MYCOLOGY
2018, VOL. 9, NO. 3, 176–188
https://doi.org/10.1080/21501203.2018.1478333

Introgression and gene family contraction drive the evolution of lifestyle and host shifts of hypocrealean fungi

Weiwei Zhang[a,b], Xiaoling Zhang[a], Kuan Li[a], Chengshu Wang [c], Lei Cai[a], Wenying Zhuang[a], Meichun Xiang[a] and Xingzhong Liu[a]

[a]State Key Laboratory of Mycology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China; [b]University of Chinese Academy of Sciences, Beijing, China; [c]Key Laboratory of Insect Developmental and Evolutionary Biology, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

ABSTRACT

Hypocrealean fungi (Ascomycota) are known for their diversity of lifestyles. Their vital influences on agricultural and natural ecosystems have resulted in a number of sequenced genomes, which provide essential data for genomic analysis. Totally, 45 hypocrealean fungal genomes constructed a phylogeny. The phylogeny showed that plant pathogens in Nectriaceae diverged earliest, followed by animal pathogens in Cordycipitaceae, Ophiocordycipitaceae and Clavicipitaceae with mycoparasites in Hypocreaceae. Insect/nematode pathogens and grass endophytes in Clavicipitaceae diverged at last. Gene families associated with host-derived nutrients are significantly contracted in diverged lineages compared with the ancestral species. Introgression was detected in certain lineages of hypocrealean fungi, and the main functions of the genes located in the introgressed regions are involved in host recognition, transcriptional regulation, stress response and cell growth regulation. These results indicate that contraction of gene families and introgression might be main mechanisms to drive lifestyle differentiation and evolution and host shift of hypocrealean fungi.

ARTICLE HISTORY
Received 26 April 2018
Accepted 15 May 2018

KEYWORDS
Phylogenomic; evolutionary process; introgression; gene contraction; host shift

1. Introduction

Hypocreales (Ascomycota) containing nine recognised families and over 2600 species (Rogerson 1970; Kirk et al. 2008) is one of the most important orders in Ascomycota. Species within hypocreales have evolved various lifestyles including saprophytism, endophytism and parasitism on plants, insects, nematodes and other fungi (Berbee 2001). The evolution of plant and animal pathogens and the origin of the grass endophytes from insect pathogens in Clavicipitaceae were documented by multigene phylogenetic analysis (Spatafora et al. 2007; Sung et al. 2008). Generally, different families display distinct host associations. Nectriaceae includes numerous important plant pathogens such as *Fusarium* that cause serious plant diseases and economical losses (De Wolf et al. 2003; Summerell et al. 2011). Cordycipitaceae represented by *Cordyceps* spp. includes well-known insect pathogens and medicinal fungi (Sung et al. 2007; Zheng et al. 2011). Ophiocordycipitaceae also includes a large number of insect and nematode pathogens and medicinal fungi, such as *Ophiocordyceps sinensis, Hirsutella minnesotensis* and *Hirsutella rhossiliensis* (Jaffee and Zehr 1982; Chen et al. 2000). Meanwhile, Clavicipitaceae is composed of grass endophytes that benefit plants but impair grass-feeding animals (Clay 1988; White et al. 2003), as well as insect and nematode pathogens such as the *Metarhizium* spp. and *Pochonia* spp.

Vital impacts of hypocrealean fungi on agriculture, ecosystems and human life have led to a number of genomes being sequenced (Rogerson 1970). Most of the research on genomics in Hypocreales are mainly focused on gene function related to phylogeny, development and pathogenesis, and has revealed the sophisticated strategies associated with the adaption to various lifestyles (Klosterman et al. 2011; Rouxel et al. 2011). The subtilisins and chitinases, for example, have been shown to be involved in the pathogenesis of insects (Gao et al. 2011). Polysaccharide lyases (PLs) and glycoside hydrolases (GHs), enzymes involved in the breakdown of pectin and cellulose in plant cell walls play a role in the

CONTACT Meichun Xiang xiangmc@im.ac.cn; Xingzhong Liu liuxz@im.ac.cn
Supplemental material for this article can be accessed here.

infection of plants (Klosterman et al. 2011). However, the evolutionary history and potential mechanisms of lifestyle changes are not comprehensively understood at the genomic level. The availability of numerous genome sequences of hypocrealean fungi provides an opportunity to examine evolutionary mechanisms in Hypocreales.

Hybridisation and interbreeding between species can lead to adaptive introgression by transmitting beneficial alleles and has the potential to influence adaptation and speciation in a variety of ways, which can happen during either sympatric speciation or the secondary contact phase of allopatric speciation (Arnold 2004). Recent studies have revealed several mechanisms, including introgression, that can lead to adaptive divergence, especially in rapidly radiating groups (Pease et al. 2016). There are numerous striking examples of adaptive introgression in plants, animals and humans that illustrate functional introgressed loci contributing to ecologically and reproductively significant traits (Whitney et al. 2006; Song et al. 2011; Sankararaman et al. 2014; Lamichhaney et al. 2015). *Epichloë* spp., grass endophytes in Hypocreales, contain many species that have evolve through a complex process of hybridisation (Moon et al. 2004), indicating there is high possibility for the occurrence of introgression.

In order to study the evolutionary history and mechanism of cross-kingdom host adaptation of fungi in Hypocreales, a total of 45 sequenced genomes were selected, including two newly sequenced genomes, e.g. the nematode endoparasite *Hirsutella rhossiliensis* (Table S1) sequenced for this project and the pathogenic fungus *Clonostachys rosea* (Toledo et al. 2006; Zhang et al. 2008). The purposes of this study are to characterise the evolutionary patterns of lifestyle shifts in Hypocreales and to illustrate mechanisms that drive the evolution of diverse lifestyles and host shift.

2. Materials and method

2.1. Fungal strains and genome sequencing

Hirsutella rhossiliensis is a dominant parasite of juveniles of the soybean cyst nematode (SCN), *Heterodera glycines*) (Liu and Chen 2000). Strain OWVT-1 was isolated from SCN juvenile in Minnesota, USA and has shown biocontrol potential against nematodes. A single spore isolate of OWVT-1 was cultured on potato dextrose agar (PDA, BD^{TM}, New Jersey, U.S.A.) plate for 4 weeks and the mycelium was harvested for genomic DNA preparation using the CTAB/SDS/Proteinase K method (Möller et al. 1992). Whole-genome shotgun sequencing of OWVT-1 was performed using Illumina next generation sequencing technology. DNA libraries with 170, 500, 2 and 5 kb inserts were constructed and sequenced with an Illumina Genome Analyzer at the Beijing Genomics Institute (BGI, Shenzhen, China). The genome was sequenced to approximately 128-fold coverage and assembled using SOAP denovo (Li et al. 2009). Assembly yielded 3543 scaffolds and a genome size of 50.39 Mb.

2.2. Gene prediction and annotation

To accurately compare the gene numbers and gene distribution in each fungus, the gene structures were predicted for *H. rhossiliensis* as well as for all other fungi with the same algorithms, Evidence Modeler (Haas et al. 2008), using *Fusarium graminearum* sequence as a reference. Finally, functional predictions were performed by BLASTX search against a protein database and InterProscan searches against protein domain databases (Zdobnov and Apweiler 2001).

