Lifestyles shape genome size and gene content in fungal pathogens

Anna Fijarczyk[1,2,3,4,*], Pauline Hessenauer[1,2], Richard C. Hamelin[1,2,5], Christian R. Landry[1,2,3,4,6]

[1]Département de Biologie, Université Laval, Québec, G1V 0A6 Québec, Canada; [2]Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, G1V 0A6 Québec, Canada; [3]PROTEO, Le réseau québécois de recherche sur la fonction, la structure et l'ingénierie des protéines, Université Laval, Québec, G1V 0A6 Québec, Canada; [4]Centre de Recherche en Données Massives (CRDM), Université Laval, Québec, G1V 0A6 Québec, Canada; [5]Department of Forest and Conservation Sciences, The University of British Columbia, V6T 1Z4 Vancouver, Canada; [6]Département de Biochimie, Microbiologie et Bioinformatique, Université Laval, Québec, G1V 0A6 Québec, Canada;

*Corresponding author: anna.fijarczyk.1@ulaval.ca

Keywords:
fungal pathogens, genome evolution, fungal lifestyles, genome size, insect vectors, Sordariomycetes

Summary

Fungi have a wide range of lifestyles and hosts. We still know little about the impact of lifestyles on their genome architecture. Here, we combined and annotated 562 fungal genomes from the class Sordariomycetes and examined the coevolution between 12 genomic and two lifestyle traits: pathogenicity and insect association. We found that pathogens tend to evolve a larger number of protein-coding genes, tRNA genes, and have larger non-repetitive genome sizes than non-pathogenic species. In contrast, species with a pathogenic or symbiotic relationship with insects have smaller genome sizes and genes with longer exons; they also have fewer genes if they are vectored by insects, compared to species not associated with insects. Our study demonstrates that pathogen genome size and complexity are the result of an interplay between drift, imposed by symbiosis and small effective population size, which leads to genome contraction, and the adaptive role of gene amplification, which leads to genome expansion.

Introduction

Variation in genome size has been explained through mutually non-exclusive hypotheses invoking neutral and adaptive processes (reviewed in[1]). For instance, neutral hypotheses are supported by the correlation of genome size with non-coding DNA content[2,3], whereas natural selection is invoked to explain the strong correlation between genome size and cell size[4]. So far, genome size has been shown to correlate with various other traits in eukaryotes, including body size in many species (e.g. salamanders[5] and amphipods[6]), life-history traits (e.g. Pinus[7], or birds[8]), flowering time across latitudinal cline in *Zea mays*[9] or higher fitness in algae[10]. Understanding the contribution of neutral and adaptive processes to genome size evolution may thus require analyzing multiple mechanisms underlying the structural complexity of genomes, such as the amount and distribution of coding and non-coding DNA[11].

Fungi display a great variety of genome sizes (from 8 to 177.6 Mbp)[12]. Observations support both neutral and adaptive mechanisms in shaping this diversity. The evolution of genome size in ascomycetes on a broad phylogenetic scale supports the mutation-hazard hypothesis (MHH) according to which stronger drift facilitates the accumulation of slightly deleterious non-coding DNA, including mobile genetic elements and of introns[2]. As a consequence, species with smaller effective population sizes (Ne) are more prone to genome expansion. As expected with MHH in eukaryotes, ascomycetes with larger genomes have more introns, lower gene density, and higher transposable elements (TE) activity than species with smaller genomes[13]. On the other hand, some host-specialized fungi and obligate pathogens, species that are also associated with small Ne and strong drift, experience genome degradation and gene loss, which is explained by the elimination of accumulated deleterious DNA in the long term due to bias towards deletions[13,14].

The role of adaptive evolution in shaping genome size has also been supported in fungi. The most obvious observation is the strong correlation between the number of genes and genome size[15]. For instance, the diversification of specific gene families is often associated with pathogenicity in fungi[16–18]. Coincidentally, fungal plant pathogens harbor some of the largest known genomes across fungi[15]. Some other genetic mechanisms that contribute to genome expansion have been attributed to rapid adaptation of pathogens to hosts and these include accessory chromosomes[19] or TE activity[20], often linking the emergence of pathogenicity to genome size expansion[21]. For instance, TE expansions are known to play a role in the emergence of new virulence genes in emerging pathogens[22,23].

The ecological aspects of fungal lifestyles, for instance the dependence on the host species, can have an effect both on the gene repertoire of the species as well as on their Ne, and as a consequence on the amount of non-coding DNA and genome size. Pathogens can undergo phases of high clonality which significantly reduces their Ne, however other mechanisms such as accessory chromosomes, or chromosome duplication can increase Ne in parts of the genomes[24]. Other pathogens, such as the opportunistic human pathogen *Aspergillus fumigatus*, persist in the environment where they can maintain high levels of nucleotide polymorphism, consistent with large Ne[25]. On the contrary, obligate endoparasites, endosymbionts, and pathogens requiring vectors have typically small Ne due to their association with a single host species and transmission bottlenecks. For instance, the endoparasites Microsporidia have one of the smallest known fungal genomes, showing signatures of strong genome contractions, including reduced gene repertoire[26]. Therefore, studying the impact of pathogenic lifestyle on genome size in fungi requires a wider ecological context, including their symbiotic or mutualistic relationships with other species.

In this study, we investigate how variation in fungal genome size and its underlying potentially neutral or slightly deleterious genomic features (e.g. intron number, repeat content) and adaptation (number of genes) coevolve with lifestyles. We consider pathogenic and symbiotic relationships with plants and insects. We aim to answer the following questions: 1) whether the fungal genome size and complexity coevolve with lifestyle, and 2) if there are common genomic features related to the evolution of pathogenicity. To answer these questions, we focus on the ascomycete class of Sordariomycetes fungi, which have rich genomic resources and include species with a wide range of lifestyles, from plant and insect pathogens, through wood decomposers, and endophytes, to opportunistic human pathogens. We studied over 560 Sordariomycetes genomes and examined how various genomic features coevolved with the diversity of lifestyles in this taxonomic group.

Results

Quality of genome assemblies of 562 Sordariomycetes fungal species

We analyzed 562 genomes from the class Sordariomycetes (Ascomycota) and 11 outgroup species together comprising fungi mostly represented by plant pathogens, saprotrophs, and entomopathogens (Table S1). All species were classified as pathogenic (including pathogens of plants, animals, and of fungi) or non-pathogenic based on a literature search. Initially, representative genome assemblies for known species of Sordariomycetes and an outgroup (n=584 in total) were downloaded from NCBI (n=580) or other resources (n=4, Table S1) and complemented with 21 genomes sequenced and assembled in this study (Table S2). In total, 605 assemblies were

filtered for contaminants and short contigs, and analyzed for quality. Gene models were inferred *ab initio* for all assemblies. Thirty-two assemblies were excluded due to lack of many single-copy conserved genes (more than 25% of conserved genes were missing) or an excessive number of inferred gene models (*Blumeria graminis*), arriving at a final number of 573 genome assemblies for further analyses. The number of genes was systematically underestimated (10% on average) compared to the gene numbers that had been submitted to NCBI for the corresponding species, but underestimation was not different between pathogenic or non-pathogenic fungi (Wilcoxon rank sum test, w = 2994, p = 0.57). Species genomes ranged from 20.7 Mbp in *Ceratocystiopsis brevicomis* (CBS 137839) to 110.9 Mbp in *Ophiocordyceps sinensis* (IOZ07) and the *ab initio* gene models ranged from 6,280 in *Ambrosiella xylebori* (CBS 110.61) to 17,878 in *Fusarium langsethiae* (Fe2391).

Genomic traits correlating with genome size

The maximum likelihood tree generated from 1000 concatenated single-copy conserved proteins has a 100% bootstrap support for all major nodes except for one (support of 82% for the split between two subclades of Ophiostomatales, Figure 1A). The topology placed the Microascales order as a sister clade to Hypocreales, which is consistent with a topology based on four nuclear genetic markers[27] but not consistent with the topology presented on the JGI MycoCosm (accessed July 10th 2022, https://mycocosm.jgi.doe.gov/mycocosm/home), where Hypocreales and Glomerellales form sister clades, and Microascales is their outgroup. To check if the discrepancy is caused by the inference method, we reconstructed the topology with two other methods: i) from 1000 individual protein trees using a coalescent-based method, and ii) with a maximum likelihood approach from concatenated 250 single-copy conserved proteins with separate partition models assigned to each protein. Both methods gave nearly identical topology to the first one and showed maximal support for Microascales and Hypocreales as sister clades. Thus, the primary maximum likelihood phylogeny (Figure 1A) based on 1000 concatenated protein alignments was used for all downstream analyses.

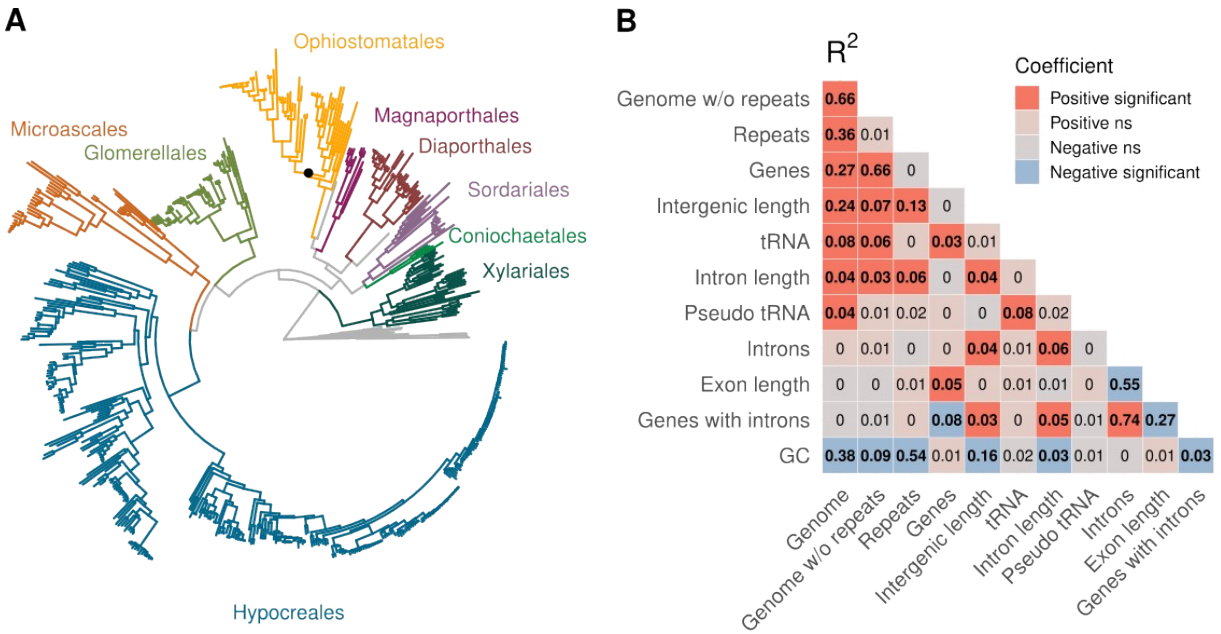

Figure 1. Several genomic traits are correlated with genome size in Sordariomycetes. **A**. Maximum likelihood tree based on 1000 concatenated protein sequence alignments calculated with IQ-TREE using ultrafast bootstrap approximation (n=573 species). Largest orders are indicated with different colors. Bootstrap support for all major clades except one within Ophiostomatales (82%, black dot) reach 100%. **B**. R-squared values derived from generalized least squares models fitted to all pairwise combinations of genomic traits. Colors indicate positive (red) or negative (blue) coefficients obtained from fitting one trait to another. Coefficients significantly different from zero (t-test, adjusted $p<0.05$) are highlighted with dark colors and bold R^2 values. Genomic traits include genome size (genome), size of the assembly excluding repeat content (genome w/o repeats), the fraction of repeat content (repeats), the number of genes (genes), mean intergenic length (intergenic length), the number of tRNA genes (tRNA), mean intron length (intron length), the number of pseudo tRNA genes (pseudo tRNA), mean number of introns per gene (introns), the mean exon length (exon length), the fraction of genes with introns (genes with introns), and GC content (GC).

