gene intervals from those that are entirely in NCBI intergenic regions,

198 we fit random forest models to predictors summarizing sequence content, expression level, and
199 whether the associated ORF had strong evidence to a match in the reference proteins of a
200 related species. We did this for the five reference species for which NCBI annotations are
201 thought to be the most complete. Overall, out-of-bag error rates (OOB) were consistently low
202 for RNA-seq assemblers and TOGA, with OOB always being < 0.05 for TOGA, and for RNA-seq
203 assemblers only exceeding 5% in *Z. mays* (Supplemental Fig. S4). Except for CGP, the class
204 (genic vs. intergenic) error rates that comprise OOB are substantially higher for intergenic
205 predictions (Supplemental Fig. S5). This result is undoubtedly due to the fact, that, for most
206 species-method combinations, there are far fewer intergenic than genic predictions
207 (Supplemental Fig. S6), making it harder for random forest to optimally classify intergenic
208 sequences. However, random forest correctly classifies the majority of intergenic predictions as
209 intergenic for methods that predict large numbers of intergenic transcripts (Supplemental Fig.
210 S6).

211 Estimates of node purity by the Gini index—an estimate of variable importance that
212 quantifies the extent to which removing a predictor reduces the frequency of predictions
213 matching the true class on either side of a split—reveal that whether or not a transcript ORF has
214 a BLAST hit is frequently the most powerful predictor of whether or not a transcript is genic or
215 intergenic (Supplemental Fig. S7), with expression level and CDS length also frequently making
216 large contributions to the models. These results suggest that, in the absence of a truth-set of
217 known CDS intervals, some CDS features are potentially useful for setting filters to
218 discriminating true from false positive intergenic predictions. The importance of individual
219 expression metrics is likely underestimated, due to the correlation between expression metrics,
220 such that the effect of removing one is mitigated by the presence of another in any one of the
221 constituent trees in the random forest. These results collectively suggest that short, lowly

222 expressed transcripts without hits to an external protein database are enriched for intergenic
223 (and likely spurious) predictions.
224
225 **Gene fusions**
226 We defined fusions as cases where for a predicted gene, the CDS of an associated transcript
227 overlapped with the CDS of > 1 NCBI gene, or different CDS transcripts of a predicted gene
228 each overlapped with the CDS of a single NCBI gene, but different predicted CDS transcripts
229 overlapped with different NCBI genes; these definitions were not mutually exclusive. With few
230 exceptions, the rate of putatively false fusions fell below 5% across species and methods, with
231 notable exceptions for a handful of MAKER and $BRAKER_{RNA}$ gene sets (Fig. 4). In general,
232 fusion rates for HMM-based methods were lower than for RNA-seq assemblers, likely due to the
233 tendency for shorter CDS lengths of the former; TOGA fusion rates were consistently among the
234 lowest (Fig. 4). Gene-level fusions were not merely due to individual predicted CDS transcripts
235 spanning the CDS of multiple NCBI genes, but were often dominated by cases where different
236 transcripts from the same predicted gene each overlapped the CDS of a different NCBI gene
237 (Supplemental Fig. S8).
238
239 **Protein sequence completeness**
240 BRAKER and CGP consistently had the highest percentage of predicted proteins that had
241 proper start and stop codon without any internal stop codons (Supplemental Fig. S9), with the
242 former usually outperforming the latter by a modest margin. TOGA predictions had up to 20%
243 fewer complete proteins than these methods, with Stringtie performing either slightly better or
244 slightly worse, depending upon the species. Scallop consistently had the smallest percentage of
245 predictions that represented complete proteins. Despite being a pipeline meant to process
246 predictions from multiple *ab inito* tools (e.g. AUGUSTUS), MAKER performed worse than

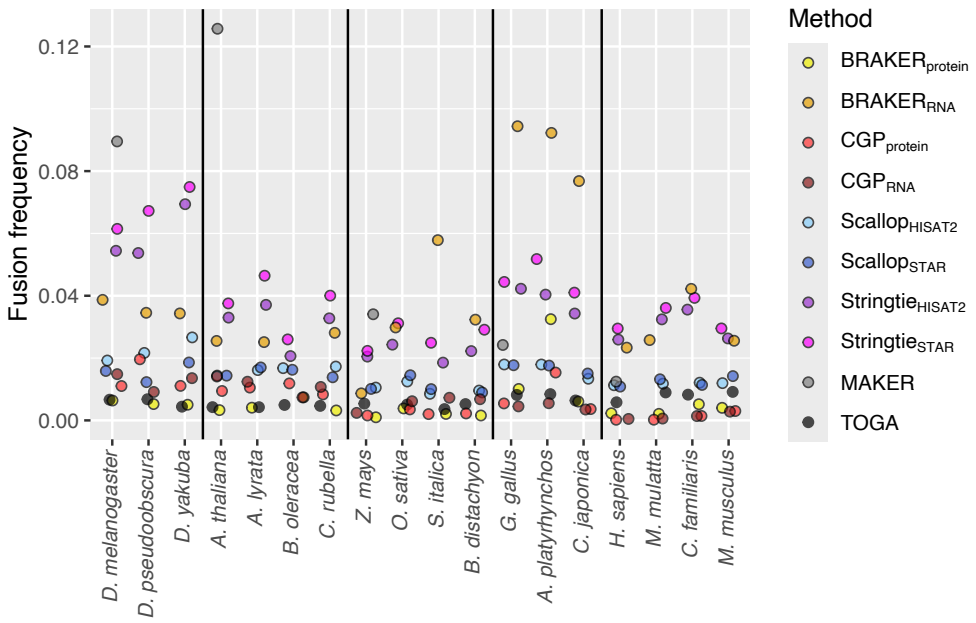


Figure 4. Frequency of gene fusions by species and method. Fusions are defined as when individual transcripts overlap the CDS of multiple NCBI genes, different transcripts from the same predicted gene each have overlaps to different genes, or a combination of both of these. Frequencies are calculated after filtering out NCBI-annotated fusion events.

247

248 BRAKER and CGP and was often, and for plants had the smallest proportion of complete

249 protein predictions. Nevertheless, most predicted transcripts in an annotation have a structure

250 consistent with complete proteins, and with the exception of Scallop, rarely dipped below 80%.

251 The utility of “completeness” may be misleading as a stand-alone measure of whether a

252 prediction is correct (or representing a true protein-coding sequence), if the wrong trinucleotide

253 sequences are classified as start and stop codons due to incorrect inference of splice sites. That

254 this may happen is highlighted by contrasting rates of completeness with the frequency with

255 which predicted protein sequences match those in high quality protein databases. For each

256 species, BLASTP was performed against a database comprised of proteins derived from the

257 NCBI annotations of the species we included in their taxonomic group, including the species in

258 question. The overall proportions of BLASTP hits were, with the exception of CGP, very high

259 (Fig. 5A). TOGA was consistently the top performer, but for dipterans and rosids, TOGA, RNA-

260 seq assemblers and BRAKER were barely distinguishable; for monocots, birds, and mammals,

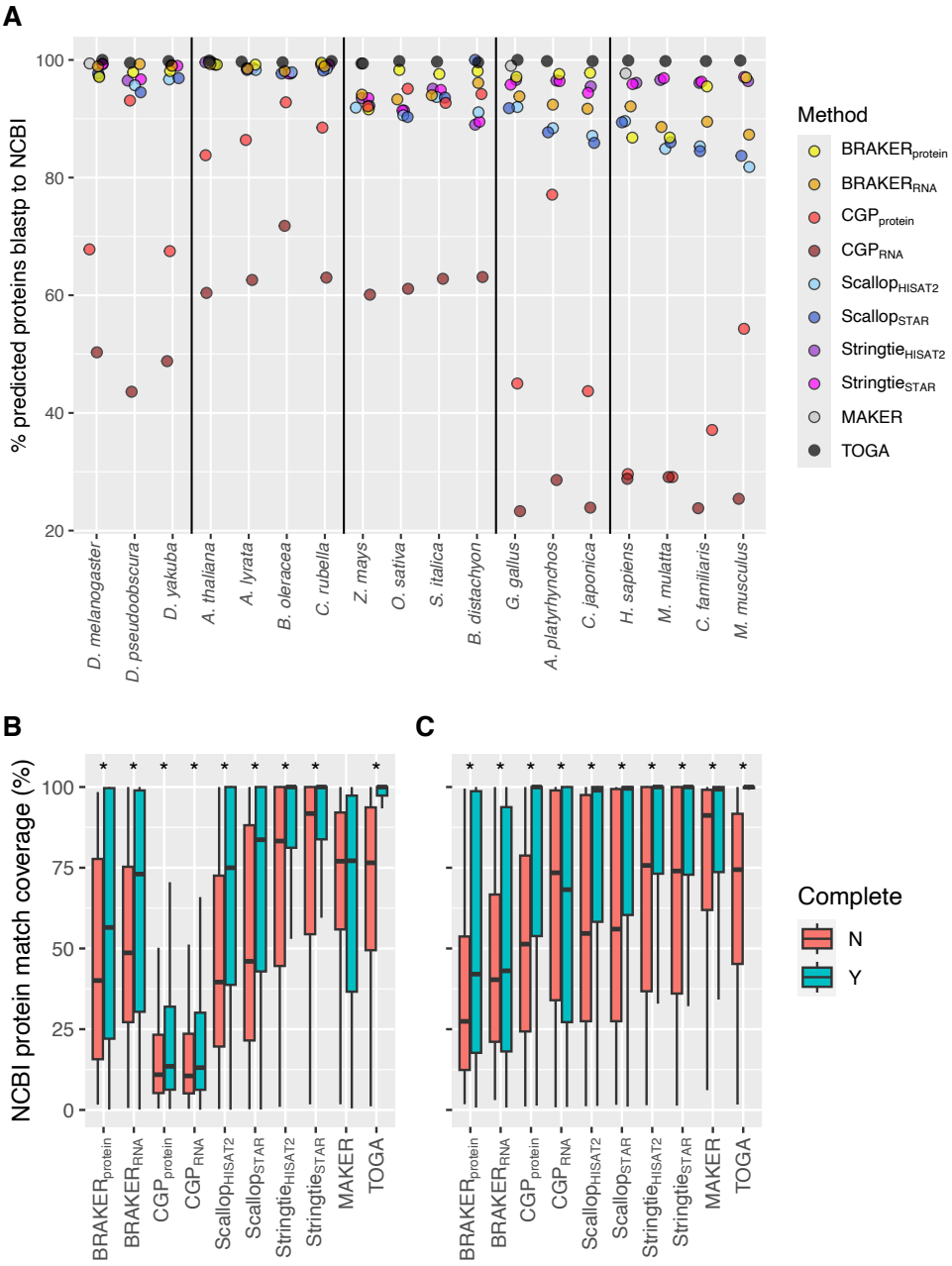


Figure 5. (A) By species and method, proportion of predicted proteins with BLASTP hit to NCBI proteins of species in taxonomic group. For predicted proteins with BLASTP hits, proportion of NCBI best hit target covered by amino acid matches with the predicted protein for (B) *H. sapiens* and (C) *Z. mays,* broken into proteins that are complete (start and stop codons present, no internal stop codons), and those that are not. Benjamini-Hochberg adjusted p-values for Wilcoxon rank-sum tests p≤0.05 indicated by *.

261

262 BRAKER with protein evidence was the second highest rate of BLASTP hits, with Stringtie
263 following close behind (Fig. 5A). However, when we focused on our reference species for which
264 NCBI annotations are of highest quality, in looking at the distributions of the coverage of NCBI
265 proteins in BLASTP hits—defined as the number of matching bases by the length of the best-hit
266 target—with the exception of *D. melanogaster*, there was broad overlap in coverage
267 distributions between complete and incomplete protein predictions for BRAKER, CGP, and
268 MAKER (Figs. 5B, 5C; Supplemental Fig. S10). This suggests these tools are frequently
269 producing truncated protein predictions by identifying the wrong start or stop codons. In
270 contrast, there was far less overlap for RNA-seq assemblers and TOGA. Stringtie and TOGA
271 appear to do a better job than other tools of reconstructing the amino acid sequences of high-
272 quality reference annotations that we defined as our truth set. The strong performance of TOGA
273 cannot be attributed solely to the fact that the annotations being transferred originate from one
274 of the species contained in our protein databases used for BLASTP searches, as excluding the
275 proteins originating from the species whose annotations are being transferred led to negligible
276 decreases in percentages of predicted proteins with hits: *H. sapiens*, 99.9% vs. 98.9%; *G.*
277 *gallus*, 100% vs. 99.7%; *D. melanogaster*, 100% vs. 99.7%; *Z. mays*, 99.4% vs. 99.3%; and *A.*
278 *thaliana*, 99.9% vs. 99.8%.
279
280 **Transcriptome representation: expression**
281 Random forest models we constructed to distinguish intergenic predictions from ones
282 overlapping known protein-coding gene intervals indicate that the former may be characterized
283 by low expression. Predictions by BRAKER and CGP (regardless of the evidence type) and to a
284 lesser extent MAKER contain larger proportions of genes with TPM < 1 (Fig. 6A; Supplemental
285 Fig. S11) relative to RNA-seq assemblers. HMM-base methods using RNA-seq evidence often
286 had larger or comparable fractions of lowly expressed genes compared to their counterparts

287

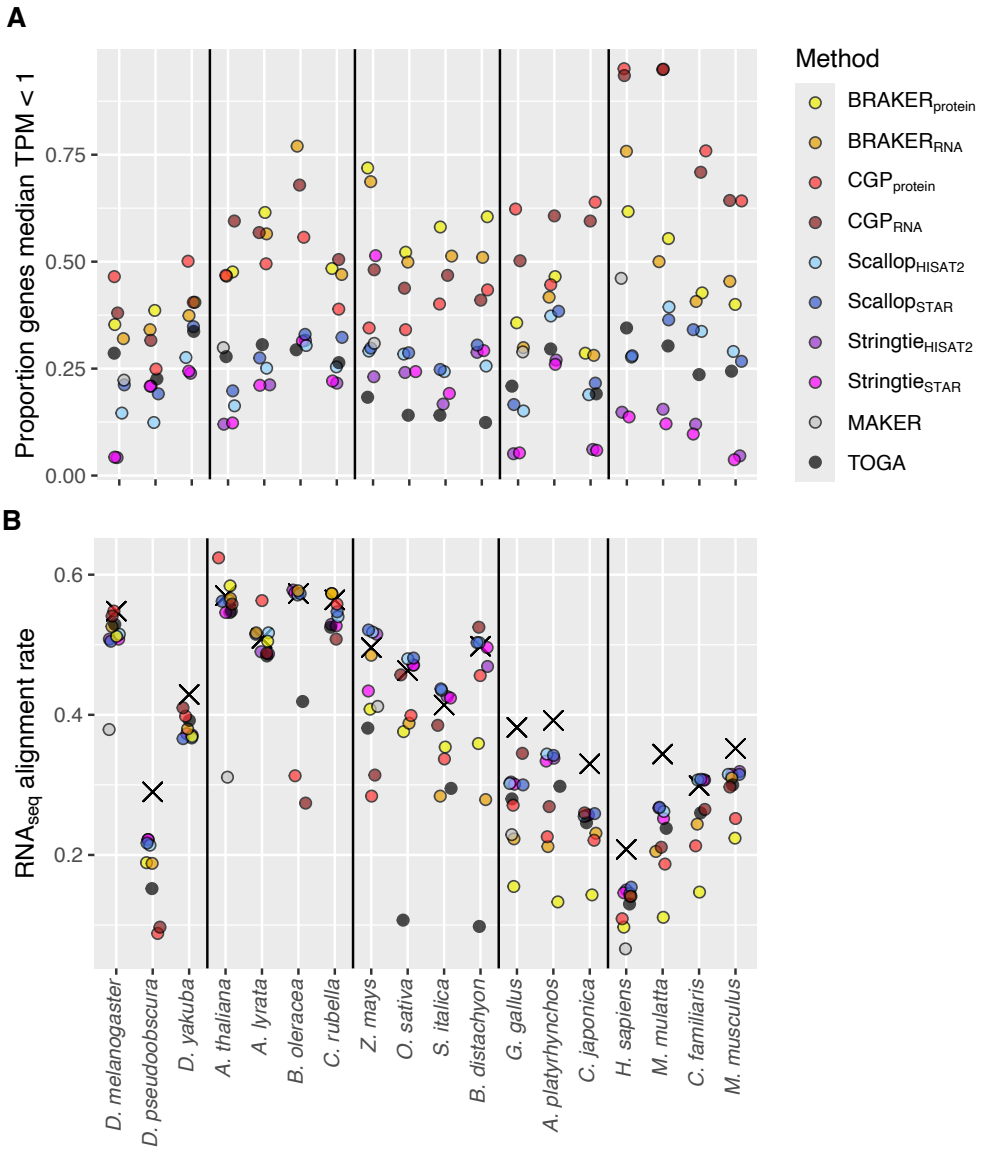


Figure 6. By species and method, (A) proportions of predicted genes for which TPM < 1 , and (B) RNA-seq read alignment rates to predicted transcripts. For Stringtie and Scallop, UTR intervals have been removed, such that all annotations are for CDS only. Xs denote rates for the NCBI annotations (also excluding UTR intervals).

288

289 using protein evidence (Fig. 6A; Supplemental Fig. S11). TOGA predictions did not have

290 substantially elevated proportions of low TPM genes (relative to RNA-seq assemblers), and for

291 monocots had the lowest fraction of lowly expressed genes (Fig. 6A; Supplemental Fig. S11).

292 This may be due to alignments of expressed sequences being part of the evidence used by the
293 NCBI annotation pipeline. In general, with their direct connection to expression, transcript
294 assemblies of RNA-seq data (particularly those based upon Stringtie) consistently had the
295 smallest fraction of lowly expressed transcripts.

296 The RNA-seq read alignment rate to an annotation provides an estimate of the extent to
297 which an annotation captures the underlying expressed transcriptome. Because BRAKER2,
298 CGP, MAKER, and TOGA do not predict UTRs, and because BRAKER1's UTR prediction
299 option was an experimental features we chose not to include, we removed UTR intervals from
300 Stringtie and Scallop annotations to make methods comparable. When UTRs were excluded
301 from consideration, alignment rates were relatively low, only ever exceeding 50% for rosids, *D.*
302 *melanogaster,* and for two of four monocots (Fig. 6B); this is partly due to the RNA-seq data
303 containing reads originating from UTR intervals. While RNA-seq assemblers had alignment
304 rates that were frequently the highest for a particular species, there were only modest
305 differences between these tools and the best performing implementations of BRAKER or CGP
306 (Fig. 6B). Rates for TOGA were lower than these, and were at approximately 10% for two of
307 four monocots, with rates for MAKER also being low relative to most other tools (Fig. 6B).
308 Regardless of method, alignment rates for human annotations always fell below 20%. Even so,
309 alignment rates varied in a manner similar to NCBI annotations without UTRs (Pearson's ρ =
310 0.86, $p=2.2 \times 10^{-16}$), albeit in most species-by-method combinations lower than those for NCBI
311 (Fig. 6B).

