

1 **Title**

2 OrthoFiller: utilising data from multiple species to improve the completeness of genome annotations.

3 **Authors**

4 Michael P. Dunne¹, Steven Kelly²

5 **Author affiliations**

6 ^{1,2}Department of Plant Sciences, University of Oxford, South Parks Road, OX1 3RB, UK

7 **Corresponding author**

8 ²E-mail: steven.kelly@plants.ox.ac.uk, phone: +44 (0) 1865 275123

9 **Abstract**

10 **Background**

11 Complete and accurate annotation of sequenced genomes is of paramount importance to their utility
12 and analysis. Differences in gene prediction pipelines mean that genome sequences for a species
13 can differ considerably in the quality and quantity of their predicted genes. Furthermore, genes that
14 are present in genome sequences sometimes fail to be detected by computational gene prediction
15 methods. Erroneously unannotated genes can lead to oversights and inaccurate assertions in
16 biological investigations, especially for smaller-scale genome projects which rely heavily on
17 computational prediction.

18 **Results**

19 Here we present OrthoFiller, a tool designed to address the problem of finding and adding such
20 missing genes to genome annotations. OrthoFiller leverages information from multiple related
21 species to identify those genes whose existence can be verified through comparison with known
22 gene families, but which have not been predicted. By simulating missing gene annotations in real
23 sequence datasets from both plants and fungi we demonstrate the accuracy and utility of OrthoFiller
24 for finding missing genes and improving genome annotation. Furthermore, we show that applying

25 OrthoFiller to existing “complete” genome annotations can identify and correct substantial numbers
26 of erroneously missing genes in these two sets of species.

27 **Conclusions**

28 We show that significant improvements in the completeness of genome annotations can be made
29 by leveraging information from multiple species.

30 **Introduction**

31 Genome sequences have become fundamental to many aspects of biological research. They provide
32 the basis for our understanding of the biological properties of organisms, and enable extrapolation
33 and comparison of information between species. Owing to the increasing availability and affordability
34 [1][2] of whole-genome sequencing technology, genomic data sets are now produced at a rate at
35 which it is infeasible to rely entirely on careful manual curation to annotate a new genome; rather it
36 is taken as given that a considerable portion of the process must be automated.

37 There has been substantial methodology development in the area of automated gene prediction,
38 with the production of several effective algorithms for identifying genes in *de novo* sequenced
39 genomes [3]. In general, these methods predict genes by learning species-specific characteristics
40 from training sets of manually curated genes. These characteristics include the distribution of intron
41 and exon lengths, intron GC content, exon GC content, codon bias, and motifs associated with the
42 starts and ends of exons (splice donor and acceptor sites, poly-pyrimidine tracts and other features).
43 These characteristics are then used to identify novel genes in raw nucleotide sequences. These
44 prediction methods vary in their performance, as demonstrated by considerable disagreement in the
45 genes and gene models that they predict [4][3]. For example, one study [4] comparing Augustus,
46 GENSCAN, Fgenesh and MAKER, looked at the number of genes predicted on a sample set of *D.*
47 *melanogaster* assemblies with varying numbers of scaffolds. At the extreme end, with 707 scaffolds,
48 the most frugal prediction (MAKER, with 12687 predicted genes) was almost doubled by the most
49 generous prediction (GENSCAN, with 22679 predicted genes). Thus it is to be expected that genome
50 annotations generated by different research groups using different methodologies will differ
51 considerably in the complement of genes that they contain.

52 Absent or inaccurate gene models can not only contribute to oversights in biological investigations,
53 they can also lead to false assertions in large-scale genome and cross-species analyses [5]. For
54 example, incorrectly missing gene annotations can be mistakenly interpreted as gene loss, and such
55 interpretations can lead to mistaken inferences about the biological or metabolic properties of an
56 organism. Similarly, missing gene models can lead to errors in gene expression analyses that map
57 and quantify RNA-seq reads using predicted gene models. Here, reads derived from erroneously
58 missing genes, as they have no reference to map to, have the potential to map to the wrong gene
59 leading to errors in transcript abundance estimation.

60 Much of the cost and effort involved in *de novo* genome annotation can be reduced by leveraging
61 data from other taxa. Moreover, data from disparate taxa have the potential to be used to
62 simultaneously improve a cohort of genome annotations in a mutualistic framework. A number of
63 approaches have been developed to utilise data from other species to improve or assist the process
64 of genome annotation. For example, an automated alignment-based fungal gene prediction
65 (ABFGP) method [6] has been developed for fungal genomes. While this method works well on
66 fungal genomes, it cannot be applied to other taxa and thus has limited general utility.

67 OrthoFiller aims to simultaneously leverage data from multiple species to mutually improve the
68 genome annotations of all species under consideration. It is designed specifically to find “missing”
69 genes in sets of predicted genes from multiple species. That is, to identify those genes that should
70 be present in a genome’s annotation, whose existence can be verified through comparison with
71 known gene families. A standalone implementation of the algorithm is available under the GPLv3
72 licence at <https://github.com/mpdunne/orthofiller>.

73 **Results**

74 **Problem definition, algorithm overview and evaluation criteria**

75 OrthoFiller aims to find genes that are present in a species’ genome, but which have no predicted
76 gene model in the genome annotation for that species. It takes a probabilistic, orthogroup-based
77 approach to gene identification, leveraging information from multiple species simultaneously to
78 improve the completeness of the genome annotations for all species under consideration. OrthoFiller

79 is not designed for *ab initio* gene prediction and requires that each genome under consideration
80 possesses a basic level of annotation, taken to be at least 100 annotated genes. The genomes
81 should ideally be from a set of related species from the same taxonomic group (genus, family, order
82 or class).

83 A workflow for OrthoFiller is shown in Figure 1. The basic input for the algorithm is a set of genome
84 annotation files in general transfer format (GTF) and a set of corresponding genome sequence files
85 in FASTA format. Protein sequences are extracted from the genome FASTA files using the
86 coordinates in the GTF files and a user-selected translation table. The predicted proteomes from the
87 submitted species are clustered into orthogroups using OrthoFinder [7], the protein sequences of
88 each orthogroup are aligned and the source nucleotide sequences for these proteins are threaded
89 back through the protein multiple sequence alignment to create multiple sequence alignments of the
90 nucleotide sequences of each orthogroup. Each nucleotide alignment is used to build a hidden
91 Markov model (HMM) that is used to search the complete genome sequence of each species under
92 consideration. The scores of these HMMs are used to learn the score distributions of true positive
93 and false positive HMM hits (see methods). Each hit to an HMM that does not overlap with an existing
94 predicted gene is subject to filtration using species-specific parameters that have been learned for
95 true and false positive hits. Each hit that survives this filtration is considered to be a potential genic
96 region, or *hint*. The algorithm then attempts to build gene models around these hints, using the
97 Augustus [8] gene finder. Gene models constructed by Augustus are subject to two successive
98 rounds of assessment and filtration. Firstly, the predicted gene models are compared against the
99 hints that were used to inform them: if the gene model and its source hint are not sufficiently similar
100 (see methods), the gene model is considered to be unrelated to the hint, and thus to the orthogroup
101 used to inform its prediction. Secondly, the newly predicted genes that satisfy the first criterion are
102 subject to orthogroup inference using the full set of existing and newly predicted genes. Those newly
103 predicted genes that are clustered in an orthogroup whose HMM was used to predict them are then
104 accepted as *bona fide* genes and added to the genome annotation. Thus genes predicted by
105 OrthoFiller satisfy stringent orthology based criteria for inclusion.

106 To demonstrate the utility of OrthoFiller on real data it was applied independently to two sets of
107 species. Set A comprised five fungal genomes (Table 1) and Set B comprised five plant genomes
108 (Table 2), sourced from the Joint Genome Institute (JGI) and the Saccharomyces Genome Database
109 (SGD) [9][10][11][12]. OrthoFiller was assessed using these datasets in two ways: first via simulating
110 an incomplete genome annotation by randomly removing entries from the genome annotation of one
111 species from each set, and assessing the accuracy of OrthoFiller in recovering the removed genes;
112 second by application of OrthoFiller to the complete datasets and validating the novel detected genes
113 through analysis of publicly available RNA-seq data.

114 Two measures were used to assess the quality of recovered genes: the protein F-score and the
115 orthogroup F-score, both defined in the methods section. These scores were calculated for all genes
116 identified by OrthoFiller, by comparing the recovered gene with the removed gene and assuming
117 that the original removed gene model was correct. Genes that are unique to the test species that
118 lack homologues in other species were not analysed in this test, as OrthoFiller was designed to find
119 evolutionarily conserved genes. As there were no publicly-available comparable methods that
120 perform the same task as OrthoFiller, the method was assessed in comparison to performing the
121 analysis without conducting the OrthoFiller evaluation and filtration steps. i.e. accepting all identified
122 gene models that did not overlap an existing gene.

123 **Evaluation of OrthoFiller on *S. cerevisiae* after removal of 10% of gene annotations**

124 Figure 2 and Table 3 show the results of running OrthoFiller on the set of fungal species shown in
125 Table 1 after random removal of 10% of “discoverable” genes (genes that were contained in an
126 orthogroup with at least one gene from another species) from the predicted complement of genes in
127 *S. cerevisiae* (i.e. 528 nuclear encoded gene annotations were deleted from a total set of 5288
128 discoverable genes).

129 After running OrthoFiller, a total of 197 genes were predicted in the genome of *S. cerevisiae* that
130 were not present in the submitted genome annotation file. Of these, 196 overlapped with genes that
131 were deleted from the original annotation and one was not present in the original annotation (37.1%
132 of 528, Figure 2A). In total, 190 of the 196 found genes (96.9%) were recovered to high accuracy
133 (protein F-score ≥ 0.95). The mean protein F-score of the remaining 6 genes of lower accuracy

134 (protein F-score < 0.95) was 0.85 (Figure 2B). All of the genes that had lower gene model accuracy
135 were placed in exactly the same orthogroup as expected when the sequences were subjected to
136 orthogroup inference. Thus, although 6 of the gene models differed from the original reference gene
137 model, this difference was not sufficient to disrupt downstream identification of orthologous genes.

138 To provide a comparison, in the absence of the OrthoFiller evaluation steps a total of 503 genes
139 were identified, of which 447 overlapped with genes that were deleted from the original annotation
140 and 56 were not predicted as genes in the original *S. cerevisiae* genome (Figure 2A). The regions
141 comprising these 447 found genes corresponded to 440 deleted genes (84.6% of 528). This
142 discrepancy in gene number is due to genes which were recovered, but whose recovered versions
143 were split into multiple parts. There were 7 such split genes. In total, 411 (91.9% of 447) of the genes
144 that overlapped with genes present in *S. cerevisiae* genome annotation were genes recovered with
145 high accuracy (protein F-score \geq 0.95) and the mean protein F-score of those recovered to a lower
146 accuracy was 0.68 (Figure 2B), considerably lower than in the filtered case. Of these 36 lower-quality
147 genes, 11 (30.5%) had an orthogroup F-score less than or equal to 0.95. Moreover, 10 of these
148 genes were sufficiently mis-predicted that they failed to be placed in an orthogroup, or were placed
149 in an orthogroup that shared no members with the orthogroup that contained the original gene. Thus
150 in the absence of OrthoFiller filtration, more genes were recovered but 6 genes were fragmented,
151 10 of the found genes bore insufficient similarity to the reference gene to facilitate orthogroup
152 inference, and 26 were sufficiently mis-predicted that the results of orthogroup inference was altered.