2.3. Orthology and phylogenomic analysis

Using HMM models, putative orthologs were identified against the BUSCO (Benchmarking sets of Universal Single-Copy Orthologs) database, as the highest full sequence HMM bit score with a minimum *E*-value of e^{-50}. Total 627 orthologous proteins were extracted and aligned with MAFFT (Kazutaka and Standley 2013). The program RAxML was used to create a maximum likelihood tree (Stamatakis 2006). The estimation of the evolution of fungal life-strategies was performed using RASP based on the results of RAxML (Yu et al. 2015).

The program BEAST v1.7.4 was used to estimate the divergence time between the compared species, using the orthologous protein sequences (Drummond et al. 2012). The maximum likelihood tree constructed in RAxML (above) was used as the phylogeny. The calibration point of the origin of Ascomycota estimated at 600 Ma was used to estimate divergence

times as soft constraints following a uniform limitation (Lücking et al. 2009).

2.4. Protein family classification and repeat analysis

Protein families of the whole genomes were classified by analysis of genes descended from a common ancestor. BLAST searching against the FUNCAT database and PHI (pathogen–host interaction) provided a global view of the gene functions (Ruepp et al. 2004; Winnenburg et al. 2006). Statistics for the gene abundance were performed by *t*-test and was corrected by false discovery rate (FDR) test with the $p < 0.01$. Putative enzymes involved in carbohydrate utilisation were identified by BLAST searching against a carbohydrate-active enzymes database (CAZymes) (http://www.cazy.org/). Putative protease families were classified by BLAST against the MEROPS database. Fungal secondary metabolite pathways were analysed with the program SMURF (http://www.jcvi.org/smurf/index.php). The evolution of the protein family sizes and expansion and contractions were analysed by CAFE (De Bie et al. 2006). Simple repeat and transposable elements (TEs) were annotated using BLAST against the RepeatMasker library (http://www.repeatmasker.org/) (Tempel 2012).

2.5. Testing for introgression and host-specific evolution

Single nucleotide polymorphisms (SNPs) were detected globally by MUMmer and combined using VCFtools (Kurtz et al. 2004). The *D*-statistic was used to test the phylogenetic distribution of SNPs that display either an ABBA or BABA allelic configuration. *D*-statistic was calculated with (Pease and Rosenzweig 2015):

$$D(P_1, P_2, P_3, O) = \frac{\sum C_{\mathrm{ABAB}}(i) - \sum C_{\mathrm{BABA}}(i)}{\sum C_{\mathrm{ABAB}}(i) + \sum C_{\mathrm{BABA}}(i)},$$

using window sizes of 5, 50 and 100 kb where $C_{\mathrm{ABBA}}(i)$ and $C_{\mathrm{ABAB}}(i)$ are counts of whether or not specified pattern (ABBA or BABA) at the *i*th site in the genome was observed and were calculated using the following equation:

$$C_{\mathrm{ABAB}}(i) = (1 - \hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3}(1 - \hat{p}_{i4})$$

$$C_{\mathrm{BABA}}(i) = \hat{p}_{i1}(1 - \hat{p}_{i2})\hat{p}_{i3}(1 - \hat{p}_{i4})$$

Under the null hypothesis of no introgression, *D* approached zero and the candidate loci were assessed for significance using a *z*-test with $p < 0.01$. To confirm candidate introgression loci, f_d was calculated using the following equation:

$$f_d = \frac{S(P_1, P_2, P_3, O)}{S(P_1, P_D, P_D, O)}$$

where among the compared the four taxa of P_1, P_2, P_3 and *O*, P_D can be either P_2 or P_3, which has higher frequency of the derived allele. We excluded the windows with f_d lower than the top 10% f_d estimators to reduce the bias of heterogeneity in genetic variation. To rule out false positive introgression loci due to incomplete lineage sorting, mean DNA sequence divergence (d_{xy}) was calculated and compared between the candidate loci with the whole scaffold regions using the following equation:

$$d_{xy} = \frac{1}{n}\sum_{n}^{i=1}\left(\hat{p}_{ix}\left(1 - \hat{p}_{iy}\right) + \hat{p}_{iy}(1 - \hat{p}_{ix})\right),$$

where p_x and p_y refer to reference allele frequency in taxa *x* and *y*. Standard error was calculated for the windows in each chromosome. The two values were compared using a *z*-test with $p < 0.01$.

2.6. Annotation of introgression loci

To investigate the functions of introgressed loci, the genes associated with the windows with significant introgression signal were identified. Genes that cover at least one introgressed locus were extracted. Functions of these genes were annotated by blasting against PFAM, FUNCAT and GO database (Harris et al. 2004; Ruepp et al. 2004; Marco et al. 2012).

3. Results

3.1. Phylogeny and host shifts of hypocrealean fungi

Phylogenomic analysis of 45 fungal genomes (Table 1) from seven families in Hypocreales was conducted using *Ustilago maydis* (Basidiomycota) and *Saccharomyces cerevisiae* (Ascomycota) as an outgroup (Rogerson 1970). A dataset comprised of 627 genes encoding single-copy homologous proteins obtained by blasting against the BUSCO database was used to construct the phylogenetic relationships using RAxML (Figure 1) (Alexandros

Table 1. Species and accession numbers of the genomic data used for phylogenomic analysis.

Families	Species	Accession numbers
Clavicipitaceae	*Aciculosporium take*	AFQZ00000000
Clavicipitaceae	*Atkinsonella hypoxylon*	JFHB00000000
Clavicipitaceae	*Balansia obtecta*	JFZS00000000
Cordycipitaceae	*Beauveria bassiana*	ADAH00000000
Clavicipitaceae	*Claviceps fusiformis*	AFRA00000000
Clavicipitaceae	*Claviceps paspali*	AFRC00000000
Clavicipitaceae	*Claviceps purpurea*	CAGA00000000
Bionectriaceae	*Clonostachys rosea*	JYFM00000000
Cordycipitaceae	*Cordyceps militaris*	AEVU00000000
Clavicipitaceae	*Epichloe amarillans*	AFRF00000000
Clavicipitaceae	*Epichloe aotearoae*	JFGX00000000
Clavicipitaceae	*Epichloe baconii*	JFGY00000000
Clavicipitaceae	*Epichloe brachyelytri*	AFRB00000000
Clavicipitaceae	*Epichloe elymi*	AMDJ00000000
Clavicipitaceae	*Epichloe festucae*	AFRX00000000
Clavicipitaceae	*Epichloe gansuensis*	AFRE00000000
Clavicipitaceae	*Epichloe mollis*	JFGW00000000
Clavicipitaceae	*Epichloe typhina*	AMDI00000000
Nectriaceae	*Fusarium circinata*	JRVE00000000
Nectriaceae	*Fusarium fujikuroi*	JRVG00000000
Hypocreaceae	*Fusarium graminearum*	AACM00000000
Nectriaceae	*Fusarium oxysporum*	AAXH00000000
Nectriaceae	*Fusarium pseudograminearum*	AFNW00000000
Nectriaceae	*Fusarium solani*	ACJF00000000
Nectriaceae	*Fusarium virguliforme*	AEYB00000000
Ophiocordycipitaceae	*Hirsutella minnesotensis*	JPUM00000000
Ophiocordycipitaceae	*Hirsutella rhossiliensis*	MPJM00000000
Ophiocordycipitaceae	*Hirsutella thompsonii*	APKU00000000
Clavicipitaceae	*Hypocrella siamensis*	JMQE00000000
Clavicipitaceae	*Metarhizium acridum*	ADNI00000000
Clavicipitaceae	*Metarhizium anisopliae*	AZNF00000000
Ophiocordycipitaceae	*Ophiocordyceps sinensis*	ANOV00000000
Clavicipitaceae	*Periglandula ipomoeae*	AFRD00000000
Clavicipitaceae	*Pochonia chlamydosporia*	AOSW00000000
Stachbotryaceae	*Stachybotrys chartarum*	LDEE00000000
Ophiocordycipitaceae	*Tolypocladium inflatum*	AOHE00000000
Hypocreaceae	*Trichoderma atroviride*	JZUQ00000000
Hypocreaceae	*Trichoderma hamatum*	ANCB00000000
Hypocreaceae	*Trichoderma harzianum*	JNNP00000000
Hypocreaceae	*Trichoderma longibrachiatum*	ANBJ00000000
Hypocreaceae	*Trichoderma reesei*	AAIL00000000
Hypocreaceae	*Trichoderma virens*	ABDF00000000
Incertae sedis	*Verticillium albo-atrum*	ABPE00000000
Incertae sedis	*Verticillium dahiliae*	ABJE00000000
Clavicipitaceae	*Villosiclava virens*	JHTR00000000

