In Silico Characterization and Molecular Evolutionary Analysis of a Novel Superfamily of Fungal Effector Proteins

Ioannis Stergiopoulos,1,2,3,4 Yiannis A.I. Kourmpetis,5,6 Jason C. Slot,2 Freek T. Bakker,7 Pierre J.G.M. De Wit,5,8 and Antonis Rokas*1,2

1Department of Plant Pathology, University of California Davis
2Department of Biological Sciences, Vanderbilt University
3Laboratory of Phytopathology, Wageningen University and Research Centrum, Wageningen, The Netherlands
4Centre for Biosystems Genomics, Wageningen, The Netherlands
5Laboratory of Bioinformatics, Wageningen University and Research Centrum, Wageningen, The Netherlands
6Functional Genomics Group, BioAnalytical Science Department, Nestec Ltd, Nestlé Research Center
7Biosystematics Group, Wageningen University and Research Centrum, Wageningen, The Netherlands
8Department of Botany and Microbiology, King Saud University, Riyadh, Saudi Arabia

*Corresponding authors: E-mail: istergiopoulos@ucdavis.edu; pierre.dewit@wur.nl; antonis.rokas@vanderbilt.edu.
†These authors contributed equally to this work.
Associate Editor: Matthew Hahn

Abstract

Most fungal plant pathogens secrete effector proteins during pathogenesis to manipulate their host’s defense and promote disease. These are so highly diverse in sequence and distribution, they are essentially considered as species-specific. However, we have recently shown the presence of homologous effectors in fungal species of the Dothideomycetes class. One such example is Ecp2, an effector originally described in the tomato pathogen *Cladosporium fulvum* but later detected in the plant pathogenic fungi *Mycosphaerella fijiensis* and *Mycosphaerella graminicola* as well. Here, using in silico sequence-similarity searches against a database of 135 fungal genomes and GenBank, we extend our queries for homologs of Ecp2 to the fungal kingdom and beyond, and further study their history of diversification. Our analyses show that Ecp2 homologs are members of an ancient and widely distributed superfamily of putative fungal effectors, which we term Hce2 for Homologs of *C. fulvum* Ecp2. Molecular evolutionary analyses show that the superfamily originated and diversified within the fungal kingdom, experiencing multiple lineage-specific expansions and losses that are consistent with the birth-and-death model of gene family evolution. Newly formed paralogs appear to be subject to diversification early after gene duplication events, whereas at later stages purifying selection acts to preserve diversity and the newly evolved putative functions. Some members of the Hce2 superfamily are fused to fungal Glycoside Hydrolase family 18 chitinases that show high similarity to the Zymocin killer toxin from the dairy yeast *Kluyveromyces lactis*, suggesting an analogous role in antagonistic interactions. The observed high rates of gene duplication and loss in the Hce2 superfamily, combined with diversification in both sequence and possibly functions within and between species, suggest that Hce2s are involved in adaptation to stresses and new ecological niches. Such findings address the need to rationalize effector biology and evolution beyond the perspective of solely host-microbe interactions.

Key words: adaptive evolution, birth-and-death, diversification, effectors, fungi, GH18 chitinases.

Introduction

Effectors are low molecular weight proteins that are secreted by bacteria, oomycetes, and fungi to manipulate and adapt to their hosts and their environment (Hogenhout et al. 2009). Unlike bacteria and oomycetes from which several effector families have been described (Jiang et al. 2006; Sarkar et al. 2006; Hajri et al. 2009), only very few homologous effector proteins are known from fungi. Indeed, one of the hallmarks of fungal effector proteins so far has been their high sequence divergence and species-specificity (Stergiopoulos and de Wit 2009). So far, Ecp6, a secreted effector protein from the tomato pathogen *C. fulvum* (class Dothideomycetes) was the only known fungal effector with homologs in several other fungal species. Homology in this case, however, is mainly based on the presence of LysM motifs in this protein, a domain that is widespread among fungi of diverse taxa and lifestyles (de Jonge and Thomma 2009). More recently, a comparative genomics study in Dothideomycetes showed the presence of highly divergent homologs of the *C. fulvum* Avr4 and Ecp2 effector proteins in the closely related banana pathogen *M. fijiensis* (Avr4 and Ecp2) and the wheat pathogen *M. graminicola* (Ecp2) (Stergiopoulos et al. 2010). Interestingly, both *M. fijiensis* and *M. graminicola* contain three highly divergent homologs (64–25% identity at the amino acid level) of *C. fulvum* Ecp2. Phylogenetic analysis of
these seven Ecp2s indicated clustering based on orthology rather than paralogy, suggesting that gene duplications took place prior to speciation (Stergiopoulos et al. 2010). Ecp2 is a 165 amino acid (aa) secreted protein that was originally identified as a virulence factor in C. fulvum, since disruption reduces virulence of the fungus on tomato plants (Lauge et al. 1997). Although the intrinsic function of this effector protein during pathogenesis is still unknown, it has been hypothesized that Ecp2 interacts with a host virulence target to induce necrosis and the release of nutrients from the host cells during infection (Stergiopoulos et al. 2010). Indeed, the closest homolog of the C. fulvum Ecp2 in M. fijiensis was shown to trigger necrosis in tomato plants, irrespectively of the presence of the cognate Cf-Ecp2 resistance gene that mediates Ecp2 recognition and subsequent resistance reactions in tomato (Stergiopoulos et al. 2010).