153 Figures 2C-D show the distribution of orthogroup F-scores versus protein F-scores obtained
154 following application of OrthoFiller to this test dataset. The majority of recovered genes had both
155 high protein and orthogroup F-scores (Figure 2C): 189 out of 196 genes (96.4%) had both F-scores
156 \geq 0.95. This indicates that the majority of predicted genes are identical (or nearly identical) to the
157 original removed gene and that when subject to orthogroup inference they were clustered in the
158 correct orthogroup. Imperfect protein F-scores can be explained by discrepancies in intron/exon and
159 start/stop codon choices between the removed and recovered gene models. Imperfect orthogroup
160 F-scores were due to fluctuations in orthogroup membership. Figure 2D shows the results in the
161 absence of OrthoFiller processing. In this case, 399 of 447 genes (89.3%) were of dually high quality.

162 In particular, there were 5 predicted genes with both a low (< 0.5) protein and orthogroup F-score,
163 indicating those predicted genes were sufficiently incorrect to cause errors in orthologous gene
164 identification. Thus, although OrthoFiller does not recover all deleted genes (37% of removed
165 genes), application of OrthoFiller resulted in the recovery of high-quality gene annotations that
166 contain few (in this example there are none) incorrectly predicted genes.

167 **Evaluation of OrthoFiller on *S. cerevisiae* after removal of 90% of gene annotations**

168 Figure 3 and Table 3 show the performance statistics for OrthoFiller using a version of *S. cerevisiae*
169 genome where 90% of gene annotations were removed. This represents an extreme case where a
170 genome has minimal annotation. The full details of detection of the deleted genes at different stages
171 in the OrthoFiller algorithm are shown in Supplemental Figure 3. Here, application of OrthoFiller
172 resulted in the identification of 1529 genes that overlapped with 1528 of the removed genes (32.1%,
173 Figure 3A). One of the genes was split into two parts. Of the found genes, 1455 (95.1%) were
174 recovered with a protein F-score of 0.95 or greater. Of the 74 genes with lower protein F-scores
175 (Figure 3B), only 6 (8.1%) had an orthogroup F-score < 0.95 . As before, although these gene models
176 differed from the original reference gene model, this difference was not sufficient to disrupt
177 downstream identification of orthologous genes.

178 In the absence of OrthoFiller filtration, 4325 genes were found, of which 4116 overlapped the
179 removed genes. Of the removed genes, 4156 were recovered, of which 64 genes were split. 3801
180 of the found genes had a protein F-score ≥ 0.95 (87.9%). Of the 355 genes with lower protein F-
181 scores, 113 had an orthogroup F-score lower than 0.95, and 97 were sufficiently mis-predicted that
182 they failed to be placed in any orthogroup at all, or in an orthogroup completely different to the one
183 that was used to find them.

184 Figures 3C-D show the distribution of orthogroup F-scores versus protein F-scores for recovery in
185 the 90% removal case. Figure 3C shows that most genes were recovered well, with 1367 of 1529
186 (89.4%) genes predicted correctly and placed in the correct orthogroup when subject to orthogroup
187 inference (protein F-score ≥ 0.95 , orthogroup F-score ≥ 0.95). Interestingly, there are many genes
188 that are predicted correctly but are placed into a slightly different orthogroup to what was expected.
189 This is due to changes in orthogroup membership caused by the many still-missing genes. Thus,

190 although the input datasets are dramatically different the performance characteristics of OrthoFiller
191 on the 10% and 90% datasets are broadly consistent (e.g. 37.1% and 32.1% recovery respectively,
192 96.9% and 95.1% high-accuracy recoveries respectively).

193 **Evaluation of OrthoFiller on *A. thaliana* after removal of 10% of gene annotations**

194 As it could be argued that fungal genomes present an easier challenge, an additional demonstration
195 of the utility of OrthoFiller on an alternative group of organisms was also conducted. Here the
196 analogous test of the method was applied to a set of five land plant genomes (Table 2). Table 4 and
197 Figure 4 show performance statistics from application of OrthoFiller to the *A. thaliana* genome with
198 10% (3168) gene annotations removed. Out of the 1097 genes that were output by OrthoFiller, 982
199 overlapped removed genes. A total of 908 of the original genes were recovered, of which 67 were
200 recovered but split into multiple parts (7.4%). Of the found genes, 416 (42.4%) had a protein F-score
201 of 0.95 or higher, and of the lower quality genes, 56.5% had orthogroup F-scores of 0.95 or higher,
202 and 52.5% were placed into exactly the same orthogroup as the one used to predict them. The mean
203 protein F-score of lower-quality genes was 0.60. Thus similar to the fungal dataset, application of
204 OrthoFiller resulted in the identification of 31.0% of the removed genes, with 42.4% being of gene
205 model accuracy (assuming the deleted gene to be true).

206 In the absence of OrthoFiller filtration 7048 genes were found, nearly twice as many as were
207 removed. Only 3484 of these overlapped removed genes, of which 491 (14.1%) had a protein F-
208 score of 0.95 or higher. 1664 genes were recovered, of which 850 (51.1%) were split into multiple
209 parts. The mean protein F-score of lower-quality genes was 0.37, and the percentage of lower-quality
210 genes which received an orthogroup F-score of 0.95 or above was 18.2%.

211 Figures 4C-D show the distribution of orthogroup F-scores versus protein F-scores for recovery in
212 the 10% removal case for *A. thaliana*. Using OrthoFiller, 325 of 902 (33%) of genes had both a very
213 high (≥ 0.95) protein and orthogroup F-score. In the unfiltered case, 324 of the genes had both a
214 high protein and orthogroup F-score, though as a percentage of the total genes found (9.2% of 3484
215 found genes), the success rate was considerably lower. Conversely, 35 out of 982 (3.6%) had both
216 scores very low (<0.5), compared with 1710 out of 3484 (49.1%) genes in the absence of OrthoFiller

217 filtration. Thus in this case using OrthoFiller considerably reduces the proportion of found genes
218 which are erroneous.

219 **Evaluation of OrthoFiller on *A. thaliana* after removal of 90% of gene annotations**

220 Performance statistics for the application of OrthoFiller to the 90% depleted *A. thaliana* genome
221 (28516 genes removed) can be seen in Table 4 and Figure 5. Of 10931 found genes, 10788
222 overlapped removed genes, 3393 of which (31.5%) had protein F-score 0.95 or above. 889 (9.0%)
223 of the recovered genes were split into multiple parts. A total of 9840 out of 28516 (34.5%) removed
224 genes were recovered, though 889 were split into parts (9.0%). Of the lower-quality genes, 50.1%
225 had orthogroup F-score ≥ 0.95 , and 46.7% were placed in exactly the right orthogroup. The mean
226 protein F-score of the lower-quality genes was 0.57. Thus having fewer gene models to serve as
227 examples for gene model training resulted in a higher error rate in gene model prediction.

228 In the absence of OrthoFiller filtration, 28793 genes were predicted, 26004 of which overlapped
229 removed genes. Of these, only 3539 (13.4%) had a protein F-score of 0.95 or above, with just 23%
230 of the lower-quality genes having orthogroup F-score ≥ 0.95 . In total 12646 of the 28516 removed
231 genes were recovered, although 6052 of them were split (47.9%). The mean protein F-score of the
232 lower-quality genes was 0.37. This shows that, although slightly more genes were recovered in the
233 unfiltered case, considerably more noise and erroneous predictions are produced.

234 Figures 5C-D show the distribution of orthogroup F-scores versus protein F-scores for recovery in
235 the 90% removal case for *A. thaliana*. Using OrthoFiller, 2427 of 10788 found genes (22.5%) had
236 both a very high (≥ 0.95) protein and orthogroup F-score, compared with 2413 out of 26004 (9.3%)
237 in the unfiltered case. Conversely, only 5.9% of genes (636 out of 10788) predicted using OrthoFiller
238 had both scores very low (<0.5), compared with 44.7% of genes (11631 out of 26004) in the absence
239 of OrthoFiller filtration. Thus, similarly to with the fungal data set, the performance characteristics of
240 OrthoFiller on the 10% and 90% plant datasets are broadly consistent (e.g. 31.0% and 34.5%
241 recovery respectively, 42.4% and 31.5% high-accuracy recoveries respectively), and both contain a
242 considerably smaller proportion of clearly erroneous genes than would be found without filtering.

243 **OrthoFiller detects hundreds of conserved genes not present in the reference** 244 **genome annotations**

245 In addition to testing the ability of OrthoFiller to recover already predicted genes, the algorithm was
246 applied to both of the sets of complete genomes listed in Table 1 and Table 2, to assess the potential
247 for novel genes to be discovered. The number of genes found for each species in each set is listed
248 in Tables 5 and 6. Application of OrthoFiller to the 5 fungal species listed in Table 1 resulted in the
249 detection of 31 novel genes distributed across the 5 species. Further rounds of OrthoFiller gene
250 prediction identified no additional genes to those already found. Application of OrthoFiller to the 5
251 plant species listed in Table 2 resulted in the identification of 570 individual novel genes in these
252 species.

253 To be detected as a novel gene OrthoFiller requires genes to pass rigorous sequence similarity tests
254 to genes in other species (including empirical evaluation of sequence similarity scores to distinguish
255 real from spurious hits), which in itself provides evidence for the existence of predicted genes through
256 homology. To provide additional evidence for the existence of the novel predicted genes they were
257 subjected to analysis using publicly available RNAseq data from the Sequence Read Archive (SRA)
258 [13]. The datasets used for this analysis are listed in Tables 7 and 8. The tables also show the
259 percentage of the novel genes found that had evidence for their existence in the RNAseq data. For
260 most genomes, most genes predicted by OrthoFiller are supported by RNAseq evidence, with the
261 average percentage of evidence-supported novel genes being 85.3% across the fungal species, and
262 55.5% across the plant species. Given that the plant RNAseq datasets come from single tissue
263 samples under a single condition it is not expected that all genes will be detected in these samples.
264 For example, similar detection statistics were obtained for the original predicted genes from the
265 source datasets, shown in Tables 7 and 8. It should also be noted that genes that are present in
266 RNAseq reads are more likely to have been annotated already, given that many genome annotation
267 pipelines rely on such data to perform their analyses [3].

268 ***Discussion***

269 Here we present OrthoFiller, an automated method for improving the completeness of genome
270 annotations. It leverages information from multiple taxa, clustering genes into orthogroups and

271 finding genes that are conserved between species but that have escaped detection. OrthoFiller is
272 designed to be stringent, conservatively identifying genes that can be confidently identified as
273 missing members of existing orthogroups. Specifically, to pass the filtration criteria for detection by
274 OrthoFiller, genes must be members of orthogroups conserved in multiple species. Thus OrthoFiller
275 will not find genes that lack homologues in other species. These stringent criteria mean that not all
276 genes that could be detected will be detected by the algorithm, but rather that the user should have
277 confidence in the validity of genes identified by the method.

278 OrthoFiller is intended to be run after a genome annotation is considered by the user to be complete
279 or near-complete. OrthoFiller is designed with small scale genome sequencing projects in mind and
280 is provided to enable users without significant resources for comprehensive RNAseq-based genome
281 annotation to leverage information from related species to improve their genome annotations.
282 However, OrthoFiller is equally suited for use in large-scale genome comparisons, reliably filling
283 gaps in gene sets prior to large scale comparative genomics investigations. Application of OrthoFiller
284 in these cases will enable genes to be analysed in downstream analysis that would otherwise have
285 been classified as absent.