2014). Using the divergence time for Ascomycota at 600 million years ago (Mya) (Lücking et al. 2009) for calibration, the divergence time for Hypocreales was estimated to be 217 Mya by BEAST (Figure S1). The resulting phylogeny had high boot-strap support for all families and was mostly consistent with results of a previous multilocus phylogeny (Spatafora et al. 2007). *Verticillium* spp. in Plectospharellaceae, highly virulent plant pathogens, were the earliest diverging lineage in the Hypocreales. *Clonostachys rosea* in Bionectriaceae evolved to infect various hosts including animals, plants, and other fungi and subsequently the families of hypocrealean fungi diverged to various lifestyles. *Fusarium* spp., as representatives of Nectriaceae, developed after Bionectriaceae to be plant pathogens or weak insect pathogens. In turn, several monophyletic lineages corresponding to Cordycipitaceae, Ophiocordycipitaceae and Clavicipitaceae diverged to insect and nematode pathogens. The mycoparasitic *Trichoderma* spp. in Hypocreaceae were nested within the animal pathogenic lineages between Cordycipitaceae and Ophiocordycipitaceae As one of the most diverged lineages, fungi in Clavicipitaceae split into two clades corresponding to insect pathogens and grass endophytes. Available data suggest that early diverging insect pathogenic lineages within Clavicipitaceae originated from other insect pathogen lineages such as Ophiocordycipitaceae and then reverted to a plant host as grass endophytes (Figure 1). Divergence of Clavicipitalean endophyte at 70 Mya was consistent with the divergence time of their plant hosts in Gramineae (Paterson et al. 2004), indicating that the grass endophytes may represent a rapid radiation resulting from adaption to and coevolution with the plant hosts.

The ancestral host-associations and nutrient requirements at the nodes of each lineage were obtained using the program RASP (Figure 1, Figure S2) (Spatafora and Bushley 2015). Results suggested that the fungi in early diverging nodes mainly utilised plant-based nutrients as pathogens and subsequently shifted to simpler nutrient resources including insect-, fungi- and nematode-based resources, and finally a reversal to plant-based nutrition as symbionts (Yu et al. 2015). The most recently diverged lineage, Clavicipitaceae, might also originate from insect/nematode pathogens from other families (probability = 0.9688). Insect/nematode pathogens in Clavicipitaceae most likely evolved from other insect pathogenic families and then reverted to plant hosts as grass endophytes (probability = 0.8893). This evolutionary scenario is also supported by the shared secondary metabolism associated with animal-toxins in insect/nematode pathogens and endophytes in Clavicipitaceae (Spatafora and Bushley 2015).

3.2. Genome characteristics and lifestyles

Genome size expansions and low gene density can result from the accumulation of repetitive sequences

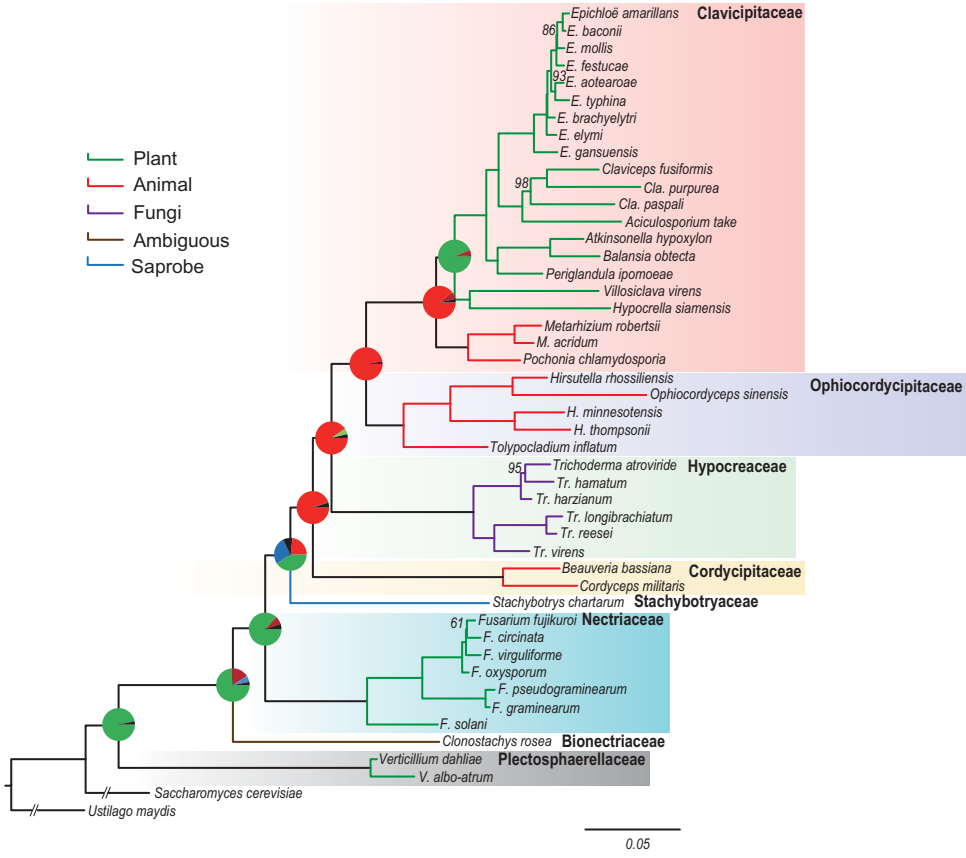

Figure 1. Phylogenetic relationships of the species of the order Hypocreales.

Bootstrap values are 100, except the marked nodes, and provide evidence for the tree structure. Pie charts at each nodes show the possibility of the ancestral life strategies with different colours.

(Wang and St Leger 2013). The genome sizes, gene numbers and gene densities were compared for the 45 fungal genomes. Genome size ranged from the largest, an average of 60.2 Mb for Ophiocordycipitaceae, to the smallest, an average of 32.3 Mb for Clavicipitaceae (Figure 2, Table S2). The largest number of genes were 16,779 predicted in the genome of *Clo. rosea* which interacts with diverse hosts (Figure 2(a)), while the smallest number of genes was 7480 predicted in the genome of *Epichloë* spp. that live as symbiotic endophytes in grasses (Table S2). Low gene density was found in host-specific fungi in both Ophiocordycipitaceae parasitizing animals and Clavicipitaceae colonising plants as symbiotic endophytes (Figure 2(b)). On the other hand, the contents of repetitive sequences and TEs were detected in both grass endophytes *Epichloë* spp., an average of 46% TEs, and Ophiocordycipitaceae with an average of 31%. Although the plant parasitic *Claviceps* spp. are phylogenetically close to *Epichloë* spp., they contained only 13% TEs. The nematode endoparasitic *Hirsutella* spp. and ghost moths parasite *O. sinensis* are both host density-dependent obligate pathogens, indicated their strong interactions with and dependence upon their hosts (Hu et al. 2013; Lai et al. 2014). The large number of TEs in the genomes of these fungi might be associated with their obligate interactions with their hosts.