312 Because these comparisons were based upon alignment of the same reads that were
313 assembled by Stringtie and Scallop, we considered the possibility that this would provide an
314 unfair advantage to the assemblers relative to tools that only used RNA-seq data to generate
315 splice hints ($BRAKER_{RNA}$ and CGP_{RNA}), or to filter HMM-based predictions post hoc (MAKER),

316 and even more so for methods that did not use RNA-seq data ($BRAKER_{protein}$ and $CGP_{protein}$).
317 Training and test data alignment rates were strongly correlated (Pearson's $\rho = 0.86$, $p=2.2 \times 10^{-}$
318 ^{16}), although for all species except human (and for the $BRAKER_{protein}$ annotation for chicken),
319 test alignment rates were lower than training rates (Supplemental Fig. S12A). Nevertheless, the
320 reduction in test data alignment rates relative to training data were modest, only exceeding 10%
321 for *D. melanogaster* and *Z. mays* for several methods (Supplemental Fig. S12B). Only for *D.*
322 *melanogaster* was there a clear drop in the test alignment rate for Stringtie and Scallop relative
323 to other methods (Supplemental Fig. S12), suggesting that the recycling of training data for
324 alignment does not generate substantial bias in the broad patterns we observe.
325 RNA-seq assemblers are agnostic to the functional role of the sequence intervals from
326 which reads originate, while HMM-based approaches and TOGA do not predict UTRs. The
327 inclusion of UTR intervals predicted by the assemblers led to large increases in their alignment
328 rates, such that they outperformed other methods (Supplemental Fig. S13). The magnitude of
329 this alignment rate difference raises questions regarding the contents of predicted UTR
330 intervals. We thus assessed whether the lengths of predicted UTRs are consistent with those
331 found in NCBI annotations, and whether the alignment rate increase they provide is due, at least
332 in part, to there being true CDS intervals contained within UTRs predicted by Transdecoder—
333 CDS that it failed to predict. We address these questions immediately below.
334
335 **UTRs in RNA-seq assemblers**
336 Ratios of UTR to CDS length are greater for Scallop and Stringtie assemblies than for NCBI
337 annotations (Supplemental Fig. S14). These disproportionately long (relative to NCBI) UTRs
338 constitute an excess of target sequence for alignment, such that their exclusion clearly
339 contributes to a large drop in alignment rates. We considered the possibility that the greater
340 proportional length of UTRs for RNA-seq assemblers relative to NCBI annotations could be due

341 to the NCBI pipeline computationally truncating UTRs when mitigating for cases of putative
342 transcriptional readthrough past stop codons. In this case the gold standard genomes (*H.*
343 *sapiens*, *M. musculus*, *D. melanogaster*, *Z. mays*, and *A. thaliana*) for which there has been
344 extensive curation would be expected to have a lower ratio of UTR to CDS length. However,
345 plotting the ratio of predicted UTR to CDS ratios for the assemblers over that for NBCI
346 predictions produces the opposite pattern, where these ratios of ratios are lower for the gold
347 standard genomes (Supplemental Fig. S14).
348 Next, we evaluated whether Transdecoder pipeline fails to predict CDS at the ends of
349 transcripts, and be default assigns them to the UTR functional class. If this were a pervasive
350 problem, then we would expect a large fraction of UTRs to have BLASTX hits to an NCBI
351 protein database for the same species. As would be expected if undetected CDS occur in the
352 UTR intervals for RNA-seq assemblers, the percentage of transcripts with a UTR BLASTX hit as
353 the length of UTRs relative to CDS for assemblers increases relative to that observed for NCBI
354 annotations (Fig. 7). Depending upon the species and particular assembler-aligner combination,
355 up to 60% of transcripts may potentially contain undetected CDS in regions annotated as UTRs.
356 This suggests the failure to detect CDS plays a substantial role in the drop in RNA-seq
357 alignment rates when UTRs are excluded. This then leads to an underestimate of the realizable
358 alignment rate for RNA-seq assemblers than if these CDS were properly incorporated into the
359 annotations. It seems likely that the frequency of such undetected CDS exons may contribute to
360 the tendency for assemblers to have lower percentages of proteins classified as complete
361 relative to HMM-based methods (Supplemental Fig. S9).
362
363 **Multi-method integration**
364 Above, we have demonstrated that MAKER, a workflow for integrating and filtering predictions
365 from multiple *ab initio* prediction tools, underperforms compared to several different stand-alone

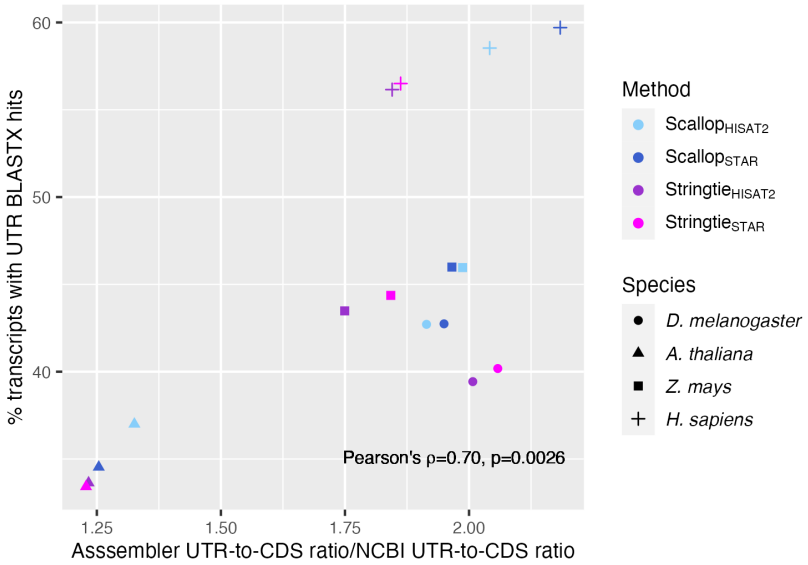


Figure 7. Evidence for undetected CDS in predicted UTR intervals for Stringtie and Scallop. Increasing percentage of RNA-seq assembler transcripts with UTRs that have a BLASTX hit to the NCBI protein database of the same species, as a function of an increase in the assembler UTR-to-CDS ratio relative to that for NCBI annotations.

366

367 annotation methods and across a variety of metrics. Another integration approach, TSEBRA
368 (Gabriel et al. 2021a) selects transcripts and merges transcript models from separate runs with
369 protein and RNA-seq evidence, respectively. TSEBRA integration produces an annotation with
370 BUSCO scores that are no better than the best of the two BRAKER runs (Supplemental Fig.
371 S15A), and in most cases have slightly worse scores. Proportions of predicted transcripts with
372 BLASTP hits to the NCBI proteins from the species group we investigated are consistently
373 worse than either BRAKER implementation (Supplemental Fig. S15B). Similarly, in four of five
374 reference species, the percentage of intergenic genes is higher for TSEBRA than either
375 BRAKER implementation (Supplemental Fig. S15C), with RNA-seq alignment rates falling
376 between the two BRAKER runs (Supplemental Fig. S15D). In short, MAKER and TSEBRA lead
377 to a loss of meaningful transcriptome information relative to the best of the alternative BRAKER
378 approaches.

379 While not necessarily an optimized method for merging annotations, we investigated
380 whether adding genes (and their constituent child features) from a second annotation to a base
381 annotation—requiring that those added genes fell entirely outside of the gene intervals of that
382 base annotation—would leverage the complementary strengths of different methods while
383 minimizing information loss. We also explored whether successive additions from different
384 methods would continue to improve the overall annotation. There was a high degree of
385 variability in the benefits of such integration. While stand-alone methods like TOGA had the
386 highest BUSCO score, adding TOGA annotations to those of Stringtie achieved the best
387 balance of maximizing both BUSCO score and RNA_{seq} alignment rate, overcoming the tradeoff
388 observed between these two metrics often seen in individual methods (Fig. 8A). Which method
389 one chooses as the base annotation can impact BUSCO scores, the number of genes that have
390 BLASTP hits to NCBI proteins, the percentage of predicted genes with such hits, and the RNA-
391 seq alignment rate. For example, using $Stringtie_{STAR}$ as a base annotation leads to lower
392 BUSCO scores and higher alignment rates than for integrations that start with TOGA (Figs. 8A,
393 8B; Supplemental Figs. S16, S17), and while integration increased, as expected by definition,
394 an increase in predicted genes, the percentage of genes with at least one transcript having a
395 BLASTP hit to an NCBI protein decreased, suggesting that both real and spurious predictions
396 get added (Supplemental Figs. S16-S19). Method integrations that did not include TOGA
397 produced high RNA_{seq} alignment rates but lower BUSCO scores than TOGA and integrations
398 that used it as the base annotation (upon which to add others), with reasonably high fractions of
399 genes with BLASTP hits (Fig. 8C; Supplemental Fig. S18). Finally, in a scenario where RNA-
400 seq data is not available, adding $BRAKER_{protein}$ annotations to those of TOGA did little else other
401 than to minimally increase the RNA_{seq} alignment rate (Fig. 8D; Supplemental Fig. S19). In these
402 four scenarios, adding $BRAKER_{protein}$ had negligible effect.
403

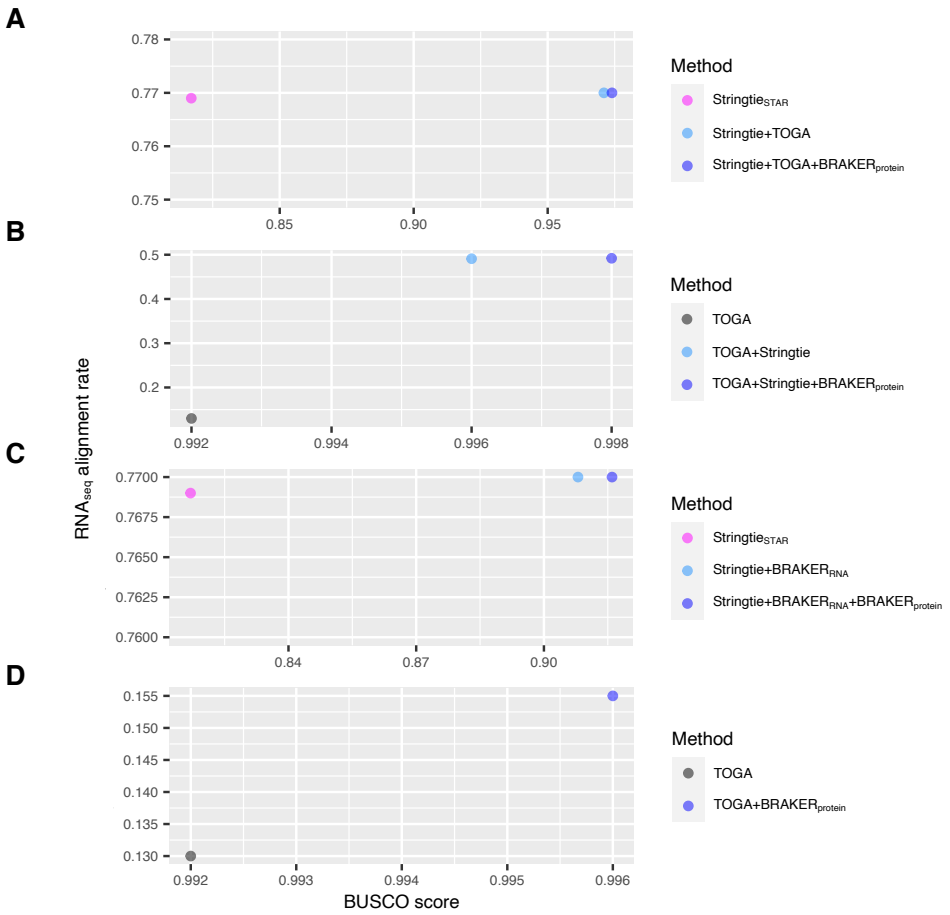


Figure 8. Bi-plots of median sample RNA-seq alignment rate versus BUSCO score for individual base annotation methods and subsequent integrations for (A) Stringtie$_{STAR}$ as the base, with subsequent integration of TOGA and BRAKER$_{protein}$ (B) same as B but with TOGA as the base annotation, (C) the absence of a closely related high quality genome annotation, precluding the use of TOGA, with Stringtie$_{STAR}$ set as the base annotation, and (D) no RNA-seq data, with TOGA as the base annotation, and an integration of BRAKER$_{protein}$. Integration involves successively adding genes from one annotation that fall entirely outside of the annotation to which those genes are being added.

DISCUSSION

With genome assembly and annotation increasingly becoming part of the workflow for researchers studying non-model organisms, the choice of an annotation method depends upon knowing whether a particular method will perform well in the species in question, as well as what data will need to be generated to generate the best quality annotation possible. Previous efforts to evaluate and compare methods have sampled a small slice of the tree of life, and often focused on small, tractable genomes with extensive genomic resources, or other model

412 organisms such as *H. sapiens* or *M. musculus*. This can make choosing a method more difficult,
413 because the species used to benchmark annotation tools may be evolutionary distant from a
414 newly assembled genome of interest. Our investigation overcomes this problem by evaluating a
415 large set of methods across the broadest taxonomic swath investigated to date, revealing both
416 cross-species and taxon-restricted patterns. By identifying methods that performed well across
417 diverse species, we believe that researchers using those methods will likely be able to generate
418 reasonably high-quality annotations for their newly assembled genome of interest. Our findings
419 have implications for study design, data collection, and annotation method choice, and highlight
420 ongoing challenges that require further methods development.

421 First and foremost, the inclusion of RNA-seq data will invariably improve annotation
422 quality, particularly when used to assemble transcripts directly from sequence alignment with
423 Stringtie. While HMM-based methods such as CGP and BRAKER using either protein or RNA-
424 seq evidence can produce high BUSCO scores, they consistently lag behind RNA-seq
425 assemblers with respect to their representation of the underlying transcriptome—as
426 characterized by read alignment rates. Furthermore, HMM-based approaches typically produce
427 thousands of false-positive predictions which need to be filtered but may be difficult to identify.
428 That HMM-based predictions may consistently be "complete", with proper start and stop codons,
429 yet also shorter than NCBI transcripts and those produced by RNA-seq assemblers, highlights a
430 shortcoming of HMM-based methods—splicing patterns that are consistent with an inferred
431 model of protein-coding sequence structure may not be the real pattern, instead representing
432 truncated open reading frames or spurious predictions. While we did not assess the utility of
433 cDNA long reads for annotation, we expect our findings to be robust to their adoption, and that
434 the performance disparities between RNA-seq assemblers and HMM-based methods will almost
435 certainly widen. The direct evidence of splicing patterns across full length reads will enable
436 reconstruction of full-length transcripts, while in the HMM context, longer reads will simply lead

437 to more accurate detection of splice sites, and more accurate model parameterization—to the
438 extent that any model can capture the diversity of sequence composition and splicing patterns
439 observed in higher organisms. We suspect that increasingly accurate model parameterization
440 will lead to diminishing returns relative to direct assembly of transcripts from reads. RNA-seq
441 assemblers are particularly valuable for larger, more complex genomes and consistently rank
442 among the top methods across diverse performance metrics. Furthermore, and while not the
443 focus of this study, assemblers permit the inclusion of non-coding RNAs in an annotation, with
444 the caveat that it is more difficult to distinguish real non-coding transcripts from spurious
445 assembly of low-coverage transcriptional noise. Overall, while the advantages of using RNA-seq
446 assembly over other methods may be diminished for smaller, less complex genomes, it clearly
447 produces better annotations than HMM-based approaches for more complex genomes, and
448 reliably produces relatively complete annotations regardless of taxonomic group or genome
449 organization.

450 The demonstrated superiority of RNA-seq assemblers to HMM-based approaches may
451 in fact be an underestimate of their relative performance. Our discovery that predicted UTRs
452 have high fractions of sequence that with BLASTX hits to NCBI proteins suggests that we failed
453 to recover many CDS exons. Our pipeline for producing CDS annotations uses Transdecoder,
454 and our results suggest that it may have a harder time correctly classifying CDS exons at the
455 termini of a transcript than those within the transcript body. While long-read technology might
456 help overcome this deficiency, we suggest that there is room for methods development to
457 improve ORF detection from predicted CDS transcripts. Improved ORF detection and CDS exon
458 boundary delineation would lead to improved performance with respect to several of the metrics
459 we used to compare methods in this study.

460 Our findings also highlight the power and limitations of annotation transfer from another
461 species with a high-quality annotation, as is done by TOGA. While TOGA often had high

462 sensitivity (BUSCO scores) accompanied by low rates of intergenic predictions and gene
463 fusions, we found that sensitivity could be much lower in plants, especially in those with larger
464 genomes such as *Z. mays*. This undoubtedly stems from known difficulties performing whole-
465 genome alignment with plant genomes (Song et al. 2024). Given the strong performance of
466 RNA-seq assemblers across all the species we surveyed, researchers should consider RNA-
467 seq assembly when there are known issues performing whole genome alignment for the
468 species in question. For some taxa, there may be few if any closely related species with high-
469 quality annotations, or genome alignments may be fragmented or contain many missing
470 intervals in the target species.

471 Our finding that, for some species, there may complementarity among methods for
472 which BUSCOs are recovered—and that RNA-seq based methods can recover genes that
473 protein-based methods fail to predict—suggests that integration of annotations across multiple
474 methods may potentially improve sensitivity. That MAKER performed poorly and TSEBRA
475 appears to filter out real annotations when integrating protein and RNA-seq iterations of
476 BRAKER, suggest that additional methods development on annotation integration is needed.
477 Our naïve approach of consecutively adding annotations from one method that did not overlap
478 with a base annotation also increased sensitivity, leading to an increase in the number of
479 protein-coding genes with BLASTP hits to NCBI proteins. Nevertheless, we did not attempt to
480 apply any filters to remove the spurious annotations that were invariably added in that workflow.

481 While we did not explore in depth how applying various filters impacted performance
482 metrics, the frequent observation of high rates of intergenic predictions, as well as the
483 occurrence of gene fusions, strongly suggest that more work is needed to identify filters that
484 strike the balance between removing as many low-quality annotations as possible, while
485 minimizing the filtering out of real sequences. It is often the case that filters applied in the
486 literature appear *ad hoc*, even if guided by intuition and experience. We suggest a more

487 quantitative approach is needed. For example, our application of random forest to classify genic
488 and intergenic predictions suggests that machine learning approaches, while not perfect, offer
489 great promise in identifying which variables should be used to set filtering thresholds. We found
490 that whether a sequence had a BLAST hit to a set of known proteins, and expression level were
491 useful discriminatory variables. Of course, a new genome does not benefit from the advantage
492 of a "truth set" of real transcripts. However, a random forest (or other) model could, in principle,
493 be trained with annotations from a related species, and that model could be applied to the
494 predicted transcripts for a new genome assembly. Future work is needed to explore the utility of
495 such an approach.

496 Even as much work remains to be done, our findings suggest some general guidelines
497 for a researcher deciding how to annotate their newly assembled genome.

498 1. Generate RNA-seq data for at least the tissues related to the most pressing project
499 needs, but ideally, across as many tissues as necessary to capture the species'
500 transcriptional complexity. Use Stringtie to assemble transcripts, and, until a better
501 option is available, the Transdecoder workflow for adding CDS features to the
502 annotation.
503 2. Consider using TOGA if RNA-seq data are not available and a well-annotated high-
504 quality genome is available for a closely related species. If there is complementarity in
505 the recovery of seemingly real protein coding sequences (determined with BUSCOs or
506 gene symbols extracted from BLAST hits to established protein databases), consider an
507 approach that integrates predictions of TOGA with $BRAKER_{protein}$, giving more weight to
508 TOGA that BRAKER.
509 3. If it is not possible to use TOGA and RNA-seq data are not available, use $BRAKER_{protein}$.

4. If RNA-seq data are available and if there appears to be complementarity in the recovery of real protein-coding sequences between the methods, consider using an approach to integrate predictions from Stringtie and TOGA, giving more weight to Stringtie.
5. Either through a statistical approach such as random forest, or through heuristically-thresholded metrics (e.g. expression level, BLAST hits to an established protein database), remove predictions with a high likelihood of being intergenic.
6. Given the non-trivial frequency of fusions detected in the methods we analyzed (with the exception of TOGA), consider flagging genes that are likely fusions, e.g. if BLAST hits of different of different transcripts are to functionally distinct genes produced by genes with clearly different symbols, or similarly, if subsequences of individual transcripts have such divergent BLAST targets. Exclude these fusion genes from downstream expression analyses.
7. Consider using CGP in edge cases where, for example, there are a handful of incomplete annotations for some related species (perhaps generated by already available RNA-seq data), and the genome of interest is of small or modest size. Expect to do extensive filtering to exclude many spurious predictions.

In conclusion, the longer-term challenge for building genome annotations across the tree of life is to make methodological advances suggested above, and to integrate them into reproducible, automated workflows that can be deployed with minimal headaches for biologists. When this happens, population and comparative genomics studies will be easy to scale to hundreds, and even thousands of species, unleashing unprecedented power to tackle long-standing questions regarding the genetic architecture of phenotypic variation and the evolutionary mechanisms that generate and maintain biodiversity.

535 **METHODS**

536 **Target taxa**

537 Genome annotation tools are typically developed and optimized using high quality genome
538 assemblies from a small suite of model organisms, e. g. *Homo* sapiens, *Caenorhabditis*
539 *elegans*, and *Drosophila melanogaster*. As a result, it is difficult to generalize their performance
540 in this narrow context to taxonomic groups that are highly divergent from those focal taxa, and
541 for which the genome assemblies may not be of comparable quality. To facilitate more accurate
542 generalizations regarding the performance of annotation methods, and, conversely, to explore
543 whether there are effects of taxonomy and genome structure on annotation quality, we
544 generated genome annotations for 21 species spanning six taxonomic groups: three species of
545 heliconiine butterflies, three *Drosophila* species (dipterans), three birds, four mammals, four
546 rosids, and four monocots (Supplemental Table 1). With the exception of the butterflies, we
547 included as a "reference" a species for which both a high-quality genome assembly and
548 annotation were available, and downloaded assemblies and annotations from NCBI. Each group
549 contains at least one species that is relatively closely related to this reference. Because the
550 NCBI genome versions and annotations for heliconiines are older than those widely used by the
551 heliconiine research community, we used lepbase assemblies that were filtered to remove all
552 scaffolds less than 1kb in length (Edelman et al. 2019). High quality annotations for these
553 species were either unavailable, or generated by tools we evaluated and thus inappropriate to
554 serve as a truth set. For example, the annotation for *H. melpomene* was generated with
555 BRAKER. Because of the unavailability of gold standard annotations for heliconines, for these
556 species we only generate a subset of performance metrics that don't rely of an annotation truth
557 set. Furthermore, *H. melpomene* is used as a reference assembly in whole-genome alignments
558 (see below) but we did not generate annotations for this species.

559

560 **Genome assembly and reference annotation quality**

561 To understand how features of genome assembly structure and quality might impact the quality
562 of annotations, we generated several summary statistics describing genome contiguity and
563 completeness. We generated standard statistics such as the total genome size, the contig and
564 scaffold N50, and the fraction of genomic nucleotides that were soft-masked. We also quantified
565 the number of single copy orthologs (BUSCOs) contained in a genome (Simão et al. 2015), and
566 calculated a BUSCO score as 1 – (number missing BUSCOs/total number of BUSCOs
567 searched).

568 To assess the quality of the reference genome annotations generated by NCBI, for each
569 genome we extracted protein-coding transcripts and generated transcriptome BUSCO scores,
570 and statistics on the minimum, maximum and median CDS length and fraction of CDS bases
571 that were soft-masked. We also calculated the median RNA-seq alignment rate (see below) for
572 each species' annotation, as well as the fraction of transcripts for which estimated transcripts
573 per million (TPM) was < 1.

