1   **FGMP: assessing fungal genome completeness and gene content**

2

3   Ousmane H. Cissé[1,#] and Jason E. Stajich[1]

4

5   [1] Department of Microbiology & Plant Pathology and Institute for Integrative Genome

6   Biology, University of California-Riverside, Riverside, CA 92521 USA

7

8   [#] Present address: Critical Care Medicine Department, NIH Clinical Center, National

9   Institutes of Health, Bethesda, Maryland, USA

10

11

12   Abstract

13   **Background:** Inexpensive high-throughput DNA sequencing has democratized access to

14   genetic information for most organisms so that research utilizing a genome or

15   transcriptome of an organism is not limited to model systems. However, the quality of the

16   assemblies of sampled genomes can vary greatly which hampers utility for comparisons

17   and meaningful interpretation. The uncertainty of the completeness of a given genome

18   sequence can limit feasibility of asserting patterns of high rates of gene loss reported in

19   many lineages.

20   **Results:** We propose a computational framework and sequence resource for assessing

21   completeness of fungal genomes called FGMP (Fungal Genome Mapping Project). Our

22   approach is based on evolutionary conserved sets of proteins and DNA elements and is

23   applicable to various types of genomic data. We present a comparison of FGMP and

24   state-of-the-art methods for genome completeness assessment utilizing 246 genome

25   assemblies of fungi. We discuss genome assembly improvements/degradations in 57

26   cases where assemblies have been updated, as recorded by NCBI assembly archive.

27   **Conclusion:** FGMP is an accurate tool for quantifying level of completion from fungal

28   genomic data. It is particularly useful for non-model organisms without reference

29   genomes and can be used directly on unassembled reads, which can help reducing

30   genome sequencing costs.

31   Keywords: assembly; conserved elements; gene model.

32

33 **BACKGROUND**

34 The recent explosion of high-throughput sequencing methods and analytic tools has made

35 sequencing easier and cheaper for nearly all species across the tree of life including

36 uncultivable organisms. However, the quality and completeness of these genomes can

37 vary due to challenges in assembling repeat rich regions and variable or insufficient

38 sequencing coverage (1). Large-scale sequencing projects such as the microbial dark

39 matter project (2), the Human Microbiome Project (3) or the 1000 fungal genomes

40 project (http://1000.fungalgenomes.org) have produced thousands of microbial genome

41 assemblies. The rapid generation and release of draft data is contributing important and

42 useful datasets that are extensively used for studies of pathology, evolution, and

43 discovery of enzymes or pathways. Variable quality and completeness of draft genomes

44 can impact the inferences drawn regarding gene content, transposable element load, and

45 genome size. There is a need to quantify a genome's completeness to provide context of

46 the quality of information that can be inferred from it. This work is also motivated by

47 observations that lineage specific gene loss is an important driving force in evolution,

48 especially in fungi (4, 5), and the accuracy of conclusions drawn about the patterns of

49 missing genes requires comparisons among similar quality genomes.

50 Approaches to assess the quality and completeness of a genome have been

51 proposed using nearly 100 different metrics (6). Unfortunately, most of these metrics are

52 generally not applicable to non-model species because they require a substantial amount

53 of additional high-quality data (e.g. fosmids, reference genomes, optical maps) that can

54 be expensive or infeasible to obtain for a large number of samples. Currently, few

55 methods attempt to estimate the amount of missing data in an assembly without prior

3

56   knowledge. One of the most popular approaches, CEGMA estimates the completeness to

57   the presence of set of 248 single copy gene markers (7, 8). Although CEGMA has been

58   used in numerous studies, a key issue is that makers were selected from only six model

59   eukaryotic species and the ubiquity and detections of these markers may not be consistent

60   as more distant lineages are sampled. CEGMA has been recently discontinued and the

61   authors recommend using alternative tools (http://www.acgt.me/blog/2015/5/18/goodbye-

62   cegma-hello-busco). The concept has been recently revisited and updated with clade-

63   focused sets of protein coding gene markers in BUSCO (9). Another set of 246 single

64   copy fungal gene families has been proposed by FUNYBASE (10). The latter provides a

65   set of conserved fungal genes but the tools are not explicitly developed to assess genome

66   completeness. Furthermore, the FUNYBASE database was generated in 2010 while a

67   broader sampling of diverse fungal genomes is now available (11).