The gene families were functionally annotated by blasting against the FUNCAT database and a total of 566 gene families were identified and compared among each group characterised as plant pathogens, animal pathogens, fungal pathogens and grass endophytes. The plant pathogens, as the earliest diverging lineages had the highest number of gene families (332 families) that were mainly associated with carbohydrate, lipid and nitrogen metabolism, facility transportation, intracellular signal transduction and stress response (Figure 3, Table S3). The mycoparasitic *Trichoderma* shared similar functional gene families (175 families) with plant pathogens but contained more families involved in stress response, G-protein signal transduction and virulence.

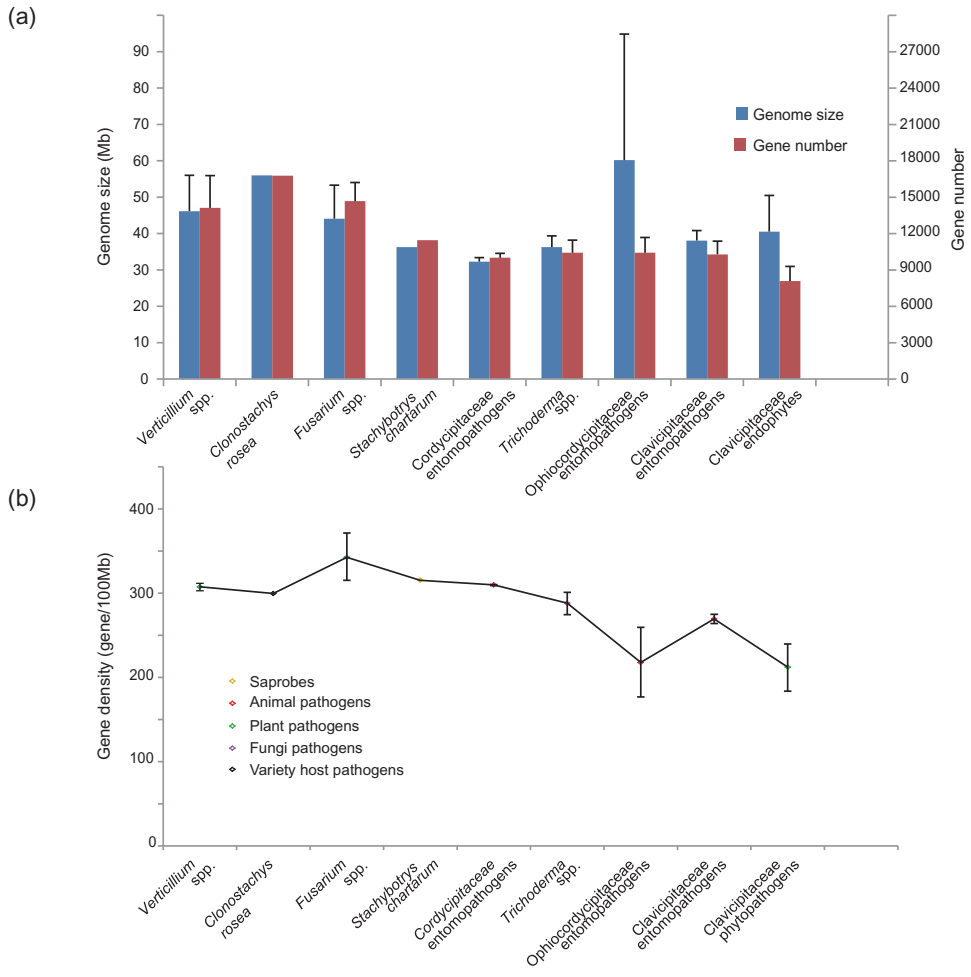

Figure 2. Average genome sizes, gene numbers and gene densities of fungal lineages employing different life strategies.

(A) Left axis shows genome sizes (Mb) and right axis shows gene numbers. Fungal lineages are classified by the same life strategies. (B) Gene densities of fungal lineages employing different life strategies.

Although the genomes of insect and nematode pathogenic fungi had fewer identified gene families than those of fungal and plant pathogens, gene families associated with metabolism, stress response and transportation were more abundant, indicating that these gene families might be involved in adaption or virulence on insect or nematode hosts.

The virulent factors are a key component in the interactions of fungi with their hosts. By blasting against PHI database, a total 521 PHI gene families were identified (Table S4). A lower number of PHI gene families in symbiotic endophytic fungi was observed, while fungi interacting with multiple hosts, such as *Clo. rosea* and *Fusarium*, have a much higher number of G-protein coupled receptors (PHI:441) (average 43) than plant pathogenic *Verticillium* (average 14) and other pathogens (average 10), indicating that fungi with multiple lifestyles required more virulent factors to adapt to different hosts. Furthermore, the typical plant pathogens had a larger number of pectinases (PHI:179, PHI:180 and PHI:222) (average 12), enzymes involved in degradation of the plant cell wall and middle lamella, while these enzymes are present in lower numbers of absent in the other types of pathogens. The utilisation of pectin is essential for plant parasitism.

3.3. Host nutrient-based evolution

The utilisation of host-based nutrition is critical for fungal parasitism and includes or carbohydrate-activated enzymes (CAZymes) and as well as proteases, which are among key virulence factors involved in parasitism in various pathogenic fungi (Gao et al.

