Technical advance

A resource for the *in silico* identification of fungal polyketide synthases from predicted fungal proteomes

JAVIER A. DELGADO¹, OMAR AL-AZZAM², ANNE M. DENTON², SAMUEL G. MARKELL¹ AND RUBELLA S. GOSWAMI^{1,*}¹Department of Plant Pathology, North Dakota State University, Fargo, ND 58108, USA²Department of Computer Sciences, North Dakota State University, Fargo, ND 58108, USA

SUMMARY

The goal of this study was to develop a tool specifically designed to identify iterative polyketide synthases (iPKSs) from predicted fungal proteomes. A fungi-based PKS prediction model, specifically for fungal iPKSs, was developed using profile hidden Markov models (pHMMs) based on two essential iPKS domains, the β -ketoacyl synthase (KS) domain and acyltransferase (AT) domain, derived from fungal iPKSs. This fungi-based PKS prediction model was initially tested on the well-annotated proteome of *Fusarium graminearum*, identifying 15 iPKSs that matched previous predictions and gene disruption studies. These fungi-based pHMMs were subsequently applied to the predicted fungal proteomes of *Alternaria brassicicola*, *Fusarium oxysporum* f.sp. *lycopersici*, *Verticillium albo-atrum* and *Verticillium dahliae*. The iPKSs predicted were compared against those predicted by the currently available mixed-kingdom PKS models that include both bacterial and fungal sequences. These mixed-kingdom models have been proven previously by others to be better in predicting true iPKSs from non-iPKSs compared with other available models (e.g. Pfam and TIGRFAM). The fungi-based model was found to perform significantly better on fungal proteomes than the mixed-kingdom PKS model in accuracy, sensitivity, specificity and precision. In addition, the model was capable of predicting the reducing nature of fungal iPKSs by comparison of the bit scores obtained from two separate reducing and nonreducing pHMMs for each domain, which was confirmed by phylogenetic analysis of the KS domain. Biological confirmation of the predictions was obtained by polymerase chain reaction (PCR) amplification of the KS and AT domains of predicted iPKSs from *V. dahliae* using domain-specific primers and genomic DNA, followed by sequencing of the PCR products. It is expected that the fungi-based PKS model will prove to be a useful tool for the identification and annotation of fungal PKSs from predicted proteomes.

INTRODUCTION

Natural products, such as polyketides, are of great pharmaceutical and agrochemical value because of their antibacterial, antifungal, immunosuppressant and other biological activities. Polyketide compounds are a large group of naturally occurring active compounds that are synthesized by a family of proteins known as polyketide synthases (PKSs). PKSs are considered to be the most successful means of discovering and creating new natural and engineered active compounds for the manufacture of novel pharmaceutical drugs and pesticides (reviewed in Shen, 2003). Fungi are prolific producers of active compounds under natural conditions. However, fungal PKSs have been studied to a much lesser extent than bacterial PKSs. Fungal PKSs are involved in the biosynthesis of secondary metabolites and pigments, which have been shown to be important in the survival, competition and pathogenicity of several fungal species (Feng and Leonard, 1995; Gaffoor and Trail, 2006; Loppnau *et al.*, 2004; Zhang *et al.*, 2003). In addition, polyketide compounds can be either beneficial or detrimental to human health (reviewed in Keller *et al.*, 2005).

PKSs are multidomain proteins that biosynthesize polyketide compounds from the precursor acyl coenzyme A (acyl CoA) by sequential decarboxylative condensations, also known as Claisen condensations (Crawford *et al.*, 2008; reviewed in Gokhale *et al.*, 2007; Schümann and Hertweck, 2006; Shen, 2003; Weissman, 2008). Three types of PKS have been described, namely types I, II and III, which differ from each other in terms of their protein structure (reviewed in Gokhale *et al.*, 2007). Type I PKSs are commonly found in both bacteria and fungi, but bacterial type I PKSs are mostly modular proteins, whereas fungal type I PKSs are iterative (iPKSs). These two types of PKS present different biosynthesis mechanisms and domain architectures. Modular PKSs are arranged in modules, with each module consisting of a set of domains. iPKSs are organized in one single module with one set of domains that acts in a repetitive fashion until the final polyketide product has been synthesized (Gaitatzis *et al.*, 2002; reviewed in Gokhale *et al.*, 2007; Hoffmeister and Keller, 2007; Shen, 2003; Shen and Hutchinson, 1996; Wiesmann *et al.*, 1995).

Functional type I iPKSs can contain up to eight domains, but must have a minimum domain architecture consisting of a

*Correspondence: Email: rubella.goswami@ndsu.edu

β -ketoacyl synthase (KS), an acyltransferase (AT) and an acyl carrier protein (ACP) domain (reviewed in Hopwood and Sherman, 1990). Type I iPKSs activate the acyl CoA substrates through the ACP domain, whereas the KS domain elongates the carbon chain yielding the final product (reviewed in Gokhale *et al.*, 2007; Shen, 2003). Polyketide compounds are classified into nonreduced, reduced and highly reduced categories. Reducing type I iPKSs produce either partially reduced or highly reduced products by having multiple AT domains and/or other reducing domains, such as ketoreductase (KR), dehydratase (DH) and enoyl reductase (ER) (reviewed in Gokhale *et al.*, 2007). Reduced polyketide compounds include 6-methylsalicylic acid-related compounds (patulin, ochratoxin, etc.) and other mycotoxins (alternanapyrone, T-toxin, etc.). Nonreduced compounds include fungal pigments (conidial pigments and melanin) and nonreduced mycotoxins (aflatoxin, cercosporin, etc.). The KS domain is the most conserved PKS domain and can be used to categorize PKS proteins phylogenetically into reducing and nonreducing types (Bingle *et al.*, 1999; Kroken *et al.*, 2003; reviewed in Langfelder *et al.*, 2003; Nicholson *et al.*, 2001). This domain is persistently located towards the N-terminus of the protein sequence, followed by the AT domain in type I iPKSs (Bingle *et al.*, 1999; reviewed in Langfelder *et al.*, 2003).

The overall diversity and complexity of PKS proteins have made the molecular analysis and heterologous expression of these proteins much more difficult than originally expected. Evolution has separated fungal iPKSs from bacterial modular PKSs in such a manner that initial attempts to clone fungal PKSs failed because of the use of genomic probes from bacterial PKSs (reviewed in Schümann and Hertweck, 2006; Weissman, 2008). In addition, our preliminary studies have shown that bacterial PKS prediction models are not suitable for fungal iPKS mining. Fungal iPKS genes and proteins have been cloned and characterized in several different ways, such as by mapping the locus of T-toxin from *Cochliobolus heterotrophus* (Yang *et al.*, 1996), polymerase chain reaction (PCR) cloning using degenerate primers for the KS and AT domains of highly reducing, partially reducing and nonreducing PKSs (Bingle *et al.*, 1999; Fujii *et al.*, 2005; Kellner and Zak, 2009), probe hybridization to localize a melanin PKS in *Ascochyta rabiei* (White and Chen, 2007) and genome-based PKS mining (Bok *et al.*, 2006; Gaffoor *et al.*, 2005). In one such study, 15 iPKSs were identified and disrupted in the proteome of *Fusarium graminearum* Schwabe [teleomorph *Gibberella zeae* (Schwein.) Petch] using a genome-based PKS mining approach (Gaffoor *et al.*, 2005). Fungal type I iPKSs have been shown to group separately from bacterial PKSs and related proteins, such as fatty acid synthases (FASs), during phylogenetic analysis of the protein sequence of the KS domain. It has also been shown that the KS domain alone is sufficient to discriminate between reducing and nonreducing fungal iPKSs (Kroken *et al.*, 2003).

Profile hidden Markov models (pHMMs) are statistical models that have been used with great success in bioinformatics (Birney,

2001; De Fonzo *et al.*, 2007; Eddy, 1998). The idea behind pHMMs is to represent positions in a set of aligned sequences as hidden states. The actual amino acid instances are considered as emissions from the hidden match states. Insertions and deletions with respect to the match states are represented through insertion and deletion states, respectively. Emission and transition probabilities from and to the match states are determined on the basis of a set of target sequences that show the protein domain of interest. The first step in building a pHMM is to assemble those sequences that are to be represented and to construct a multiple alignment. The sequences that are to be included in a pHMM represent a design choice that depends on the objectives of the use of the pHMM. A common use of pHMMs is to search databases for members of protein families (Schuster-Böckler *et al.*, 2004). pHMMs can, however, also be built to be more specific and to distinguish subtypes of protein domains. A match in a pHMM search is characterized by a bit score (*S*), i.e. a statistical value based on the raw alignment score that has been normalized for comparison among different searches or databases. When the bit score of a match with a pHMM for one subtype is higher than that of the other subtype, this information can be used to differentiate such subtypes.