The identification of homologous effectors within the fungal class of Dothideomycetes suggests that, as in bacteria and oomycetes, core effector proteins that facilitate basic virulence functions on a diverse set of hosts are also present in fungi. Thus, differences in the effector repertoire amongst fungal species could reflect adaptation to specific hosts or cultivars (Stergiopoulos et al. 2010). Theoretically, both convergent evolution driven by the need to cope with similar environmental challenges and divergent evolution driven by shared ancestry between different species, could potentially shape the effector repertoire of fungal species. In this respect, two studies in oomycete species of Phytophthora spp. showed that their effector repertoire has been largely shaped by divergent evolution, where high levels of genetic diversity are driven by selection pressure imposed by the host(s) (Jiang et al. 2008; Jiang et al. 2006; Jiang et al. 2009). However, although a few studies have described the sequence diversity and distribution of effector alleles in populations of a particular species (Schuch et al. 2004; Ma et al. 2006; Stergiopoulos et al. 2007; Hajri et al. 2009; Stukenbrock and McDonald 2009), detailed studies addressing the presence of effector families in fungi and their evolutionary paths after speciation are mostly lacking (de Jonge and Thomma 2009; Chuma et al. 2011).

To address this issue, we systematically searched for the presence of putative Ecp2 effector homologs across the entire fungal kingdom and beyond, and analyzed their history of evolutionary diversification. We show that Ecp2 is a member of a novel and widely distributed within the fungal kingdom multigene superfamily, which we have now designated Hce2, for Homologs of C. fulvum Ecp2 effector. Detailed evolutionary analysis of this superfamily of putative effectors in fungi shows that it most likely originated in a Pezizomycotina ancestor and has subsequently diverged extensively, experiencing multiple lineage-specific expansions and losses that are consistent with the birth-and-death model of gene family evolution. We also identified within the Hce2 subfamily members with a unique fusion to fungal Glycoside Hydrolase family 18 (GH18) chitinases that show highest similarity to the Zymocin α- and β-subunits of the heterotrimeric (αβγ) killer toxin from the dairy yeast K. lactis, suggesting a functional linkage between the two domains. Thus, we propose that in addition to gene duplication and rapid diversification, new or enhanced specificities of putative effectors can be generated by acquisition of novel protein domains as well. In summary, we identify and characterize a novel superfamly in fungi of putative core effector proteins, shedding light on its origin and diversification by showing that gene duplication, rapid diversification and recombination with novel protein domains, all contribute in creating remarkably diverse putative effector molecules required for plant parasitism and likely adaptation to stressful environments as well. Based on our findings, we further propose that the biology and evolution of effector proteins and virulence traits in general should be studied in a broader ecological context that will allow us to fully understand microbial pathogenicity beyond the current perspective of host–microbe interactions alone.

**Materials and Methods**

A detailed description of methods applied is provided as supplementary material.

**Identification of Putative Ecp2 Homologs**

We searched for homologs of Ecp2 against the non-redundant protein database nr (GenBank) and a local database of 135 fungal genomes that was constructed from genomes available in public databases (supplementary fig. S1 and table S1, Supplementary Material online). Searches were performed using PSI-Blast (e-value cutoff: 10−5) iterated till saturation with the previously identified Ecp2 orthologs from C. fulvum (CAA78401.1), M. fijiensis (Mfij_52972, Mfij_60658, Mfij_198160), and M. graminicola (Mgra_104404, Mgra_107904, Mgra_106176) (Stergiopoulos et al. 2010). The reference set of seven Ecp2 orthologs, together with their five best BlastP hits against nr and our local database were further exploited as a training set for HMM built and motif discovery, using the HMMER (v3.0) (http://hmmer.janelia.org) software package and the GLAM2SCAN program (Frith et al. 2008) of the MEME suite (v.4.6.1) (Bailey et al. 2006), respectively. These programs were used to filter the list of PSI-Blast hits from false positives that might have been obtained during the PSI-Blast step. To maximize output, further BlastP and tBlastN searches were performed using Hce2 sequences from close related species as queries. Finally, local synteny was utilized to carefully examine genomic loci in closely related species that were likely to hold homologs overlooked or mis-annotated by the automatic annotations provided in public genome databases.

Pairwise similarity scores among retrieved Hce2 homologs were calculated using the Needleman-Wunsch algorithm (Needleman and Wunsch 1970), as implemented in the “Needleall” software package available from EMBOSS, while pairwise e-value scores were computed from an all-versus-all BlastP analysis, executed using a locally installed BlastALL application available from NCBI. The query database consisted of all 153 Hce2 proteins identified in this study. Matrices of pairwise similarity scores and BlastP e-values were generated using the MultiExperiment Viewer (MeV: v4.7.1) software (http://www.tm4.org/mev).
Protein sequence alignments were performed using the GLAM2 (Gapped Local Alignment of Motifs) algorithm (Frith et al. 2008) available by the MEME suite (Bailey et al. 2006). The parameters used were z 150, a 120, b 250, w 120 r 10, and n 25000, where z is the minimal number of sequences in the alignment, a, b, and w the minimum, maximum, and initial number of aligned columns, respectively, r the number of alignment runs and n the number of iterations without improvement for each run.

Phylogenetic Analyses
Bayesian inference (BI) and Maximum Likelihood (ML) methods as implemented in MrBayes (v3.1.2) (Ronquist, Huelsenbeck 2003) and RAxML (v7.2.8) (Stamatakis 2006), respectively, were used to infer the phylogenetic trees presented in this study and estimate clade support. For each BI analysis 10 independent runs were performed using a mixed amino acid substitution model. Every run consisted of 2,000,000 iterations, a burn-in of 2,000 iterations, four chains, and a sampling frequency of every 100 generations. Convergence was examined using Tracer (v1.5) (http://tree.bio.ed.ac.uk/software/tracer/) and sampled trees from all runs were concatenated into a single tree using the sumt function. For each ML analysis we performed five independent fastML bootstrap heuristic searches that began with parsimony-derived trees (Stamatakis 2006; Stamatakis et al. 2008). Each run consisted of 2,000 bootstrap replicates and trees were inferred under the DAYHOFF+I amino acid substitution matrix. A bootstrap consensus tree from all runs was obtained using the Extended Majority Rule as implemented in the program CONSENSE from the Phylip software package (v3.67) (Felsenstein 2005).

Maximum Likelihood Ancestral State Reconstruction and Reconciled Tree Analyses
Correlation between Hce2 numbers and species ecology was performed using pairwise comparisons (Read and Nee 1995; Maddison 2000) as implemented in the Mesquite software package (v2.73) (Maddison and Maddison 2010). All correlation analyses were performed with the “most pairs” option, under which the choice of which pairs of states in the two characters are to be compared is aimed to maximize the number of pairs compared, regardless of the states in the characters.

