Genome analysis

FGMP: assessing fungal genome completeness and gene content

Ousmane H. Cissé[1,□] and Jason E. Stajich[1*]

[1] Department of Plant Pathology & Microbiology and Institute for Integrative Genome Biology, University of California-Riverside, Riverside, CA 92521 USA

*To whom correspondence should be addressed.

Present address: National Institutes of Health, Bethesda, MD 20814 USA

Abstract

Motivation: Inexpensive high-throughput DNA sequencing has democratizing access to genetic information for most organisms so that access to a genome or transcriptome of an organism is not limited to model systems. However, the quality of the sampled genomes can vary greatly which hampers utility for comparisons and meaningful interpretation. The uncertainty of the completeness of a given genome sequence can limit feasibility of asserting patterns of high rates of gene loss reported in many lineages.

Results: We propose a computational framework and sequence resource for assessing completeness of fungal genomes called FGMP (Fungal Genome Mapping Project). Our approach is based on evolutionary conserved sets of proteins and ultra conserved DNA elements and is applicable to various types of genomic data. We present a comparison of FGMP with state-of-the-art methods utilizing 246 genome assemblies of fungi. We discuss genome assembly improvements/degradations in 56 two-point fungal genome assemblies, as recorded by NCBI assembly archive.

Availability and Implementation: FGMP software and datasets are freely available from https://github.com/stajichlab/FGMP or biocluster.ucr.edu/~ocisse/manuscript/FGMP.v.1.0.tar.gz

Contact: ousmanecis@gmail.com or jason.stajich@ucr.edu

Supplementary information: Supplementary data are available at biocluster.ucr.edu/~ocisse/manuscript.

1 Introduction

The recent explosion of high-throughput sequencing methods and analytic tools has made sequencing easier and cheaper for nearly all species across the tree of life including uncultivated organisms. However, the quality and completeness of these genomes is not perfect. Large-scale sequencing projects have emerged such as the microbial dark matter project (Rinke *et al.*, 2013), the Human Microbiome Project (Turnbaugh *et al.*, 2007) or the 1000 fungal genomes project (http://1000.fungalgenomes.org). The rapid generation and release of draft data is without a doubt beneficial and have been extensively used for many purposes. However, the quality and the level of completion of draft genomes can vary greatly and there is a need to quantify a genome's completeness to provide context for how much be inferred from it. Another important motivation of this work is that lineage specific gene loss is an important driving force in evolution, especially in fungi (Spanu *et al.*, 2010; Kohler *et al.*, 2015).

Approaches to assess the quality and completeness of a genome have been proposed using nearly 100 different metrics (Bradnam *et al.*, 2013). Unfortunately, most of these metrics are generally not applicable for non-model species because they require a substantial amount of high quality data (e.g. fosmids, reference genomes, optical maps) that can be expensive or infeasible to obtain for a large number of samples. So far extremely few methods attempt to estimate the amount of missing data in an assembly without prior knowledge. One of the most popular approaches, CEGMA estimates the completeness to the presence of set of 248 single copy gene markers (Parra *et al.*, 2007). Although CEGMA has been used in numerous studies, a key issue is that makers were selected from only six model eukaryotic species and the ubiquity and detections of these markers may not be consistent as more distant lineages are sampled. The concept has been recently revisited and updated with clade-focused sets of protein coding gene markers in BUSCO (Simao *et al.*, 2015). Alternatively, for Fungi, another set of 246 single copy gene families has been proposed by FUNYBASE (Marthey *et al.*, 2008). Typically, multicopy gene families are systematically filtered out in these selections, but their utility, as well as that of alternative, non-protein coding gene markers has not been fully explored. Some of the most simple and often used statistics for assembly measurement is N50 and L50 (Salzberg *et al.*, 2012), which describes the level of fragmentation of the assembly. Both statistics utilize a sorted list of largest to smallest sizes of contigs, where L50 is the length (in bases) of the shortest contig for which 50% of the genome can be contained within contigs of that size or larger, and N50 is the number of contigs that when summed their length is half of the assembly size (Yandell and Ence, 2012;). Note that for some tools the meaning two statistics are swapped, where N50 means length and L50 means the count. Other methods measure the number error per bases or assembly inconsistencies (Hunt *et al.,* 2013; Gurevich *et al.,* 2013).