286 The utility of OrthoFiller is demonstrated on both plant and fungal genome datasets, both in its ability
287 to successfully find missing genes, and in the effectiveness of its filters in eliminating low-quality
288 gene predictions. Application of this method to small groups of plant and fungal genomes resulted
289 in the identification of 570 and 31 genes respectively. These genes are conserved in one or more
290 species but were absent from the genome annotation in which they were predicted. We anticipate
291 that application of OrthoFiller to larger datasets will likely result in further genome annotation
292 improvement. The quality of genes found by OrthoFiller was assessed by artificial removal and
293 recovery of subsets of genes from a single genome, treating those original gene models as true, and
294 evaluating the quality of those genes that were recovered by comparison to the removed genes. In
295 the absence of the OrthoFiller filtration steps, the proportion of poor-quality genes that are recovered
296 is considerably higher.

297 OrthoFiller is mainly designed for use on genomes that have already undergone some basic level of
298 annotation. As can be seen by comparing the 10% and 90% removal cases in the two data sets,

299 application to very poorly annotated genomes can result in more genes of dubious quality, both from
300 a sequence and an orthogroup perspective. It is worth noting that many of the genes with lower-
301 quality scores, particularly those with only one of the scores being low, can be explained by alternate
302 gene models (in the protein F-score) and shifting of orthogroups due to expansion of proteome sets
303 (in the orthogroup F-score case). In all cases, in the absence of OrthoFiller filtration considerably
304 higher numbers of genes were predicted that didn't resemble the genes that they were supposed to,
305 indicating that they are erroneous.

306 The OrthoFiller algorithm is designed to run on a Unix system with python and a minimal number of
307 standard additional tools (HMMer, BedTools, Augustus, R). The speed of the algorithm is principally
308 dependent on the speed of Augustus and HMMer, however processing time can be decreased by
309 parallelising these steps of the method over multiple CPUs.

310 Accurate and complete genome annotation is of paramount importance to the effective analysis of
311 genomic and transcriptomic data, as well as for phylogenetic inference from genomic data. As the
312 quantity of published genomes increases, care must be taken to ensure accuracy and quality of
313 genome annotations are maintained. Automated methods that leverage publicly available
314 information from multiple species to improve the annotation of newly sequenced genomes will help
315 improve the accuracy and completeness of these resources and thus the quality of all analyses that
316 utilise them.

317 **Methods**

318 **Data sources**

319 For algorithm development and evaluation, a set of five small, well-annotated fungal genomes (Table
320 1) and a set of five well-annotated plant genomes (Table 2) were selected. Evaluation of the
321 algorithm focussed on *S. cerevisiae* and *A. thaliana*, as the gene models in these genomes have
322 historically been subject to extensive improvement and revision and are the most likely to be correct.

323 **Algorithm overview**

324 OrthoFiller proceeds in five stages summarised in Figure 1 and described in detail in the following
325 sections. In brief, the algorithm begins by inferring a set of orthogroups from the protein coding genes

326 of the set of species submitted to OrthoFinder (Figure 1A). The protein sequences in these
327 orthogroups are subject to multiple sequence alignment, converted to nucleotide sequences and
328 used to build HMMs. These HMMs are used to search the genomes of each species under
329 consideration (Figure 1B) and the resultant HMM hits are subject to stringent filtering (Figure 1C)
330 before being used as hints for gene model construction (Figure 1D). The gene models are subject
331 to additional filtering (Figure 1E) and only those gene models that pass all filters are added to the
332 revised genome annotation. The revised genome annotations are then subject to orthogroup
333 inference (Figure 1F) and resultant orthogroups are analysed to confirm the identity of the newly
334 predicted genes. The complete details for each step of this algorithm are described in the sections
335 below.

336 **Inference of Orthogroups and construction of HMMs**

337 Orthogroups are inferred using OrthoFinder [7]. If a gene from the source annotation is not included
338 in an orthogroup with at least one other sequence, it is classed as a *singleton*, and is not considered
339 in downstream analyses. This is consistent with the problem definition of OrthoFinder, that is to identify
340 unannotated genes that are conserved between species. Amino acid sequences from the
341 orthogroups are aligned with MAFFT [14], using the L-INSI algorithm, and the resultant multiple
342 sequence alignments are back-translated using the source nucleotide sequences. The resulting
343 nucleotide alignments are converted to Hidden Markov models (HMMs) using HMMer [15], each of
344 which is then searched against each input genome in turn to generate a set of hits per HMM per
345 species.

346 **Evaluation of HMM search results**

347 Due to the probabilistic nature of HMM searches, there is considerable variation in the quality of the
348 relationship between a hit region and the set of sequences used to generate the source HMM. One
349 expects a large amount of “background noise”, that is sequence regions which pass the thresholds
350 of the HMM but whose relevance is dubious. Each HMM hit has an associated bit score, an
351 aggregated base-by-base similarity score between the hit and the aligned sequences used to
352 generate it: we use this score to assess the quality of the hit. The bit score is strongly dependent on

353 the hit length, thus to prevent gene length from biasing downstream analyses the bit score of a hit is
354 divided by the hit length, to generate the *adjusted score* for a hit h :

355
$$score_{adj}(h) = \frac{score(h)}{length(h)}$$

356 The adjusted score is related to the e-value. However, the e-value calculation enforces a strict lower
357 limit of 1×10^{-200} , all lower scores being rounded down to zero. Thus use of e-values would
358 introduce irreversible length bias and would lead to downstream errors, as has been shown
359 previously [7]. As bit scores do not have a threshold value, and they have been previously shown to
360 be capable of facilitating accurate inference of phylogenetic trees [16], and length-corrected bit
361 scores are used as the basis of the scoring scheme in OrthoFinder [7], they were used here.

362 For each species, a threshold value for hit acceptance or rejection based on a hit's adjusted score
363 is created, by considering the distribution of hits which overlapped known genes. Anything above
364 this threshold is considered to be genuine, and anything below this threshold is considered to be
365 noise. An HMM hit is classed as *good* if it overlaps any gene from the orthogroup used to create the
366 HMM, *bad* if it only overlaps genes from orthogroups other than the one used to create the HMM,
367 and *candidate* if it overlaps no known gene at all. Here candidate hits are the potential new genes
368 of interest, and the *good* and *bad* genes are used to inform our judgement about the reliability of the
369 candidate hits.

370 Distributions of adjusted scores for good and bad hits to the *S. cerevisiae* genome from all HMMs
371 generated by the species in Table 1 are shown in Supplemental Figure 1. Distributions for good and
372 bad hits are clearly delineated into two distinct distributions. Note that in this case there are relatively
373 few candidate hits, since the genome under inspection is already well annotated and is expected to
374 have few missing gene predictions. Skew-t distributions are fit separately to the good and bad score
375 distributions using *gam/ss* [17]. Skew distributions were chosen because they allow flexibility in
376 location, shape and scale of the underlying data and are commonly used for estimating parameters
377 such as location and scale, while allowing the same distribution type to be used to fit both the good
378 and bad hits. A separate skew-t distribution for the good and bad hits is fit for each species. In the

379 event that there are insufficient good and bad hits to fit distributions, good and bad hits from the
380 other species are aggregated and a threshold value is calculated from this.

381 For a given adjusted score x , the distributions of the *good* and *bad* hits are used to estimate both
382 the absolute probabilities of a hit being genuine or being a mistake. We can estimate

$$383 \quad P(\text{genuine}|x) = \frac{P(x|\text{genuine})P(\text{genuine})}{P(x|\text{genuine})P(\text{genuine}) + P(x|\text{mistake})P(\text{mistake})}$$

$$384 \quad P(\text{mistake}|x) = \frac{P(x|\text{mistake})P(\text{mistake})}{P(x|\text{genuine})P(\text{genuine}) + P(x|\text{mistake})P(\text{mistake})}$$

385 and then retain the hit depending on whether it has a higher probability of being genuine than being
386 a mistake, based on its adjusted score. The probabilities $P(\text{genuine})$ and $P(\text{false})$ are estimated by
387 considering the proportion of good/bad hits which are good and bad respectively. The probability
388 density functions $P(x|\text{genuine})$ and $P(x|\text{false})$ are determined using the fitted distributions as
389 described above.

390 **Acquisition and evaluation of putative predicted genes**

391 Hits which survive the hit filtration step are passed to the gene-finding program Augustus as *hints*
392 specified as exon parts. Only predicted genes that have a nonzero overlap with these hints are
393 retained. These predicted genes are then subjected to a *hint filter*, which aims to separate those
394 genes which have genuinely arisen from the hint from those that overlap the hint by chance. The
395 hint filter evaluates a *hint F-score* for each predicted gene, by comparing against the hints from a
396 particular orthogroup which overlap it. The hint F-score is a measure of how well the found gene
397 corresponds to the hints used to inform its discovery. Each predicted gene G will have at least one
398 *hint region* corresponding to it, which is a set of non-overlapping coordinates obtained from merging
399 all hints that overlap G , and which are all derived from the same orthogroup. For a hint region H and
400 a predicted gene G , the hint F score is defined as:

$$401 \quad hf(H, G) = \frac{2 \cdot hP(H, G) \cdot hR(H, G)}{hR(H, G) + hP(H, G)}$$

402 where

403
$$hP(H, G) = \frac{|H \cap G|}{|H|}; \quad hR(H, G) = \frac{|H \cap G|}{|G|}$$

404

405 The filter uses a threshold hint F-score value of 0.8 (i.e. on average 80% of the length of the predicted
406 gene is covered by the hit and vice versa), below which potential gene models are discarded. This
407 value was chosen based on an analysis of hint F-scores of good and bad hits (as defined above)
408 versus the Augustus output corresponding to them. Distributions for hint F-scores for the good and
409 bad hits can be seen in Supplemental Figure 4, in which it can be clearly seen that practically all
410 genuine hints pass the threshold value of 0.8.

411 Once gene models have been filtered, they are fed once again into OrthoFinder, to cluster them into
412 orthogroups. The orthogroup of each newly predicted gene is compared with the orthogroup(s) which
413 were used to predict that gene. It is possible that multiple orthogroups informed the prediction of the
414 same gene; similarly, there may be fluctuations in orthogroup membership between the original and
415 new genomes. It is therefore only required that the new orthogroup into which the gene is clustered
416 has non-zero overlap with at least one of the orthogroups used to predict it, and genes which do not
417 fulfil this criterion are discarded.

418 **Algorithm evaluation**

419 **Recovery of removed genes**

420 The test set of species from Table 1 was used to analyse the effectiveness of OrthoFiller for genomes
421 of various levels of completion. Altered versions of the *S. cerevisiae* genome annotation were
422 constructed with 10% and 90% of genes randomly removed, and the level of recovery of the removed
423 genes upon implementation of OrthoFiller was assessed, where a gene G was considered to be
424 *recovered* if OrthoFiller predicted a gene G' such that G and G' have non-zero overlap.