With the increasing number of fungi having whole genome sequences and predicted proteomes already available, a tool to identify PKS proteins for further studies would be very valuable, particularly when complete annotation of the genome is not available. Computational tools are already available on the web. However, they are limited in their ability to detect fungal PKSs, mainly because these models were constructed using bacterial modular PKSs (Li *et al.*, 2009; Starcevic *et al.*, 2008; Tae *et al.*, 2009). Foerstner *et al.* (2008) manually curated PKS domains retrieved from PKSDB, a modular bacterial PKS database (Yadav *et al.*, 2003), to build eight pHMMs for the following protein domains: AT, ACP, KS, DH, ER, KR, C-methyltransferase (MeT) and thioesterase (TE). These domain pHMMs were shown to be better predictors for PKS domains in metagenomics samples than those available at Pfam and TIGRFAM. However, bit score values were not found to be sufficient to discriminate between true PKS and non-PKS domains, and reliable PKS and non-PKS discrimination was only achieved by phylogenetic analyses (Foerstner *et al.*, 2008). The ACP domain, although essential for creating a functional PKS (reviewed in Hopwood and Sherman, 1990), was shown to be the least discriminative of all domains based on bit scores and phylogenetic analysis (Foerstner *et al.*, 2008). Khaldi *et al.* (2010) created a web-based software called SMURF (Secondary Metabolite Unique Regions Finder) for the prediction of secondary metabolite clusters, which included PKSs, nonribosomal peptide synthetases (NRPSs), hybrid PKS–NRPSs and prenyltransferases. SMURF predicts the backbone enzyme of each cluster and the decorating enzymes. The backbone enzymes are the main enzymes that catalyse the biosynthesis of the main product intermediate. The decorating enzymes modify the product intermediate, catalys-

ing the final product. Khaldi *et al.* (2010) used the pHMMs for the PKS domains available in Pfam and TIGRFAM to search for backbone enzymes. The search for PKSs was performed using the AT domain, the KS C-terminal domain and the KS N-terminal domain. The predictions carried out in this work were found to be accurate. However, this model was found to yield false positives and false negatives, which could be a result of the fact that the authors used pHMMs for PKS domains available at Pfam and TIGRFAM. As mentioned earlier, the PKS pHMMs from Pfam and TIGRFAM were found to be less accurate than the pHMMs developed by Foerstner *et al.* (2008). We believe that the approaches suggested by Foerstner *et al.* (2008) and Khaldi *et al.* (2010) are valid and effective, but suffer from the limitations posed by predictions obtained by mixed models. We believe that these limitations can be overcome by constructing protein sequence models based on the KS and AT domains of fungal PKSs. These could then be used to identify and study type I iPKSs within fungal genomes, predicted transcriptomes and proteomes. Therefore, the objectives of the present study were as follows: (i) to assemble the aforementioned databases to create two pHMMs corresponding to fungi-based KS and AT domains; (ii) to apply these two newly built fungi-based pHMMs to the identification of putative type I iPKSs from fungal proteomes; and (iii) to compare these fungi-based pHMMs with mixed-kingdom KS and AT pHMMs built using the KS and AT domain sequences from different taxonomic groups.

RESULTS

Data retrieval and model building

A total of 543 PKS protein sequences was compiled from GenBank. The presence of complete KS and AT domains in all the sequences was confirmed by conserved domain search (CDS) at the National Center for Biotechnology Information (NCBI), Bethesda, MD, USA. The accession numbers and sequences of all 543 PKS protein sequences are given in Notes S1 (see Supporting Information). PKS protein sequences with incomplete domains were not included in the pHMMs. Eight pHMMs were built: six fungi-based iPKS pHMMs containing the KS and AT domains from reducing and nonreducing iPKSs, both separately and in combination (FungalKS.hmm, FungalAT.hmm, RedFungalKS.hmm, NrFungalKS.hmm, RedFungalAT.hmm and NrFungalAT.hmm), and two mixed-kingdom PKS pHMMs (MixedKS.hmm and MixedAT.hmm). The mixed-kingdom KS and AT pHMMs were built using domain sequences obtained from Foerstner *et al.* (2008), which included PKS protein domains from the following taxonomic groups: Actinobacteria, Proteobacteria, Fungi, Cyanobacteria, Firmicutes, Animal, Mycetozoa, Alveolata, Virioplantae, Chloroflexi, Planctomycetes, Bacteroidetes/Chlorobi, among others. The six pHMMs based on the reducing, nonreducing and combined KS and AT domain sequences of fungal origin were considered for the fungi-

based iPKS model. The fungi-based and mixed-kingdom models were used to identify potential iPKSs from the predicted proteomes of all fungal species included in this study.

PKS predictions using fungi-based KS and AT pHMMs

The fungi-based PKS model predicted a total of 54 true KS domains and 56 true AT domains, which represented a total of 53 putative iPKS proteins with both KS and AT domains (Table 1). Further evaluation of the entire protein sequence showed that 45 of the 53 predicted iPKSs were expected to be functional based on the presence of the ACP domain in their protein sequence (Table 2). In the predicted proteome of *F. graminearum*, 14 of 15 iPKSs were found to contain the ACP domain and were expected to be functional. Likewise, six of 10 iPKS proteins in the proteome of *Verticillium albo-atrum* were expected to be functional because of the presence of the ACP domain.

True PKSs and non-PKS proteins were determined on the basis of their length, domain architecture, previous annotation and BLASTP matches using the nonredundant protein database at NCBI. Predicted domains were considered to be true PKS domains when they yielded bit scores higher than zero ($S > 0$). However, predicted domains that yielded bit scores lower than zero ($S < 0$) when using the fungi-based PKS model were considered to be non-PKS domains. The fungi-based KS and AT pHMMs were able to separate true PKS proteins from non-PKS proteins in all studied fungal proteomes. Among all five proteomes, 26 non-PKS proteins were predicted by the fungi-based KS pHMM, and 39 by the fungi-based AT pHMM (Fig. 1).

Model predictions showed nine KS and nine AT domains in the proteome of *Verticillium dahliae*. Among these, eight predicted iPKSs had both domains, together with the ACP domain, and were therefore expected to be functional. The protein sequence VDAG_04539 (1114 amino acids) was predicted only by AT pHMM with a bit score of 73.7. This sequence was also predicted by the fungi-based KS pHMM, but the bit score was -239.7. Domain architecture analysis showed the presence of a truncated KS domain (91 amino acids). In addition, the protein sequence VDAG_05633 (402 amino acids) was predicted only by the KS pHMM with a bit score of 506.5, which was determined to be a true KS domain protein by CDS at NCBI.

In the proteome of *Fusarium oxysporum* f.sp. *lycopersici*, 12 of 13 predicted PKSs were considered to be functional. Sequence FOXT_03945 (3802 amino acids) was predicted only by AT pHMM with a bit score of 467.3, but KS pHMM predicted it with a bit score of -9.1. Further domain architecture analysis showed the presence of a truncated KS domain (166 amino acids). Its domain architecture resembled a PKS-NRPS protein. As a result of the presence of a truncated KS domain, this sequence was considered to be a distorted PKS. Such distortion could be caused by several factors, including, among others, the state of scaffold assembly, gene prediction errors and mutations.

Table 1 Summary of polyketide synthase (PKS) and non-PKS predictions by the fungi-based and mixed-kingdom PKS models.