Reconciliation of the inferred phylogenetic Hce2 trees with the Ascomycota species tree was performed using Notung (v2.6) (Durand et al. 2006; Vernot et al. 2008). The species tree was deduced from the available published phylogenetic data (Fitzpatrick et al. 2006; James et al. 2006; Schoch et al. 2009). To limit the impact of error in the Hce2 phylogenies on the estimation of gain and loss events, all clades with <70% bootstrap support or 99% posterior probability were collapsed prior to reconciliation with the species tree. In all cases, a rearrangement step was followed after the initial reconciliation of the Hce2 phylogenies with the Ascomycota species tree, in order to minimize the number of duplication and loss events (Durand et al. 2006).

Detection of Local Synteny
The progressive MAUVE algorithm as implemented in the MAUVE software package (v2.3.1) (Darling 2004) was used to align genomic regions of 10 kb on either side of Hce2 genes (20 kb in total) and identify areas of local synteny among them. Default seed weight was set to 11 in order to increase the sensitivity of the alignment and all other parameters were left at default.

Tests for Selection and Diversification in Putative Functions
The pattern of molecular evolution at individual sites of the Hce2 sequences was investigated using the site models (M0, M1a, M2a, M3, M7, and M8) implemented in the CODEML program of the PAML software package (v4.4) (Yang 1997, 2007). Positive selection was inferred only when the models that allow for positively selected sites (M2a, M8) fitted the data better than their nested null models (M1a, M7, respectively) based on Likelihood Ratio Tests (LRT). If this was the case then positively selected sites were assigned based on posterior probabilities (>95%) calculated using the Bayes Empirical Bayes (BEB) and Naïve Empirical Bayes (NEB) estimation methods in models M2a and M8 (Nielsen and Yang 1998; Yang et al. 2005).

Functional divergence between groups of orthologous Hce2 proteins was detected on the basis of shifts in site-specific evolutionary rates (Type I functional divergence) and group-specific changes in amino acid biochemical properties (Type II functional divergence) after gene duplication, using DIVERGE (v2.0) (Gu 2001). The method uses a maximum likelihood approximation to measure the coefficient of functional divergence ($\theta$) (Gu 1999, 2001, 2006). Rejection of the null hypothesis that $\theta = 0$ in favor of $\theta > 0$ indicated functional divergence (Gu 1999, 2001). Detection of amino acid residues responsible for Type I and/or Type II functional divergence was based on the posterior probabilities calculated for each position in the alignment that indicates the likelihood of a site to contributing to Type I and/or Type II divergence between groups.

Results
In Silico Identification of Ecp2 Homologs
In silico sequence-similarity searching of 135 fungal genomes and the GenBank nr database (supplementary fig. S1 and table S1, Supplementary Material online), identified 153 putative homologs of Ecp2, including three pseudogenes (supplementary table S2, Supplementary Material online). Pairwise sequence comparisons indicated that most homologs were highly divergent, sharing ~20–40% of amino acid similarity. Despite the high sequence divergence among the identified homologs, pairwise similarities and blast-hit e-values computed in an all-versus-all BlastP analysis indicated an asymmetric but nearly fully connected network of query-hit pairs with e-values $\geq10^{-6}$ and >40% similarity (supplementary fig. S2, Supplementary Material online). As expected, e-values and similarity scores were higher between protein pairs from closely related species but dramatically
The high sequence divergence among members of the Hce2 superfamily and lack of similarity to any known protein in sequence databases suggests that they share few common features. These include a putative N-terminal signal peptide that is present in nearly all members, indicating that this is a superfamily of secreted proteins (supplementary table S2, Supplementary Material online). Further surveys of protein domains and sequence length distribution indicated a set of widespread features among these proteins that allowed us to group them in three classes (fig. 1A and supplementary table S2, Supplementary Material online).

Class I contains 117 cysteine-rich, small secreted proteins of ~80–400 amino acid long that match our current concept of an extracellular effector (Stergiopoulos and de Wit 2009). The modular architecture of proteins from this class is relatively simple consisting of a signal peptide and the so-called “mature” part of the protein that corresponds to the actual Ecp2 effector protein, hereafter referred to as “Ecp2 domain.” This domain has been deposited in Pfam as a new protein family (PF14856).

Class II contains just 8 proteins with a modular architecture similar to class I proteins, except that they are significantly longer (up to ~800 amino acids). Sequence alignments show that members of Class I and Class II appear homologous only with respect to the Ecp2 domain located at the C-terminus of the two protein sets.

Class III contains 28 proteins that show a composite architecture, in which the Ecp2 domain is fused to the C-terminus of fungal GH18 chitinases from Subgroup C (Seidl et al. 2005). The typical architecture of Class III proteins consists of a signal peptide, a substrate-binding segment assembled by two LysM peptidoglycan-binding domains (InterPro Acc. No: IPR002482) and a chitin-binding domain 1 (ChtBD1) (IPR001002), followed by the catalytic GH18 domain with chitinolytic activity (IPR001223), and then by the Ecp2 domain (fig. 2). The chitinase domain of class III proteins shows highest similarity to the Zymocin alpha (α)- and beta (β)-subunits of the heterotrimeric (αβγ) Zymocin killer toxin from the dairy yeast K. lactis (Butler et al. 1991). However, no sequence similarity was found between the independently encoded γ-subunit of the yeast killer toxin, in which toxicity resides (γ-toxin), and the Ecp2 domain. Class III proteins primarily contain two LysM domains but six proteins with a single LysM domain were also identified, as well as one class member from which this domain was absent (supplementary table S2, Supplementary Material online).

