Early diverging insect-pathogenic fungi of the order entomophthorales possess diverse and unique subtilisin-like serine proteases
Arnesen, Jonathan A.; Malagocka, Joanna; Gryganskyi, Andrii; Grigoriev, Igor V.; Voigt, Kerstin; Stajich, Jason E.; Licht, Henrik H. De Fine

Published in:
G3: Genes, Genomes, Genetics (Bethesda)

DOI:
10.1534/g3.118.200656

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Early Diverging Insect-Pathogenic Fungi of the Order Entomophthorales Possess Diverse and Unique Subtilisin-Like Serine Proteases

Jonathan A. Arnesen,* Joanna Malagocka,* Andrii Gryganskyi,† Igor V. Grigoriev,‡ Kerstin Voigt,§ ** Jason E. Stajich,†† and Henrik H. De Fine Licht*,1

*Section for Organismal Biology, Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsevej 40, 1871 Frederiksberg, Denmark, †Department of Biology, Duke University, Durham, North Carolina, USA, ‡US Department of Energy Joint Genome Institute, Walnut Creek, California, USA, §Jena Microbial Resource Collection (JMRC), Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, Adolf-Reichwein-Str. 23, 07745 Jena, Germany, **Institute of Microbiology, Friedrich Schiller University, Neugasse 25, 07743 Jena, Germany, and ††Department of Plant Pathology and Microbiology, University of California, Riverside, California

ORCID IDs: 0000-0002-6037-0092 (A.G.); 0000-0002-7591-0020 (J.E.S.); 0000-0003-3326-5729 (H.H.D.F.L.)

ABSTRACT Insect-pathogenic fungi use subtilisin-like serine proteases (SLSPs) to degrade chitin-associated proteins in the insect procuticle. Most insect-pathogenic fungi in the order Hypocreales (Ascomycota) are generalist species with a broad host-range, and most species possess a high number of SLSPs. The other major clade of insect-pathogenic fungi is part of the subphylum Entomophthoromycotina (Zoopagomycota, formerly Zygomycota) which consists of high host-specificity insect-pathogenic fungi that naturally only infect a single or very few host species. The extent to which insect-pathogenic fungi in the order Entomophthorales rely on SLSPs is unknown. Here we take advantage of recently available transcriptomic and genomic datasets from four genera within Entomophthoromycotina: the saprobic or opportunistic pathogens Basidiobolus meristosporus, Conidiobolus coronatus, C. thromboides, C. incongruus, and the host-specific insect pathogens Entomophthora muscae and Pandora formicae, specific pathogens of house flies (Musca domestica) and wood ants (Formica polyctena), respectively. In total 154 SLSP from six fungi in the subphylum Entomophthoromycotina were identified: E. muscae (n = 22), P. formicae (n = 6), B. meristosporus (n = 60), C. thromboides (n = 18), C. coronatus (n = 36), and C. incongruus (n = 12). A unique group of 11 SLSPs was discovered in the genomes of the obligate biotrophic fungi E. muscae, P. formicae and the saprobic human pathogen C. incongruus that loosely resembles bacillosipeptidase F-like SLSPs. Phylogenetics and protein domain analysis show this class represents a unique group of SLSPs so far only observed among Bacteria, Oomycetes and early diverging fungi such as Cryptomycota, Microsporidia, and Entomophthoromycotina. This group of SLSPs is missing in the sister fungal lineages of Kickxellomycotina and the fungal phyla Mucoromycota, Ascomycota and Basidiomycota fungi suggesting interesting gene loss patterns.
(Meerupvati et al. 2013), and dermatophytic fungi such as Arthroderma benhamiae and Trichophyton verrucosum that can cause nail and skin infections in humans and animals (Burmester et al., 2011; Desjardins et al., 2011; Martinez et al., 2012; Sharpton et al., 2009). Fungi that are able to utilize chitin-rich substrates, including many insect pathogenic fungi, thus appear to often be associated with a diversified and expanded set of SLSPs.