68   To build a dataset of independent markers to assess completeness, typically,

69   single copy orthologous genes are chosen. Multi-copy gene families are systematically

70   filtered out in these selections, but their utility, as well as that of alternative, non-protein

71   coding gene markers has not been fully explored in assessing genome completeness. Two

72   summary statistics of genome assemblies are frequently used to evaluate quality

73   completeness. The N50 and L50 statistics (12) which describe the level of fragmentation

74   of the assembly are computed based on the lengths of assembly scaffolds or contigs. Both

75   statistics utilize a sorted list of largest to smallest sizes of contigs, where L50 is the length

76   (in bases) of the shortest contig for which 50% of the genome can be contained within

77   contigs of that size or larger, and N50 is the number of contigs that when summed their

78   length is half of the assembly size (13). Note that unfortunately these two concepts are

4

79   swapped in some tools, where N50 means length and L50 means the count. Still other

80   methods measure the number error per bases or assembly inconsistencies to prediction

81   genome quality (14, 15).

82   In the present study, we focused on the fungal kingdom. Fungal genome sizes

83   vary from several megabases (Mb) to nearly 1,000 Mb (11). A primary motivation of this

84   work is to provide a realistic estimation of assembly completeness for fungal genomes.

85   The precision dependents on the ability to accurately identify genes, which can appear

86   artifactually fragmented by an incomplete assembly or appear lost due to more rapidly

87   evolving loci in some lineages. The nature, evolutionary trajectory and loss likelihood of

88   genes need to be considered when calculating genome completeness from gene content.

89   We propose a novel set of markers and build a pipeline to assess their presence in

90   genome assemblies called FGMP (Fungal Genome Mapping Project). Our multistep

91   approach extends previous approaches by integrating identifiable fungal protein and

92   highly conserved non-coding regions. The protein markers selected include both single

93   and multi-copy markers and have only a 50% overlap with previously published datasets

94   providing a different dimension of sequence evolution to evaluate the completeness.

95   Highly conserved non-coding regions of fungal genomes are a novel resource we have

96   developed and incorporated into assessment of genome completeness in FGMP. Lastly,

97   we use a multisampling approach coupled to a rarefaction analysis to search for markers

98   in unassembled sequencing reads, which bypass the need for an assembly. Therefore,

99   using FGMP, a researcher can quickly assess the quality of a set of reads in hand before

100   attempting an assembly, which can be computationally expensive. Finally, we described a

101   side-by-side comparison of our tool with state-of arts methods over 246 fungal species

5

102   with genome assemblies of varying ranges of quality. We captured assembly

103   improvements/degradations in 57 fungal species with more than one released assembly,

104   as recorded in NCBI assembly archive. The modular construction of this work can be a

105   valuable tool for genome completion estimation that can be easily incorporated in more

106   complex pipeline.

107

108   **IMPLEMENTATION**

109   A typical run of FGMP consists of three steps. First, a set of raw gene models (proteins)

110   is generated from the queried assembly which are further filtered down to high

111   confidence genes in subsequent steps. Second, the presence of highly conserved non-

112   coding fungal DNA elements (>200 nucleotides) is estimated. Third, the copy number of

113   ubiquitous multi-copy protein families is determined to track possible mis-assemblies or

114   collapsed duplicated regions. The FGMP workflow is diagrammed in Figure 1 and

115   methodology further detailed in the following sections.

116

117   **Reference data preparation**

118   FGMP is primarily designed for assessment of fungal genome quality using defined sets

119   of conserved proteins, noncoding highly conserved DNA elements (HDE) and multi-copy

120   protein families. All the datasets are included in FGMP package and a stable released

121   version is available at DOI: 10.5281/zenodo.1453438. Installation is available via

122   Bioconda package system as "conda install -c bioconda fgmp" (16). Alternatively, a step

123   by step manual installation guide is provided at https://github.com/stajichlab/FGMP.

124   To generate the protein markers, we analyzed a phylogenomic dataset of 25 fungi

125   with complete genomes covering from all major fungal lineages (Additional file 1). In

126   total, 164,232 proteins were analyzed. Our goal is to capture the fungal protein diversity

127   rather than focusing on universally conserved proteins. We focused on obtaining a set of

128   diverse proteins to be used for initial identification of candidate regions and training sets

129   for gene predictions. Our assumption is that irrespective to the phylogenetic classification

130   of species under analysis, our diversified set of proteins would contain a homolog with

131   sufficient protein similarity to generate a valid gene model. With these concepts in mind,

132   we identified orthologous protein families using OMA (17) followed by inspection of the

133   ortholog clusters using BLAST (18) and full-length pairwise alignments generated using