**Fig. 1.** (A) Hce2 proteins can be classified into three distinct classes based on their domain organization. ORFs are indicated as brown boxes and signal peptides as gray ones. Cysteine residues are shown as vertical yellow lines. The Ecp2 domain of Hce2 proteins that is homologous to the mature secreted Ecp2 effector protein from the tomato pathogen C. fulvum is shown as a light brown box. The four conserved cysteine residues (1–4) present in the Ecp2 domain are also shown, as well as their relative positioning in this domain. Class I Hces are small secreted proteins of ~80–400 amino acids in length, while class II Hce proteins have a similar modular architecture but are much longer. Class III Hces are multimodular proteins that show a composite architecture in which the Ecp2 domain is fused to a catalytic Glycoside Hydrolase family 18 (GH18) chitinase domain. A substrate-binding segment assembled by two LysM peptidoglycan-binding domains and a chitin-binding domain 1 (ChtBD1) are also parts of the multimodular architecture of class III proteins. Similar domain architecture is found in the heterotrimeric (αβγ) Zymocin killer toxin from the dairy yeast K. lactis and interestingly the chitinase domain of the class III proteins shows considerable similarity to the Zymocin alpha (α)- and beta (β)-subunits. Figure is not drawn to scale. (B) Hce2 sequence alignment with arbitrary insertions and deletions using GLAM2 showed that these proteins share similarity only in the Ecp2 domain and revealed the presence of at least four conserved cysteine residues (1–4), as well as a few additional amino acids within this domain.
incomplete sampling and the considerably lower taxonomic representation of some of these classes in sequence databases. All six members of the Hce2 superfamily identified in Basidiomycota were from the subphylum of Agaromycotina (29 species examined), although the absence of Hce2 genes in Pucciniomycotina (4 species examined) and Ustilaginomycotina (2 species examined) might also be due to the underrepresentation of genomes from these species in public databases. Finally, the highly discontinuous distribution of Hce2 superfamily members was also evident at the order, family and genus levels (supplementary fig. S3 and table S1, Supplementary Material online).

Within species, the number of Hce2 paralogs per genome varied from one to as many as 14, suggesting several rounds of gene duplications and losses (fig. 3). Most species had one (20 species), two (10 species), or three (10 species) Hce2 genes in their genomes, but a substantial increase in gene copy number was observed in species of Sordariaceae, including Neurospora discreta (9 genes), N. crassa (10 genes), N. tetrasperma (10 genes), and Sordaria macrospora (14 genes), as well as in the Sordariales Podospora anserina (8 genes) and the Basidiomycete Auricularia delicata (7 genes) (supplementary fig. S3 and tables S1 and S2, Supplementary Material online). Evidence of lineage-specific expansions and contractions of the Hce2 superfamily were evident, even when comparing close related species within the same genus. For example, the distribution of Hce2 genes in species of Trichoderma varied from one (T. reesei), to two (T. atroviride) or three (T. virens) copies per genome.

This highly variable distribution of Hce2 genes in the different fungal lineages suggests a complex and dynamic evolutionary history consistent with the birth-and-death model of gene family evolution (Nei et al. 1997; Nei and Rooney 2005). To infer ancestral gene copy number and identify key events in the evolutionary history of the Hce2 superfamily, we performed parsimony based ancestral character state reconstruction on a species tree phylogeny of Ascomycota that was constructed with reference to available published phylogenies. Our analysis shows that the Hce2 superfamily most likely arose in Pezizomycotina and that the most recent common ancestor had a single Hce2 (supplementary fig. S4, Supplementary Material online). The subsequent evolution of the superfamily consists of multiple independent episodes of expansion and contraction across different fungal lineages. For example, multiple independent rounds of gene duplications have resulted in a large-scale expansion in species of Sordariomycetes, whereas Hce2 genes have been largely lost in Eurotiomycetes. Similarly, within Dothideomycetes an expansion of the superfamily occurred in species of Capnodiales, whereas a contraction occurred in species of Pleosporales.

To test whether the distribution and lineage-specific expansions and contractions of the Hce2 superfamily were associated with the species ecology, we classified species as 1) plant pathogens, 2) pathogens or parasites of animals, human, insects, nematodes, or other fungi, and 3) non-pathogens (saprophytic), and used a phylogenetic correlation test (Read and Nee 1995; Maddison 2000) to evaluate the relationship between Hce2 numbers and lifestyle (fig. 4 and
correlation could be found between a particular lifestyle and gene expansion or contraction across fungal species ($P > 0.05$). However, Class III Hce2 proteins were mostly associated with non-plant pathogenic (46.5%) and saprophytic (35.7%) fungi rather than plant pathogens (17.8%), whereas 85% of Class I and II proteins were almost equally distributed between plant pathogens and saprophytes (supplementary table S3, Supplementary Material online). This proportional representation of Hce2 genes from the three classes across fungi with different lifestyles could reflect functional divergence and lineage-specific expansions and contractions related to the ecological niches that the different species occupy.

**Molecular Evolution of the Hce2 Superfamily**

To characterize the evolution of the Hce2 superfamily we examined the phylogenetic relations of the Ecp2 domain between its members, using Bayesian inference (BI) and Maximum Likelihood (ML) approaches. Tree topologies derived with BI and ML methods were nearly identical and had generally low clade support, especially for deep internodes (supplementary fig. S5 and S6, Supplementary Material online), suggesting that the history of the Hce2 superfamily cannot be confidently resolved beyond the level of closely related species. This lack of support is likely due to the limited phylogenetic signal contained in the short but highly divergent aa sequence alignment of Hce2 superfamily members. Perhaps the only exception is the robust support (100% BI posterior probability and 85% ML bootstrap support) for the internode connecting Basidiomycota and Ascomycota. In contrast, the topology of the shallow internodes is largely consistent with the consensus fungal species phylogeny and is generally fairly strongly supported (>65% ML bootstrap support and >95% BI probability). Shallow internodes show extensive clustering in the Hce2 superfamily based on orthology rather than paralogy across the entire phylogeny, suggesting that gene duplications occurred prior to speciation. Exceptions are a weakly supported clade of three Hce2 paralogs from the plant pathogen *M. graminicola*, a well-supported two-protein clade from *S. macrospora*, and a six-paralog clade from the Basidiomycete *A. delicata*. These duplications are likely recent, indicating that “birth” is still an ongoing process in the evolution of the Hce2 superfamily.