Although SLSPs are expanded among insect pathogenic fungi, this group of proteases is ubiquitous among eukaryotic organisms. Most SLSPs are secreted externally or localized to vacuoles, and especially in saprobic and symbiotic fungi SLSPs constitute an important component of the secretome (Li et al., 2017). According to the MEROPS peptidase classification, the S8 family of SLSPs together with the S53 family of serine-carboxyl proteases make up the SB clan of subtilases (Rawlings et al., 2016). The S8 family of SLSPs is characterized by an Asp-His-Ser catalytic triad (DHS triad), which forms the active site and is further divided into two subfamilies S8A and S8B. Subfamily S8A contains most S8 representatives, including the well-known Proteinase K enzyme that is widely used in laboratories as a broad-spectrum protease. The S8B SLSPs are kexins and furins which cleave peptides and protohormones (Jalving et al., 2000; Muszewska et al., 2017, 2011). Based on characteristic protein domain architectures and protein motifs surrounding the active site residues, the large S8A subfamily of SLSPs is further divided into several groups such as proteinase-K and pyrolisin. Besides these two major groups of proteinase K-like and pyrolisin subfamilies, six new groups of subtilase genes designated new 1 to new 6 have recently been found (Li et al., 2017; Muszewska et al., 2011). The analysis of fungal genome data from a wide taxonomic range has shown that the size of the proteinase K gene family has expanded independently in fungi pathogenic to invertebrates (Hypocreales) and vertebrates (Onygenales) (Muszewska et al., 2017; Sharpton et al., 2009). Closely related invasive human-pathogenic fungi, however, do not show the same expansions and related pathogenicity in non-pathogens and non-pathogens can show the same expansions (Muszewska et al., 2011; Whiston and Taylor 2016). This suggests that the number of SLSPs that a fungus possess is not directly related to pathogenicity, but instead is associated with the use of dead or alive animal tissue as growth substrate (Li et al., 2017; Muszewska et al., 2011).

Most anamorphic insect-pathogenic fungi in the order Hypocreales (Ascomycota) are generalist species with a broad host-range capable of infecting most major orders of insects (e.g., M. robertsi and B. bassiana) or specific to larger phylogenetic groups (e.g., the locust-specific M. acridum or the coleopteran pathogen B. brongniartii) (Boomsma et al., 2014; Hu et al., 2014). The above-mentioned inferences of fungal SLSP evolution rely almost exclusively on insights from Ascomycota, and consequently have strong sampling bias toward generalist insect-pathogenic fungi. In contrast, the other major clade of insect-pathogenic fungi in the subphylum Entomophthoromycotina (Zoopagomycota, the saprobic Basidiobolus meristosporus, the saprobic and opportunistic pathogens, Conidiobolus coronatus, C. incongruus, and the host-specific insect pathogens Entomophthora muscae and Pandora formicae, specific pathogens of house flies (Musca domestica) and wood ants (Formica polyctena), respectively. We use phylogenetics and protein domain analysis to show that the obligate biotrophic fungi E. muscae, P. formicae and the saprobic human pathogen C. incongruus, in addition to more “classical” fungal SLSPs, harbor a unique group of SLSPs that loosely resembles bacillopeptidase F-like SLSPs.

**MATERIALS AND METHODS**

**Sequence database searches for subtilisin-Like serine proteases**

We identified putative subtilisin-like serine proteases (SLSPs) from six fungi in the subphylum Entomophthoromycotina: Entomophthora muscae, Pandora formicae, Basidiobolus meristosporus, Conidiobolus coronatus, C. incongruus and C. thromboides. First, protein family (pfam) domains were identified in the de-novo assembled transcriptomes of E. muscae KVL-14-117 (De Fine Licht et al., 2017) and P. formicae (Malagocka et al., 2015) using profile Hidden Markov Models with hmmsearch scans (E-value < 1e-10) against the PFAM-A database ver. 31.0 using HMMER ver. 3 (Eddy 1998; Finn et al., 2016). All sequences in the transcriptome datasets containing the S8 subtilisin-like protease domain (PF00082) were identified and included in further analyses. Second, all sequences that contain the PF00082 domain were obtained from the genomes of B. meristosporus CBS 931.73 (Mondo et al., 2017), C. coronatus NRR128638 (Chang et al., 2015), and C. thromboides FSU 785 from the US Department of Energy Joint Genome Institute MycoCosm genome portal (http://igi.doc.gov/fungi). Third, predicted coding regions in the genome sequence of C. incongruus CDC-B7586 (Chibucos et al., 2016), were searched for the presence of the S8 subtilisin-like protease domain (PF00082) as described above.

Sequences encoding an incomplete Asp-His-Ser catalytic triad (DHS triad) characteristic of S8 family proteases were regarded as potential pseudogenes and excluded from further analysis. Although a subset may represent neofunctionalizations, the risk of including false-positive SLSPs in down-stream analyses were considered too high. A preliminary protein alignment made with ClustalW (Larkin et al., 2007) using default parameters and construction of a Neighbor-Joining tree using Geneious 4.8.5 (Kearse et al., 2012) revealed a highly divergent group of P. formicae SLSP-sequences that had significant homology with insect proteases (blastp, E-value < 1e-6, ncbi-nr protein database, accessed June 2017). These putative insect-sequences likely originate from the ant host, Formica polyctena, and represents host contamination that were not filtered out from the dual-RNAseq reads used to construct the P. formicae transcriptome (Malagocka et al., 2015). These divergent sequences were therefore removed and excluded from further analysis.

