Genes & Genomics (2019) 41:337–342
https://doi.org/10.1007/s13258-018-0766-1

Online ISSN 2092-9293
Print ISSN 1976-9571

RESEARCH ARTICLE

TaF: a web platform for taxonomic profile-based fungal gene prediction

Sin-Gi Park[1] · DongSung Ryu[1] · Hyunsung Lee[1] · Hojin Ryu[2] · Yong Ju Ahn[1] · Seung il Yoo[1] · Junsu Ko[1] · Chang Pyo Hong[1]

Received: 18 April 2018 / Accepted: 13 November 2018 / Published online: 19 November 2018

Abstract

Introduction The accurate prediction and annotation of gene structures from the genome sequence of an organism enable genome-wide functional analyses to obtain insight into the biological properties of an organism.

Objectives We recently developed a highly accurate filamentous fungal gene prediction pipeline and web platform called TaF. TaF is a homology-based gene predictor employing large-scale taxonomic profiling to search for close relatives in genome queries.

Methods TaF pipeline consists of four processing steps; (1) taxonomic profiling to search for close relatives to query, (2) generation of hints for determining exon–intron boundaries from orthologous protein sequence data of the profiled species, (3) gene prediction by combination of *ab inito* and evidence-based prediction methods, and (4) homology search for gene models.

Results TaF generates extrinsic evidence that suggests possible exon–intron boundaries based on orthologous protein sequence data, thus reducing false-positive predictions of gene structure based on distantly related orthologs data. In particular, the gene prediction method using taxonomic profiling shows very high accuracy, including high sensitivity and specificity for gene models, suggesting a new approach for homology-based gene prediction from newly sequenced or uncharacterized fungal genomes, with the potential to improve the quality of gene prediction.

Conclusion TaF will be a useful tool for fungal genome-wide analyses, including the identification of targeted genes associated with a trait, transcriptome profiling, comparative genomics, and evolutionary analysis.

Keywords *Ab initio* · Exon–intron boundary · Filamentous fungal genome · Homology-based gene prediction · Taxonomic profile · Web platform

Introduction

A primary goal of determining the genome sequence of an organism is to obtain information about its genes. The accurate prediction and annotation of gene structures enable genome-wide analyses, including the identification of target genes associated with a trait, transcriptome profiling, comparative genomics, and evolutionary analysis, thus providing insights into the biological properties of an organism. As whole-genome sequence data have been rapidly accumulating with the advance of sequencing technologies (Hayden 2014), the development of methods for complete and accurate gene annotation is of paramount importance (Yandell and Ence 2012). However, gene prediction in newly sequenced genomes has become a bottleneck because of assembly errors resulting from short read lengths and difficulties in the training, optimization and configuration of

Sin-Gi Park and DongSung Ryu have contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13258-018-0766-1) contains supplementary material, which is available to authorized users.

✉ Junsu Ko
junsuko@gmail.com

✉ Chang Pyo Hong
changpyo.hong@theragenetex.com

[1] TheragenEtex Bio Institute, Suwon 16229, Republic of Korea

[2] Department of Biology, Chungbuk National University, Cheongju 28644, Republic of Korea

gene prediction, due to a lack of pre-existing gene models and small-scale efforts in bioinformatics (Yandell and Ence 2012). For example, incorrectly missing gene annotations can lead to false interpretations, such as incorrect predictions of gene loss, and errors in gene expression profiles that are employed to map and quantify RNA-Seq reads using predicted gene models (Dunne and Kelly 2017).

The development of gene prediction software has been advancing, with a number of effective algorithms being generated over the last fifteen years (Yandell and Ence 2012). In general, genes can be predicted *ab initio* by detecting probabilistic species-specific signals in DNA sequences using hidden Markov models (HMMs) (Stanke and Waack 2003), conditional random fields (DeCaprio et al. 2007), and support vector machines (Schweikert et al. 2009). The signals include codon frequencies, the distribution of intron and exon lengths, intron and exon GC contents, and motifs associated with the beginning and end of exons (e.g., splice sites and start/stop codons). Genes can also be predicted by accepting transcriptome-based and protein-based evidence identified by aligning expressed sequence tags (ESTs), RNA-Seq data, and protein sequences to a genome (Yandell and Ence 2012). In particular, RNA-Seq can yield detailed information on the structure of genome-wide mature transcripts in an organism (Marioni et al. 2008). These evidence-driven methods can be used to train gene predictors, even in the absence of pre-existing reference gene models, thus improving the accuracy of gene prediction. Therefore, the integration of all the available *ab initio* and evidence-driven gene predictions provides great potential to improve the quality of gene prediction in newly sequenced genomes.