Proteome	Score of the KS domain pHMM			Score of the AT domain pHMM		
	Predicted domains	Mean	Range	Predicted domains	Mean	Range
Fungal-based PKS model						
Prediction of true PKS proteins						
<i>Fusarium graminearum</i>	15	791	603.9–935.4	15	412.9	204.7–497.4
<i>Alternaria brassicicola</i>	7	795.9	449.7–910.7	8	353.3	46.3–530.0
<i>Fusarium oxysporum</i>	13	767	579.6–958.4	14	354.3	27.0–516.1
<i>Verticillium albo-atrum</i>	10	749.1	515.3–928.1	10	386.2	9.1–525.0
<i>Verticillium dahliae</i>	9	751.4	506.5–928.8	9	340.3	73.7–527.1
Total	54	770.8	449.7–958.4	56	369.4	9.1–530.0
Prediction of non-PKS proteins						
<i>F. graminearum</i>	4	–242.6	–309.0––77.7	12	–149.8	–175.3––37.0
<i>A. brassicicola</i>	6	–276.9	–314.0––112.5	14	–156.8	–178.9––69.2
<i>F. oxysporum</i>	4	–180.2	–296.1––9.1	9	–143.1	–172.7––70.2
<i>V. albo-atrum</i>	6	–253.2	–312.1––84.5	9	–149.2	–177.3––8.7
<i>V. dahliae</i>	6	–272.2	–311.7––48.9	8	–149.1	–176.8––67.2
Total	26	–245.0	–314.0––9.1	52	–149.6	–178.9––8.7
Mixed-kingdom PKS model						
Prediction of true PKS proteins						
<i>F. graminearum</i>	15	178.4	123.8–216.0	15	278.8	197.7–347.1
<i>A. brassicicola</i>	7	171.1	65.9–212.9	8	238.1	29.5–354.6
<i>F. oxysporum</i>	13	148.2	78.5–242.8	14	241.8	27.7–497.7
<i>V. albo-atrum</i>	10	159.9	43.3–243.8	10	254.7	43.9–379.1
<i>V. dahliae</i>	9	151.7	60.5–243.9	9	221.6	52.1–380.8
Total	54	161.9	43.3–243.9	56	247	27.7–497.7
Prediction of non-PKS proteins						
<i>F. graminearum</i>	2	12	–5.4–30	13	–59.7	–112.8–173.2
<i>A. brassicicola</i>	13	–7.2	–14.3–7	20	–91.9	–116.3–82.2
<i>F. oxysporum</i>	2	–1.8	–1.8	8	–9.9	–108.6–92.1
<i>V. albo-atrum</i>	8	–4.7	–11.8–29.1	29	–90.7	–114.5–198.9
<i>V. dahliae</i>	14	–5.5	–11.6–37.1	34	–98.4	–114.5–128.0
Total	39	–1.4	–14.3–37.1	104	–70.1	–116.3–198.9

True PKSs were separated from non-PKS proteins according to their protein length, previous annotation, BLASTP annotation using the National Center for Biotechnology Information (NCBI) protein database and protein domain architecture.

The publicly available genome sequence of *Alternaria brassicicola* had a total of 838 scaffolds that were analysed using FGESH to predict a total of 7840 genes. The translated proteins were subsequently analysed by the KS and AT models to identify putative PKS sequences. Five of seven predicted PKSs were expected to be functional based on the presence of all KS, AT and ACP domains. The sequence Ab003seq693 (460 amino acids) was predicted only by AT pHMM with a bit score of 46.3. Its domain architecture showed the presence of a true AT domain and a truncated KS domain (71 amino acids). This sequence was also considered to be distorted as in the case of *F. oxysporum* f.sp. *lycopersici* mentioned above.

Discrimination between reducing and nonreducing iPKSs

Two approaches were tested to determine the efficacy of these pHMMs to discriminate between reducing and nonreducing iPKSs. The first approach involved phylogenetic analysis of the KS domain based on Kroken *et al.* (2003), and the second approach

involved the use of bit scores from predictions using the reducing and nonreducing models (RedFungalKS.hmm, NrFungalKS.hmm, RedFungalAT.hmm and NrFungalAT.hmm).

The KS domains of all 53 putative PKS proteins were analysed phylogenetically by neighbour-joining analysis (Fig. 2). This analysis showed that the predicted fungal iPKSs were grouped in clusters that were different from the modular bacterial PKSs and animal FASs. Moreover, within the main grouping, the fungal iPKSs separated into two distinctive clusters. The separation of the fungal iPKSs occurred according to their reducing nature. The reducing fungal iPKS cluster contained 41 of the predicted sequences, whereas the nonreducing fungal iPKS cluster grouped 12 sequences. According to this grouping, 10 iPKSs were reducing and five were nonreducing in *F. graminearum*, six iPKSs were reducing and one was nonreducing in the proteome of *A. brassicicola*, 11 were reducing and two were nonreducing in *F. oxysporum* f.sp. *lycopersici*, eight were reducing and two were nonreducing in *V. albo-atrum* and six were reducing and two were nonreducing in *V. dahliae*.

Table 2 Protein domain architecture of predicted polyketide synthase (PKS) protein sequences from all five fungal predicted proteomes.

Protein	Protein size (amino acids)	Domain architecture††	Functional status*
<i>Alternaria brassicicola</i>			
Ab001seq371	2376	KS-AT-DH-ER-KR-ACP	Functional PKS
Ab001seq648	1985	KS-AT-ACP-TE	Functional PKS
Ab003seq685	2905	KS-AT-DH-KR-TE	Nonfunctional PKS
Ab003seq693	460	[KS]-AT	Distorted PKS
Ab004seq368	2370	KS-AT-DH-MeT-ER-KR-ACP	Functional PKS
Ab004seq469	2998	MeT-KS-AT-DH-MeT-ER-KR	Nonfunctional PKS
Ab005seq505	2493	KS-AT-DH-MeT-ER-KR-ACP	Functional PKS
Ab008seq352	2200	KS-AT-DH-MeT-ER-KR-ACP	Functional PKS
<i>Fusarium graminearum</i>			
FG01790	2466	KS-AT-DH-MeT-ER-KR-ACP	Functional PKS
FG02395	1344	KS-AT	Nonfunctional PKS
FG03340	2566	KS-AT-DH-MeT-ER-KR-ACP	Functional PKS
FG03964	2030	KS-AT-DH-ACP-TE	Functional PKS
FG04588	2173	KS-AT-ACP-MeT	Functional PKS
FG05794	3178	KS-AT-DH-MeT-ER-KR-ACP	Functional PKS
FG08795	2351	KS-AT-DH-ER-KR-ACP	Functional PKS
FG10548	2464	KS-AT-DH-ER-KR-ACP	Functional PKS
FG12040	2068	KS-AT-ACP-TE	Functional PKS
FG12055	2346	KS-AT-DH-ER-KR-ACP	Functional PKS
FG12100	3921	KS-AT-DH-MeT-KR-ACP-CD-LuxE-ACP-KR	Functional PKS-NRPS
FG12109	2567	KS-AT-DH-MeT-ER-KR-ACP	Functional PKS
FG12121	2643	KS-AT-DH-MeT-KR-ACP	Functional PKS
FG12125	2288	AT-KS-AT-ACP-ACP-TE	Functional PKS
FG12977	2564	KS-AT-DH-MeT-ER-KR-ACP	Functional PKS
<i>Fusarium oxysporum</i> f.sp. <i>lycopersici</i>			
FOXT_02741	2553	KS-AT-DH-MeT-ER-KR-ACP	Functional PKS
FOXT_02884	1982	KS-AT-DH-ER-KR-ACP	Functional PKS
FOXT_03051	2358	KS-AT-DH-ER-KR-ACP	Functional PKS
FOXT_03945	3802	[KS]-AT-DH-MeT-KR-ACP-CD-LuxE-KR	Distorted PKS-NRPS
FOXT_04757	1241	KS-AT	Nonfunctional PKS
FOXT_05816	2084	KS-AT-ACP-ACP-TE	Functional PKS
FOXT_10805	2524	KS-AT-DH-MeT-ER-KR-ACP	Functional PKS
FOXT_11892	2439	KS-AT-DH-MeT-KR-ACP	Functional PKS
FOXT_11954	2454	KS-AT-DH-MeT-ER-KR-ACP	Functional PKS
FOXT_14587	3585	KS-AT-DH-MeT-KR-CD-LuxE-ACP-KR	Functional PKS-NRPS
FOXT_14850	2183	KS-AT-DH-MeT-ER-KR-ACP	Functional PKS
FOXT_15248	2377	KS-AT-DH-ER-KR-ACP	Functional PKS
FOXT_15296	3895	KS-AT-DH-MeT-KR-CD-LuxE-ACP-KR	Functional PKS-NRPS
FOXT_15886	2166	KS-AT-DH-MeT-ER-KR-ACP	Functional PKS
<i>Verticillium dahliae</i>			
VDAG_00190	2190	KS-AT-ACP-ACP-TE	Functional PKS
VDAG_01835	2624	KS-AT-DH-MeT-ER-KR-ACP	Functional PKS
VDAG_01856	2403	KS-AT-DH-ER-KR-ACP	Functional PKS
VDAG_03466	2161	KS-AT-DH-ER-KR-ACP	Functional PKS
VDAG_04539	1114	[KS]-AT-DH	Distorted PKS
VDAG_07270	2240	KS-AT-DH-MeT-KR-ACP	Functional PKS
VDAG_07928	4043	KS-AT-DH-MeT-KR-ACP-CD-LuxE-ACP-KR	Functional PKS-NRPS
VDAG_08448	2211	KS-AT-DH-ER-KR-ACP	Functional PKS
VDAG_09534	1857	KS-AT-DH-ACP-ACP-TE	Functional PKS
<i>Verticillium albo-atrum</i>			
VDBG_00162	2167	KS-AT-DH-ER-KR-ACP	Functional PKS
VDBG_00580	2006	KS-AT-ACP-ACP-TE	Functional PKS
VDBG_01329	1365	KS-AT-DH	Nonfunctional PKS
VDBG_01693	1971	KS-AT-DH-ER-KR	Nonfunctional PKS
VDBG_01714	2496	KS-AT-DH-MeT-ER-KR	Nonfunctional PKS
VDBG_04992	596	KS-AT	Nonfunctional PKS
VDBG_04998	3999	KS-AT-DH-MeT-KR-ACP-CD-LuxE-ACP-KR	Functional PKS-NRPS
VDBG_06312	1756	KS-AT-DH-ER-KR-ACP	Functional PKS
VDBG_09122	2207	KS-AT-DH-MeT-KR-ACP	Functional PKS
VDBG_09801	1912	KS-AT-DH-ACP-ACP-TE	Functional PKS

*Functional protein status was predicted according to the presence or absence of the acyl carrier protein (ACP) domain in the protein architecture.