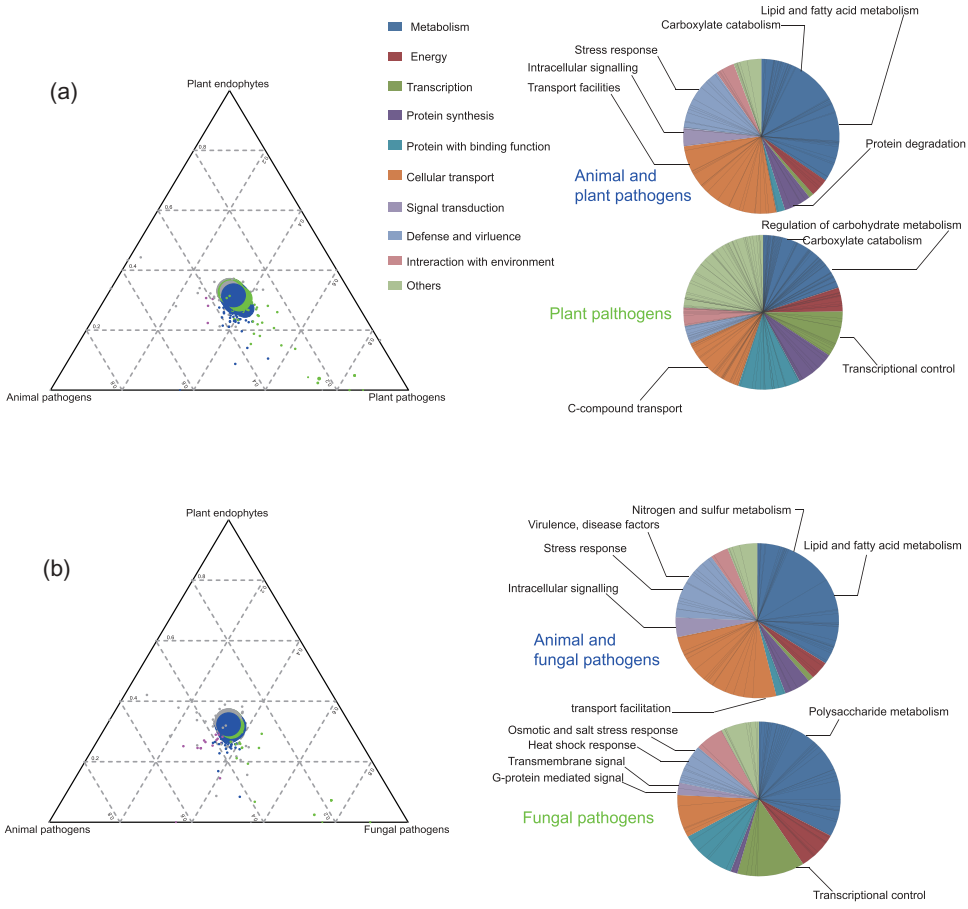

Figure 3. Significant expansion of gene families base on FUNCAT database compared with grass endophytes.

The significance was tested by FDR < 0.05. (A) The expanded gene families shared by plant pathogens and animal pathogens were represented with blue circles in the triangle; the expanded gene families of plant pathogens were coloured green; the expanded gene families of animal pathogens were coloured red; gene families without significance were coloured grey; detailed gene families were represented in the pies. (B) The expanded gene families shared by fungal pathogens and animal pathogens coloured blue; the expanded gene families of fungal pathogens were coloured green; the expanded gene families of animal pathogens were coloured red.

2011; Klosterman et al. 2011; Xiao et al. 2012). The abundance and types of these enzymes should be key to the adaption to different host-based nutrition. The expansion and contraction of CAZymes and MEROPS, as key enzymes responsible for substrate utilisation, were detected by the programme CAFÉ (Figure 4, Table S5) (De Bie et al. 2006; Rawlings et al. 2006; Cantarel et al. 2009). One significant gene family contractions occurred during the divergence of animal pathogens in Cordycipitaceae with 101 contracted families of CAZymes and 29 of MEROPS. Similarly, a total of 64 CAZyme families and 13 MEROPS families were found contracted in Ophiocordycipitaceae. Though Clavicipitaceae appears to be a dichotomous group with both animal pathogens and endophytes, gene family contraction also occurs during the divergence of grass endophytes.

The distribution of families was analysed by a heatmap of the most differentially presented gene families (Figure 4). The degradation of plant cells are associated with a series of CAZymes including numerous cellulase encoding genes containing carbohydrate-binding module 1, GH3, GH5, GH7, GH11 and GH61 domains as well as pectinases coding genes containing GH28, GH43, PL1, PL2 and PL4 domains presented in our matrix analysis of gene families. The genes that process the degradation of plant cells are abundant in the ancestral plant pathogenic *Verticillium* spp. and saprophytic *Stachybotrys chartarum* genomes, while protease-coding genes were less abundant. The largest numbers of enzyme families were identified in the genomes of *Fusarium* spp. and *Clo. Rosea*, fungi with multiple life-strategies. After the transition from plant pathogens to animal pathogens, the number of genes responsible for plant degradation sharply decreased

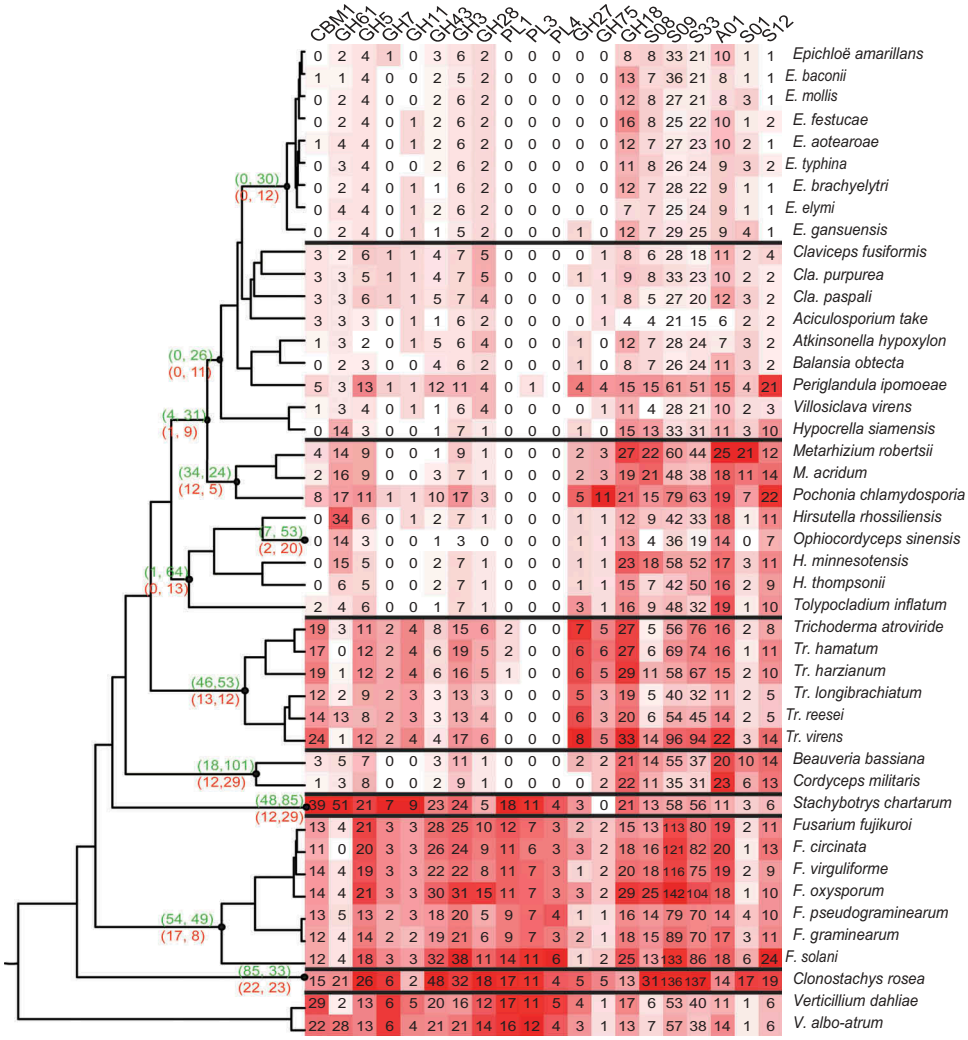

Figure 4. Gene expansion/contraction and gene numbers of corresponding gene families.

The numbers of expanded and contracted gene families are shown at the node of divergence by CAFE: upper brackets (expanded families, contracted families) of CAZy database; below brackets (expanded families, contracted families) of MEROPS database.

and the pectinase PLs almost disappear. On the other hand, the number of protease encoding genes also decreased when the lifestyle changed from insect/nematode pathogens to grass endophytes within Clavicipitaceae (Figure 4, Table S5).

3.4. Evidence for introgression among species

Introgression lines and introgression have been documented to be common during rapid speciation and host adaption (Arnold 2004). Although introgression has been rarely investigated in the evolution of fungi (Pease and Rosenzweig 2015; Zhang et al. 2015), molecular phylogenetic analysis has demonstrated that the speciation of *Epichloë* species has often involved a complicated processes of hybridisation (Moon et al. 2004). In addition, the close relationships between grass endophytes and insect pathogens in Clavicipitaceae and our results suggest that they also share similar genomic features. Thus, introgression may contribute to the evolution of host-shifts and diverse lifestyles.