While many of the newly sequenced fungal genomes have been recently assembled through next-generation sequencing, using single-molecule real-time (SMRT) sequencing in particular, which produces long-reads (approximately 10 kb read length) and improves assembly errors (Choo et al. 2016; Shim et al. 2016), there is a need for the development of gene prediction software for fungal genomes. Fungal genomes exhibit a number of differences from large complex plant and animal genomes, such as a compact genome structure, shorter intergenic spaces and introns, (Galagan et al. 2005), and diverse genetic codons (Nakagawa et al. 2008; Riley et al. 2016). Most importantly, the availability of well-characterized genome annotations for only a few fungal species limits comparative gene predictions for species in other clades. In addition, many of the predicted fungal proteins in sequence databases lack experimental verification across a variety of species. Several fungal gene prediction and annotation tools, such as SnowyOwl (Reid et al. 2014), ABFGP (van der Burgt et al. 2014), and OrthoFiller (Dunne and Kelly 2017), have been developed recently. SnowyOwl is a gene prediction pipeline that uses RNA-Seq data to train and provide evidence (or hints) for the generation of HMM-based gene predictions (Reid et al. 2014). ABFGP is a sequence alignment-based gene prediction tool that assesses gene models on a gene-by-gene basis and is suitable for the plastic genomes (van der Burgt et al. 2014). OrthoFiller is intended to identify missing annotations for evolutionarily conserved genes (orthogroups).

We recently developed the homology-based gene prediction pipeline TaF, coupled with an *ab initio* method, for filamentous fungal genomics. Homology-based gene prediction can be very useful for predicting efficient gene models from a fungal draft genome for which the annotated gene set or transcriptome data (i.e., RNA-Seq) are not supported. However, the use of data from distantly related orthologs in such a method may cause false-positive predictions of exon–intron boundaries within a gene, or even fusion between/among neighboring gene models (i.e., paralogous genes). To improve such false-positive predictions, we employed taxonomic profiling in TaF, which searches for close relatives showing high homology with a query fungal genome, and generated protein-based hints for determining exon–intron boundaries based on the data of orthologous protein sequences of the profiled species. Therefore, TaF suggest a new approach for homology-based gene prediction based on newly sequenced or uncharacterized fungal genomes.

Methods

Development of the TaF pipeline

The TaF pipeline uses a genome sequence file as input. To profile close relatives showing high homology with an input genome, the input is searched against the fungal genomic sequence database of NCBI (https://www.ncbi.nlm.nih.gov/) using BLASTN (Altschul et al. 1990), and taxonomic classification is performed using KronaTools (Ondov et al. 2011). Protein sequences of the profiled species are extracted from the NCBI non-redundant (NR) protein sequence database. The collected protein sequences are aligned to the input using Exonerate (Slater and Birney 2005), and exon and intron models are then generated in GFF format. Repeat sequences, which may tend to confuse *ab initio* predictions, are masked in the input genomic sequence using RepeatMasker and RepeatModeler (http://www.repeatmasker.org/). Gene structures are predicted using Augustus (Stanke et al. 2004), and protein-based evidence hints are generated. Finally, the resulting predicted genes are searched against the UniProt database (http://www.uniprot.org/) using BLASTP (Altschul et al. 1990) for functional annotation.

Sensitivity and specificity of TaF

To evaluate the accuracy of TaF, the sensitivity (Sn) and specificity (Sp) of TaF were calculated based on a filamentous fungal genome of *Aspergillus oryzae* (RIB40 ASM18445v3 in ENSEMBLE release 36), which contains gene models with exon–intron structures, and were compared with those of different programs using Augustus, employing a trained dataset for *A. oryzae* and GeneMark-ES (Borodovsky and Lomsadze 2011) with a fungus self-training option; however, evidence hints that were generated based on transcriptome or protein sequence data were not used. Sn and Sp included the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) according to the sequence annotation for the annotated region (Supplementary Fig. 1).