†Square brackets around KS indicate the presence of a truncated domain.

‡AT, acyltransferase; CD, condensation domain of NRPSs; DH, dehydratase; ER, enoyl reductase; KR, ketoreductase; KS, β -ketoacyl synthase; LuxE, acyl-protein synthetase; MeT, C-methyltransferase; NRPS, nonribosomal peptide synthase; PKS-NRPS, hybrid proteins; TE, thioesterase.

Fig. 1 Box plots of the bit score distribution of profile hidden Markov model (pHMM) searches. (A) Bit score distribution of true β -ketoacyl synthase (KS) domains predicted by the fungi-based KS model and mixed-kingdom KS model. (B) Bit score distribution of true acyltransferase (AT) domains predicted by the fungi-based KS model and mixed-kingdom KS model. (C) Bit score distribution of non-KS domains predicted by the fungi-based KS model and mixed-kingdom KS model. (D) Bit score distribution of non-AT domains predicted by the fungi-based KS model and mixed-kingdom KS model. The vertical bars represent the range of bit scores for each prediction. The bottom horizontal bar of the boxplot represents the 25th percentile, the middle bar represents the median and the top bar represents the 75th percentile.

Comparison of the bit scores obtained from the searches performed using the reducing and nonreducing models showed that 42 of the predicted fungal iPKSs were reducing and 11 were nonreducing. The sequence of the predicted protein FG04588 of *F. graminearum* was predicted differently by the two methods. It was predicted as nonreducing by the phylogenetic analysis, but as reducing by bit score comparison of the reducing and nonreducing model predictions. Previous phylogenetic studies had concluded that this protein is a nonreducing iPKS (Gaffoor *et al.*, 2005; Kroken *et al.*, 2003). The bit scores for this sequence were 569.9 and 520.9 when using RedFungalKS and NrFungalKS, respectively, and 180.4 and 69.6 when using RedFungalAT and NrFungalAT pHMMs, respectively (Notes S4, see Supporting Information). Comparison of the reducing and nonreducing models for each domain suggested a reducing nature for this sequence, because higher bit scores were obtained for both domains when using the reducing pHMM. The sequence VDBG_00162 of *V. dahliae* has been annotated as a FAS *S*-acetyltransferase through automated annotation by the Broad Institute (Cambridge, MA, USA); however, our fungi-based model predicted this sequence to be a PKS. Domain architecture analysis showed that this sequence con-

tained true PKS domains. Furthermore, phylogenetic analysis of the KS domain showed that this sequence grouped with reducing fungal iPKSs rather than with animal FASs or other fungal fatty acid *S*-acetyltransferases. The Broad Institute predicted these fungal fatty acid *S*-acetyltransferases through automated annotation. BLASTP analysis of this sequence showed that one of its closest relatives is, in fact, a PKS from the fungus *Botryotinia fuckeliana* (GenBank accession no. AAR90244) with an *E* value of 0.0 and a bit score value of 5185, previously characterized by Kroken *et al.* (2003). The best matches for all predicted iPKSs in GenBank were determined to be PKSs by BLASTP (Notes S2, see Supporting Information).

Comparison of the fungi-based PKS model with a mixed-kingdom model

The performance of the fungi-based PKS model was evaluated in comparison with the mixed-kingdom PKS model by calculation of the accuracy, sensitivity, specificity and precision parameters (Table 3). Both the fungi-based PKS model and mixed-kingdom PKS model yielded the same number of true PKS predictions in the

Fig. 2 Phylogenetic analysis of the β -ketoacyl synthase (KS) domain of type I iterative polyketide synthase (iPKS) proteins identified from all five fungal proteomes. *, functional PKSs; b, bacterial PKSs; fas, fatty acid synthase; fassat, fatty acid synthase S-acetyltransferase; nr, nonreducing; r, reducing.

proteome of *F. graminearum* and in all five fungal proteomes, but differed in their performance as measured by the aforementioned parameters and statistical analyses of predictions using bit scores. The fungi-based KS and AT models performed with higher accuracy, specificity and precision than the mixed-kingdom KS and AT models, respectively, when evaluated using the *F. graminearum* proteome alone and all five fungal proteomes.

Paired *t*-test analyses showed that the mean bit scores of true KS domains were significantly different ($P < 0.0001$) between the fungi-based KS model and the mixed-kingdom KS model. Overall, the bit score values for the fungi-based KS model ranged from 449.7 to 958.4, with a mean of 771.5, whereas, for the mixed-kingdom KS model, the values ranged from 43.3 to 243.9, with a mean of 162.5 (Fig. 1A). Likewise, the paired *t*-test analyses determined that the mean bit scores of true AT domains were significantly different ($P < 0.0001$) between the fungi-based AT model and the mixed-kingdom AT model. Overall, the bit score values for the fungi-based AT model ranged from 9.1 to 530.0, with a mean of 372.5, whereas, for the mixed-kingdom AT model, the values ranged from 27.7 to 497.7, with a mean of 249.6 (Fig. 1B).

Independent group analyses (*t*-test) showed that the mean bit scores of true PKS domains were significantly different ($P < 0.0001$) from those of non-PKS domains for all fungi-based KS, fungi-based AT, mixed-kingdom KS and mixed-kingdom AT models. There was no bit score overlap between the bit score values of true PKS domains and non-PKS domains when the fungal proteomes were analysed with both the fungi-based KS and AT models (Table 1, Fig. 1). However, there was an overlap between the bit scores of true PKS domains and non-PKS domains when the fungal proteomes were analysed with both the mixed-kingdom KS and AT models (Table 1, Fig. 1). Moreover, the non-PKS domains predicted by both fungi-based models did not present positive bit score values ($S < 0$), in contrast with the mixed-kingdom models, where some non-PKS domains showed positive bit score values ($S > 0$), leading to the prediction of false positives (Fig. 1C,D). Overall, the fungi-based KS and AT models performed better than the mixed-kingdom KS and AT models, respectively.

Validation of the reducing/nonreducing fungi-based models

Both KS and AT models were cross-validated using a two-fold cross-validation approach with the reducing and nonreducing databases. In the KS model, 403 true positives, 139 true negatives, one false positive and no false negatives were determined, showing that the KS model had an accuracy of 99.8%, sensitivity

Table 3 Performance of fungi-based and mixed-kingdom β -ketoacyl synthase (KS) and acyltransferase (AT) profile hidden Markov models (pHMMs)*.

Parameters	Proteome of <i>Fusarium graminearum</i>				Proteome of all five fungal species†			
	Fungal-based pHMM		Mixed-kingdom pHMM		Fungal-based pHMM		Mixed-kingdom pHMM	
	KS domain	AT domain	KS domain	AT domain	KS domain	AT domain	KS domain	AT domain
True positives	15	15	15	15	54	56	54	56
True negatives	4	12	1	10	26	52	34	91
False positives	0	0	1	3	0	0	5	13
False negatives	0	0	0	0	0	0	0	0
Accuracy (%)	100.0	100.0	94.1	89.3	100.0	100.0	94.6	91.9
Sensitivity (%)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Specificity (%)	100.0	100.0	50.0	76.9	100.0	100.0	87.2	87.5
Precision (%)	100.0	100.0	93.8	83.3	100.0	100.0	91.5	81.2

*PKS domains identified as PKS domains ($S > 0$) were considered to be true positives, non-PKS domains identified as non-PKS domains were considered to be true negatives ($S < 0$), non-PKS domains identified as PKS domains ($S > 0$) were considered to be false positives and PKS domains identified as non-PKS domains were considered to be false negatives ($S < 0$).

†*Fusarium graminearum*, *Fusarium oxysporum* f.sp. *lycopersici*, *Alternaria brassicicola*, *Verticillium albo-atrum* and *Verticillium dahliae*.

of 100%, specificity of 99.3% and precision of 99.8%. Similarly, the AT model was found to have an accuracy of 99.8%, sensitivity of 99.8%, specificity of 100% and precision of 100% with 403 true positives, 139 true negatives, no false positives and one false negative.