We used phylogenetic generalized least squares to estimate the correlation between genome size (bp) and 11 genomic traits, including the number of genes, the fraction of repeat content, size of the assembly excluding repeat content (bp), GC content, the mean number of introns per gene, mean intron length (bp), mean exon length (bp), the fraction of genes with introns, mean intergenic length (bp) and the number of tRNA and pseudo tRNA genes. As expected from the MHH in eukaryotes, in Sordariomycetes, genome size is positively correlated with the fraction of repeat content, intergenic length, and mean intron length (Figure 1B, Figure S1). Genome size is also negatively correlated with the GC content, as a consequence of the negative correlation between repeat and GC content. Similar to other fungi, genome size is also strongly positively correlated with the number of genes (including tRNA genes) and consequently to genome size without the repeat content. The numbers of introns and exon length are

strongly negatively correlated with each other, but weakly with genome size, and they explain mostly the second principal component in the PCA analysis (Figure S1). Similarly, the fraction of genes with introns is also not correlated with genome size (Figure 1B, Figure S1). These results support that drift has shaped genome size of Sordariomycetes fungi through expansion of repeat content and intergenic DNA, and that adaptive evolution has shaped genome size through the proliferation of genes. However, some signatures of drift, such as frequency of introns and size of exons do not show a clear relationship with genome size, suggesting opposing consequences of drift on small and large genomes.

Pathogenicity coevolves with the number of genes

Our dataset comprises 357 pathogens, 202 non-pathogens and 4 species with undetermined pathogenicity traits. Pathogens can be found in every order except for Sordariales (Figure 2A). Even though the trait could not be inferred with high accuracy at the root, the dispersion of pathogenic species across the phylogeny implies that the evolution and loss of pathogenicity have occurred multiple times in Sordariomycetes history (Figure 2A).

We used three methods (trait coevolution in BayesTraits, phylogenetic logistic regression, and machine learning-based classification of pathogens vs. non-pathogens) to test which genomic traits coevolve with pathogenicity. Using a discrete reversible jump mcmc (rjMCMC) method in BayesTraits, we tested dependent versus independent models of evolution of genomic traits with pathogenicity, by transforming all genomic traits into binary traits (high/low) with respect to their median. All genomic traits except for total genome size, fraction of genes with introns, and exon length show correlated evolution with pathogenicity (Figure 2B, Table S3). Posterior probabilities of transition rates show that the most frequent gain of pathogenicity occurs when the number of genes, introns, and the genome size without repeats are high (Figure 2B, Figure S2). We also observe gains of pathogenicity for low values of repeats (Figure 2B, Figure S2). We fitted a covarion model to test if dependent evolution with pathogenicity is present across the tree or is rather limited to some branches. There is a significantly better fit for varying rates of evolution than the strictly dependent model for all genomic traits, except for the number of genes, in which the support for the varying rates of evolution was the weakest (average log bayes factor across five runs is 4.5, Table S3).

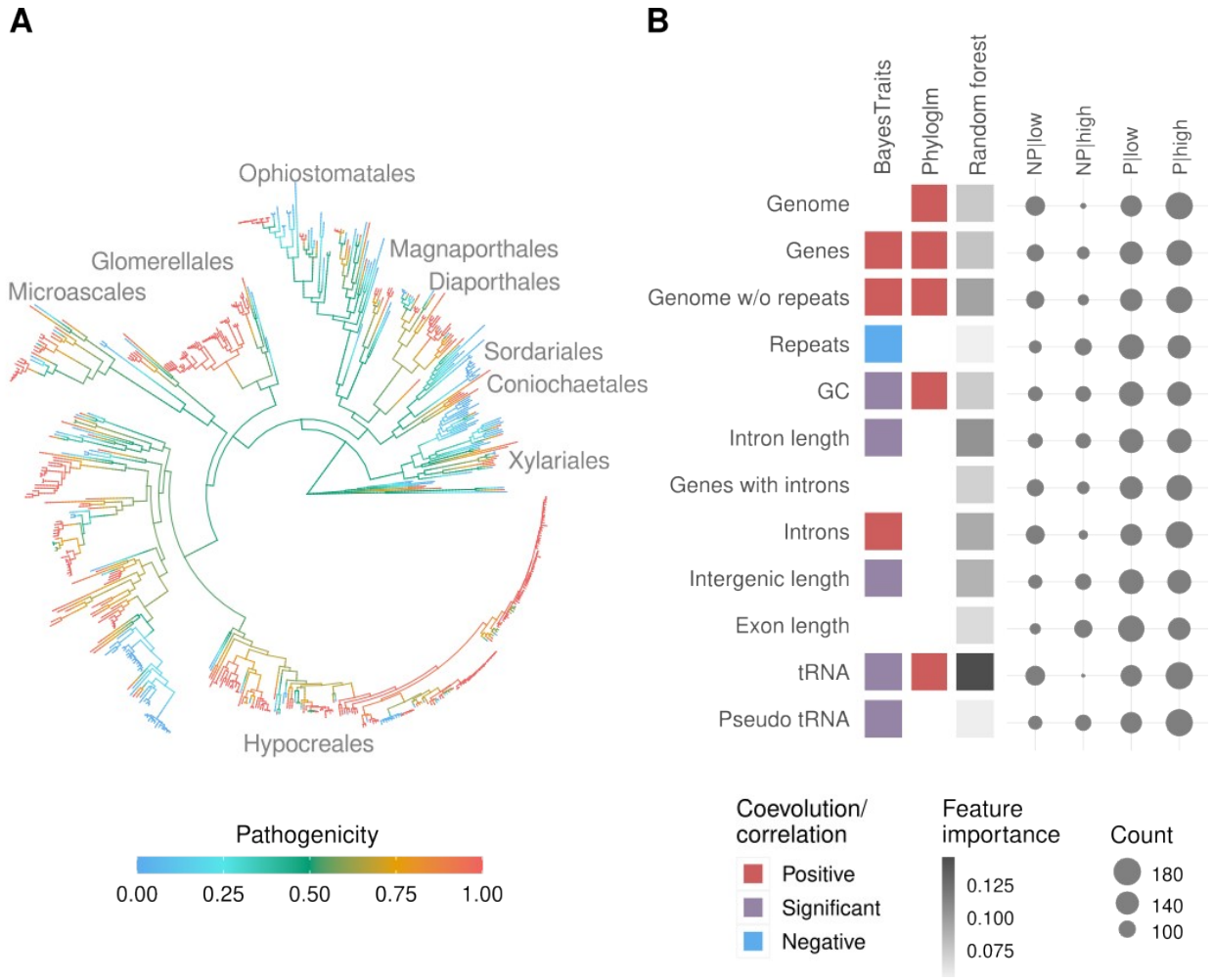

Figure 2. Genomic traits are evolving in concert with pathogenicity. A. Ancestral state reconstruction of pathogenicity. Colors correspond to estimated likelihoods of pathogenicity, calculated with ace function from the phytools package in R, using the "ER" model of discrete evolution with a single rate for the transition from pathogen to non-pathogen and vice versa. **B.** Results of the three methods for detecting correlation between pathogenicity and genomic traits: BayesTraits, phylogenetic logistic regression (phyloglm), and random forest classifier. Red and blue colors indicate positive and negative, respectively, coevolution/correlation of the trait with pathogenicity. Violet indicates those traits which coevolve with pathogenicity according to BayesTraits but a single direction of correlation could not be determined. Circles indicate frequencies of pathogenic (P) and nonpathogenic (NP) species with low (below median) or high (above median) values for 12 genomic traits.

In the second approach, the phylogenetic logistic regression was fitted to the pathogenicity trait. This analysis confirmed a positive correlation between pathogenicity and the number of genes and genome size without repeats, similarly to the analysis with BayesTraits, and revealed a positive correlation with GC content, and the number of tRNA genes (Figure 2B, Table S4). This method also positively correlated genome size with pathogenicity.

Finally, we used machine learning classification to rank the most important among 12 genomic traits for predicting pathogenicity, with the random forest classifier providing the highest accuracy predictions. We tested the performance of the classifier including phylogenetic relationships, which did not substantially improve the accuracy (AUC=0.7 vs AUC=0.71, see Methods, Figure S2). Top ranked genomic features included the number of tRNAs, intron length, assembly size without repeats, and the number of introns (Figure 2B, Figure S2).

All analyses were repeated for a dataset that excluded repeated species (n=563 distinct species in total), giving nearly identical results, except that BayesTraits inferred weaker coevolution between pathogenicity and the number of introns, and the phyloglm model gave no support for the significant correlation with GC content (Tables S3 and S4, Methods).

Overall, several lines of evidence indicate that most pathogens evolve through an increase in gene number, genome size excluding repeats, number and/or length of introns, and the number of tRNAs. This is consistent with the observation that most non-pathogenic species very rarely show high values for these traits (Figure 2B).

Genome reductions in species associated with insects

After reconstructing ancestral gene numbers in the Sordariomycetes phylogeny, as predicted with phylogenetic approaches (Figure 2B), several clades enriched for plant pathogens show an increased number of genes (Figure 3). These include Diaporthales (D), Magnaporthales (Ma), Glomerellales (G), and many *Fusarium* pathogens within the Hypocreales (H1) clade. On the other hand, a couple of clades including insect pathogens (eg. H2.4) and a few plant pathogens (M1, O) carry some of the smallest gene numbers and genome sizes without repeats (Figure 3, Figure S3).