425 The quality of the predicted genes was assessed by considering two scores: the orthogroup F-score
426 and the protein F-score. The protein F-score is defined as

427
$$pF(S, S') = \frac{2 \cdot pP(S, S') \cdot pR(S, S')}{pR(S, S') + pP(S, S')}$$

428 where S is the original amino acid sequence and S' is the amino acid sequence of the recovered
429 gene, and

$$430 \quad pP(S, S') = \frac{|S \cap S'|}{|S|}; \quad pP(S, S') = \frac{|S \cap S'|}{|S'|}$$

431 where the intersection is defined to be the sum of identical amino acids in an alignment (MAFFT L-
432 INSI) of the two sequences. The orthogroup F-score is defined as

$$433 \quad oP(S, S') = \frac{2 \cdot oP(O, O') \cdot oR(O, O')}{oR(O, O') + oP(O, O')}$$

434 where O is the orthogroup that the gene is placed when no deductions have been made, O' is the
435 orthogroup into which the gene is placed when OrthoFinder is run on the OrthoFiller results, and

$$436 \quad oP(O, O') = \frac{|O \cap O'|}{|O|}$$

$$437 \quad oR(O, O') = \frac{|O \cap O'|}{|O'|}$$

438 where cardinality of the orthogroups takes into account only genes which were present in the input
439 set of genome annotations, i.e. not counting the newly discovered genes.

440 **Evaluation of novel predicted genes**

441 RNA-seq data was downloaded from the Sequence Read Archive, and aligned to the genome with
442 BowTie2 using default parameters. Coverage was calculated using BedTools coverage.

443 **Availability of data and materials**

444 The software is available under the GPLv3 licence at <https://github.com/mpdunne/orthofiller>.

445 ***Competing Interests***

446 The authors declare that they have no competing interests.

447 ***Acknowledgements***

448 NA.

449 **Funding**

450 SK is a Royal Society University Research Fellow. This work was supported by the European Union's
451 Horizon 2020 research and innovation programme under grant agreement number 637765. MPD is
452 supported by an EPSRC studentship through EP/G03706X/1.

453 **Author's Contributions**

454 SK conceived the project. MPD developed the algorithm. SK and MPD analysed the data and wrote
455 the manuscript. Both authors read and approved the final manuscript.

456 **Figure Legends**

457 **Figure 1: Workflow diagram for the OrthoFiller algorithm. A)** Proteomes are subdivided into
458 orthogroups using OrthoFinder. **B)** Protein sequences in each orthogroup are subject to multiple
459 sequence alignment, back-translated to DNA and used to create hidden Markov models (HMMs).
460 These HMMs used to search each genome in the set. **C)** The set of hits are evaluated and filtered
461 to remove low quality hits. **D)** Gene models are constructed around each retained hit using Augustus.
462 **E)** The new gene models are compared to the hints that were used to generate them, and filtered to
463 remove those which bear in sufficient similarity to the hints. **F)** The filtered genes are clustered into
464 orthogroups and genes that are successfully placed into the orthogroup that was used to identify
465 them are retained. **G)** The process may be run once, or iteratively until no further genes are found.

466

467 **Figure 2: Performance of OrthoFiller on *S. cerevisiae* genome with 10% of annotated genes**
468 **removed.**

469 **A)** Using OrthoFiller 197 genes were found whose genomic locations matched any of the 528 deleted
470 genes. In the absence of OrthoFiller filtration this increased to 447 genes identified that overlap any
471 part of a deleted gene. **B)** A boxplot of protein F-scores for genes predicted using OrthoFiller, or in
472 the absence of OrthoFiller filtration, that had a protein F-score of ≤ 0.95 . **C)** Density plot showing the
473 protein and orthogroup F-scores for all recovered genes using OrthoFiller. **D)** Density plot showing
474 the protein and orthogroup F-scores for all recovered genes in the absence of OrthoFiller filtration.

475 **Figure 3: Performance of OrthoFiller on *S. cerevisiae* genome with 90% of annotated genes**
476 **removed.**

477 **A)** Using OrthoFiller 1529 genes were found which overlapped any of the 4759 deleted genes. In
478 the absence of OrthoFiller filtration this increased to 4156 genes. **B)** A boxplot of protein F-scores
479 for genes predicted using OrthoFiller, or in the absence of OrthoFiller filtration, that had a protein F-
480 score of ≤ 0.95 . **C)** Density plot showing the protein and orthogroup F-scores for all recovered genes
481 using OrthoFiller. **D)** Density plot showing the protein and orthogroup F-scores for all recovered
482 genes in the absence of OrthoFiller filtration.

483 **Figure 4: Performance of OrthoFiller on *A. thaliana* genome with 10% of annotated genes**
484 **removed.**

485 **A)** Using OrthoFiller 982 genes were found which overlapped any of the 3168 deleted genes. In the
486 absence of OrthoFiller filtration this increased to 3484 genes. **B)** A boxplot of protein F-scores for
487 genes predicted using OrthoFiller, or in the absence of OrthoFiller filtration, that had a protein F-
488 score of ≤ 0.95 . **C)** Density plot showing the protein and orthogroup F-scores for all recovered genes
489 using OrthoFiller. **D)** Density plot showing the protein and orthogroup F-scores for all recovered
490 genes in the absence of OrthoFiller filtration.

491 **Figure 5: Performance of OrthoFiller on *A. thaliana* genome with 90% of annotated genes**
492 **removed.**

493 **A)** Using OrthoFiller 10788 genes were found which overlapped any of the 28516 deleted genes. In
494 the absence of OrthoFiller filtration this increased to 26204 genes. **B)** A boxplot of protein F-scores
495 for genes predicted using OrthoFiller, or in the absence of OrthoFiller filtration, that had a protein F-
496 score of ≤ 0.95 . **C)** Density plot showing the protein and orthogroup F-scores for all recovered genes
497 using OrthoFiller. **D)** Density plot showing the protein and orthogroup F-scores for all recovered
498 genes in the absence of OrthoFiller filtration.

499 **Figure 6: Coverage plots and orthogroup trees for a selection of new genes.** Five
500 representative examples of RNAseq coverage on genes predicted using OrthoFiller. Phylogenetic
501 trees demonstrate the relationship of the newly predicted gene to other genes in the orthogroup.

502 **Supplemental Figure Legends**

503 **Supplemental Figure 1: hit score distributions for good, bad and candidate hits.** Hits are to the
504 *S. cerevisiae* genome, using HMMs from all orthogroups. **A)** Length normalised bit scores of HMM
505 hits to regions of the genome that contained genes that were not part of the orthogroup used to
506 generate the HMM (bad hits). **B)** Length normalised bit scores of HMM hits to regions of the genome
507 that do contain the gene used to generate the HMM (good hits). **C)** Length normalised bit scores of
508 HMM hits to regions of the genome that do not contain any previously annotated genes (candidate
509 novel gene hits). **D)** All distributions overlaid.

510 **Supplemental Figure 2: Recovery of removed genes from *S. cerevisiae* after 10% removal:**
511 **Representation of removed genes at each stage, filtered vs. unfiltered cases.** **A)** The number
512 of deleted genes that obtained hits from one or more orthogroup HMMs. **B)** The number of deleted
513 genes that had hits after OrthoFiller hint filtration. **C)** No hint filtration. **D)** The number of deleted
514 genes for which a gene prediction was made using Augustus that satisfied OrthoFiller filtration tests.
515 **E)** The number of deleted genes that for which a gene prediction was made using Augustus in the
516 absence of OrthoFiller filtration. **F)** The number newly predicted genes that were retained or
517 discarded based on the orthogroup assignment filter step in OrthoFiller.

518 **Supplemental Figure 3: Recovery of removed genes from *S. cerevisiae* after 90% removal:**
519 **Representation of removed genes at each stage, filtered vs. unfiltered cases.** **A)** The number
520 of deleted genes that obtained hits from one or more orthogroup HMMs. **B)** The number of deleted
521 genes that had hits after OrthoFiller hint filtration. **C)** No hint filtration. **D)** The number of deleted
522 genes for which a gene prediction was made using Augustus that satisfied OrthoFiller filtration tests.
523 **E)** The number of deleted genes that for which a gene prediction was made using Augustus in the
524 absence of OrthoFiller filtration. **F)** The number newly predicted genes that were retained or
525 discarded based on the orthogroup assignment filter step in OrthoFiller.

526 **Supplemental Figure 4: Distribution of hint F-scores for good vs. bad hints.** Here, Augustus
527 has been allowed to predict genes that are already present in the input genome, hence we can
528 consider separately the good and bad hits as hints. Shown are the distributions of hint F-scores for

529 good (green) and bad (red) hits respectively, demonstrating that practically all of the genuine hints
530 have a hint F-score of 0.8 or higher.

531

532 **Tables**

533 **Table 1: Species Set A, fungal species used for algorithm validation**

Species Name	Source	Strain	Taxonomy ID	References
<i>Eremothecium gossypii</i>	JGI ¹	ATCC10895	284811	[9]
<i>Debaromyces hansenii</i>	JGI	CBS767	284592	[10] [11]
<i>Kluveromyces lactis</i>	JGI	CLIB210	284590	[10]
<i>Saccaromyces cerevisiae</i>	SGD ²	S288C	559292	[18]
<i>Yarrowia lipolytica</i>	JGI	CLIB122	284591	[10]

534 ¹Joint Genome Institute; ²Saccaromyces Genome Database

535

536 **Table 2: Species Set B, plant species used for algorithm validation**

Species Name	Source	Version	Taxonomy ID	References
<i>Arabidopsis thaliana</i>	JGI	TAIR10	3702	[12]
<i>Brassica rapa</i>	JGI	v1.3	3711	[12]
<i>Carica papaya</i>	JGI	ASGPBv0.4	3649	[12]
<i>Capsella rubella</i>	JGI	V1.0	81985	[12]
<i>Theobroma cacao</i>	JGI	V1.1	3641	[12]

537

538 **Table 3: Recovery of removed genes in *S. cerevisiae***

	10% annotations removed		90% annotations removed	
	OrthoFiller	Unfiltered	OrthoFiller	Unfiltered
No. genes removed	528	528	4759	4759
Total genes found	197	503	1529	4325
Found genes which overlap removed genes	196	447	1529	4156
Total recovered genes	196	440	1528	4116
Number of split genes	0	7	1	34
Mean pF score of found genes	0.99	0.97	0.99	0.97
Mean oF score of found genes	1.00	0.98	0.98	0.96
High-quality found genes (pF≥0.95)	190	411	1455	3801
Lower-quality found genes (pF<0.95)	6	36	74	355

539

540

Table 4: Recovery of removed genes in *A. thaliana*

	10% annotations removed		90% annotations removed	
	OrthoFiller	Unfiltered	OrthoFiller	Unfiltered
Mean pF-score of lower-quality genes	0.85	0.68	0.87	0.69
% of lower-quality genes with $\text{oF} \geq 0.95$	100.0%	69.4%	91.9%	68.2%
No. genes removed	3168	3168	28516	28516
Total genes found	1097	7048	10931	28793
Found genes which overlap removed genes	982	3484	10788	26004
Total recovered genes	908	1664	9840	12646
Number of split genes	67	850	889	6052
Mean pF score of found genes	0.77	0.46	0.71	0.46
Mean oF score of found genes	0.87	0.50	0.84	0.57
High-quality found genes ($\text{pF} \geq 0.95$)	416	491	3393	3539
Lower-quality found genes ($\text{pF} < 0.95$)	566	2993	7395	22465
Mean pF-score of lower-quality genes	0.60	0.37	0.57	0.37
% of lower-quality genes with $\text{oF} \geq 0.95$	56.5%	18.2%	50.1%	23.0%

541

542

Table 5: Novel genes in fungal species

Species Name	Genome size (Mbp)	No. pre-existing genes	No. new genes.
<i>E. gossypii</i>	9.10	4768	2
<i>D. hansenii</i>	12.15	6272	13
<i>K. lactis</i>	10.69	5076	6
<i>S. cerevisiae</i>	12.16	6572	2
<i>Y. lipolytica</i>	20.50	6447	8