PCR assays for confirmation of *V. dahliae* PKS sequences

Domain-specific primers (Table 4) were designed for the KS and AT domains of each *V. dahliae* PKS sequence in order to confirm the presence of the genes in four isolates obtained from potato plants affected by *Verticillium* wilt. Initially, sets of primers were tested to determine their ability to amplify the KS and AT domains of VDAG_00190, VDAG_01856, VDAG_03466 and VDAG_07928 genes from the isolates M5 and U2 obtained from infected potato stems. The PCR products were found to match the predicted sizes for each gene and were subsequently sequenced (Notes S1). The sequences of these amplicons were analysed using BLASTX. This sequence analysis confirmed either the KS or AT identity for each PCR product (Notes S1). All *V. dahliae* isolates, namely H5, M5, U1 and U2, showed the presence of seven of the eight predicted PKSs (Fig. 3). The KS and AT domains of VDAG_01835 were not amplified in any of the isolates, suggesting that it may be unique to strains most closely related to the sequenced isolate.

DISCUSSION

The present study explored the possibility of detecting type I iPKSs in predicted fungal proteomes using fungal PKS-specific domain pHMMs to provide a simple method to identify iPKSs from the predicted proteomes of unannotated genome sequences or transcript data. In order to achieve this goal, we manually selected and curated intact, complete protein sequences of KS and AT domains to create domain databases. Two compiled pHMMs (containing

sequences of each domain from reducing and nonreducing fungal iPKSs) and four pHMMs (containing only reducing or nonreducing sequences for each domain) were built based on the two protein domains, and used to identify fungal PKS proteins from the *F. graminearum*, *V. albo-atrum*, *F. oxysporum* f.sp. *lycopersici*, *V. dahliae* and *A. brassicicola* proteomes. Similar studies have demonstrated the advantages of identifying specific types of PKS, including type I iPKSs, through a bioinformatics-based genomics approach versus classical molecular cloning using degenerate primers (Bingle *et al.*, 1999; Gaffoor *et al.*, 2005; Kroken *et al.*, 2003). Among the available approaches, the most recently published resource was developed by Khaldi *et al.* (2010). This approach, called SMURF, is a web-based resource for the identification of gene clusters of secondary metabolites, including PKSs. This software has been used successfully to identify fungal iPKSs. However, it has been shown to yield false positives and, more importantly, false negatives. We believe that this might be a result of the use of Pfam and TIGRFAM pHMMs to predict PKS proteins. The pHMMs available at Pfam and TIGRFAM are mixed-kingdom models and have been shown to perform less accurately than the mixed models used by Foerstner *et al.* (2008), which we employed for comparisons. To our knowledge, there are no publicly available pHMMs solely based on protein domains of fungal PKSs.

The *F. graminearum* genome was sequenced in 2003 (Cuomo *et al.*, 2007) and extensive studies have since been conducted for gene prediction and characterization, including functional evaluation of all 15 iPKSs found in this fungus through gene disruption (Gaffoor *et al.*, 2005). Therefore, the proteome of *F. graminearum* was used as a control for the validation and evaluation of our fungi-based PKS model. Our goal was to develop a method for the identification of type I iPKSs from fungal plant pathogens for further analysis of their function, and their classification into reducing and nonreducing types, to study their regulation in association with disease initiation and development.

Table 4 Domain-specific primers for predicted polyketide synthase (PKS) protein sequences from *Verticillium dahliae*.

Protein designation*	PKS domain	Domain-specific PKS primers		Size of PCR product (bp)†
		Forward sequence (5'–3')	Reverse sequence (5'–3')	
VDAG_00190	KS	GAGCTTCTGTCGACGGCTT	AACCATTCTGCCGACGCCAT	752
	AT	GCAAGGACCTCTACGCCAC	CCCGTCAGCCTGTGGACTTC	716
VDAG_01835	KS	GGTCACGGGCTCCAAGACTG	TACTCGTGTGGAGCACGGC	772
	AT	CTTTGTACGGCGCTCCAGGT	ATCAGCAGCAGCGGCTAAG	708
VDAG_01856	KS	TTGTCTGATCCCCGGCACAG	TCGCCCCAAGACTCTCCAA	1334
	AT	CCGCAGCCTATGCTTGTGGA	AGCCCTTGGGGAAGCTCTGGT	708
VDAG_03466	KS	CCCGAACCCCGAGAAAAGGG	ACCGGCTTTGTTGAGGCACA	960
	AT	GCTCGGAGCCACTTTCTCCC	TACGTCCCTTCGCTGAAGCG	822
VDAG_07270	KS	TACGCGACGCACTATGACGG	CCTGTTGACCAGCCGCTCAT	1022
	AT	GACTCGGCTCTGCAGGCATT	TGCCGGTCCTTATCGCAGA	769bp
VDAG_07928	KS	GTCGAGGCCACAGCATTGA	ACACCTGCTGAGACCCCAT	963bp
	AT	GCTTCTCGTTGGAGCGGACA	CCGTATCCCCTACACTGGCG	806bp
VDAG_08448	KS	CTGCCGAGTTCTGGGAGCTG	GACGGCGAAAGATCCGCTCT	1079bp
	AT	TAGCCAGCATCTTGCACGG	CCCGCAAAAAGGACGCCATC	766bp
VDAG_09534	KS	AGCATTTCTGAGCGTCTGCA	ACTCCCCTAGCCAACGAGCA	778bp
	AT	TATGCGTTGACGCGCTCTT	TCGTACTACCATCGCTGCG	828bp

AT, acyltransferase; KS, β -ketoacyl synthase.

*Broad Institute protein designation.

†Polymerase chain reaction (PCR) product from genomic DNA.

Fig. 3 Polymerase chain reaction (PCR) assay for polyketide synthase (PKS) genes of four *Verticillium dahliae* isolates using β -ketoacyl synthase (KS) and acyltransferase (AT) domain-specific primers. Rows represent the different PKS genes. Isolates are arranged by columns.

Three PKS domains are considered to be the minimal domain architecture for a PKS protein to be functional, namely KS, AT and ACP domains (reviewed in Hopwood and Sherman, 1990). Our model was built using only the KS and AT domains, as it has been demonstrated previously that the ACP domain is the least suitable domain for discrimination between true PKS and non-PKS proteins through phylogenetic and pHMM analyses because of its short length (Foerstner *et al.*, 2008). Later, each protein sequence predicted by both the fungi-based KS and AT pHMMs was analysed by protein domain architecture searches to determine whether or not the predicted proteins were functional, based on the presence of the ACP domain in their domain architecture. Our models predicted 53 iPKSs, 45 of which were predicted to be functional according to their domain architecture. Of the 15 iPKSs identified

by our model in the predicted proteome of *F. graminearum*, 14 were found to be functional. The only iPKS believed to be non-functional because of the lack of the ACP domain was the protein FG02395. The length of this protein, according to the *Fusarium graminearum* Genome Database at the Munich Information Centre for Protein Sequences, was 1344 amino acids, and its domain architecture was KS–AT. However, further evaluation showed that this sequence has been experimentally determined to be 1995 amino acids in length and its domain architecture is KS–AT–ACP (Gaffoor *et al.*, 2005). This iPKS is functional and encodes for a PKS involved in the biosynthesis of zearalenone in *F. graminearum* isolates (Gaffoor *et al.*, 2005; Gaffoor and Trail, 2006). This is an example of an atypical protein for which an earlier gene prediction program did not predict the complete gene.

These and other results support the conclusion that our model is well suited for the identification of putative iPKSs in fungal genomes, and that our strategy of including only the KS and AT domains in our fungi-based PKS model is well justified.

The fungi-based KS model also showed the ability to exclude truncated KS domains. Truncated PKS proteins were only predicted by the fungi-based AT model. It has been reported by Wight *et al.* (2009) that the genome of *A. brassicicola* has nine iPKSs, but details of these iPKSs were not available. The only iPKS published as being associated with virulence (GenBank accession no. ACZ57548) was common with the sequence Ab001seq371 identified by the fungi-based PKS model. Our model only identified seven PKS sequences plus one truncated PKS sequence. We attribute this disagreement to differences in the gene prediction and to scaffold assembly. We found 13 PKS sequences in the proteome of *F. oxysporum* f.sp. *lycopersici* that had previously been annotated as conserved hypothetical proteins. However, the domain architecture and phylogenetic analysis of the KS domain clearly defines them as iPKS proteins. Their closest relatives were also identified as PKSs by stand-alone BLASTP against the NCBI nonredundant protein database.