134   needle from EMBOSS package (19). We extracted 7,773 protein families present in at

135   least four species and use them to construct Hidden Markov Models (HMMs) using

136   HMMER3 (20). In parallel, we selected a single most informative protein in each of these

137   families using M-COFFEE (21), which corresponds to the sequence containing

138   information that is lacking in other sequences of a multiple sequence alignment. We

139   computed the significance scores as follows: each protein of each cluster was compared

140   to its corresponding HMM and the threshold corresponds to 80% of the score of the

141   protein with the lowest score. The use of the full set of 7,773 proteins appeared to be

142   excessively demanding in term of computational resources, even for small sized genome

143   assemblies (e.g. < 8 megabases; data not shown). To reduce the computational burden,

144   we then filtered out potentially paralogous sequences using PHMMER with an E-value of

145   10^{-50} as cut off (20), and applied the following rules: (i) a marker should be present in at

146   least 99% of the species and (ii) should be unambiguously identifiable based on the

7

147   alignment score of the protein against the HMM of the family. This filtering reduces our

148   markers dataset to 593 proteins of which 60.3% are from single copy genes contrasting

149   other published strategies, which exclude multi-copy gene families (BUSCO and

150   CEGMA). These 593 representative protein sequences are aligned to the queried genome

151   assembly to identify genomic regions that encode homologous genes using tBLASTn

152   (18). Once candidate regions are narrowed down by these translated alignment searches,

153   fine-grain alignments of the proteins to these homologous regions in the target genome

154   are generated using splice-site aware protein2genome alignment with EXONERATE

155   (22). These alignments-based gene models are used as training sets for AUGUSTUS (23).

156   The predicted proteins including both AUGUSTUS gene models and translated

157   EXONERATE alignment matches are then searched against 593 HMMs to identify the

158   originating genes. FGMP assigns confidence in these predictions based on pre-defined

159   thresholds. We benchmarked FGMP using the full (7,773) and reduced set of proteins

160   (593) on our subsequent analyses and found no significant differences in completeness

161   estimates between the two sets of proteins.

162   To identify highly conserved non protein-coding fungal DNA elements, we

163   performed pairwise whole genome alignments of nine fungi using LAST (24). The

164   phylogeny of the selected species is presented in the additional file 2. Coding regions

165   were removed from alignments based on NCBI annotations using BEDtools (25). The

166   filtering was carried out using a computational pipeline combining enriched motifs and

167   alignments from MEME (26), BLASTn (18) and EMBOSS 'needle' (19). A total of 31

168   non-coding highly conserved regions in each species were extracted with a requirement

169   that loci be at least 200 nucleotides long with a minimum of 70% global identity. These

8

170   alignments were converted into HMMs using NHMMER from HMMER version 3.1b2.

171   These genomic segments are specific to fungi because no significant hit was detected to

172   species outside the kingdom using NHMMER with an E-value of 0.1 as cut off against a

173   custom eukaryotic genome database containing *Oomycetes*, *Leishmania* and *Plasmodium*

174   species.

175   To identify ubiquitous multi-copy protein families within our set of 593 proteins,

176   we surveyed 345 fungal genomes. Thirty-three protein families appear to have more than

177   one copy in all the genomes (HMMSEARCH, E-value $< 10^{-50}$). For each of these 33

178   proteins, we consider the minimum number of copies. FGMP records the number of

179   copies of these 33 proteins and reports when the copy number is lower than expected.

180   Lastly, FGMP can estimate the level of completion directly from raw sequences

181   using an iterative reservoir sampling approach. The process starts by splitting the set of

182   reads by chunks of 10^4 sequences. Then, FGMP randomly selects 1000 chunks using a

183   reservoir sampling approach. This parameter can be modified by the user. Chunks of

184   sequences are iteratively screened for presence of 593 protein makers using BLASTx

185   (24). The number of markers detected is recorded at each iteration. FGMP will stop after

186   20 successive unsuccessful attempts to detect new markers.

187

188   **RESULTS**

189   **Protein markers comparison**

190   To determine if there is overlap among protein markers, we compared FGMP protein

191   markers to proteins used by other tools. Noting that CEGMA has been recently

192   discontinued and FUNYBASE is outdated, this comparison is only for an historical

193   perspective. A total of 7,773 FGMP markers were originally obtained, which was

194   reduced to 593 after the removal of ambiguous markers (see Reference data preparation).