In the present study, we focused on the fungal kingdom, which has been estimated to originate approximately 1.5 billion years ago (Brundrett, 2002; Berbee and Taylor, 2006; Stajich 2009). The primary motivation of this work is to provide a realistic estimation of assembly completeness for fungal genomes. The precision dependents on the ability to accurately identify genes, which can appear artifactually fragmented by an incomplete assembly or due to rapidly evolving entities in some lineages. The nature, evolutionary trajectory and loss likelihood of genes need to be considered when calculating genome completeness from gene content. We propose a different set of markers and build a pipeline around them called FGMP (Fungal Genes Mapping Project). Our multistep approach extends previous approaches by integrating identifiable fungal protein and ultraconserved genomic regions. Our markers include both single and multigenic markers, and has only 50% overlap with previously published datasets providing a different dimension of sequence evolution to evaluate the completeness. Additionally, we use a multisampling approach coupled to a rarefaction analysis to search for markers in unassembled sequencing reads, which bypass the need for an assembly. Therefore, a researcher can quickly assess the quality of a set of reads in hand before attempting an assembly, which can be computationally expensive. Finally, we described a side-by-side comparison of our tool with state-of arts methods over 246 fungal species. We captured assembly improvements/degradations in 56 different released versions of assemblies of species, as recorded in NCBI assembly archive. This work can be a valuable source for genome completion estimation that can be easily incorporated in more complex pipeline because of its modular structure.

2 Methods

2.1 Species and data sources

We selected 40 fungal species covering the major fungal phylogenetic clades for analysis to select and initial seed set of conserved markers. Our approach differs from previous approaches because (i) at least two species per subphylum are selected which increases the likelihood of capturing lineage specific markers (ii) a credible homolog needed to be present in 99% species tested and (iii) no filtering for single copy genes is required. Proteomes and genomes data are from UniProtKB release 2015_07 (UniProt Consortium, 2014), JGI (Grigoriev *et al.,* 2014) and GeneBank (Benson *et al.,* 2014).

2.2 Orthology, hidden markov models and markers

Orthologous groups were inferred using OMA (Roth *et al.,* 2008). Clusters with less than 5 species were excluded. To verify the consistency of orthology predictions, we systemically surveyed related proteins placed in different orthologous cluster using BLASTP (e-value 10^{-10}) and architecture using HMMER3 and Pfam A (e-value of 0.1) (Eddy, 2011; Punta *et al.,* 2012). For each cluster, the most informative sequence was selected using T-coffee (Wallace *et al.,* 2006). Alignment filtered criteria were as follows: the alignment score should be higher or equal to 80 and 70% of the sequence should be covered by the alignment. For multigenic families, we generated a global alignment and select the protein with fewer gaps and more similar to the consensus sequence from multiple alignment. Hidden Markov Models (HMMs) were generated using HMMER3. We controlled the inclusion of related proteins by selecting only gene models that can unambiguously distinguished from each other. Gene identification is assessed by comparison to HMMs using optimized pre-computed thresholds allowing the identification of highly divergent models.

2.3 Fungal ultra conserved elements

We selected ten representative fungi: Ascomycetes: *Aspergillus nidulans*, *Candida albicans*, *Neurospora crassa*, *Saccharomyces cerevisiae*, *Saitoella complicata* Basidiomycetes: *Coprinopsis cinerea*, *Puccinia gramianis*, *Sporobolomyces. roseus* and *Ustillago maydis*, and Mucoromycota: *Rhizopus delemar*. These species were selected based on empirical tests of multi species whole genome alignments. In practice, attempts with more than 10 species or heavily fragmented genome assemblies were unsuccessful. We then built a phylogenetic tree on basis of 95 conserved markers (Cisse *et al.,* in preparation) using RAxML v8.2 (Stamatakis, 2014). Protein model was inferred automatically using PROTGAMMAAUTO option and support values were obtained from 100 bootstraps. The tree and genome sequences were given as input for alignment to progressive Cactus v.0.0 (Paten *et al.,* 2011). Conserved alignment blocks were converted in psl format using Maf2psl tool (https://github.com/ENCODE-DCC/kentUtils). In each genome, conserved blocks (size >= 200 bps) were extracted, merged, aligned using MUSCLE (Edgar, 2004) and used to prepare HMMs with HMMER3 v3.1b2. Assignments were performed using NHMMER (e-value < 1e-5).