574

575 **RNA-seq data acquisition and processing**

576 For use with annotation methods that either build transcripts from RNA-seq reads or use read
577 alignments to generate splice hints, for each species we downloaded 15-20 fastq file accessions
578 from NCBI's short read archive (SRA). We used the following criteria to choose accessions. We
579 only considered 1) paired-end reads with a release data of 2011 or later, 2) at most one run per
580 biosample, 3) Illumina reads sequence on HiSeq 2000, NextSeq or newer instruments (i.e. no
581 Genome Analyzer II), 4) we excluded experimental treatments such as knockdowns, infections,
582 and CRISPR modifications. If these criteria resulted in > 20 possible biosamples, we further
583 required a minimum read length of 100bp, and for Metazoans, preferentially selected brain or
584 head samples. If < 20 samples were available, we relaxed read length, release date, and

585 instrument criteria with the goal of retaining 15 biosamples; with the exception of the heliconiine
586 butterfly *D. plexippus* (for which 8 of 19 libraries had 36bp reads), we strictly excluded paired-
587 end libraries where the read length was < 50bp. In a small number of cases where libraries
588 contained hundreds of millions of reads, we down-sampled libraries to approximately 20 million
589 read pairs with seqtk (https://github.com/lh3/seqtk).
590 To process the reads prior to sequence alignment, we stripped adapters with TrimGalore
591 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
592). We did not trim low-quality bases at the ends of reads because the short read aligners used in
593 this study soft-clip such bases such that they do not impact sequence alignment, and because
594 such trimming can bias expression estimates (Williams et al. 2016). So as to avoid having to
595 make inferences from SRA metadata regarding the strandedness of RNA-seq libraries, and to
596 avoid the likely variable effectiveness of stranding protocols, we treated all libraries as
597 unstranded, under the assumption that such information, if used would result in modest
598 improvements in performance for annotation methods that leverage RNA-seq alignments.
599
600 **Annotation tools**
601 *RNA-seq assembly*
602 We used RNA-seq reads to directly assembly transcripts with StringTie v. 2.1.2 and Scallop v.
603 0.10.5. These assemblers take as input spliced alignments of reads to the genome. We
604 evaluated the impact of aligner by generating assemblies with two different aligners: HISAT2 v.
605 2.2.1 and STAR v. 2.7.9a. Following best practice, and to leverage evidence for splice sites
606 across multiple samples, we used a 2-pass approach to generate STAR alignments. In the first
607 pass, we generated an initial set of alignments for each sample. In the second pass, we
608 concatenated the splice site tables generated for each sample's first-pass alignment, and
609 supplied the concatenated table as splice-site evidence for the second pass alignment of each

610 sample. To generate a merged transcript assembly combining information across all individual-
611 level assemblies, for StringTie and Scallop we use the stringtie-merge function and TACO
612 (Niknafs et al. 2017) v. 0.7.3, respectively. With our heliconiine species, we initially evaluated
613 two additional assemblers: PsiCLASS (Song et al. 2019) and Scallop2 (Zhang et al. 2022, 2).
614 Poor performance relative to StringTie and Scallop, as well as excessive run times for
615 PsiCLASS, led us to not consider these two tools further.
616 StringTie and Scallop annotations contain transcript and exon features: they do not
617 predict CDS. Therefore, to incorporate CDS predictions into the merged annotations, we use a
618 TransDecoder v. 5.5.0 (https://github.com/TransDecoder/TransDecoder) and an associated
619 workflow (https://github.com/TransDecoder/TransDecoder/wiki) that predicts orfs, and then
620 leverages orf predictions to predict CDS and UTR intervals associated with the gtf format input
621 annotation file. After initial prediction of likely candidate orfs, we run blastp (Altschul et al. 1990)
622 v. 2.12.0 searches against a protein database consisting of Uniprot and Trembl entries from all
623 the species that we are attempting to annotate in the species group, e.g. for dipterans, the
624 database consists of entries for *D. melanogaster*, *D. pseudoobscura*, and *D. yakuba*. We
625 provide the search results as an input to transdecoder-predict, such that given two similarly
626 scoring orfs, we preferentially keep the one with a blastp hit. In the interest of minimizing the
627 filtering out of real orfs, we set the maximum e-value threshold for these blastp searches to 1 x
628 10^{-4}. It should be noted that the workflow as described filters out of the final annotation any
629 transcript without a retained orf prediction. The filtered orfs contain an unknown fraction of real
630 orfs that TransDecoder failed to discover, as well as ncRNAs. While the focus of our research is
631 on prediction of protein-coding genes, we provide as part of our code repository an accessory
632 script for adding back into the final annotation these putative false negatives and ncRNA
633 annotations.
634

635 *Single-species ab initio methods*

636 In contrast to transcript assembly-from-reads approaches, a long-established approach for
637 predicting genes (and coding sequences in particular) is to parameterize hidden Markov models
638 (HMMs) that are designed to traverse scaffolds, identify exon boundaries, and connect exons
639 into transcript and gene-level features. The most sophisticated single-species versions of this
640 approach use external evidence to parameterize HMMs and identify specific genomic locations
641 where exon splice junctions are located. We evaluate BRAKER1 and BRAKER2 (both v. 2.1.6) ,
642 which conduct iterative training and gene prediction using RNA-seq read and protein alignment
643 evidence, respectively. Both BRAKER flavors wrap *ab initio* prediction with AUGUSTUS (Stanke
644 et al. 2006) and GeneMark, with BRAKER1 using GeneMark-ET (Lomsadze et al. 2014) and
645 BRAKER2 using GeneMark-EP+ (Brůna et al. 2020). Following developer recommendations,
646 we provide protein evidence to BRAKER2 in the form of a protein fasta from OrthoDB v. 10
647 (Kriventseva et al. 2019) for the relevant taxonomic group, generated from pre-partitioned raw
648 files as provided by the BRAKER developers (https://bioinf.uni-greifswald.de/bioinf),
649 downloaded on 14 Septermber, 2018. For BRAKER1, we provide a bam file of RNA-seq STAR
650 alignments merged across all libraries from the species being annotated.

651 While we consider multi-method annotation integration below, we also evaluate TSEBRA
652 (Gabriel et al. 2021b), a python script-based tool to combine BRAKER1 and BRAKER2 runs,
653 and select a well-supported subset of transcripts. We run TSEBRA following guidelines
654 available at the TSEBRA github repository (https://github.com/Gaius-Augustus/TSEBRA),
655 running it on the braker.gtf files (that include AUGUSTUS and GeneMark predictions) rather
656 than on the AUGUSTUS-only annotations.

657

658 *Single-species exon-aware liftover: TOGA*

659 Using whole-genome alignments to transfer annotations across species from well-annotated to
660 poorly or un-annotated species has a long history, e.g. with the UCSC Genome Browser
661 LiftOver tool first becoming available in 2006 (Hinrichs et al. 2006). To perform such "liftovers",
662 we use TOGA (Kirilenko et al. 2023), which transfers CDS annotations across genomes in an
663 exon-aware fashion that minimizes disruptions of ORFs. TOGA takes as input a whole genome
664 alignment, and involves several steps, the details of which we provide at
665 https://github.com/harvardinformatics/GenomeAnnotation-TOGA, and are an adaptation of the
666 workflow described at https://github.com/hillerlab/TOGA. To remove potential spurious or bad
667 annotation transfers, we filter out any transcripts in the primary annotation output
668 (query_annotation.bed) for which there was not a corresponding entry in
669 orthology_classification.tsv, i.e. transcripts for which TOGA could not determine an orthology
670 class. Within each taxonomic group, we transfer annotations from the high-quality reference to
671 all other species within the group, and from the species most closely related to the reference
672 back to the reference species. For example, for dipterans we carry out three TOGA analyses,
673 transferring *D. melanogaster* to both *D. pseudoobscura* and *D. yakuba*, and from *D. yakuba* to
674 *D. melanogaster*.
675
676 *Multi-species ab initio annotation*
677 In studies seeking to perform phylogenetic comparative analyses, annotate multiple genome
678 assemblies from related organisms, or where annotations or evidence (protein or RNA-seq)
679 already exist for a subset of species of interest, methods that transfer evidence between
680 lineages offer, in principle, a promising approach for performing genome annotation. We
681 evaluate the most well-established approach for doing this, AUGUSTUS run in comparative
682 mode (König et al. 2016), referred to hereafter as CGP. CGP relies on whole-genome alignment
683 (WGA). Thus, as a first step, for each taxonomic group of genomes, we use Progressive Cactus

684 (Armstrong et al. 2020) to produce a WGA. We then use an AUGUSTUS accessory script,
685 *hal2maf_split.pl*, to split the hal-format cactus output file into multiple sub-files in multiple
686 alignment (MAF) format; in doing so, we set as the "reference" genome (with which to provide
687 coordinate anchors) a species with both a highly contiguous assembly and a high-quality
688 annotation, and split in such a way so as to avoid splits that bisect the genomic coordinates of
689 annotated genes in the reference. For each taxonomic group of species, we run CGP twice,
690 once with splice site evidence from protein alignments, and once from RNA-seq alignments. For
691 analysis with protein evidence, similar to analysis with BRAKER2, we use OrthoDB v.100 data
692 representing the taxonomic group. For analysis with RNA-seq we used the merged STAR
693 alignments across samples. In both instances, following guidelines from the developers (Hoff
694 and Stanke 2019), we generate splice hints files for each species using scripts and code
695 provided as part of the AUGSTUS package. In both modes, we do not predict UTRs, nor do we
696 predict alternative isoform, i.e. one transcript prediction is made per putative gene. Detailed
697 instructions regarding how we generated hints and run CGP are found at
698 https://github.com/harvardinformatics/GenomeAnnotation-ComparativeAugustus.
699
700 *MAKER*
701 MAKER (Cantarel et al. 2008) is a genome annotation pipeline that has the ability to integrate
702 multiple *ab initio* gene prediction packages, and to use protein and RNA-seq derived external
703 evidence to perform post hoc curation of predictions. Because results with MAKER usually
704 involve > 1 runs in order to retrain gene-prediction models, it is not a fully automated pipeline.
705 Nevertheless, it has been used extensively due to its purported ease of use. MAKER also has
706 the option to perform quality filtering and integration of annotations with EVidenceModeler
707 (EVM) (Haas et al. 2008). For initial testing with three heliconiine species, we ran MAKER v.
708 3.01.03 four different ways that integrate predictions from AUGUSTUS (Stanke et al. 2006),

709 SNAP (Korf 2004), and Genemark-ES (Lomsadze et al. 2005): 1) protein evidence only, without
710 EVM; 2) protein and RNA-seq evidence, without EVM; 3) protein evidence only, with EVM; and
711 4) protein and RNA-seq evidence, with EVM. For protein evidence, we used the protein
712 accessions associated with the lepbase (lepbase.org) Hmel2 genome assembly, which are
713 proteins derived from BRAKER predictions. RNA-seq evidence was included as a gff3 file
714 generated from the Stringtie assembly using STAR alignments of the species' RNA-seq
715 samples. We used default settings for the EVM configuration scoring file. To produce
716 annotations, we ran MAKER twice closely following Daren Card's detailed workflow
717 (https://gist.github.com/darencard/bb1001ac1532dd4225b030cf0cd61ce2); see also
718 https://github.com/harvardinformatics/GenomeAnnotation-Maker . Because with these test runs
719 the use of EVM frequently produced lower quality annotations, and because we wished to
720 evaluate the potential of MAKER as a full-service annotation tool, runs for other taxa were only
721 performed with both protein and RNA-seq evidence and without EVM. Furthermore, because
722 MAKER is computationally intensive and can take a considerable amount of time to run, for the
723 other taxonomic groups we only generate MAKER annotations for the "reference" species of
724 each group, with protein evidence being represented by the NCBI protein accession associated
725 with the genome for the next closely related species in the set of species we annotated within
726 each taxonomic group.
727
728 **Annotation quality metrics**
729 Accurate annotation of UTRs is challenging, and even more so for non-model organisms for
730 which RNA-seq data are typically sparse. Similarly, long non-coding RNAs are also difficult to
731 annotate (Uszczynska-Ratajczak et al. 2018). In most contexts, neither feature is crucial to
732 genome-enabled evolutionary studies in non-model organisms. Thus, we focus our evaluation of
733 genome annotation methods on protein coding sequences, filtering out UTRs and non-coding

734 loci. For annotation methods that include the UTR portions of mRNAs, we strip UTR exons (and
735 any UTR-labelled features) from annotation gtf or gff3 files. We also do this for NCBI gff3 files
736 prior to comparisons with the annotation methods we evaluate.

737 We use the NCBI genome annotations as putative sets of "true" annotations. Although
738 the quality of these annotations certainly vary due to many factors – the quality of the genome
739 assembly, the amount and kind of experimental evidence available at the time the annotation
740 was generated, challenges in annotating larger, more complex genomes—we believe they are a
741 reasonable approximation to annotations one would hope to achieve with a new genome
742 assembly, using stand-alone annotation tools deployed on local HPC clusters. Nevertheless,
743 while we make comparisons to NCBI annotations for all species-tool combinations, we pay
744 particularly close attention to those species for which we know the genomes and annotations
745 are of the highest quality: *H. sapiens*, *D. melanogaster*, *Z. mays*, *Arabidopsis thaliana*, and
746 *Gallus gallus*.
747 Details on bioinformatics package command lines and custom python scripts are available in the
748 relevant GitHub repositories detailed in the DATA ACCESS section at the end of this paper.
749
750
751 *Annotation completeness: transcriptome BUSCOs and expression*
752 To assess transcriptome completeness, we calculate transcriptome BUSCO scores and
753 compare them to scores for the NCBI transcriptomes. For comparisons across all species-
754 method combinations, and in order to normalize for varying degrees of genome assembly
755 completeness and quality, we calculate ratios of transcriptome BUSCO score to that for the
756 respective genome. To evaluate the extent to which the predicted transcriptome represents the
757 expressed transcriptome, we then use RSEM (Li and Dewey 2011) v. 1.3.3 to wrap bowtie2
758 (Langmead and Salzberg 2012) alignment of each RNA-seq library to the predicted

759 transcriptome, and estimate both gene and isoform-level expression. From those alignments, for
760 each annotation we calculate the median alignment rate (across the set of samples), and the
761 proportion of genes and transcripts for which TPM < 1. Because using the same RNA-seq
762 libraries to generate transcriptome assemblies with Stringtie and Scallop may bias alignment
763 rates upwards relative to tools that don't leverage evidence from those RNA-seq libraries, we
764 also perform alignments on an additional test set of six RNA-seq paired-end SRA accessions for
765 each of our five reference species.
766
767 *Protein-level statistics*
768 For each genome annotation we report the number of protein-coding genes, and the number of
769 CDS transcripts. We use GetProteinFastaStats.py, a custom python script that leverages
770 biopython (Cock et al. 2009) modules to calculate mean and median protein lengths, the fraction
771 of predicted proteins that have internal stop codons, and the fraction that are complete, where
772 complete is defined as having proper start and stop codons, and no internal stop codons. For
773 TOGA annotations, we generate these statistics from the protein sequences the software
774 outputs (after filtering out those without ortholog classifications). For all other annotation tools
775 we used the version of gffread distributed with cufflinks (Trapnell et al. 2010) v. 2.2.1, to extract
776 the protein sequences.
777 To estimate the fraction of protein predictions that were real, we used blastp (Camacho
778 et al. 2009) v. 2.12.0 to search for matches, with a maximum e-value of 1×10^{-5}, against a
779 database consisting of the NCBI protein accessions for all of the species that were used in our
780 study for the taxonomic group of interest.
781
782 *False positives: intergenic predictions*

783 Motivated, in part, by some tools predicting far more CDS transcripts that are recorded in NCBI
784 annotations, we assessed whether this could be due to (presumably incorrect) intergenic
785 predictions, where intergenic is defined as falling entirely outside of the CDS intervals for all
786 protein-coding genes annotated by NCBI. To do this, for each new annotation, we generate
787 transcript interval and gene interval bed files, where each entry represented the genomic
788 boundaries of the transcript or gene (excluding UTRs), respectively. We then use bedtools v.
789 2.26.0 (Quinlan and Hall 2010) to intersect these files with a bed file consisting of UTR-stripped
790 NCBI gene boundaries, recording the number of bases of overlap such that only same-strand
791 overlaps were counted as overlaps, e.g. *intersectBed -s -wao -a newannotation_intervals.bed -b*
792 *NCBI_gene_intervals.bed*. We then count the number of predicted transcripts and genes lacking
793 any overlap with NCBI gene coordinates.
794
795 *Gene fusions*
796 Real fusion events in which a transcript contains CDS from different annotated genes should be
797 extremely rare. Thus, the presence of non-trivial frequencies of predicted transcripts for which
798 exons span multiple NCBI genes most likely represent bioinformatics pipeline-induced errors.
799 We evaluate fusions at the gene level for which we define three types: 1) an individual predicted
800 CDS overlapping with CDS from multiple NCBI genes, 2) different CDS originating from the
801 same predicted gene overlapping with the CDS of different NCBI genes, and 3) cases where
802 both of these types of fusions occur. To quantify the frequency of these fusion events, we first
803 converted the CDS features of a species' NCBI annotation and the CDS annotations of a
804 method being evaluated to bed format. Next, we used bedtools to perform a "left outer join" of
805 NCBI CDS features to those of the new annotation, e.g. *bedtools -loj -a newannotation_cds.bed*
806 *-b ncbi_cds.bed > loj_overlaps.bed.* We then evaluate each gene in the new annotation relative
807 to the first two classes of fusions using a custom python script,

808 BuildProteinCodingGeneCdsFusionSummaryTable.py and including the -filter-nested switch
809 that excludes from calculations known fusions that are present in the NCBI annotation. We
810 calculate the frequency of all three classes of fusions using basic awk commaxnds. Our
811 bedtools operation does not enforce same-strand matching, but our python script does, such
812 that we do not consider overlaps between new predictions and NCBI predictions on opposite
813 strands.
814
815 *Undetected CDS in UTRs*
816 Our RNA-seq assembly pipelines integrate ORF finding with Transdecoder such that exon
817 features are decomposed further into CDS and UTR features. While we exclude UTRs to make
818 assembler performance metrics comparable to tools that do not predict UTRs, we observed a
819 substantial drop-off in RNA-seq read alignment rates to the Stringtie and scallop annotations
820 when UTRs were filtered out: unfiltered annotations had much higher alignment rates than all
821 other methods, but were on par with those methods after excluding UTR intervals. To examine
822 the cause of this phenomenon, we considered three possible causes. First, because for
823 relatively complete high-quality annotations the NCBI annotation pipeline will computationally
824 truncate UTRs to prevent stop-codon readthrough, we contrasted the length distributions RNA-
825 seq assembler UTRs and those from NCBI, expecting that the disparity would be greater for the
826 reference genomes for each of our taxonomic groups than for other, more recent genome
827 assemblies for which truncation would not be as severe. Under this scenario, the reduction in
828 alignment rate after UTR removal would be due to an excess of reads originating from
829 transcriptional readthrough (or because the NCBI UTR truncation was overly conservative).
830 Next, we considered the possibility that Transdecoder consistently fails to predict CDS orfs at
831 the terminal ends of transcripts, such that a large proportion of real CDS sequences is being
832 incorrectly filtered out when we strip out UTRs. To test this hypothesis, we extracted the UTR

833 sequences from the Stringtie and scallop annotations and used blastx (Camacho et al. 2009) v.
834 2.12.0 to search for matches against the NCBI protein sequences from the same species' NCBI
835 accession, with a maximum e-value of 1 x 10^{-5}. We calculated the fraction of transcripts for
836 which at least one of the UTRs had a hit to the protein database.
837
838 *Mulit-method integration*
839 To determine whether integration of multiple annotation methods could harness the
840 complementary strengths of individual methods, we tested an approach where we 1) set one
841 annotation as the "base annotation", 2) add features from a second annotation methods that fall
842 entirely outside of the gene intervals for the base annotation, and 3) iterate this process for
843 additional methods, e.g. adding features from a third method that do not overlap genes from the
844 integration of the first two annotations. We demonstrated this approach for *Homo sapiens*,
845 investigating the effects of different choices for base annotation, exclusion of RNA-seq
846 assembler or TOGA, and whether there are decreasing returns with increasing number of
847 integrated methods. For these comparisons we focused on four metrics: BUSCO score, RNA-
848 seq alignment rate, the total number of genes with BLASTP hits to NCBI reference proteins, and
849 the percentage of all integrated genes that have such hits. While we do not compare this
850 approach directly to the other integration methods evaluated here (MAKER and TSEBRA,) our
851 results demonstrate that both of these annotation tools under-perform relative to other
852 standalone methods.
853 **REFERENCES**

854 Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li
855 PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of Drosophila
856 melanogaster. *Science* **287**: 2185–2195.

857 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J*
858 *Mol Biol* **215**: 403–410.

859 Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S,
860 Stiller J, et al. 2020. Progressive Cactus is a multiple-genome aligner for the thousand-
861 genome era. *Nature* **587**: 246–251.

862 Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic
863 eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a
864 protein database. *NAR Genomics Bioinforma* **3**: lqaa108.

