Lifestyles shape genome size and gene content in fungal pathogens

Anna Fijarczyk[1,2,3,4,*], Pauline Hessenauer[2,5], Richard C. Hamelin[2,5,6], Christian R. Landry[2,3,4,7]

[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]Département des sciences du bois et de la forêt, Université Laval, Québec, G1V 0A6 Québec, Canada;
[6]Department of Forest and Conservation Sciences, The University of British Columbia, V6T 1Z4 Vancouver, Canada;
[7]Département de biochimie, microbiologie et bio-informatique, Université Laval, Québec, G1V 0A6 Québec, Canada;

*corresponding author: anna.fijarczyk@uqo.ca

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

Summary

Fungi display a wide range of lifestyles and hosts. We still know little about the impact of lifestyles, including pathogenicity, on their genome architecture. Here, we combined and annotated 552 fungal genomes from the class Sordariomycetes and examined the association between 13 genomic and two lifestyle traits: pathogenicity and insect association. We found that pathogens on average tend to have a larger number of protein-coding genes, including effectors, and tRNA genes. In addition, the non-repetitive size of their genomes is larger than that of non-pathogenic species. However, this pattern is not consistent across all groups. Insect endoparasites and symbionts have smaller genome sizes and genes with longer exons; moreover, insect-vectored pathogens possess fewer genes compared to those not transmitted by insects. Our study shows that genes are the main contributors to genome size variation in Sordariomycetes and that pathogens can exhibit distinct genome architectures, depending on their host and vector interactions.

Introduction

The growing emergence of severe infectious diseases in humans, crops, and wildlife caused by fungi, prompts strengthened efforts to uncover the molecular mechanisms of pathogenic traits. This requires understanding the forces shaping the evolution of fungal pathogen genomes (Fisher et al. 2020; Rokas 2022). Fungal genome sizes span from 3 Mb in microsporidians (Biderre et al. 1997) up to 892 Mb reported in a rust fungus (Mohanta and Bae 2015; Tavares et al. 2014), and fungal pathogens score both among the smallest and the largest of them. Effectors, including small secreted proteins and RNAs, are fundamental in host colonization and defense against host immune system, and are also known to play a key role as virulence factors, i.e. molecular mechanisms responsible for causing damage to the host (Stergiopoulos and de Wit 2009). Other genomic characteristics or processes correlated with or implicated in the evolution of pathogens include presence of specialized gene families (Raffa and Keller 2019), horizontal gene transfer (Sahu et al. 2023; McDonald et al. 2019), copy number variation of pathogenicity-relevant genes (Bergin et al. 2022), expansion of transposable elements (Oggenfuss and Croll 2023), and genome compartmentalization (Raffaele and Kamoun 2012; Wacker et al. 2023). Yet we still have limited knowledge of how important and frequent different genomic processes are in the evolution of pathogenicity across phylogenetically distinct groups of fungi and whether we can use genomic signatures left by some of these processes as predictors of pathogenic state.

Fungal pathogen genomes—particularly those of plant pathogens—are often characterized by large sizes, with expansions of TEs, and a unique presence of a compartmentalized genome with fast and slow evolving regions or chromosomes (Raffaele and Kamoun 2012; Möller and Stukenbrock 2017). Such accessory genomic compartments could facilitate the fast evolution of effectors (Dong, Raffaele, and Kamoun 2015). Similarly, TE expansions have been implicated in the evolution of new virulence genes in emerging pathogens (Bao et al. 2017; Wacker et al. 2023). Even though such architecture can facilitate pathogen evolution, it is currently recognized that its origin is more likely a side effect of a species evolutionary history rather than being caused by pathogenicity (Torres et al. 2020). Effectors are considered pivotal for pathogen evolution, along with the diversification of other gene families (Muszewska et al. 2011; Baroncelli et al. 2016; Sipos et al. 2017). As the number of genes is strongly correlated with fungal genome size (Stajich 2017), such expansions could be a major contributor to fungal genome size. Notably, not all pathogenic species experience genome or gene expansions, or show compartmentalized genome architecture. While gene family expansions are important for some pathogens, the contrary can be observed in others, such as Microsporidia. Due to transition to obligatory intracellular lifestyle these fungi show signatures of strong genome contractions and reduced gene

repertoire (Katinka et al. 2001) without compromising their ability to induce disease in the host. This raises questions about universal genomic mechanisms of transition to pathogenic state.

