A genetic and biochemical map for the biosynthesis of occidiofungin, an antifungal produced by *Burkholderia contaminans* strain MS14

Ganyu Gu a, Leif Smith b*, Aixin Liu c, and Shi-En Lu a*

a Department of Entomology and Plant Pathology, Mississippi State University, 32 Creelman St., Mississippi State, MS 39762

b Department of Biological Sciences, Texas A&M University, College Station, TX 77843

c Department of Plant Pathology, Shandong Agricultural University, Taian, China, P.R. 271018

*Corresponding authors: Shien Lu, Department of Entomology and Plant Pathology, Mississippi State University, 32 Creelman St., Mississippi State, MS 39762, Phone: (662) 325-3511, Fax: 662 325 8955, E-mail: sl332@msstate.edu; Leif Smith, Department of Biological Sciences, Texas A&M University, College Station, TX 77843, Phone: 979-845-2417, Fax: 979-845-2891, Email: jsmith@bio.tamu.edu
Abstract

A striking feature of *Burkholderia contaminans* strain MS14 is the production of a glyco-lipopeptide, named occidiofungin. Occidiofungin has a broad range of antifungal activities against plant and animal pathogens. In this study, a complete covalent structure characterization and identification of the whole genomic DNA region for the occidiofungin gene (*ocf*) cluster are described. Discovery of the presence of 2,4-diaminobutyric acid, 3-chloro-β-hydroxytyrosine and elucidation of the structure of a novel C18 fatty amino acid residue has been achieved. In addition, seven additional putative open-reading frames (*ocfJ-N* and ORF16) were identified. Transcription of all the putative genes *ocfJ-N* identified in the region except ORF16 was regulated by both *ambR1* and *ambR2*. Elucidation of the structure and the *ocf* gene cluster provides insight into the biosynthesis of occidiofungin and promotes future aims at understanding the biosynthetic machinery. This work provides new avenues for optimizing the production and synthesis of structural analogs of occidiofungin.

Keywords

*Burkholderia contaminans*; antifungal activity; nonribosomal peptide synthetase; polyketide synthetase
Introduction

*Burkholderia contaminans* strain MS14 showing a broad range of antifungal activities against plant and animal pathogens was isolated from a disease-suppressive soil (23). A 45.2-kb genomic DNA fragment harboring 11 open-reading frames (ORFs) including the biosynthetase genes and two LuxR regulatory genes *ambR1* and *ambR2* (13,14) was identified. Analysis of transcription demonstrated that both *ambR1* and *ambR2* are essential for the expression of all the ORFs except for ORF1 and the production of the antifungal activity. Production of a glycolipopeptide, named occidiofungin, is responsible for the antifungal activities of strain MS14 (21).

Initial structural characterization of occidiofungin revealed two variants, Occidiofungin A and Occidiofungin B. Both are composed of eight amino acids differing by an addition of oxygen to occidiofungin B forming a β-hydroxy asparagine. The two structurally related antifungal compounds have a mass of 1119.5 and 1215.5 Da, respectively.

Nonribosomal peptide synthetase (NRPS) and polyketide synthetase (PKS) are large multimodular enzymes, which are involved in natural product synthesis in many microorganisms (5). NRPS, involved in the biosynthesis of oligopeptide, is grouped by active sites termed modules, in which each module is required for catalyzing one single cycle of product length elongation. The order and number of the modules of a NRPS protein are mainly followed by the “collinearity rule” (12). There are three main domains in each module: the adenylation (A) domain, responsible for amino acid recognition; the thiolation (T) domain, which is the carrier of thioesterified amino acid intermediates; and the condensation (C) domain, which catalyzes peptide bond formation between two consecutive amino acids (5). The epimerization (E) domain as one of the modification domains catalyzes the conversion of L-amino acids to their D-isomers (25). Cyclization and release of the peptide product are catalyzed by the C-terminal thioesterase.
(Te) domain (35). Polyketide synthetases are a family of enzymes or enzyme complexes that produce polyketides, a large class of secondary metabolites, in bacteria, fungi, plants, and a few animal lineages (17). The biosynthesis of polyketides shares striking architectural and organizational similarities with nonribosomal peptide biosynthesis, and their modules can be integrated to produce the hybrid NRPS-PKS products. Some pharmaceutical antimicrobial agents, such as the precursor of penicillin and erythromycin, are synthesized through these mechanisms (6,7,31,40).

*Burkholderia* bacteria, which are widely distributed in nature, are reported to be important for plant growth promotion (4). However, taxonomic distinctions have not enabled biological control strains to be clearly distinguished from human pathogenic strains, which has lead to the reassessment of the risk of the registered *Burkholderia* strains as biological control agents (26). For example, some strains of *B. contaminans* were isolated from sputum and blood samples of debilitated patients such as cystic fibrosis patients (42), while some of them have shown significant antifungal activities (13). Analysis of the genetic elements and molecular mechanisms in *Burkholderia* may benefit the development of biologically based management approaches, while potentially providing means to eliminate the deleterious effects of the microorganism. In this paper, we describe a genomic region that contributes to biosynthesis of occidiofungin, as well as the structure of the unique antifungal compound.

Till now, the whole genomic sequences of 22 strains of *Burkholderia* are available in NCBI (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). According to the annotation of the ExPASy Proteomics Server, eight of them are human pathogens including three strains of *B. cenocepacia*, three strains of *B. pseudomallei* and two strains of *B. mallei*. *B. cenocepacia* is the opportunistic pathogen of human cystic fibrosis or chronic granulomatous diseases (24).
*pseudomallei* and *B. mallei* are the causal agents of melioidosis and glanders diseases, respectively (3,15). The other 12 strains were not considered to be human pathogens, such as *B. lata* strain 383 (http://www.expasy.ch/sprot/hamap/BURM1.html).