195   We compared the markers selected for FGMP (593 proteins) to those used in CEGMA

196   (248 families, 1,488 proteins), BUSCO fungi (1,438 proteins) and FUNYBASE (246

197   families, 5,166 proteins). Using reciprocal best BLASTp (E-value $< 10^{-5}$), 49.5% of

198   FGMP protein markers are not found in the other datasets whereas the proportions of

199   unique markers using the same criterion in CEGMA, FUNYBASE and BUSCO are

200   respectively 21.7%, 10.5% and 69.8% (see additional file 3). FGMP proteins tend to be

201   conserved in other eukaryotes but their utility outside the fungal kingdom is not explored

202   in the present study. Transferases and transporters are common (13%). Kinases and

203   helicases are overrepresented in FGMP protein dataset where they represent 10% and 5%

204   of 593 protein markers, respectively as compared to 0.8% and 2% of CEGMA makers;

205   3.3% and 2% in FUNYBASE markers; 3.3% and 0.7% of BUSCO fungi markers.

206   Kinases and helicases are multi-copy protein families in nearly all fungi and likely this

207   multicopy property is why these genes are not present in other datasets, which actively

208   restrict gene duplicates. Most FGMP kinases have homologs in bacteria and archaea,

209   suggesting that they are ancient. Most of the helicases also have archaeal or bacterial

210   homologs as well and are likely a mix of ancient and derived forms.

211

212   **Comparison with related tools**

213   To investigate the ability of different software to detect changes in genome assemblies'

214   quality, we analyzed the initial and subsequently updated genome assemblies of 45 fungi.

215   As FGMP measures the genome completeness using protein and highly conserved non-

10

216 coding elements, we refer these modules as FGMP_PROT and FGMP_HCE,

217 respectively. This is because FGMP_PROT is roughly equivalent to BUSCO or CEGMA,

218 but FGMP_HCE is unique to FGMP. FGMP_PROT and FGMP_HCE predict increases

219 in genome completeness in 35% and 31% of the 45 species, respectively whereas

220 BUSCO-fungi predicts increases in 53% and CEGMA in 60% of the species (Figure 2

221 panels a to d). Overall, the compared methods agree on 16 out of 45 species (Figure 2

222 panel e). FGMP is the most conservative method at assigning increase in completeness

223 between versions and CEGMA is the most permissive. No statistically significant

224 correlation between FGMP and BUSCO results and various genome statistics was

225 observed (i.e. N50, sequencing coverage, or the sequencing technology used), which is

226 consistent with Assemblathon results (6) showing that completeness metrics are not

227 necessarily correlated among genome assembly statistics. However, CEGMA results

228 appeared to be correlated to the sequencing technology used (Spearman rho = 0.2), that

229 is, assemblies generated exclusively with short reads (e.g. Illumina) tend to have lower

230 rates of increased completeness between versions than those built using long reads (e.g.

231 PacBio). FGMP_PROT and FGMP_HCE results are correlated (Spearman rho = 0.39)

232 but clearly independent, which further highlights the utility of interrogating different

233 genomic regions to assess completeness.

234   To further assess the ability of FGMP to detect missing genes and gene loss, we

235 evaluated the impacts of randomly removing ~10% of genomic sequence using 57 fungal

236 genome assemblies (completeness estimates are provided in additional file 4).

237 FGMP_PROT successfully captures the degradations in all assemblies, the average loss

238 rate was estimated at 5% instead of the original 10%, which means the full extent of the

239   simulated loss is not recovered. FGMP_HDE detected the loss of genomic regions in 54

240   assemblies with an average loss rate of 5.2%. A search with BUSCO (fungi models)

241   captures degradations in 55 assemblies with an average loss rate of 4.3%. The removal of

242   genomic regions prevented CEGMA pipeline from completing in many cases without

243   apparent reasons. Therefore, we have concluded that CEGMA could only be run on the

244   original genome assemblies. Examination of CEGMA completeness estimates found they

245   are not statistically different from BUSCO (fungi models) estimations (Wilcoxon test, p-

246   value 0.73), but differ significantly from FGMP results (p-value = 0.01). These findings

247   indicate FGMP and BUSCO perform relatively well on genome assemblies of varying

248   degrees of completeness and are similar in estimations detecting degradations in genomes

249   (Figure 3).