The size of genomes is dictated by the balance of mutation, selection and genetic drift. Neutral evolution hypotheses generate predictions regarding the size of the genomes, the size of their deleterious non-coding parts, the number of genes, and the mutation rate (Lynch and Conery 2003; Petrov 2002; Lynch 2007). In eukaryotes, the efficiency of purifying selection depends on the size of the species' effective population size (Ne) and will thus determine the rate of loss of slightly deleterious non-coding DNA, including mobile genetic elements or introns, leading to larger genomes with more repetitive content in species with smaller Ne. In bacteria, bias towards deletion rates will lead to shortening of the genome in small Ne (Mira, Ochman, and Moran 2001). These hypotheses also explain distinct genome evolution in species experiencing Ne bottlenecks due to reproductive and lifestyle strategies, such as endosymbionts, endoparasites or species with specialized vectors (Mira and Moran 2002).

In ascomycetes, the evolution of genome size on a broad phylogenetic scale supports the neutral hypothesis of genome evolution, but in different ways. In species with larger genomes, indicators of drift such as increased intron frequency and decreased gene density are associated with genome expansions, but in species with small and medium-sized genomes, those indicators are associated with genome size reductions (Kelkar and Ochman 2012). Other phylogenomic studies investigating a wide range of Ascomycete species, while not explicitly focusing on the neutral evolution hypothesis, have found strong phylogenetic signals in genome evolution, reflected in distinct genomic traits (e.g., genome size, gene number, intron number, repeat content) across lineages or families (Shen et al. 2020; Hensen et al. 2023). Variation in genome size has been shown to correlate with the activity of the repeat-induced point mutation (RIP) mechanism (Hensen et al. 2023; Badet and Croll 2025), by which repeated DNA is targeted and mutated. RIP can potentially lead to a slower rate of emergence of new genes via duplication (Galagan et al. 2003), and hinder TE proliferation limiting genome size expansion (Badet and Croll 2025). Variation in genome dynamics across lineages has also been suggested to result from environmental context and lifestyle strategies (Shen et al. 2020), with Saccharomycotina yeast fungi showing reductive genome evolution and Pezizomycotina filamentous fungi exhibiting frequent gene family expansions. Given the strong impact of phylogenetic membership, demographic history (Ne) and host-specific adaptations of pathogens on their genomes, we reasoned that further examination of genomic sequences in groups of species with various lifestyles can generate predictions regarding the architecture of pathogenic genomes.

In this study, we investigate how fungal genome size and complexity associates with pathogenic lifestyle in the diverse ascomycete class of Sordariomycetes fungi. Sordariomycetes have rich genomic resources and include species with a wide range of lifestyles. Apart from plant pathogens and saprotrophs, this class contains several groups of insect endoparasites and species vectored by insects. Because endoparasitism and presence of vectors can impact genome evolution, we considered two non-exclusive lifestyle traits: pathogenicity and insect-association. We studied 552 genomes, including 14 newly sequenced and assembled genomes and their associated 13 genomic traits (genome size, genome size excluding repeats, the number of genes, repeat content, GC content, number of introns, intron length, exon length, the fraction of genes with introns, intergenic length, the number of tRNA and pseudo tRNA genes) to answer the following questions: 1) which genomic traits predict genome size, 2) does larger genome size correlate with pathogenic lifestyle, and 3) are there genomic traits that are shared across all pathogens.

