Genomic and transcriptomic survey of an endophytic fungus *Calcarisporium arbuscula* NRRL 3705 and potential overview of its secondary metabolites

Jintao Cheng
Zhejiang University
Fei Cao
Zhejiang University
Xinai Chen
Zhejiang University
Yongquan Li
Zhejiang University
Xuming Mao (xmmao@zju.edu.cn)
https://orcid.org/0000-0003-0669-7630

Research article

Keywords: Endophytic Fungus, Calcarisporium arbuscula, Genome, Transcriptome, Secondary Metabolite

Posted Date: January 16th, 2020

DOI: https://doi.org/10.21203/rs.2.16691/v2

Version of Record: A version of this preprint was published at BMC Genomics on June 24th, 2020. See the published version at https://doi.org/10.1186/s12864-020-06813-6.

Abstract

Background: Endophytic fungi can produce many active secondary metabolites, which are important resources of natural medicines. However, there is currently little understanding of endophytic fungi at the omics levels. *Calcarisporium arbuscula*, an endophytic fungus from the healthy fruit of Russulaceae, can produce a variety of secondary metabolites with anti-cancer, anti-nematode and antibiotic bioactivities. Comprehensive survey of the endophytic fungi genome and transcriptome will help to understand their capacity to biosynthesize secondary metabolites and lay the foundation for the development of these precious resources.

Results: In this study we reported the high-quality genome sequence of a strain *C. arbuscula* NRRL 3705 based on Single Molecule Real-Time sequencing technology. The genome of this fungus is over 45 Mb in size, relatively larger than other typical filamentous fungi, and comprises 10,001 predictable genes, encoding at least 762 secretory-proteins, 386 carbohydrate-active enzymes and 177 P450 enzymes. 398 virulence factors and 228 genes related to pathogen-host interactions were also predicted in this fungus. Moreover, 65 secondary metabolite biosynthetic gene clusters were revealed, including the gene cluster for mycotoxins aurovertins. In addition, several gene clusters were predicted to produce various mycotoxins, including aflatoxin, alternariol, destruxin, citrinin and isoflavipucine. Notably, two independent gene clusters were shown possibly involved in the biosynthesis of alternariol. Furthermore, RNA-Seq assay showed that only the expression of aurovertin gene cluster is much stronger than the housekeeping genes under laboratory conditions, consistent with that aurovertins are the predominant metabolites. The gene expression of the remaining 64 gene clusters for compound backbone biosynthesis was all lower than the housekeeping genes, which might partially explain poor production of other secondary metabolites in this fungus.

Conclusions: Our omics data along with bioinformatics analysis indicated that *C. arbuscula* NRRL 3705 contains a large number of biosynthetic gene clusters and has a huge potential to produce profound secondary metabolites. This work also provides the basis for development of endophytic fungi as a new resource of natural products with promising biological activities.

Background

Fungi are important resources of natural product-derived drugs, such as penicillin, cephalosporins, lovastatin and cyclosporin A [1, 2]. Endophytic fungi are those that live in various tissues and organs of healthy hosts at a certain stage or all stages of their life history, and generally do not confer external symptoms to the infected hosts [3]. They can be developed as biopesticides by artificial introduction into other plants and thus inherit by the host seeds. Endophytic fungi are also gradually attracting scientists' attention due to profound production of natural products, especially some bioactive compounds, such as taxol, sequoiatone A and B (antitumor) [4, 5], cryptocandin and cryptocin (antibiotics) [6, 7], peramine, loline, lolitrem B and ergovaline (insecticides) [8, 9], IAA, acetonitrile (plant growth regulators) and other types of active products subglutinol A and B [10, 11]. However, there is currently little research of the

biosynthesis capacity of endophytic fungi and its secondary metabolites, especially at the omics levels, which has limited our understanding and development of these resources.

Calcarisporium, a genus of fungus founded by Preuss, is characterized by a transparent, conical, spore-like sporophyte with spores [12]. Most researches on this genus focus on species classification and biomorphological studies. Some reports also showed some bioactive natural products from the fermentation of this fungal genus, such as 15G256α, 15G256α-2, 15G256β, 15G256β-2 and calcarides A-E [13], suggesting that fungi in this genus might be new promising resources of natural products. However, no details about the genomic information of the relevant species in this genus have been reported.

Calcarisporium arbuscula is an endophytic fungus from the fruit-bodies of Russulaceae, aiding the resistance to other fungi by producing certain antibiotics [14]. It can produce a large number of aurovertin-type mycotoxins as inhibitors of F0F1-ATP synthase [15, 16], such as aurovertin B as a potential therapeutic against cancer [15], and aurovertin D with strong toxicity towards the root-knot nematode *Meloidogyne incognita* [17]. *C. arbuscula* is also considered as a myco-parasite to kill the pathogen of coffee plantations - *Hemileia vastatrix* [18]. In addition, the draft genome sequence has shown a rich repertoire of natural products in this fungus, and intriguing compounds with attractive structures and bioactivities were discovered after epigenetic activation [16, 19]. These findings suggested that *C. arbuscula* is of great potential significance for biological control and new drug development. However, the lack of detail information about its genome and transcriptome has limited our further understanding and development of this fungus as a representative species of *Calcarisporium* genus.

Recently, a large number of fungal genome programs have been launched (1000 fungal genomes project, http://1000.fungalgenomes.org) to facilitate the access to more secondary metabolites at the genomic level. Genomic studies have shown that fungi contain a larger number of biosynthetic gene clusters than ever expected for secondary metabolite production [20]. However, most gene clusters are silent under laboratory conditions and the fungi are unable to produce corresponding secondary metabolites.

In this study, we reported the genome sequence of *C. arbuscula* NRRL 3705 in high quality by Single Molecule Real-Time (SMRT) sequencing technology. The genome annotation and transcriptome assays revealed that *C. arbuscula* NRRL 3705 harbors many secreted proteins and virulence factors. This genome contains a large number of gene clusters involved in biosynthesis of secondary metabolites, including aurovertins and other mycotoxins. We demonstrated that silencing of most gene clusters in *C. arbuscula* NRRL 3705 is most likely due to the low level of transcription as revealed from RNA-Seq assay. Moreover, the genome information can be further used for comparative genomic studies and discovery of more novel secondary metabolites.

Results And Discussion

Genome sequencing and annotation

The genome of *C. arbuscula* NRRL 3705 was sequenced by Illumina Miseq technology and the third generation sequencing technology (Single-Molecule Real-Time sequencing technology) with the coverage of over 100X [21]. This method can efficiently decode difficult but important genomic areas, such as methylated regions, repetitive elements and non-coding regions for possible gap-free eukaryotic genome assembly. Sub-read distribution analysis confirmed the high quality of the 20-kb library (Additional file 1: Fig S1). Moreover, we had RNA-Seq results serving as a reference for genome annotation (Additional file 2：Table S1 and Table S2).The genome was finally assembled in size approximately 45.01 Mb comprising 91 contigs as displayed by circos-plot (Fig. 1) with an N50 length of 1,530,317 bp. This genome is larger than that of *Calcarisporium sp.* KF525 (Table 1) [22]. Totally 10,001 genes were predicted and the average gene length is 1,365 bp. The total coding regions of *C. arbuscula* are 13.6 Mb, accounting for 30.2% of the entire genome. Statistics analysis for gene length distribution of *C. arbuscula* showed that about 752 genes have the length over 2,500 bp (Additional file 1: Fig S2). Notably, compared to other commonly studied filamentous fungi, *C. arbuscula* has a relatively larger genome, while most other fungal genomes are less than 40 Mb in size (Table 2) [23-26]. In addition, there are also sections about non-coding proteins (Additional file 2：Table S3).

Among all coding genes, 9,397 genes could be annotated by various databases (Additional file 1: Fig S3). Using the Non-Redundant Protein Database for protein annotation, we found that 8816 genes were protein-encoding genes (Additional file 1: Fig S3), which account for 88.16% of the total coding genes. KEGG analysis revealed that the products of most genes are involved in metabolism: carbohydrates (≈ 321 proteins), amino acids (≈ 252 proteins), or lipids (≈ 161 proteins) (Additional file 1: Fig S4). These data suggested that *C. arbuscula* may produce a large number of enzymes involved in its rich metabolic processes. Gene Ontology (GO) functional classification of *C. arbuscula* (Additional file 1: Fig S5) also showed that most genes are involved in catalytic activity (≈ 3649 proteins / strain) and metabolic processes (≈ 3654 proteins / strain). In addition, Eukaryotic Orthologous Groups (KOG) functional classification showed that many genes are involved in posttranslational modifications (Additional file 1: Fig S6).

Taxonomy

C. arbuscula NRRL 3705 is an endophytic fungus in fruit-bodies of Russulaceae, producing aurovertin-type mycotoxins that are potent against F0F1-ATPase and breast cancers [14-16]. According to the fungus taxonomy, it belongs to calcarisporium, hypocreales, pezizomycotina, ascomycota. Spore of *C. arbuscula* NRRL 3705 culture on potato dextrose agar (PDA) medium for 5 days at 25°C. The filamentous fungus displays high sporulation and the conidial heads are yellow (Additional file 1: Fig S7).