250

251   **Genome completion and ecological or lifestyle traits**

252   We estimated the genome completeness of 166 fungal genomes (species details are

253   described in additional file 7). Only one version of the genome assembly was considered

254   for each species. Each species was classified according to its lifestyle based on published

255   literature (e.g. saprotroph, parasite; references are presented in additional file 7). Parasites

256   are typically characterized by a reduced genome size usually attributed their reliance,

257   either partially or obligately, on nutrients scavenged from hosts. Their genomes are often

258   enriched with transposable and repetitive elements, which in some extreme case

259   composed more than 80% of the genomes (27). Our dataset includes 34 pathogenic

260   species which genome sizes range from 177.6 Mb for the ectomycorrhizal fungus

261   *Cenococcum geophilum* to 2.1 Mb for the microsporidia *Encephalitozoon romaleae*. The

12

262   remaining 132 fungi were classified as saprotopic and their genome sizes vary from 177

263   Mb for the ectomycorrhizal fungus *Cenococcum geophilum* to 9.8 Mb for the xerophilic

264   fungus *Wallemia sebi*. Taking the whole set of genomes, the average N50 is 126.7 Mb for

265   an average number of scaffolds per genome of 1,029; an average genome size of 38 Mb

266   and the average fraction of Ns per genome is 3.2%.

267   Analysis of our set of 166 genome assemblies found that 92% have a CEGMA

268   value > 95% whilst only 58.7% of these assemblies have a 95% completeness with

269   BUSCO fungi, 40% with FGMP_HDE and 54.2% with FGMP_PROT. Genomes labeled

270   as incomplete are typically parasites, which suggest that gene losses from genome

271   streamlining and missing sequence in assemblies might be confounded. Overall

272   completeness predictions correlated with the N50: CEGMA (Spearman rho = 0.35, *P*-

273   value = 1.3 x 10^{-8}), BUSCO fungi (R = 0.40; *P* = 2.1 x 10^{-11}), FGMP_HDE (R = 0.17; *P*

274   = 0.005) but FGMP_PROT (R = -0.05; *P* = 0.4). FGMP_PROT predictions are not

275   correlated with N50 as this metric incorporates gene fragments, which allow the partial

276   detection of markers even when reliable gene models cannot be built. To avoid

277   overcounting false positives and inflating the estimate, gene fragment sequences are

278   required to score above a predetermined threshold to be accepted as valid hit. However,

279   because short fragments are still required to display a significant similarity versus FGMP

280   protein markers (scores), the likelihood of inflated completeness estimates is expected to

281   be negligible.

282

283

284

13

285 **Runtimes**

286 We tracked the running times for 90 fungal genome assemblies (additional file 6). Using

287 six CPU cores each with 8 GB memory (AMD Opteron clock speed 2.1Ggz), FGMP

288 runtimes are proportional to the size and the levels of fragmentation of the genomes

289 under analysis. Runtimes are more influenced by the level of fragmentation of the

290 assembly than its size (additional file 5). For example, FGMP analysis is completed in 39

291 minutes for the 118 Mb genome size of *B. graminis* (N50 = 2,030. 3 kb) whereas the

292 analysis of the 41 Mb genome of *Magnaporthe oryzae* (N50 = 153 kb) requires three

293 hours (additional file 6). The fastest runtime observed was that of *Cryptococcus gattii*

294 (assembly version 1, size 17.1 Mb, N50 = 44 kb) completed in 22 minutes and the longest

295 with the genome of analysis of the 49.9 Mb genome of *Hortea werneckii* (N50 153 kb)

296 which required four hours.

297

298 **DISCUSSION**

299 FGMP is a useful tool for automated assessment of genome assembly completeness of

300 fungal genomes that incorporates measures of gene content covering both protein coding

301 and noncoding regions. The tool combines multilevel analysis by scanning of both coding

302 and non-coding regions of a given genome and provides a detailed reported describing

303 the recovery of multiple types of genomic features in a genome assembly. Compared to

304 existing methods BUSCO and CEGMA, FGMP fills a unique niche by assessing non-

305 coding highly conserved segments and collapsed gene family's content in addition to

306 measure of protein coding gene conservation. Additionally, FGMP does not rely

307 exclusively on *ab inito* gene predictions with tools like AUGUSTUS which require

14

308   parameter training. FGMP reports complete, partial and aberrant gene models. FGMP

309   also includes an experimental module, which allow a user to query raw reads using a

310   reservoir sampling approach. This module is currently optimized for low input long reads

311   similar to PacBio or Nanopore sequences. Future versions will include support for

312   estimation from Illumina reads. FGMP has a modular architecture and thus can be easily

313   incorporated into existing genome annotation pipelines.