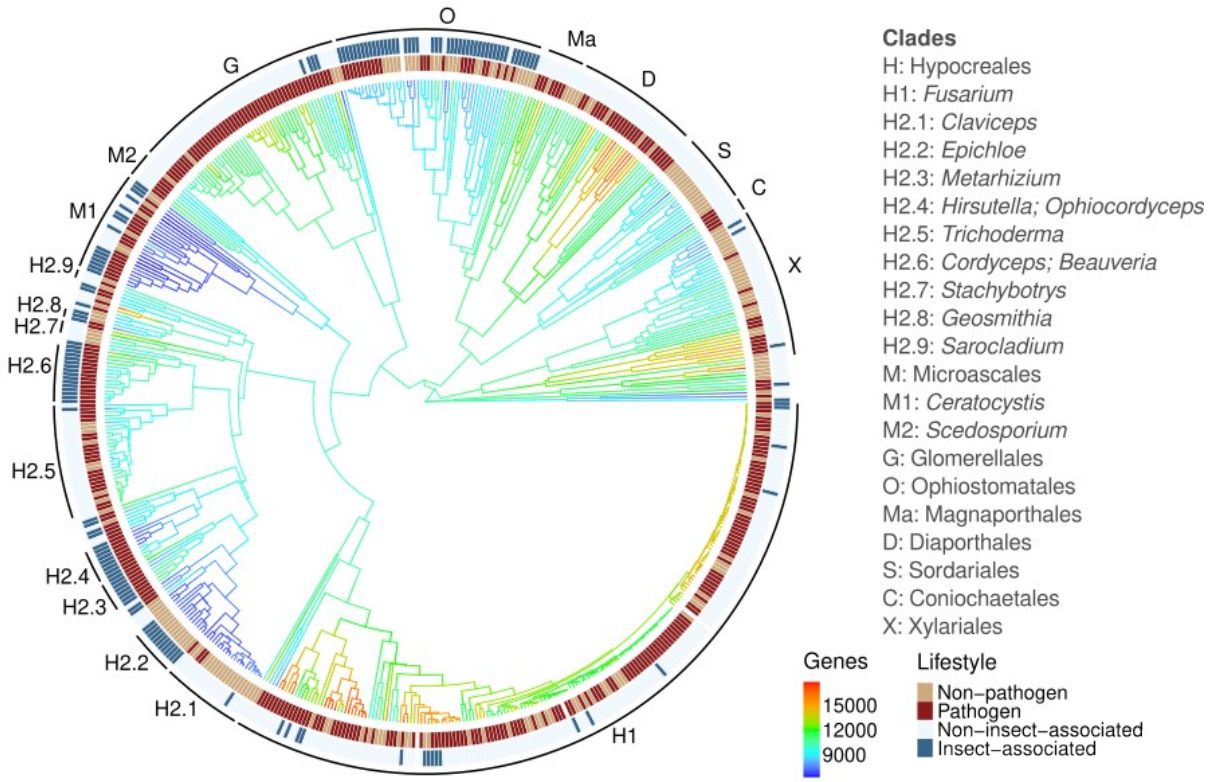

Figure 3. Substantial gene losses in some Sordariomycetes clades. Ancestral gene numbers were estimated with fastAnc function in R package phytools. A circular heatmap on the inside indicates species annotated as pathogens (red) or non-pathogenic (yellow), and the heatmap on the outside indicates species annotated as insect-associated (blue) or not (yellow). Black lines mark selected clades. The list on the right shows a corresponding order name or a representative genera for highlighted clades.

Symbiosis is one trait that often leads to streamlining of microbial genomes[26], therefore, we checked if association with insects can explain gene and genome reductions in some clades. We consider species to be associated with insects if they are symbionts, mutualists, or pathogens of insects, and we annotated a subset of clades in the phylogeny that are either composed of mostly insect-associated species or not (Figures 3-4, Table S5). First, we compared genomic traits between several pairs of insect- and non-insect-associated clades (groups a to e, Figure 4). Clades O (group a), M1 (group b), and H2.8 (group c) comprise insect vector-transmitted tree endophytes and pathogens. Clades H2.6 (group d), H2.4, and H2.3 (group e) are specialized entomopathogens. Clade H2.2 comprises fly-transmitted grass symbionts (group e). In four groups (a to d), insect-associated species have smaller genomes, smaller genomes without repeats, and longer exons than their sister clades (Figure 4). Smaller genomes are accompanied in most cases by fewer introns and fewer genes (groups b, c, and a except for clade S) or/and smaller intergenic sizes (group d). All clades from group e

show overall fewer differences between insect and non-insect-associated clades, which can be explained by the fact that, unlike other groups, they originated from an insect-associated ancestor (probability of insect-association in the ancestor = 0.97, 95% CI [0.96 - 0.99], Table S5). Consequently, all clades in group e except for one (H2.3) have among the smallest genomes without repeats and the fewest genes out of all groups. By comparing current values of genomic traits with estimated ancestral states in the most recent common ancestor for each group, we found a consistent decrease in the total size of the genome, genome size without repeats, number of genes, and number of introns in all clades which are in symbiotic/mutualistic relationship with insects (clades O, M1, H2.8, and H2.2, Figure S4).

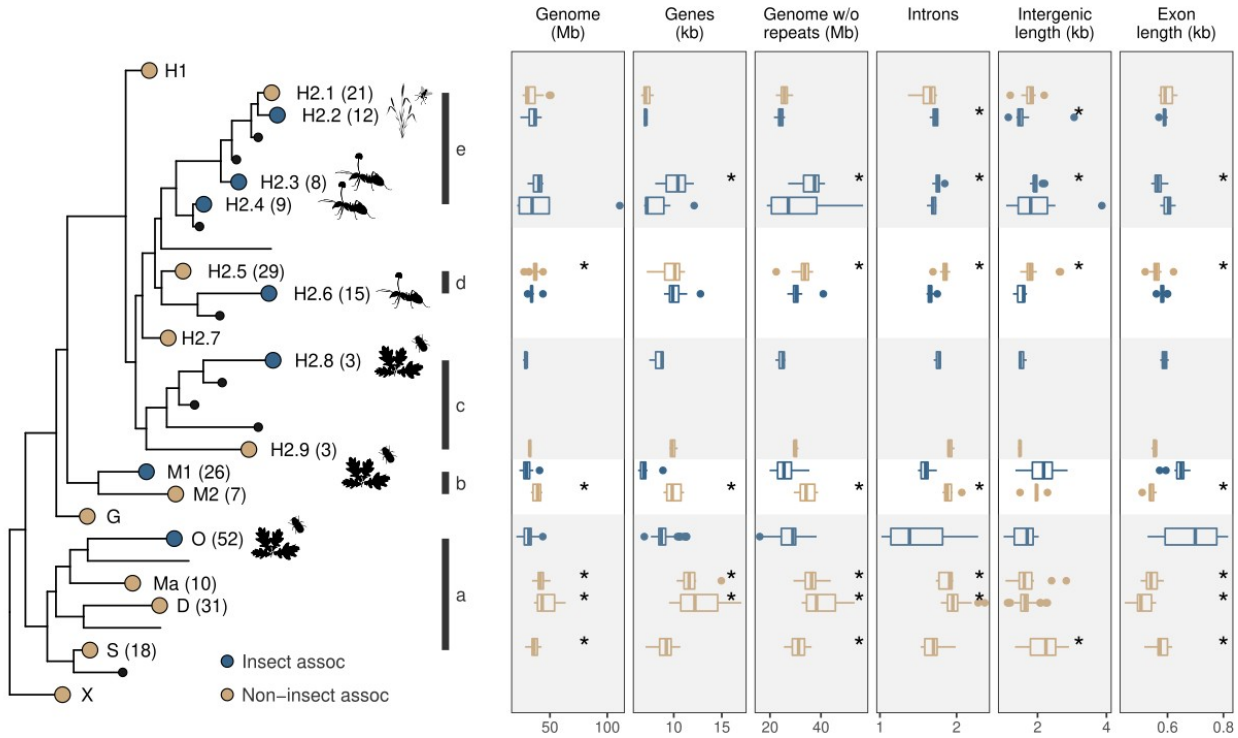

Figure 4. Genomic traits vary across insect and non-insect-associated clades. Five groups (a-e marked with vertical black lines) comparing insect-associated (blue nodes) and non-insect-associated (yellow) clades are shown. Black filled circles indicate clades composed of >1 species. Numbers shown in the parentheses indicate clade abundances. Asterisks indicate the comparisons with statistically significant pairwise differences, between blue and brown clades within each group (Mann-Whitney test, adjusted p<0.05). Clades in group c have too few species for testing. Clades O (group a), M1 (group b), H2.8 (group c), and H2.2 (group e) are insect mutualists or symbionts, whereas clades H2.6 (group d) and clades H2.4 and H2.3 (group e) are insect pathogens.

Next, we used three methods as described in the previous section, to test the coevolution of genomic traits with insect association. All three methods consistently suggest the presence of longer exons in insect-associated species, and two out of three methods confirm smaller genome size but not fewer genes in insect-associated species (Figure 5, Tables S6-S7). Analyses conducted on a dataset that excluded repeated species (n=563 in total), gave nearly identical results (Tables S6-S7, Methods), with the exception that BayesTraits inferred additional inverse coevolution between insect association and genome size without repeats, whereas phyloglm model gave no support for negative correlation of insect association with genome size without repeats and the number of introns.

Overall, these results show that insect-associated fungi are characterized by genes with longer exons and usually fewer introns and have smaller genomes. In addition, clades in symbiotic/mutualistic relationships with insects experience a reduction in genome size excluding repeats and a reduction in the number of genes.