In this study, the right border of the *ocf* gene cluster was sequenced, and further sequence analysis revealed the presence of another seven additional ORFs, which include *ocfJ*, another NRPS gene, and *ocfN*, a putative thioesterase. Nuclear Magnetic Resonance (NMR), mass spectrometry, and amino analysis data, have helped to resolve regions of structural uncertainty. We have clarified the structure of a novel amino acid, and identified the presence of a chloro-β-hydroxy tyrosine (chloro-BHY) and 2,4-diaminobutyric acid (DABA). Characterization of the covalent structure and elucidation of the gene cluster provide a solid basis for modeling the biosynthesis of the unique antifungal compound, as well as provide avenues for improving the yield of occidiofungin during fermentation.

**MATERIALS AND METHODS**

**NMR spectroscopy.** Occidiofungin was produced and purified as previously described (13). A 3.5 mg sample of occidiofungin was dissolved in 600 µL of dimethyl sulfoxide (DMSO-d6, Cambridge Isotopes). The NMR data were collected on a Bruker Avance DRX spectrometer, equipped with a CryoProbe, operating at a proton frequency of 600 MHz. The ¹H resonances were assigned according to standard methods (45) using COSY, TOCSY and NOESY experiments. ROESY and ¹³C-HSQC experiments were used to clarify some areas of ambiguity in the TOCSY and NOESY spectra. NMR experiments were collected at 25°C. The carrier frequency was centered on the residual water resonance (3.333 ppm), which was suppressed minimally using standard presaturation methods. A 2.0 s relaxation delay was used between
scans. The TOCSY experiment was acquired with a 60 ms mixing time using the Bruker DIPSI-2 spinlock sequence. The NOESY and ROESY experiments were acquired with 400 ms mixing times. The parameters for collecting the HSQC spectrum were optimized to observe aliphatic and aromatic CH groups. The spectral sweep width for the TOCSY, NOESY, and ROESY was 11.35 ppm in both dimensions. The spectral sweep widths for HSQC were 11.35 ppm in the proton dimensions and 100 and 150 ppm for the carbon dimension. All 2D data were collected with 2048 complex points in the acquisition dimension and 256 complex points for the indirect dimensions, except for the HSQC which was collected with 2048 and 128 complex points in the direct and indirect dimension, respectively. Phase sensitive indirect detection for NOESY, ROESY, TOCSY, and COSY experiments was achieved using the standard Bruker pulse sequences. $^1$H chemical shifts were referenced to the residual water peak (3.33 ppm). Data were processed with nmrPipe (8) by first removing the residual water signal by deconvolution, multiplying the data in both dimensions by a squared sinebell function with 45 or 60 degree shifts (for the $^1$H dimension of HSQC), zero-filling once, Fourier transformation, and baseline correction. Data were analyzed with the interactive computer program NMRView (16). The NOE cross-peak intensities were measured in NMRView. Distances were calibrated using the relationship $r_{ab}^6 = r_{cal}^6(V_{cal}/V_{ab})$, where $r_{ab}$ is the distance between atoms a and b, $V_{ab}$ is the NOESY a to b cross-peak volume, $r_{cal}$ is a known distance, and $V_{cal}$ is the corresponding volume of the NOESY calibration cross-peak. The distance used for calibrations was the $\beta$-hydroxy Tyr4 H$^\delta$ and H$^\varepsilon$ aromatic protons (2.46 Å).

Mass Spectrometry. Occidiofungin (10 µg) was evaporated to dryness in a Speed Vac Concentrator (ThermoScientific, San Jose, CA) and the residue was taken up in 50 µl methanol and analyzed by direct infusion at 3 µl/minutes into an LCQ DecaXP (ThermoScientific, San
Jose, CA) operated with a capillary voltage of 20 volts, spray voltage of 3000 volts and a capillary temperature of 150°C. Data were acquired for 3 minutes over a mass range of m/z 200 to 2000.

**Amino Acid Analysis.** Amino acid analysis was performed at the Molecular Structure Facility, UC Davis. This facility uses a "post-ion-exchange column" ninhydrin reaction detection system. Amino acid analysis was also performed at Texas A&M Protein Chemistry Lab using a Hewlett Packard AminoQuant II system. Occidiofungin samples were aliquoted, mixed with Internal Standards [Norvaline (Int1), Sarcosine (Int2), and DABA], dried in glass tubes in a vacuum concentrator and subjected to vapor phase hydrolysis. The samples were subsequently reconstituted in 0.4 N borate buffer to bring the pH to 10 for optimum derivitization. Primary amino acids tagged with o-phthaldialdehyde (OPA) were detected by the Diode Array (UV) detector at 338/390 nm and the fluorometric detector monitored the primaries at excitation/emission 340/450. Pre-column derivatized amino acids were eluted from a narrow bore 5 µm reverse phase column.

**Bacterial strains, plasmids and culture conditions.** The bacterial strains and plasmids used in this study are described in Table 1. *Escherichia coli* strain JM109 was grown in Luria-Bertani medium at 37°C (34). *Burkholderia* strains were cultured at 28°C on nutrient broth–yeast (NBY) extract agar medium (43). When required, antibiotics were added at the following concentrations: ampicillin (100 µg ml⁻¹), trimethoprim (50 µg ml⁻¹), chloramphenicol (12.5 µg ml⁻¹), and kanamycin (100 µg ml⁻¹ for *E. coli* and 300 µg ml⁻¹ for the MS14 mutants).

**Random mutagenesis, genomic library screening and sequencing.** To identify more genes associated with production of antifungal compound of strain MS14, the mutant MS14MT24 (Table 1) was generated and plasmid pSL604 was obtained using an EZ-Tn5™
<R6Kory/KAN-2>Tnp Transposome™ Kit (Epicentre Biotechnologies, Madison, WI) as described previously (13). Fosmid 4G5 was identified from the MS14 genomic library using the 1-kb PCR product from pSL604 as described previously (13). The DNA insert of fosmid 4G5 was sequenced using a random shotgun approach (37). At least triple coverage of sequencing reactions was achieved, and the Lasergene software package (DNASTAR, Inc., Madison, WI) was used for generation of consensus sequence. Open-reading frames (ORFs) and genes were subsequently predicted by the Softberry FGENESB program (Softberry, Inc., Mount Kisco, NY) and the identified ORFs and genes were analyzed using Blastx in the NCBI database. Putative promoter sequences were identified by the Softberry BPROM program. The InterProScan program was used for prediction of functional domains of proteins (32). NRPSpredictor (33) and the NRPS-PKS web-based software (2) were used for specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPSs). The program “IslandPick” was used to analyze the sequence characteristics as genomic islands (18).