314

315   **CONCLUSION**

316   A realistic estimation of level of genome completeness is a critical metric for accurate

317   comparative genomics studies. This is particularly relevant as the sequencing costs

318   decrease and whole genome assembly is attempted as daily routine for many purposes.

319   BUSCO is currently the only maintained tool for such purpose. FGMP fills a unique

320   niche in the sense that it has modules that assay additional feature types in genomes with

321   no equivalent in existing methods. By applying FGMP to real and simulated datasets, we

322   show that FGMP predictions are reliable and extended that of other software. The tool

323   allows a deeper analysis in the context of evolutionary biology by quickly providing key

324   metrics such the presence of potentially collapsed regions or can be used to screen reads

325   before computationally costly genome assembly is attempted.

326

327   **Availability and requirements**

328   Project name: FGMP

329   Project home page: https://github.com/stajichlab/FGMP

330   Operating system(s): Linux, Mac OS

331   Programming language: Perl 5

332   License: MIT Open Source License

333   Archived release: DOI: 10.5281/zenodo.1453438

334   Package system availability: Bioconda

335

336   **Availability of data and materials'**

337   Additional file 1: List of fungal species used for phylogenomic analysis.

338   Additional file 2: Phylogeny of nine fungal species used for the detection of highly

339   conserved nucleotide elements. The divergence times were obtained from

340   http://www.timetree.org (28)

341   Additional file 3: Comparison of protein markers used for genome completeness

342   estimation.

343   Additional file 4: Assessment of genome completeness in 57 fungal genome assemblies.

344   Additional file 5: Scatterplot showing the relationship between FGMP running times and

345   the level of fragmentation for different genome assemblies expressed as N50.

346   Additional file 6: Genome characteristics, completeness estimates and run times of 90

347   fungal genomes.

348   Additional file 7: Lifestyle, genome characteristics and completeness estimates of 166

349   fungi.

350   **Competing interests**

351   The authors declare that they have no competing interests.

16

352   **METHODS**

353   FGMP is written in Perl 5, and is designed for a command line interface. The code is

354   organized into five distinct modules, which are stored in the main library "FGMP.pm".

355   1.   Identify candidate regions: this module scans the genome assembly using 593

356   fungal proteins with TBLASTn (18). In parallel, the assembly is translated using

357   SIXPACK (19) and compared to 593 Hidden Markov models of the 593 protein

358   markers using HMMER3 (20). Long FASTA headers are discouraged.

359   2.   Process alignments: aligns 593 protein makers to candidate regions using

360   EXONERATE (22). Alignments are converted in protein sequences and training

361   sets for AUGUSTUS (23).

362   3.   Annotation of candidate regions: uses AUGUSTUS to annotate the candidate

363   regions. The module merges AUGUSTUS predictions with translated proteins

364   from module 2 into a single FASTA file. The module compares raw predictions

365   (proteins or peptides) to 593 HMMs using HMMSEARCH. Lastly, FGMP scans

366   the original assembly for 31 universally conserved fungal elements using

367   NHMMER.

368   4.   Check the status of multi-copy protein families: scans the raw predictions and

369   identify markers that are expected to be in multiple copies. Markers with a lower

370   number of copies than expected are tagged as potentially collapsed regions.

371   5.   Generate final report: gather all raw predictions, filter aberrant predictions (at

372   least twice the average length of the reference makers) and choose the longest

373   gene model for each protein markers.

374   **6.** Infer genome completeness from long reads: is triggered when reads are provided.

17

375   FGMP uses a reservoir sampling approach and BLASTx (18) to search for the

376   593 proteins markers in the reads.

377   **Competing interests**

378   The authors declare that they have no competing interests.

379   **Author's contributions**

380   OHC and JES conceived of the project. OHC wrote the code. OHC and JES unit tested

381   and wrote documentation. OHC generated the data utilized for testing. JES directed the

382   project. OHC and JES wrote the manuscript.

383   **Funding**

388   **Acknowledgements**

393

394   **Figure legends**

395   Figure 1 | The FGMP workflow. A typical workflow consists of three sequential modules

396   (indicated by the boxes). The first module (FGMP_PROT) automates the use of different

397   programs to evaluate the genome completeness based on pre-defined protein and

398   nucleotide markers. Additional modules evaluate the patterns of conservation fungal

399   multi-copy protein families. FGMP protein and nucleotide datasets are derived from 25

400   and nine fungal species, respectively (indicated as dotted arrows).