Both duplication and loss were frequently inferred in clades of close relatives, thus supporting a mode of birth-and-death evolution for the Hce2 superfamily. To obtain a more thorough picture of the pattern of birth-and-death events, we performed phylogenetic reconciliation analysis of the Ecp2 domain topology with the Ascomycota species tree (fig. 5). The analysis identified 25 duplication (D) and 54 loss events (L) ($D/L$ Score = 91.5), whereas reconciliation after rearrangement of the Hce2 topology to obtain the minimal score for duplication and losses, identified 23 duplication and 37 loss events ($D/L$ Score = 60). These numbers should be considered as approximate since the history of alternative events from reconciliation assuming equal weights for duplications and losses indicates a range of 29–37 duplications and 23–31 losses ($D/L$ Score = 60.0), respectively. Nevertheless, the analysis exposed once more the dynamic evolutionary history of the Hce2 superfamily, consisting of recurrent episodes of...
duplications and losses taking place across the entire Pezizomycotina clade and which continue to occur even to the present day. The reconciliation analysis also showed a large expansion of the Hce2 superfamily in the Sordariales clade as a result of 11 duplication events, five of which took place before the divergence of the Chaetoniaceae, Lasiosphaeriaceae and Sordariaceae families. Since then, the Hce2 superfamily has continued to expand in Sordariaceae (N. discreta, N. crassa, N. tetrasperma, and S. macrospora), as indicated by the three recent duplication events suggested to have taken place in S. macrospora, but has contracted in Chaetoniaceae (Chaetomium globosum, Sporothrix thermophila, Thielavia terrestris). Similar observations can be made for other fungal lineages as well, for example the expansion of the Hce2 superfamily in Glomerellales (including Verticillium and Colletotrichum spp.).

Despite the fact that the deeper nodes of the Hce2 phylogeny are not well supported, it appears that there are two major groupings, of Class I/Class II proteins on the one hand and Class II proteins on the other (supplementary fig. S5 and S6, Supplementary Material online). Both BI and ML methods consistently reproduce the class III protein clade, although with weak support, thus suggesting a monophyletic origin for this clade. A pattern of interspecies clustering was also evident within this clade but a general arrangement according to the consensus fungal species phylogeny was not as clear.

Fig. 4. The number of Hce2s per genome is not correlated with the fungal species ecology. Maximum parsimony-based ancestral character state reconstruction tracing variation in Hce2 numbers in species of Pezizomycotina (A) and their pathogenic lifestyle (B). The Pezizomycotina species topology upon which the multistate character states were traced was assembled manually with reference to available published phylogenies. Branches on the left panel are color-coded based on the present (end nodes) and inferred (deeper nodes) numbers of Hce2 genes in each species or its ancestor. In a similar way, branches on the right panel are color-coded according to the species life style as plant pathogens (green nodes), human, and/or animal pathogens (blue nodes) and non-pathogenic or saprophages (orange nodes). In both panels, branches supported by more than one character states are indicated by two-color patterns. For easier interpretation of the image on the left panel the number of states (range of Hce2 gene numbers) on each branch is also indicated, as well as gain (G) and losses (L) along specific lineages. The maximum parsimony reconstruction shows that the Hce2 superfamily most likely arose in Pezizomycotina and followed a complex evolutionary path with multiple independent expansions and contractions across the different fungal lineages that are consistent with a birth-and-death model of evolution. However, no specific correlation could be found between the fungal species ecology and Hce2 gene numbers in the different fungal species.
evident as with Class I and Class II proteins. This resulted in a number of apparent, albeit unsupported, differences between the Hce2 and the species phylogeny in this clade that could be either an artifact of the fairly unresolved phylogenetic relationships or indicative of extensive gene loss in an array of fungal lineages. 

**Birth-and-Death Evolution within Specific Fungal Lineages**

To better understand the tempo and mode of evolution within the Hce2 superfamily, we examined in greater detail Ecp2 proteins from Sordariaceae (Sordariomycetes) and Arthrodermataceae (Eurotiomycetes) (presented in Supplementary Material online). We chose these fungal families because 1) each represented a different order of Ascomycota from which Hce2 genes were identified, 2) several different highly similar Hce2 ortholog groups were present in at least four closely related species of the family that would allow the use of codon-based models for the detection of changes in evolutionary pressure in members of the Hce2 superfamily, and 3) they exhibited other interesting features, such as a particularly large lineage-specific expansion (Sordariaceae),
and, unlike other fungal families, 4) they comprised mainly of Class III chitinase-associated proteins (Arthrodermataceae).

**Sordariaceae**

Four species were analyzed from this family, each of which contained a large number of Hce2 paralogs in their genome (N. crassa: 10 paralogs, N. discreta: 9 paralogs, N. tetrasperma: 10 paralogs, and S. macrospora: 14 paralogs). BI phylogenetic analysis of the Ecp2 domain of Sordariaceae Hce2 proteins identified 10 distinct groups of orthologs (Groups I–X; fig. 6A). All but one of the ortholog groups (Group IV), were supported by very high posterior probability clade support values. Groups I and II contain exclusively class III proteins and form a robustly supported (100% BI posterior probability) clade. The sister relationship of these two groups supports our previous assumptions (see supplementary results, Supplementary Material online) on the early radiation and diversification of class III proteins. Nine of the groups contained orthologs from all four species, with only group I containing orthologs from N. crassa and N. tetrasperma, indicating the loss of this ortholog in N. discreta and S. macrospora. Indeed, reconciliation of the Sordariaceae Hce2 gene tree with the Ascomycota species tree confirmed this loss (L) and further suggested the occurrence of 13 duplications (D) in the evolutionary history of the Hce2 superfamily in Sordariaceae (D/L Score = 21.5 D = 13, L = 2, 99% BI posterior probability Edge Weight Threshold) (fig. 6B). Four of these duplications were associated exclusively with S. macrospora, thus contributing to the expansion of this superfamily in this species. The other nine duplications all took place in the four species’ common ancestor.