Site-directed mutagenesis and complementation of the ocfJ gene. The wild-type ocfJ gene was disrupted by the insertion of a kanamycin cassette into its open reading frame as described previously (22). Plasmids pBR325 and pBSL15 were digested by EcoRI, blunted by T4 DNA polymerase, and self ligated to generate plasmids pBR325[R1] and pBSL15[R1] (1). To mutate ocfJ, the 4-kb fragment obtained by PCR using primers MoccEF and MoccER (Supplementary Table 1), was cloned into the pGEM-T Easy Vector System I (Promega Corporation, Madison, WI) resulting in plasmids pGG20. The nptII gene was inserted into pGG20 at BamHI, generating plasmids pGG21. The 5-kb EcoRI fragment of pGG21 harboring the ocfJ gene disrupted by insertion of nptII was cloned into pBR325 (30) at the EcoRI site to generate pGG22. Mutagenesis of the ocfJ gene was conducted via a marker exchange procedure.
as described previously (22), to generate the mutant MS14GG78 (Table 1). PCR analysis and sequencing were used to verify the double crossover mutants.

To obtain the intact wild type gene *ocfJ*, which was disrupted in the mutant MS14GG78, a 4.5-kb fragment was amplified by PCR using primers EoccEF2 and EoccER2 (Supplementary Table 1), both of which contain the HindIII site (Table S1), and cloned into the pGEM-T Easy Vector, resulting in the plasmid pGG23. The presence of the intact *ocfJ* gene in the 4.5-kb DNA fragment was verified by sequencing. The 4.5-kb HindIII fragment harboring the intact *ocfJ* gene was inserted into the *Burkholderia* gene expression vector pMLS7 (20), to generate the plasmid pGG24. The plasmid pGG24 were electroporated into cells of the mutant MS14GG78, an *ocfJ* mutant (Table 1). Colonies acquiring the trimethoprim resistance were confirmed to contain the plasmid construct pGG24 by plasmid extraction and restriction enzyme digestion.

Complementation experiments were conducted using the plate assays to evaluate antifungal activity against *G. candidum* as described previously (13).

**RNA extraction and quantitative real-time PCR.** Total RNA of strain MS14 and its mutants was extracted using an RNeasy Protect Bacteria kit (Qiagen, Valencia, CA) as recommended by the manufacturer. Transcription of the ORFs and genes was analyzed using quantitative real-time PCR (Q-PCR) as described previously (12,13). Primers for Q-PCR were provided in supplemental data (Supplementary Table 1). Three replicates of Q-PCR were conducted independently, and statistical significant differences were determined for RQ values by analysis of variance (*P* < 0.05) followed by the Bonferroni and Dunnet post hoc multiple comparisons (ANOVA, SAS Institute, Inc., NC, USA).

**RESULTS AND DISCUSSION**
Occidiofungin contains non-proteinogenic amino acids. The covalent structure of occidiofungin was determined to contain a β-hydroxy modification of asparagine (BHN), β-hydroxy modification of tyrosine (BHY) and a 3-chloro-addition to β-hydroxytyrosine (chloro-BHY), as well as a novel amino acid (NAA) derived from fatty acid synthesis (Figure 1). Expansion of the TOCSY 2D NMR spectra showing the amide to alpha, and amide to side chain spin systems for each assigned residue in occidiofungin is shown in Figure 2. Chemical shifts values are provided in supplemental data (Supplementary Table 2). Observed inter-residue NOEs are summarized in Supplementary Figure 1 and an expansion of the NOESY spectra showing inter-residue NOEs are shown in the supplemental data (Supplementary Figure 2 and 3).

Initial structural characterization was done using an aqueous solvent (50% acetonitrile-d3(ACN):50% Water) and limited NMR experiments that provided a preliminary understanding of the antifungal compounds structure (21). In these analyses, the antifungal peptide was determined to be an eight residue cyclic glyco-lipopeptide via amide to alpha and amide to amide proton sequential walk along the backbone atoms. Furthermore, the compound was determined to contain a xylose attached to a novel amino acid. Further analyses, using a non-aqueous solvent along with additional NMR experiments provided means to refine the structural predictions. In this study, we have determined that position 2 is a C18 glyco amino acid, position 3 is a serine, position 4 is a mixture of BHY and chloro-BHY, position 5 is a 2,4-diaminobutyric acid (DABA).

To help elucidate the structure of the NAA2, COSY NMR spectra were collected to sequentially assign proton couplings within this residue. An almost complete sequential walk for the proton couplings could be observed in the COSY data set for NAA2 (Supplementary Figure 4), providing evidence that the novel amino acid is a C18 fatty acid residue, similar to what has
been reported for cepacidine (19). The only break in the proton coupling occurs between C8 and C9, which is visible in the TOCSY data set. The attachment site for the xylose sugar was determined to be at the C7 position (Supplementary Figure 5 and Figure 3). C5, C6, and C7 all have chemical shift values supporting the presence of an oxygen in their vicinity and could support the attachment of the xylose. However, C5 and C6 have a correlation to a hydroxyl proton, while C7 lacks a correlation to a proton on a hydroxyl, thus, supporting an ether linkage to the sugar. In addition, NOEs are observed between the proton on C7 and C8 of NAA2 to the proton on C1 of the xylose sugar.

The chemical shift values that were found along the amide frequency for the amino acid at position 3 could only account for a few amino acids. Many of which could be ruled out based on their absence in the amino acid analysis (AAA) data. AAA revealed the presence of asparagine, glycine, serine and lysine in the molecule (Supplementary Figure 6). The beta proton chemical shift values (~3.27 ppm) for the residue at position 3 are out of the normal range for a serine (~3.88 ppm). The proximity of the tyrosine ring presumably shields the beta proton on Ser3 causing them to have a lower than predicted chemical shift (~3.88 ppm), which accounts for the ser β-proton chemical shifts of 3.27 ppm (Supplementary Table 2). Further examination of the NMR data sets revealed coupling between the beta protons to the hydroxyl proton of Ser3 could be seen (Supplementary Figure 7), further supporting the assignment of a serine at this position.