2.4 Identification of markers in unassembled sequences

Subsets of unassembled sequences are randomly selected using reservoir sampling (1,000 chunks of 10,000 reads per sample) and screened for protein makers using BLASTx with an e-value of 10^{-4} (Altschul *et al.,* 1997). For each sample, the number of detected markers is recorded at each iteration and used as input for rarefaction analysis. The process is stopped when no more new markers are detected after 20 unsuccessful trials. The choice of sample size as well as the number of trials is arbitrary and based on empirical tests. These parameters might not be realistic but can be set manually depending on user needs. Future developments will focus on implementing a better statistical approach to select the sample size accounting for the population size, the confidence interval, margins errors and the expected coverage.

To test this approach, we generated simulated PacBio-like sequences from the genomes of *Neurospora crassa* OR74A (NCBI assembly accession: GCA_000182925.2), *Botryotinia fuckeliana* strain B05.10 (GCA_000143535.2) *and Acremonium alcalophilum* strain JCM 7366 (PRJNA33785) with 30x coverage depth using PBSIM (Ona and Hamada, 2013). Real Pacbio sequences were obtained from NCBI sequence archive for *Pyronema confluens* (SRR3110858) and *Rozella allomycis* (SRR834607).

2.5 Evaluation of FGMP

We compared the genome completeness assessments of FGMP to three other tools: CEGMA, BUSCO and FUNYBASE. CEGMA uses 248 single copy genes from six eukaryotes but for comparative purpose, we used only *Schizosaccharomyces pombe* proteins. FUNYBASE data correspond to 236 families including 5,166 proteins. One representative sequence per family were retained based on T-Coffee multiple alignment of each gene family. Only one representative sequence per family in CEGMA and FUNYBASE datasets was used. BUSCO provides 1,438 fungi specific markers that were downloaded from http://busco.ezlab.org/ (last accessed 9-10-2015). Consensus sequences were extracted from HMMs using hmmemit.

Protein sequences were annotated using Blast2GO (Conesa *et al.*, 2005), InterProScan ver. 5.16-55.0 (Jones *et al.*, 2014) and Priam release of 04-Mar-2015 (Claudel-Renard *et al.*, 2003).

3 Results

3.1 Workflow

FGMP is primarily designed for assessment of fungal genomes. It works on assembled sequences or raw shotgun reads and generates completeness statistics and gene models based on the presence of a pre-selected set of markers (Figure 1). To generate the consensus protein markers we analyzed a phylogenomic dataset of 164,232 proteins from 42 taxa including 40 fungi and two outgroup eukaryotes (*Homo sapiens* and *Arabidopsis thaliana*). The major fungal lineages were sampled, from the early diverging fungi (EDF), Ascomycota and Basidomycota. The clustering yields 6,845 gene families across these clades. Lineage specific gene families (i.e. present in less than 5 species) are excluded. Makers were selected according to the following rules: (i) a marker should be present in at least 99% of the species and (ii) should be unambiguously identifiable (Supplementary material). In contrast with other published strategies, multicopy gene families are not excluded. The final set of markers included 593 proteins of which 60.3% are single copy genes. We also generated 172 ultraconserved fungal genomic segments derived from whole genome alignment of ten fungal species covering the major phylogenetic clades.

3.2 Comparative analysis of markers

A total of 2,004 FGMP markers were originally obtained, which was reduced to 593 after the removal of ambiguous markers. We compared the markers selected for FGMP (593 proteins) to those used in CEGMA (248 families, 1,488 proteins), BUSCO fungi (1,438 proteins) and FunyBase (236 families, 5,166 proteins. Using reciprocal best BLASTp (e-value < 10^{-5}), 49.5% of FGMP protein markers are not found in other datasets whilst this proportion is 21.7% for CEGMA, 10.5% for FunyBase and 69.8% for BUSCO (Figure S1). Transferases and transporters are common (13%). Kinases and helicases are overrepresented in FGMP-protein dataset where they represent 10% and 5% of 593 protein markers, respectively as compared to 0.8% and 2% of CEGMA makers; 3.3% and 2% in FunyBase markers; 3.3% and 0.7% of BUSCO fungi markers. Kinases and helicases are multicopies gene families in nearly all fungi, which might explain why these genes are not present in other datasets which actively restrict gene duplicates. Most of FGMP kinases have homologs in bacteria and archeae, suggesting that they are ancient. Most of the helicases also have archeal or bacterial homologs as well and are likely a mix of ancient and derived forms.