Results

Genome assemblies and phylogeny of Sordariomycetes

We analyzed 552 genomes from the class Sordariomycetes (Ascomycota) together comprising fungi mostly represented by plant pathogens, saprotrophs, and entomopathogens (Supplementary Table 1). Most genomes were downloaded from NCBI or other sources (Supplementary Table 1) and complemented with 14 genomes sequenced and assembled in this study (Supplementary Table 2). Genome size ranged from 20.7 Mbp in *Ceratocystiopsis brevicomis* (CBS 137839) to 110.9 Mbp in *Ophiocordyceps sinensis* (IOZ07) and the number of *ab initio* gene models ranged from 6,280 in *Ambrosiella xylebori* (CBS 110.61) to 17,878 in *Fusarium langsethiae* (Fe2391). An overview of genome assemblies and their annotations, in relation to different sequencing technologies and other genomic databases, is provided in the Supplementary Results. A maximum likelihood tree was generated from 1000 concatenated single-copy conserved proteins and had a 100% bootstrap support for all major nodes except for one (support of 82% for the split between two subclades of Ophiostomatales, Figure 1A).

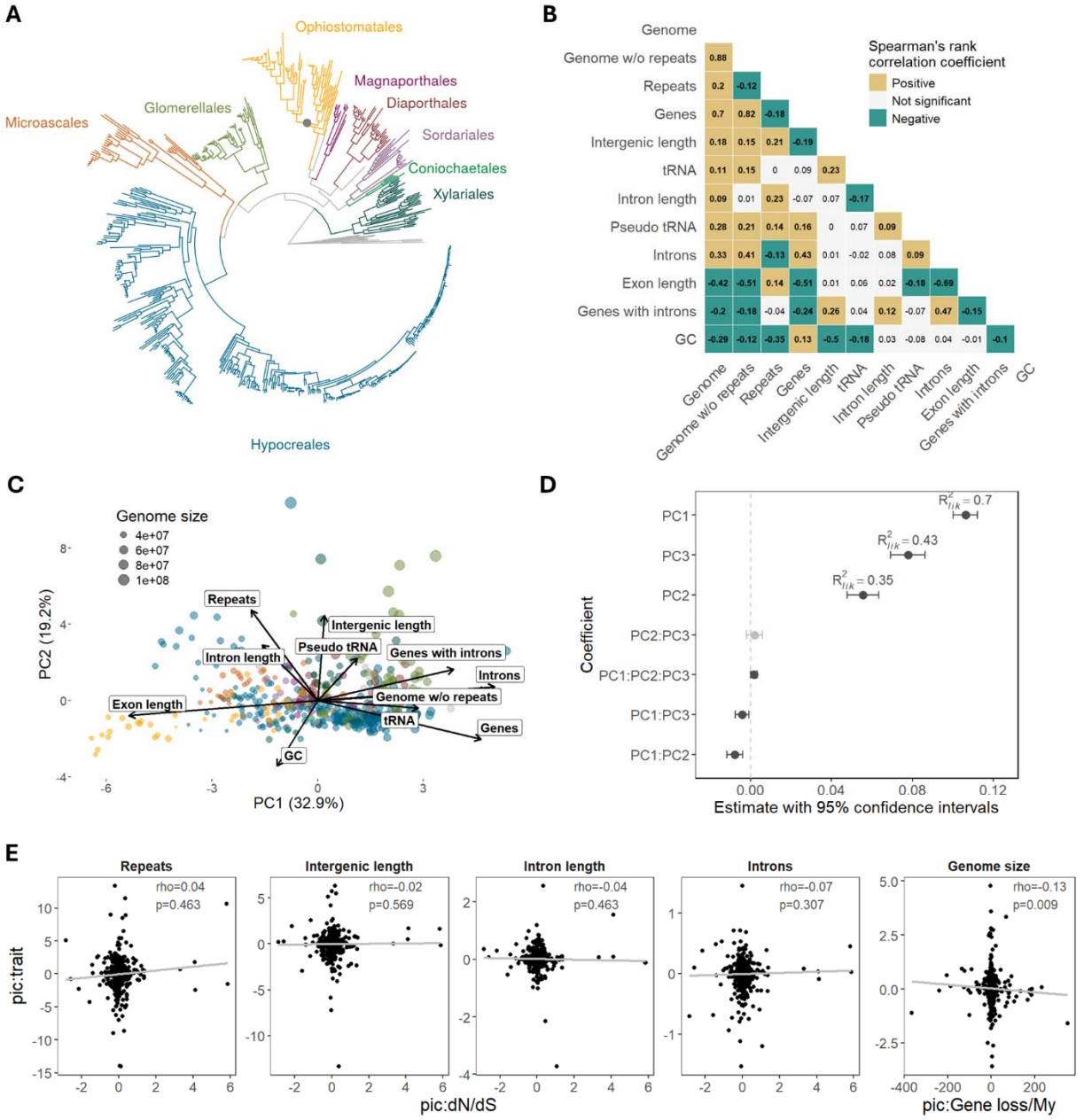
Genome size predictors

We examined the correlation between genome size (bp) and 11 genomic traits, including the number of genes, the repeats proportion in genome assembly, the 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.

Most genomic traits were correlated with genome size and with each other, with the strongest positive correlation observed between the genome size, the genome size excluding repeats and the number of genes (Figure 1B). Exon length, number of genes with introns and GC content were negatively correlated with genome size. We extracted three major principal components from all 11 traits, which together explained 66% of variance among species. The first PC was mostly driven by coding traits (exon length, number of introns, genome without repeats, genes), whereas the second PC was mostly driven by non-coding parts of the genome (repeats, intergenic length, GC and intron length, Supplementary Figure 1). The third PC was mainly driven by the proportion of genes with