We compared our fungi-based PKS model against a mixed-kingdom PKS model derived from the mining of PKSs within metagenomic libraries (Foerstner *et al.*, 2008). This mixed-kingdom model contained PKS domains from different taxonomic groups, including bacteria and fungi, among others. Models developed by Foerstner *et al.* (2008) have been proven to perform better than the models available at Pfam and TIGRFAM, which are also considered to be mixed-kingdom models. Similarly, our fungi-based KS and AT models were shown to perform even better than the mixed-kingdom KS and AT models in the prediction of iPKSs from fungal species. Although overall iPKS predictions between the two models were similar, our fungi-based PKS models were found to predict fewer false positives and false negatives. Moreover, the fungi-based model was significantly better in identifying iPKS proteins from closely related proteins. The performance was evaluated using the *F. graminearum* proteome and all five fungal proteomes. We observed that the fungi-based PKS models predicted all true PKS domains with positive bit score values ($S > 0$), whereas non-PKS domains were predicted with bit score values below zero ($S < 0$). The mixed-kingdom KS and AT models failed to discriminate between closely related PKS domains, yielding false positives as a result of some non-PKS domains being predicted with bit scores above zero ($S > 0$). In addition, the fungi-based KS model predicted true KS domains with higher bit scores than the mixed-kingdom model, with no overlap of bit score values. The fungi-based AT model had a bit score mean significantly different from that of the mixed-kingdom AT model, but there was overlap of bit score values. This can be justified by the fact that the KS domain is more conserved than the AT domain within the kingdoms (Bingle *et al.*, 1999; reviewed in Langfelder *et al.*, 2003;

Nicholson *et al.*, 2001). The only discrepancies observed in the predictions were in the sequences VDBG_00162 and VDAG_03466 from *V. albo-atrum* and *V. dahliae*, respectively, which have been annotated by the Broad Institute as FAS *S*-acetyltransferases, and the sequence VDAG_01835, which has been annotated as FAS. All three sequences were not discriminated as non-PKSs by the fungi-based PKS models, the mixed-kingdom PKS models or phylogenetic analysis. Further analysis showed that these proteins were also indistinguishable from PKSs according to their domain architecture. Therefore, it is likely that their true identity can only be resolved by gene disruption coupled with metabolite screening analyses, which are yet to be conducted. Indeed, fungal FAS *S*-acetyltransferases have only been identified in the predicted proteomes of *V. albo-atrum*, *Paracoccidioides brasiliensis* and *Pyrenophora tritici-repentis*, all through automated annotation by the Broad Institute.

During our search for currently available PKS prediction tools, we came across a computational tool called MAPSI (Management and Analysis for Polyketide Synthase Type I), which is an updated version of the previous tool named ASMPKS (Analysis System of Modular PKSs) (Ansari *et al.*, 2004; Tae *et al.*, 2009). According to the information available, MAPSI is able to analyse not only modular type I PKSs, but also iterative type I PKSs, from bacterial and fungal genomes previously uploaded to the software or manually uploaded by the user. This software works under the assumption that several organisms may share the same PKS gene clusters. This software identifies PKS genes, describes their domain architectures and associates the PKS genes with polyketide products. In the database, each PKS protein sequence is associated to a GenBank accession number as well as a polyketide product. The database consists of a total of 215 PKS protein sequences, 193 of which are bacterial PKSs and only 22 are fungal PKSs. The database MAPSI consists of 65 different polyketide products, 45 of which are synthesized by modular PKSs and 20 by iPKSs. There are only 17 polyketide products of fungal origin, all from iPKSs. We have not been able to upload and evaluate the fungal genomes included in this study as the 'user genome' option seems to be under construction. However, we feel that a more diverse fungal PKS protein database is needed in order to identify novel PKS proteins that have yet to be associated with a polyketide product. We also explored the possibility of using other bacterial-based PKS predictors before building our fungi-based PKS model. In this regard, we evaluated the software ClustScan (Starcevic *et al.*, 2008) and NP.searcher (Li *et al.*, 2009) with all five fungal genomes. The latter was initially developed for nonribosomal peptide synthases. ClustScan was not appropriate for fungal genomes as the software input asks for nucleotide sequences from which PKS genes are predicted using bacterial gene predictors. NP.searcher failed to predict any PKS from all five fungal genomes.

We observed that the comparison of bit scores from reducing and nonreducing models of either the KS or AT domains could be

used to determine the reducing nature of the predicted proteins before conducting a phylogenetic analysis of the KS domain. This observation was confirmed with phylogenetic analyses of the identified type I iPKSs from all five fungal proteomes using KS domain sequences reported previously (Kroken *et al.*, 2003). For example, the PKS sequence FG04588 has been classified previously as a nonreducing PKS (Gaffoor *et al.*, 2005; Kroken *et al.*, 2003), but our model classified it as reducing. The bit score comparison was a good estimator of the reducing nature of PKSs; however, phylogenetic analysis shows even more discriminatory power, as concluded by other authors (Foerstner *et al.*, 2008; Kroken *et al.*, 2003). Previous authors have also concluded that phylogenetic analyses are more powerful in distinguishing true PKS and non-PKS domains (Foerstner *et al.*, 2008). The reducing and nonreducing pHMMs presented high accuracy, sensitivity, specificity and precision (99–100%). Therefore, we believe that bit scores could be used effectively for initial categorization into reducing and nonreducing PKSs immediately after screening through the predicted proteome.

We feel that the use of the two fungi-based KS and AT models would be helpful in the identification and annotation of iPKSs from predicted fungal proteomes, and that fungal genomes and predicted transcriptomes could be screened to narrow down the number of sequences to study when dealing with whole-genome BLAST searches. This fungi-based PKS model has been used successfully in the genome of the legume fungal pathogen *Ascochyta rabiei*, at a scaffold stage, from which we have predicted its iPKSs and confirmed them through PCR and reverse transcription (RT)-PCR analyses. The domain sequences and alignment files developed and used in this work are given in Notes S1. User instructions for the use of HMMER3 software with the fungi-based KS and fungi-based AT alignments are given in Notes S3 (see Supporting Information). HMMER3 can be downloaded from <http://hmmer.janelia.org> (Eddy, 1998, 2008). The alignments are also available at SMART (Letunic *et al.*, 2009), which is a web-based protein domain database (http://smart.embl.de/help/additional_alignments.shtml). The submitted alignments contain the 543 KS and AT domains obtained from fungal PKSs from NCBI that were manually curated for this work, and the KS and AT domains obtained from the fungal proteomes included in this study, i.e. KS and AT domains from *A. brassicicola*, *F. graminearum*, *F. oxysporum* f.sp. *lycopersici*, *V. albo-atrum* and *V. dahliae*.

EXPERIMENTAL PROCEDURES

Fungal genomes

The predicted proteomes of five plant pathogens with fully sequenced genomes were selected for this study. They included the cereal pathogen *F. graminearum*, which has a fully sequenced and annotated genome, in which all the PKSs have been identified and characterized. Based on

the PKS identification status, *F. graminearum* was used for validation of the model. The predicted proteome of the *F. graminearum* isolate PH-1 (NRRL 31084) was obtained from the *F. graminearum* Genome Database (<http://mips.helmholtz-muenchen.de/genre/proj/fusarium>). This study also included the use of the predicted proteomes of the pathogens *F. oxysporum* f.sp. *lycopersici*, *Verticillium albo-atrum* and *V. dahliae* for which automated annotation of the genome is available. The proteomes of these three pathogens were retrieved from the Broad Institute (<http://www.broadinstitute.org/>). The genome sequence of fungus *A. brassicicola* (ATCC 96836) was downloaded from the Genome Sequencing Center at Washington University School of Medicine (<http://genome.wustl.edu>). Gene identification was carried out using FGENESH, the gene-finding algorithm from Softberry, with organism parameters specific to *Alternaria*. This algorithm yielded an output indicating the number of predicted genes, predicted exons, predicted mRNAs and predicted proteins.

Data compilation, model building and PKS prediction

Fungal PKS protein sequences used to build the PKS identification model were downloaded from GenBank (<http://www.ncbi.nlm.nih.gov/>), excluding proteins derived from *A. brassicicola*, *F. graminearum*, *F. oxysporum* f.sp. *lycopersici*, *V. albo-atrum* and *V. dahliae*. The KS and AT domains were manually curated from each PKS protein sequence using CDS at NCBI (<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>). The protein sequences that had complete amino acid sequences for both KS and AT domains were trimmed and edited to create two fasta files containing the KS and AT domains, respectively. Each domain sequence was given the same accession number as the original protein sequence. In order to build pHMMs, multiple alignments were performed for the sequences of each domain. The pHMMs were named fungi-based KS model (FungalKS.hmm) and fungi-based AT model (FungalAT.hmm). The pHMMs were first tested against the predicted proteins from *F. graminearum* to determine the performance of the pHMMs, and then against the predicted proteins of *V. albo-atrum* to further validate the performance of the pHMMs. Later, the predicted proteins of *A. brassicicola*, *F. oxysporum* f.sp. *lycopersici* and *V. dahliae* were searched against the pHMMs to identify potential PKS sequences by searching for KS and/or AT domain matches. The models were built using HMMER (version 2.3.2) software and the search was performed using the default parameters (i.e. local alignment) of the software. A new version of HMMER (HMMER3) can be found at <http://hmmer.janelia.org>. The fungi-based KS and AT sequences were further separated according to their predicted reducing nature to create four additional pHMMs: reducing fungi-based KS model (RedFungalKS.hmm), nonreducing fungi-based KS model (NrFungalKS.hmm), reducing fungi-based AT model (RedFungalAT.hmm) and nonreducing fungi-based AT model (NrFungalAT.hmm). These pHMMs were used to predict the reducing nature of the putative PKS sequences by comparison of the bit scores (S).