The phylogenetic analysis performed in this study with several reference aligned sequence matrix (ITS, SSU, LSU, TEF1-α and RPB2) and revealed the close relationship between *C. arbuscula* NRRL 3705 and other strains. The multilocus analysis was performed on our isolate with 17 reference strains (NCBI accession numbers were available in Additional file 3). For species delimitation, the aligned sequences matrix of (ITS, SSU, LSU, TEF1-α and RPB2) sequences data for *Calcarisporium* and for *Cordyceps*

militaris, and *C. brongniartii* was the outgroup taxa. The phylogenetic tree was constructed with maximum likelihood and Bayesian analysis, and resulted in high bootstrap values (Fig. 2). This tree showed that *C. arbuscula* NRRL 3705 was most closely related to *C. arbuscula* 111.57 and *C. arbuscula* 144.52.

Repetitive elements

Repeated sequence plays an important role in maintaining the spatial structure of chromosomes, gene expression regulation and genetic recombination of fungi [29]. Totally 1,387,508 bp repeat sequences were identified in *C. arbuscula* genome, including LTR retrotransposons, DNA transposons, long interspersed repeated elements (LINEs), tandem repeat sequences (TR) and mini-satellite DNA. Interestingly, the majority of repetitive sequences (63.6%) are tandem repeat sequences, whereas the dispersed repetitive-sequence just accounts for 30.39%. Notably, the highest percentage of all repeat sequences is TR, at 38.29% (Fig. 3).

Predicted candidate secreted effectors involved in virulence and pathogenicity

Secreted effectors play critical roles in virulence and pathogenicity [30, 31]. Signal peptide prediction tool SignalP was used to determine the proteins containing the cross-membrane structures [32]. Totally, there were 762 possible secretion proteins identified, accounting for 7.62% of all predicted proteins (10,001). Based on the alignment analysis of all predicted proteins in the genome against the pathogen-host interaction (PHI) database [33], 228 out of 10,001 (2.29%) predicted proteins were related to pathogenicity, of which 21 (0.21%) putative PHI-related proteins were potential secreted effectors.

After whole proteome BLAST against the database of fungal virulence factor (DFVF) [34], 398 out of 10,001 (39.8%) predicted proteins encoded within the *C. arbuscula* genome were identified to share homology with the genes implicated in virulence, of which 63 (0.63 %) putative DFVF-related proteins were predicted to be secreted. And 62 of them were predicted involved in pathogen-host interaction (Fig. 4).

P450 enzymes not only participate in the production of important internal metabolites, but also play an important role in adaptation to different environments by modifying harmful chemicals [35]. By BLASTP, the amino acid sequences of all *C. arbuscula* proteins were compared to the Fungal Cytochrome P450 Database (FCPD), 177 out of 10,001(1.77%) was identified as putative CYP450 enzymes, part of which are involved in fungal virulence factor and pathogen-host interaction (Fig. 4).

Carbohydrate-active enzymes

Carbohydrate-active enzymes (CAZy) play an important role in carbohydrate degradation, modification and biosynthesis in fungi [36]. CAZy is known as Carbohydrate-Active enZYmes Database [37], a specialized database of carbohydrate enzymes, including a family of related enzymes that catalyze the degradation, modification, and biosynthesis of carbohydrates. It contains five main categories: Glycoside Hydrolase (GH) [38], Glycosyl Transferase (GT) [39], Polysaccharide Lyases (PL) [40], Carbohydrate

Esterases (CE) [41], and Oxidoreductase (Auxiliary Activitie, AA). In addition, it also contains Carbohydrate-Binding Module (CBM) [42]. In *C. arbuscula* genome, 386 proteins were identified as CAZymes，part of which are involved in pathogen-host interaction (Fig. 4). GH is the highest proportion (62.29%) of all CAZymes (Additional files 1: Fig S8). Based on the genomic information, we compared the potential for hydrolysis with eight *Aspergillus* species. Notably, *C. arbuscula* NRRL 3705 contains a large amount of glycoside hydrolases GH18 and GH2, much more than other fungi (Additional file 4). The fungal glycoside hydrolase GH18 is mainly responsible for the remodeling and recovery of the fungal cell wall and other cell wall degrading enzymes [43]. The fungal glycoside hydrolases GH2 can specifically hydrolyze the β-glycosidic bond between d-glucuronidose and aglycon, and have important applications in the diagnosis and drug development of metabolic diseases [44].

Secondary metabolite biosynthetic gene clusters in *C. arbuscula* NRRL 3705

The secondary metabolites of fungi constitute a rich resource of bioactive compounds with potential pharmaceutical values as antibiotics, cholesterol-lowering drugs and antitumor drugs [1]. Interestingly, the genes encoding the biosynthetic pathway responsible for the production of such a secondary metabolite are very often spatially clustered together at a certain position on the chromosome; such a compendium of genes is referred to as a 'secondary metabolite biosynthesis gene cluster' [45]. Based on profile hidden Markov models of genes that are specific for certain types of gene clusters and antiSMASH 4.0, we identified 65 gene clusters for secondary metabolites in *C. arbuscula* NRRL 3705. Among them, 23 and 12 gene clusters containing genes encoding polyketide synthases (PKS) and non-ribosomal peptides synthases (NRPS) were identified, respectively. In addition, there are gene clusters for terpenes, PKS/NRPS hybrids, indoles and other types of natural products (Table 3 and Additional file 5). Some of these gene clusters are highly similar to known gene clusters (Table 4).

Aurovertin biosynthetic gene cluster

Aurovertins are a class of toxic polyketides harboring a unique structure of a 2, 6-dioxabicyclo-[3.2.1]-octane (DBO) ring system and a conjugated a-pyrone moiety [16, 46]. Due to the unusual polyketide-derived structure, aurovertins have been shown to have potent antiviral, antitumor and antibacterial activities. *C. arbuscula* is capable of predominantly producing aurovertins (Fig. 5a), and the biosynthetic gene cluster for these mycotoxins has been identified [16]. In addition, LC-MS analysis was performed on a methanol extract obtained from 7-days culture of *C. arbuscula* NRRL 3705 grown on PDA plate at 25°C (Additional files 1: Fig S9 and Fig S10). We have also analyzed a phylogenetic analysis about the aurovertin-related gene cluster of different strains (Additional files 1: Fig S11). The aurovertin biosynthetic gene cluster was composed mainly of seven genes including *aurA, aurB, aurC, aurD, aurE, aurF* and *aurG*. However, some genes are missing in the cluster after the genome annotation with the automatic bioinformatics.

According to antiSMASH 4.0, there are totally 4 genes in this gene cluster 23 (aurovertin biosynthetic gene cluster). These genes mainly encode a PKS (A05996), a SAM-dependent methyltransferase (A05995), a FAD-dependent monooxygenase (A05994), and an acetyltransferase (A05993) (Fig. 5b). We found that

there is a 7 kb spacer between gene A05993 and A05994, which was re-predicted by the web-based software Softberry（http://www.softberry.com/）and found that this spacer contains three known genes *aurD*, *aurE* and *aurF*. This is consistent with the gene cluster reported previously [16]. This also indicates that there are certain defects in genome sequencing and automatic NR annotation.

Other SM clusters for mycotoxin biosynthesis

Based on antiSMASH predictions, *C. arbuscula* NRRL 3705 has a potential to produce variety of mycotoxins. Aflatoxin (AFT), a class of toxic secondary metabolites originally produced by *Aspergillus parasiticus* and highly toxic, carcinogenic, mutagenic and teratogenic [47]. Cluster 60 is composed of 15 genes and contains a PKS (A09345), a putative ketoreductase (A09348), a transcription factor (A0934) and two cytochrome P450 monooxygenases (A09346 and A09350). PKS (A09345) of cluster 60 is in high identity with AflC (a polyketide synthase involved in aflatoxin biosynthesis) in *A. ochraceoroseus* (protein coverage: 98%; identity: 79%). Moreover, two cytochrome P450 monooxygenases of cluster 60 have highest homology with AflV (protein coverage: 99%; identity: 83%) and AflG (protein coverage: 96%; identity: 79.8%) of *A. ochraceoroseus* (Fig. 6). These *in silico* data suggested that *C. arbuscula* potentially produces compounds and derivatives structurally related to aflatoxin.

In addition, we found that *C. arbuscula* has the potential to produce alternariol (AOH) [48]. Alternarolis, a secondary metabolite produced by *Alternaria* and other fungi, is harmful to animals and plants. One polyketide synthase (PKS 19) from *Parastagonospora nodorum* has been shown responsible for AOH production in this fungus. Surprisingly, we found two candidate gene clusters (cluster 35 and cluster 44) in high similarity to alternariol biosynthetic gene cluster from *P. nodorum* SN15. Cluster 35 is composed of 9 genes. The backbone gene encodes a PKS (A07007) and other genes encode an NAD^+-binding protein (A07006), an acyl-CoA-acyltransferase (A07005), an aldehyde dehydrogenase (A07004), an integral membrane (A07003), an arginosuccinate synthetase (A07002), an ABC transporter (A07001), a putative capsule polysaccharide biosynthesis protein (A07008) and a transcriptase (A07009) (Fig. 7a). Cluster 44 is also composed of 9 genes for a PKS, four putative signal sequence proteins and a transcription factor (Fig. 7b). PKS from cluster 35 has higher identity to PKS 19 (72%) than PKS from cluster 44 (44%). Considering that the cluster for AOH biosynthesis in *P. nodorum* contains one PKS, four tailoring enzymes (*O*-methyl transferase OmtI, monooxygenase MoxI, short chain dehydrogenase like protein SdrI and an estradiol dioxygenase DoxI), cluster 35 is more likely for the biosynthesis of AOH. This is not the case in cluster 44, which lacks multiple enzymes as shown above. So we hypothesized that cluster 35 could be more likely the putative SM cluster for biosynthesis of AOH. However, it needs to be validated by further genetic and biochemical analysis.