401   Figure 2 | Estimation of genome completeness in fungal genomes. A comparison of

402   genome completeness by multiple software tools using initial and latest genome assembly

403   versions of 45 fungi. Genome completeness expressed as a percentage of expected

404   markers (y-axis) is plotted against assembly size in megabases (x-axis). (A) FGMP

405   completeness estimates based exclusively on protein markers are shown; (B) FGMP

406   estimation based exclusively on highly conserved nucleotide segments; (C,D) BUSCO

407   fungi and CEGMA completeness estimates respectively. In each plot, the dots represent

408   distinct assemblies and color represents their status (red for the initial version and light

409   blue for the latest version), and the diameters are proportional to the N50, a measure of

410   assembly contiguity. Three representative species are highlighted in all panels (ellipses)

411   to show the evolution of genome completeness between assembly versions. (E) Heat map

412   of FGMP, BUSCO and CEGMA completeness estimates.

413   Figure 3 | Genomic loss simulations. The genomes of 57 fungal species were randomly

414   truncated to evaluate accuracy of FGMP, BUSCO and CEGMA. The density plots

415   represent the differences (expressed as percentages) between completeness estimates

416   from the full and truncated assemblies.

19

References

417

418 1. Sohn JI, Nam JW. The present and future of de novo whole-genome assembly.
419   *Briefings in bioinformatics* 2018; 19: 23-40.
420 2. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A,
421   Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert
422   SM, Liu WT, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM,
423   Hugenholtz P, Woyke T. Insights into the phylogeny and coding potential of
424   microbial dark matter. *Nature* 2013; 499: 431-437.
425 3. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The
426   human microbiome project. *Nature* 2007; 449: 804-810.
427 4. Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stuber K, Ver Loren van
428   Themaat E, Brown JK, Butcher SA, Gurr SJ, Lebrun MH, Ridout CJ, Schulze-
429   Lefert P, Talbot NJ, Ahmadinejad N, Ametz C, Barton GR, Benjdia M, Bidzinski
430   P, Bindschedler LV, Both M, Brewer MT, Cadle-Davidson L, Cadle-Davidson
431   MM, Collemare J, Cramer R, Frenkel O, Godfrey D, Harriman J, Hoede C, King
432   BC, Klages S, Kleemann J, Knoll D, Koti PS, Kreplak J, Lopez-Ruiz FJ, Lu X,
433   Maekawa T, Mahanil S, Micali C, Milgroom MG, Montana G, Noir S, O'Connell
434   RJ, Oberhaensli S, Parlange F, Pedersen C, Quesneville H, Reinhardt R, Rott M,
435   Sacristan S, Schmidt SM, Schon M, Skamnioti P, Sommer H, Stephens A,
436   Takahara H, Thordal-Christensen H, Vigouroux M, Wessling R, Wicker T,
437   Panstruga R. Genome expansion and gene loss in powdery mildew fungi reveal
438   tradeoffs in extreme parasitism. *Science* 2010; 330: 1543-1546.
439 5. Kohler A, Kuo A, Nagy LG, Morin E, Barry KW, Buscot F, Canback B, Choi C,
440   Cichocki N, Clum A, Colpaert J, Copeland A, Costa MD, Dore J, Floudas D, Gay
441   G, Girlanda M, Henrissat B, Herrmann S, Hess J, Hogberg N, Johansson T,
442   Khouja HR, LaButti K, Lahrmann U, Levasseur A, Lindquist EA, Lipzen A,
443   Marmeisse R, Martino E, Murat C, Ngan CY, Nehls U, Plett JM, Pringle A, Ohm
444   RA, Perotto S, Peter M, Riley R, Rineau F, Ruytinx J, Salamov A, Shah F, Sun H,
445   Tarkka M, Tritt A, Veneault-Fourrey C, Zuccaro A, Mycorrhizal Genomics
446   Initiative C, Tunlid A, Grigoriev IV, Hibbett DS, Martin F. Convergent losses of
447   decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal
448   mutualists. *Nature genetics* 2015; 47: 410-415.
449 6. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S,
450   Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou WC, Corbeil J, Del Fabbro
451   C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy
452   G, Gibbs RA, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler
453   D, Hiatt JB, Ho IY, Howard J, Hunt M, Jackman SD, Jaffe DB, Jarvis ED, Jiang
454   H, Kazakov S, Kersey PJ, Kitzman JO, Knight JR, Koren S, Lam TW, Lavenier
455   D, Laviolette F, Li Y, Li Z, Liu B, Liu Y, Luo R, Maccallum I, Macmanes MD,
456   Maillet N, Melnikov S, Naquin D, Ning Z, Otto TD, Paten B, Paulo OS, Phillippy
457   AM, Pina-Martins F, Place M, Przybylski D, Qin X, Qu C, Ribeiro FJ, Richards
458   S, Rokhsar DS, Ruby JG, Scalabrin S, Schatz MC, Schwartz DC, Sergushichev A,
459   Sharpe T, Shaw TI, Shendure J, Shi Y, Simpson JT, Song H, Tsarev F, Vezzi F,
460   Vicedomini R, Vieira BM, Wang J, Worley KC, Yin S, Yiu SM, Yuan J, Zhang