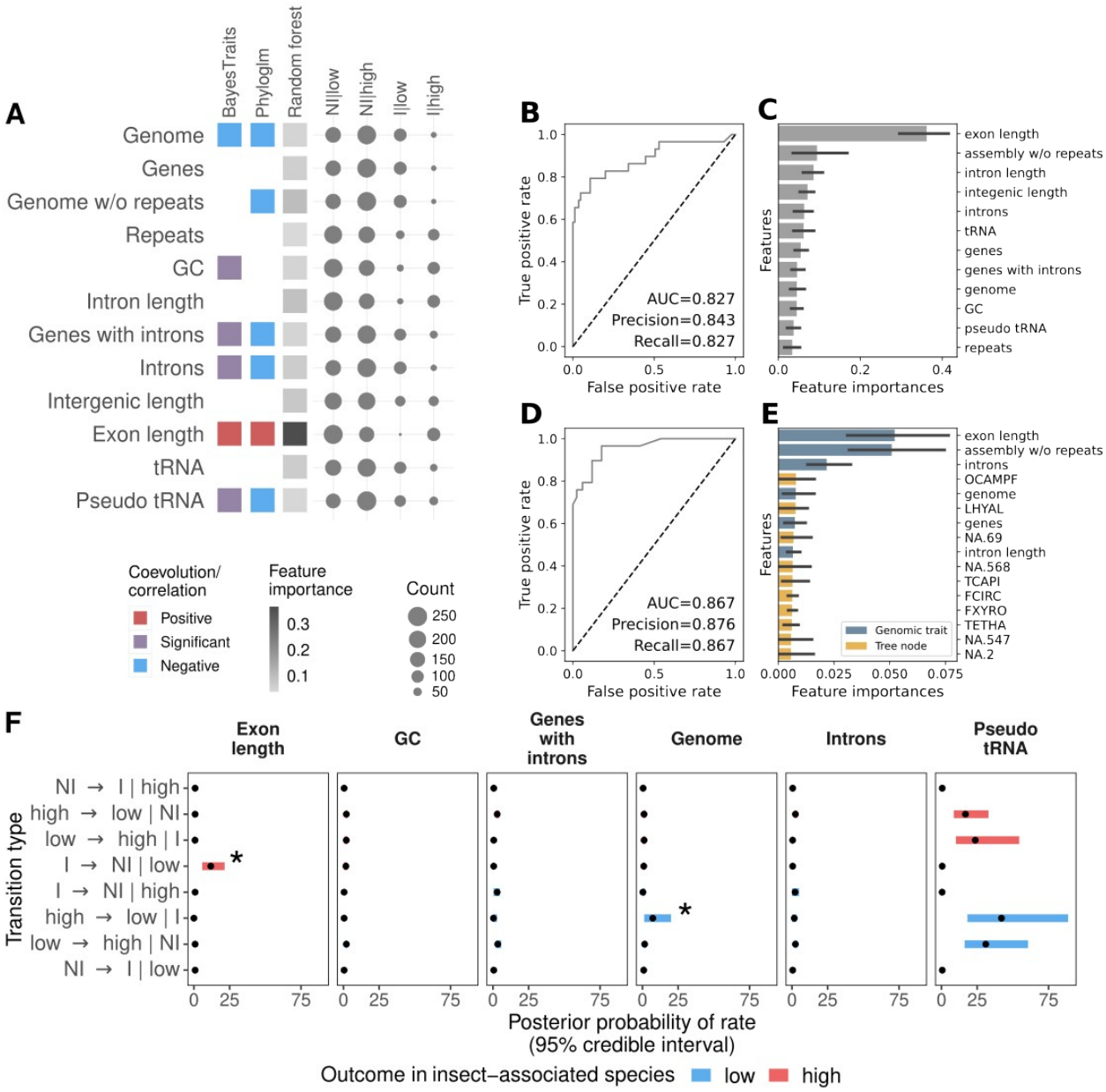
A
BayesTraits
Phyloglm
Random forest
NI|low
NI|high
I|low
I|high
Genome
Genes
Genome w/o repeats
Repeats
GC
Intron length
Genes with introns
Introns
Intergenic length
Exon length
tRNA
Pseudo tRNA
Coevolution/ correlation
Positive
Significant
Negative
Feature importance
0.3
0.2
0.1
Count
250
200
150
100
50
B
True positive rate
False positive rate
AUC=0.827
Precision=0.843
Recall=0.827
C
Features
Feature importances
exon length
assembly w/o repeats
intron length
integenic length
introns
tRNA
genes
genes with introns
genome
GC
pseudo tRNA
repeats
D
True positive rate
False positive rate
AUC=0.867
Precision=0.876
Recall=0.867
E
Features
Feature importances
exon length
assembly w/o repeats
introns
OCAMPF
genome
LHYAL
genes
NA.69
intron length
NA.568
TCAPI
FCIRC
FXYRO
TETHA
NA.547
NA.2
Genomic trait
Tree node
F
Exon length
GC
Genes with introns
Genome
Introns
Pseudo tRNA
Transition type
NI → I | high
high → low | NI
low → high | I
I → NI | low
I → NI | high
high → low | I
low → high | NI
NI → I | low
Posterior probability of rate
(95% credible interval)
Outcome in insect−associated species
low
high

Figure 5. Genomic traits coevolve with insect-associated taxa. A. Results of the three methods used for detecting correlation between insect-associated species and genomic traits: BayesTraits, phylogenetic logistic regression (phyloglm), and random forest classifier. Circles indicate frequencies of insect-associated (I) and non-insect-associated (NI) species with low (below median) or high (above median) values for 12 genomic traits. **B.** ROC curves (where the positive label is the insect-associated) and performance metrics of random forest classifier trained to distinguish insect-associated from non-insect-associated species. **C.** Top feature importances (and credible intervals) according to random forest classifier. **D** and **E** are the same as **B** and **C**, but genomic traits were combined with internode phylogenetic distances. **F.** Posterior probabilities of transition rates estimated with rjMCMC (BayesTraits) from one run of a dependent model of the evolution of insect association with 12 genomic traits. Asterisks indicate traits for which two dominant rates are not equal. I - insect association, NI - no insect association, low - genomic trait value below the median, high - genomic trait value above the median.

Losses in gene families are more frequent than gains

To evaluate the dynamics of gene losses and gains in Sordariomyctes, we estimated rates of gains and losses of gene families (groups of orthologous genes with >= 1 member) from the root of the tree for a subset (112) of species. We found that gene family losses are overall more frequent than gene family gains, in particular at the deeper branches of the tree (Figure S5). The strongest gene family losses are located on deep branches leading to subclades of H2 clade, on the branch leading to clade M1, and on the branches leading to clades O, Ma, and D. Gene family losses dominate over gains in most species but are overcome by gains only in pathogenic species from clades H1, M2, G, D, and some X. The clades H1, M2 and G did not undergo major ancient losses in their genomes. This analysis shows that gene losses are prevalent across most Sordariomycete clades and are not restricted to insect-associated fungi. Net gains of genes since the Sordariomycete common ancestor are observed only in strictly pathogenic clades.

Insect-associated plant pathogens preferentially lose genes important for host colonization

To resolve whether gene loss in insect-associated fungi is uniform across all genes or constrained to specific functional classes, we annotated *ab initio* gene models with KOG functions[28], annotated genome assemblies with Secondary Metabolite Clusters[29] (SMC), and searched gene models against databases of CAZymes[30], peptidases[31] (MEROPS, M), and transcription factor pfam domains[32] (TF). The CAZymes and SMC genes are critical for overcoming external host barriers and for entry into the host[30,33]. Therefore, selection pressures acting on them are expected to vary with lifestyles. Ancestral content of annotated gene classes was estimated for nodes at the split of each clade

with its sister clade (most recent common ancestor with a sister clade) and compared with mean content observed in the clade to obtain a fold change.

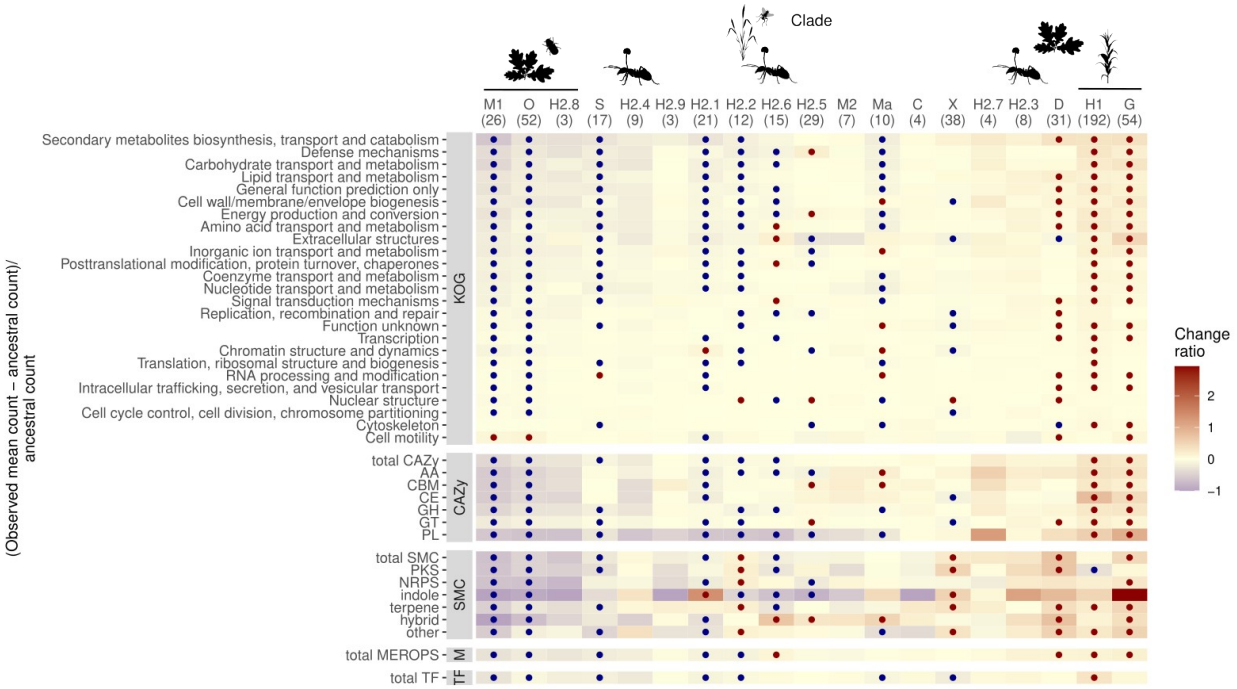

Figure 6. Insect-vectored clades lose genes involved in breaking plant host barriers. Heatmap shows the fold change of genes/clusters relative to the ancestral state ((observed - ancestral)/ ancestral state). Clades are shown in columns (clade names correspond to the ones in Figure 3) with the number of clade members in parentheses, functional classes are shown in rows. The dots indicate significant gain (red) or loss (blue) of genes/clusters across clade members estimated from 100 rounds of bootstrapping of 10 species in clades with >= 10 members. SMC: secondary metabolite clusters, M: Merops, TF: transcription factors. Drawings indicate clades which are associated with insects (M1, O, H2.8, H2.4, H2.2, H2.6, H2.3), and three right-most clades, which are specialized tree (D) and crop pathogens (H1, G).

The insect-vectored clades Ophiostomatales (O), Microascales (M1) and *Geosmithia* (H2.8) exhibit the highest gene losses across nearly all KOG groups (Figure 6). To a lesser degree gene loss is also observed in non-pathogenic clades Sordariales (S), H2.1 as well as other insect-associated clades H2.2, H2.4, and H2.6. Only one insect-pathogenic clade H2.3 has an opposite pattern. On the opposite side of the spectrum, strictly pathogenic *Fusarium* from Hypocreales (H1), Glomerellales (G), and Diaporthales (D), exhibit gene gains across most KOG groups. Loss of genes is the strongest in fungal clades in symbiotic/mutualistic relationships with insects (M1, O, H2.8), with the most prominent loss in genes involved in secondary metabolites synthesis, transport, and catabolism, defense mechanisms, and carbohydrate transport and metabolism. Indeed, all analyzed types of CAZymes, SMC, as well as peptidases, and transcription factors have contracted in these clades (Figure 6).

Only genes related to the cytoskeleton and cell motility exhibit no gain or change in number in insect-vectored O, M1, and H2.8. These clades include some important emerging tree plant pathogens, therefore, to test whether losses are experienced equally by pathogenic and non-pathogenic members of these clades, gene gain/loss analyses were repeated for pathogenic species only. Results confirm extensive losses in similar clades and KOG groups (Figure S6). Notably, KOG groups most commonly lost in insect-associated clades (and clades derived from them), include (apart from the ones mentioned above) lipid transport and metabolism, cell wall/membrane/envelope biogenesis, or energy production and conservation, and are the same KOG groups which undergo expansion in plant pathogenic clades (Figure 6, Figure S6). These results show that plant pathogens and those vectored by insects have distinct repertoires of genes and experience contrasting dynamics in genes important for host colonization.