The non-ribosome polypeptide synthase is responsible for the synthesis of peptide secondary metabolites, such as surugamides and ferricrocin [49, 50]. In *C. arbuscula*, 12 putative NRPS genes were found. According to antiSMASH and MIBiG, cluster 37 and cluster 59 have similarity with the biosynthetic gene cluster of destruxin, a secondary metabolite of non-ribosomal cyclic hexapeptides with insecticidal and pharmaceutically active activities (Additional file 4).

Gene clusters for biosynthesis of other secondary metabolites

In *C. arbusclua* genome, seven hybrid NRPS / PKS gene clusters were also predicted. Among them, cluster 26, cluster 41 and cluster 58 have similarity to the gene clusters of aculeacin A, citrinin and isoflavipucine, respectively (Additional file 4) [51-53]. Moreover, we also predicted eleven terpene genes, whose products remain to be determined. In addition to the PKS, NRPS, and hybrid NRPS/PKS gene clusters, we also identified 12 gene clusters likely to produce indoles, one hybrid indole-t1PKS and one hybrid t1PKS-terpene.

Gene cluster expression by RNA-Seq analysis

Although *C. arbuscula* contains a large number of biosynthetic gene clusters, this fungus rarely produces secondary metabolites, except aurovertins [16]. It is very likely that under laboratory culture conditions, the expression of the core genes of most gene clusters is low. RNA-Seq assay was further performed and the gene expression was evaluated based on FPKM from RNA-Seq. With three housekeeping genes (*gpdA*, *tubC* and *actA*) as reference genes, we found only the PKS gene in cluster 23 (aurovetin biosynthetic gene cluster) was expressed in the highest level in core genes of all gene clusters (Fig. 8). These results confirmed that most gene clusters are expressed in low levels or silenced. However, nine gene clusters contain pathway-specific transcription factors (Additional file 4), which provides the possibility to activate these gene clusters by overexpression of these transcription factors to obtain secondary metabolites.

Conclusions

In this study, we reported the high-quality genome sequence of *C. arbuscula* NRRL 3705. Phylogenetic tree analysis showed that this fungal strain is unique from the same species. This fungus contains 65 gene clusters involved in biosynthesis of a variety of secondary metabolites, including aurovertins and other putative mycotoxins. We also demonstrated that most gene clusters are silenced or in low expression, most likely due to the low levels of transcription based on RNA-Seq assay. The genomic and transcriptomic survey of *C. arbuscula* NRRL 3705 will help us in further development of this fungus for discovery of new bioactive natural products.

Methods

Fungal culture and genomic DNA extraction

The isolated spores of *C. arbuscula* NRRL 3705 were inoculated on potato dextrose agar (PDA) (Sigma) medium. Fungal mycelia were grown at 25°C for 5 days, collected and grounded into powder with liquid nitrogen. Genomic DNA was extracted with the cetyltrimethylammonium bromide (CTAB) method [54]. The harvested DNA was detected by the agarose gel electrophoresis and quantified by Nanodrop.

Genome sequencing and assembly

The genome of *C. arbuscula* NRRL 3705 was sequenced by Single Molecule Real-Time (SMRT) technology [21]. Sequencing was performed at Beijing Novogene Bioinformatics Technology. The low quality reads were filtered by the SMRT Link v5.0.1 and the filtered reads were assembled to generate contigs without gaps. DNA libraries with 350 bp and 20 kb inserts were constructed. The 350-bp library was constructed following Illumina's standard protocol, including fragmentation of genomic DNA, end repair, adaptor ligation and PCR amplification [55]. The 20-kb library was constructed following PacBio's standard methods, including fragmentation of genomic DNA, end repair, adaptor ligation, and templates purification [55].

Genome annotation

Genome annotation was performed based on *de novo* prediction and transcriptome-assisted gene prediction. Seven databases were used to predict gene functions, including GO (Gene Ontology), KEGG (Kyoto Encyclopedia of Genes and Genomes), KOG (Clusters of Orthologous Groups), NR (Non-Redundant Protein Database databases), TCDB (Transporter Classification Database), P450, and Swiss-Prot. A whole genome BLAST search (e-value less than 1e-5, minimal alignment length percentage larger than 40%) was performed against above seven databases. The secretory proteins were predicted by the Signal P database. For pathogenic fungi, we added the pathogenicity and drug resistance analysis. We used PHI (Pathogen Host Interactions), DFVF (database of fungal virulence factors) to perform the above analysis. Carbohydrate-Active enzymes were predicted by the Carbohydrate-Active enZYmes Database. BLAST alignment of predicted genes with various functional databases (BLASTP, e-value ≤ 1e-5); BLAST result filtering: For each BLAST result of the sequence, select the highest score alignment (default identity >= 40%, coverage >= 40%) for annotation of *C. arbuscula*.

PCR amplification and DNA sequencing

Sequences of ITS and partial LSU ribosomal RNA, partial small-subunit (SSU) ribosomal RNA, translation elongation factor 1 alpha (TEF1-α), and the second largest subunit of RNA polymerase II (RBP2) were amplified by polymerase chain reaction (PCR) with the primer pairs ITS1–ITS4, LR5F–LROR, NS1–NS4, EF983F–EF2218R and RPB2-5F–RPB2-7cR, respectively. Each amplification reaction included 2 mM each dNTP, 0.4 mM each primer, 1 U of KOD FX (Takara, China), 2 μL of genomic DNA solution, 2 × KOD FX buffer in a 50 μL-reaction volume. A typical reaction included initial denaturation at 94°C for 5 min, followed by 30 cycles of denaturation at 98 °C for 10s, annealing at 52 °C for 30 s, and extension at 68°C for 60 s. A final extension was demonstrated at 68 °C for 10 min. Automated sequencing was performed by TsingKe Biological Technology (Hangzhou, China).

Phylogenetic analysis

The SSU, ITS, LSU, TEF1-α and RPB2 data sets of *Cacarisporium* species determined from recent studies [12], were downloaded from GenBank (Additional file 3) and used in the phylogenetic analysis. Single and combined genes were analyzed with maximum likelihood (ML) performed in RAxML implemented in raxmlGUI v.1.3 with rapid bootstrap analysis with 1000 replicates. For Bayesian analyses, the posterior

probabilities were determined by Markov chain Monte Carlo sampling (MCMC) in MrBayes v3.2 based on the models from MrModeltest.

PKS sequences were aligned by BLASTN to obtain higher homology sequences in different stranis. Phylogenetic tree about aurA was generated with MEGA7.0 based on Neighbor-Joining method [56]. Bootstrap values were calculated from 500 replications of the bootstrap procedure using phylogeny.fr and added to the phylogenetic tree.

Repetitive sequences

Repetitive sequences were predicted with RepeatMasker software (version 4.0.5) and Tandem Repeats Finder software [57].

Biosynthetic gene cluster prediction

Gene clusters were predicted by the web-based software antiSMASH database (antibiotics and Secondary Metabolite Analysis 4.0) [45]. The core genes were annotated using stand-alone BLAST against Swiss-Prot database. Gene cluster domain prediction can be obtained by PKS/NRPS Analysis web-site (http://nrps.igs.umaryland.edu/).

LC-MS analysis

LC–MS analysis was performed in an Agilent 1200HPLC system (Agilent, Santa Clara, CA, United States) and a Termo Finnigan LCQDeca XP Max LC/MS system (Termo Finnigan, Waltham, MA, United States). Poroshell 120 SB C18 was used as the column, H2O (containing 0.1% formic acid) and acetonitrile (containing 0.1% formic acid) were used as the mobile phase A and B performing a linear gradient from 30 to 100% (v/v) B over 30 min.

Transcriptome analysis

Library preparation and bioinformatics analysis were performed according to the method of Chen et al [55]. All three RNA samples were prepared from *C. arbuscula* mycelia as above and subjected to RNA-Seq on the Illumina HiSeq 2000 platform (Illumina, San Diego, CA, USA). *C. arbuscula* wild type was grown on solid PDA media (sigma) at 25°C for 5 days. Mycelia were and sent to the company for RNA-Seq assay after being treated with liquid nitrogen. FPKM (fragments per kilobase of transcript per) value was used to evaluate gene expression, and the upper-quartile algorithm was used to correct the gene expression.

Declarations

Acknowledgments

We gratefully thank Dr. Zhifeng Zhang in Institute of Microbiology, Chinese Academy of Sciences, for his kind helps in the evolutionary tree analysis

Funding
This work was financially supported by National Key R&D Program of China (2018YFA0903202) and National Natural Science Foundation of China (31770071).

Availability of data and materials
This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank

under the accession WBSA00000000. The version described in this paper is versionWBSA01000000.

Authors' contributions
JTC, YQL and XMM conceived the study. JTC created its design and performed bioinformatics analysis. JTC and XMM drafted the manuscript. CF and XAC performed the experiments and genome sequencing preparations. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details

[1]Institute of Pharmaceutical Biotechnology, School of Medicine, Zhejiang University, Hangzhou 310058, China

[2]Zhejiang Provincial Key Laboratory for Microbial Biochemistry and Metabolic Engineering, Hangzhou 310058, China.