introns, tRNAs and number of introns. Modeling genome size using these 3 PCs showed that all axes were important predictors of genome size, with PC1 being the strongest (Figure 1C). PC1 also showed negative interaction with PC2 and PC3. This suggests that fungi with large genomes have in general either an excess of non-coding elements or an excess of coding elements but rarely both of them at the same time.

We calculated genome-wide ratio of non-synonymous to synonymous substitution rates (dN/dS) as a proxy for an effective population size (Ne) (Kelkar and Ochman 2012; Lefébure et al. 2017) to determine its association with the non-coding portion of the genome. We found no correlation between dN/dS and non-coding traits such as repeat content, intergenic length, intron length or number of introns (Figure 1D), therefore we found no confirmation of accumulation of slightly deleterious DNA in species that have high dN/dS indicative of small Ne. As the amount of non-coding elements is generally minimal in our dataset compared to other eukaryotes, such correlation may be difficult to detect. The rate of gene loss was negatively correlated with genome size (Figure 1D), but was not correlated with dN/dS (Spearman's Rho=-0.05, P-value=0.44). This suggests that variations in deletion rates may influence genome size dynamics in this group of fungi.


A
Ophiostomatales
Magnaporthales
Diaporthales
Sordariales
Coniochaetales
Xylariales
Glomerellales
Microascales
Hypocreales
B
Spearman's rank correlation coefficient
Positive
Not significant
Negative
Genome
Genome w/o repeats
Repeats
Genes
Intergenic length
tRNA
Intron length
Pseudo tRNA
Introns
Exon length
Genes with introns
GC
C
Genome size
4e+07
6e+07
8e+07
1e+08
Repeats
Intergenic length
Intron length
Pseudo tRNA
Genes with introns
Introns
Exon length
Genome w/o repeats
tRNA
Genes
GC
PC2 (19.2%)
PC1 (32.9%)
D
PC1
PC3
PC2
PC2:PC3
PC1:PC2:PC3
PC1:PC3
PC1:PC2
Coefficient
Estimate with 95% confidence intervals
E
Repeats
rho=0.04
p=0.463
Intergenic length
rho=-0.02
p=0.569
Intron length
rho=-0.04
p=0.463
Introns
rho=-0.07
p=0.307
Genome size
rho=-0.13
p=0.009
pic:trait
pic:dN/dS
pic:Gene loss/My