**Protein domain architecture and sequence clustering**

The domain architectures of all putative SLSPs identified within Entomophthoromycotina were predicted using Pfam domain annotation. The presence of putative signal peptides for extracellular secretion were predicted using SignalP (Petersen et al., 2011). A Markov Clustering
Algorithm (MCL) was used to identify clusters of similar proteins among putative SLSPs identified within Entomophthoromycotina. Clustering using MCL is based on a graph constructed by an all-vs-all-BLAST of SLSPs (BLASTP, E-value < 1e-10). The Tribe-MCL protocol (Enright et al., 2002) as implemented in the Spectral Clustering of Protein Sequences (SCPS) program (Nepusz et al., 2010) was used with inflation = 2.0. The inflation parameter is typically set between 1.2 – 5.0 (Nepusz et al., 2010), and controls the “tightness” of the sequence matrix, with lower values leading to fewer clusters and higher values to more sequence clusters. To putatively assign protease function to the newly identified Entomophthoromycotina sequence clusters, the Tribe-MCL protocol was used to identify clusters of similar proteins between the putative SLSPs identified within Entomophthoromycotina and 20,806 protease sequences belonging to the peptidase subfamily S8A obtained from the MEROPS database, accessed November 2017 (Rawlings et al., 2016). Investigation of the MEROPS protease sequences.
that clustered together with the identified Entomophthoromycotina sequence clusters allowed putative protease holotype information to be assigned to the identified clusters.

Phylogenetic analysis
All identified putative Entomophthoromycotina SLSP coding nucleotide sequences were aligned in frame to preserve codon structure using MAFFT (Larkin et al., 2007). Unreliable codon-columns with a Guidance2 score below 0.90 in the multiple sequence alignment were removed (Penn et al., 2010). The best model for phylogenetic analysis was selected by running PhyML with GTR as substitution model and with-or-without Gamma parameter and a proportion of invariable sites (Guindon and Gascuel 2003). The optimal substitution model based on the Bayesian Information Criterion (BIC) score (GTR+G) was determined using TOPALI 2.5 (Milne et al., 2009) and used in maximum likelihood analysis using RaxML with 10,000 bootstrap runs (Stamatakis 2014).

To identify branches that potentially contain signatures of positive selection among SLSP sequences we used maximum likelihood estimates of the dN/dS ratio (ω) for each site (codon) along protein sequences. A specific lineage (branch) was tested independently for positive selection (ω > 1) on individual sites by applying a neutral model that allows ω to vary between 0 – 1 and a selection model that also incorporates sites with ω > 1 using the software codeml implemented in PAML 4.4 (Yang 2007). Statistical significance was determined with a likelihood ratio test of these two models for the tested lineage.

To infer phylogenetic relationship of identified putative Entomophthoromycotina SLSPs with fungal SLSPs from other invertebrate-associated ascomycete fungi, two approaches were used. First, all S8A SLSPs in the MEROPS database from the insect-pathogenic genera Metarhizium (n = 240), Cordyceps (including Beauveria) (n = 44), Ophiocordyceps (n = 11), and the nematode-trapping genera Arthrobostrys (n = 33), and Monachrosporium (n = 8), were clustered by sequence identity with the identified SLSPs from Entomophthoromycotina (n = 152) using the MCL approach previously described. Second, from the three identified SLSP groups (A, B, and C) within Entomophthoromycotina, 10 representative group A SLSPs and all sequences from group B (n = 11) and group C (n = 11) were searched against the entire S8A MEROPS database (blastp, e-value = 1e-10). The non-redundant top-ten hits from this search (n = 149) were aligned with the Entomophthoromycotina query SLSPs (n = 32) and representative invertebrate-associated SLSPs from the clustering analysis (n = 350) using MAFFT. Unreliable columns with a Guidance2 score below 0.80 in the multiple sequence alignment were removed (Penn et al., 2010). The optimal protein substitution model

Figure 2 Active site and domain co-occurrence variability of the three Tribe-MCL clusters identified among 152 Entomophthoromycotina subtilisin-like serine proteases. The columns DTG, GHGTH, and SGTS represents the closest amino acid sequence for each of the amino acids from the DHS catalytic triad. A. Amino acid alignment of the active site residues for the three identified groups (A-C) of SLSPs within Entomophthoromycotina. Accession codes are color coded as: Orange – C. incongruus, Blue – E. muscae, and Purple – P. formicae. B. Sequence motifs of the active site residues for each group.
based on the Bayesian Information Criterion (BIC) score (\(LG+i\)) was determined using ModelFinder (Kalyaanamoorthy et al., 2017) and used in maximum likelihood analysis using RaxML with 100 bootstrap runs (Stamatakis 2014).