Abbreviations

SMRT, Single Molecule RealTime; antiSMASH, Antibiotics & Secondary Metabolite Analysis Shell; BLAST, Basic local alignment search tool; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; KOG, Clusters of Orthologous Groups; NR, Non-Redundant Protein Database databases; PHI, Pathogen Host Interactions; DFVF, database of fungal virulence factors; CAZyme, Carbohydrate activity enzyme; CBM, Carbohydrate binding module; CYP450, Cytochrome P450; GH, Glycoside hydrolases; GT, Glycosyl transferase; PL, Polysaccharide Lyases; CE, Carbohydrate Esterases; AA, Auxiliary Activitie. PKS, Polyketide synthase; NRPS, Non-ribosomal peptide synthase; SM, Secondary metabolite; T1PKS, Type I PKS.

References

1. Alberti F, Foster GD, Bailey AM. Natural products from filamentous fungi and production by heterologous expression. Appl Microbiol Biotechnol. 2017;101(2):493-500.
2. Hoffmeister D, Keller NP. Natural products of filamentous fungi: enzymes, genes, and their regulation. Nat Prod Rep. 2007;24(2):393-416.
3. Fernandes EG, Pereira OL, Silva CCd, Bento CBP, Queiroz MVd. Diversity of endophytic fungi in Glycine max. Microbiol Res. 2015;181:84-92.
4. Stierle A, Strobel G, Stierle D. Taxol and taxane production by Taxomyces andreanae, an endophytic fungus of *Pacific yew*. Science. 1993;260(5105):214.
5. Stierle AA, Stierle DB, Bugni T. Sequoiatones A and B: novel antitumor metabolites isolated from a redwood endophyte. J Org Chem. 1999;64(15):5479-5484.
6. Strobel GA, Miller RV, Martinez-Miller C, Condron MM, Teplow DB, Hess WM. Cryptocandin, a potent antimycotic from the endophytic fungus *Cryptosporiopsis cf. quercina*. Microbiology. 1999;145(8):1919-1926.
7. Li JY, Strobel G, Harper J, Lobkovsky E, Clardy J. Cryptocin, a potent tetramic acid antimycotic from the endophytic fungus *Cryptosporiopsis cf. quercina*. Org Lett. 2000;31(27).
8. Strickland JR, Cross DL, Birrenkott GP, Grimes LW. Effect of ergovaline, loline, and dopamine antagonists on rat pituitary cell prolactin release in vitro. Am J Vet Res. 1994;55(5):716-721.
9. Fuchs B, Krischke M, Mueller MJ, Krauss J. Peramine and lolitrem B from endophyte-grass associations cascade uUp the food chain. J Chem Ecol. 2013;39(11):1385-1389.
10. Lee JC, Lobkovsky E, Pliam NB, Strobel G, Clardy J. Subglutinols A and B: immunosuppressive compounds from the endophytic fungus *Fusarium subglutinans*. Cheminform. 1996;27(12).
11. Hong LU, Wen XZ, Meng JC, Jun HU, Ren XT. New bioactive metabolites produced by *Colletotrichum* sp., an endophytic fungus in *Artemisia* annua. Plant Sci. 2000;151(1):67-73.
12. Sun J-Z, Liu X-Z, Hyde KD, Zhao Q, Maharachchikumbura SSN, Camporesi E, Bhat J, Nilthong S, Lumyong S. *Calcarisporium xylariicola* sp. nov. and introduction of *Calcarisporiaceae* fam. nov. in Hypocreales. Mycol Prog. 2017;16(4):433-445.
13. Silber J, Ohlendorf B, Labes A, Erhard A, Imhoff JF. Calcarides A-E, antibacterial macrocyclic and linear polyesters from a *Calcarisporium* strain. Mar Drugs. 2013;11(9):3309-3323.
14. Watson P. Further observations on *Calcarisporium arbuscula*. Mycol Res. 1965;48(1):9.
15. Huang T-C, Chang H-Y, Hsu C-H, Kuo W-H, Chang K-J, Juan H-F. Targeting Therapy for breast carcinoma by ATP synthase inhibitor aurovertin B. J Proteome Res. 2008;7(4):1433-1444.
16. Mao XM, et al. Efficient biosynthesis of fungal polyketides containing the dioxabicyclo-octane ring system. J Am Chem Soc. 2015;137(37):11904-11907.
17. Wang Y-l, Li L-f, Li D-x, Wang B, Zhang K, Niu X. Yellow pigment aurovertins mediate interactions between the pathogenic fungus *Pochonia chlamydosporia* and its nematode host. J Agric Food Chem. 2015;63(29):6577-6587.
18. Carrion G, Rico-Gray V. Mycoparasites on the coffee rust in Mexico. Fungal Divers. 2002; 11:49-60.

19. Mao XM, Xu W, Li D, Yin WB, Chooi YH, Li YQ, Tang Y, Hu Y. Epigenetic genome mining of an endophytic fungus leads to the pleiotropic biosynthesis of natural products. Angew Chem Int Ed Engl. 2015;54(26):7592-7596.
20. Sanchez JF, Somoza AD, Keller NP, Wang CC. Advances in *Aspergillus* secondary metabolite research in the post-genomic era. Nat Prod Rep. 2012;29(3):351-371.
21. Chin C-S, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Method. 2013;10:563.
22. Kumar A, et al. Genome sequencing and analyses of two marine fungi from the north sea unraveled a plethora of novel biosynthetic gene clusters. Sci Rep. 2018;8(1):10187.
23. Galagan JE, et al. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. Nature. 2005;438(7071):1105-1115.
24. Gao Q, et al. Genome sequencing and comparative transcriptomics of the model entomopathogenic fungi *Metarhizium anisopliae* and *M. acridum*. PLoS Genet. 2011; 7(1):e1001264.
25. Machida M, et al. Genome sequencing and analysis of *Aspergillus oryzae*. Nature. 2005;438(7071):1157-1161.
26. Pel HJ, et al. Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. Nat Biotechnol. 2007;25(2):221-231.
27. Dubey A, Jeon J. Epigenetic regulation of development and pathogenesis in fungal plant pathogens. Mol Plant Pathol. 2017;18(6):887-898.
28. Luo X, Cao J, Huang J, Wang Z, Guo Z, Chen Y, Ma S, Liu J. Genome sequencing and comparative genomics reveal the potential pathogenic mechanism of *Cercospora sojina* Hara on soybean. DNA Res. 2017;25(1):25-37.
29. Ye P, Luan Y, Chen K, Liu Y, Xiao C, Xie Z. MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. Nucleic Acids Res. 2017;45(D1):D85-D89.
30. Koeck M, Hardham AR, Dodds PN. The role of effectors of biotrophic and hemibiotrophic fungi in infection. Cell Microbiol. 2011;13(12):1849-1857.
31. Gonzalez-Fernandez R, Jorrin-Novo JV. Contribution of proteomics to the study of plant pathogenic fungi. J Proteome Res. 2012;11(1):3-16.
32. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8(10):785-786.
33. Urban M, et al. PHI-base: a new interface and further additions for the multi-species pathogen-host interactions database. Nucleic Acids Res. 2017;45(D1):D604-D610.
34. Waseem H, Williams MR, Stedtfeld T, Chai B, Stedtfeld RD, Cole JR, Tiedje JM, Hashsham SA. Virulence factor activity relationships (VFARs): a bioinformatics perspective. Environ Sci Process Impacts. 2017;19(3):247-260.

35. Ladda MA, Goralski KB. The effects of CKD on cytochrome P450–mediated drug metabolism. Adv Chronic Kidney Dis. 2016;23(2):67-75.
36. Garron M-L, Henrissat B. The continuing expansion of CAZymes and their families. Curr Opin Chem Biol. 2019;53:82-87.
37. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42(Database issue):D490-D495.
38. Henrissat B. A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J. 1991;280:309-316.
39. Lairson LL, Henrissat B, Davies GJ, Withers SG. Glycosyltransferases: structures, functions, and mechanisms. Annu Rev Biochem. 2008;77(1):521-555.
40. Sutherland IW. Polysaccharide lyases. FEMS Microbiol Rev. 1995;16(4):323-347.
41. Armendáriz-Ruiz M, Rodríguez-González JA, Camacho-Ruíz RM, Mateos-Díaz JC. Carbohydrate esterases: nn Overview Methods Mol Biol. 2018;39-68.
42. Carvalho CC, Phan NN, Chen Y, Reilly PJ. Carbohydrate-binding module tribes. Biopolymers. 2015;103(4):203-214.
43. Alsina C, Faijes M, Planas A. Glycosynthase-type GH18 mutant chitinases at the assisting catalytic residue for polymerization of chitooligosaccharides. Carbohydr Res. 2019;478:1-9.
44. Domingues MN, et al. Structural basis of exo-β-mannanase activity in the GH2 family. J Biol Chem. 2018;293(35):13636-13649.
45. Blin K, et al. AntiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017;45(W1):W36-W41.
46. Azumi M, Ishidoh K-i, Kinoshita H, Nihira T, Ihara F, Fujita T, Igarashi Y. Aurovertins F–H from the entomopathogenic fungus *Metarhizium anisopliae*. J Nat Prod. 2008;71(2):278-280.
47. Srour AY, Fakhoury AM, Brown RL. Targeting aflatoxin biosynthetic genes. Methods Mol Biol. 2017;159-171.
48. Scott PM. Analysis of agricultural commodities and foods for Alternaria mycotoxins. J AOAC Int. 2001;84(6):1809-1817.
49. Zhou Y, Lin X, Xu C, Shen Y, Wang S-P, Liao H, Li L, Deng H, Lin H-W. Investigation of penicillin binding protein (PBP)-like peptide cyclase and hydrolase in surugamide non-ribosomal peptide biosynthesis. Cell Chem Biol. 2019;26(5):737-744.e734.
50. Wallner A, Blatzer M, Schrettl M, Sarg B, Lindner H, Haas H. Ferricrocin, a siderophore involved in intra- and transcellular iron distribution in *Aspergillus fumigatus*. Appl Environ Microbiol. 2009;75(12):4194-4196.
51. Torres-Bacete J, Hormigo D, Stuart M, Arroyo M, Torres P, Castillón MP, Acebal C, García JL. Newly discovered penicillin acylase activity of aculeacin A acylase from Actinoplanes utahensis. Appl Environ Microbiol. 2007; 73(16):5378-5381.