The presence of indole-like resonances in the NMR TOCSY and NOESY data sets were identified. A peak at 10.03 ppm (Supplementary Figure 8), which corresponds to the frequency of an indole epsilon proton, was observed. In addition, the characteristic indole proton frequencies were also observed in the aromatic region of the TOCSY spectra. These resonances
were characteristic of a tryptophan residue. Additional NMR data was collected to determine the
location and structure of the indole-like residue in the compound. An aromatic HSQC data
revealed the presence of three additional proton carbon couplings, which is one less than
expected for an indole ring (Figure 4A). Mass spectrometry data helped to characterize the
nature of this residue. ESI mass spectrometry data revealed the existence of four structural
variants of the antifungal peptide; having a mass \([M + 1 \ (H)]\) of 1,200.39 Da, 1216.41 Da,
1234.17 Da, and the other having a mass of 1250.41 Da which corresponds to the addition of
oxygen and/or chlorine to the first compound (Figure 4B). These masses are in accordance with
the elucidated structure (calculated monoisotopic masses of 1,199.592, 1215.587, 1233.553, and
1,249.548) shown in Figure 1. Natural abundance of chlorine isotope 35 and 37 exists in a 3:1
ratio, which helps to confirm chlorination. As shown in Figure 4B, labeled with “abc”, the
relative abundance for each isotope of a non-halogenated compound will have a typical stair step
pattern for \(M + 1\), \(M + 2\), and \(M + 3\), while the chlorinated compound, labeled with “xyz”, will
have an increase in the \(M + 3\) isotope. The increase in the \(M + 3\) isotope is due to the addition
of \(^{37}\)Cl. The identification of a chlorination taking place in the molecule helped clarify the nature
of the other aromatic residue observed in the NMR data. The three additional proton carbon
couplings observed in the aromatic HSQC data are actually resonances for \(C_2\), \(C_5\), and \(C_6\) of the
BHY ring and the chlorination of the tyrosine ring occurs at the \(C_3\) position. The addition of the
chlorine to the \(C_3\) ring position is responsible for the observed downfield shift of the hydroxyl
proton of Tyr to 10.03 ppm and is responsible for the downfield shift of the other aromatic
protons. The presence of three distinct proton resonances on the aromatic ring for chloro-BHY4,
instead of two proton resonances for BHY4, arises from the chlorination at the \(C_3\) position
causing the environments for the protons at position \(C_2\), \(C_5\), and \(C_6\) to be chemically distinct.
The presence of these four structural variances is similar to what has been reported for the non-ribosomally synthesized antibiotic balhimycin and vancomycin (6,44). Furthermore, the presence of 4 structural variants of occidiofungin explains the multiple amide proton chemical shift values observed for each amino acid residue (Figure 2).

Further analysis of the COSY data set revealed that the proton couplings of the predicted Lysine at position 5 (prediction based on amino acid analysis data) did not match this assignment. Alternatively, the residue appeared to be a 2,4-diaminobutyric acid. This was unexpected given the identification of Lys by amino acid analysis (Supplementary Figure 6A). Possibly as was previously done, lysine and 2,4-diaminobutyric acid may have the same retention time when ion-exchange chromatography is used to separate amino acids using a "post-column" ninhydrin reaction detection system. Therefore, an alternative method was used along with a DABA standard. Following acid hydrolysis, HPLC separation followed by ultraviolet detection (UV) and fluorescent detection (FLD) was used to see if we could distinguish a DABA from a lysine. Using this method, we did not see a lysine, but we did observe concomitant increase in glycine volume ratio relative to serine (Supplementary Figure 6B). An internally loaded standard of DABA (5 nm) was shown to co-eluted with glycine using this procedure (Supplementary Figure 6C). The amino acid analysis data is in agreement with the NMR data supporting the assignment of a DABA at position 5.

As described above, an extensive amount of new data has been analyzed providing a definitive assignment of the structure of occidiofungin. The assignment of the fatty glyco amino acid at position 2, serine at position 3, presence of a β-hydroxy tyrosine and a 3-chloro-β-hydroxy tyrosine at position 4, and 2,4-diaminobutyric acid at position 5 satisfy all chemical
shifts present in the NMR data sets, ESI mass spectrometry data, and the amino acid analysis

data. Furthermore, our structure is in agreement with the genomic data in the ocf gene cluster.

The ocf gene cluster. Seven new ORFs (ocfl-N and ORF16) were identified (Figure 5; Table 2) upstream of the 45.2-kb genomic fragment that was previously reported (14). This region was sequenced through the sequencing of the fosmid 4G5. A 58.1-kb genomic DNA fragment, named as the ocf gene cluster, was obtained and deposited into GenBank with the accession number: EU938698.

Sequence analysis revealed that the ocf gene cluster, excluding ORF1 and ORF16, shares high similarity (99% nucleotide coverage and 91% identity) to an uncharacterized DNA region of B. ambifaria strain AMMD, which was used for biological control of plant diseases (27). The gene cluster was not found in other Burkholderia strains or any other strains available in GenBank (nucleotide coverage < 24%), including some identified opportunistic human pathogens B. cenocepacia strains J2315, AU1054, HI2424, B. pseudomallei strains 668, 1710b, 1106a, and B. mallei strains NCTC 10247, NCTC 10229. The sequences of the flanking regions of the ocf gene cluster in strain MS14 share highest identities (89% nucleotide identity) with the homologs C7522 (left) and C7511 (right) in chromosome 3 of B. lata 383. The average G+C content of the ocf gene cluster is 67.74%, which is closer to those of flanking region sequences in strain AMMD (left: bam_6475, 68.5% and right: bam_6483, 65.89%) but different from those of strain MS14 (55.16% left and 61.81% right). Besides, sequence analysis using the program “IslandPick” indicated that the uncharacterized exclusive homolog of the ocf gene cluster in strain AMMD possesses the characteristics of a genomic island. These data suggest that the ocf gene cluster may be horizontally transferred from a strain similar to B. ambifaria AMMD and had integrated into a strain similar to B. lata 383. More importantly, the absence of the gene
cluster in clinical strains of *Burkholderia* suggests that the gene cluster is not required for potential human pathogenesis, and presumably is not involved in pathogenesis. This finding has provided insights for the usage of the gene cluster for engineering bio-control agents for agricultural use. Furthermore, a detailed understanding of the gene cluster provides avenues for improving the production of occidiofungin, thus, promoting studies aimed at understanding the compounds therapeutic potential.