**Data availability**

All data used are available through the US Department of Energy Joint Genome Institute MycoCosm genome portal (http://jgi.doe.gov/-fungi), or has been previously published (Malagocka et al., 2015; Chibucos et al., 2016; De Fine Licht et al., 2017). SLSP sequence data and Tribe-MCL outputs analyzed in this manuscript are provided as a zip-compressed single file ArnesenEtalSupplementaryData.zip in the supplementary material. Supplemental material available at Figshare: https://doi.org/10.25387/g3.6949037.

**RESULTS**

We identified 154 SLSP sequences from six fungi in the subphylum Entomophthoromycotina: *E. muscae* (n = 22), *P. formicae* (n = 6), *B. meristosporus* (n = 60), *C. thromboides* (n = 18), *C. coronatus* (n = 36), and *C. incongruus* (n = 12). Close inspection of the active site residues revealed two *C. incongruus* sequences (Ci7229 and Ci12055), which contained the active site DHS residues in the motifs Asp-Asp-Gly, His-Gly-Thr-Arg, and Gly-Thr-Ser-Ala/Val-Ala/Ser-Pro characteristic of the S8B subfamily of S8 proteases. These two sequences also

Figure 3 Mid-point rooted maximum likelihood phylogeny calculated with RAxML and based on a (479 amino acid) alignment of 413 protein subtilisin-like serine protease sequences, which belonged to group C in the Tribe-MCL analysis (see text for details). Bootstrap values >70 from 1000 iterations are shown for non-terminal deeper nodes.
Subtilisin-like serine proteases (SLSPs) have many roles in fungal biology and are known to be involved in host–pathogen interactions. Independent expansion of copy number and diversification of SLSPs is widespread among animal pathogenic Dikarya (Ascomycota and Basidiomycota) (Li et al., 2017). The repeated expansion of SLSPs among the generalist insect–pathogenic hypocrealean fungi has been interpreted as an adaptation to enable infection of insect hosts (Muszewska et al., 2011), whereas comparatively little is known about
the evolution and diversification of SLSPs among the early diverging fungal clades. To understand the evolution of SLSPs among the vertebrate and arthropod pathogenic fungi in the subphylum Entomophthoraomycotina, we searched available genomic and transcriptomic sequence data to identify all Entomophthoromycotina genes with SLSP domains. We found 154 Entomophthoromycotina SLSPs, of which two copies were classified as S8B kexin SLSPs. The remaining 152 S8A SLSPs were clustered by sequence similarity and compared by phylogenetic analysis to show that the majority of the SLSPs (n = 130) are similar to and cluster together with “classical” proteinase-K-like fungal S8A SLSPs (Figure 1). A statistical test for a significant expansion of SLSP copy number among the insect-pathogenic Entomophthoromycotina was not explicitly performed in the present analysis due to uncertainty of total gene numbers from transcriptomic data sets of the specialist insect-pathogens E. muscae and P. formicae. In the sampled transcriptomes, the number of transcripts is likely larger than the genome gene count due to splice variants, post-transcriptional modifications, and allelic variants assembling into multiple transcripts per gene. In addition, the assembled transcripts only reflect actively transcribed genes expressed in the sampled conditions and time points, and may underestimate the actual number of genes. These confounding factors impact the estimated number of genes and make quantitative comparative analyses of gene family size between transcriptomes unreliable.

We did identify 11 SLSPs that cluster together with 402 un-annotated or Bacillopeptidase F-like SLSPs primarily from bacteria and Oomycetes.