Gene structure of lost genes

Genomes of insect-vectored species are characterized by fewer genes and genes with longer exons and fewer introns, the last two being strongly negatively correlated (Figure 1B). This pattern can be explained either by the more frequent loss of intron-rich genes or by a general trend towards a less complex gene structure. To test this, we first compared gene structure (number of exons, exon, and intron lengths) in single copy one-to-one orthologs between species belonging to two insect-vectored clades (O and M1) and their non-insect-vectored sister clades (D and M2). Longer exons among one-to-one orthologs of insect-vectored species would suggest a general trend toward less complex gene structure. Indeed, one-to-one orthologs of insect-vectored clades have longer exons (by 286 bp in O, Figure 7A, and by 140 bp in M1, Figure S7), and fewer introns (by 0.58 in O, Figure 7A, and by 0.33 in M1, Figure S7). Intron length is either longer (by 9 bp in M1, Figure S7) or shorter (by 10 bp in O, Figure 7A) compared to non-insect-vectored species.

As a second test, we looked at the gene structure of gene families relative to the occupancy of the gene family in the clade. If complex genes are more likely to get lost, we should observe rare genes (present in a few clade members) to have more introns (and exons) and shorter exons than common genes. Gene families that are frequently lost since the common ancestor of the clade, indeed have shorter exons, however, they also have fewer exons, contrasting our first hypothesis that more complex genes are lost more often (Figures 7B and S7, Poisson regression with 2 predictors: exon length and count, and species occupancy as a response, $p<3.61e-08$ for all clades). This trend is significant in both insect- and non-insect-vectored clades (Figure 7B and S7).

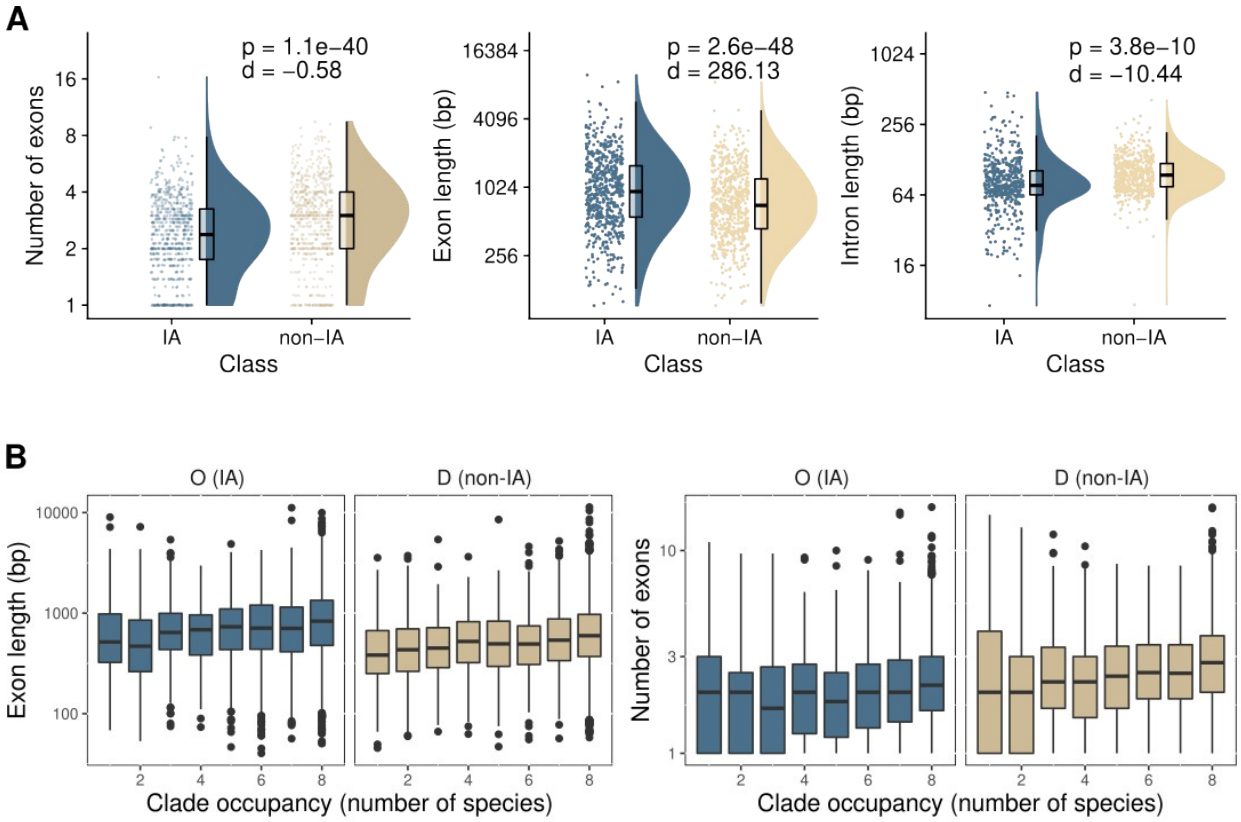

Figure 7. Gene structure changes in insect-associated clades affect all genes. A. Comparison of gene structures between one-to-one single-copy orthologs (n=583) from insect-associated (IA) clade O (blue) and the corresponding non-insect-associated (non-IA) clade D (yellow). Intron length was compared only between orthologs with at least one intron (n=476). Ortholog features were averaged across eight species. Boxplots show medians, first and third quartiles, and lines span minimum to maximum values excluding outliers. p - p-values estimated with the paired two-sided Wilcoxon signed-rank test; d - mean differences between IA and non-IA orthologs. **B.** Average exon length and the number of exons of gene families (orthogroups) present in different frequencies across up to eight clade members.

Discussion

We investigated the impact of fungal lifestyle on the evolution of genome size and complexity using 562 Sordariomycetes genomes. We analyzed 12 genomic traits and two lifestyle traits: pathogenicity and association with insects and found several patterns of genome evolution in this group of fungi. First, genome size is strongly positively correlated with the repeat content and the number of genes in the genome. Second, fungi with the most streamlined genomes (smaller genomes, less introns and genes) belong to groups that associate with insects that serve as hosts, vectors or symbionts to fungi, or to groups directly derived from them. Third, pathogens exhibit distinct patterns in gene dynamics, with most plant and animal pathogens having increased gene numbers, tRNAs and number of introns per gene, and plant pathogens vectored by insects showing opposite patterns with frequent losses of genes and introns.

Impact of the lifestyle on genome size in Sordariomycetes

The strongest predictor of Sordariomyctes genome size is repeat content and to a lesser degree other non-coding DNA, such as intergenic length and intron length. Accumulation and expansion of non-coding DNA support mutation-hazard hypothesis in explaining genome size expansions, according to which small effective population size (Ne) leads to stronger drift and fixation of deleterious DNA. Eukaryotic species with smaller Ne accumulate more noncoding DNA[34]. In the case of fungi, drift can affect the amount of the repeat content and other non-coding DNA in some species or clades, but this trend is not associated with any particular lifestyle (note however that pathogens tend to have genes with more introns). Both pathogens (*Monosporascus*, *Ophiocordyceps*) and non-pathogens (*Claviceps*, *Epichloe*) experienced bouts of repeat expansions, which could be linked to some stochastic events of low Ne in these species, rather than a specific occupied ecological niche.

As one of the important contributors to repeat content, transposable elements (TEs) are often associated with pathogens. There are well-documented cases of virulence factors emerging due to transposable element activity. Despite a few cases of pathogens with an expanded repeat content, suggesting an increased TE activity, in general, pathogenic Sordariomycetes do not carry many repeats. In contrast, an average pathogen experiences a decrease in the number of repeats. This pattern can be caused by an effective selection in removing deleterious noncoding DNA, and/or by a strong bias toward deletions in the genome. Pathogens with an expanded TE content may indicate rare cases of fungi that developed means to accommodate fast-evolving deleterious DNA, for instance in accessory chromosomes, chromosomal duplications,

separate genome compartments, or experienced events of low Ne, due to transmissions, or multiple rounds of population extermination[24].

The second major predictor of genome size is gene content. Expansion and diversification of genes is usually linked to adaptation to a specific niche, for instance in mycorhizal[35] or wood decomposing fungi[36]. We find that an overall increase in gene repertoire is linked to a pathogenic lifestyle, and this pattern is visible in pathogens from different Sordariomycetes clades, except for vector-transmitted tree pathogens. The genomes of pathogens are in the upper part of the genome size scale in our dataset, a pattern that is mostly driven by gene expansions since the repeat content is limited in most pathogens. Nevertheless, the largest genomes belong to species that exhibit both high repeat content and many genes (for instance pathogenic *Colletotrichum* species).

Sordariomycetes with the smallest genomes are almost exclusively endosymbionts, insect endoparasites, are vectored by insects, or are directly derived from clades associated with insects. They evolved smaller genomes, genes with longer exons (and fewer introns), and in the case of insect-vectored species, they have also lost many genes. These patterns of genome streamlining are reminiscent of the reductive genome evolution in endosymbiotic and endoparasitic prokaryotes, as well as some endoparasitic fungi. Gene-rich genomes of prokaryotes show an inverse correlation between drift and genome size, explained mainly by the bias towards deletions[14,37]. Loss of genes in endosymbiotic and some free-living bacteria has been also explained by selection against non-essential genes[38], a loss of redundant genes with drift[39], or increased mutation rate[40]. Our results show that in two major insect-vectored clades (Ophiostomatales and Microascales) the change in gene architecture towards longer exons and fewer introns is consistent across all the genes and is not caused by the retention of genes with such structure. Moreover, these clades exhibit the longest branches in the phylogeny (measured by amino-acid substitutions), an observation that can be explained by the fixation of deleterious substitutions when Ne is small. Consistent gene structure and long branches support the impact of drift on the genome complexity in species with the reduced Ne.

Gene expansions in Sordariomycetes pathogens

Pathogenicity has evolved multiple times across Sordariomycetes in nearly every order. As we were unable to assign the probability of pathogenicity with high accuracy for deeper branches in the tree, our findings confirm that this is a fast-evolving trait. Despite the emergence of pathogenicity across independent taxonomic groups, several genomic traits proved to be reappearing among most pathogens, and these are high overall numbers of genes, tRNA genes, larger genomes without repeats, genes with longer introns, and less repeat content. Expanded gene numbers are most evident in plant

pathogenic species such as *Fusarium* (Hypocreales), *Colletotrichum,* and *Verticillium* (Glomerellales) or *Diaporthe* (Diaporthales), but also in some (not all) entomopathogens (*Metarhizium*), mycopathogens (*Trichoderma*) and opportunistic human pathogens (*Scedosporium*, *Sporothrix*). This implies that most pathogens (excluding those that are vectored by insects) benefit from expanded gene repertoires, for example in a more effective establishment in host tissues and/or defense against host immunological mechanisms. Indeed, we see that the largest gains of genes in pathogens correspond to genes involved in secondary metabolite synthesis, carbohydrate metabolism, and defense mechanisms.

One group of genes that coevolves with pathogenicity are tRNAs. These observations are more difficult to interpret. However, in fungi and other microbes, post-translational modifications of tRNAs are known to play a role in triggering virulence and evading host defense[41–43]. Duplications of existing tRNAs may be an alternative way to maintain a variety of tRNAs to cope with the host immune system.

Another trait observed in pathogens is an increased number of introns per gene. Previous studies have shown that genes of fungal ancestors were intron-rich and several events of massive intron loss occurred across fungi, leading to intron-poor groups including Saccharomycotina as well as Pezizomycotina, which comprise Sordariomycetes[44]. Despite an overall low number of introns, we observe an increasing trend of introns in pathogens. Introns play an important role in alternative splicing, and this process has been found to occur more frequently in pathogenic species, affecting genes involved in dimorphism and stress response, essential functions when entering host environment[45,46].