543

544

Table 6: Novel genes in plant species

Species Name	Genome size (Mbp)	No. pre-existing genes	No. new genes.
<i>A. thaliana</i>	119.67	35386	116
<i>B. rapa</i>	315.05	43370	10
<i>C. papaya</i>	342.68	27793	382

<i>C. rubella</i>	134.83	28447	228
<i>T. cacao</i>	346.16	44404	94

545

546

547 **Table 7: SRA RNA-seq data coverage for novel genes in fungal genomes**

Species	SRA ID	Instrument/details	Genes in original annotation			Novel genes		
			Total	W/ reads	%	Total	W/ reads	%
<i>E. gossypii</i>	N/A ¹	N/A	N/A	N/A	N/A	N/A	N/A	N/A
<i>D. hansenii</i>	SRR899423	Helicos Heliscope, single	5106	6739	75.77	13	7	53.8
<i>K. lactis</i>	SRR1200528	Illumina Genome Analyzer II, single	5248	5251	99.9	6	6	100
<i>S. cerevisiae</i>	SRR539284	Illumina HiSeq 2000, paired end	6498	6572	98.87	2	2	100
<i>Y. lipolytica</i>	SRR868669	Illumina HiSeq 2000, single	7199	7520	95.73	8	7	87.5

548 ¹No publically available data found for this species

549

550 **Table 8: SRA RNA-seq data coverage for novel genes in plant genomes**

Species	SRA ID	Instrument/details	Genes in original annotation			Novel genes		
			Total	W/ reads	%	Total	W/ reads	%
<i>A. thaliana</i>	SRR3932355	Illumina HiSeq 2500, paired end. Wild type Columbia rep1	162699	197160	82.52	116	54	46.6
<i>B. rapa</i>	SRR2984945	Illumina HiSeq 2000, paired end. ga-deficient dwarf (gad1-2) +GA rep2	180358	218457	83.56	10	3	30.0
<i>C. papaya</i>	SRR3509576	Illumina HiSeq 2500, paired end. SunUp/Sunset cultivar, young hermaphrodite leaf	93345	112604	82.90	382	309	80.1
<i>C. rubella</i>	SRR797557	Illumina Genome Analyzer IIx, paired end	135008	148564	90.88	228	154	67.5
<i>T. cacao</i>	SRR3217315	Illumina HiSeq 2000, paired end. Flower/leaf sample	209407	264870	79.06	94	50	53.2

551

552

553 **References**

- 554 [1] E. C. Hayden, "The \$1,000 genome," *Nature*, vol. 507, p. 295, 2014.
- 555 [2] K. Wetterstrand, "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing
556 Program (GSP)," Mar-2016. [Online]. Available: www.genome.gov/sequencingcosts.
- 557 [3] M. Yandell and D. Ence, "A beginner's guide to eukaryotic genome annotation," *Nat. Rev.*
558 *Genet.*, vol. 13, no. May, pp. 329–342, 2012.
- 559 [4] J. F. Denton, J. Lugo-Martinez, A. E. Tucker, D. R. Schrider, W. C. Warren, and M. W. Hahn,
560 "Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies," *PLOS*
561 *Comput. Biol.*, vol. 10, no. 12, 2014.
- 562 [5] E. V. Koonin and M. Y. Galperin, "Genome Annotation and Analysis," in *Sequence - Evolution*
563 *- Function: Computational Approaches in Comparative Genomics.*, Boston: Kluwer Academic.
- 564 [6] A. van der Burgt, E. Severing, J. Collemare, and P. de Wit, "Automated alignment-based
565 curation of gene models in filamentous fungi," *BMC Bioinformatics*, vol. 15, no. 1, p. 19, 2014.
- 566 [7] D. M. Emms and S. Kelly, "OrthoFinder: solving fundamental biases in whole genome
567 comparisons dramatically improves orthogroup inference accuracy," *Genome Biol.*, vol. 16,
568 no. 1, p. 157, 2015.
- 569 [8] M. Stanke and B. Morgenstern, "AUGUSTUS: A web server for gene prediction in eukaryotes
570 that allows user-defined constraints," *Nucleic Acids Res.*, vol. 33, no. SUPPL. 2, pp. 465–467,
571 2005.
- 572 [9] F. S. Fred S. Dietrich, S. Voegeli, S. Brachat, A. Lerch, K. Gates, S. Steiner, C. Mohr, P.
573 Luedi, S. Choi, R. a Wing, A. Flavier, T. D. Gaffney, P. Philippsen, F. S. Dietrich, S. Voegeli,
574 S. Brachat, A. Lerch, K. Gates, S. Steiner, C. Mohr, R. Pöhlmann, P. Luedi, S. Choi, R. a
575 Wing, A. Flavier, T. D. Gaffney, and P. Philippsen, "The *Ashbya gossypii* Genome as a Tool
576 for Mapping the Ancient *Saccharomyces cerevisiae* Genome," *Science (80-.)*, vol. 304, no.
577 April, pp. 304–7, 2004.
- 578 [10] B. Dujon, D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny,
579 S. Blanchin, J.-M. Beckerich, E. Beyne, C. Bleykasten, A. Babour, J. Boyer, L. Cattolico, F.
580 Confanioleri, A. De Daruvar, L. Despons, E. Fabre, J. De Montigny, C. Marck, C. Neuvéglise,

- 581 E. Talla, N. Goffard, L. Frangeul, M. Aigle, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S.
582 Blanchin, J.-M. Beckerich, E. Beyne, C. Bleykasten, A. Boisramé, J. Boyer, L. Cattolico, F.
583 Confanioleri, A. De Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi,
584 F. Hantraye, C. Hennequin, N. Jauniaux, P. Joyet, R. Kachouri, A. Kerrest, R. Koszul, M.
585 Lemaire, I. Lesur, L. Ma, H. Muller, J.-M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-
586 Kalogeropoulos, S. Pellenz, S. Potier, G.-F. Richard, M.-L. Straub, A. Suleau, D. Swennen,
587 F. Tekaia, M. Wésolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M.
588 Bolotin-Fukuhara, A. Thierry, C. Bouchier, B. Caudron, C. Scarpelli, C. Gaillardin, J.
589 Weissenbach, P. Wincker, and J.-L. Souciet, "Genome evolution in yeasts," *Nature*, vol. 430,
590 no. 6995, pp. 35–44, 2004.
- 591 [11] C. Sacerdot, S. Casaregola, I. Lafontaine, F. Tekaia, B. Dujon, and O. Ozier-kalogeropoulos,
592 "Promiscuous DNA in the nuclear genomes of hemiascomycetous yeasts," *FEMS Yeast Res.*,
593 vol. 8, no. 6, pp. 846–857, 2008.
- 594 [12] D. M. Goodstein, S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks,
595 U. Hellsten, N. Putnam, and D. S. Rokhsar, "Phytozome: A comparative platform for green
596 plant genomics," *Nucleic Acids Res.*, vol. 40, no. D1, pp. 1178–1186, 2012.
- 597 [13] R. Leinonen, H. Sugawara, and M. Shumway, "The sequence read archive," vol. 454, pp. 1–
598 3, 2010.
- 599 [14] K. Katoh and D. M. Standley, "MAFFT Multiple Sequence Alignment Software Version 7 :
600 Improvements in Performance and Usability Article Fast Track," *Mol. Biol. Evol.*, vol. 30, no.
601 4, pp. 772–780, 2013.
- 602 [15] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity
603 searching," *Nucleic Acids Res.*, p. gkr367, 2011.
- 604 [16] D. H. Huson and C. Scornavacca, "Dendroscope 3: An interactive tool for rooted phylogenetic
605 trees and networks," *Syst. Biol.*, vol. 61, no. 6, pp. 1061–1067, 2012.
- 606 [17] D. M. Stasinopoulos and R. A. Rigby, "Generalized additive models for location scale and
607 shape (GAMLSS) in R," *J. Stat. Softw.*, vol. VV, no. li.
- 608 [18] S. Gnerre, I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, L. Williams,
609 R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, D. B. Jaffe, T. Sharpe, G. Hall, T. P. Shea, S.

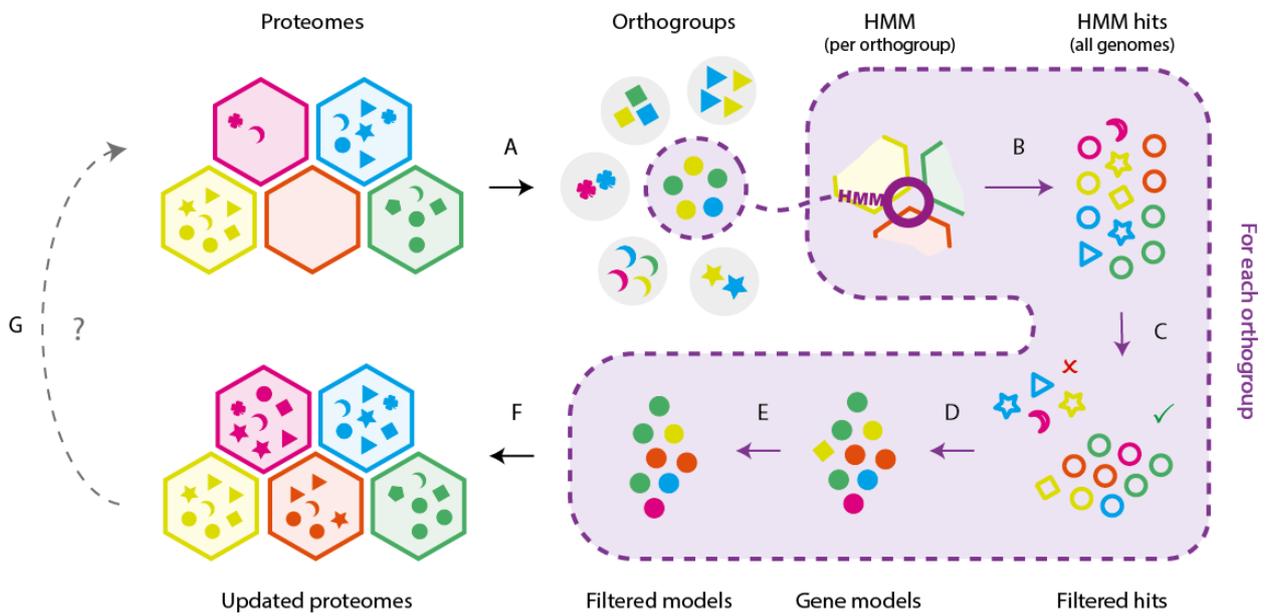
610 Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C.
611 Nusbaum, E. S. Lander, and D. B. Jaffe, “High-quality draft assemblies of mammalian
612 genomes from massively parallel sequence data,” *PNAS*, vol. 108, no. 4, pp. 1513–1518,
613 2011.