Therefore, inter-species introgression in Hypocreales was analysed. A sliding window of 100, 50 and 5 kb was used to compute the *D*-statistic (Hudson 1983) as a signal of introgression between *Epichloë* spp. (*E. festucae, E. aotearoae* and *E. gansuensis*) that originate from sexual hybridisation (Durand et al. 2011). *Claviceps purpurea* was designed as an outgroup. There were 28 among 243 the 100-kb windows detected to be exchanged between *E. gansuensis* and *E. festucae* with significant *D*-statistics through *z*-text ($|D| > 0.59$, $p < 1 \times 10^{-3}$ and |

ABBA − BABA| > 10) (Figure 5(b)). Similarly, 74 out of the 582 50-kb windows and 337 out of the 5816 5-kb windows also had significant *D*-values. These data indicate a high probability of introgression during the speciation of *Epichloë* spp. A 5-kb slide window was also applied to further examine genome-wide introgression and identify the candidate introgressed loci and their functions. The *f*-statistic was calculated for the focal intervals and values significantly lower than 90% f_d were excluded (Pease and Rosenzweig 2015). The DNA sequence divergence (d_{xy}) was calculated for each candidate introgression interval to distinguish introgression and ancestral variation. There were 59 introgressed 5-kb windows with significantly lower d_{xy} than that of the whole scaffold region (Table S6). These results reveal a significant signal of introgression between *E. gansuensis* and *E. festucae*, which is to equivalent to the level of introgression observed in distantly related butterfly species (Zhang et al. 2015).

On the other hand, the introgression among the endophytes *E. festucae, Cla. purpurea* and *Villosiclava virens* for which no sexual hybridisation has been reported was examined using the insect pathogenic *Metarhizium robertsii* in Clavicipitaceae as an outgroup. Only 2 out of 253 100-kb windows and 2 out of the 6709 5-kb windows were identified with significant *D* values between *E. festucae* and *Villosiclava virens*, indicating a low possibility for gene introgression (Figure 5(b), Table S6). However, 13 out of 93 100 kb windows and 12 out of the 1842 5-kb windows were identified between the genomes of *Cla. purpurea* and *M. robertsii* (Figure 5(b), Table S6), suggesting that the grass endophytes have genetic resources shared with insect pathogens.

Overall, a coarse inter-subclade frequency of introgression and inferences of the sources of genetic variation were estimated from the *D*-statistics across all the branches within the hypocrealean fungi using 500-kb windows. A total of 100 out of the 112 500-kb windows showed evidence of introgression for at least one clade in the tree (Figure 5(c)). The subclade of grass endophytes in Clavicipitaceae had the largest numbers of windows (average of 9.08%) showing significant inter-species introgression compared with *Trichoderma* spp. (average of 3.59%). Evidence of introgression was also identified between the subclades in Hypocreales. Clavicipitaceae showed

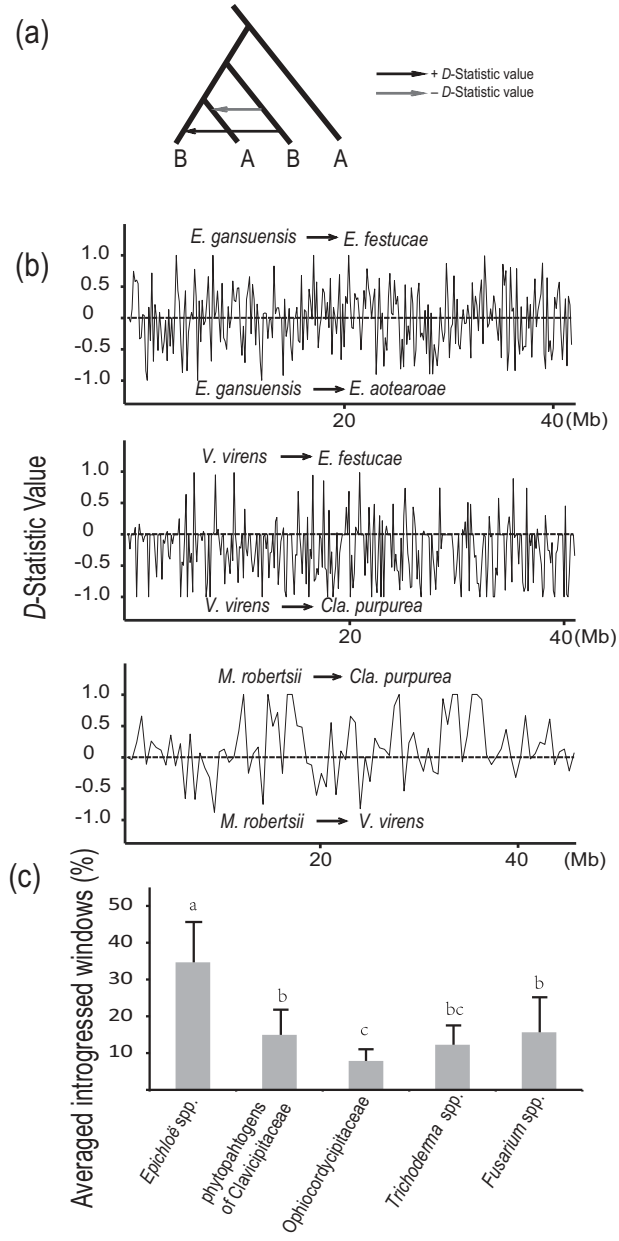

Figure 5. Patterson's *D*-statistic scans along the whole genome sequences.

(A) For the tree topology above, clear introgression patterns are observed. (B) Patterson's *D*-statistic along the whole genome using 500 kb adjacent windows. (C) Averaged windows with introgressive signal detected by significant *D*-statistic.

approximately 10% of the total number of windows introgressed from Hypocreaceae, Cordycipitaceae and Nectriaceae, respectively. Introgression is often associated with sexual hybridisation during speciation (Geiser et al. 1996; Holder et al. 2001). Hypocreales contains many sexual lineages as teleomorphs for many species are found in nature and the evolution of lifestyle shifts and flexible lifestyle traits could result from introgression during sexual

hybridisation. This hypothesis has been proposed previously for Cordycipitaceae (Wang and St Leger 2013). Although our estimates of introgression are based on simple calculations of *D*-statistics, they suggest that hypocrealean fungal species might arise more frequently that previously recognised through sexual hybridisation.

3.5. Annotated functions of the introgressed loci

The functions of genes identified within the introgressed 5-kb windows were further investigated. The genes within introgressed regions were extracted and annotated by blasting them against the Pfam and GO databases. Among the total of 34 genes identified as introgressed between *E. gansuensis* and *E. festucae*, genes that encode proteins involved in transcriptional regulation and stress resistance were highly represented among the introgressed genes as well as three genes homologous to CPUR_06320, CPUR_03992 and CPUR_06368 in that are involved in cell growth (Table S7). These genes could be involved in helping these grass endophytes living inside of plant to protect themselves from oxidative and other stresses activated by host defense responses. On the other hand, introgressed genes between *M. robertsii* and *Cla. purpurea* were mainly involved in regulation of metabolism and host recognition. For example, the introgressed THAM_008331, a GTPase-activator protein and THAM_06628, a cAMP dependent protein kinase, highlights the importance of host recognition for these taxa (Table S7).