Different modes of genome evolution in Sordariomycetes pathogens

Our results show that the genome complexity of Sordariomyctetes fungi is influenced by an interplay between drift and adaptation, both of which are affected by a specific lifestyle of the pathogen. The distinctions between lifestyles are visible in particular in the evolution of the gene content and gene structure. Based on these traits alone, three groups of pathogens can be distinguished: i) specialized plant pathogens with expanded gene repertoires and intron-rich genes, ii) insect pathogens with low to high gene expansions counteracting strong ancient gene contractions, and with long-exon genes, and iii) insect-vectored plant pathogens with overall gene contractions including many genes responsible for host adaptation and pathogenicity, and long-exon genes. Considering these observations, one can imagine a hypothetical scenario of genome evolution in Sordariomycetes, in which species constantly reduce their genome size due to the overall bias towards deletions. This trend would be exacerbated in species that evolve in symbiosis, where drift is strong and many genes become redundant and thus

dispensable due to presence of host gene products, eventually leading to genome size contraction. But the trend of gene loss would be reversed if the species evolved as pathogens or switched from insect-symbiosis to insect-pathogenicity, in which case their gene repertoires would increase through duplications. Finally, species that are neither pathogenic nor symbiotic, would constantly reduce their genomes due to an overall high deletion rate (but not small Ne), though not to the same degree as symbiotic species (as can be exemplified by non-pathogenic Sordariales).

Methods

Genome assemblies

14 strains (11 x *Ophiostoma* and 3 x *Leptographium*, Table S2) were selected for high-depth short-read Illumina sequencing. Isolates were grown on Malt Extract Agar medium (15 g l^{-1} agar, 30 g l^{-1} malt extract, and 5 g l^{-1} mycological peptone), and DNA was extracted with acetyl trimethylammonium bromide chloroform protocol. Library preparation and sequencing were conducted at the Génome Québec Innovation Center (Montréal, Canada). One strain (*O. quercus*) was sequenced with Illumina NovaSeq (paired-end 150 bp), and the rest with Illumina HighSeq X (paired-end 150 bp). Data quality was inspected with Fastqc v0.11.2/8[47]. Ten strains were used for *de novo* genome assembly (Table S2). The depth of coverage of the generated data was between 32x and 555x (Table S2). Reads were trimmed for adapters with Trimmomatic v0.33/0.36[48] using options 'ILLUMINACLIP:adapters.fa:6:20:10 MINLEN:21' and overlapping reads were merged with bbmerge from BBTools v36/v37[49]. *De novo* genome assemblies were generated with SPAdes v3.9.1[50] with options '-k 21,33,55,77,99 --careful'. Mitochondrial DNA was searched in contigs with NOVOPlasty v3.8.3[51] using the mitochondrial sequence of *O. novo-ulmi* as a bait (CM001753.1) and matching contigs were removed. Reads were remapped to the nuclear assembly with bwa mem v0.7.17[52], and contigs with normalized mean coverage < 5% and shorter than 1000 bp were also removed.

A total of 11 *Ophiostoma* strains were selected for long-read sequencing with PacBio (Table S2). DNA extraction was done in the same way as for Illumina libraries, except that no vortexing and shaking was done to avoid DNA fragmentation. Library preparation and sequencing with the PacBio SMRTcell Sequel system were conducted at the Génome Québec Innovation Center (Montréal, Canada). Average depth of coverage of the generated data was between 47x and 269x. *De novo* genome assemblies were generated with pb-falcon v2.2.0[53]. The range of parameters was tested and the final configuration files with the best-performing parameters for each species are on https://github.com/aniafijarczyk/Fijarczyk_et_al_2022. *O. quercus* was sequenced in two runs and read data from the two runs were combined together. *O. novo-ulmi* (H294) was sequenced in two runs but only read data from one run was used for an assembly due to sufficient coverage. Assemblies were ordered according to *O. novo-ulmi* H327 genome (GCA_000317715.1) using Mauve snapshot-2015-02-13[54]. Assemblies were polished between two to four times by remapping long reads with minimap2 (pbmm v1[55]) and correcting assemblies using arrow (pbgcpp v1[56]). We also performed one round of polishing with pilon v1.23[57] after mapping short-read Illumina reads (bwa v0.7.17[52]) from this study (three assemblies) or from the previous study[58] (seven assemblies, Table S2). The only exception is *O. quercus*, for which we had no

corresponding short-read data, therefore we performed four rounds of correction using arrow. The effectiveness of polishing was assessed by analyzing the completeness of genes with Busco v3[59]. Mitochondrial genomes were assembled using Illumina assemblies as baits (or those of related species). Long reads were mapped to bait assembly with pbmm2[55], and a subsample of mapped reads was used for mtDNA assembly with mecat v2[60]. Consecutive rounds of mapping and assembly were conducted until circularized assemblies were obtained. Nuclear assembly contigs with more than 50% of low-quality bases, those mapping to mtDNA, or with no mapping of Illumina reads (standardized mean coverage < 5%) were filtered out. The genome assembly of *O. montium* was very fragmented and had a high percentage of missing conserved genes (74.3%) so instead, an Illumina *de novo* assembly for this species was considered in further analysis.

In both short-read and long-read assemblies, contigs were searched for viral or bacterial contaminants by BLAST searches against bacterial or viral UniProt accessions and separately against fungal UniProt accessions. Contigs having more bacterial/viral hits in length than fungal hits were marked as contaminants and removed. Finally, repeats were identified with RepeatModeller v2.0.1[61] using option -LTRStruct, and recovered repeat families together with fungal repeats from RepBase were used for masking assemblies with RepeatMasker v4.1.0[62].

Genes were annotated with Augustus v.3.3.2[63] and Breaker v2.1.2[64]. To obtain *ab initio* gene models in *Ophiostoma novo-ulmi, ulmi, himal-ulmi, quercus* and *triangulosporum*, a training file was generated using RNA-seq reads from *O. novo-ulmi* H327 (SRR1574322, SRR1574324, SRR2140676) mapped onto a *O. novo-ulmi* H294 genome with STAR v2.7.2b[65]. A training set of genes was generated with GeneMark-ES-ET v4.33[66]. Final gene models were merged from *ab initio* models, models inferred from the alignment of RNA-seq reads (except *O. triangulosporum*) and *O. novo-ulmi* H327 proteins. The rest of the genome assemblies were annotated only *ab initio* using Magnaporthe grisea species training file.

Sordariomycetes genomes

580 Sordariomycetes genome assemblies (one reference genome per species) were downloaded from NCBI (21/05/2021), the two assemblies were downloaded from JGI Mycocosm (*S. kochii, T. guianense*), and two from other resources (*O. ulmi* W9[67], *L. longiclavatum*[68]). All assemblies were filtered for short contigs (< 1000 bp), and contaminants as described above. Gene completion was assessed with Busco v3[59] using an orthologous gene set from Sordariomycetes. *Ab initio* gene models were obtained with Augustus v.3.3.2[63] with different species-specific training files: Magnaporthe grisea, Fusarium, Neurospora, Verticillium longisporum1 or Botrytis cinerea (Table S1). 31 assemblies were removed based on the low percentage of

complete conserved genes (Busco score < 85%) and one due to an excessive number of gene models (*Botrytis cinerea*).

Phylogeny

The maximum likelihood tree was built from 1000 concatenated genes (retrieved with Busco v3[59]) with the highest species representation. Protein alignments were generated with mafft v7.453[69] with E-INS-i method (option '--genafpair --ep 0 --maxiterate 1000'), trimmed with trimal v1.4.rev22[70] with option '-automated1' and converted into a matrix. Best protein evolution model JTT+I+G4 was chosen as the most frequent model across all protein alignments, based on the BIC score in IQ-TREE v1.6.12[71,72]. The maximum likelihood tree was inferred with ultrafast bootstrap[73], seed number 17629, and 1000 replicates implemented in IQ-TREE. Time-scaled phylogeny was inferred with program r8s v1.81[74], setting the calibration point of 201 My at the split of *Neurospora crassa* and *Diaporthe ampelina*, retrieved from TimeTree[75].

To confirm the branching order of Hypocreales, Microascales, and Glomerellales, another maximum likelihood tree was built, based on a subset of 250 protein alignments (250 longest sequences out of 1000), with a separate model partition assigned to each protein[76]. The maximum likelihood tree was inferred with ultrafast bootstrap, seed number 17629, and 1000 replicates implemented in IQ-TREE v1.6.12[71,73].

In the third method for phylogeny inference, we used the multispecies coalescent-based method to obtain a consensus topology based on 1000 separate protein-based trees. The maximum likelihood tree was generated for each of 1000 protein alignments, with their corresponding protein evolution model, ultrafast bootstrap, and 1000 replicates using IQ-TREE v1.6.12[71,73]. Tree nodes with support of less than 30 were contracted using newick-utils v1.6[77]. Consensus topology was obtained with Astral v5.7.8[78].

Genomic and ecological traits

Pathogenicity and insect association were assigned based on a literature search. Pathogenicity was assigned if the species caused a well-recognized disease, or pathogenicity was experimentally documented at least on one host species. We considered pathogens of plants, animals, and fungi, both obligatory and opportunistic. Insect association included all types of relationships with insects, including pathogenic, symbiotic, and mutualistic. Insect-vectored species were limited to documented cases of insect transmission. Analyzed genomic traits included genome size (bp), number of genes, a fraction of repeat content, size of the assembly excluding repeat content (bp), GC content, the mean number of introns per gene, mean intron size (bp), mean exon size (bp), a fraction of genes with introns, mean intergenic length (bp) and number of tRNA and pseudo tRNA genes. Genome size is equivalent to assembly size after filtering. Genome size excluding repeat content is assembly size excluding regions

masked by RepeatMasker (in this study or downloaded from NCBI), and repeat content is the proportion of masked bases compared to total assembly size. Gene number corresponds to the number of all *ab initio* gene models obtained with Augustus. GC content is the proportion of GC bases in the total assembly. The mean number of introns per gene, intron and exon size, a fraction of genes with introns and intergenic length were estimated from gff files with annotated gene models. tRNA and pseudo tRNA genes were obtained with tRNAscan-SE v2.0.9[79].

Trait correlations

Correlations between genomic traits were calculated using phylogenetic generalized least squares (gls) function from the R package nlme and independent contrasts using pic function from the R package ape[80]. P-values in phylogenetic gls models were obtained using anova() function, and adjusted for multiple tests with Benjamini-Hochberg method at false discovery rate of 0.05. Some genomic traits were rescaled: genome (bp x 10^{-7}), genes (x 10^{-4}), assembly wo repeats (bp x 10^{-7}), intron length (bp x 10^{-4}), exon length (bp x 10^{-3}), intergenic length (bp x 10^{-4}), tRNAs (x 10^{-3}), and pseudo tRNAs (x 10^{-3}).