52. Pascual-Ahuir A, Vanacloig-Pedros E, Proft M. Toxicity mechanisms of the food contaminant citrinin: application of a quantitative yeast model. Nutrients. 2014; 6(5):2077-2087.
53. Guo CJ, Wang CCC. Recent advances in genome mining of secondary metabolites in *Aspergillus terreus*. Front Microbiol. 2014;5:717-717.
54. Liu L, Wang CL, Peng WY, Yang J, Lan MQ, Zhang B, Li JB, Zhu YY, Li CY. Direct DNA extraction method of an obligate parasitic fungus from infected plant tissue. Genet Mol Res. 2015; 14(4):18546-18551.
55. Chen J, Zeng X, Yang YL, Xing YM, Zhang Q, Li JM, Ma K, Liu HW, Guo SX. Genomic and transcriptomic analyses reveal differential regulation of diverse terpenoid and polyketides secondary metabolites in *Hericium erinaceus*. Sci Rep. 2017, 7(1):10151.
56. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016, 33(7):1870-1874.
57. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573-580.

Tables

Table 1 Genome characteristics and predicted features of *Calcarisporium* strains

Genome	*Calcarisporium arbuscula* NRRL 3705 (this study)	*Calcarisporium sp.* KF0525[22]
Assembly size (Mb)	45.01	36.8
G + C content (%)	49.75	50.6
N50 length (kbp)	153.03	95.7
N50 (num of scaffolds)	10	115
Average length (bp)	1,365	571
Assembled contigs	91	2274
Predicted proteins	10,001	15,459
Sequencing technology	Pacbio+Illumina	Illumina

Table 2 Summary of main genome features of *C. arbuscula* NRRL 3705 and four sequenced fungi

Species	Genome size (Mb)	%GC	proteins	Ref
C. arbuscula	45.01	49.75	10,001	This study
Aspergillus nidulans	30.1	50.32	10,560	[23]
Aspergillus niger	33.9	50.36	8,592	[26]
Aspergillus oryzae	36.7	48.24	12,063	[25]
Metarhizium robertsii	39.04	51.49	10,582	[24]

Table 3 Classification of putative secondary metabolite gene clusters of *C. arbuscula* NRRL 3705 predicted by antiSMASH

Type	Clusters number
Indole	1
indole-t1pks	1
Lantipeptide	1
t1pks-nrps	7
Terpene	11
Nrps	12
Other	8
t1pks-terpene	1
t1pks	23
Total	65

Table 4 Predict the possible products of secondary metabolite gene clusters in *C. arbuscula* NRRL 3705

Cluster	Type	smiliarty to known clusters
Cluster1	other	
Cluster2	terpene	
Cluster3	t1pks-nrps	
Cluster4	terpene	
Cluster5	t1pks-terpene	Sordarin (32% of genes show similarity)
Cluster6	nrps	
Cluster7	nrps	
Cluster8	terpene	
Cluster9	t1pks	
Cluster10	t1pks	
Cluster11	t1pks	
Cluster13	t1pks	
Cluster14	nrps	
Cluster16	terpene	
Cluster17	t1pks-nrps2385	
Cluster18	nrps	
Cluster19	terpene	
Cluster20	nrps	
Cluster21	nrps	
Cluster22	indole-t1pks	
Cluster23	t1pks	Citreoviridin (40% of genes show similarity)
Cluster24	t1pks	
Cluster25	terpene	
Cluster26	t1pks-nrps	Aculeacin A(NRPS 100% only)
Cluster27	t1pks	
Cluster28	lantipeptide	
Cluster29	terpene	
Cluster30	nrps	
Cluster31	t1pks	
Cluster32	t1pks	
Cluster33	nrps	Dimethylcoprogen (100% of genes show similarity)
Cluster34	t1pks	
Cluster35	t1pks	Alternariol (100% of genes show similarity)
Cluster36	nrps	
Cluster37	nrps	Destruxin (66% of genes show similarity)
Cluster38	t1pks	
Cluster40	t1pks-nrps	
Cluster41	t1pks-nrps	Citrinin(18% of genes show similarity)
Cluster42	terpene	Copalyl_diphosphate(28% of genes show similarity)
Cluster43	t1pks-nrps	Leucinostatins (10% of genes show similarity)
Cluster44	t1pks	Alternariol (100% of genes show similarity)
Cluster45	t1pks	
Cluster46	t1pks	
Cluster47	t1pks	
Cluster48	t1pks	
Cluster51	t1pks	
Cluster52	t1pks-nrps	
Cluster53	t1pks	
Cluster54	indole	
Cluster55	t1pks	
Cluster56	nrps	
Cluster57	terpene	

Cluster58	nrps	Isoflavipucine(12% of genes show similarity)
Cluster59	other	Destuxin (66% of genes show similarity)
Cluster60	t1pks	Aflatoxin (46% of genes show similarity)
Cluster61	t1pks	
Cluster62	t1pks	
Cluster63	terpene	
Cluster64	terpene	
Cluster65	other	

Figures

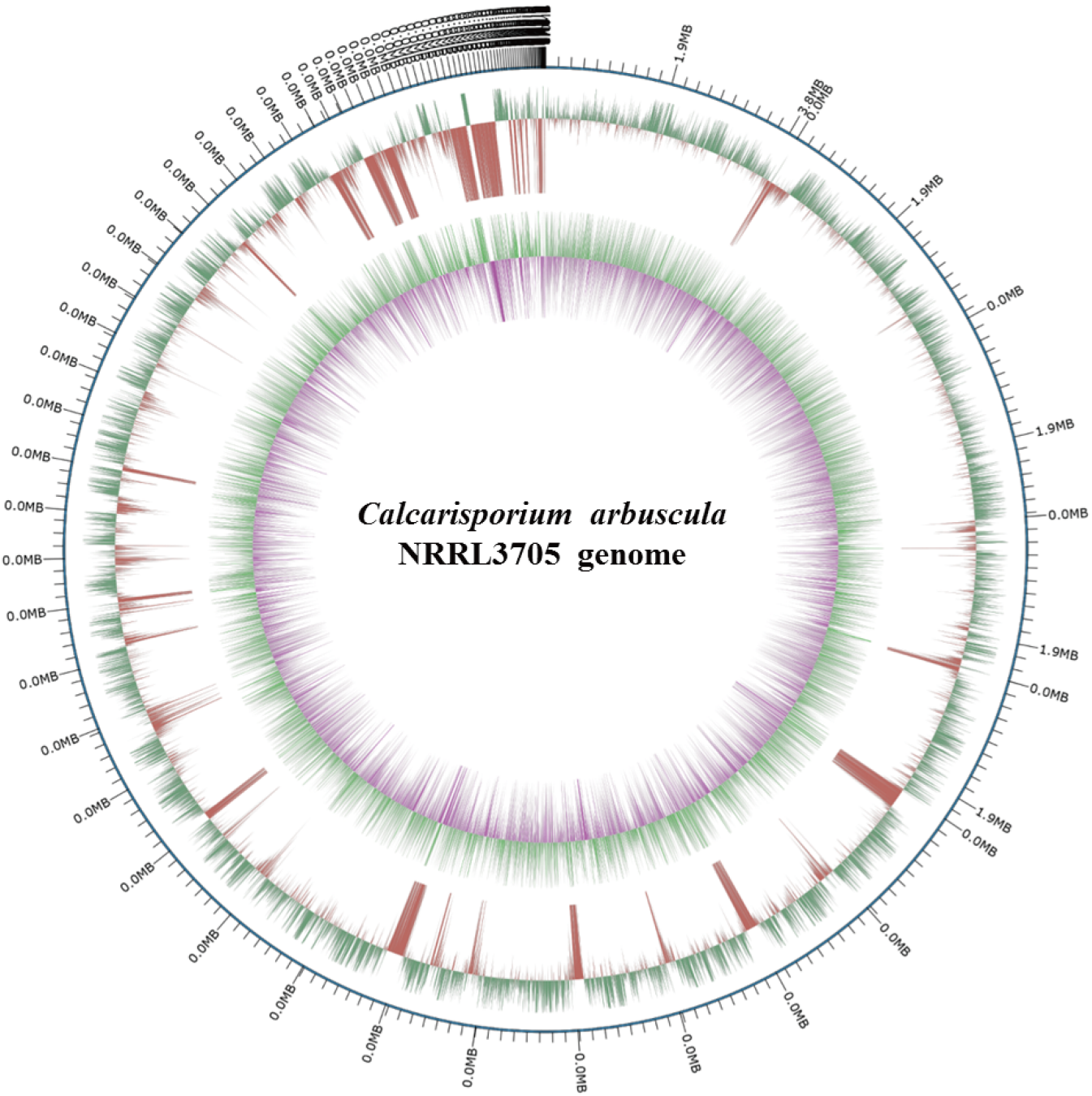


Figure 1

Circos-plot of C. arbuscula NRRL 3705. The 91 contigs of C. arbuscula NRRL 3705 are displayed by circos-plot (Mb scale). The circos from outside to inside are: (a) 99 contigs; (b) DNA methylations (+); (c) DNA methylations (-); (d) GC content; (e) GC preference.