4. Discussion

The 45 genomes included in this study belong to seven out of the nine families currently recognised in Hypocreales. Several species were included in some genera of some families (e.g. *Fusarium, Trichoderma*, etc.). However, due to the limited availability of genomes, only one species in each family of Stachybotryaceae and Bionectriaceae was included in the present analysis, which is less than that of the previous phylogenetic study using multilocus sequencing of ribosomal and protein coding genes (Spatafora et al. 2007). Although insufficient genomic data limited the depth of analysis of some evolution mechanisms such as the host jumping in these fungi (Sung et al. 2007), this study highlights evolutionary processes involved in nutrient-based lifestyles and a comprehensive understanding of the phylogenetic relationships of different lineages. The results suggest that the hypocrealean fungi might originate from a plant-based nutrition. The RASP yielding ancestral lifestyle and the identification of numerous plant parasitism-related genes in the genomes of *Metarhizium* spp. also support this hypothesis (Gao et al. 2011).

To adapt to distinct host-based nutrients, gene families significantly contracted or expanded. Generally, genomic analysis revealed that a higher abundance of CAZymes is associated with the utilisation of plant-based nutrition, while greater numbers of proteases and chitinases are associated with insect-based nutrition, while only proteases are associated with nematode-based nutrition (Dimitrios et al. 2012; Xiao et al. 2012; Lai et al. 2014; Liu et al. 2014). A series of CAZymes involved in carbohydrate degradation were significantly contracted during the lifestyle transition from plant pathogens to other animal pathogenic, mycoparasitic or endophytic lifestyles. Proteases, were mainly associated with animal-based nutrition and may also function as virulence factors, were also significantly contracted during the lifestyle transition from animal pathogens to grass endophytes. Meanwhile, contraction of genes involved in nutrient utilisation might further narrow the host range. Analysis of global distributions of gene functions showed that functions that show a decrease in animal pathogens and grass endophytes are mainly associated with metabolism and nutritional transport. The contraction of G-protein receptors might also limit the host recognition and contribute to limiting the host range. The global contraction of gene families in *O. sinesis* also likely contributes to its narrow host range and obligate parasitism (Hu et al. 2013).

The *D*-statistic test in genomic analysis was first introduced by Green et al. (2010) to evaluate formally whether humans harbour some Neandertal ancestry (Green et al. 2010). Recently, it has been used as a convenient statistic for studying locus-specific introgression of genetic material controlling coloration in *Heliconius butterflies* (Zhang et al. 2015). *D*-statistic analysis requires four species including two sister species, a third species potentially involved in

introgression and an outgroup species (Martin et al. 2014). Most investigations of introgression focus on animals and plants, such as horse, butterfly and tomato, that have sexual reproduction during their life-cycles (Pease and Rosenzweig 2015; Zhang et al. 2015). Generally, it is believed that sexual reproduction could lead to a higher occurrence of introgression (Geiser et al. 1996; Holder et al. 2001). However, introgression has rarely been investigated in the evolutionary studies of fungi.

The global analysis of introgression in hypocrealean fungi was conducted and *Clo. Rosea*, an ancestral species that displays diverse life-strategies without significant gene contraction, was used as an outgroup to guarantee the maximum amount of homologous sequences. However, only 40% of the 500-kb windows are homologous to the genome sequences of *Epichloë* spp. that have global gene contractions and large number of TEs. A large number of sequences with a signal of introgression were identified in *Epichloë* spp, supporting previous observations that speciation among *Epichloë* spp. is often associated with sexual reproduction and hybridisation (Moon et al. 2004). Numerous genes are introgressed between *M. robertsi* and *Cla. purpurea*, indicating that the endophytes and animal pathogens share very close ancestors and that speciation in the Clavicipitaceae has involved in frequently introgression. The high frequency of introgression identified among species in Hypocreales provide evidence that adaptive introgression and gene flow among fungi living on similar hosts may contribute to the evolution of the diverse and flexible lifestyles notable for this group of fungi.

In summary, the evolution of distinct host nutrient-based lifestyles of hypocrealean fungi is supported by the contraction and expansion of nutrient utilisation related gene families corresponding to the lifestyle adaptions to various hosts. Plant pathogens appear to be the earliest group from which animal and fungal pathogens evolved, and finally reverted back to a plant-based nutrition as plant endophytes. The observation of global gene family contractions, especially in cellulases encoding genes in the transition from plant pathogens to animal and fungal pathogens, and pectinases encoding genes in the transition from animal pathogens to endophytes. Introgression signals were significantly detected in certain lineages of hypocrealean fungi and the main functions of the genes located in the introgressed regions were related to host recognition, transcriptional regulation, stress response and cell growth regulation. Introgression and gene family contraction/expansion are evolutionary mechanisms that may drive rapid speciation and diverse host shift observed in hypocrealean fungi, one of the most impact group on ecosystem, agriculture and human health.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant# 31430071) and the National Key Basic Research Program of China 973 Program (2013CB127506). The authors thank Prof. Manhong Sun at Institute of Plant Protection (IPP), Chinese Academy of Agricultural Sciences (CAAS) for the valuable providing of the *Clonostachys rosea* genome.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Natural Science Foundation of China (grant# 31430071) and the National Key Basic Research Program of China 973 Program (2013CB127506)

ORCID

Chengshu Wang http://orcid.org/0000-0003-1477-1466

References

Alexandros S. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.

Arnold ML. 2004. Transfer and origin of adaptations through natural hybridization: were Anderson and Stebbins right? Plant Cell. 16:562–570.

Berbee ML. 2001. The phylogeny of plant and animal pathogens in the Ascomycota. Physiol Mol Plant P. 59:165–187.

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. 37:D233–D238.

Chen S, Liu X, Chen F. 2000. *Hirsutella minnesotensis* sp. nov., a new pathogen of the soybean cyst nematode. Mycologia. 92:819–824.

Clay K. 1988. Fungal endophytes of grasses: a defensive mutualism between plants and fungi. Ecology. 69:10–16.

De Bie T, Cristianini N, Demuth J, Hahn M. 2006. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 22:1269–1271.

De Wolf E, Madden L, Lipps P. 2003. Risk assessment models for wheat *Fusarium* head blight epidemics based on within-season weather data. Phytopathology. 93:428–435.

Dimitrios F, Manfred B, Robert R, Kerrie B, Robert B, Bernard H, Angel M, Otillar R, Spatafora JW. 2012. The paleozoic origin of enzymatic mechanisms for lignin degradation reconstructed using 31 fungal genomes. Science. 1715:336.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 29:1969–1973.

Durand EYN, Patterson D, Reich SM. 2011. Testing for ancient admixture between closely related species. Mol Biol Evol. 28:2239–2252.

Gao Q, Jin K, Ying SH, Zhang Y, Xiao G, Shang Y, Duan Z, Hu X, Xie XQ, Zhou G. 2011. Genome sequencing and comparative transcriptomics of the model entomopathogenic fungi *Metarhizium anisopliae* and *M. acridum*. PLoS Genet. 7: e1001264.

Geiser DM, Timberlake WE, Arnold ML. 1996. Loss of meiosis in *Aspergillus*. Mol Biol Evol. 13:809–817.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Heng L, Zhai W, Fritz MH-Y, et al. 2010. A draft sequence of the neandertal genome. Science. 328:710–722.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 9:r7.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R. 2004. The gene ontology (go) database and informatics resource. Nucleic Acids Res. 32:258–261.

Holder MT, Anderson JA, Holloway AK. 2001. Difficulties in detecting hybridization. Systematic Biol. 50:978–982.