614

615

616 **Figures**

617 **Figure 1: OrthoFiller workflow**

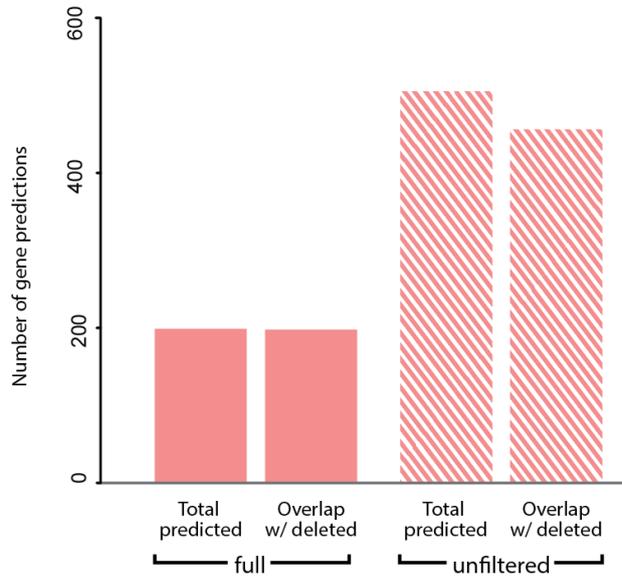


618

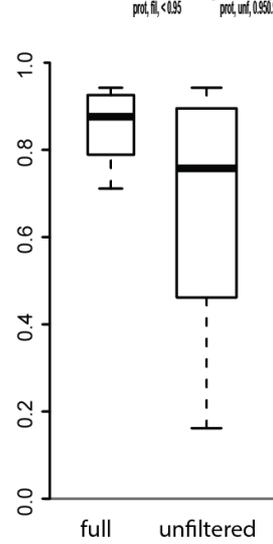
619

620 **Figure 2: Performance of OrthoFiller on *S. cerevisiae* genome with 10% of annotated**
621 **genes removed.**

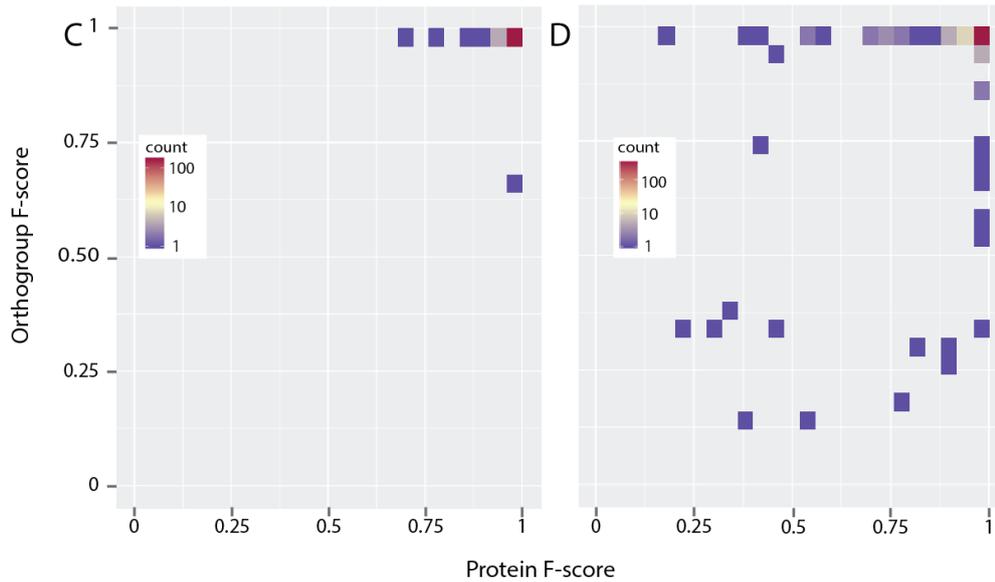
A: 10% annotations deleted: number of recovered genes



B: Distr. of low quality proteins



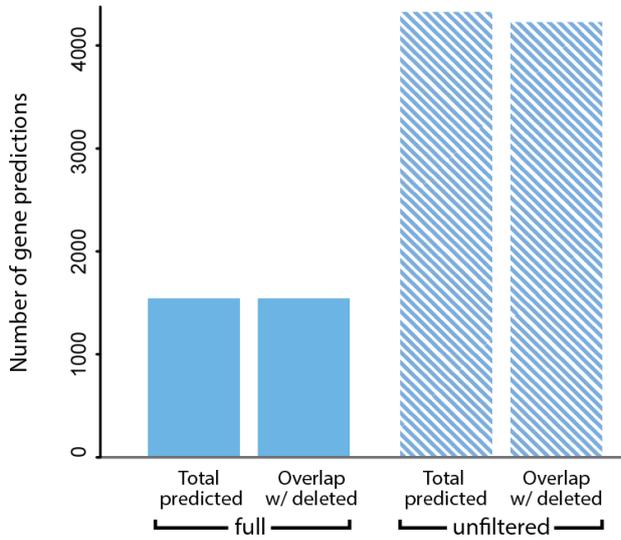
C-D: Orthogroup F-score & protein F-score comparison, filtered vs unfiltered.



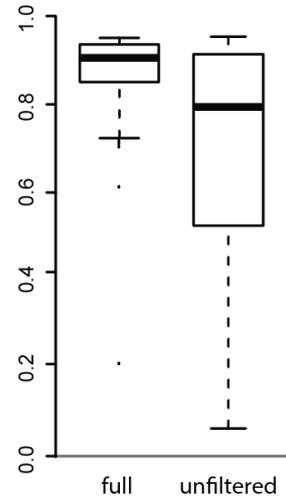
622

623 **Figure 3: Performance of OrthoFinder on *S. cerevisiae* genome with 90% of annotated**
624 **genes removed.**

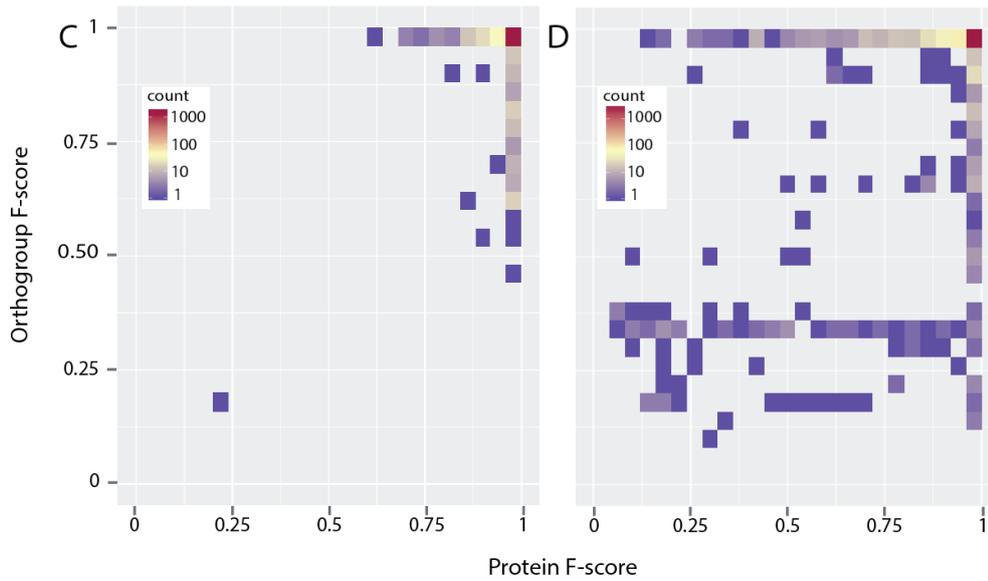
A: 90% annotations deleted: number of recovered genes



B: Distr. of low quality proteins



C-D: Orthogroup F-score & protein F-score comparison, filtered vs unfiltered.

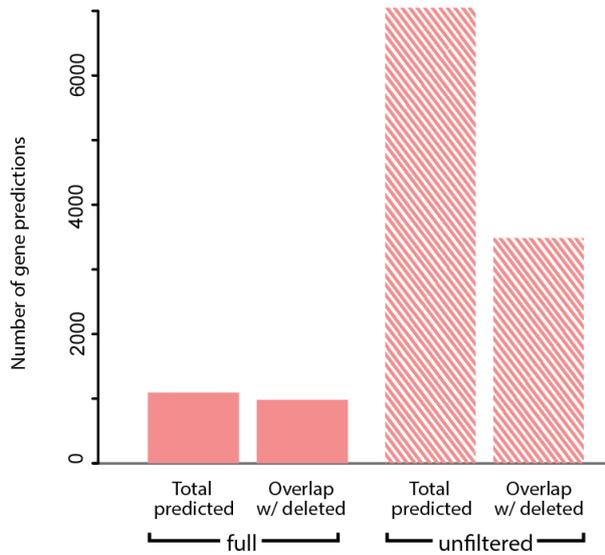


625

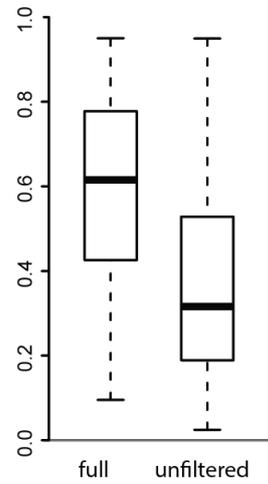
626

627 **Figure 4: Performance of OrthoFiller on *A. thaliana* genome with 10% of annotated**
628 **genes removed.**

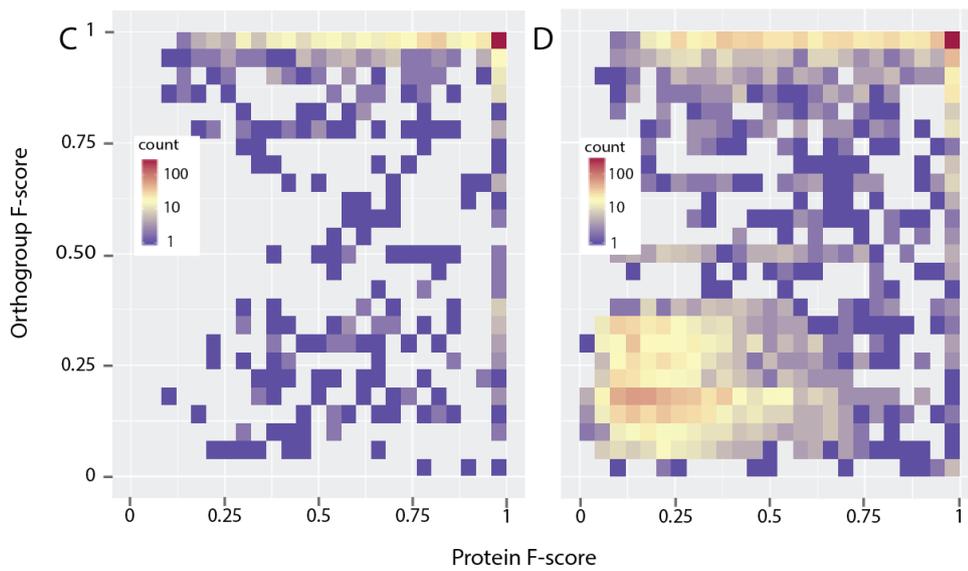
A: 10% annotations deleted: number of recovered genes



B: Distr. of low quality proteins

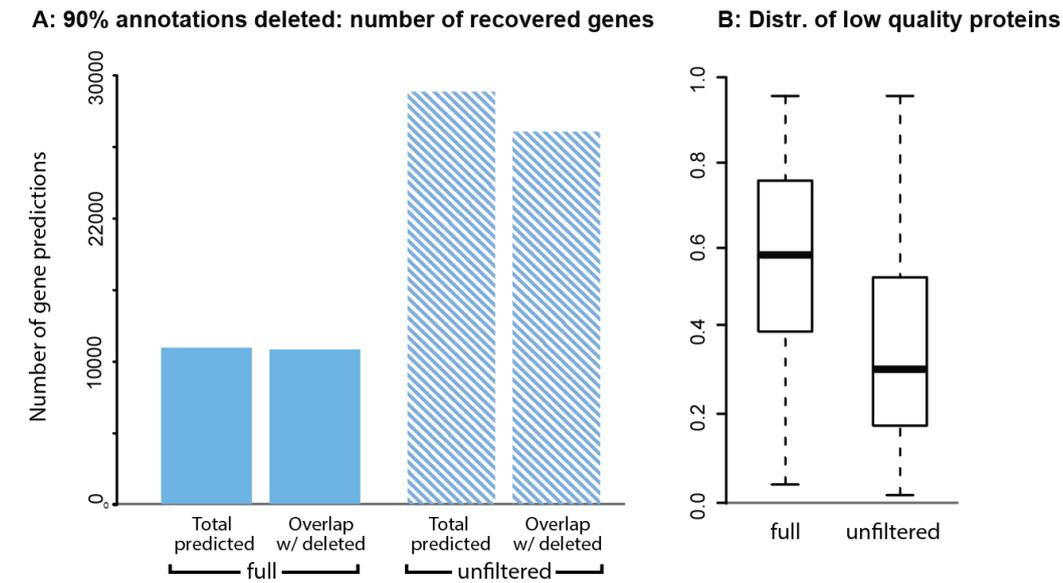


C-D: Orthogroup F-score & protein F-score comparison, filtered vs unfiltered.