865 Brůna T, Lomsadze A, Borodovsky M. 2020. GeneMark-EP+: eukaryotic gene prediction with
866 self-training in the space of genes and proteins. *NAR Genomics Bioinforma* **2**: lqaa026.

867 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
868 BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

869 Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M.
870 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model
871 organism genomes. *Genome Res* **18**: 188–196.

872 Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff
873 F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational
874 molecular biology and bioinformatics. *Bioinforma Oxf Engl* **25**: 1422–1423.

875 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras
876 TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

877 Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, Dikow RB, García-Accinelli G, Van
878 Belleghem SM, Patterson N, Neafsey DE, et al. 2019. Genomic architecture and
879 introgression shape a butterfly radiation. *Science* **366**: 594–599.

880 Freedman AH, Clamp M, Sackton TB. 2021. Error, noise and bias in de novo transcriptome
881 assemblies. *Mol Ecol Resour* **21**: 18–29.

882 Gabriel L, Hoff KJ, Brůna T, Borodovsky M, Stanke M. 2021a. TSEBRA: transcript selector for
883 BRAKER. *BMC Bioinformatics* **22**: 566.

884 Gabriel L, Hoff KJ, Brůna T, Borodovsky M, Stanke M. 2021b. TSEBRA: transcript selector for
885 BRAKER. *BMC Bioinformatics* **22**: 566.

886 Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR.
887 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the
888 Program to Assemble Spliced Alignments. *Genome Biol* **9**: R7.

889 Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey
890 TS, Harte RA, Hsu F, et al. 2006. The UCSC Genome Browser Database: update 2006.
891 *Nucleic Acids Res* **34**: D590-598.

892 Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: Unsupervised
893 RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinforma*
894 *Oxf Engl* **32**: 767–769.

895 Hoff KJ, Stanke M. 2019. Predicting Genes in Single Genomes with AUGUSTUS. *Curr Protoc*
896 *Bioinforma* **65**: e57.

897 Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT,
898 Fiddes IT, et al. 2018. Nanopore sequencing and assembly of a human genome with
899 ultra-long reads. *Nat Biotechnol* **36**: 338–345.

900 Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and
901 genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907–915.

902 Kirilenko BM, Munegowda C, Osipova E, Jebb D, Sharma V, Blumer M, Morales AE, Ahmed A-
903 W, Kontopoulos D-G, Hilgers L, et al. 2023. Integrating gene annotation with orthology
904 inference at scale. *Science* **380**: eabn3107.

905 König S, Romoth LW, Gerischer L, Stanke M. 2016. Simultaneous gene finding in multiple
906 genomes. *Bioinformatics* **32**: 3388–3395.

907 Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**: 59.

908 Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. 2019.
909 OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral
910 genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* **47**:
911 D807–D811.

912 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M,
913 FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature*
914 **409**: 860–921.

915 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**:
916 357–359.

917 Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or
918 without a reference genome. *BMC Bioinformatics* **12**: 323.

919 Lomsadze A, Burns PD, Borodovsky M. 2014. Integration of mapped RNA-Seq reads into
920 automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* **42**: e119.

921 Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in
922 novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* **33**: 6494–6506.

923 Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J,
924 Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, et al. 2002. Initial
925 sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.

926 Nachtweide S, Stanke M. 2019. Multi-Genome Annotation with AUGUSTUS. *Methods Mol Biol*
927 *Clifton NJ* **1962**: 139–160.

928 Niknafs YS, Pandian B, Iyer HK, Chinnaiyan AM, Iyer MK. 2017. TACO produces robust multi-
929 sample transcriptome assemblies from RNA-seq. *Nat Methods* **14**: 68–70.

930 Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie
931 enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*
932 **33**: 290–295.

933 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
934 features. *Bioinformatics* **26**: 841–842.

935 Rhesus Macaque Genome Sequencing and Analysis Consortium, Gibbs RA, Rogers J, Katze
936 MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter
937 JC, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome.
938 *Science* **316**: 222–234.

939 Shao M, Kingsford C. 2017. Accurate assembly of transcripts through phase-preserving graph
940 decomposition. *Nat Biotechnol* **35**: 1167–1169.

941 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO:
942 assessing genome assembly and annotation completeness with single-copy orthologs.
943 *Bioinforma Oxf Engl* **31**: 3210–3212.

944 Song B, Buckler ES, Stitzer MC. 2024. New whole-genome alignment tools are needed for
945 tapping into plant diversity. *Trends Plant Sci* **29**: 355–369.

946 Song L, Sabunciyan S, Yang G, Florea L. 2019. A multi-sample approach increases the
947 accuracy of transcript assembly. *Nat Commun* **10**: 5000.

948 Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio
949 prediction of alternative transcripts. *Nucleic Acids Res* **34**: W435-439.

950 Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron
951 submodel. *Bioinforma Oxf Engl* **19 Suppl 2**: ii215-225.

952 Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ,
953 Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals
954 unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*
955 **28**: 511–515.

956 Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. 2018. Towards a
957 complete map of the human long non-coding RNA transcriptome. *Nat Rev Genet* **19**:
958 535–548.

959 Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans
960 CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* **291**: 1304–
961 1351.

962 Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J,
963 Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus
964 long-read sequencing improves variant detection and assembly of a human genome.
965 *Nat Biotechnol* **37**: 1155–1162.

966 Williams CR, Baccarella A, Parrish JZ, Kim CC. 2016. Trimming of sequence reads alters RNA-
967 Seq gene expression estimates. *BMC Bioinformatics* **17**: 103.

968 Zhang Q, Shi Q, Shao M. 2022. Accurate assembly of multi-end RNA-seq data with Scallop2.
969 *Nat Comput Sci* **2**: 148–152.

970

971 **DATA ACCESS**

972 Detailed explanation of steps in annotation pipelines, along with associated python scripts for
973 data processing are provided in several repositories listed below

974 - RNA-seq transcript assembly: https://github.com/harvardinformatics/GenomeAnnotation-
975 RNAseqAssembly/
976 - BRAKER: https://github.com/harvardinformatics/GenomeAnnotation-Braker
977 - TOGA: https://github.com/harvardinformatics/GenomeAnnotation-TOGA
978 - CGP: https://github.com/harvardinformatics/GenomeAnnotation-ComparativeAugustus
979 - MAKER: https://github.com/harvardinformatics/GenomeAnnotation-MAKER

980 In addition, HPC slurm job scripts and specific command lines used to run annotation tools in
981 this paper, as well as python scripts to generate annotation quality metrics are available at
982 https://github.com/harvardinformatics/GenomeAnnotation.

Building Better Genome Annotations Across the Tree of Life

Adam H. Freedman[1]*, Timothy B. Sackton[1]

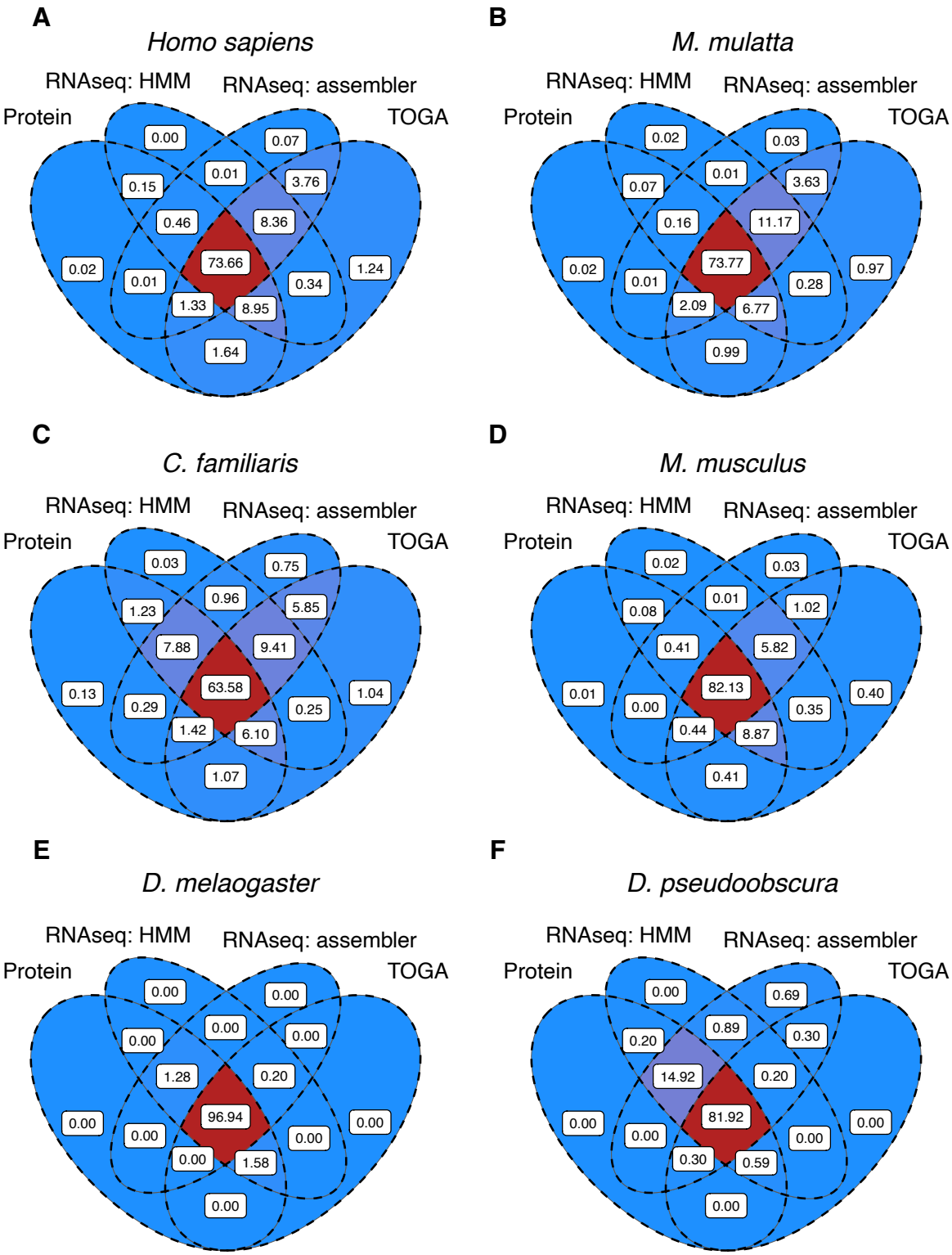
[1] Faculty of Arts and Sciences, Informatics Group, Harvard University, 52 Oxford Street, Cambridge, MA 02138

* Corresponding author: adamfreedman@fas.harvard.edu

Supplementary Materials

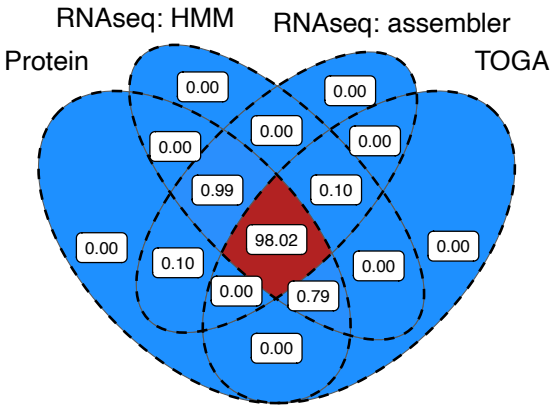
Figure Page

Supplementary Figure S1 ……1
Supplementary Figure S2 ……4
Supplementary Figure S3 ……8
Supplementary Figure S4 ……10
Supplementary Figure S5 ……11
Supplementary Figure S6 ……12
Supplementary Figure S7 ……13
Supplementary Figure S8 ……14
Supplementary Figure S9 ……15
Supplementary Figure S10 ……16
Supplementary Figure S11 ……17
Supplementary Figure S12 ……19
Supplementary Figure S13 ……20
Supplementary Figure S14 ……21
Supplementary Figure S15 ……22
Supplementary Figure S16 ……23
Supplementary Figure S17 ……24
Supplementary Figure S18 ……25
Supplementary Figure S19 ……26


A
Homo sapiens
RNAseq: HMM
RNAseq: assembler
Protein
TOGA
0.00
0.07
0.15
0.01
3.76
0.46
8.36
73.66
0.02
0.01
0.34
1.24
1.33
8.95
1.64
B
M. mulatta
RNAseq: HMM
RNAseq: assembler
Protein
TOGA
0.02
0.03
0.07
0.01
3.63
0.16
11.17
73.77
0.02
0.01
0.28
0.97
2.09
6.77
0.99
C
C. familiaris
RNAseq: HMM
RNAseq: assembler
Protein
TOGA
0.03
0.75
1.23
0.96
5.85
7.88
9.41
63.58
0.13
0.29
0.25
1.04
1.42
6.10
1.07
D
M. musculus
RNAseq: HMM
RNAseq: assembler
Protein
TOGA
0.02
0.03
0.08
0.01
1.02
0.41
5.82
82.13
0.01
0.00
0.35
0.40
0.44
8.87
0.41
E
D. melaogaster
RNAseq: HMM
RNAseq: assembler
Protein
TOGA
0.00
0.00
0.00
0.00
0.00
1.28
0.20
96.94
0.00
0.00
0.00
0.00
0.00
1.58
0.00
F
D. pseudoobscura
RNAseq: HMM
RNAseq: assembler
Protein
TOGA
0.00
0.69
0.89
0.20
0.30
14.92
0.20
81.92
0.00
0.00
0.00
0.00
0.30
0.59
0.00