Sn and Sp of TaF were also compared with those obtained through evidence-based prediction methods using transcriptome and/or protein sequence data. In addition, the efficiency of taxonomic profiling employing orthologous protein sequence data of close relatives was also assessed. For this purpose, the draft genome of *Lentinula edodes* (Shim et al. 2016) was employed, with the corresponding transcriptome data, including Illumina RNA-Seq and PacBio Iso-Seq data. To generate transcriptome-based evidence hints, TopHat2 (Kim et al. 2013) and GMAP (Wu and Watanabe 2005) were used for short- and long-read RNA sequence alignments, respectively. To generate protein-based evidence hints, fungal orthologous protein sequences from target species that were searched using KronaTools were collected from the NCBI NR protein database, and Exonerate was employed for protein sequence alignment. Using the generated evidence hints, Augustus was employed to perform de novo prediction. The Sn and Sp of different evidence-based prediction methods were calculated as described above. Additionally, the classification of splicing junctions was performed using RSeQC (v2.6.3) (Wang et al. 2012).

Implementation

TaF is operated on a Linux server with an Intel (R) Xeon (R) CPU E7-8850 and 256 Gb of RAM. The TaF pipeline is implemented using Python and bash shell scripts, and the web interface is implemented based on APM (Apache, PHP, and MySQL). The size of the input sequence is limited to ~30 Mbp, with a *.gz compressed format, and the analytical process is checked with an assigned job ID. TaF is freely available at http://taf.genome-report.com/.

Results and discussion

Overview of TaF

The workflow of TaF consists of four steps: (1) taxonomic profiling to search for close relatives to query; (2) generation of hints for determining exon–intron boundaries from orthologous protein sequence data of the profiled species; (3) gene prediction using the resulting hints and *ab initio* information; and (4) homology searches for predicted gene models (Fig. 1). TaF searches for close relatives of the query genome through large-scale taxonomic profiling using KronaTools. Taxonomic profiling reveals the abundance of sequence regions conserved between the query and relatives (Fig. 1) and selects the top 6 relatives. If related species in which the abundance is greater than 5% for the queried sequence length at the species level are selected, we empirically recommend the application of TaF. TaF generates hints for determining exon–intron boundaries in gene predictions, by aligning orthologous protein sequences derived from the selected species to the queried genomic sequence (Fig. 1). Thus, TaF can improve the false-positive prediction of gene structure and suggest new approaches for homology-based gene prediction, based on newly sequenced or uncharacterized fungal genomes with taxonomic profiling. Gene prediction is performed using the resulting protein sequence-based hints and *ab initio* information, and known homologous genes are searched against the UniProt database, providing a resource for orthologous gene cluster analysis. The outputs of TaF include the distribution of the quantitative abundance of sequence regions conserved between queries and relatives through taxonomic profiling; predicted gene models in GFF and FASTA formats; and BLASTP search results.

Processing time of TaF depends on the length of genomic sequence and sequence homology. For example, TaF took 2.63 h, 2.63 h, 3.58 h, and 9.93 h to predict gene models in the genomes of *Saccharomyces cerevisiae* (12.2 Mb), *Schizosaccharomyces pombe* (12.6 Mb), *Aspergillus fumigatus* (28.8 Mb), and *Aspergillus oryzae* (37.1 Mb), respectively.

Assessment of the accuracy of gene prediction

The Sn and Sp of TaF were first compared with the Sn and Sp obtained through *ab initio* prediction using Augustus and GeneMark-ES, with no evidence hints. For this purpose, the genome of *Aspergillus oryzae*, which has been well annotated, with 12,074 protein-coding genes, was employed as a reference. Augustus, GeneMark-ES, and