190

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=563 species). The largest orders are indicated with different colors. Bootstrap support for all major clades except one within Ophiostomatales (82%, black dot) reached 100%. **B**. Spearman's rank correlation for phylogenetic independent contrasts of all pairwise combinations of genomic traits. **C**. Principal component analysis of genomic traits. Colors correspond to orders depicted in A. **D**. Model of genome size. On the y axis are the model coefficients with 95% confidence intervals obtained with phylogenetic generalized least squares (pgls). Principal components correspond to the three main principal components based on 11 genomic traits. **E**. Testing hypotheses of genome size evolution. The first four plots show correlation of dN/dS as a proxy of Ne with the non-coding elements of the genome. The last plot is a correlation of gene loss rate with genome size. Correlations were tested using Spearman's rank correlation on phylogenetic independent contrasts. Loadings of each principal component from figure C are shown in Supplementary Figure 1. Raw data underlying figures are in Figure 1-Source Data 1-3.

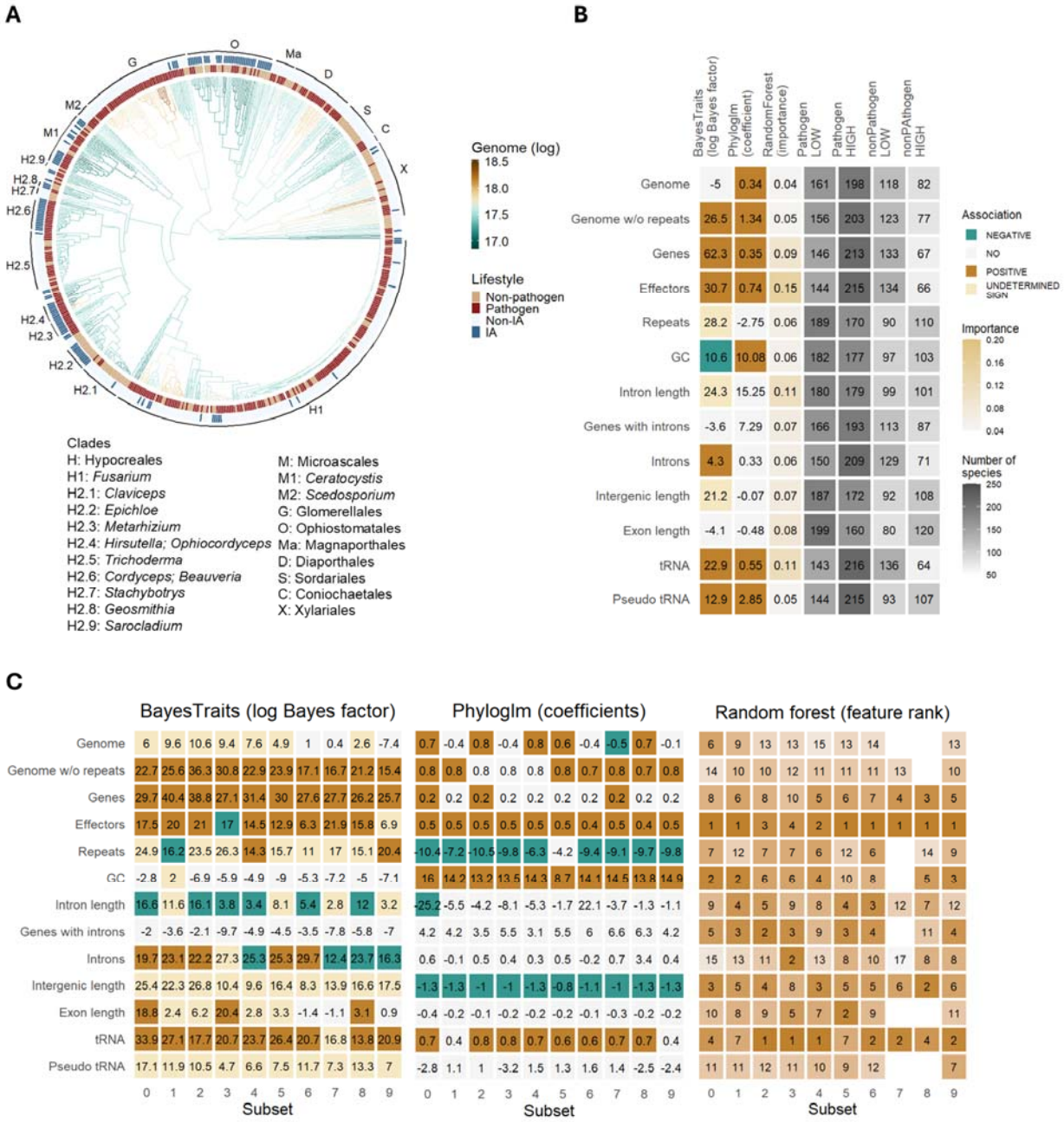
Genomic traits associated with pathogenicity