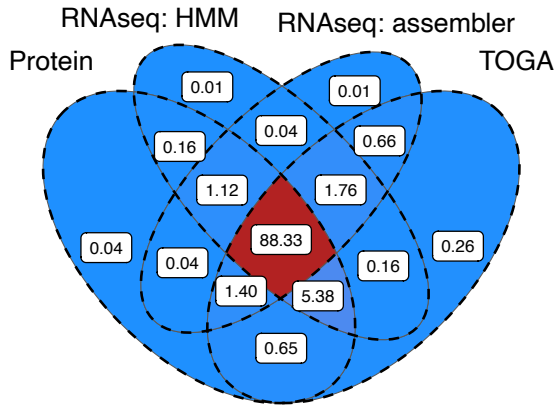
G

D. yakuba



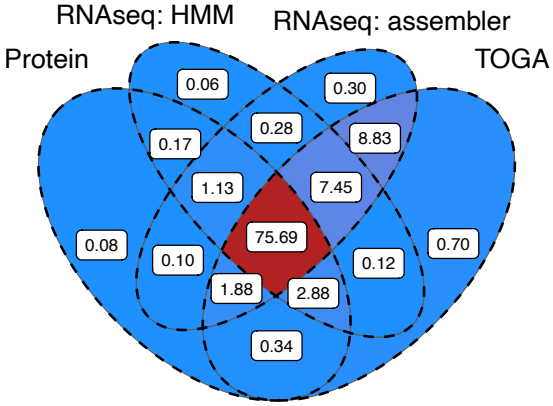
H

G. gallus



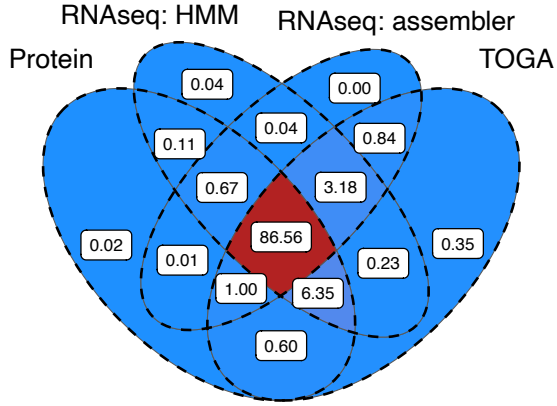
I

A. platyrhynchos



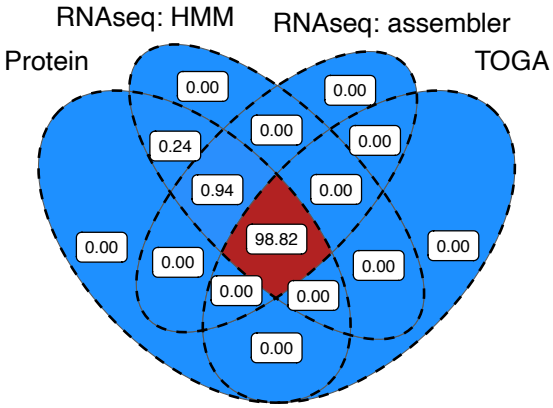
J

C. japonica



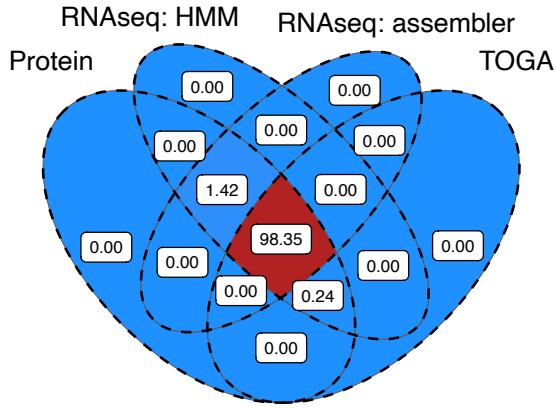
K

A. lyrata



L

A. thaliana



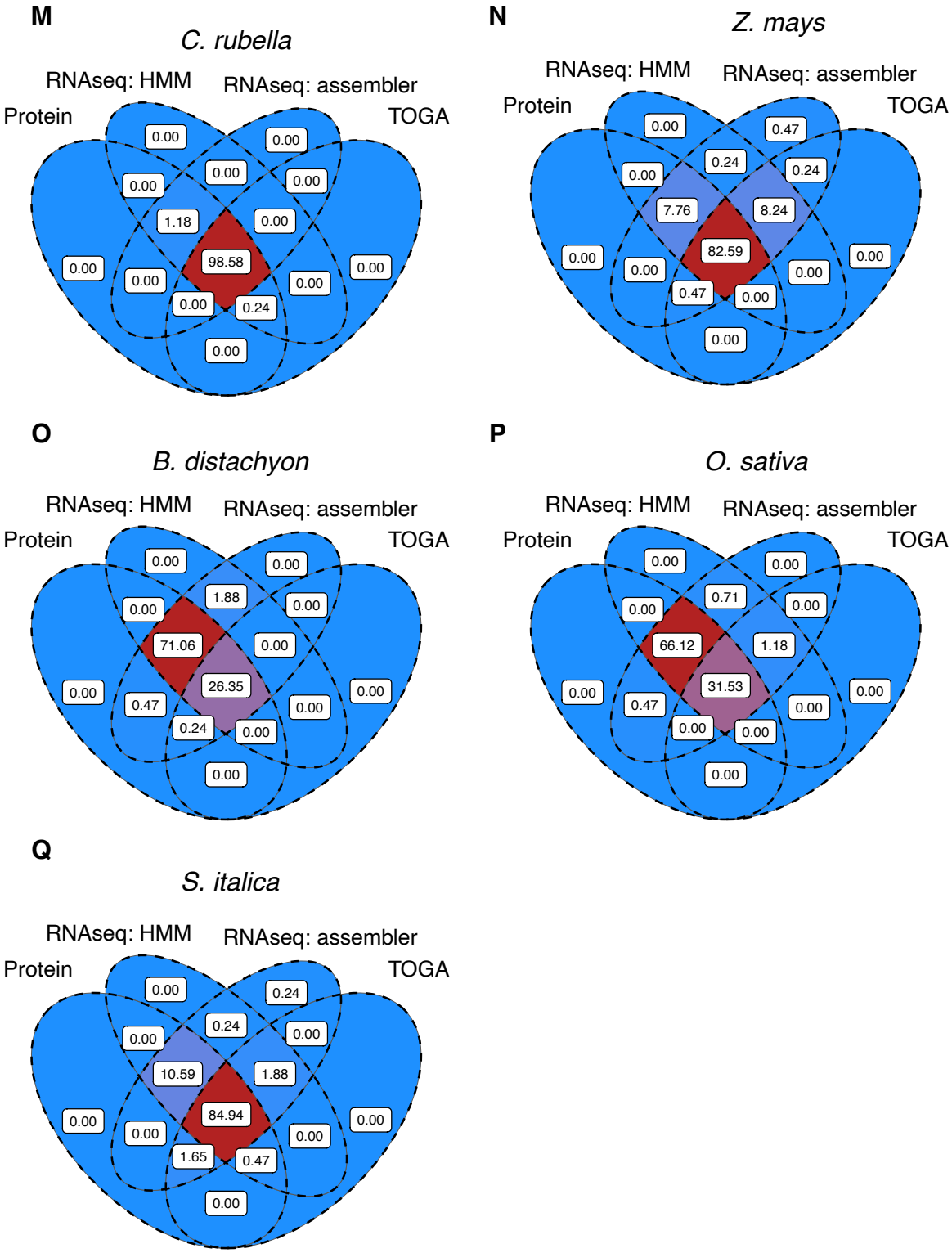
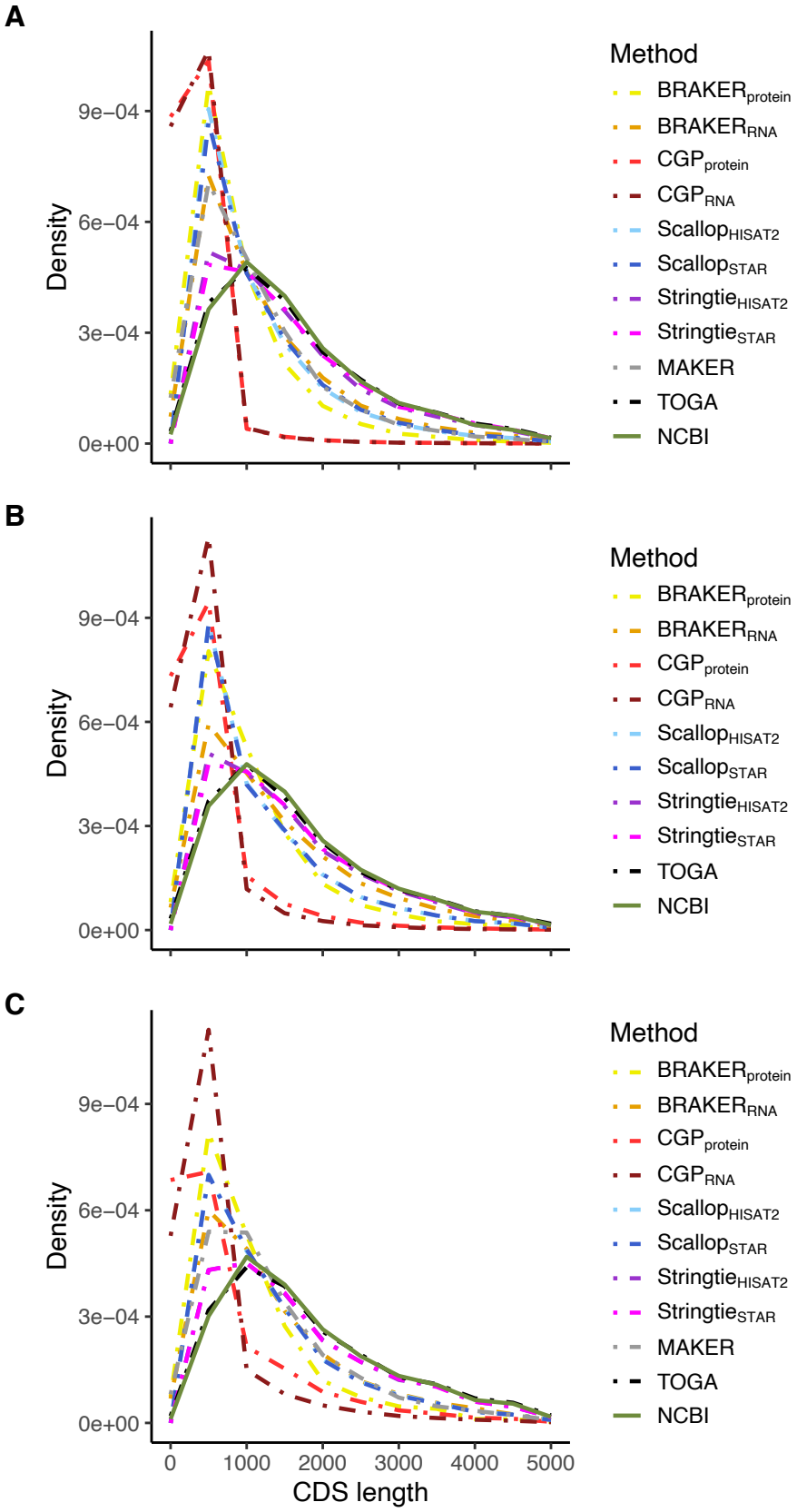
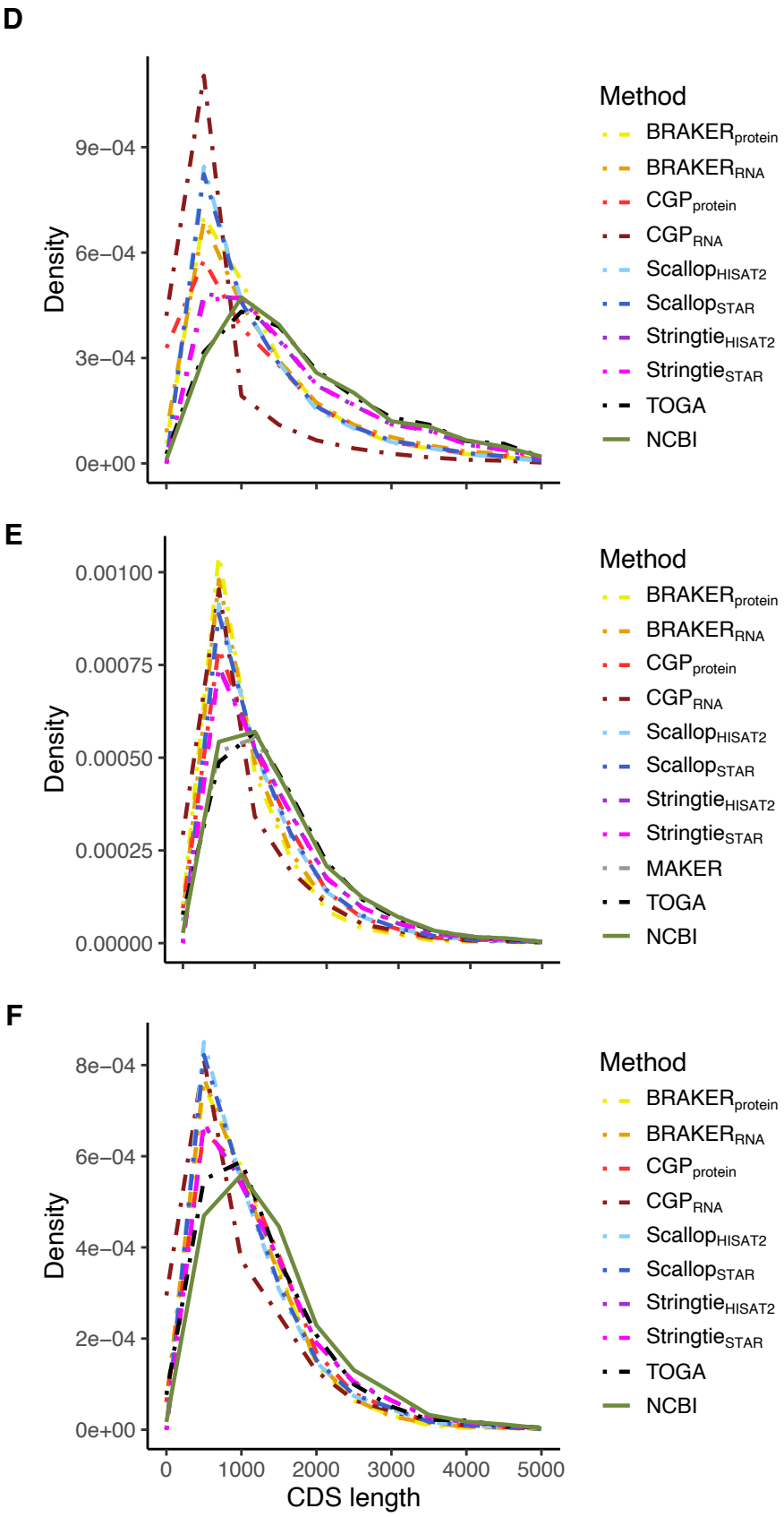
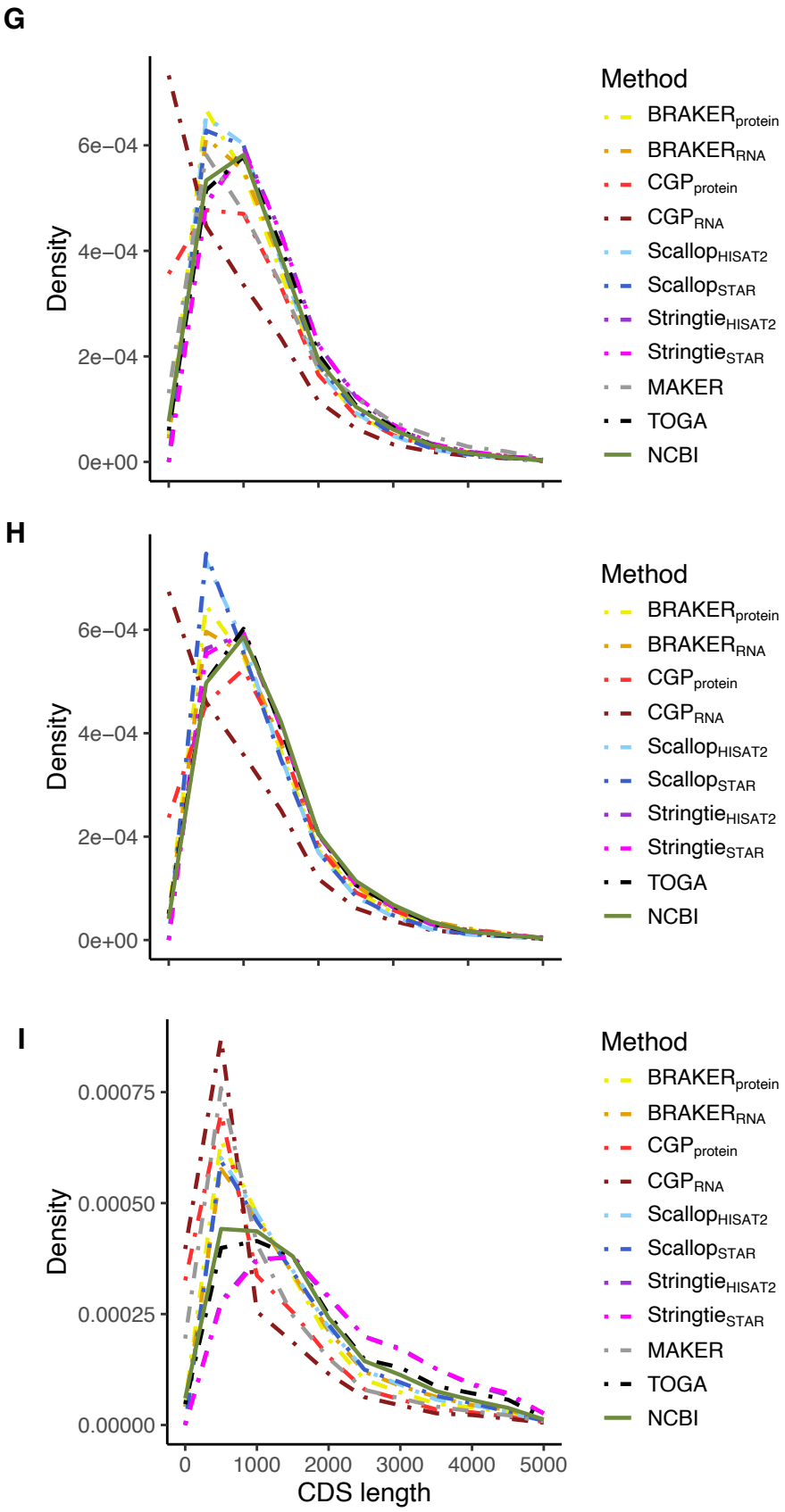


Figure S1. Percent of recovered BUSCOs that overlap between methods for species in addition to Figure 2. (A-D) mammals, (E-G) dipterans, (H-J) birds, (K-M) rosids, and (N-Q) monocots.


A
B
C
Density
9e−04
6e−04
3e−04
0e+00
0
1000
2000
3000
4000
5000
CDS length
Method
BRAKER protein
BRAKER RNA
CGP protein
CGP RNA
Scallop HISAT2
Scallop STAR
Stringtie HISAT2
Stringtie STAR
MAKER
TOGA
NCBI


D
Density
9e−04
6e−04
3e−04
0e+00
Method
BRAKER protein
BRAKER RNA
CGP protein
CGP RNA
Scallop HISAT2
Scallop STAR
Stringtie HISAT2
Stringtie STAR
TOGA
NCBI
E
Density
0.00100
0.00075
0.00050
0.00025
0.00000
Method
BRAKER protein
BRAKER RNA
CGP protein
CGP RNA
Scallop HISAT2
Scallop STAR
Stringtie HISAT2
Stringtie STAR
MAKER
TOGA
NCBI
F
Density
8e−04
6e−04
4e−04
2e−04
0e+00
0
1000
2000
3000
4000
5000
CDS length
Method
BRAKER protein
BRAKER RNA
CGP protein
CGP RNA
Scallop HISAT2
Scallop STAR
Stringtie HISAT2
Stringtie STAR
TOGA
NCBI


G
Density
6e−04
4e−04
2e−04
0e+00
Method
BRAKER protein
BRAKER RNA
CGP protein
CGP RNA
Scallop HISAT2
Scallop STAR
Stringtie HISAT2
Stringtie STAR
MAKER
TOGA
NCBI
H
Density
6e−04
4e−04
2e−04
0e+00
Method
BRAKER protein
BRAKER RNA
CGP protein
CGP RNA
Scallop HISAT2
Scallop STAR
Stringtie HISAT2
Stringtie STAR
TOGA
NCBI
I
Density
0.00075
0.00050
0.00025
0.00000
0
1000
2000
3000
4000
5000
CDS length
Method
BRAKER protein
BRAKER RNA
CGP protein
CGP RNA
Scallop HISAT2
Scallop STAR
Stringtie HISAT2
Stringtie STAR
MAKER
TOGA
NCBI

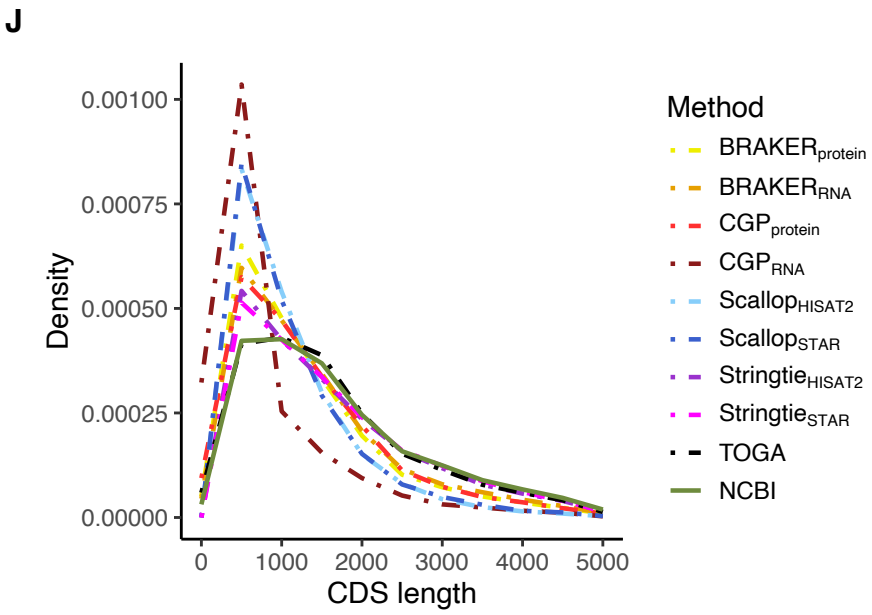
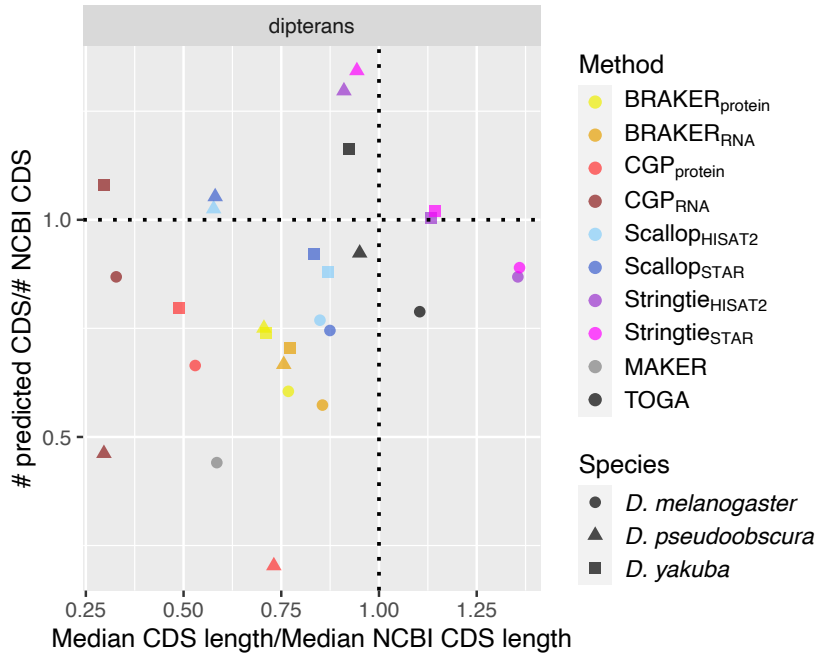
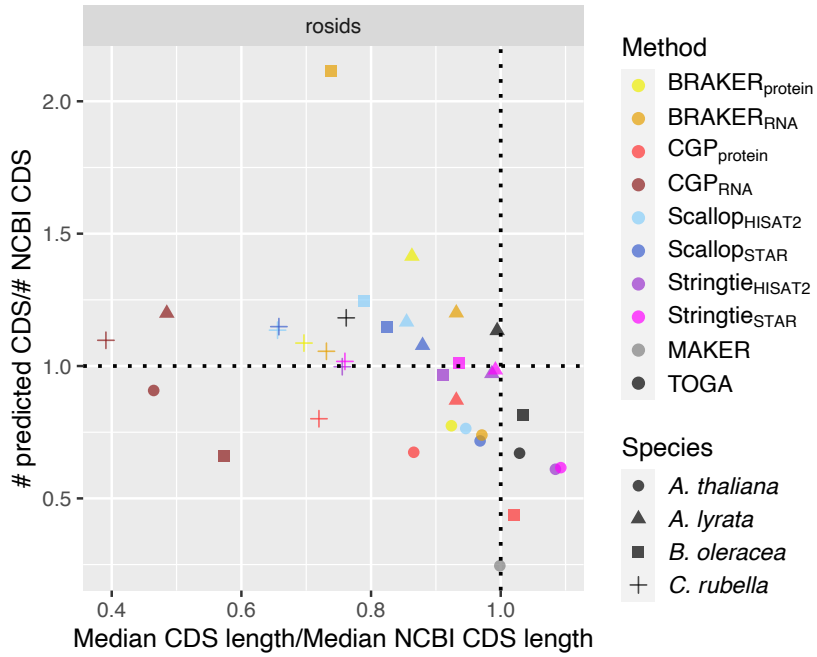


Figure S2. CDS length distributions for annotation methods and NCBI benchmark for (A) *H.* sapiens, (B) C*. familiaris*, (C) *G.* gallus, (D) *A. platyrhynchos*, (E) *Z. mays* (F) *O. sativa*, (G) *A. thaliana*, (H) *C. rubella*, (I) *D. melanogaster*, and (J) *D. pseudoobscura*.

A


dipterans
predicted CDS/# NCBI CDS
1.0
0.5
0.25
0.50
0.75
1.00
1.25
Median CDS length/Median NCBI CDS length
Method
BRAKER protein
BRAKER RNA
CGP protein
CGP RNA
Scallop HISAT2
Scallop STAR
Stringtie HISAT2
Stringtie STAR
MAKER
TOGA
Species
D. melanogaster
D. pseudoobscura
D. yakuba

B


rosids
predicted CDS/# NCBI CDS
2.0
1.5
1.0
0.5
0.4
0.6
0.8
1.0
Median CDS length/Median NCBI CDS length
Method
BRAKER protein
BRAKER RNA
CGP protein
CGP RNA
Scallop HISAT2
Scallop STAR
Stringtie HISAT2
Stringtie STAR
MAKER
TOGA
Species
A. thaliana
A. lyrata
B. oleracea
C. rubella

C

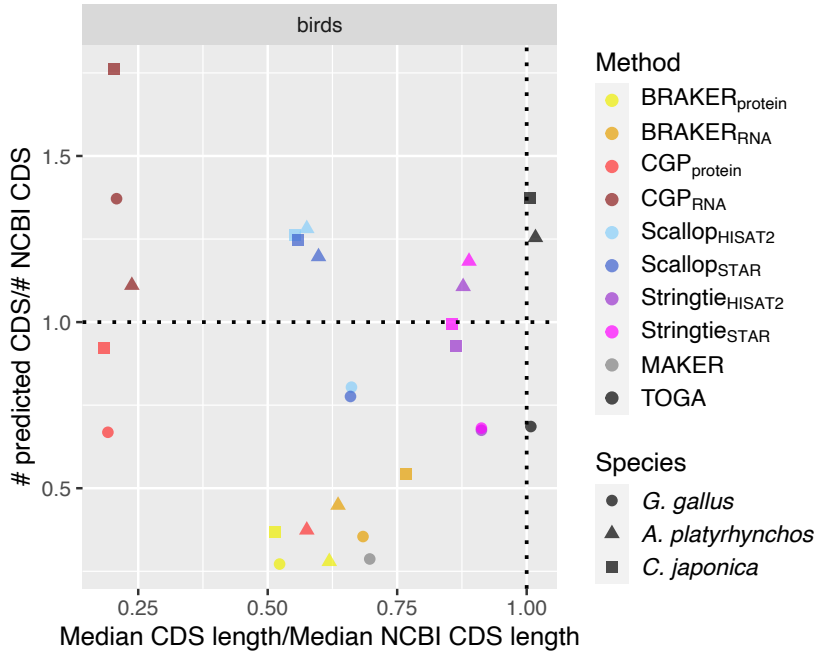


Figure S3. Joint distributions of number of predicted CDS (normalized by number of NCBI predictions) over median predicted CDS length (normalized by median NCBI CDS length) for (A) dipterans, (B) rosids, and (C) birds. Dotted lines indicated equivalence to NCBI annotation, such that methods that are closest to the intersection of those lines best approximate CDS length and number of NCBI annotations.