Hu X, Zhang Y, Xiao G, Zheng P, Xia Y, Zhang X, St Leger RJ, Liu X, Wang C. 2013. Genome survey uncovers the secrets of sex and lifestyle in caterpillar fungus. Chin Sci Bull. 58:2846–2854.

Hudson RR. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution. 37:203–217.

Jaffee B, Zehr E. 1982. Parasitism of the nematode *Criconemella xenoplax* by the fungus *Hirsutella rhossiliensis*. Phytopathology. 72:1378–1381.

Kazutaka K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:102–343.

Kirk P, Cannon P, Minter D, Stalpers J. 2008. Dictionary of the fungi. Wallingford, UK: CABI.

Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BP, Chen Z, Henrissat B, Lee YH, Park J. 2011. Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens. PLoS Pathog. 7: e1002137.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg S. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5:R12.

Lai Y, Liu K, Zhang X, Zhang X, Li K, Wang N, Shu C, Wu Y, Wang C, Bushley KE, et al. 2014. Comparative genomics and transcriptomics analyses reveal divergent lifestyle features of nematode endoparasitic fungus *Hirsutella minnesotensis*. Genome Biol Evol. 6:3077–3093.

Lamichhaney S, Berglund J, Almen MS, Maqbool K, Grabherr M, Martinez-Barrio A, Promerová M, Rubin CJ, Wang C, Zamani N, et al. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature. 518:371–375.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 25:1966–1967.

Liu K, Zhang W, Lai Y, Xiang M, Wang X, Zhang X, Liu X. 2014. *Drechslerella stenobrocha* genome illustrates the mechanism of constricting rings and the origin of nematode predation in fungi. BMC Genomics. 151:452–457.

Liu X, Chen SY. 2000. Parasitism of *Heterodera glycines*, by *Hirsutella* spp. in Minnesota soybean fields. Biol Control. 19:161–166.

Lücking R, Huhndorf S, Pfister DH, Plata ER, Lumbsch HT. 2009. Fungi evolved right on track. Mycologia. 101:810–822.

Marco P, Coggill PC, Eberhardt RY, Jaina M, John T, Chris B. 2012. The pfam protein families database. Nucleic Acids Res. 40:290–301.

Martin S, Davey J, Jiggins C. 2014. Evaluating the use of ABBA-BABA statistics to locate introgrerssed loci. Mol Biol Evol. 32:244–257.

Möller EM, Bahnweg G, Sandermann H, Geiger HH. 1992. A simple and efficient protocol for isolation of high molecular weight DNA from filamentous fungi, fruit bodies, and infected plant tissues. Nucleic Acids Res. 20:6115–6116.

Moon CD, Craven KD, Leuchtmann A, Clement SL, Schardl CL. 2004. Prevalence of interspecific hybrids amongst asexual fungal endophytes of grasses. Mol Ecol. 13:1455–1467.

Paterson A, Bowers J, Chapman B. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA. 101:9903–9908.

Pease J, Rosenzweig B. 2015. Encoding data using biological principles: the multisample variant format for phylogenomics and population genomics. IEEE ACM Trans Comput Biol. 9623:1.

Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. Plos Biol. 14:e1002379.

Rawlings ND, Morton FR, Barrett AJ. 2006. MEROPS: the peptidase database. Nucleic Acids Res. 34:D270–D272.

Rogerson CT. 1970. The hypocrealean fungi (ascomycetes, Hypocreales). Mycologia. 62:865–910.

Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, Couloux A, Dominguez V, Anthouard V, Bally P, Bourras S. 2011. Effector diversification within compartments of the

Leptosphaeria maculans genome affected by repeat-induced point mutations. Nature Commun. 2:202.

Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Güldener U, Mannhaupt G, Münsterkötter M, et al. 2004. The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. 32:5539–5545.

Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 507:354–357.

Song Y, Endepols S, Klemann N, Richter D, Matuschka FR, Shih CH, Nachman MW, Kohn MH. 2011. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. Curr Biol. 21:1296–1301.

Spatafora JW, Bushley KE. 2015. Phylogenomics and evolution of secondary metabolism in plant-associated fungi. Curr Opin Plant Bio. 26:37–44.

Spatafora JW, Sung G, Sung J, Hywel-Jones NL, White JF. 2007. Phylogenetic evidence for an animal pathogen origin of ergot and the grass endophytes. Mol Ecol. 16:1701–1711.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690.

Summerell BA, Leslie JF, Liew ECY, Laurence MH, Bullock S, Petrovic T, Bentley AR, Howard CG, Peterson SA, Walsh JL. 2011. *Fusarium* species associated with plants in Australia. Fungal Divers. 46:1–27.

Sung GH, Hywel-Jones NL, Sung JM, Luangsa-Ard JJ, Shrestha B, Spatafora JW. 2007. Phylogenetic classification of *Cordyceps* and the clavicipitaceous fungi. Stud Mycol. 57:5–59.

Sung GH, Poinar GO, Spatafora JW. 2008. The oldest fossil evidence of animal parasitism by fungi supports a Cretaceous diversification of fungal–arthropod symbioses. Mol Phylogenet Evol. 49:495–502.

Tempel S. 2012. Using and understanding repeatmasker. Methods Mol Biol. 859:29–51.

Toledo AV, Virla E, Humber RA, Paradell SL, Lastra CC. 2006. First record of *Clonostachys rosea*, (ascomycota: hypocreales) as an entomopathogenic fungus of oncometopia tucumana, and sonesimia grossa, (hemiptera: cicadellidae) in argentina. J Invertebr Pathol. 92:7–10.

Wang CSt Leger RJ. 2013. In: editor. The ecological genomics of fungi. John Wiley & Sons, New York; 243–260 p.

White JF Jr, Bacon CW, Hywel-Jones NL, Spatafora JW. 2003. Clavicipitalean fungi: evolutionary biology, chemistry, biocontrol and cultural impacts. CRC Press, New York.

Whitney KD, Randell RA, Rieseberg LH. 2006. Adaptive introgression of herbivore resistance traits in the weedy sunflower *Helianthus annuus*. Am Nat. 167:794–807.

Winnenburg R, Baldwin TK, Urban M, Rawlings C, Köhler J, Hammond-Kosack KE. 2006. Phi-base: a new database for pathogen host interactions. Nucleic Acids Res. 34:459–464.

Xiao G, Ying SH, Zheng P, Wang ZL, Zhang S, Xie XQ, Shang Y, St Leger RJ, Zhao GP, Wang CS. 2012. Genomic perspectives on the evolution of fungal entomopathogenicity in *Beauveria bassiana*. Sci Rep. 2:483.

Yu Y, Harris AJ, Blair C, He X. 2015. RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography. Mol Phylogenet Evol. 87:46–49.

Zdobnov E, Apweiler R. 2001. InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 17:847–848.

Zhang L, Yang J, Niu Q, Zhao X, Ye F, Liang L, Zhang KQ. 2008. Investigation on the infection mechanism of the fungus *Clonostachys rosea* against nematodes using the green fluorescent protein. Appl Microbiol Biot. 78:983–990.

Zhang W, Dasmahapatra KK, Mallet J, Moreira GRP, Kronforst MR. 2015. Genome-wide introgression among distantly related Heliconius butterfly species. Genome Biol. 17:1–15.

Zheng P, Xia Y, Xiao G, Xiong C, Hu X, Zhang S, Zheng H, Huang Y, Zhou Y, Wang C. 2011. Genome sequence of the insect pathogenic fungus *Cordyceps militaris*, a valued traditional Chinese medicine. Genome Biol. 12:R116.