We found that pathogenic fungi did not have larger genomes, but instead had greater total number of genes, effectors, tRNAs, and fewer non-coding regions (Figure 2B). However, these patterns were not consistent across the phylogeny (Supplementary Table 2). To explore genomic traits associated with pathogenicity, we applied three methods that account for phylogenetic non-independence: an MCMC evolutionary model of binary traits (BayesTraits), phylogenetic logistic regression (Phyloglm), and random forest analysis. We assessed the same 11 genomic traits as before, adding genome size and the number of effectors. The distributions of genomic traits across pathogenic and non-pathogenic species are shown in Supplementary Figure 2. All three methods consistently pointed to the number of effectors and tRNA genes as predictors of pathogenicity (Figure 2B, Supplementary Figure 3). There was no support for a positive association between repeats and pathogenicity, and only one method indicated that pathogens tend to have larger genome sizes (Figure 2B). The total number of genes, genome size without repeats, and the number of pseudo tRNAs were also positively associated with pathogenicity in at least two methods.

Our dataset has more pathogenic species than non-pathogenic ones (1.8:1), and this is true in particular for species with medium and large genomes (>50 Mb, 4.5:1). This could represent a genuine pattern or a bias due to the focus on sequencing pathogenic species with larger genomes. Even though we didn't find an association between pathogenicity and genome size, we accounted for potential sampling bias to test the robustness of our findings. Species were grouped by genome size, and we subsampled

equal numbers of pathogenic and non-pathogenic species from each group, repeating the process 10 times. These 10 subsets were analyzed using the same methods in parallel, and nearly all confirmed that the number of effectors and tRNA genes were the strongest predictors of pathogenicity (Figure 2C, Supplementary Figure 3). Genome size analyses yielded mixed results. Out of the 10 random subsets, BayesTraits identified co-evolution of genome size with pathogenicity but with no clear directional trend. In contrast, Phyloglm coefficient estimates for genome size varied, showing either positive or negative associations depending on the subset. The analysis also revealed some evidence of a negative association between pathogenicity and non-coding elements such as repeats, intron length, number of introns, and intergenic regions. This could explain the absence of consistent positive correlation between pathogenicity and genome size, as the predicted increase in gene number in pathogens is likely offset by a reduction in non-coding elements.

We also examined whether associations with pathogenicity are consistent across the entire phylogenetic tree or vary between branches. The model with varying rates of evolution provided a significantly better fit for all genomic traits (Supplementary Table 2).