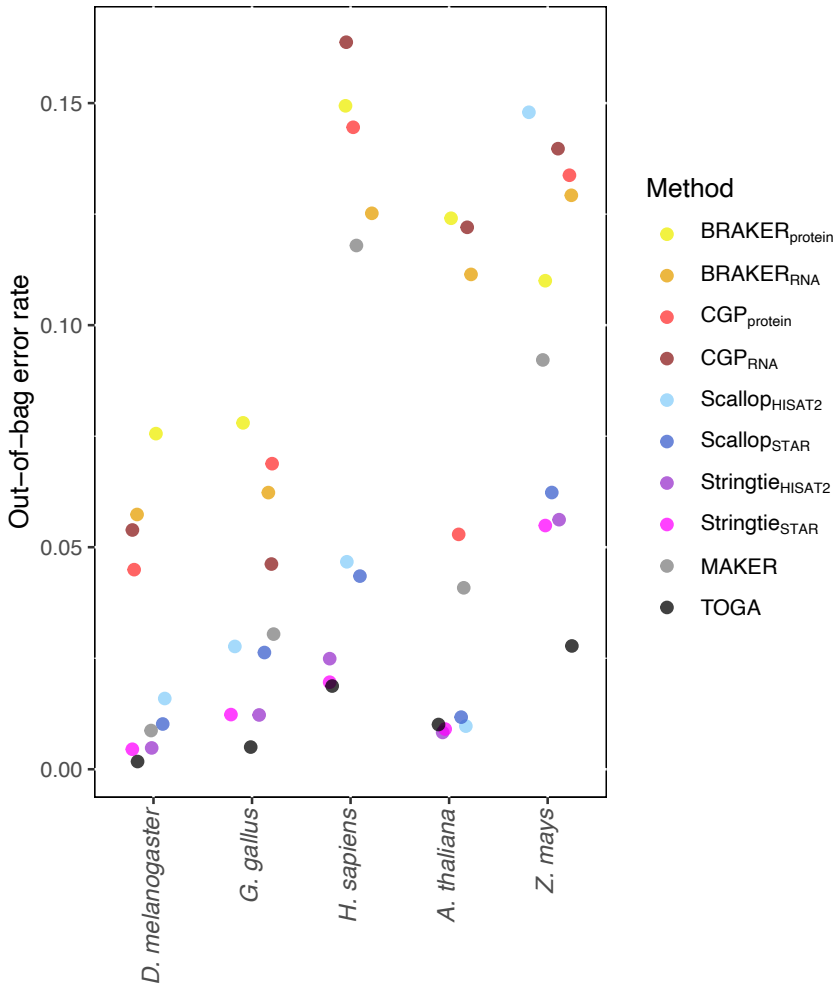


Figure S4. Out-of-bag error rates for random forest models classifying CDS transcripts as either overlapping NCBI gene intervals, or falling outside of those intervals, i.e. intergenic. Models are generated for the reference species for each taxonomic group for which annotations are thought to be complete or nearly so.

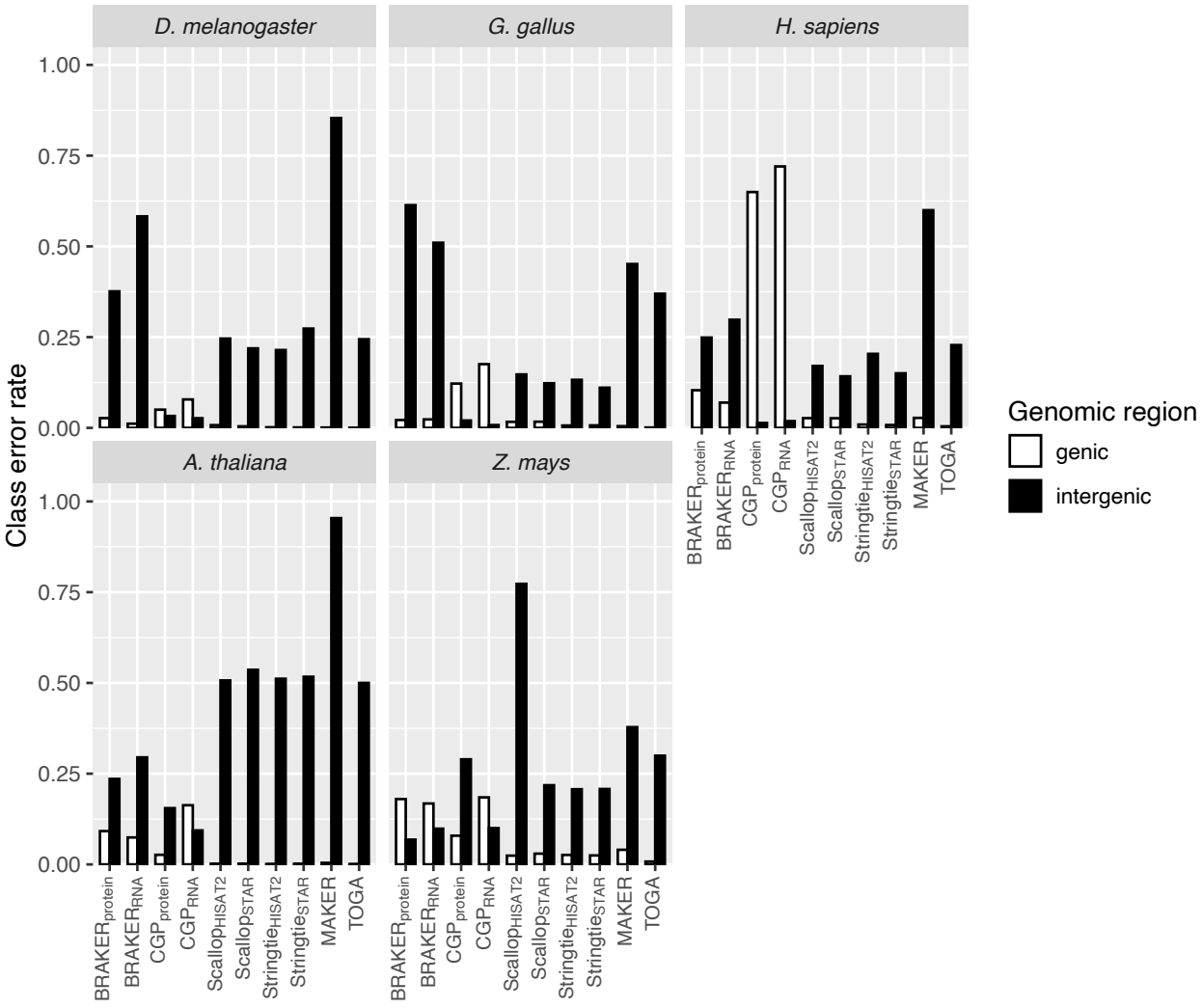


Figure S5. For reference species, out-of-bag error rates broken down by predicted class, for random forest predictions of CDS overlapping NCBI protein-coding genes, vs. those entirely outside of NCBI protein coding gene intervals.

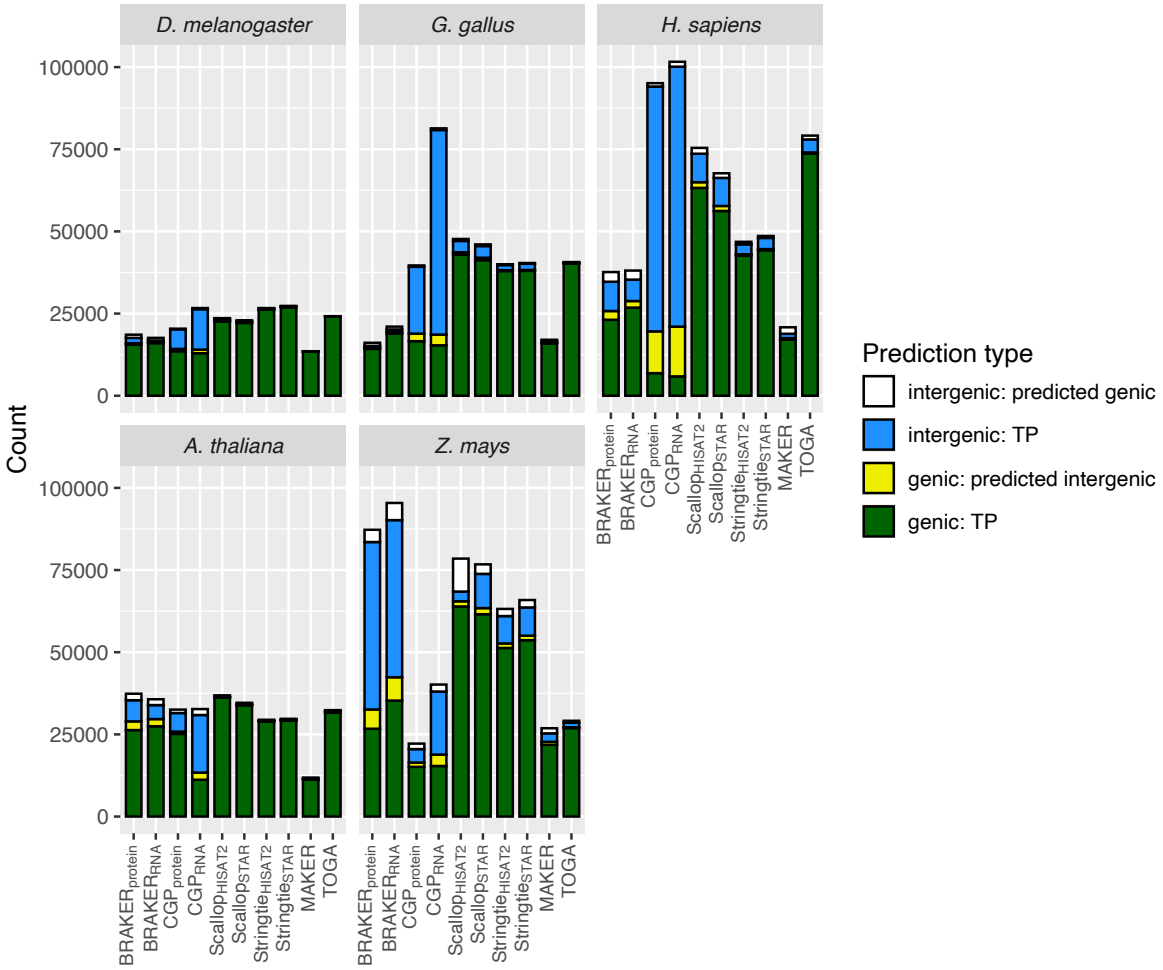


Figure S6. Random forest confusion matrices for reference species, for prediction accuracy of models predicting the genic versus intergenic status of CDS transcripts based upon transcript features.

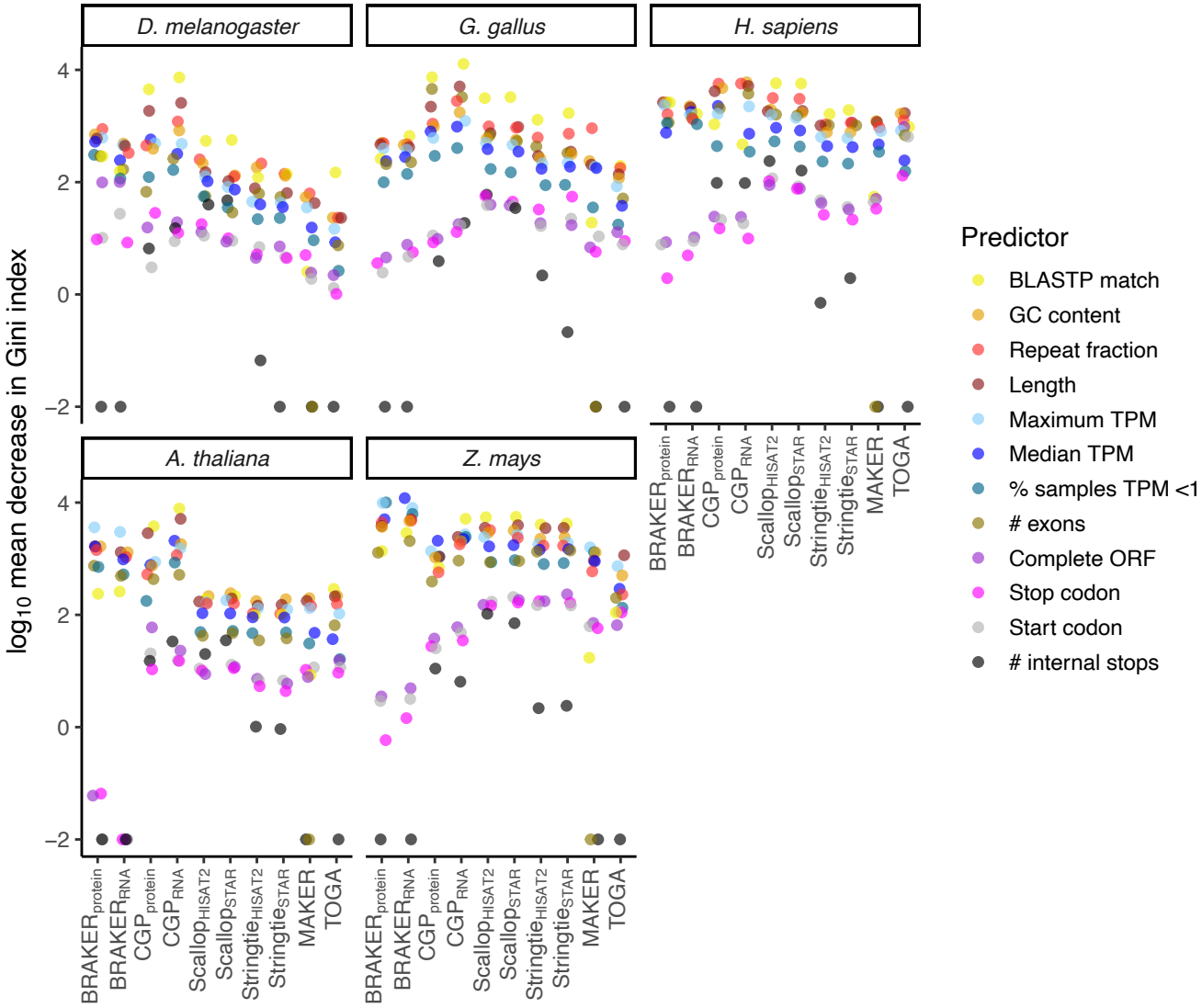


Figure S7. Decrease in Gini index, an estimator of variable importance for random forest predictions of intergenic versus genic region location of predicted transcripts for reference species.

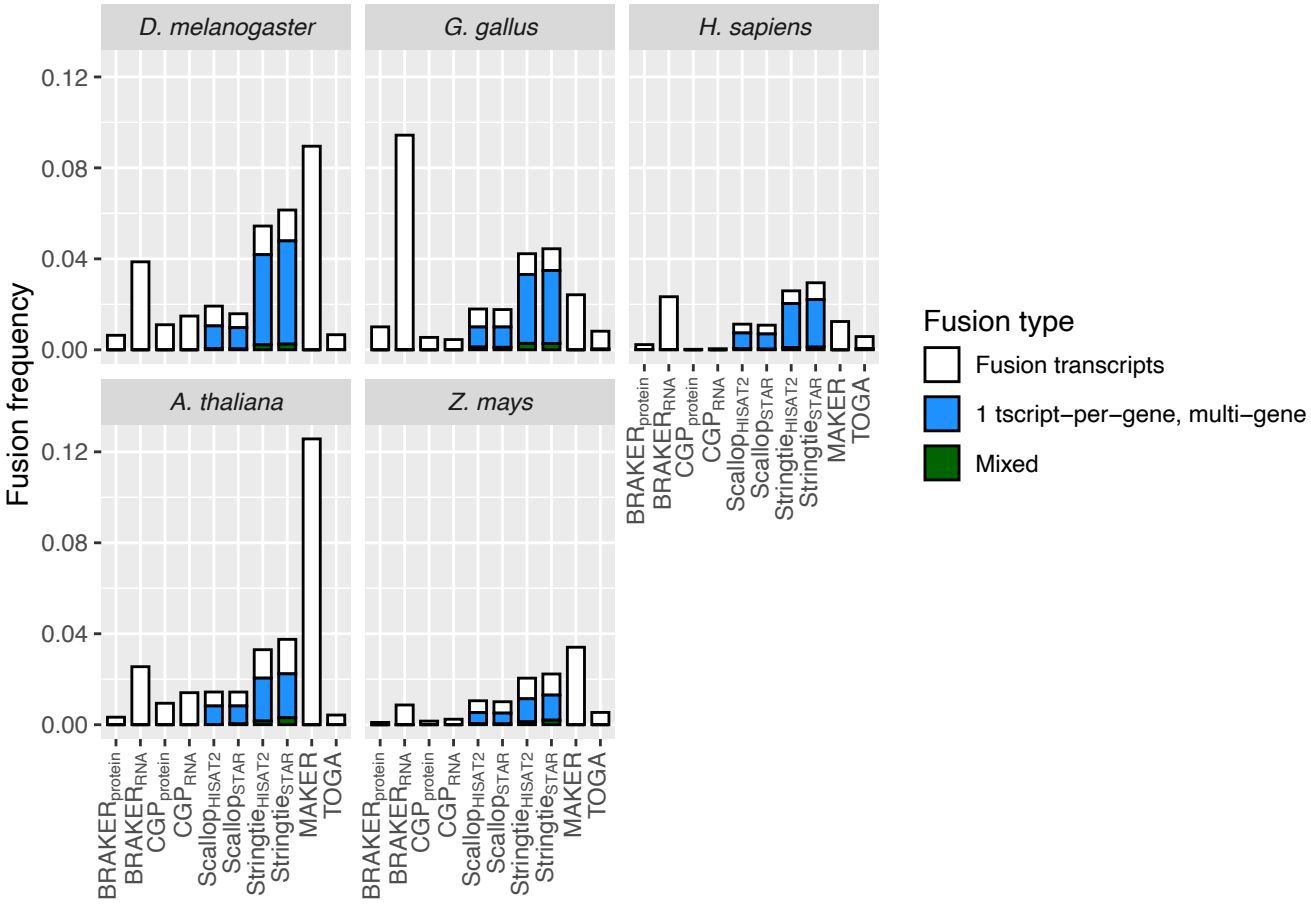


Figure S8. Frequency of gene fusions, broken down by fusion type for five reference species Fusions are defined as when individual transcripts overlap the CDS of multiple NCBI genes (fusion transcripts), different transcripts from the same predicted gene each have overlaps to different genes (1 tscript-per-gene, multi-gene), or a combination of both of these (mixed). Frequencies are calculated after filtering out NCBI-annotated fusion events.

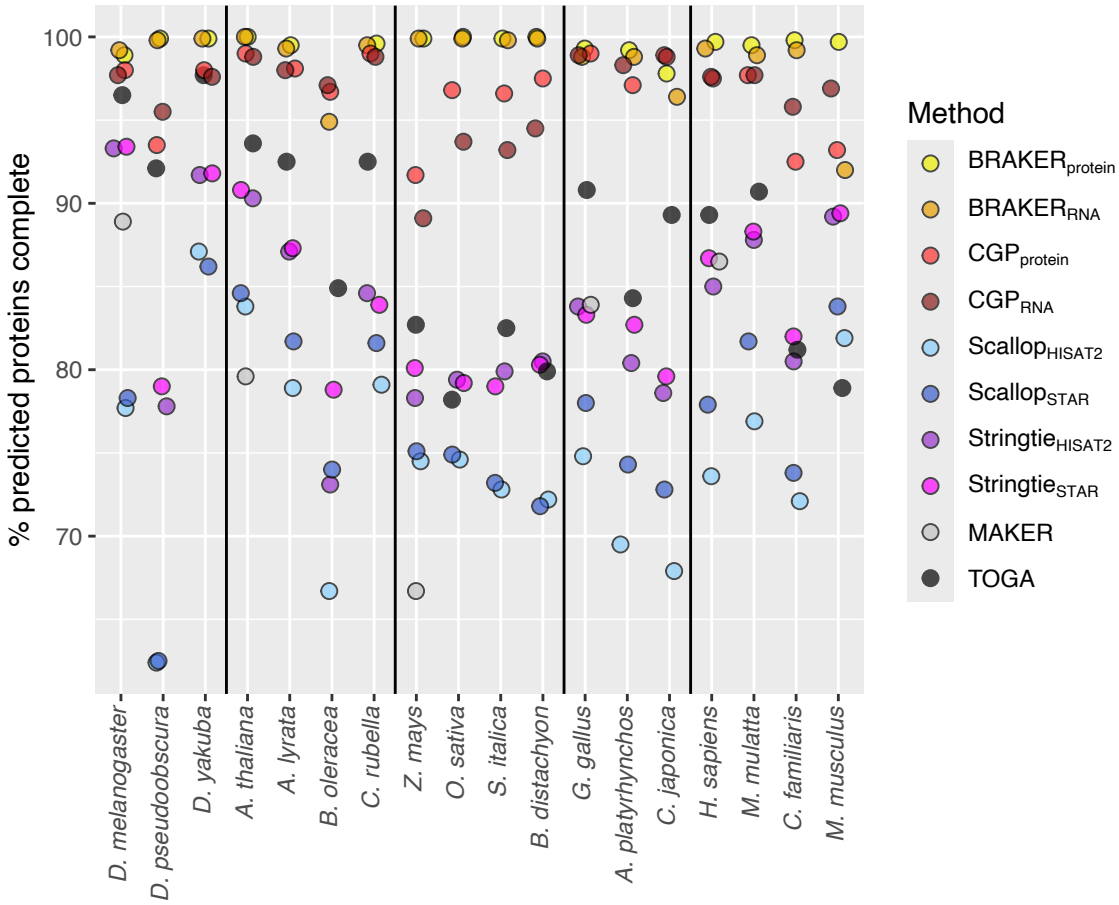


Figure S9. Grouped by species and method, percentage of predicted proteins that are complete, having both initiating start and terminating stop codons, and no internal stop codons.