It should be noted that the placement of sequence Ntet_B5717 from N. tetrasperma differed between the phylogeny constructed using only the sequences from the four Sordariaceae species and the phylogeny using all Hce2 superfamily sequences. In the Sordariaceae-specific phylogeny this gene is placed in group IV, while in the overall phylogeny it appears together with sequences from group II. However, by analyzing syntenic information within 10 kb regions (supplementary fig. S7, Supplementary Material online) of the Sordariaceae Hce2 genes it was finally determined that Ntet_B5717 should be regarded as orthologous to group IV Hce2 proteins. The only genomic loci that were not syntenic to any other were the ones flanking genes CB157005.1, CB157631.1, CB157148.1, and CB154907.1, which were inferred by the reconciliation analysis to be the products of gene duplications within S. macrospora, thus confirming these results and indicating a rapid dispersal of the Hce2 genes in genomes after duplication.

Gene duplication can generate new genes with novel or altered functions and functional divergence of paralog genes is a major factor promoting their retention in the genome. In such cases, adaptive evolution or relaxed selection in early stages after duplication plays a critical role towards the functional diversification of the two copies, whereas at later stages purifying selection can ensure the maintenance of their distinct functions. To investigate changes in evolutionary pressure on members of Hce2 superfamily in Sordariaceae that could be indicative of diversification in putative functions, we conducted a series of likelihood ratio tests between alternative models of codon evolution for each group of orthologs. Because the per-site frequency of synonymous substitutions was saturated (dS > 1.0) on some internal branches of the Sordariaceae Hce2 gene phylogeny, thus preventing accurate estimation of the ω (dN/dS) ratio, we restricted our analysis within the orthologous groups of Hce2s. Group I was excluded from the analysis as it consists of only two members. Estimates of ω under the Model M0 showed low overall values ranging from 0.06 to 0.23 depending on the group examined (table S3). Interestingly, LRT of site-specific models showed (irrespective of dS saturation) that for all nine groups examined, model M3 fitted the data significantly better than model M0, thus suggesting heterogeneous selection pressure among codon sites that could have promoted divergence of the Hce2 superfamily members across clades. For groups III, V, VIII, and IX all three classes of ω under M3 had values <1, suggesting purifying selection. However, for groups II, IV, VI, VII, and X, a small proportion of sites (1–3%) were predicted to show ω > 1. To identify the specific codon sites evolving with ω > 1, we further tested whether models M2a and M8 that allow for positively selected sites fit the data better than their null models M1a and M7, respectively. In eight of the nine groups, LRTs between models M1a and M2a and between M7 and M8 were not significantly different, suggesting it is unlikely that these are sites under strong positive selection. In contrast, in group VII all three LRT tests showed that models M3, M2a, and M8 fitted the data significantly better than their nested null models M0, M1a, and M7, respectively, leading to the identification of four sites undergoing positive selection. Two of these sites, though, were consistently detected with significant posterior probabilities (P > 95%), and they involved transitions from Alanine-to-Lysine or vice-versa and from a Tryptophan-to-Leucine and Threonine (supplementary fig. S8, Supplementary Material online). These sites were located in highly variable regions of the overall alignment. Nevertheless, despite the presence of two positively selected sites in Group VII, purifying selection is the dominant evolutionary force shaping the evolution of the Hce2 ortholog groups in Sordariaceae.

The codon-based analysis is restricted to within group variation and does not provide any information for changes in evolutionary rates between groups. Furthermore, it is unlikely that positive selection would have operated on large numbers of sites of the duplicated Hce2 genes over prolonged periods of time. Rather, diversification in putative functions might have happened soon after gene duplication. Changes in functional constraints at individual sites could reflect a change in rates of evolution (known as Type I divergence) and result in amino acid sites that are highly conserved in one group of paralogs but highly variable in another (Gu 1999, 2001). Alternatively, these changes in functional constraints at individual sites could reflect a change in amino acid properties (known as Type II divergence) and result in amino acid sites that, although conserved within each group of paralogs, when compared between such groups differ radically in their biochemical properties (Gu 2006). To test for evidence of
Type I and/or Type II functional divergence in the paralog groups of Hce2 proteins in Sordariaceae, we conducted all possible pairwise comparisons between Groups II-to-X and estimated their coefficient of Type I ($\#_I$) and Type II ($\#_II$) functional divergence (Gu 1999, 2001, 2006).

Estimates of $\#_I$ for almost all examined pairs were moderate ($0.5 \leq \#_I \leq 0.8$) to high ($\#_I > 0.5$), suggesting significant shifts in evolutionary rates between the groups (supplementary table S4, Supplementary Material online). However, all pairwise comparisons of group IV with the rest of the groups resulted in $\#_I = 0$, indicating absence of Type I functional divergence. This is surprising given that Group IV is the most divergent in sequence when compared to all other groups (supplementary table S4, Supplementary Material online). It is possible that this lack of significance is an artifact caused by the high degree of sequence divergence within Group IV, which is almost 2-fold higher (0.43%) as compared to all other groups (0.18% on average). From the 28 pairwise comparisons, $\#_I$ was significantly higher than 0 in 24 ($P < 0.05$). Consideration of the site-specific posterior probabilities did not identify any specific protein domains with elevated rates of evolutionary divergence. Rather a large number of amino acid sites distributed across the entire alignment display high $\#_I$ Type I divergence (supplementary fig. S9A, Supplementary Material online). Evidence of Type II functional divergence was also present. Of the 36 comparisons 28 were statistically significant. Again, comparisons involving group IV were not significant. Type II functional divergence detected a large number of amino acid sites distributed across the alignment with radical and statistically significant changes ($\#_II > 1$) in their biochemical properties (supplementary fig. S9B, Supplementary Material online). Collectively, the data for Type I and Type II functional divergence provide strong evidence for early diversification in putative functions, both in evolutionary rates and amino acid properties, between duplicated genes in the Hce2 superfamily.