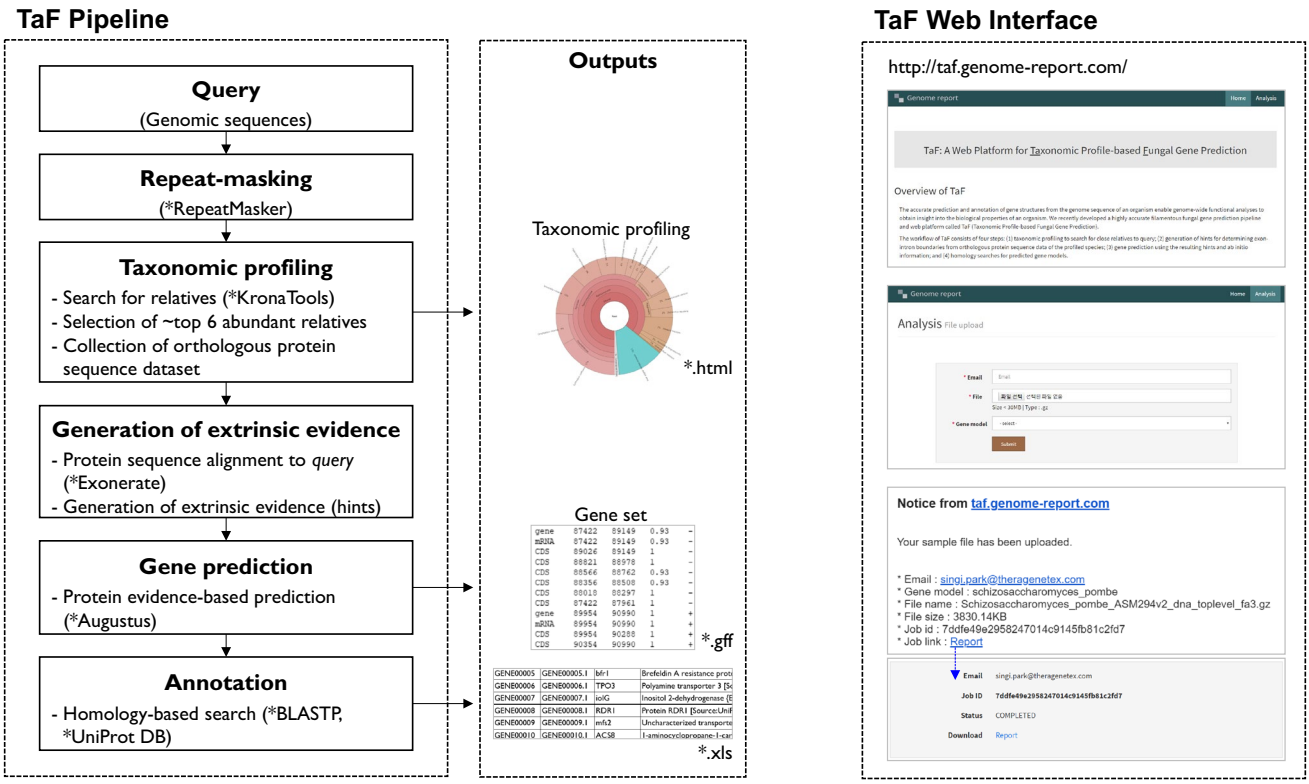

Fig. 1 Workflow of the TaF pipeline and web interface. The TaF pipeline (on the left) is processed in the following order: uploading of assembled genomic sequences, repeat masking, taxonomic profiling, generation of extrinsic evidence, gene prediction, and annotation. TaF outputs (in the center) include the results of taxonomic profiling (*.html), predicted gene models (GFF and FASTA formats), and BLASTP searches. The web interface of TaF (on the right) consists of introduction, analysis, and job status pages

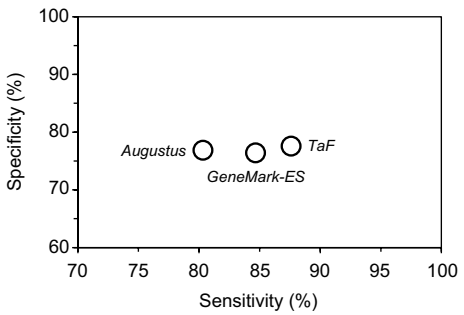

Fig. 2 Comparison of the sensitivity and specificity of gene predictions by Augustus, GeneMark-ES, and TaF. For comparison of the accuracy of the three tools, the reference genome of *Aspergillus oryzae* including gene introns in was employed, and Augustus and GeneMark-ES without evidence hints were applied for *ab initio* prediction

TaF predicted 11,442, 12,791, and 12,722 gene models, respectively. The Sn and Sp for these predicted gene sets were calculated based on the reference gene set. Sn and Sp were found to be 80.33% and 76.86%, respectively, for Augustus; 84.67% and 76.41% for GeneMark-ES; and 87.60% and 77.58% for TaF (Fig. 2), suggesting that prediction based on the combination of taxonomic profiling and *ab initio* methods shows an improved accuracy over *ab initio* prediction alone.