Organization and orientations of the genes *ocfI*-N are the same as those of the *B. ambifaria* strain AMMD genome with an averaged amino acid identity of 91%, and ORF16 has the highest similarity (93%) to its homolog C7511 of *B. lata* strain 383 (Figure 5; Table 2), while its homolog was not found in the genome of the strain AMMD with BLAST. The putative protein (1107 a.a.) from *ocfI* was predicted to encode a hybrid monooxygenase-PKS. The gene *ocfJ* was predicted to encode a NRPS (1475 a.a.). The genes *ocfK*, *ocfL*, and *ocfM* were predicted to encode oxygenase or halogenase, transaminase, and an epimerase or dehydratase respectively (Table 2). Putative protein sequence of *ocfN* is 219 amino acids in length and carries a Te domain. Putative promoters were identified upstream of ORF1, *ambR1*, *ocfL* and ORF16 and their locations are shown in Figure 5.

**Disruption of the NRPS gene *ocfJ* eliminated the antifungal activity of strain MS14.**

A nonpolar mutation was constructed by insertion of an *nptII* cassette into BamHI of *ocfJ*, and the mutant MS14GG78 was generated by marker exchange mutagenesis. PCR analysis demonstrated that *ocfJ* was disrupted by insertion of *nptII* in the genome of the mutant. The mutant exhibited negligible antifungal activity towards *G. candidum* (inhibitory zone radius ± SEM: 0.33 ± 0.33 mm), which is similar to the mutants MS14MT13, MS14MT18 (13), MS14MT16 (14), and MS14MT24 (Supplementary Figure 9). As expected, the wild-type strain
MS14 showed a strong inhibition to the fungus (13.00 ± 0.58 mm). The wild-type level of antifungal activities against *G. candidum* was observed for MS14GG78 complemented in trans with plasmid pGG24 (12.67 ± 0.33 mm). As expected, the presence of the empty vector pMLS7 had no effect on antifungal activities of either the wild type strain MS14 or the mutant MS14GG78 (data not shown). These results further confirmed the essential role of the *ocf* gene cluster in the production of antifungal activity of strain MS14.

**Transcription of the genes ocfI-N except ORF16 was regulated by both *ambR1* and *ambR2*, the LuxR-type regulatory genes.** To test the relationship between the newly identified genes or ORFs with the LuxR regulatory genes *ambR1* and *ambR2*, Q-PCR analyses were performed with RNAs isolated from strain MS14 and the mutants MS14GG44 and MS14MT15. Previous studies demonstrated that both *ambR1* and *ambR2* positively regulated transcription of *ocfa-H* (14). As expected, significant differences (P < 0.05) of expression levels of *ocfI-N* in the mutants MS14GG44 and MS14MT15 were observed as compared with those in the wild-type strain MS14. Transcript levels of the genes in mutants MS14GG44 and MS14MT15, mutants in LuxR-type regulatory genes, were reduced 13.2 and 12.0 fold, respectively (Figure 6). These data suggest that transcription of the gene *ocfI-N* is promoted by both the *ambR1* and *ambR2* genes.

Transcription of ORF16 was not significantly affected by the mutations in either *ambR1* or *ambR2*, indicating that the ORF16 is not regulated by either of them. In addition, the ORF16 gene, which codes for a hypothetically conserved protein, has highest identity (93%) to the protein encoded by ORF C7511 of *B. lata* strain 383; however, no significant homolog was found in the genome of strain AMMD, which has an uncharacterized genomic region sharing the highest similarity to the *ocf* gene cluster (*ambR1 - ocfN*). These data imply that ORF16, similar
to ORF1 (14), may not be part of the gene cluster required for the antifungal activity of strain MS14. Therefore, *ocfN* is most likely the last gene at the right border of the gene cluster, while *ambR1* is the first gene at the left border of the gene cluster. These results indicate that the whole length of the *ocf* gene cluster responsible for the production of occidiofungin has been identified (Figure 5).

In summary, the whole length of the *ocf* gene cluster has been characterized, which is composed of 16 ORFs. Among the 16 members of this cluster, *ocfD*, *ocfE*, *ocfF*, *ocfH* and *ocfJ* were predicted to encode NRPS or NRPS-PKS, and are directly related to the biosynthesis of the antifungal compound occidiofungin. The genes *ocfC*, *ocfK*, *ocfL*, *ocfM*, and *ocfN* were predicted to be involved in the modification of occidiofungin or its components. AmbR1 and AmbR2 are the regulators controlling the transcription and expression of occidiofungin. The predicted functions of the *ocf* gene cluster are basically in agreement with the covalent structure.

**Occidiofungin Biosynthesis.** Completion of the sequencing of the *ocf* gene cluster and the completion of the detailed structural characterization of occidiofungin provides an excellent opportunity to model the biosynthesis of occidiofungin (Figure 7). Gene products required for the synthesis of the novel amino acid are all located in the *ocf* gene cluster (Figure 5). The order in which the modules function, so that they are in agreement with the structure are as follows: OcfJ-OcfI-OcfH-OcfF-OcfE-OcfD. Through NRPS-PKS web-based software and interProScan software in EMBL-EBI, the predicted domains in these modules are A-T-KS (ketosynthase; OcfJ) FDM (flavin-dependent monooxygenase)-KS-T (OcfI) KS-AT (acyl transferase)-KR (ketoreductase)-T-KS-AT-T-AmT (aminotransferase)-C-A-T-C (OcfH) A-T-E (OcfF) C-A-T-E-C-A-T-C (OcfE) A-T-C-A-T-E-C-T-Te (OcfD). These predictions are in agreement with the structural analyses of occidiofungin.
The *ocfJ* gene encodes a hybrid NRPS-PKS peptide containing domains characteristic of both NRPSs and PKSs. As described above, the potential domains are A-T-KS. There is no AT domain to accompany the KS domain in ORF11. Therefore, it is possible OcfJ functions solely as a NRPS. This NRPS domain is presumably important for the addition of Asn1/BHN1 to NAA2. Evidence for this assumption comes from the lack of a C domain, which can occur in the first module of NRPS (9). A condensation module is present in OcfH, which can join Asn1/BHN1 to NAA2.