**Discussion**

Hce2 Is an Ancient Superfamily within Fungi

The taxonomic circumscription of the Hce2 superfamily within fungi and the presence of multiple paralogs per species suggest that this superfamily originated and diversified within the fungal kingdom. Furthermore, the presence of the Hce2 superfamily in both Ascomycota and Basidiomycota indicates that its origin predates the divergence of Dikarya. However, this scenario requires a complex pattern of gene maintenance and loss along many intervening lineages and taxa, implying, on one hand, the maintenance of Hce2 genes for long periods of time in Pezizomycotina (Ascomycota) and Agaricomycotina (Basidiomycota) and, on the other, loss of Hce2 genes on numerous independent lineages (e.g., Saccharomycotina). Given the clear overrepresentation of these genes in Pezizomycotina and the presence of multiple paralogs in species of this subphylum as compared to Agaricomycotina, an alternative hypothesis is that the Hce2 superfamily originated prior to, or early in the evolution of Pezizomycotina, from where it was horizontally transferred to Agaricomycotina. This hypothesis relies solely on the taxon...
distribution of this gene superfamily and it is not supported by the observed phylogeny, and should therefore be treated with caution until more extensive and uniform sampling of genomes across fungi has been achieved.

Recurrent Gene Birth-and-Death Defines the Evolution of the Hce2 Superfamily

The Hce2 superfamily is characterized by frequent independent duplications and losses, even among closely related species, consistent with the birth-and-death model (Nei et al. 1997; Nei and Rooney 2005). Clear evidence of birth-and-death evolution in the Hce2 superfamily was found by examining lineage-specific variations in gene copy numbers of putative orthologs from closely related species within the same fungal genus or family. In almost all such cases examined we observed 1) high rates of copy number variation even among closely related species, 2) clustering of Hce2 genes based on orthology rather than paralogy, 3) conservation of gene structure among orthologous Hce2 genes, and in some cases, 4) presence of pseudogenes. All these are consistent with the birth-and-death model of evolution (Nei et al. 1997; Nei and Rooney 2005).

The most likely scenario for the evolution of the Hce2 superfamily in Ascomycota is that a single gene was present in a common ancestor of Pezizomycotina, which independently underwent multiple expansions and contractions in a lineage-specific manner. In this respect, species of Sordariomycetes have experienced a large-scale expansion in Hce2 numbers since their divergence from Eurotiomycetes. In contrast, Hce2 genes were mostly lost from species in the order of Eurotiomycetes, whereas in Dothideomycetes they underwent both gains and losses. The ecological pressures driving such lineage-specific expansions and contractions are unclear, since no correlation was found between Hce2 numbers and the species ecology. The largest lineage-specific expansion took place in Sordariaceae, which is rather surprising given that in N. crassa, and presumably its close relatives as well, the fixation of duplicated genes is greatly reduced by the Repeat Induced Point mutation (RIP) mechanism that largely prevents evolution through gene duplication (Galagan et al. 2003). The presence of so many paralogs that have escaped the effects of RIP indicates that these are ancient copies, whose duplication and diversification predate the emergence of RIP (Galagan et al. 2003).

Both Divergent Evolution and Purifying Selection Have Shaped the Evolution of Hce2 Genes

Gene duplication is a major force of genomic innovation, constantly creating new genes, whose retention in the genome depends on their functional diversification (Lynch and Conery 2000; Conant and Wolfe 2008). In such cases, positive or relaxed selection in one of the two copies in early stages after duplication drives functional diversification, whereas at later stages purifying selection maintains the newly acquired function(s) (Lynch and Conery 2000). This seems to be the case for duplicated Hce2 genes as well. Our analyses on the Hce2 superfamily support a model where Hce2 duplicates diversify rapidly early after duplication as a result of positive or relaxed selection, most likely acquiring new functions as well. Indeed, the high levels of sequence divergence among paralogs and their dispersal in the genome are indicative of an accelerated rate of evolution after duplication. Moreover, tests for Type I and Type II functional divergence between almost all pairs of paralogous Hce2 groups in Sordariomycetes and Arthrodermataceae indicate that the newly generated paralogs diversified in their putative functions early in their evolution. Finally, our molecular evolutionary analyses show that, once diversified, Hce2 genes within ortholog groups are under purifying selection. Thus, any changes that contributed to the diversification in putative functions were likely fixed early in their evolution.

Hce2 Proteins Are Putatively Involved in Stress-Responses

The driving force behind the patchy distribution of Hce2 genes in different fungal species is unknown but likely relates to their yet uncharacterized intrinsic function. While gene deletions and duplications are frequently observed in multifamily genes, the diverse and large-scale lineage-specific expansion and losses observed here suggest that Hce2 genes perform taxon-specific roles providing conditional advantages in specific environmental niches. One possibility is that Hce2 genes are gained or lost after speciation events according to niche-specific selection pressures, thus fine-tuning environmental and/or parasitic fitness (Conant and Wolfe 2008).