To evaluate the performance of taxonomic profiling-based prediction in TaF, the results of gene prediction using the combination of different methods, including the *ab initio* (*Abi*) strategy, transcriptome-based prediction (*THint*) employing RNA-Seq data as extrinsic evidence, homologous protein-based prediction with taxonomic profiling (*TP-PHint*), and homologous protein-based prediction without taxonomic profiling (*PHint*), were compared: (1) Hint (H) 1: *Abi* + *THint*, (2) H2: *Abi* + *THint* + *TP-PHint*, (3) H3: *Abi* + *THint* + *PHint*, (4) H4: *Abi* + *TP-PHint*, (5) H5: *Abi* + *PHint*, and (6) H6: *Abi*. For these assessments, we used the draft genome of *Lentinula edodes*, with transcriptome datasets including RNA-Seq and Iso-Seq data and 1002 representative gene models that were verified based on full-length cDNAs from Iso-Seq data and contained exon–intron structures. Although prediction methods that make use of transcriptome data showed a high gene identification ability and accuracy (sensitivity and specificity), with relatively small differences (H1, H2, and H3 in Fig. 3a, b), the H2 method exhibited the best performance (Fig. 3a, b).

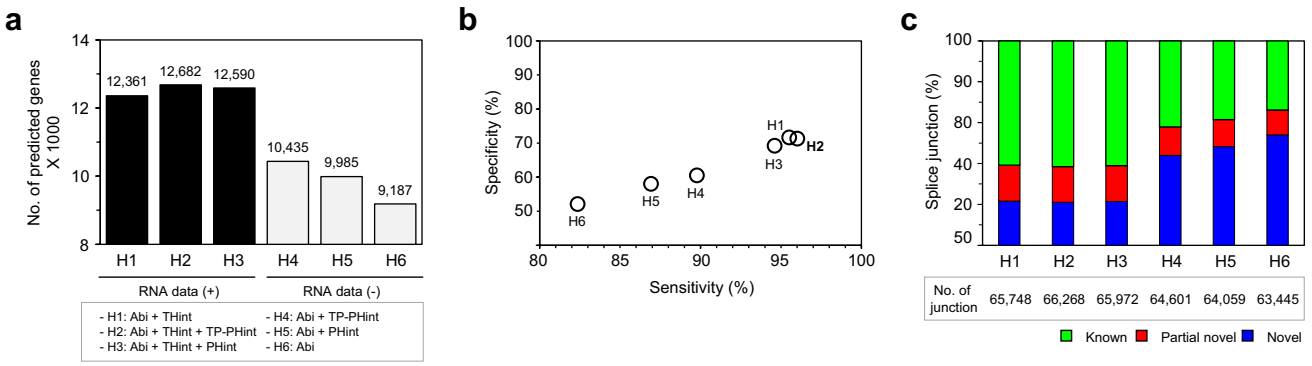

Fig. 3 Assessment of the application of taxonomic profiling-based prediction in gene prediction. **a** Gene prediction via the combination of different methods, including the *ab initio* (*Abi*) strategy, transcriptome-based prediction (*THint*) using RNA-Seq data as extrinsic evidence, homologous protein-based prediction with taxonomic profiling (*TP-PHint*), and homologous protein-based prediction without taxonomic profiling (*PHint*). **b** Comparison of the sensitivity and specificity of different gene prediction results. **c** Abundance of known splice junctions for the accurate detection of exon–intron boundaries. For the assessment, the draft genome of *Lentinula edodes* with transcriptome datasets was used

Moreover, the assessment showed that gene prediction (H2) employing taxonomic profiling presented better accuracy than did (H3), which employs a homology-based prediction method that considers all the homologous fungal protein sequence data to generate extrinsic hints: H2 > H3 in terms of the number of predicted gene models, Sn, and Sp (Fig. 3a, b). In the comparison between two homology-based predictions, H4 and H5, which provide no transcriptome data, the accuracy of the H4 method was also found to be superior to the H5 method. Additionally, known splice junctions were identified as more abundant by H4 than by H5 (although the difference was small, 3.63%) (Fig. 3c), likely because of the more accurate detection of exon–intron boundaries. Our results suggest that the taxonomic profiling employed in TaF and the use of the resulting orthologous protein sequence dataset will improve the accuracy of gene prediction.

Conclusion

The remarkable features of TaF are as follows: it searches for close relatives of query genomes based on large-scale taxonomic profiling, and it can generate extrinsic evidence from the profiled species-derived orthologous protein sequence dataset. This approach can reduce the rate of false-positive predictions. Thus, TaF provides a new approach for homology-based gene prediction based on newly sequenced or uncharacterized fungal genomes. Furthermore, we will upgrade TaF integrated with transcriptome-based prediction.