Based on interProScan and NRPS-PKS web-based software predictions described above, the *ocfH* gene has the following domains KS-KR-T-AT-T-AmT-C-A-T-C. OcfH has the characteristics of a hybrid PKS-NRPS and is presumably responsible for the partial synthesis of NAA2. The ketoreductase domain is likely to be involved in the oxidation of the carbonyl ketones on the fatty amino acid forming alcohols, while the aminotransferase domain is likely to be involved in the transfers of an amino group to C3 of the fatty acid forming a fatty amino acid. *ocfC*, *ocfl*, and *ocfM* are also predicted to be involved in the synthesis of NAA2. C5, C6, and C7 all carry a hydroxyl group. The presence of a hydroxyl group on C5 and C7 could arise from the reduction of the ketone by the ketoreductase present in OcfH during fatty acid synthesis. The hydroxyl group on C6 requires an independent hydroxylation. OcfI has a flavin-dependent monooxygenase domain which may be responsible for the addition of the hydroxyl group to C6 (28). OcfC and OcfM are presumably important for the addition of xylose to C7. OcfM codes for a homolog of a NAD-dependent epimerase/dehydratase called UDP-D-glucuronate 4-epimerase (39). This enzyme is involved in converting UDP-glucuronate to UDP-xylose. The *ocfC* gene codes for the putative glycosyl transferase and is predicted to catalyze the transfer of a xylose to C7 forming the glyco-amino acid (Figure 3 and Supplementary Figure 5). There was
no predicted dehydratase domain, which is needed for the subsequent removal of the hydroxyl
groups from the polyketide chain [from C8 to C18 carbons] forming the fatty acid chain
observed in the NMR data. Presumably, this activity is being harnessed from fatty acid
biosynthesis or possibly OcfB, a hypothetical protein. However, gene knockout and
complementation experiments should provide confirmation to the function of this gene product.

Using the NRPS-PKS web-based software, the C-terminal part of OcfH is predicted to
catalyze the addition of Ser having an identity of 54% and similarity of 67% to the other NRPS
domains that catalyze addition of Ser. Therefore, it is predicted that the NRPS adds a serine to
the Asn-NAA2 product. Subsequently OcfF is predicted to be responsible for the addition of the
BHY to the Asn/BHN1-NAA2-Ser3 product. The NRPS-PKS web-based software predicts OcfF
to catalyze the addition of Tyr with a 38% identity and 55% similarity to those responsible for
Tyr addition.

There were no confirmed modular predictions for OcfE using the NRPS-PKS web-based
software. The two modules are predicted to catalyze the addition of DABA5 to the Asn/BHN1-
NAA2-Ser3-BHY4 product, and subsequently add Gly6 to the the Asn/BHN1-NAA2-Ser3-
BHY4-DABA5 product. The NRPS-PKS web-based software predicts that OcfD catalyzes the
addition of Asn and Ser. Therefore, the first module would add an Asn to the Asn1/BHN1-
NAA2-Ser3-BHY4-DABA5-Gly6 product and the second module would add a ser to the
Asn/BHN1-NAA2-Ser3-BHY4-DABA5-Gly6-Asn7 product.

The gene ocfL, predicted to code an α-ketoglutarate 4-aminotransferase, is presumably
responsible for the transfers of an amino group during the biosynthesis of DABA (41).
Substrates are predicted to be L-glutamate and L-aspartate and the resulting products would be
L-2,4-diaminobutyric acid and L-α-ketoglutarate. The ocfG gene encodes a homolog of an
enzyme known to be involved in the formation of beta-hydroxy-tyrosine, StaM of the A47934 cluster (29). This enzyme oxidizes the beta carbon while Trp is tethered to the T domain. It is likely that the enzyme is responsible for the beta-hydroxylations of Try4 and Asn1. Given the presence of Asn and BHN at position 1, it is likely that the hydroxylation of Asn on the T domain of module 1 is inefficient. Based on the structural predictions presented above, three of the eight modules in the ocf gene cluster harbor an internal E-domain for the conversion of L to D isomers. Therefore, it is likely that BHY4, DABA5, and Ser8 residues in occidiofungin are D form.

The C-terminal region of OcfD, is predicted to have a Te domain, which is presumably important for the condensation reaction of Ser8 to Asn/BHN1, which terminates synthesis and forms the cyclic peptide. The ocfN gene also codes for a thioesterase. Given that the N-terminal end of the linear peptide is an Asn or BHN. It is possible that two thioesterases are required to form the Asn and BHN structural variants of the cyclic peptide.

The gene ocfA is predicted to be an ATP-binding cassette (ABC) transporter. This large superfamily of integral membrane proteins are known to carry various substrates across cellular membranes. Given that it’s regulated by ambR1 and ambR2, this is presumably the transporter for effective secretion of the antifungal compound. The ocfK gene codes for a homolog of nonheme iron dioxygenases that can catalyze oxidations, as well as chlorinations (29). Potentially this enzyme catalyzes the chlorination that is observed on BHY. Another possibility is that the halogenase activity is attributed to another gene outside of the ocf gene cluster. In Figure 7, the chlorination event is arbitrarily shown to occur at the end. Previous studies have shown that chlorination takes place sometime during peptide synthesis (31).
The genetic information of the \textit{ocf} gene cluster along with the covalent structure has provided a solid foundation to engineer strains that will enhance the production of occidiofungin and to engineer new chemical variants with increased antifungal activity, while minimizing possible toxicity to plants and animals. For example, overexpression of \textit{ocfA} and/or \textit{ambR1} (regulator) may increase occidiofungin production by strain MS14. The study of hybrid NRPS-PK natural products, which further expands the perspective of combinatorial biosynthesis, is attracting more attention (11). Some hybrid products with medicinal importance have been reported, such as rapamycin, bleomycin and leinamycin (10,36,38). Understanding the genetic structures for the production of occidiofungin, especially the biosynthesis of the novel amino acid will benefit the optimization and production of novel products.