A
B
C
Genome (log)
18.5
18.0
17.5
17.0
Lifestyle
Non-pathogen
Pathogen
Non-IA
IA
Clades
H: Hypocreales
H1: Fusarium
H2.1: Claviceps
H2.2: Epichloe
H2.3: Metarhizium
H2.4: Hirsutella; Ophiocordyceps
H2.5: Trichoderma
H2.6: Cordyceps; Beauveria
H2.7: Stachybotrys
H2.8: Geosmithia
H2.9: Sarocladium
M: Microascales
M1: Ceratocystis
M2: Scedosporium
G: Glomerellales
O: Ophiostomatales
Ma: Magnaporthales
D: Diaporthales
S: Sordariales
C: Coniochaetales
X: Xylariales
BayesTraits (log Bayes factor)
PhyloGlm (coefficient)
RandomForest (importance)
Pathogen LOW
Pathogen HIGH
nonPathogen LOW
nonPAthogen HIGH
Genome -5 0.34 0.04 161 198 118 82
Genome w/o repeats 26.5 1.34 0.05 156 203 123 77
Genes 62.3 0.35 0.09 146 213 133 67
Effectors 30.7 0.74 0.15 144 215 134 66
Repeats 28.2 -2.75 0.06 189 170 90 110
GC 10.6 10.08 0.06 182 177 97 103
Intron length 24.3 15.25 0.11 180 179 99 101
Genes with introns -3.6 7.29 0.07 166 193 113 87
Introns 4.3 0.33 0.06 150 209 129 71
Intergenic length 21.2 -0.07 0.07 187 172 92 108
Exon length -4.1 -0.48 0.08 199 160 80 120
tRNA 22.9 0.55 0.11 143 216 136 64
Pseudo tRNA 12.9 2.85 0.05 144 215 93 107
Association
NEGATIVE
NO
POSITIVE
UNDETERMINED SIGN
Importance
0.20
0.16
0.12
0.08
0.04
Number of species
250
200
150
100
50
BayesTraits (log Bayes factor)
Genome 6 9.6 10.6 9.4 7.6 4.9 1 0.4 2.6 -7.4
Genome w/o repeats 22.7 25.6 36.3 30.8 22.9 23.9 17.1 16.7 21.2 15.4
Genes 29.7 40.4 38.8 27.1 31.4 30 27.6 27.7 26.2 25.7
Effectors 17.5 20 21 17 14.5 12.9 6.3 21.9 15.8 6.9
Repeats 24.9 16.2 23.5 26.3 14.3 15.7 11 17 15.1 20.4
GC -2.8 2 -6.9 -5.9 -4.9 -9 -5.3 -7.2 -5 -7.1
Intron length 16.6 11.6 16.1 3.8 3.4 8.1 5.4 2.8 12 3.2
Genes with introns -2 -3.6 -2.1 -9.7 -4.9 -4.5 -3.5 -7.8 -5.8 -7
Introns 19.7 23.1 22.2 27.3 25.3 25.3 29.7 12.4 23.7 16.3
Intergenic length 25.4 22.3 26.8 10.4 9.6 16.4 8.3 13.9 16.6 17.5
Exon length 18.8 2.4 6.2 20.4 2.8 3.3 -1.4 -1.1 3.1 0.9
tRNA 33.9 27.1 17.7 20.7 23.7 26.4 20.7 16.8 13.8 20.9
Pseudo tRNA 17.1 11.9 10.5 4.7 6.6 7.5 11.7 7.3 13.3 7
0 1 2 3 4 5 6 7 8 9
Subset
PhyloGlm (coefficients)
Genome 0.7 -0.4 0.8 -0.4 0.8 0.6 -0.4 -0.5 0.7 -0.1
Genome w/o repeats 0.8 0.8 0.8 0.8 0.8 0.8 0.7 0.8 0.7 0.8
Genes 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2
Effectors 0.5 0.5 0.5 0.5 0.5 0.5 0.4 0.5 0.4 0.5
Repeats -10.4 -7.2 -10.5 -9.8 -6.3 -4.2 -9.4 -9.1 -9.7 -9.8
GC 16 14.2 13.2 13.5 14.3 8.7 14.1 14.5 13.8 14.9
Intron length -25.2 -5.5 -4.2 -8.1 -5.3 -1.7 22.1 -3.7 -1.3 -1.1
Genes with introns 4.2 4.2 3.5 5.5 3.1 5.5 6 6.6 6.3 4.2
Introns 0.6 -0.1 0.5 0.4 0.3 0.5 -0.2 0.7 3.4 0.4
Intergenic length -1.3 -1.3 -1 -1 -1.3 -0.8 -1.1 -1 -1.3 -1.3
Exon length -0.4 -0.2 -0.1 -0.2 -0.2 -0.2 -0.1 -0.3 -0.2 -0.2
tRNA 0.7 0.4 0.8 0.8 0.7 0.6 0.6 0.7 0.7 0.4
Pseudo tRNA -2.8 1.1 1 -3.2 1.5 1.3 1.6 1.4 -2.5 -2.4
0 1 2 3 4 5 6 7 8 9
Subset
Random forest (feature rank)
Genome 6 9 13 13 15 13 14 13
Genome w/o repeats 14 10 10 12 11 11 11 13 10
Genes 8 6 8 10 5 6 7 4 3 5
Effectors 1 1 3 4 2 1 1 1 1 1
Repeats 7 12 7 7 6 12 6 14 9
GC 2 2 6 6 4 10 8 5 3
Intron length 9 4 5 9 8 4 3 12 7 12
Genes with introns 5 3 2 3 9 3 4 11 4
Introns 15 13 11 2 13 8 10 17 8 8
Intergenic length 3 5 4 8 3 5 5 6 2 6
Exon length 10 8 9 5 7 2 9 11
tRNA 4 7 1 1 1 7 2 2 4 2
Pseudo tRNA 11 11 12 11 10 9 12 7
0 1 2 3 4 5 6 7 8 9
Subset