Acknowledgements This work was supported by the Strategic Initiative for Microbiomes in Agriculture and Food (Grant no. 914008-04) and by the Golden Seed Project (Grant no. 213007-05-1-SBH20) of the Ministry of Agriculture, Food and Rural Affairs of the Republic of Korea.

Authors' contributions SGP, DR, and HL developed the TaF server, evaluated the accuracy of gene prediction, and drafted the paper. HR and YJA performed the RNA isoform sequencing and gene prediction for *L. edodes*. JK and CPH conceived the study, participated in its design and coordination, and drafted the manuscript. All the authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest Sin-Gi Park, DongSung Ryu, Hyunsung Lee, Hojin Ryu, Yong Ju Ahn, Seung il Yoo, Junsu Ko, and Chang Pyo Hong declare that they do not have conflict of interest.

References

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

Borodovsky M, Lomsadze A (2011) Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. Curr Protoc Bioinform Chap 4: Unit 4(6):1–10

Choo JH, Hong CP, Lim JY, Seo JA, Kim YS, Lee DW, Park SG, Lee GW, Carroll E, Lee YW, Kang HA (2016) Whole-genome de novo sequencing, combined with RNA-Seq analysis, reveals unique genome and physiological features of the amylolytic yeast Saccharomycopsis fibuligera and its interspecies hybrid. Biotechnol Biofuels 9:246

DeCaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M, Galagan JE (2007) Conrad: gene prediction using conditional random fields. Genome Res 17:1389–1398

Dunne MP, Kelly S (2017) OrthoFiller: utilising data from multiple species to improve the completeness of genome annotations. BMC Genom 18:390

Galagan JE, Henn MR, Ma LJ, Cuomo CA, Birren B (2005) Genomics of the fungal kingdom: insights into eukaryotic biology. Genome Res 15:1620–1631

Hayden EC (2014) Technology: the $1,000 genome. Nature 507:294–295

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the

presence of insertions, deletions and gene fusions. Genome Biol 14:R36

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–1517

Nakagawa S, Niimura Y, Gojobori T, Tanaka H, Miura K (2008) Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. Nucleic Acids Res 36:861–871

Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a Web browser. BMC Bioinform 12:385

Reid I, O'Toole N, Zabaneh O, Nourzadeh R, Dahdouli M, Abdellateef M, Gordon PM, Soh J, Butler G, Sensen CW, Tsang A (2014) SnowyOwl: accurate prediction of fungal genes by using RNA-Seq and homology information to select among ab initio models. BMC Bioinform 15:229

Riley R, Haridas S, Wolfe KH, Lopes MR, Hittinger CT, Goker M, Salamov AA, Wisecaver JH, Long TM, Calvey CH, Aerts AL, Barry KW, Choi C, Clum A, Coughlan AY, Deshpande S, Douglass AP, Hanson SJ, Klenk HP, LaButti KM, Lapidus A, Lindquist EA, Lipzen AM, Meier-Kolthoff JP, Ohm RA, Otillar RP, Pangilinan JL, Peng Y, Rokas A, Rosa CA, Scheuner C, Sibirny AA, Slot JC, Stielow JB, Sun H, Kurtzman CP, Blackwell M, Grigoriev IV, Jeffries TW (2016) Comparative genomics of biotechnologically important yeasts. Proc Natl Acad Sci U S A 113:9882–9887

Schweikert G, Zien A, Zeller G, Behr J, Dieterich C, Ong CS, Philips P, De Bona F, Hartmann L, Bohlen A, Kruger N, Sonnenburg S, Ratsch G (2009) mGene: accurate SVM-based gene finding with an application to nematode genomes. Genome Res 19:2133–2143

Shim D, Park SG, Kim K, Bae W, Lee GW, Ha BS, Ro HS, Kim M, Ryoo R, Rhee SK, Nou IS, Koo CD, Hong CP, Ryu H (2016) Whole genome de novo sequencing and genome annotation of the world popular cultivated edible mushroom, Lentinula edodes. J Biotechnol 223:24–25

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinform 6:31

Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19 Suppl 2:ii215_ii225

Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 32:W309–W312

van der Burgt A, Severing E, Collemare J, de Wit PJ (2014) Automated alignment-based curation of gene models in filamentous fungi. BMC Bioinform 15:19

Wang L, Wang S, Li W (2012) RSeQC: quality control of RNA-seq experiments. Bioinformatics 28:2184–2185

Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859–1875

Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13:329–342