Figure 6 Maximum likelihood phylogeny calculated with RAxML and based on a (276 amino acid) alignment of 500 protein subtilisin-like serine protease sequences. SLSPs within Entomophthoromycotina from group B and C, and 10 representative SLSPs from group A are included together with the 10 most similar SLSPs in the MEROPS database for each group A, B, and C, respectively. To show the relationship between the three Entomophthoromycotina SLSP groups, representative SLSPs from insect-pathogenic and nematode-trapping fungi were included (see text for details). Bootstrap values >70 from 100 iterations are shown for non-terminal nodes. The tree is rooted with the S8B subfamily type Kexin from S. cerevisiae (UniProt: P13134)
(Figure 3), but also including two fungal protease sequences from the early-diverging Cryptomycota R. allomyces and microsorum. M. daphnia lineages (Figure 4). These observations remained consistent even when exploring variation in the inflation parameter, which controls the “tightness” in the cluster analysis. The entomophthoralean and oomycetous S8A SLSPs form separate clades within this 402-sequence cluster of primarily bacterial proteases. Instead, the 11 Entomophthora/SLSLs form a distinct lineage together with the two proteases from R. allomyces and M. daphnia (Figure 3, 6). This indicates that within this group, the Entomophthorales and Oomycetes SLSPs both share a more recent common ancestor with independent bacterial proteases and the 11 entomophthoralean proteases, together with the two SLSPs from Cryptomycota and Microsporidia, are a unique group of proteases exclusive to some of the early diverging fungal lineages.

Functional annotation indicates apparent protease activity based on sequence similarity, but molecular function of the novel 11 SLSPs in group C is unknown. Eight of these SLSPs possess a signal peptide that suggest external secretion and thus indicative of a function on the immediate environment of the fungi, whereas the remaining three SLSPs might not be secreted or represent incomplete sequence models. Apart from the canonical protease S8 domain (PF00082), no other Pfam domains were found among this group C SLSPs. Searches against InterPro databases similarly did not reveal any other protein domains apart from the protease S8 SLSP domain (PRINTS: subtilisin serine protease family (S8) signature (PR00723), InterPro: peptidase S8, subtilisin-related (IPR015500), and ProSitePatterns: serine proteases, subtilase family (PS00138)). Notably, none of these SLSPs contain the proteinase inhibitor i9 domain (PF05922) often found among the classical protease K-like SLSPs (Muszewska et al., 2011). Extensive diversification of the amino acids immediately surrounding the active site residues in the DHS triad (Figure 2), further suggests that the new group C of SLSPs has evolved a different function than the “classical” fungal SLSPs in group A. Out of the six Entomophthoromycota species analyzed here, only three: C. incongruus, E. muscae and P. formicaceae within the order Entomophthorales contained members in the new group C SLSPs (Figure 1). The unequal phylogenetic presence of the group C SLSPs could be indicative of specific functions related to niche adaptation. The two insect-pathogenic fungi specialized on house flies (E. muscae) and wood ants (P. formicaceae) contain five and two of the novel group C SLSPs, respectively (Figure 1). However, the soil saprobe and opportunistic human pathogen C. incongruus also contains four group C SLSPs, which provide evidence that group C SLSPs are unrelated to host-specific evolution of the specialist insect-pathogenic entomophthoralean fungi.

The novel group C SLSPs were absent from the sequenced genome of C. thromboideis, implying that the uneven taxonomic presence and absence of particular SLSPs within Entomophthoromycota taxa are unlikely to be due to sequencing or sampling artifacts of data. This is further supported by the absence of group C SLSP S8A-proteases in Basidiomycota or Ascomycota (Ascomycota) (Figure 3). However, the genomes of ascomycete insect-pathogenic hocuspocus and nematode-trapping fungi within Ascomycota contain different SLSP’s missing in Entomophthoromycota (Figure 5). The present analysis shows that the two major groups of insect-pathogenic fungi within Ascomycota and Entomophthoromycota contain a similar complement of SLSPs, but each clade also possesses unique sets of these proteases that most likely evolved independently (Figure 6). Further studies including genomic comparisons of the host-specific insect-pathogenic fungi within Entomophthoromycota will likely shed interesting new light on the gene content of these early diverging fungi (De Fine Licht et al., 2016). The presence of unusual genome organization, polyploidy and large genomes in many host-specific insect-pathogenic species within Entomophthoromycota has previously been a hindrance to genome sequencing (Gryganski and Muszewska 2014). However, this study highlights examples of many new proteins and enzymes that may be discovered through genome sequencing and data mining within Entomophthoromycota.

ACKNOWLEDGMENTS

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Work by JES and AG was partially supported by funding from the National Science Foundation (DEB-1441715) to JES. KV expresses her gratitude to generous financial support from CRC/TR 124 Pathogenic fungi and their human host: Networks of Interaction - FungiNet. (Z1 and A6) granted by the Deutsche Forschungsgemeinschaft. HHDFL was supported by the Villum Foundation (grant no. 10122).

LITERATURE CITED


Communicating editor: A. Rokas