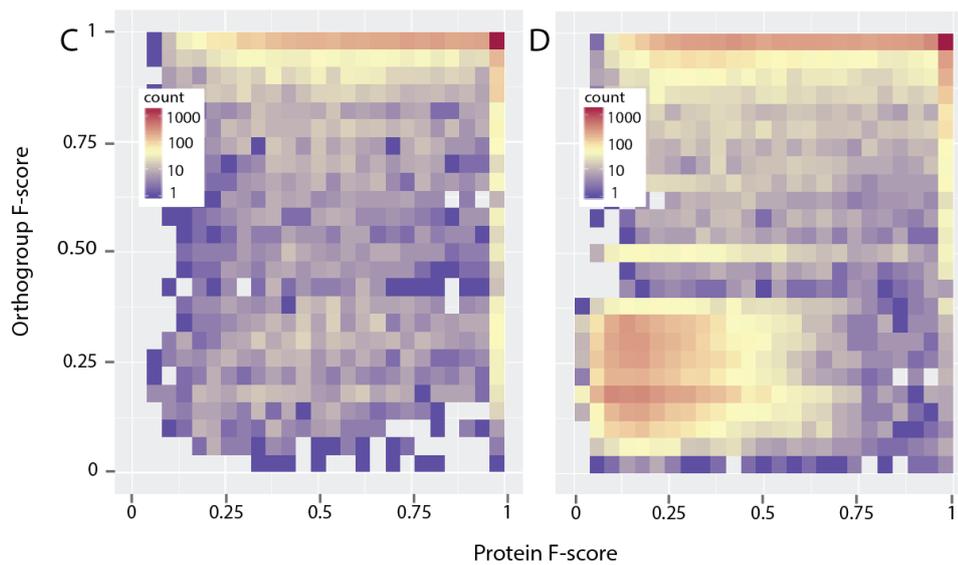


629

630 **Figure 5: Performance of OrthoFiller on *A. thaliana* genome with 90% of annotated**
631 **genes removed.**



C-D: Orthogroup F-score & protein F-score comparison, filtered vs unfiltered.

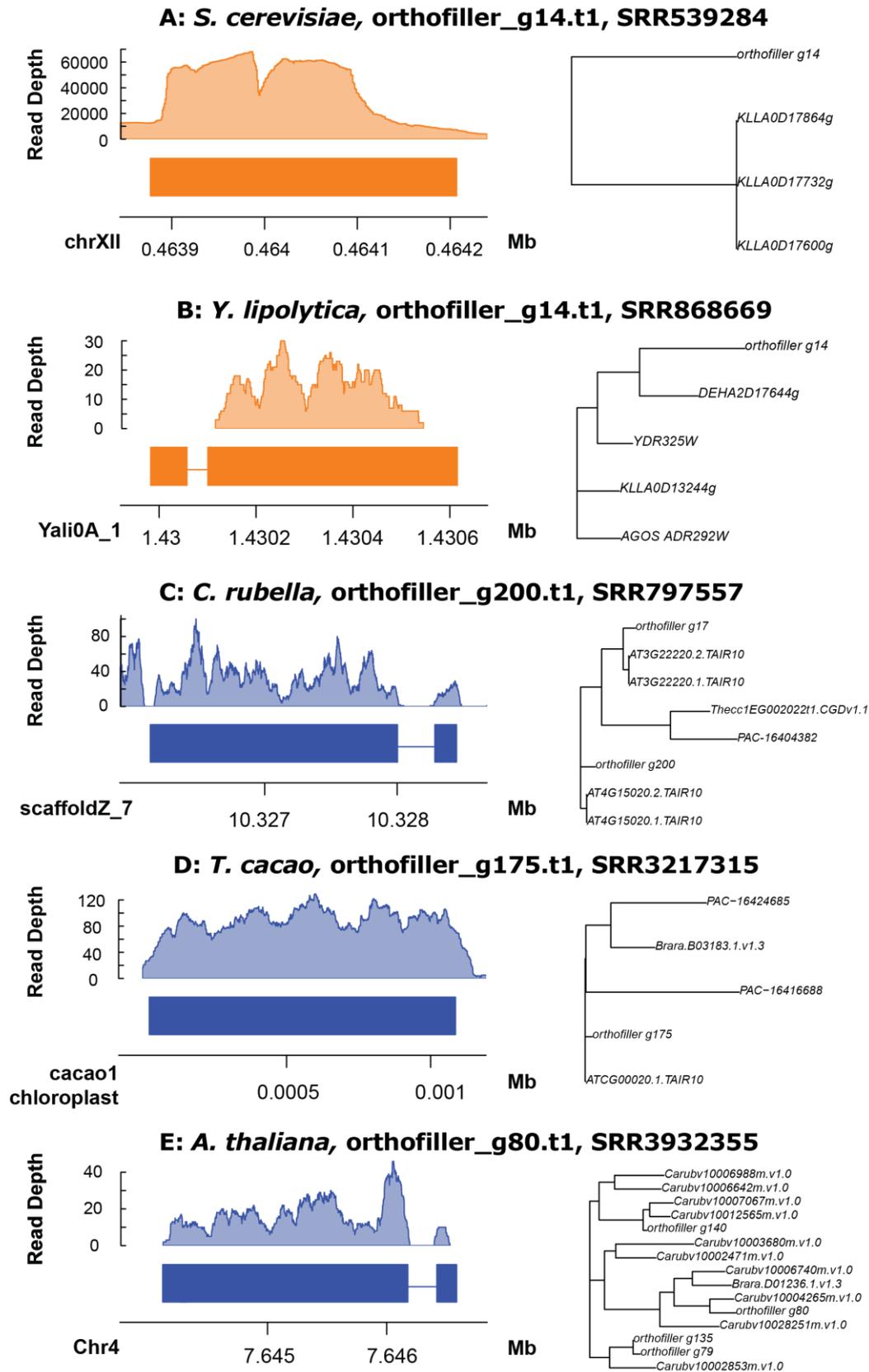


632

633

634

635 **Figure 6: Coverage plots and orthogroup trees for a selection of new genes.**