3.3 Evaluation of 1FKG data

We estimated the genome completeness of 246 fungal genomes from 1FKG. Each species was classified according to its lifestyle based on published literature (e.g. saprotroph, parasite). Parasites are characterized by a reduced genome size and rely partially or entirely on their hosts for survival. The average N50 is 126.7 Mb for an average number of scaffolds per genome of 1,029; an average genome size of 38 Mb and the average fraction of Ns per genome is 3.2%. The level of fragmentation of these genomes is high as compared to well-sequenced clades such as vertebrates.

The predictions from FGMP, CEGMA and BUSCO-fungi for 246 species suggest that most of the genomes have a completeness value above 80% (Figure S2; Table 1). A total of 92% of these genomes have CEGMA value > 95% whilst only 58.7% of these assemblies reach this cut-off with BUSCO-fungi, 40% with FGMP-fUCEs and 54.2% with FGMP-protein markers. Genomes labeled as incomplete are typically parasitic species with reduced metabolic capabilities. These results suggest that parasites represent a twilight zone where gene losses and assembly holes are confounded. Overall completeness predictions correlated with the N50: CEGMA (spearman rho = 0.35, *P*-value = 1.3 x 10^{-8}), BUSCO fungi (R = 0.40; *P* = 2.1 x 10^{-11}), FGMP-fUCEs (R = 0.17; *P* = 0.005) but FGMP-protein (R = -0.05; *P* = 0.4). The fact that FGMP-protein predictions are not correlated with N50 is due to the integration of translated DNA sequences as well as gene fragments, which allow the partial detection of markers even when reliable gene models cannot be built. Nevertheless, these sequences have to score above a precompiled threshold to be accepted as valid hit. False positive hits might inflate the estimation, but the likelihood of integrating such elements is expected to be negligible. These results suggest that the assembly continuity is not the most critical factor for the genome completeness estimation.

We found no significant correlations between CEGMA completeness estimates and the genome size (R = -0.03; *P* = 0.5) or GC content (R = -0.004; *P* = 0.9). The same is true for BUSCO fungi where the estimations are not correlated with the genome size (R = 0.07, *P*-value = 0.2) but the GC content shows a weak correlation (R = 0.19; *P* = 0.001). A different pattern is seen for FGMP-fUCEs where the completeness values are correlated with the genome size (R = 0.30, *P* = 1.0 x 10^{-6}) and the GC content (R = 0.20; *P* = 0.001). FGMP-protein predictions are not significantly correlated with genome size (R = 0.10, *P*-value = 0.1) but the correlation with the GC content is significant (R = 0.15; *P* = 0.014). The GC content is an important factor for gene prediction but not necessary for the estimation of completeness.

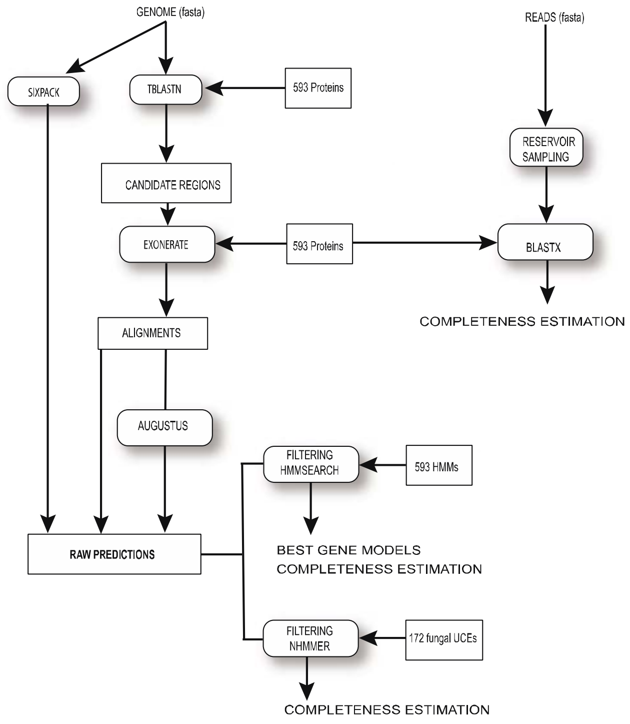

Fig. 1. Overview of FGMP workflow.