\section*{ACKNOWLEDGEMENT}

We are grateful to Jinny Johnson and Larry Dangott at Texas A&M Protein Core for their proteomic assistance. We are also grateful to Kuan-Chih Chen and Sonya M. Baird for laboratory assistance. Lastly we would like to think Ron Shin and Rama Krishna at the UAB High-Field NMR Facility for NMR data collection. This research was supported by the Special Research Initiative Program, Mississippi Agricultural and Forestry Experiment Station (to S.-E. Lu), and funds from Texas A&M University. This work was approved for publication as journal article No. J-11999 of the Mississippi Agricultural and Forestry Experiment Station (SEL).
REFERENCES


<table>
<thead>
<tr>
<th>Strains or plasmids</th>
<th>Relevant characteristics</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Escherichia coli</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JM109</td>
<td><em>recA1</em>, <em>endA1</em>, <em>gyrA96</em>, <em>phi</em>, <em>bsdR17</em>, <em>supE44</em>, <em>relA1</em>, Δ(<em>lac-proAB</em>)F' [traD36, proAB+,<em>lacIq</em>, <em>lacZAM15]</em></td>
<td>Promega</td>
</tr>
<tr>
<td><strong>Burkholderia contaminans</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS14</td>
<td>Wild type strain</td>
<td>(13)</td>
</tr>
<tr>
<td>MS14MT15</td>
<td><em>ambR2</em>:Tn5 derivative of MS14; Km'</td>
<td>(13)</td>
</tr>
<tr>
<td>MS14GG44</td>
<td><em>ambR1</em>:nptII derivative of MS14; Km'</td>
<td>(14)</td>
</tr>
<tr>
<td>MS14MT24</td>
<td><em>ocfH</em>:Tn5 derivative of MS14; Km'</td>
<td>This study</td>
</tr>
<tr>
<td>MS14GG78</td>
<td><em>ocfJ</em>:nptII derivative of MS14; Km'</td>
<td>This study</td>
</tr>
<tr>
<td><strong>Plasmid /fosmid</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pSL604</td>
<td>EZ-Tn5 carrying the ~1 kb genomic DNA of MS14MT24; Km'</td>
<td>This study</td>
</tr>
<tr>
<td>pBluescript II SK</td>
<td>Cloning vector; Ap'</td>
<td>Strategene</td>
</tr>
<tr>
<td>pBR325</td>
<td>Cloning vector; Cm', Tc', Ap'</td>
<td>(30)</td>
</tr>
<tr>
<td>pMLS7</td>
<td>Expression vector of <em>Burkholderia</em>; Tp'</td>
<td>(20)</td>
</tr>
<tr>
<td>pGEM-T Easy</td>
<td>Cloning vector; Ap'</td>
<td>Promega</td>
</tr>
<tr>
<td>pBSL15</td>
<td>Kanamycin resistance gene cassette; Km'</td>
<td>(1)</td>
</tr>
<tr>
<td>pGG20</td>
<td>pGEM-T Easy carrying 4-kb PCR product containing the <em>ocfJ</em> gene; Ap'</td>
<td>This study</td>
</tr>
<tr>
<td>pGG21</td>
<td>pGG20 with the nptII insertion in <em>ocfJ</em> at the BamHI site; Km', Ap'</td>
<td>This study</td>
</tr>
<tr>
<td>pGG22</td>
<td>pBR325 carrying the 5-kb EcoRI fragment containing <em>ocfJ</em> gene disrupted by nptII; Cm', Tc', Km'</td>
<td>This study</td>
</tr>
<tr>
<td>pGG23</td>
<td>pGEM-T Easy carrying 4.5-kb PCR product containing the intact <em>ocfJ</em> gene; Ap'</td>
<td>This study</td>
</tr>
<tr>
<td>pGG24</td>
<td>pMLS7 carrying 4.5-kb HindIII fragment harboring the intact <em>ocfJ</em> gene; Tp'</td>
<td>This study</td>
</tr>
<tr>
<td>4G5</td>
<td>pCC1FOS carrying 40-kb genomic DNA of MS14; Cm'</td>
<td>This study</td>
</tr>
</tbody>
</table>

*Km*: Kanamycin resistance; *Ap*: Ampicillin resistance; *Tp*: Trimethoprim resistance; *Cm*: Chloramphenicol resistance; *Tc*, Tetracycline resistance

*Table 1. Bacterial strains and plasmids*
### Table 2  The putative genes identified