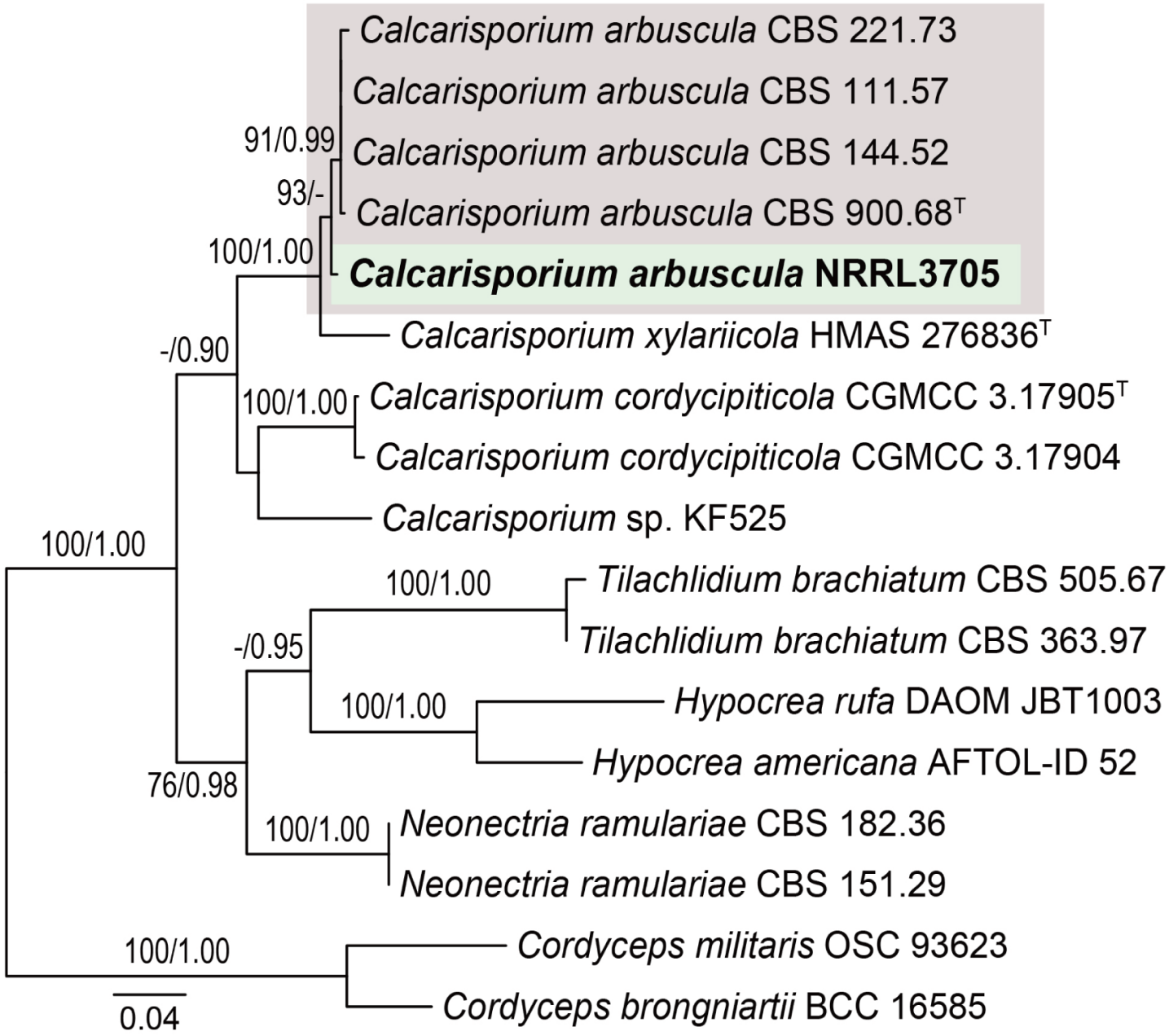


Figure 2

Phylogenetic and synteny analysis of C. arbuscula NRRL 3705 with other fungal species. Multilocus phylogenetic analysis of Calcarisporium based on a combined SSU, ITS, LSU, TEF1-α and RPB2 data set. The tree is rooted with Cordyceps militaris and Cordyceps brongniartii. Bootstrap values higher than 50% from RAxML (BSML) (left) are given above the nodes. Bayesian posterior probabilities greater than 0.90 are indicated (BYPP) (right). T indicates type.

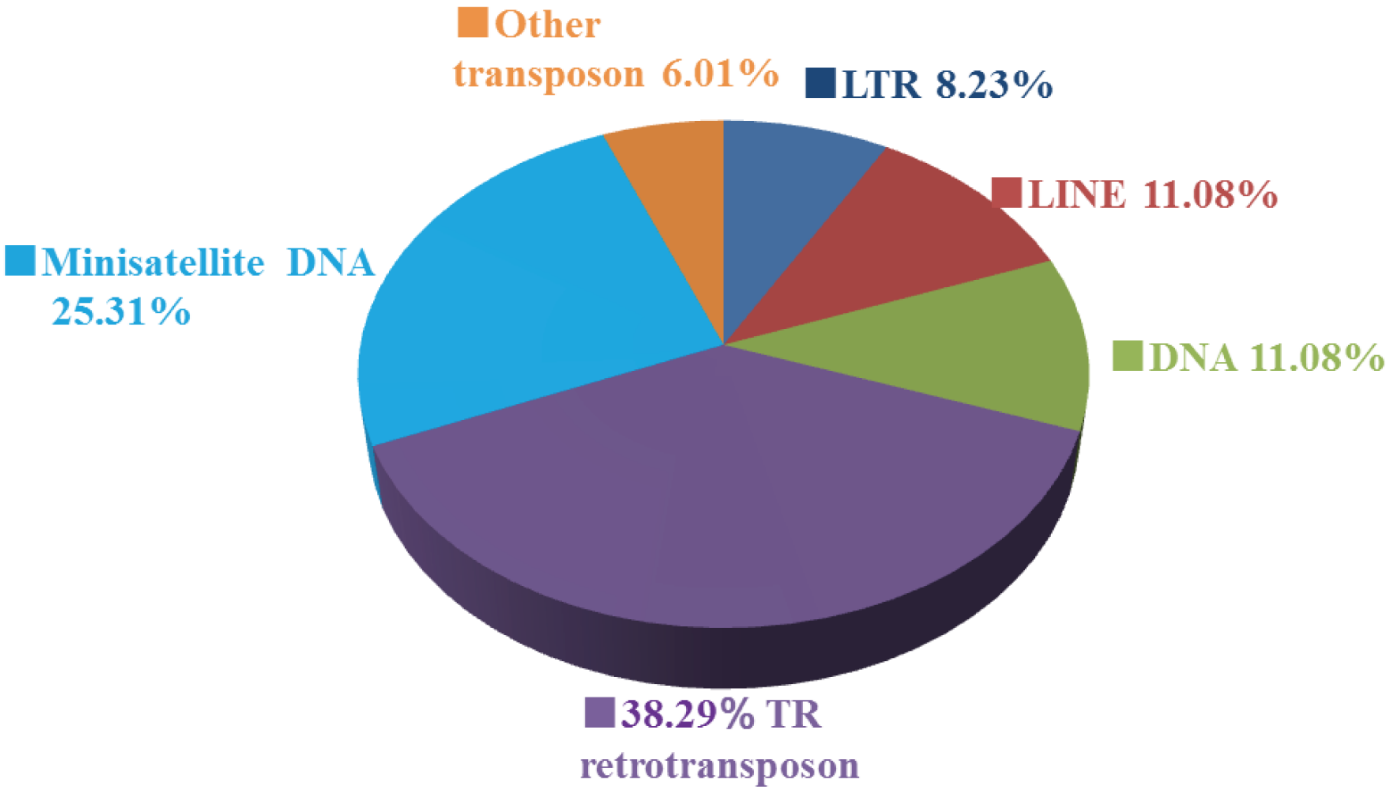


Figure 3

Repeat elements of C. arbuscula NRRL 3705. The percentage of different types of repetitive sequences in the C. arbuscula genome.