3.4 Capturing assemblies improvement/degradations

We screened the initial and latest genome assemblies for 56 fungal species (Supplementary material: dataset_1). For each species, we retrieved the initial (V1) and the latest version (V2) of the assembly and computed the completeness values using FGMP, CEGMA and BUSCO fungi. The ratio between the completeness values in the V2 versus V1 is called assembly index (AI), which is expressed as percent of the total number of markers detected. A negative AI implies a degradation of the assembly (Figure 2). CEGMA predicts negative AIs in 12 species (median AI = -1.0%), null AIs in 12 species, positive AIs in 33 species (0.81%). AIs are positively correlated with the assembly size (R = 0.52, *P* = 2.3 x 10^{-5}) and the coverage depth (R = 0.3; *P* = 0.02).

FGMP protein predicts negative AIs in 36 species (median AI = -1.9%), null AIs in 9 species and positive AIs in 12 species (+0.5%). AIs are correlated with the assembly size (R = 0.34, *P* = 0.007) but not with the coverage depth.

FGMP UCEs predicts negative AIs in 12 species (median AI = -1.4%), null AIs in 31 species and positive AIs in 14 species (+1.15%). AIs are correlated with the assembly size (R = 0.39, *P* = 0.002) and the coverage depth (R= 0.2; *P* = 0.06).

BUSCO-fungi predicts negative AIs in 11 species (median AI = -1.4%), null AIs in 7 species and positive AIs in 39 species (+0.7%). AIs are correlated with the assembly size (R = 0.47; *P* = 0.0002) and the coverage depth (R = 0.33; *P* = 0.01).

Among the conflicting results is *Sordaria macrospora* genome assembly: FGMP protein and BUSCO fungi predict a negative AI of 0.3% whilst CEGMA and FGMP-UCEs predict no improvement. The version 2 of the assembly has lost 1 Mb as compared to the first version but N50 has increased, which suggest a collapse of repeated gene families in the latest version of the assembly. In *Kluyveromyces marxianus*, FGMP protein predicts a positive AI of 0.3% whilst CEGMA and FGMP-UCEs predict no improvement and BUSCO fungi even predicts a negative AI (-0.3%). The N50 value has dropped but the size of the assembly remains unchanged. Overall, the gain of sequence is the most robust predictor for the genome improvement. However the lack of standardized protocols for assembly improvements make the backtracing of inconsistencies difficult. Indeed, most of the improved assemblies are mix of different assembly algorithms and/or sequencing technologies.

3.5 Estimation of genome completeness from unassembled sequences

Using a 30-fold simulated coverage depth, we found that *N. crassa* and *A. alcalophilum* have completeness of 100% and 97.3% respectively using FGMP-reads. Using real Pacbio sequences, we obtained lower values for *P. confluens* and *R. allomycis* with 63.4% and 81.2%, respectively. These results suggest that this method is convenient to estimate the genome completeness. However, the method is sensitive to the coverage depth.

Table 1. Genome completeness statistics of selected fungi.

Species	Size (Mb)	N50 (Kb)	FGMPp	FGMPu	CEGMA	BUSCO
S. macr	37	524	99.7	99.6	99.1	99.8
A. niger	35	1937	97.5	98.8	99.3	99.8
C. neo	19	1438	90.9	90.7	97.5	93.2
Y. lipo	21	3633	94.6	93.6	99.6	99.3
N. ire	15	16	94.4	91.9	97.1	42.0
R. ire	91	4	94.4	91.9	96.1	89.0
R. all	11	61	92.7	87.2	87.9	19.0
E. int	2	204	46.5	29.1	45.3	28.0

FGMPp are estimations based on 593 protein markers, FGMPu are based on 172 fungal conserved DNA segments, CEGMA on 293 protein markers and BUSCO on 1438 fungal protein markers. Species names are as follows: *Sordaria macrospora* strain k-hell, *Aspergillus niger* strain ATCC_1015, *Cryptococcus neoformans* var. *neoformans* JEC21, *Yarrowia lipolytica* strain CLIB122, *Rhizophagus irregularis* strain DAOM_181602, *Neolecta irregularis* strain DAH-1.v1, *Rozella allomycis* strain CSF55 and *Encephalitozoon intestinalis* strain ATCC_50506.