The domain architecture of the predicted PKSs was analysed using the Conserved Domain Architecture Retrieval Tool (CDART) and CDS at NCBI. A predicted PKS was considered to be functional when the minimum domain architecture requirements were met, i.e. all KS, AT and ACP domains were present. The closest relative of each functional PKS was determined by BLASTP at NCBI with an *E* value cut-off of 1.0×10^{-5} .

Fig. 4 Flow chart for profile hidden Markov model (pHMM) building, polyketide synthase (PKS) validations and predictions in fungal proteomes. AT, acyltransferase; BLASTP, protein basic local alignment search tool; KS, β -ketoacyl synthase; MIPS, Munich Information Centre for Protein Sequences; NCBI, National Center for Biotechnology Information; Nred, non-reducing; Red, reducing.

(<http://blast.ncbi.nlm.nih.gov/Blast.cgi>). Stand-alone BLASTP was performed against the nonredundant protein sequence database using the default general and scoring parameters. Figure 4 describes the work flow carried out from the PKS protein collection to model validation and predictions. The reducing nature of the PKSs predicted in this study was determined by phylogenetic analysis of the KS domain (Kroken *et al.*, 2003; Tamura *et al.*, 2007). The neighbour-joining phylogeny algorithm was used with the Poisson correction model, 20 000 replications in the bootstrap test and the gaps were treated with pairwise deletion.

Comparison of the fungi-based PKS model with a mixed-kingdom model

The performance of our fungi-based PKS model was evaluated against a mixed-kingdom PKS model for both the KS and AT domains developed by Foerstner *et al.* (2008). The performance was first evaluated using the predicted proteome of *F. graminearum* alone as reference, and later using all five fungal proteomes. Bit scores (S) were selected as a predictor. The class labels were as follows: true PKS domain sequences were scored as unity, and non-PKS domains as zero. Thus, true PKS domains predicted with

$S > 0$ were considered to be true positive (TP), non-PKS domains predicted with $S < 0$ were considered to be true negative (TN), true PKS domains predicted with $S < 0$ were considered to be false negative (FN) and non-PKS domains predicted with $S > 0$ were considered to be false positive (FP). The performance of the model was evaluated as follows: accuracy = $[TP + TN]/[TP + TN + FP + FN]$; sensitivity = $TP/[TP + FN]$; specificity = $TN/[TN + FP]$; precision = $TP/[TP + FP]$ (Baldi *et al.*, 2000).

Validation of the reducing and nonreducing pHMMs was performed using a two-fold cross-validation. Prior to validation, the PKS sequences were screened for duplicate sequences and even multiple versions of the same sequence to ensure the independence of the training set and test set from one another. The KS domain sequences were separated into two files: the first contained 202 reducing KS sequences and 70 nonreducing KS sequences; the second contained 202 reducing KS sequences and 69 nonreducing KS sequences. The AT domain sequences were also separated into two files as described previously. The class labels were as follows: reducing sequences were scored as unity, and nonreducing sequences as zero. Thus, reducing sequences predicted as reducing were considered to be true positive (TP), nonreducing sequences predicted as nonreducing were considered to be true negative (TN), reducing sequences predicted as

nonreducing were considered to be false negative (FN) and nonreducing sequences predicted as reducing were considered to be false positive (FP). The performance of the models was also evaluated in comparison with the phylogenetic analysis. The performance of the models was evaluated using the parameters of accuracy, sensitivity, specificity and precision, calculated as mentioned above.

To further compare the fungi-based PKS model with the mixed-kingdom PKS model, statistical analyses were performed using Statistical Analysis Software (SAS) version 9.1 (SAS Institute, Inc., Cary, NC, USA). Comparisons of the prediction of true PKSs between the fungi-based and mixed-kingdom PKS models for both the KS and AT domains were analysed using paired *t*-test analysis (PROC TTEST) under the hypothesis that the mean difference between the models is equal to zero. Comparisons between the predictions of true PKS versus non-PKS by the models were analysed using independent group analysis (PROC TTEST) under the hypothesis that the means of each group are equal. Box plot representations were obtained using PROC BOXPLOT. All analyses were performed using bit scores as the variable ($\alpha = 0.05$).

DNA extraction, PCR assays and DNA sequencing of *V. dahliae* PKSs

Four isolates of *V. dahliae* were provided by Dr Neil Gudmestad at North Dakota State University, Fargo, ND, USA. Isolates H5, M5, U1 and U2 were obtained from potato plants affected by Verticillium wilt from central Minnesota, USA. The isolates were grown in 100 mL of potato dextrose broth (EMD Chemicals, Darmstadt, Germany) for 10 days at room temperature. Fifty milligrams of freeze-dried mycelium were placed in lysing matrix A (MP Biomedicals, Solon, OH, USA) with 700 μ L of nuclei lysis solution and homogenized using a FastPrep® homogenizer (MP Biomedicals) set at a speed of 6.0 for 40 s. DNA extraction from fungal isolates was performed using the Wizard® Genomic DNA Purification Kit (Promega, San Luis Obispo, CA, USA) according to the manufacturer's instructions. The predicted PKS genes from the *V. dahliae* isolates were amplified using specific PCR primers designed using Primer-BLAST (<http://www.ncbi.nlm.nih.gov/tools/primer-blast>) (Table 4). PCR assays of *V. dahliae* PKSs were carried out using a Top Taq DNA Polymerase Kit (Qiagen, Valencia, CA, USA) with a reaction volume of 25 μ L containing 25 μ g of fungal DNA, 1 $\times$ PCR buffer, pH 8.7, 2.0 μ M of $MgCl_2$, 200 μ M of deoxy-nucleoside triphosphates (dNTPs), 200 μ M of oligonucleotide primers and 0.5 units of Top Taq DNA polymerase. The amplicons were separated on a 1% agarose gel with a GelRed™ nucleic acid gel stain (Biotium, Inc., Hayward, CA, USA) for 45 min at 90 mV. Selected amplicons were ligated to the pGEM-T easy vector (Promega) and cloned into One Shot® Mach1™ competent cells (Invitrogen, Carlsbad, CA, USA). The plasmids were isolated using Wizard® Plus SV Miniprep DNA Purification Systems (Promega). The cloned insert was sequenced using M13 forward and reverse primers at McLab DNA sequencing services (San Francisco, CA, USA).

ACKNOWLEDGEMENTS

We would like to thank the *Fusarium* Comparative Project and the *Verticillium* Group Database (<http://www.broadinstitute.com>), both funded by the US Department of Agriculture's Cooperative State Research Education

and Extension Service, for making the genomes of *Fusarium oxysporum* f.sp. *lycopersici*, *Verticillium albo-atrum* and *Verticillium dahliae* publicly available. We would also like to thank Dr Christopher Lawrence (Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA) for allowing the use of the *Alternaria brassicicola* genome, and Dr Neil Gudmestad (North Dakota State University, Fargo, ND, USA) for providing the *V. dahliae* isolates used in this study.