461   G, Zhang H, Zhou S, Korf IF. Assemblathon 2: evaluating de novo methods of
462   genome assembly in three vertebrate species. *Gigascience* 2013; 2: 10.
463 7. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in
464   eukaryotic genomes. *Bioinformatics* 2007; 23: 1061-1067.
465 8. Parra G, Bradnam K, Ning Z, Keane T, Korf I. Assessing the gene space in draft
466   genomes. *Nucleic acids research* 2009; 37: 289-297.
467 9. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G,
468   Kriventseva EV, Zdobnov EM. BUSCO applications from quality assessments to
469   gene prediction and phylogenomics. *Mol Biol Evol* 2017.
470 10. Marthey S, Aguileta G, Rodolphe F, Gendrault A, Giraud T, Fournier E, Lopez-
471   Villavicencio M, Gautier A, Lebrun MH, Chiapello H. FUNYBASE: a FUNgal
472   phYlogenomic dataBASE. *BMC Bioinformatics* 2008; 9: 456.
473 11. Stajich JE. Fungal Genomes and Insights into the Evolution of the Kingdom.
474   *Microbiology spectrum* 2017; 5.
475 12. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ,
476   Schatz MC, Delcher AL, Roberts M, Marcais G, Pop M, Yorke JA. GAGE: A
477   critical evaluation of genome assemblies and assembly algorithms. *Genome*
478   *research* 2012; 22: 557-567.
479 13. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. *Nat Rev*
480   *Genet* 2012; 13: 329-342.
481 14. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a
482   universal tool for genome assembly evaluation. *Genome biology* 2013; 14: R47.
483 15. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for
484   genome assemblies. *Bioinformatics* 2013; 29: 1072-1075.
485 16. Gruning B, Dale R, Sjodin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R,
486   Koster J, Bioconda T. Bioconda: sustainable and comprehensive software
487   distribution for the life sciences. *Nat Methods* 2018; 15: 475-476.
488 17. Altenhoff AM, Skunca N, Glover N, Train CM, Sueki A, Pilizota I, Gori K,
489   Tomiczek B, Muller S, Redestig H, Gonnet GH, Dessimoz C. The OMA
490   orthology database in 2015: function predictions, better plant support, synteny
491   view and other improvements. *Nucleic acids research* 2015; 43: D240-249.
492 18. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.
493   Gapped BLAST and PSI-BLAST: a new generation of protein database search
494   programs. *Nucleic Acids Res* 1997; 25: 3389-3402.
495 19. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open
496   Software Suite. *Trends Genet* 2000; 16: 276-277.
497 20. Eddy SR. Accelerated Profile HMM Searches. *Plos Comput Biol* 2011; 7.
498 21. Wallace IM, O'Sullivan O, Higgins DG, Notredame C. M-Coffee: combining
499   multiple sequence alignment methods with T-Coffee. *Nucleic acids research*
500   2006; 34: 1692-1699.
501 22. Slater GS, Birney E. Automated generation of heuristics for biological sequence
502   comparison. *BMC bioinformatics* 2005; 6.
503 23. Stanke M, Schoffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes
504   with a generalized hidden Markov model that uses hints from external sources.
505   *BMC bioinformatics* 2006; 7: 62.

506   24. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic
507   sequence comparison. *Genome Res* 2011; 21: 487-493.
508   25. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
509   features. *Bioinformatics* 2010; 26: 841-842.
510   26. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW,
511   Noble WS. MEME SUITE: tools for motif discovery and searching. *Nucleic acids*
512   *research* 2009; 37: W202-208.
513   27. Stajich JE. Fungal Genomes and Insights into the Evolution of the Kingdom.
514   *Microbiol Spectr* 2017; 5.
515   28. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines,
516   Timetrees, and Divergence Times. *Mol Biol Evol* 2017; 34: 1812-1819.
517
518