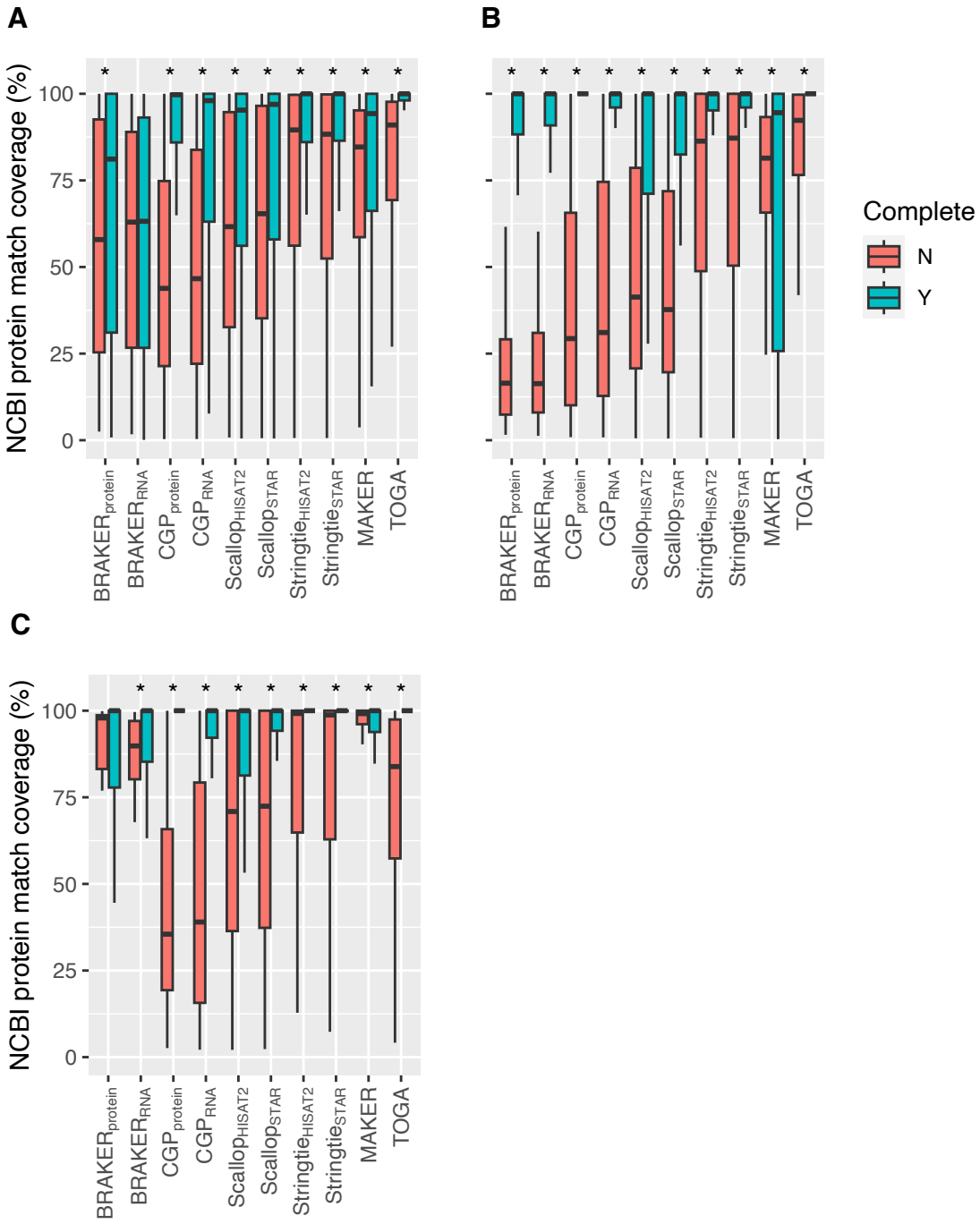
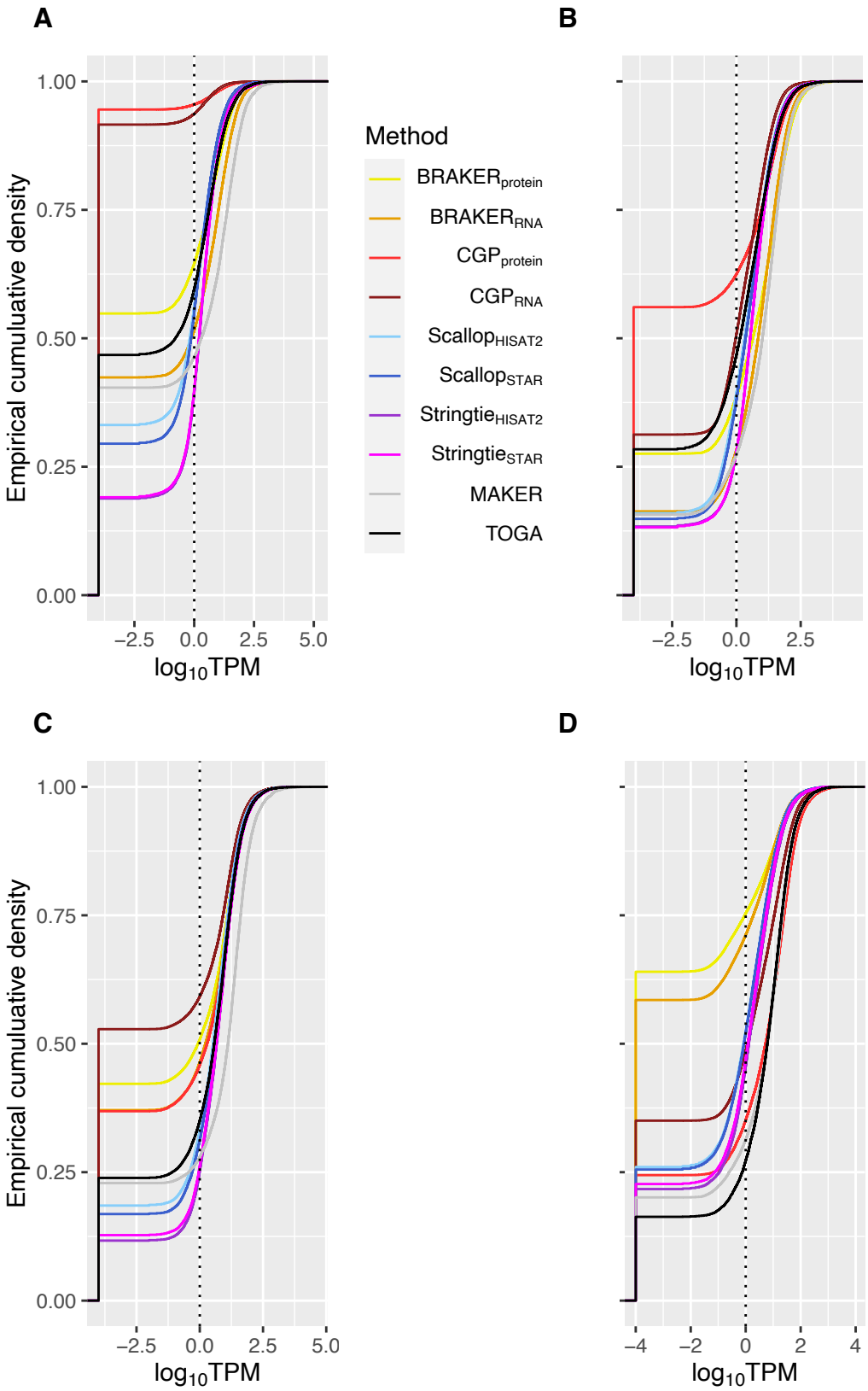


Figure S10. For predicted proteins with BLASTP hits, proportion of NCBI best hit target covered by amino acid matches with the predicted protein for (A) *G. gallus*, (B) *D. melanogaster*, and (C) *A. thaliana,* broken into proteins that are complete (start and stop codons present, no internal stop codons), and those that are not. Benjamini-Hochberg adjusted p-values for Wilcoxon rank-sum tests, p≤0.05 indicated with *.


A
B
C
D
Empirical cumulative density
1.00
0.75
0.50
0.25
0.00
−2.5
0.0
2.5
5.0
−4
−2
0
2
4
log10TPM
Method
BRAKERprotein
BRAKERRNA
CGPprotein
CGPRNA
ScallopHISAT2
ScallopSTAR
StringtieHISAT2
StringtieSTAR
MAKER
TOGA

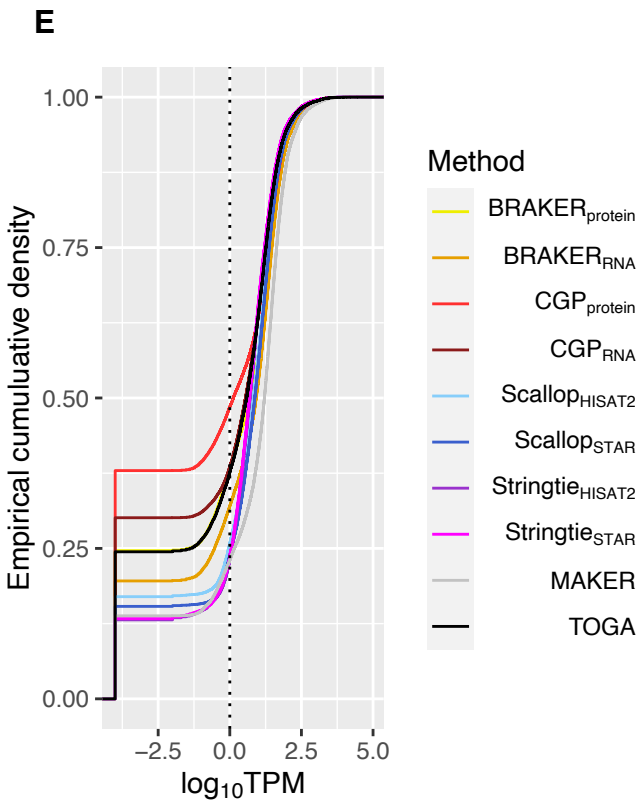


Figure S11. Empirical cumulative distributions of TPM for (A) *H. sapiens*, (B) *G. gallus*, (C) *A. thaliana*, (D) *Z. mays*, and (E) *D. melanogaster.* 0.0001 is added to all TPM values before log-transformation. Dashed line indicates TPM of 1.

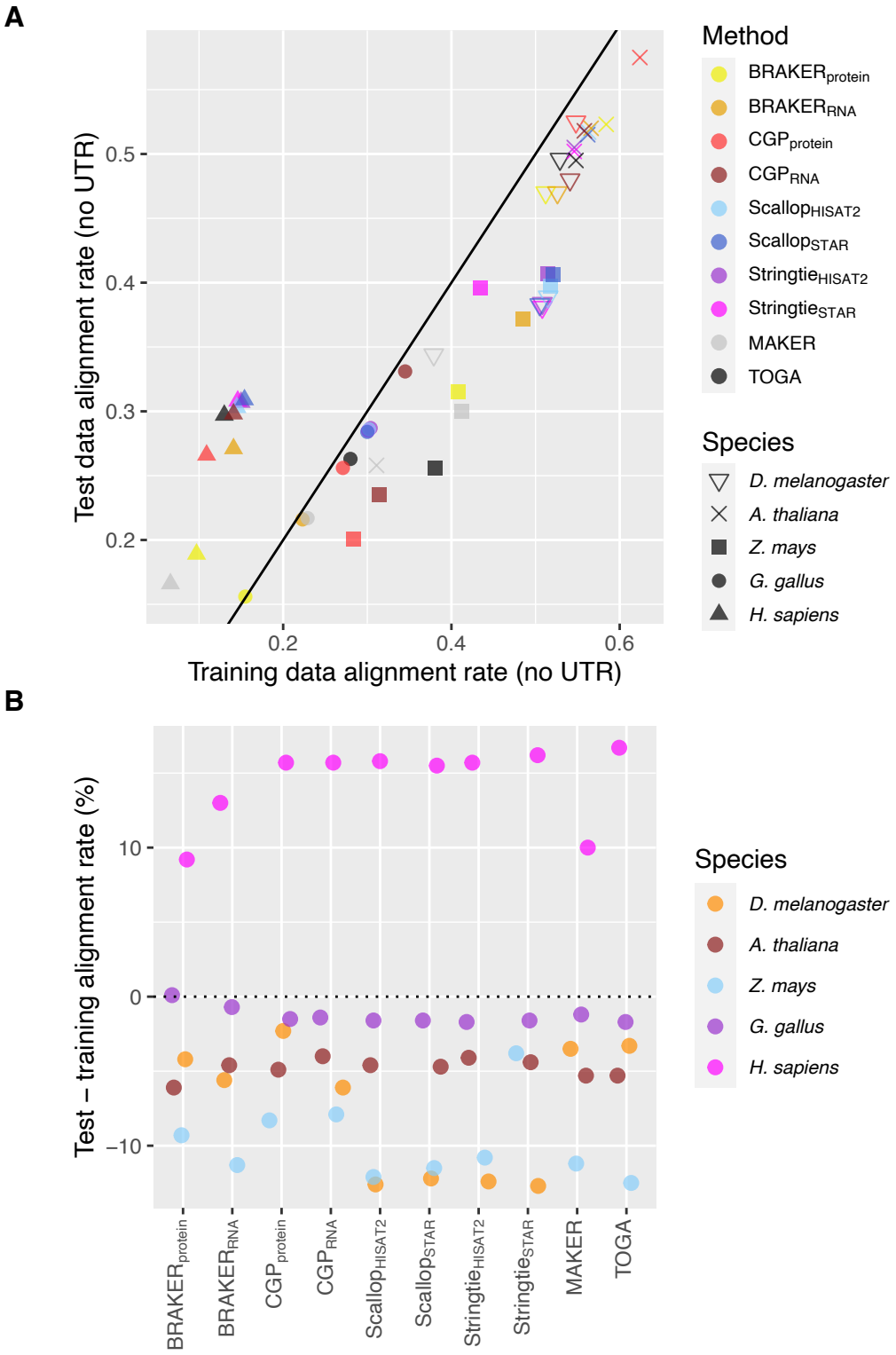


Figure S12. Comparison of RNA-seq alignment rates between training data used in annotations, and test data for the five reference species. (A) Pairwise plots of test again training rates, and (B) Rate differences by method a species.

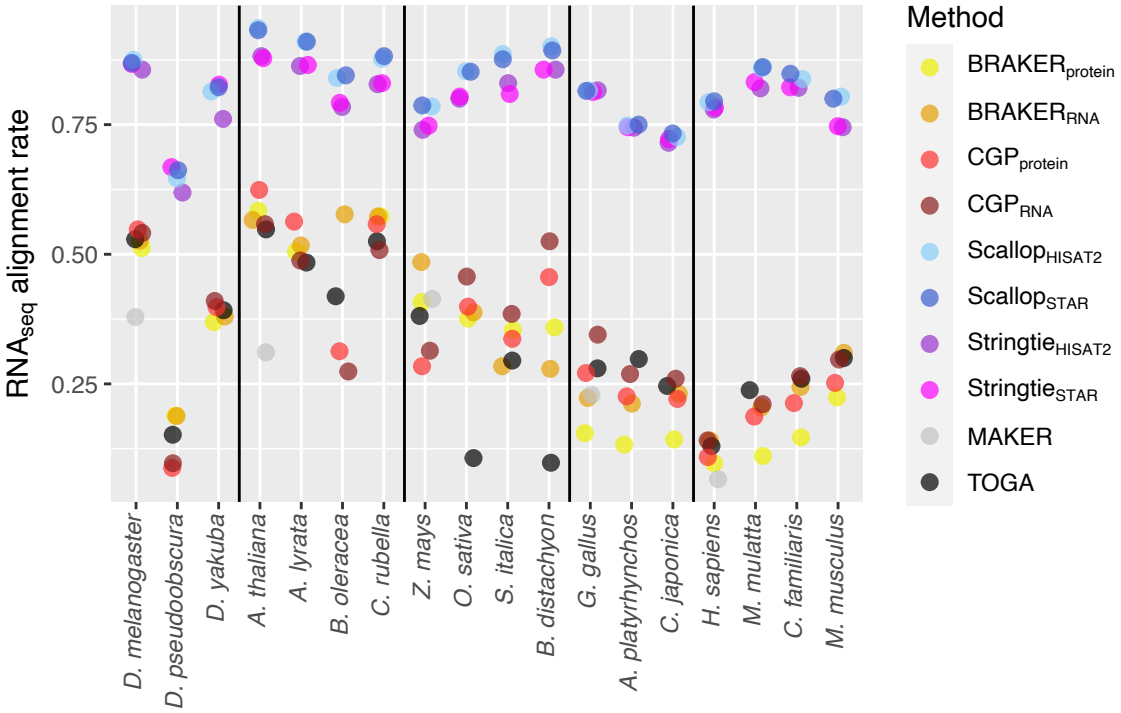


Figure S13. By method and species, RNA-seq read alignment rates to predicted transcripts, analogous to Figure 11, but for Stringtie and Scallop UTR intervals are included in the transcript predictions.

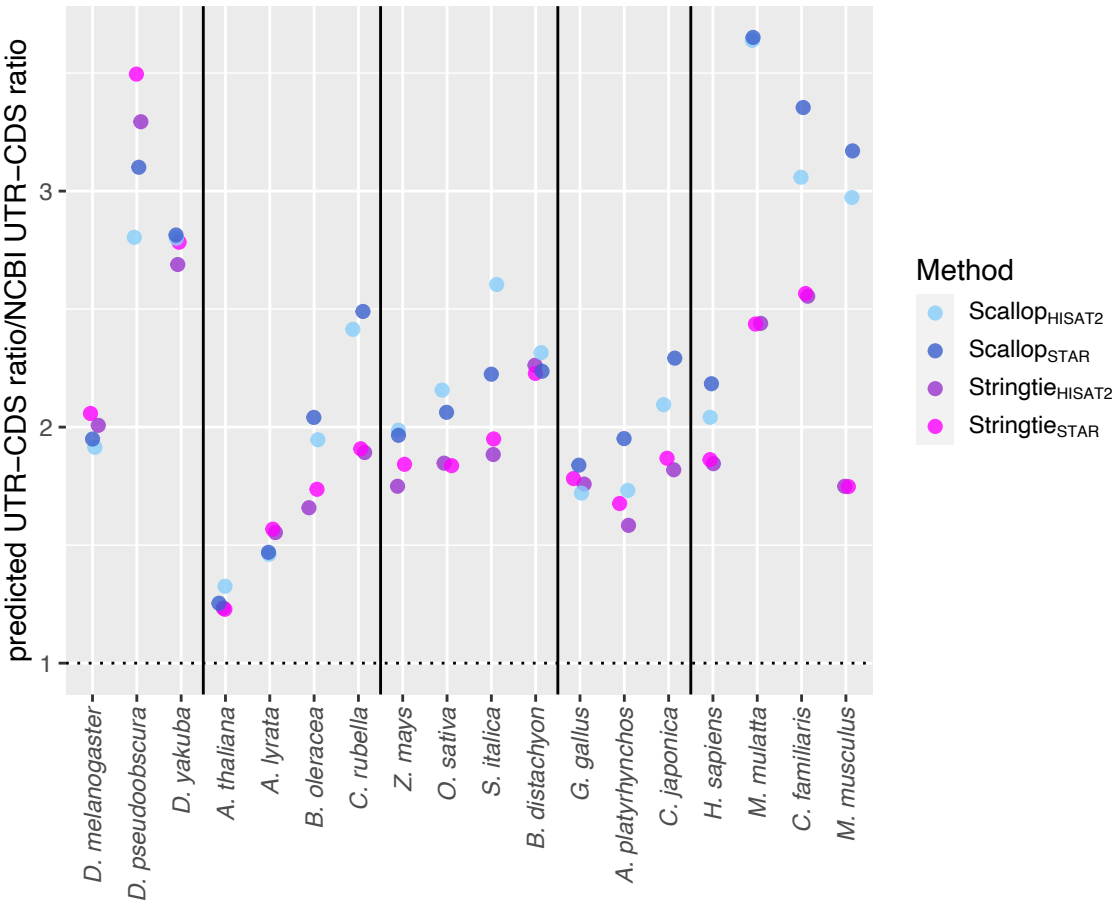


Figure S14. Ratios of the ratio of the median predicted UTR length/median predicted CDS length for RNA-seq assemblers over that for NCBI protein coding transcripts. No clear increase is observed for the most complete and curated annotations (*H. sapiens*, *D. melanogaster*, *G. gallus*, *Z. mays*, *A. thaliana*) relative to other genomes, indicating that the filtering out of cases of transcriptional readthrough (of the sort that NCBI will filter out for high quality annotations) does not explain the reduction in alignment rates for RNA-seq assemblers when UTRs are not included.