REFERENCES

- Ansari, M.Z., Yadav, G., Gokhale, R.S. and Mohanty, D. (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res.* **32**, W405–W413.
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F. and Nielsen, H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinform. Rev.* **16**, 412–424.
- Bingle, L.H., Simpson, T.J. and Lazarus, C.M. (1999) Ketosynthase domain probes identify two subclasses of fungal polyketide synthase genes. *Fungal Genet. Biol.* **26**, 209–223.
- Birney, E. (2001) Hidden Markov models in biological sequence analysis. *IBM J. Res. Dev.* **45**, 449–454.
- Bok, J.W., Hoffmeister, D., Maggio-Hall, L.A., Murillo, R., Glasner, J.D. and Keller, N.P. (2006) Genomic mining for *Aspergillus* natural products. *Chem. Biol.* **13**, 31–37.
- Crawford, J.M., Thomas, P.M., Scheerer, J.R., Vagstad, A.L., Kelleher, N.L. and Townsend, C.A. (2008) Deconstruction of iterative multidomain polyketide synthase function. *Science*, **320**, 243–246.
- Cuomo, C.A., Güldener, U., Xu, J.-R., Trail, F., Turgeon, B.G., Di Pietro, A., Walton, J.D., Ma, L.-J., Baker, S.E., Rep, M., Adam, G., Antoniw, J., Baldwin, T., Calvo, S., Chang, Y.-L., DeCaprio, D., Gale, L.R., Gnerre, S., Goswami, R.S., Hammond-Kosack, K., Harris, L.J., Hilburn, K., Kennell, J.C., Kroken, S., Magnuson, J.K., Mannhaupt, G., Mauceli, E., Mewes, H.-W., Mitterbauer, R., Muehlbauer, G., Münsterkötter, M., Nelson, D., O'Donnell, K., Ouellet, T., Qi, W., Quesneville, H., Roncero, M.I.G., Seong, K.-Y., Tetko, I.V., Urban, M., Waalwijk, C., Ward, T.J., Yao, J., Birren, B.W. and Kistler, H.C. (2007) The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science*, **317**, 1400–1402.
- De Fonzo, V., Aluffi-Pentini, F. and Parisi, V. (2007) Hidden Markov models in bioinformatics. *Curr. Bioinform.*, **2**, 49–61.
- Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
- Eddy, S.R. (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput. Biol.* **4**, e1000069.
- Feng, G.H. and Leonard, T.J. (1995) Characterization of the polyketide synthase gene (*pksL1*) required for aflatoxin biosynthesis in *Aspergillus parasiticus*. *J. Bacteriol.* **177**, 6246–6254.
- Foerster, K.U., Doerks, T., Creevey, C.J., Doerks, A. and Bork, P. (2008) A computational screen for type I polyketide synthases in metagenomics shotgun data. *PLoS ONE*, **3**, e3515.
- Fujii, I., Yoshida, N., Shimomaki, S., Oikawa, H. and Ebizuka, Y. (2005) An iterative type I polyketide synthase PKS catalyzes synthesis of the decaketide alternapyrone with region-specific octa-methylation. *Chem. Biol.* **12**, 1301–1309.
- Gaffoor, I. and Trail, F. (2006) Characterization of two polyketide synthase genes involved in zearalenone biosynthesis in *Gibberella zeae*. *Appl. Environ. Microbiol.* **72**, 1793–1799.
- Gaffoor, I., Brown, D.W., Plattner, R., Proctor, R.H., Qi, W. and Trail, F. (2005) Functional analysis of the polyketide synthase genes in the filamentous fungus *Gibberella zeae* (Anamorph *Fusarium graminearum*). *Eukaryot. Cell*, **4**, 1926–1933.
- Gaitatzis, N., Silakowski, B., Kunze, B., Nordsiek, G., Blöcker, H., Höfle, G. and Müller, R. (2002) The biosynthesis of the aromatic myxobacterial electron transport inhibitor stigmatellin is directed by a novel type of modular polyketide synthase. *J. Biol. Chem.* **277**, 13082–13090.
- Gokhale, R.S., Sankaranarayanan, R. and Mohanty, D. (2007) Versatility of polyketide synthases in generating metabolic diversity. *Curr. Opin. Struct. Biol.* **17**, 736–743.
- Hoffmeister, D. and Keller, N.P. (2007) Natural products of filamentous fungi: enzymes, genes, and their regulation. *Nat. Prod. Rep.* **24**, 393–416.
- Hopwood, D.A. and Sherman, D.H. (1990) Molecular genetics of polyketides and its comparison to fatty acid biosynthesis. *Annu. Rev. Genet.* **24**, 37–66.

- Keller, N.P., Turner, G. and Bennett, J.W. (2005) Fungal secondary metabolism—from biochemistry and genomics. *Nat. Rev. Microbiol.* **3**, 937–947.
- Kellner, H. and Zak, D.R. (2009) Detection of expressed fungal type I polyketide synthase genes in a forest soil. *Soil Biol. Biochem.* **41**, 1344–1347.
- Khalidi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H. and Fedorova, N.D. (2010) SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.* **47**, 736–741.
- Kroken, S., Glass, N.L., Taylor, J.W., Yoder, O.C. and Turgeon, B.G. (2003) Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. *PNAS*, **100**, 15670–15675.
- Langfelder, K., Streibel, M., Jahn, B., Haase, G. and Brakhage, A.A. (2003) Biosynthesis of fungal melanins and their importance for human pathogenic fungi. *Fungal Genet. Biol.* **38**, 143–158.
- Letunic, I., Doerks, T. and Bork, P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.* **37**, D229–D232.
- Li, M.H.T., Ung, P.M.U., Zajkowski, J., Garneau-Tsodikova, S. and Sherman, D.H. (2009) Automated genome mining for natural products. *BMC Bioinformatics*, **10**, 185.
- Loppnau, P., Tanguay, P. and Breuil, C. (2004) Isolation and disruption of the melanin pathway polyketide synthase gene of the softwood deep stain fungus *Ceratocystis resinifera*. *Fungal Genet. Biol.* **41**, 33–41.
- Nicholson, T.P., Rudd, B.A.M., Dawson, M., Simpson, T.J. and Cox, R.J. (2001) Design and utility of oligonucleotide gene probes for fungal polyketide synthases. *Chem. Biol.* **8**, 157–178.
- Schümann, J. and Hertweck, C. (2006) Advances in cloning, functional analysis and heterologous expression of fungal polyketide synthase genes. *J. Biotechnol.* **124**, 690–703.
- Schuster-Böckler, B., Schultz, J. and Rahmann, S. (2004) HMM logos for visualization of protein families. *BMC Bioinformatics*, **5**, 7.
- Shen, B. (2003) Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Curr. Opin. Chem. Biol.* **7**, 285–295.
- Shen, B. and Hutchinson, C.R. (1996) Deciphering the mechanism for the assembly of aromatic polyketides by a bacterial polyketide synthase. *PNAS*, **93**, 6600–6604.
- Starcevic, A., Zucko, J., Simunkovic, J., Long, P.F., Cullum, J. and Hranueli, D. (2008) ClustScan: an integrated program package for the semi-annotation of modular biosynthetic gene clusters and *in silico* prediction of novel chemical structures. *Nucleic Acids Res.* **36**, 6882–6892.
- Tae, H., Sohng, J.K. and Park, K. (2009) Development of an analysis program of type I polyketide synthase gene clusters using homology search and profile hidden Markov model. *J. Microbiol. Biotechnol.* **19**, 140–146.
- Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599.
- Weissman, K.J. (2008) Anatomy of a fungal polyketide synthase. *Science*, **320**, 186–187.
- White, D. and Chen, W. (2007) Towards identifying pathogenic determinants of the chickpea pathogen *Ascochyta rabiei*. *Eur. J. Plant. Pathol.* **119**, 3–12.
- Wiesmann, K.E.H., Cortes, J., Brown, M.J.B., Cutter, A.L., Staunton, J. and Leadlay, P.F. (1995) Polyketide synthesis in vitro on a modular polyketide synthase. *Chem. Biol.* **2**, 583–589.
- Wight, W.A., Kim, K.H., Lawrence, C.B. and Walton, J.D. (2009) Biosynthesis and role in virulence of the histone deacetylase inhibitor depudecin from *Alternaria brassicicola*. *MPMI*, **22**, 1258–1267.
- Yadav, G., Gokhale, R.S. and Mohanty, D. (2003) SEARCHPKS: a program for detection and analysis of polyketide synthase domains. *Nucleic Acids Res.* **31**, 3654–3658.
- Yang, G., Rose, M.S., Turgeon, B.G. and Yoder, O.C. (1996) A polyketide synthase is required for fungal virulence and production of the polyketide T-toxin. *Plant Cell*, **8**, 2139–2150.
- Zhang, A., Lu, P., Dahl-Roshak, A.M., Pares, P.S., Kennedy, S., Tkacz, J.S. and An, Z. (2003) Efficient disruption of a polyketide synthase gene (pks1) required for melanin synthesis through *Agrobacterium*-mediated transformation of *Glarea lozoyensis*. *Mol. Genet. Genomics*, **268**, 645–655.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Notes S1 Nucleotide and protein sequences used in the present study and generated during the study.

Notes S2 BLASTP of predicted iterative polyketide synthases (iPKSs) from all five fungal proteomes against the National Center for Biotechnology Information (NCBI) nonredundant protein database.

Notes S3 User instructions on how to build and use HMMER for polyketide synthase (PKS) protein sequence search.

Notes S4 Outputs of all iterative polyketide synthase (iPKS) predictions, using the fungi-based PKS model and the mixed-kingdom PKS model.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.