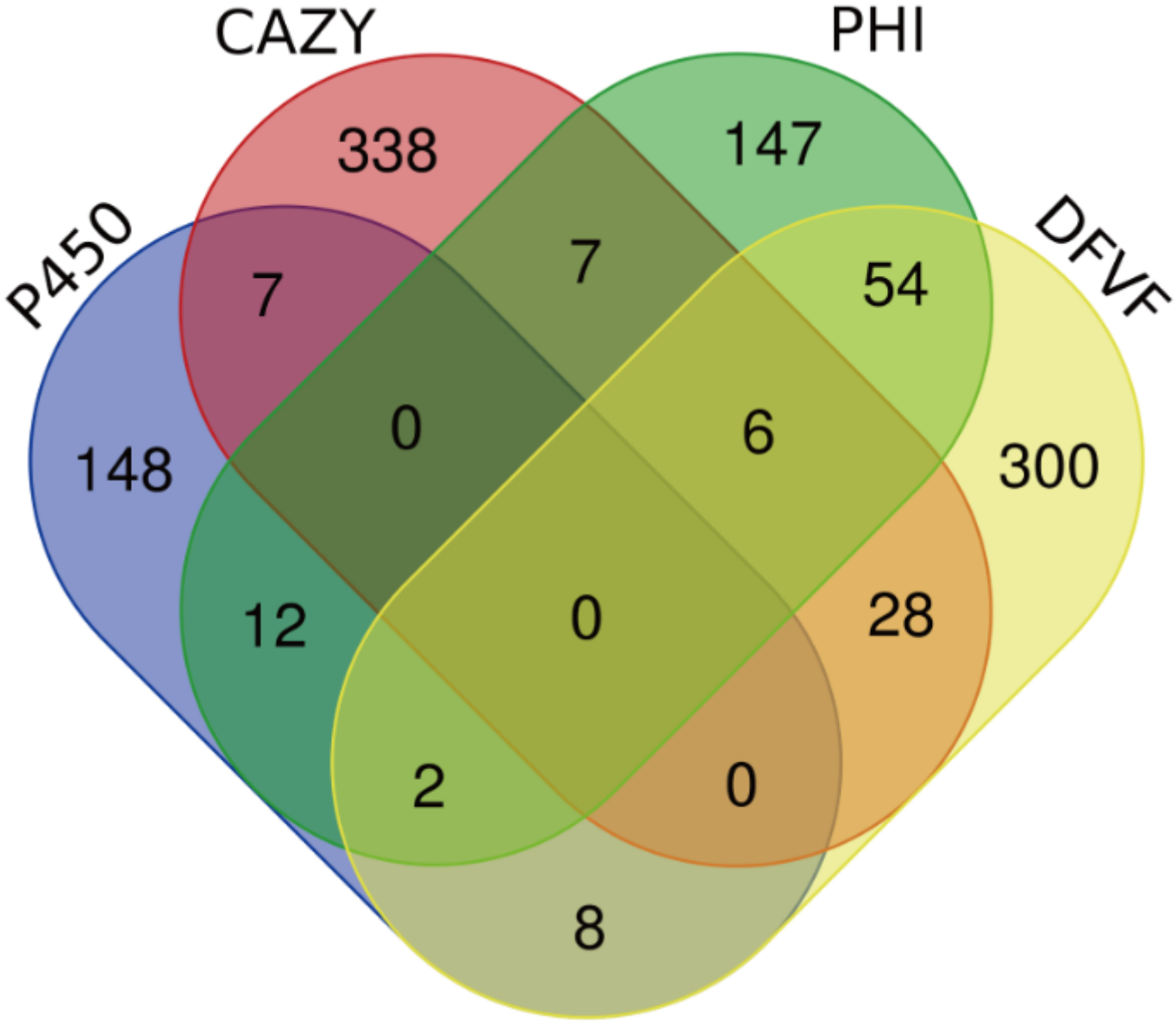
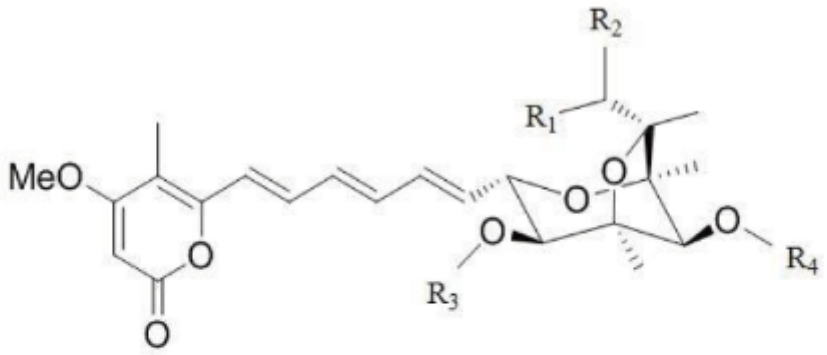


Figure 4

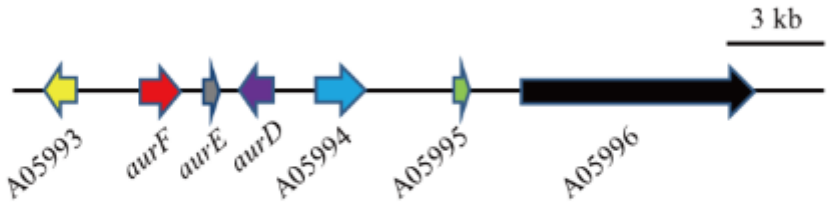
Venn-plot showing the intersections among the secreted PHI proteins (green), secreted DFVF (yellow), secreted CYP450 enzymes (purple), and secreted CAZymes (red).

a



Aurovertin A: R_1=H, R_2=H, R_2=Acetyl, R_4=Acetyl
Aurovertin B: R_1=H, R_2=H, R_2=H, R_4=Acetyl
Aurovertin D: R_1=OH, R_2=H, R_2=H, R_4=acetyl
Aurovertin E: R_1=H, R_2=H, R_2=H, R_4=H
Aurovertin J: R_1=Acetyl, R_2=H, R_2=H, R_4=H
Aurovertin M:R_1=H, R_2=H, R_2=H, R_4= propionyl

b



Gene in *C. arbuscula*	Gene function description	Related gene
A05993	tat pathway signal sequence	*aurG*
A05994	FAD-dependent monooxygenase	*aurC*
A05995	SAM-dependent methyltransferase	*aurB*
A05996	polyketide synthase	*aurA*

Figure 5

Aurovertin biosynthesis in C. arbuscula NRRL 3705 a. Chemical structures of aurovertins. b. Schematic representation of the putative aurovertin gene cluster (cluster 23) and the description of each gene in a gene cluster.

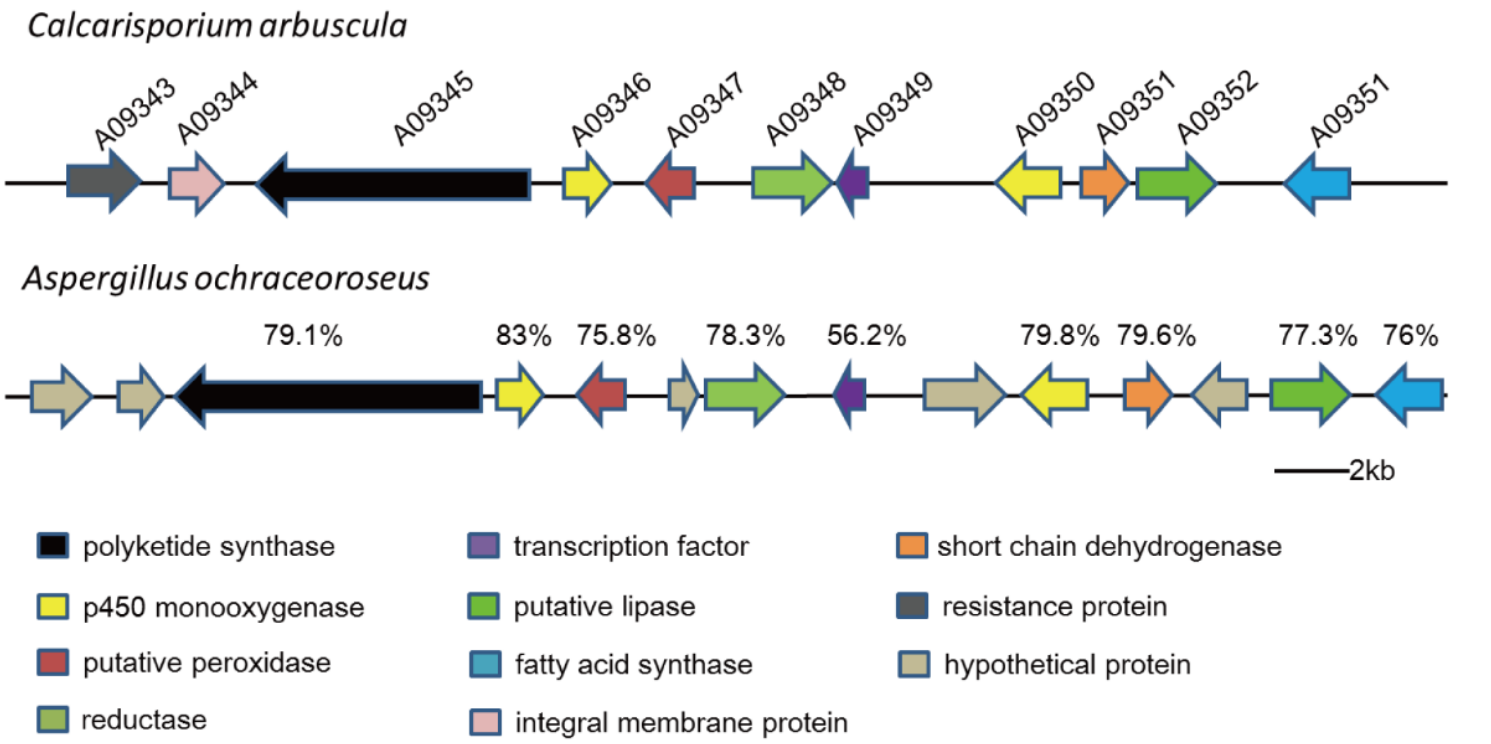


Figure 6

The putative aflatoxin biosynthetic gene cluster of C. arbuscula NRRL 3705 found in this study and the comparison of this cluster with the aflatoxin cluster reported for Aspergillus ochraceoroseus. The identity of each homolog to C. arbuscula NRRL 3705 counterparts is shown.