Such a dynamic response to the environmental stimuli involving high rates of gains and losses is frequently observed in stress-response genes, such as the ones implicated in interactions with other organisms, adaptive immunity or pathoadaptation to new hosts (Ota, Nei 1994; Nei and Rooney 2005; Wapinski et al. 2007; Korb et al. 2008; Pujol et al. 2008). Effector proteins in particular from fungi, bacteria and oomycetes are known for their rapid diversification and accelerated rates of birth-and-death evolution that lead to highly discontinuous distributions, even between taxa belonging to the same species complex (Jiang et al. 2006; Stergiopoulos et al. 2007; Jiang et al. 2008; Stavrinides et al. 2008). The intrinsic function of Hce2 genes remains unknown but in the plant pathogenic fungi C. fulvum and M. fijiensis they may function as effector proteins that both promote virulence in susceptible hosts and induce R-gene-mediated resistance in resistant ones (Laube et al. 1997; Stergiopoulos et al. 2010). It is possible that Hce2 proteins from other plant pathogenic fungi have a similar role in promoting virulence as well. If this is the case, then the broad distribution of Hce2 genes in fungi pathogenic on a variety of hosts suggests that these genes are part of the pathogenic core that targets broadly conserved defense components (Stergiopoulos et al. 2010). As such, it is unlikely that they contribute directly to host specialization but rather facilitate pathogenicity on a wide range of hosts by providing basic virulence functions. Maintenance or loss of Hce2 genes following speciation and potentially host jumps would depend both on their role during pathogenesis and the
conservation of their virulence targets in different hosts, but also on the suite of cognate resistance genes that mediate immune responses in their hosts (Stergiopoulos et al. 2010). The preservation of multiple highly diverse paralogs per species would further suggest that the basic function of Hce2 proteins can be manipulated to fine-tune interactions with the host, either by targeting diverse defense components (and acting perhaps even synergistically) or by being expressed at different stages and levels during infection.

The study of fungal effector proteins has so far been approached from a predominately host-microbe interactions perspective (Morris et al. 2009). Pathogens, however, do not only have to infect a host in order to be evolutionarily successful but also need to compete with numerous other microbial species present in their environment and survive saprophytically for large periods of their life-cycle. Thus, both pathogen biology on the host and the environment in which the organism is living-in will strongly influence the evolution of a pathogen’s virulence (Wolinska and King 2009; Allen, Little 2011). In this study, a large number of Hce2 genes were identified from fungal species that are parasitic on human, animals, insects and other fungi or saprophytic. In such cases, the function of Hce2s could have been adapted to meet their demands in these environments, including activities as toxigenic peptides in antagonistic interactions with other microorganisms. The fusion of the Ecp2 domain of some Hce2 proteins to fungal GH18 chitinases and their association with the Zymocin killer toxin from yeast favors this hypothesis. Overall, the presence of Hce2s outside plant pathogens and their implication in possible antagonist interactions with other microorganisms in their environment emphasizes the necessity to study the biology and evolution of virulence traits and of putative effectors beyond the current perspective of host-microbe interactions and in the broader context of pathogen ecology. Thus, a more holistic understanding of microbial ecology is needed in order to fully understand effector biology and microbial pathogenicity in general. In this respect, studies in human pathogens have shown that many virulence traits have dual roles in parasitic and environmental fitness and their evolution has been shaped by forces outside the narrow context of human-pathogen interactions (Pallen and Wren 2007; Morris et al. 2009). The recent characterization of the Type VI secretion system (T6SS) in bacteria, for example, has led to the identification of T6SS effectors that are not only important for pathogenesis but also mediate competitive interactions with other bacteria (Jani and Cotter 2010). Similarly, a number of virulence traits are shared between plant pathogenic fungi and saprophytes (e.g., LysM effectors) or have dual roles in parasitic fitness (e.g., several mycotoxins) and ecological survival (e.g., melanin) (Nosanchuk and Casadevall 2003; Berestetskiy 2008; de Jonge and Thomma 2009).

Hce2 Genes Show Unique Associations with Fungal GH18 Chitinases

GH18 is an ancient family of chitinases, widely distributed in Archaea, Bacteria, and Eukaryota, including humans and fungi, whose members catalyze the hydrolysis of chitin, a structural component of the cell wall of fungi and the exoskeleton of arthropods. In fungi, GH18 chitinases are involved in diverse physiological functions, including mycoparasitism through lyases of the antagonist’s cell wall. The chitinase domain of the class III proteins shows considerable similarity to the Zymocin alpha (α)-subunit of the yeast killer toxin, a heterotrimeric (αβγ) eukaryotic rRNase toxin from the dairy yeast K. lactis that inhibits cell cycle proliferation by arresting the G1 phase in Saccharomyces cerevisiae (Stergiopoulos et al. 2010). Toxicity of Zymocin resides in the intracellularly targeted γ-subunit (γ-toxin), the import of which into host cells is facilitated by the larger α- and β-subunits that act from the cell exterior to promote contact with the cell-surface and association with the plasma membrane (Stergiopoulos et al. 1990; Jablonowski et al. 2001). The α-subunit is an exochitinase that binds to chitin, thereby facilitating docking to and subsequent chitinolysis of the cell-wall chitin, whereas the β-subunit is predicted to be associated with the cell-membrane. Based on the presence of similar domains in class III proteins and Zymocin and the high sequence similarity of the chitinase domains, it is tempting to speculate that the Ecp2 domain of fungal GH18 chitinases and the γ-subunit of the yeast killer toxin have an analogous role in antagonistic interactions with other microorganisms, despite the lack of any sequence similarity between these two domains. Although testing this hypothesis requires functional assays, it is intriguing that both toxigenic peptides have recruited the same set of proteins for delivery into their host cells. The fusion of the Ecp2 domain to GH18 chitinases further suggests that, in addition to gene duplication and rapid diversification, new effector specificities in fungi might be generated by the recruitment of unrelated protein domains. This domain recruitment by putative fungal effectors to enhance or alter their function is a novel finding for fungal effectors that changes our traditional views on these proteins and reveals their evolutionary and structural plasticity.

Supplementary Material

Supplementary tables S1–S5, figures S1–S14, Materials and Methods, and Results are available Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work used resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, TN, USA. This work was supported by the Searle Scholars Program (to A.R.) and the National Science Foundation (DBI-0805625 to J.C.S. and DEB-0844968 to A.R.). I.S. was in part financially supported by the NWO-ERA-PG project ARAPGFP/06.002A entitled “RLP-and RLK-mediated innate immune responses in Arabidopsis and tomato triggered by pathogen-associated molecular patterns and avirulence factors”. This project was co-financed by the Centre for BioSystems Genomics (CBSG), which is part of the Netherlands Genomics Initiative, a Dutch organization
for scientific research. P.J.G.M.deW. was co-financed by the Royal Netherlands Academy of Arts and Sciences.

References