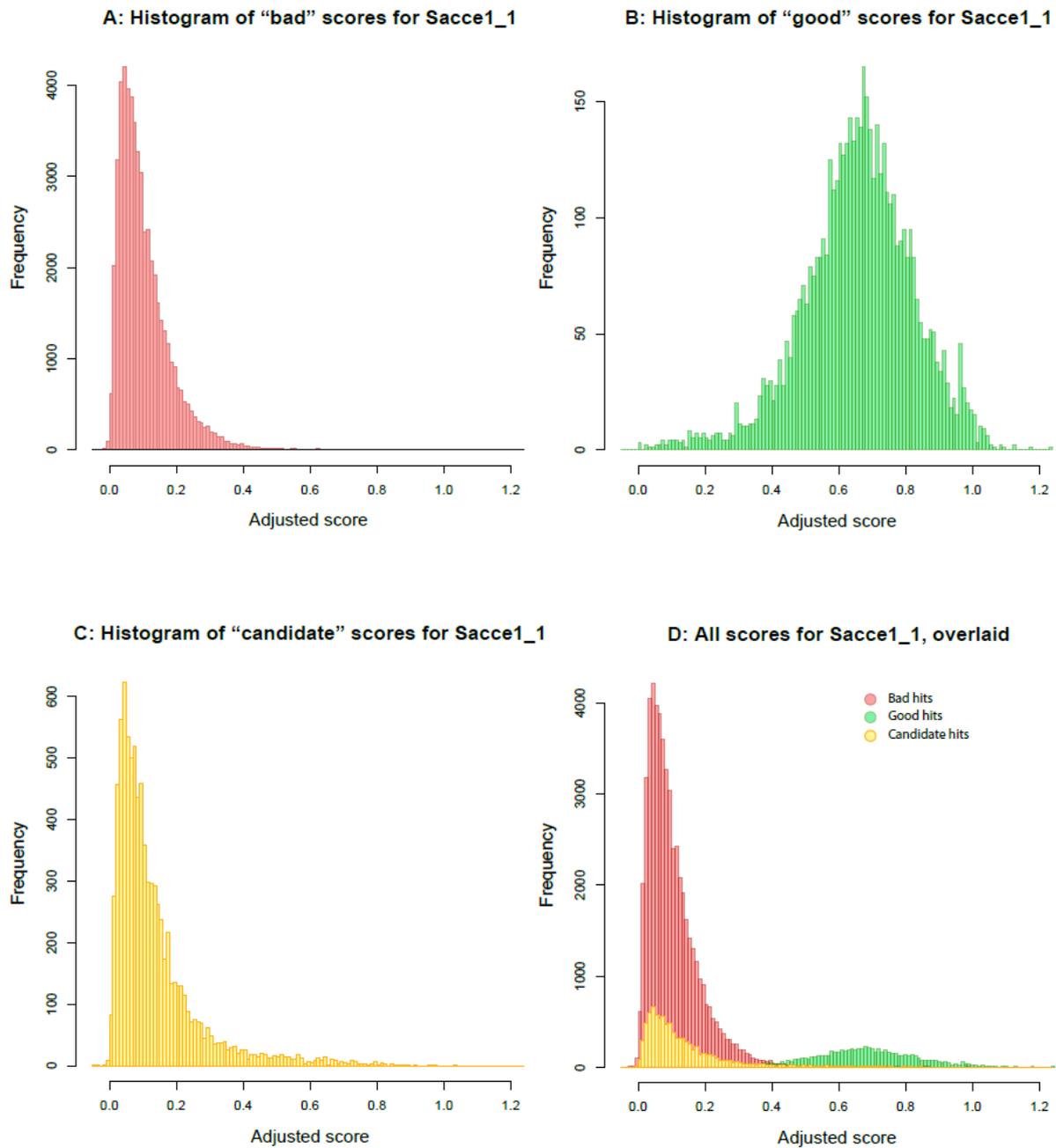


636

637

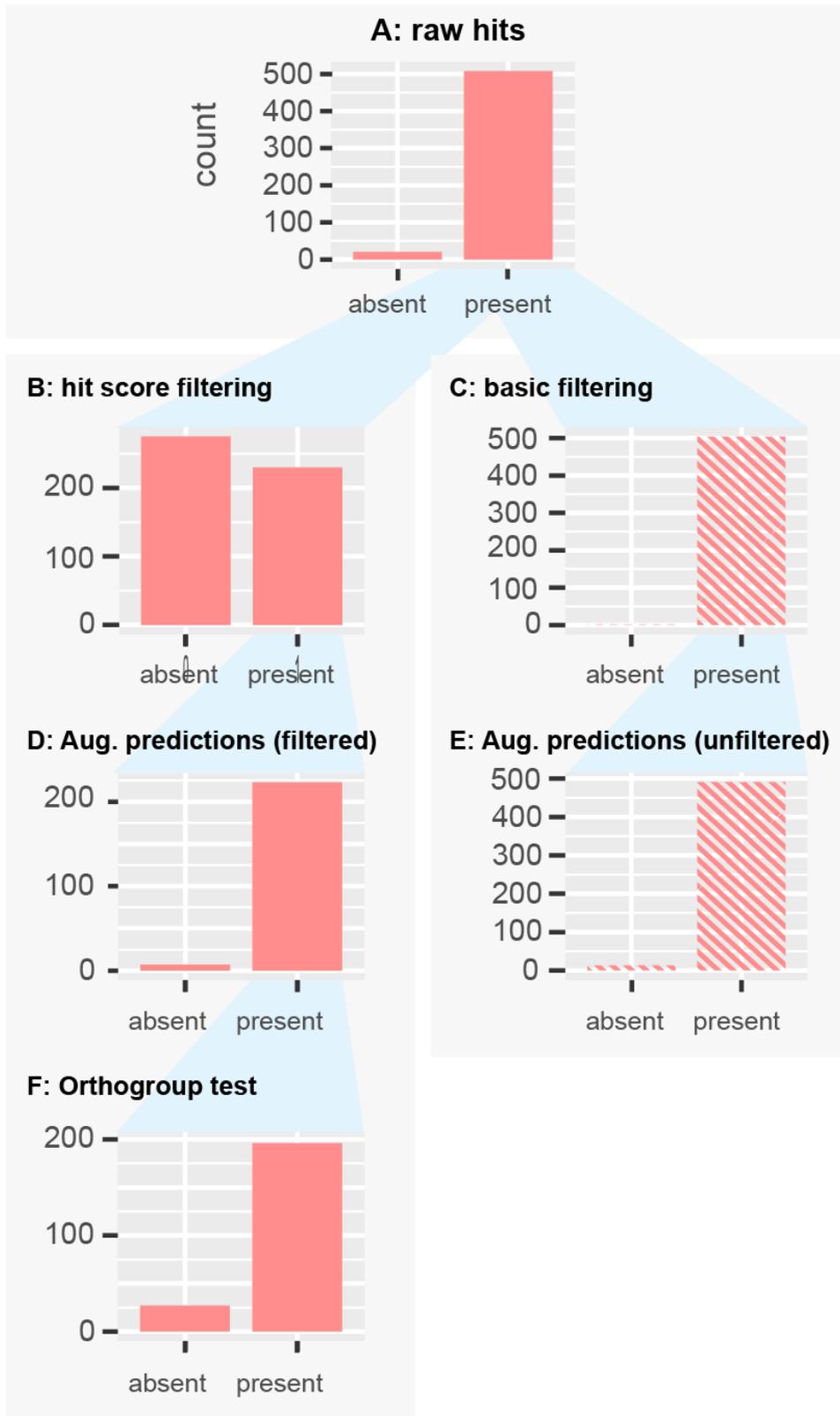
638 **Supplemental Figures**

639 **Supplemental Figure 1: hit score distributions for good, bad and candidate hits**



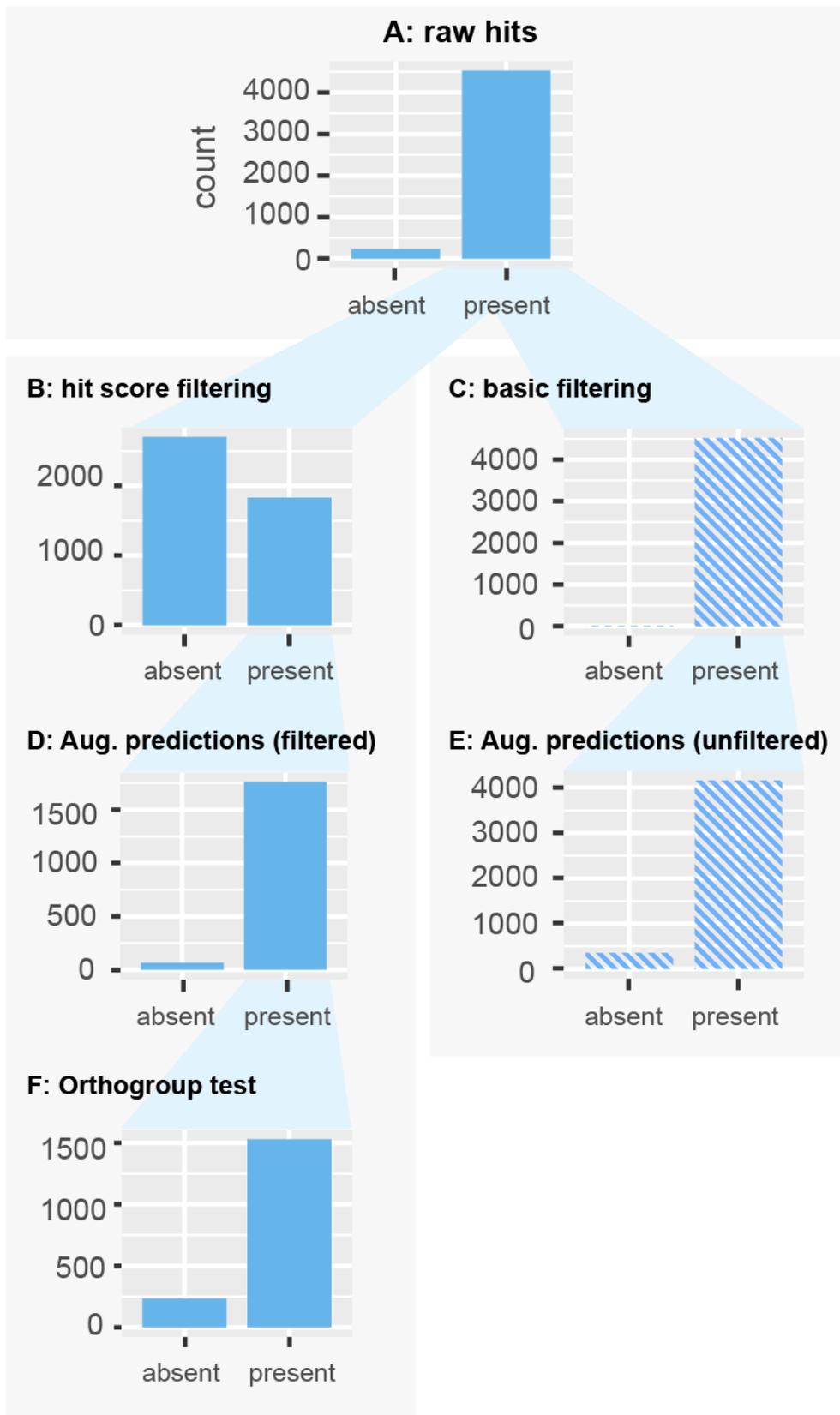
640

641 **Supplemental Figure 2: Recovery of removed genes from *S. cerevisiae* after 10% removal: Representation of removed genes at each stage, filtered vs. unfiltered cases**
642



643

644 **Supplemental Figure 3: Recovery of removed genes from *S. cerevisiae* after 90% removal: Representation of removed genes at each stage, filtered vs. unfiltered cases**
645



646

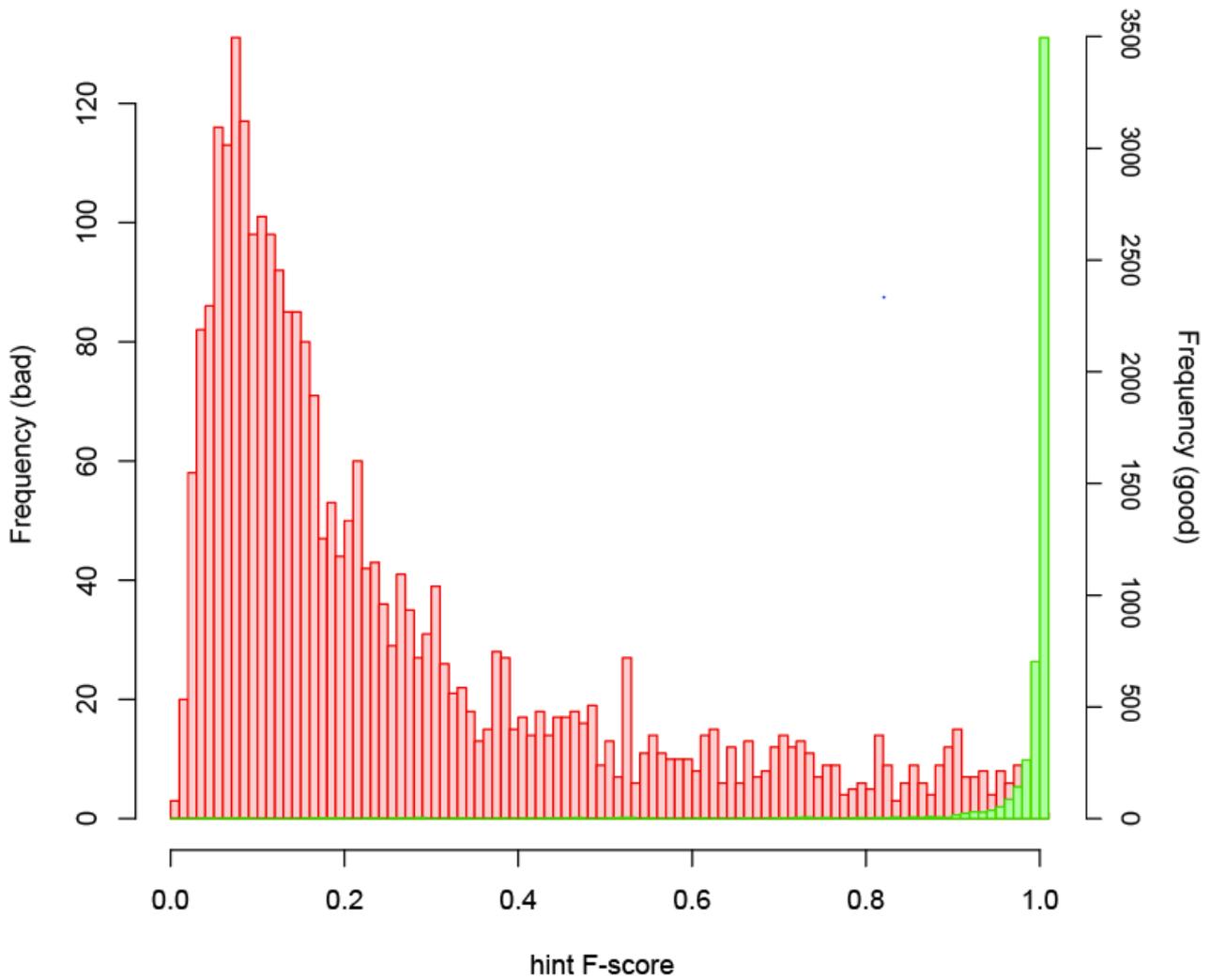
647

648

649

650

651 **Supplemental Figure 4: Distribution of hint F-scores for good vs. bad hints**
652



653