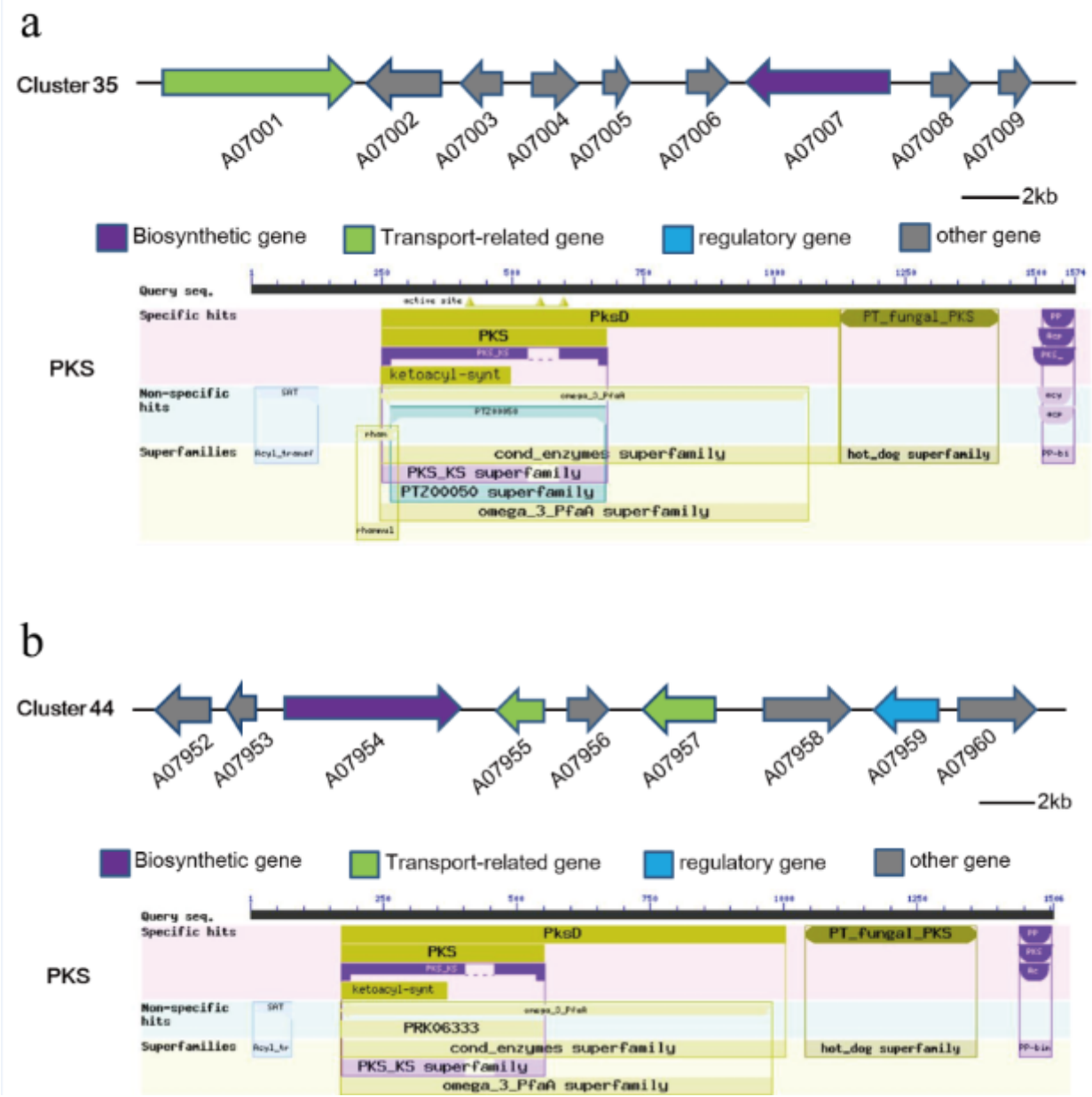


Figure 7

Two AOH biosynthesis-related gene clusters for C. arbuscula NRRL 3705. a Cluster 35, with 7 genes. b Cluster 44, with 7 genes. The PKS and domains were determined by the BLAST searches against the CDD database.

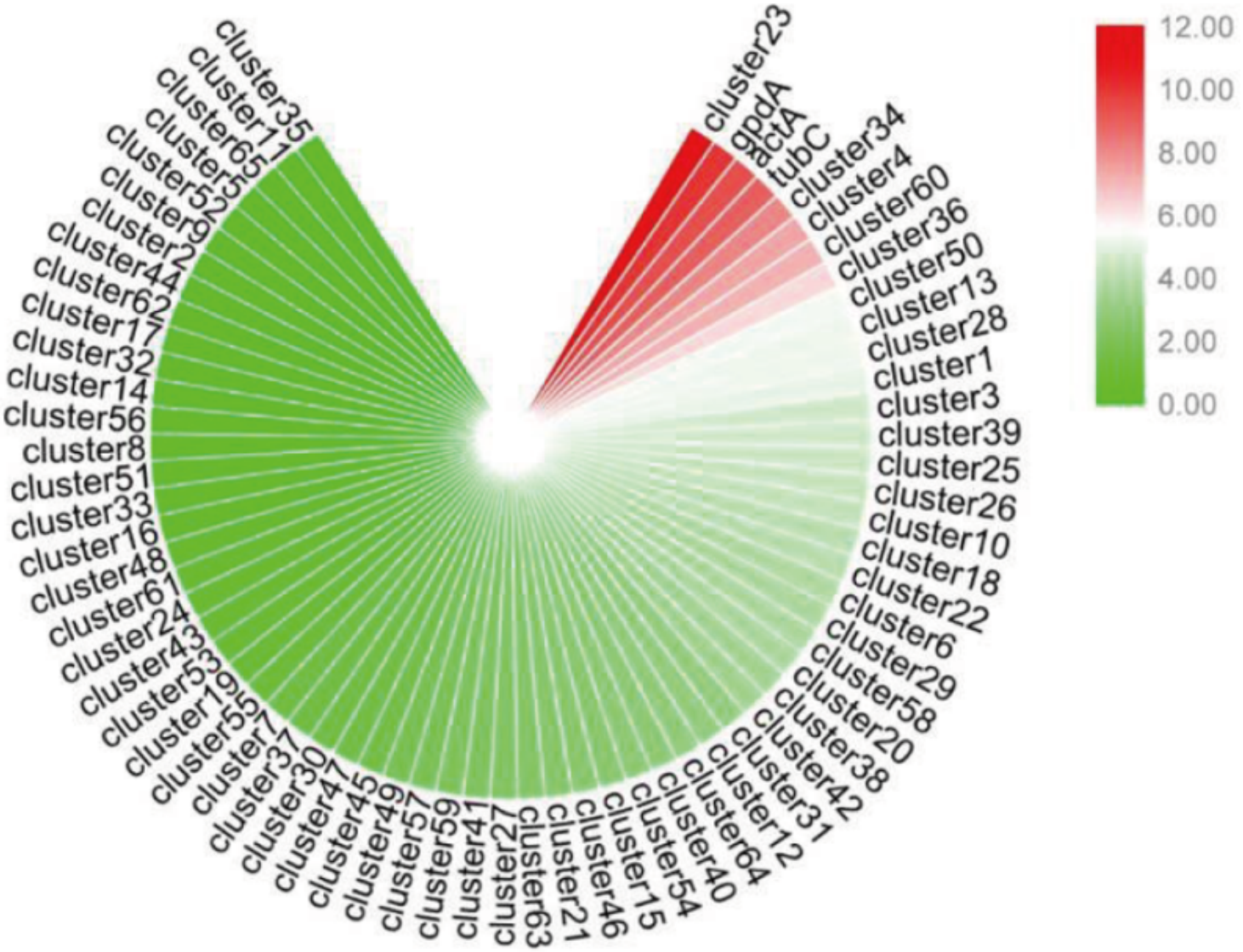


Figure 8

Comparison the backbone gene experssion of every gene cluster in C. arbuscula NRRL 3705. Heat map of gene expression for the every gene cluster in C. arbuscula NRRL 3705. The house-keeping genes actA, tubC and gpdA are the reference genes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile5.xls
- Additionalfile4.xls
- Additionalfile1.pdf
- Additionalfile2.xls
- Additionalfile3.xls