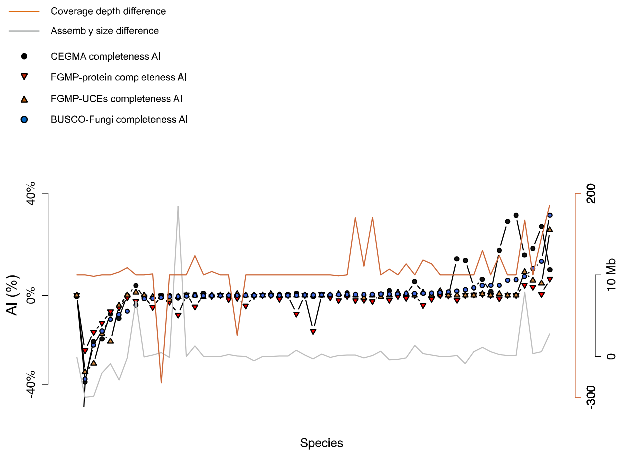

Fig. 2. Completeness statistics of 56 two-point assemblies. Assembly indexes (AI) represent the difference in term of completeness percentage between the last version of an assembly and the initial version for each species. Orange and grey lines show variations in assembly size and coverage depth for each AI, respectively.

Acknowledgements

Computations were performed on the University of California-Riverside Institute for Integrative Genome Biology high performance bioinformatics cluster (http://www.bioinformatics.ucr.edu/) supported by NSF MRI DBI 1429826 and NIH S10-OD016290.

Funding

This work was supported by the Swiss National Science Foundation fellowship grant no. 151780 to O.H.C. NIH grant GM108492 and NSF grant DEB1441715 to JES, partially supported these efforts. This work was supported by the USDA National Institute of Food and Agriculture Hatch project CA-R-PPA-5062-H.

Conflict of Interest: none declared.

References

Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-402.

Benson, D.A. *et al.* (2014) GenBank. *Nucleic Acids Res* **42**: D32-7.

Bradnam, K.R. *et al.* (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**: 10.

Brundrett MC. (2002) Coevolution of roots and mycorrhizas of land plants. *New Phytologist* 154: 275–304.

Claudel-Renard, C. *et al.* (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res* **31**: 6633-9.

Conesa, A., *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**: 3674-6.

Eddy, S.R. (2011) Accelerated Profile HMM Searches. *Plos Computational Biology* **7**(10).

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-7.

Grigoriev, I.V., *et al.* (2014) MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res* **42**: D699-704.

Kohler, A., *et al.* (2015) Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. *Nat Genet* **47**: 410-5.

Jones, P., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*

Hunt, M., *et al.* (2013) REAPR: a universal tool for genome assembly evaluation. Genome Biol **14**: R47.

Gurevich, A., *et al*. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072-5.

Marthey, S., *et al.* (2008) FUNYBASE: a FUNgal phYlogenomic dataBASE. *BMC Bioinformatics* **9**: 456.

Ono, Y., K. Asai, and M. Hamada (2013) PBSIM: PacBio reads simulator--toward accurate genome assembly. *Bioinformatics* **29**: 119-21.

Paten, B., *et al.* (2011) Cactus: Algorithms for genome multiple sequence alignment. *Genome Res* **21**: 1512-28.

Punta, M., *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res* **40**: D290-301.

Rinke, C., *et al.* (2003) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431-7.

Roth, A.C., G.H. Gonnet, and C. Dessimoz (2008) Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* **9**: 518.

Simao, F.A., *et al.* (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*

Salzberg, S.L., *et al.* (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res* **22**: 557-67.

Spanu, P.D., *et al.* (2010) Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* **330**:1543-6.

Stajich, J.E., *et al.* (2009) The fungi. *Curr Biol* **19**: R840-5.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312-3.

Taylor, J.W. and M.L. Berbee. (2006) Dating divergences in the Fungal Tree of Life: review and new analyses. *Mycologia* **98**: 838-49.

Turnbaugh, P.J., *et al.* (2007) The human microbiome project. *Nature* **449**: 804-10.

UniProt Consorsium. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **42**: D191-8.

Yandell, M. and D. Ence. (2012) A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13**: 329-42.

Wallace, I.M. *et al.* (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* **34**: 1692-9.