<table>
<thead>
<tr>
<th>Gene or ORF</th>
<th>Size (bp)</th>
<th>Homologue</th>
<th>Identity (%)</th>
<th>Predicted Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORF1(partial)</td>
<td>1175</td>
<td>bamb_6465</td>
<td>93</td>
<td>FAD linked oxidase domain protein</td>
</tr>
<tr>
<td>ambR1</td>
<td>822</td>
<td>bamb_6466</td>
<td>89</td>
<td>LuxR-type regulator</td>
</tr>
<tr>
<td>ambR2</td>
<td>891</td>
<td>bamb_6468</td>
<td>77</td>
<td>LuxR-type regulator</td>
</tr>
<tr>
<td>ocfA</td>
<td>1704</td>
<td>bamb_6469</td>
<td>90</td>
<td>Cyclic peptide transporter</td>
</tr>
<tr>
<td>ocfB</td>
<td>483</td>
<td>bamb_6470</td>
<td>82</td>
<td>Hypothetic protein</td>
</tr>
<tr>
<td>ocfC</td>
<td>657</td>
<td>bamb_6471</td>
<td>94</td>
<td>Glycosyl transferase</td>
</tr>
<tr>
<td>ocfD</td>
<td>9495</td>
<td>bamb_6472</td>
<td>88</td>
<td>NRPS</td>
</tr>
<tr>
<td>ocfE</td>
<td>9066</td>
<td>bamb_6473</td>
<td>89</td>
<td>NRPS</td>
</tr>
<tr>
<td>ocfF</td>
<td>3921</td>
<td>bamb_6474</td>
<td>90</td>
<td>NRPS</td>
</tr>
<tr>
<td>ocfG</td>
<td>1617</td>
<td>bamb_6475</td>
<td>93</td>
<td>Hydroxylase</td>
</tr>
<tr>
<td>ocfH</td>
<td>13410</td>
<td>bamb_6476</td>
<td>91</td>
<td>Hybrid NRPS-PKS</td>
</tr>
<tr>
<td>ocfI</td>
<td>3324</td>
<td>bamb_6477</td>
<td>92</td>
<td>Flavin-dependent monooxygenase</td>
</tr>
<tr>
<td>ocfJ</td>
<td>4428</td>
<td>bamb_6478</td>
<td>91</td>
<td>NRPS</td>
</tr>
<tr>
<td>ocfK</td>
<td>987</td>
<td>bamb_6479</td>
<td>91</td>
<td>Halogenase</td>
</tr>
<tr>
<td>ocfL</td>
<td>1371</td>
<td>bamb_6480</td>
<td>91</td>
<td>Transaminase</td>
</tr>
<tr>
<td>ocfM</td>
<td>951</td>
<td>bamb_6481</td>
<td>94</td>
<td>Epimerase</td>
</tr>
<tr>
<td>ocfN</td>
<td>720</td>
<td>bamb_6482</td>
<td>90</td>
<td>Thioesterase</td>
</tr>
<tr>
<td>ORF16</td>
<td>288</td>
<td>C7511</td>
<td>93</td>
<td>Conserved hypothetical protein</td>
</tr>
</tbody>
</table>

*a:* Homolog to the putative proteins of *Burkholderia ambifaria* AMMD (GenBank: NC_008392);  
*b:* Predicted functions are based on annotation of strain AMMD and strain 383;  
*c:* homolog to C7511 of *B. lata* 383 (GenBank: NC_007509).  
Shaded grey region represents previously published work (13, 14).  
Annotations of a 12.9-kb genomic DNA fragment upstream of *ocfH* is not shaded.
**Fig. 1** Covalent structure of occidiofungin. A. Monoisotopic masses for each variant, Occidiofungin A-D, are provided along with the location of each variance designated by R1 and R2. B. Representative circle diagram of the covalent structure of occidiofungin.
Fig. 2 Expansion of the TOCSY 2D NMR spectra showing the amide to alpha, and amide to side chain spin systems for each assigned residue in occidiofungin. Horizontal lines show the multiple amide proton chemical shift values for each amino acid.
Fig. 3 Covalent structure of novel amino acid 2. The proton chemical shift values (in ppm) are shown next to their respective atoms. NOE values (in grey) are written next to the circles and arrows designate the proton interactions. The thick lines between the atoms represent proton couplings observed in the COSY 2D NMR data set.
Fig. 4 Identification of a chloro-BHY. A. Expansion of the HSQC 2D NMR spectra shows the proton and carbon chemical shift values for the chloro-BHY and BHY. B. ESI mass spectrometry data shows the presence of four variants differing in mass by the addition of an oxygen and chlorine. “abe” shows a typical stair step pattern for each isotope, while “xyz” shows an increase in the M + 3 isotope corresponding to the addition of $^{37}$Cl.
Fig. 5 Map of 58.1-kb occidiofungin gene cluster region of *Burkholderia contaminans* strain MS14. The positions and orientations of the known genes and potential ORFs are shown as horizontal arrows. Vertical arrows and flags indicate the insertion positions of the Tn5 transposon and the *nptII* gene cassette, respectively. The open and solid vertical arrows/flags represent reduction and elimination of antifungal activity, respectively. The arrows indicate putative promoter sequence. The genotypes of the MS14 mutants are: GG44(*ambR1::nptII*); MT15(*ambR2::Tn5*); MT13(*ocfD::Tn5*); MT18(*ocfE::Tn5*); MT14(*ocfH::Tn5*); MT16(*ocfH::Tn5*); and GG78(*ocfJ::nptII*).
**Fig. 6** Expression of the ORFs identified in the *ocf* gene cluster in strain MS14 and its mutants MS14MT15 and MS14GG44 (Table 1). Transcript levels of the tested genes and ORFs are presented relative to the transcript levels in the wild type MS14. Mean values for three biological replicates are given, and error bars represent the standard errors of the means.
Fig. 7 Representative diagram for the biosynthesis of occidiofungin. Some steps are demarcated with an asterisk, given the high potential of alternative synthetic routes in the model. For instance, it is conceivable that glycosylation can occur before the synthesis of the peptide. In addition, no homology exists in the database for the proposed dehydratase activity of OcfB, and this activity may be attributed to another gene outside of the gene cluster.
**Occidiofungin**

Occidiofungin A: $C_{102}N_{11}O_{31}H_{165}$; 1199.59 Da
Occidiofungin B: $C_{102}N_{11}O_{32}H_{166}$; 1215.59 Da
Occidiofungin C: $C_{102}N_{11}O_{31}H_{164}Cl_i$; 1233.56 Da
Occidiofungin D: $C_{102}N_{11}O_{32}H_{164}Cl_i$; 1249.56 Da

---

**B**

- Ser8
- Asn1
- BHN1
- NAA2
- Ser3
- BHY4
- Gly6
- DABAS
- cBHY4

Occidiofungin A: $R_1 = H; R_2 = H$
Occidiofungin B: $R_1 = OH; R_2 = H$
Occidiofungin C: $R_1 = H; R_2 = Cl$
Occidiofungin D: $R_1 = OH; R_2 = Cl$