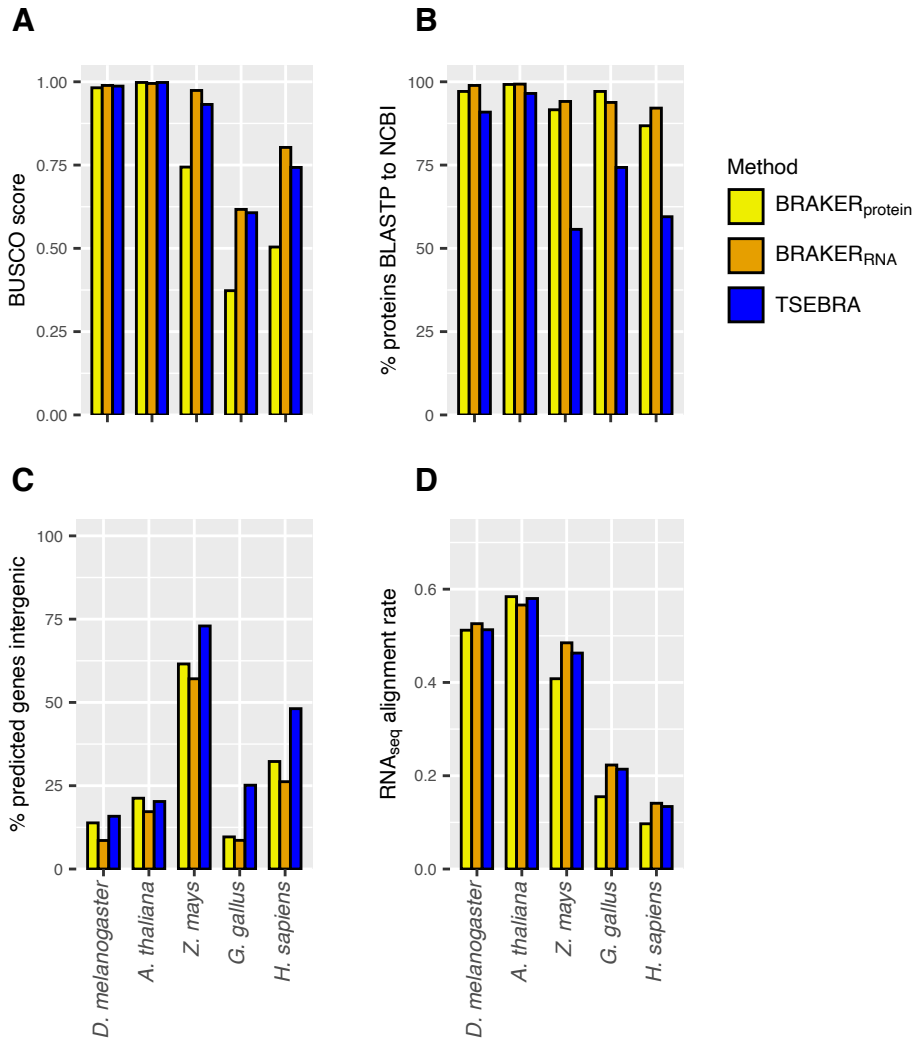


Figure S15. Comparison of TSEBRA integration to BRAKER$_{protein}$ and BRAKER$_{RNA}$ with respect to (A) BUSCO scores, (B) percentage of predicted proteins that have BLASTP hits to NCBI proteins for the species group, (C) the percentage of predicted genes that are intergenic relative to NCBI annotations, and (D) the RNA-seq alignment rates.

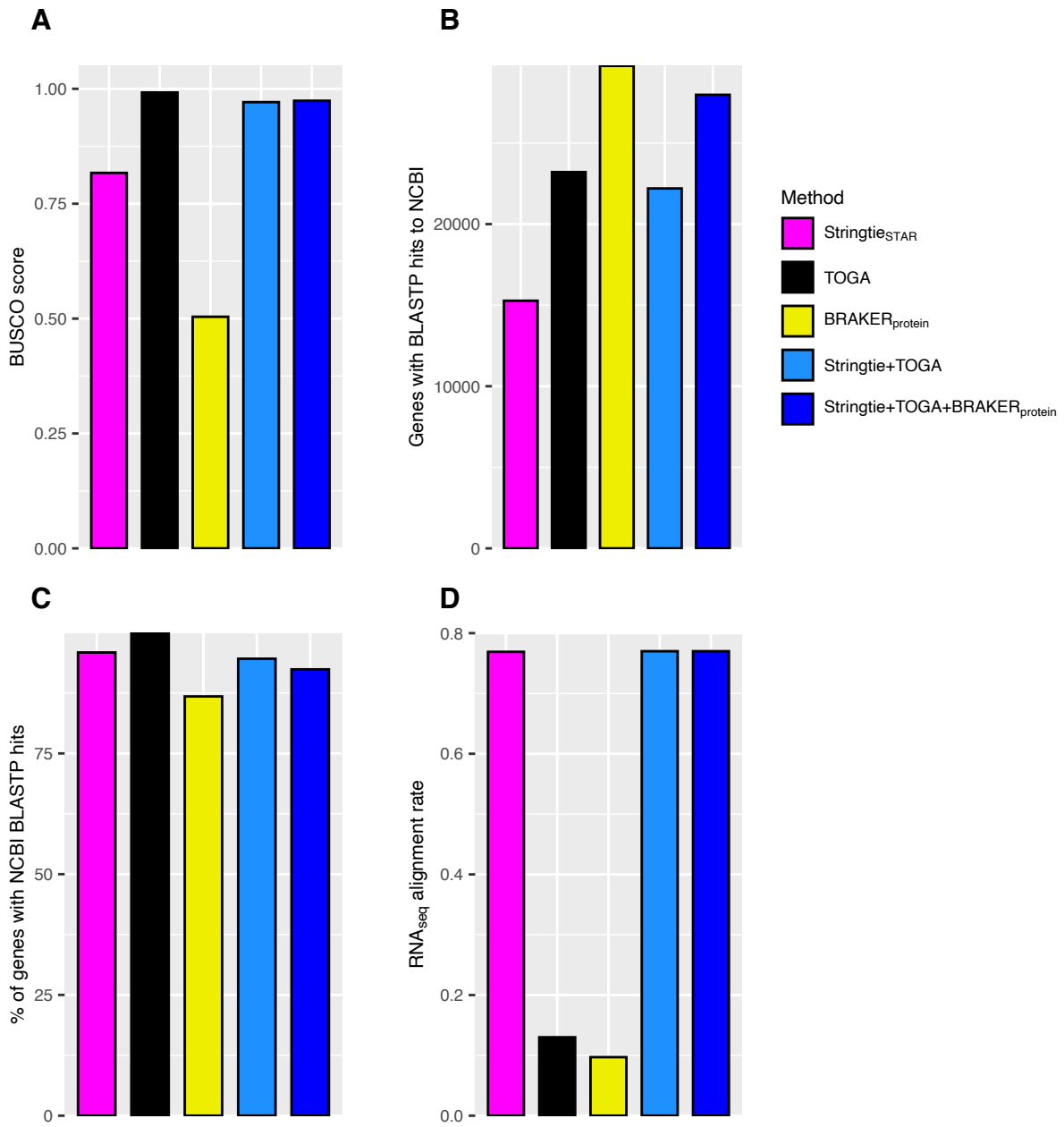


Figure S16. For *Homo sapiens*, Comparisons of (A) BUSCO scores, (B) the number of protein-coding genes with BLASTP hits to NCBI proteins from the mammal species group (C) the proportion of protein-coding genes with BLASTP hits, and (D) the median RNA-seq alignment rate across samples for individual methods and integrations of methods that start with $\text{Stringtie}_{\text{STAR}}$, then add TOGA, and finally $\text{BRAKER}_{\text{protein}}$. For each annotation that is added to the base annotation, integration involves adding genes that fall outside of the base annotation. For example, Stringtie+TOGA is built by adding TOGA genes that fall outside of Stringtie gene intervals.

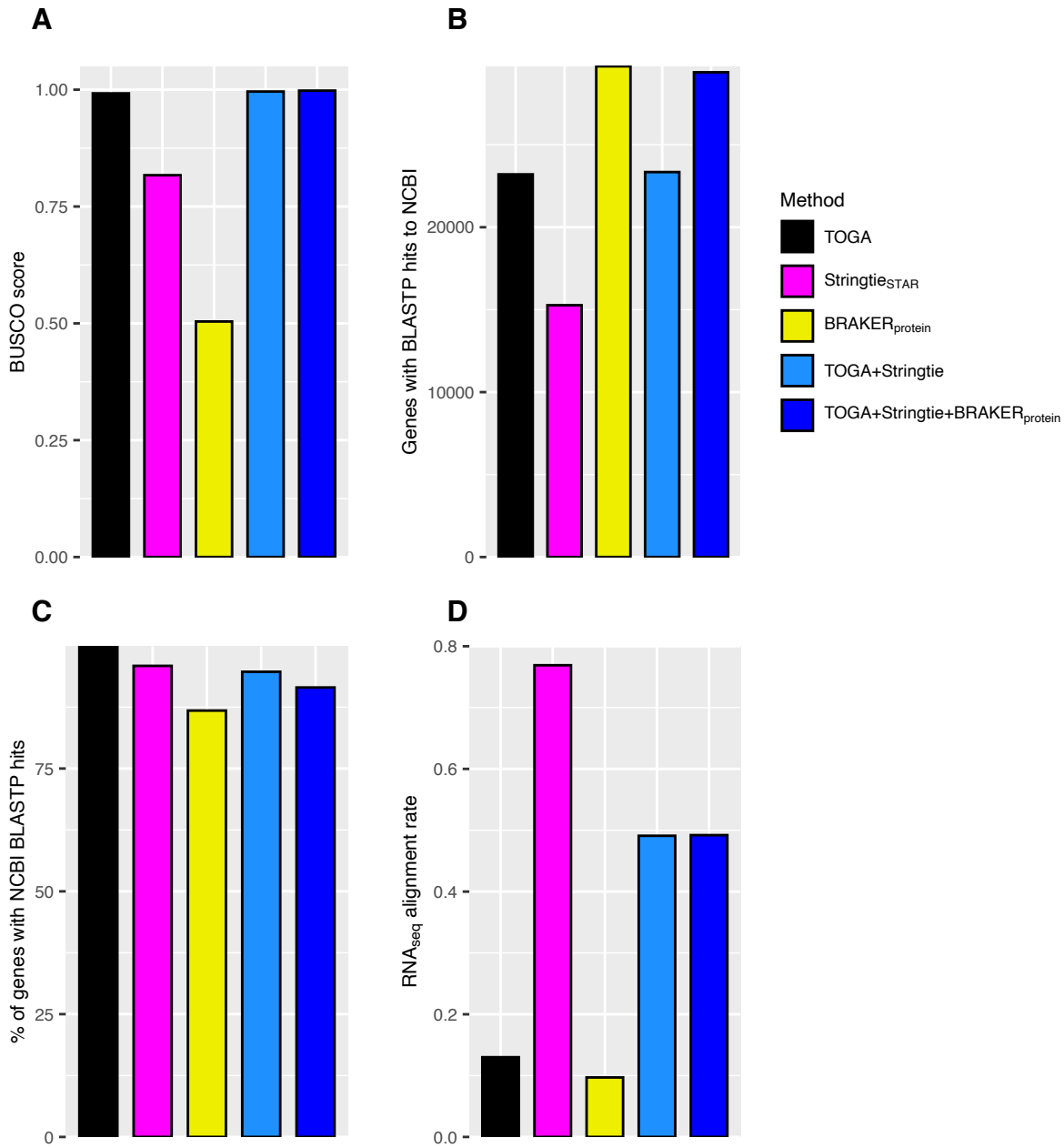


Figure S17. For *Homo sapiens*, Comparisons of (A) BUSCO scores, (B) the number of protein-coding genes with BLASTP hits to NCBI proteins from the mammal species group (C) the proportion of protein-coding genes with BLASTP hits, and (D) the median RNA-seq alignment rate across samples for individual methods and integrations of methods that start with TOGA, then add StringtieSTAR, and finally BRAKERprotein. For each annotation that is added to the base annotation, integration involves adding genes that fall outside of the base annotation. For example, TOGA+Stringtie is built by adding Stringtie genes that fall outside of TOGA gene intervals.

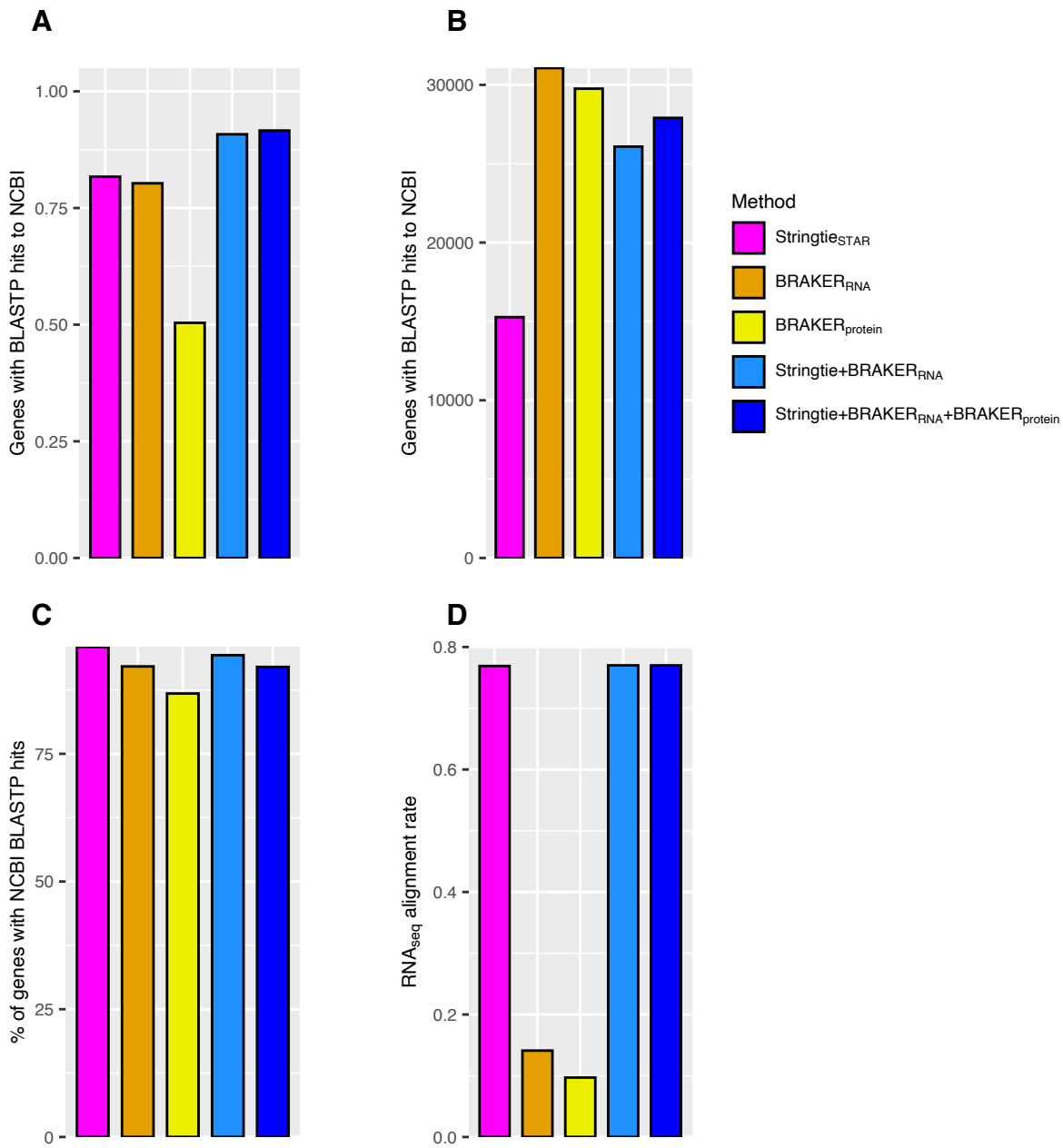


Figure S18. For *Homo sapiens*, Comparisons of (A) BUSCO scores, (B) the number of protein-coding genes with BLASTP hits to NCBI proteins from the mammal species group (C) the proportion of protein-coding genes with BLASTP hits, and (D) the median RNA-seq alignment rate across samples for individual methods and integrations of methods that use both RNA-seq and protein evidence but that don't include annotation transfer with TOGA. For each annotation that is added to the base annotation, integration involves adding genes that fall outside of the base annotation. For example, TOGA+Stringtie is built by adding Stringtie genes that fall outside of TOGA gene intervals.

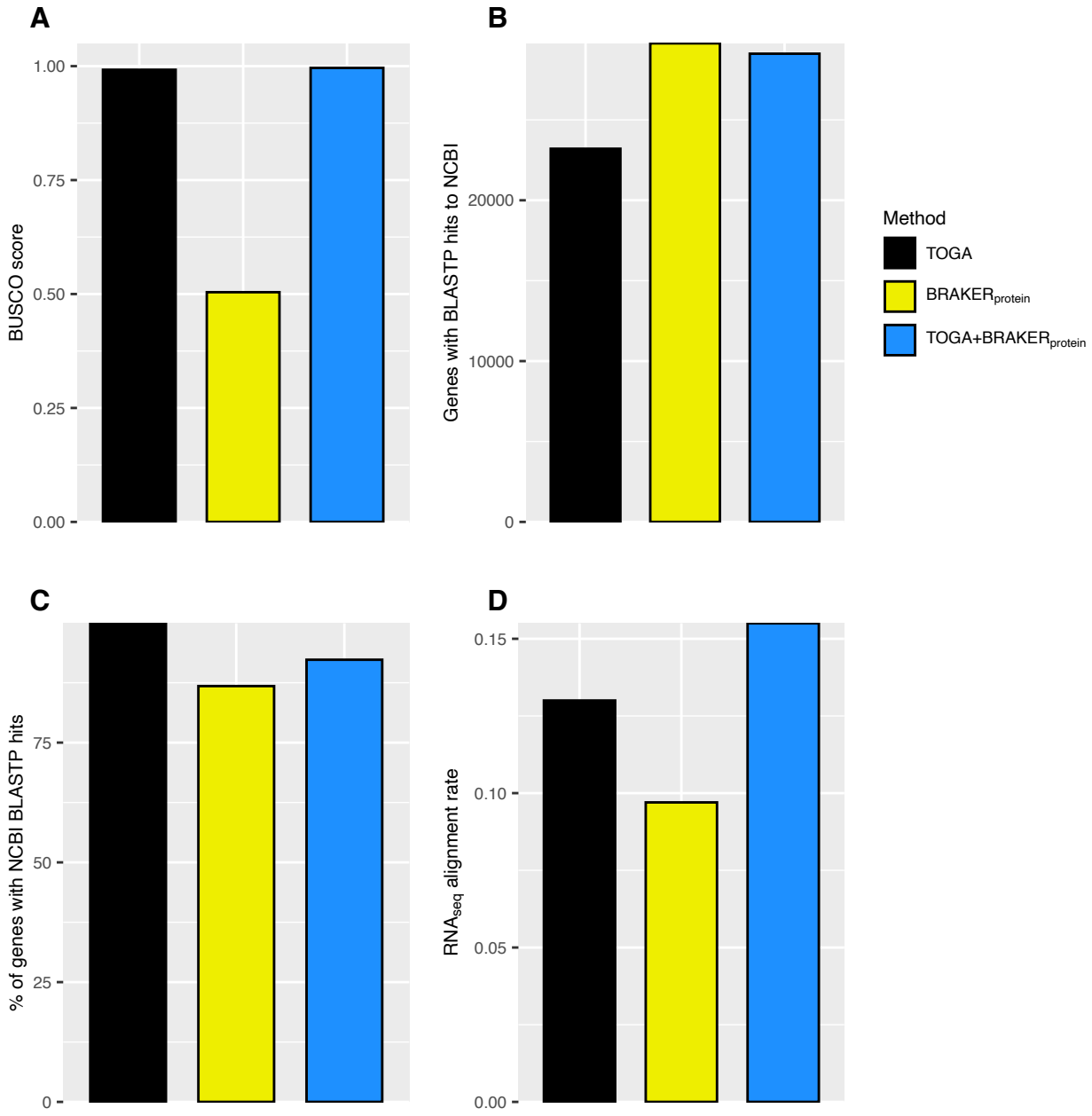


Figure S19. For *Homo sapiens*, Comparisons of (A) BUSCO scores, (B) the number of protein-coding genes with BLASTP hits to NCBI proteins from the mammal species group (C) the proportion of protein-coding genes with BLASTP hits, and (D) the median RNA-seq alignment rate across samples for individual methods and integrations of methods that do not include RNA-seq Integration of BRAKER$_{protein}$ with TOGA involves adding BRAKER genes that fall outside of TOGA gene intervals.