Coevolution of ecological traits (pathogenicity and insect association) with genomic traits was estimated using three approaches, i) a reversible jump MCMC discrete model of evolution implemented in BayesTraits v3[81], ii) a phylogenetic logistic regression with phyloglm function in R package phylolm[82], iii) an investigation of feature's importances from machine learning classification using scikit-learn python (v3) package. All analyses were additionally run for the subset of species (n=563), in which any repeating species were excluded.

For analysis in BayesTraits, genomic traits were converted into binary traits based on their median (0 if below median and 1 if above median). Four different models were investigated: independent evolution, dependent evolution, covarion model where dependent evolution varies across the tree[83], and for a subset of significant genomic traits additional model testing equal transition rates. To determine the direction of coevolution with ecological traits (positive or negative), two models were compared, one in which all transition rates were allowed to differ (dependent model), and the second one in which the two most frequent transition rates were set to be equal. If the model with differing transition rates performed better, the direction of coevolution was determined by the most frequent transition rate, otherwise, no specific direction was concluded. Models were tested by comparing complex to simpler models with log bayes factor, i.e., dependent vs. independent, covarion vs. dependent, dependent vs. dependent with selected equal transition rates. A complex model was chosen if the log bayes factor surpassed 5. Each model was run three to five times to check for consistency with 21 mln iterations, 1 mln burn-in, and thinning of 1000. Analyses for the

subset of species (n=563), in which repeating species were excluded, were run in three independent runs for the dependent, independent model, and for the dependent model with constrained rates.

In the second approach, a phylogenetic logistic regression was used to fit each genomic trait to ecological trait using phyloglm function in R with the “logistic_MPLE” method, btol option (searching space limit) set to 30, and 1000 independent bootstrap replicates. Benjamini-Hochberg correction was applied to p-values. Species with missing information on the ecological trait were filtered out.

In the last approach, genomic traits (features) were used to train several machine learning classifiers for the prediction of ecological traits. Species with missing data were filtered out, and data were rescaled (between 0 and 1) for classification with SVC. Twenty percent of data was selected as held-out data. First, the performance of several classifiers was investigated (KNeighbors, SVC, DecisionTree, RandomForest, and GradientBoosting) with a balanced accuracy score, and 5-fold cross-validation, in which train set (80%) was split into 5 parts (with two classes of a predicted trait in same proportions), and accuracy measured 5 times, each time a different part being set as a test set and the remaining four parts as the training set. Hyperparameters for the best-performing classifier were chosen using grid search. Performance on held-out data was evaluated with a balanced accuracy score, precision, recall, and ROC curve. To account for the role of phylogeny, a matrix of internode distances was used in combination with other genomic features. The classification models were evaluated in the same way and compared with those not integrating phylogenetic distances. Feature importances were obtained with a mean decrease impurity method from the RandomForest classifier.

For the trait of pathogenicity, the RandomForest classifier performed best with an accuracy of 0.7. Phylogeny had a negligible effect on model performance, improving it only to 0.71. Top features included all genomic features, and two phylogenetic nodes corresponding to two overlapping clades with *Claviceps* species excluding *C. paspali*, and *C. citrina*. The same model run on a filtered dataset (n=563), excluding repeating species, gave an accuracy below 0.7, therefore we did not report resulting feature importances.

For the trait of insect association, the RandomForest classifier also gave the best performance, giving an accuracy of 0.83, and 0.87 when the phylogenetic matrix was included. Both approaches revealed exon length and genome size without repeats as the two most important features, and intron length or number of introns as a third feature, respectively. The model ran on a filtered dataset (n=563), excluding repeating species, gave an accuracy of 0.82% and 91% for the approaches without and with the

phylogenetic matrix, respectively. The most important features of the first approach were: exon length, assembly w/o repeats, intergenic length, and intron length, and of the second approach: exon length, assembly without repeats, number of genes, and number of introns.

To compare the ancestral with the current states of genomic traits at focal nodes in the tree, ancestral states were estimated with a continuous model for genomic traits implemented in BayesTraits v3[84] using the same MCMC parameters as in the case of discrete traits models. Similarly, ecological traits were reconstructed with a discrete model in BayesTraits v3[81]. To visualize evolution of traits on the tree, ancestral states of genomic traits were reconstructed with "ER" model implemented in ace function from R package phytools[85], and discrete ecological traits were reconstructed with fastAnc function from R package phytools[85]. Traits were visualized on the tree with ggtree R package[86] using the continuous option.

Functional gene classes

Functional gene annotations were determined using several databases. Protein sequences were searched against KOG (release 2003[28]) and MEROPS Scan Sequences (release 12.1[31]) using diamond v2.0.9[87] with options '--more-sensitive -e 1e-10' and against CAZymes (download 24/09/2021[30]) with options '--more-sensitive -e 1e-102'. Pfam domains were searched with pfam_scan.pl[32] script with hmmer v3.2.1[88] and matched with Sordariomycetes transcription factors obtained from JGI Mycocosm database (download 9/12/2021). Secondary Metabolite Clusters were inferred from genome assemblies using antiSMASH v4.0.2[29]. Ancestral states of the gene and gene cluster numbers were inferred using the fastAnc function from R package phytools[85]. Per clade change ratio was calculated as a mean difference between the observed number of genes in each clade member and ancestral state divided by ancestral state. Mean gains or losses of genes per clade were tested with a bootstrap of 10 species with 100 replicates for clades with >= 10 members.

Orthologs and orthogroups

A subset of 112 species was selected, by randomly selecting 1-8 species per clade. Orthologs were searched using Orthofinder v2.5.2[89] with no *a priori* defined tree. Orthogroup expansions and contractions were detected with CAFE v4.2.1[90] with a p-value threshold of 0.05. Genes belonging to one-to-one orthologs and orthogroups were determined from the Orthofinder output, and the presence of orthogroups in the ancestral clades was determined from CAFE output.

Limitations of the study

The major challenge of this study is the assignment of lifestyle traits. Two reasons account for this. First, information on the given lifestyle is limited, and second, species

can potentially exert a given phenotype only in specific conditions, but that has not been reported or tested. These reasons may account for some inevitable uncertainty in both traits: pathogenicity and insect association. The trait of insect association is typically characteristic of individual clades; therefore, most members of the clade are expected to carry this trait or to be closely related to species with that trait. Therefore, we can alleviate the misannotation by comparing insect-associated clades with their corresponding non-insect-associated sister clades, as was done in the fourth part of the Results. Pathogenicity on the other hand is a fast-evolving trait, exerting a large variety of effects on the host. Nevertheless, the inferences of coevolution between pathogenicity and genomic traits are generally consistent with observations in clades enriched for well-studied pathogenic species. Because of the lack of ecological information for many sequenced fungal genomes, many details on the fungal lifestyles could not be included. One trait that is known to impact genome evolution is the host range of pathogens[91]. The availability of the host, as well as the ability to switch from pathogenic to saprotrophic lifestyle, can potentially have a great impact on Ne, and therefore genome evolution.

Gene annotations and genome processing was performed in a uniform way for all the species, to remove any software-related biases. Repeat content was obtained from NCBI repeat annotations together with downloaded assemblies, therefore the quality of annotations is expected to vary mostly with the assembly quality. Short-read assemblies which dominate in the dataset could cause an underestimation of the true number of repeats, however, the cases of repeat expansions in our dataset can be detected in both contiguous and fragmented assemblies. Although the annotations are mostly underestimated, our dataset allows us to detect general trends in both genes and repeat dynamics.

Several more detailed aspects of genome architecture could be informative about pathogen evolution. For instance, we did not consider the level of compartmentalization of the genome, which plays an important role in the evolution of virulence genes[92]. Accessory chromosomes or chromosome duplication are also important factors in generating variation and aiding in adaptation to the host. In fact, we know little about the frequency of these processes in non-pathogenic species. Similarly, the distribution of specific genes, such as tRNAs or repeats in the genome can inform us about the architecture of pathogenic genomes. Finally, whole genome duplications are processes that can lead to major lifestyle transitions and at the same time can influence the level of gene duplication and loss leading to major genome rearrangements.

Data Availability

Raw short and long reads from sequenced genomes are available at NCBI under project number PRJNA841745. Genome assemblies have been deposited at GenBank under the accessions JANSLN000000000-JANSMH000000000. The versions described in this paper are versions JANSLN010000000-JANSMH010000000. Code used in this study is available on github (https://github.com/aniafijarczyk/Fijarczyk_et_al_2022).

Acknowledgments

We thank Erika Dort for an access to lifestyle database to confirm pathogenicity traits of 90 species from our dataset. We thank Rohan Dandage, Ilga Porth and Louis Bernier for comments and discussions about the manuscript. This project was funded by Genome Canada and Genome Québec BioSAFE project and a NSERC Discovery grant to CRL. CRL holds the Canada Research Chair in Cellular Systems and Synthetic Biology.

References

1. Blommaert, J. (2020). Genome size evolution: towards new model systems for old questions. Proc. Biol. Sci. *287*, 20201441.

2. Lynch, M., and Conery, J.S. (2003). The origins of genome complexity. Science *302*, 1401–1404.

3. Petrov, D.A. (2002). Mutational equilibrium model of genome size evolution. Theor. Popul. Biol. *61*, 531–544.

4. Gregory, T.R. (2002). A bird's-eye view of the C-value enigma: genome size, cell size, and metabolic rate in the class aves. Evolution *56*, 121–130.

5. Decena-Segarra, L.P., Bizjak-Mali, L., Kladnik, A., Sessions, S.K., and Rovito, S.M. (2020). Miniaturization, Genome Size, and Biological Size in a Diverse Clade of Salamanders. Am. Nat. *196*, 634–648.

6. Hultgren, K.M., Jeffery, N.W., Moran, A., and Gregory, T.R. (2018). Latitudinal variation in genome size in crustaceans. Biol. J. Linn. Soc. Lond. *123*, 348–359.

7. Grotkopp, E., Rejmánek, M., Sanderson, M.J., and Rost, T.L. (2004). Evolution of genome size in pines (Pinus) and its life-history correlates: supertree analyses. Evolution *58*, 1705–1729.

8. Yu, J.P., Liu, W., Mai, C.L., and Liao, W.B. (2020). Genome size variation is associated with life-history traits in birds. J. Zool. *310*, 255–260.

9. Bilinski, P., Albert, P.S., Berg, J.J., Birchler, J.A., Grote, M.N., Lorant, A., Quezada, J., Swarts, K., Yang, J., and Ross-Ibarra, J. (2018). Parallel altitudinal clines reveal trends in adaptive evolution of genome size in Zea mays. PLoS Genet. *14*, e1007162.

10. Malerba, M.E., Ghedini, G., and Marshall, D.J. (2020). Genome Size Affects Fitness in the Eukaryotic Alga Dunaliella tertiolecta. Curr. Biol. *30*, 3450–3456.e3.

11. Petrov, D.A. (2001). Evolution of genome size: new approaches to an old problem. Trends Genet. *17*, 23–28.

12. Mohanta, T.K., and Bae, H. (2015). The diversity of fungal genome. Biol. Proced. Online *17*, 8.

13. Kelkar, Y.D., and Ochman, H. (2012). Causes and consequences of genome expansion in fungi. Genome Biol. Evol. *4*, 13–23.

14. Mira, A., Ochman, H., and Moran, N.A. (2001). Deletional bias and the evolution of bacterial genomes. Trends Genet. *17*, 589–596.

15. Stajich, J.E. (2017). Fungal Genomes and Insights into the Evolution of the Kingdom. Microbiol Spectr *5*.

16. Muszewska, A., Taylor, J.W., Szczesny, P., and Grynberg, M. (2011). Independent

subtilases expansions in fungi associated with animals. Mol. Biol. Evol. *28*, 3395–3404.

17. Baroncelli, R., Amby, D.B., Zapparata, A., Sarrocco, S., Vannacci, G., Le Floch, G., Harrison, R.J., Holub, E., Sukno, S.A., Sreenivasaprasad, S., et al. (2016). Gene family expansions and contractions are associated with host range in plant pathogens of the genus Colletotrichum. BMC Genomics *17*, 555.

18. Sipos, G., Prasanna, A.N., Walter, M.C., O'Connor, E., Bálint, B., Krizsán, K., Kiss, B., Hess, J., Varga, T., Slot, J., et al. (2017). Genome expansion and lineage-specific genetic innovations in the forest pathogenic fungi Armillaria. Nat Ecol Evol *1*, 1931–1941.

19. Croll, D., and McDonald, B.A. (2012). The accessory genome as a cradle for adaptive evolution in pathogens. PLoS Pathog. *8*, e1002608.

20. Mat Razali, N., Cheah, B.H., and Nadarajah, K. (2019). Transposable Elements Adaptive Role in Genome Plasticity, Pathogenicity and Evolution in Fungal Phytopathogens. Int. J. Mol. Sci. *20*.

21. Raffaele, S., and Kamoun, S. (2012). Genome evolution in filamentous plant pathogens: why bigger can be better. Nat. Rev. Microbiol. *10*, 417–430.

22. Bao, J., Chen, M., Zhong, Z., Tang, W., Lin, L., Zhang, X., Jiang, H., Zhang, D., Miao, C., Tang, H., et al. (2017). PacBio Sequencing Reveals Transposable Elements as a Key Contributor to Genomic Plasticity and Virulence Variation in Magnaporthe oryzae. Mol. Plant *10*, 1465–1468.

23. Wacker, T., Helmstetter, N., Wilson, D., Fisher, M.C., Studholme, D.J., and Farrer, R.A. (2021). Two-speed genome expansion drives the evolution of pathogenicity in animal fungal pathogens. bioRxiv, 2021.11.03.467166.

24. Stukenbrock, E.H., and Croll, D. (2014). The evolving fungal genome. Fungal Biol. Rev. *28*, 1–12.

25. Rhodes, J., Abdolrasouli, A., Dunne, K., Sewell, T.R., Zhang, Y., Ballard, E., Brackin, A.P., van Rhijn, N., Chown, H., Tsitsopoulou, A., et al. (2022). Population genomics confirms acquisition of drug-resistant Aspergillus fumigatus infection by humans from the environment. Nat Microbiol *7*, 663–674.

26. Katinka, M.D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., et al. (2001). Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature *414*, 450–453.

27. Hongsanan, S., Maharachchikumbura, S.S.N., Hyde, K.D., Samarakoon, M.C., Jeewon, R., Zhao, Q., Al-Sadi, A.M., and Bahkali, A.H. (2017). An updated phylogeny of Sordariomycetes based on phylogenetic and molecular clock evidence. Fungal Divers. *84*, 25–41.

28. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., et al. (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics *4*, 41.

29. Blin, K., Pascal Andreu, V., de Los Santos, E.L.C., Del Carratore, F., Lee, S.Y., Medema,

M.H., and Weber, T. (2019). The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. Nucleic Acids Res. *47*, D625–D630.

30. Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. *37*, D233–8.

31. Rawlings, N.D., Barrett, A.J., Thomas, P.D., Huang, X., Bateman, A., and Finn, R.D. (2018). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res. *46*, D624–D632.

32. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res. *49*, D412–D419.

33. Scharf, D.H., Heinekamp, T., and Brakhage, A.A. (2014). Human and plant fungal pathogens: the role of secondary metabolites. PLoS Pathog. *10*, e1003859.

34. Lynch, M. (2006). The origins of eukaryotic gene structure. Mol. Biol. Evol. *23*, 450–468.

35. Miyauchi, S., Kiss, E., Kuo, A., Drula, E., Kohler, A., Sánchez-García, M., Morin, E., Andreopoulos, B., Barry, K.W., Bonito, G., et al. (2020). Large-scale genome sequencing of mycorrhizal fungi provides insights into the early evolution of symbiotic traits. Nat. Commun. *11*, 5125.

36. Franco, M.E.E., Wisecaver, J.H., Arnold, A.E., Ju, Y.-M., Slot, J.C., Ahrendt, S., Moore, L.P., Eastman, K.E., Scott, K., Konkel, Z., et al. (2022). Ecological generalism drives hyperdiversity of secondary metabolite gene clusters in xylarialean endophytes. New Phytol. *233*, 1317–1330.

37. Kuo, C.-H., Moran, N.A., and Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. Genome Res. *19*, 1450–1454.

38. Giovannoni, S.J., Cameron Thrash, J., and Temperton, B. (2014). Implications of streamlining theory for microbial ecology. ISME J. *8*, 1553–1565.

39. Moran, N.A., McCutcheon, J.P., and Nakabachi, A. (2008). Genomics and evolution of heritable bacterial symbionts. Annu. Rev. Genet. *42*, 165–190.

40. Bourguignon, T., Kinjo, Y., Villa-Martín, P., Coleman, N.V., Tang, Q., Arab, D.A., Wang, Z., Tokuda, G., Hongoh, Y., Ohkuma, M., et al. (2020). Increased Mutation Rate Is Linked to Genome Reduction in Prokaryotes. Curr. Biol. *30*, 3848–3855.e4.

41. Hinsch, J., Galuszka, P., and Tudzynski, P. (2016). Functional characterization of the first filamentous fungal tRNA-isopentenyltransferase and its role in the virulence of Claviceps purpurea. New Phytol. *211*, 980–992.

42. Morrison, E.N., Emery, R.J.N., and Saville, B.J. (2017). Fungal derived cytokinins are necessary for normalUstilago maydisinfection of maize. Plant Pathol. *66*, 726–742.

43. Chen, H., Raffaele, S., and Dong, S. (2021). Silent control: microbial plant pathogens

evade host immunity without coding sequence changes. FEMS Microbiol. Rev. *45*.

44. Lim, C.S., Weinstein, B.N., Roy, S.W., and Brown, C.M. (2021). Analysis of Fungal Genomes Reveals Commonalities of Intron Gain or Loss and Functions in Intron-Poor Species. Mol. Biol. Evol. *38*, 4166–4186.

45. Grützmann, K., Szafranski, K., Pohl, M., Voigt, K., Petzold, A., and Schuster, S. (2014). Fungal alternative splicing is associated with multicellular complexity and virulence: a genome-wide multi-species study. DNA Res. *21*, 27–39.

46. Muzafar, S., Sharma, R.D., Chauhan, N., and Prasad, R. (2021). Intron distribution and emerging role of alternative splicing in fungi. FEMS Microbiol. Lett. *368*.

47. FastQC: A Quality Control Tool for High Throughput Sequence Data http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

48. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

49. Bushnell, B., Rood, J., and Singer, E. (2017). BBMerge - Accurate paired shotgun read merging via overlap. PLoS One *12*, e0185056.

50. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.

51. Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. *45*, e18.

52. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

53. Chin, C.-S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods *13*, 1050–1054.

54. Darling, A.C.E., Mau, B., Blattner, F.R., and Perna, N.T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. *14*, 1394–1403.

55. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

56. PacBio Secondary Analysis Tools on Bioconda https://github.com/PacificBiosciences/pbbioconda.

57. Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9*, e112963.

58. Hessenauer, P., Fijarczyk, A., Martin, H., Prunier, J., Charron, G., Chapuis, J., Bernier, L., Tanguay, P., Hamelin, R.C., and Landry, C.R. (2020). Hybridization and introgression drive

genome evolution of Dutch elm disease pathogens. Nat Ecol Evol *4*, 626–638.

59. Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. *35*, 543–548.

60. Xiao, C.-L., Chen, Y., Xie, S.-Q., Chen, K.-N., Wang, Y., Han, Y., Luo, F., and Xie, Z. (2017). MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat. Methods *14*, 1072–1074.

61. Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., and Smit, A.F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. U. S. A. *117*, 9451–9457.

62. Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics *Chapter 4*, Unit 4.10.

63. Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. *33*, W465–7.

64. Hoff, K.J., Lomsadze, A., Borodovsky, M., and Stanke, M. (2019). Whole-Genome Annotation with BRAKER. In Gene Prediction: Methods and Protocols, M. Kollmar, ed. (Springer New York), pp. 65–95.

65. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

66. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. *33*, 6494–6506.

67. Ophiostoma ulmi resource browser http://www.moseslab.csb.utoronto.ca/o.ulmi/.

68. Wong, B., Leal, I., Feau, N., Dale, A., Uzunovic, A., and Hamelin, R.C. (2020). Molecular assays to detect the presence and viability of Phytophthora ramorum and Grosmannia clavigera. PLoS One *15*, e0221742.

69. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

70. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

71. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

72. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589.

73. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. *35*, 518–522.

74. Sanderson, M.J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics *19*, 301–302.

75. TIMETREE 5: The timescale of life http://www.timetree.org/.

76. Chernomor, O., von Haeseler, A., and Minh, B.Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Syst. Biol. *65*, 997–1008.

77. Junier, T., and Zdobnov, E.M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics *26*, 1669–1670.

78. Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics *19*, 153.

79. Chan, P.P., Lin, B.Y., Mak, A.J., and Lowe, T.M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. *49*, 9077–9096.

80. Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics *20*, 289–290.

81. Pagel, M., and Meade, A. (2006). Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am. Nat. *167*, 808–825.

82. Ho, L. si T., and Ané, C. (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Syst. Biol. *63*, 397–408.

83. Venditti, C., Meade, A., and Pagel, M. (2011). Multiple routes to mammalian diversity. Nature *479*, 393–396.

84. Organ, C.L., Shedlock, A.M., Meade, A., Pagel, M., and Edwards, S.V. (2007). Origin of avian genome size and structure in non-avian dinosaurs. Nature *446*, 180–184.

85. Revell, L.J. (2011). phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution *3*, 217–223.

86. Yu, G. (2020). Using ggtree to Visualize Data on Tree-Like Structures. Curr. Protoc. Bioinformatics *69*, e96.

87. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

88. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. *41*, e121.

89. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238.

90. Han, M.V., Thomas, G.W.C., Lugo-Martinez, J., and Hahn, M.W. (2013). Estimating gene

gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol. Biol. Evol. *30*, 1987–1997.

91. Badet, T., Peyraud, R., Mbengue, M., Navaud, O., Derbyshire, M., Oliver, R.P., Barbacci, A., and Raffaele, S. (2017). Codon optimization underpins generalist parasitism in fungi. Elife *6*.

92. Möller, M., and Stukenbrock, E.H. (2017). Evolution and genome architecture in fungal plant pathogens. Nat. Rev. Microbiol. *15*, 756–771.