250

Figure 2. Genomic traits associated with pathogenicity. A. Time-scaled phylogeny of Sordariomycetes colored by the reconstructed genome size displayed in log scale. The outside heatmap indicates species which are pathogenic and insect-associated (IA). Letters and numbers correspond to clades used in subsequent analyses, with representative genera from each clade listed in the legend. **B.** Association of 13 genomic traits with pathogenicity estimated with BayesTraits, phylogenetic logistic regression (Phyloglm), and random forest classifier. In BayesTraits and Phyloglm, colors correspond to statistically significant associations, either positive (brown) or negative (green). In Phyloglm these are based on coefficient sign, in BayesTraits they were inferred based on transition rates between binary traits estimated from the model. "YES" indicates that dependent co-evolution was detected with BayesTraits but a uniform direction of association could not be deduced from transition rates. Grey color scale depicts prevalence of species classified as pathogenic and non-pathogenic with traits size below median (LOW) or above median (HIGH). **C.** Same three methods repeated for 10 subsets of the data with an equal number of pathogens and non-pathogens from each genome size bin. In random forest analysis, rank of the trait, instead of importance is given. Distributions of genomic traits are shown in Supplementary Figure 2. Transition rates (modeled gains and losses of traits) are shown in Supplementary Figure 3. Raw data underlying figures are in Figure2-Source Data 1-2.

Genomic traits of insect-associated (IA) species

Our analyses revealed that insect-associated (IA) species tend to have smaller genomes and genes with longer exons on average (Figure 3A, Supplementary Figure 2). All three methods consistently showed a positive association between exon length and IA species (Figure 3A). Additionally, there was support for a negative association between IA species and genome size, genome size excluding repeats, and gene number. Similar to pathogenicity, the evolution of IA traits varied across the phylogenetic tree, as indicated by the covarion models (Supplementary Table 3).

To ensure the association with exon length was not driven by a specific group of genes or species, we compared exon length and number in one-to-one orthologs between IA and non-IA species from two phylogenetic groups. In both phylogenetic groups, IA species had longer and fewer exons compared to non-IA species, while intron length was not different (Figure 3B). We also tested whether genes with longer exons were more likely to be retained in IA species compared to non-IA species. While genes present in most species within a clade (i.e. genes less likely to be deleted) were enriched for longer exons, this pattern was observed in both IA and non-IA species (Figure 3C).

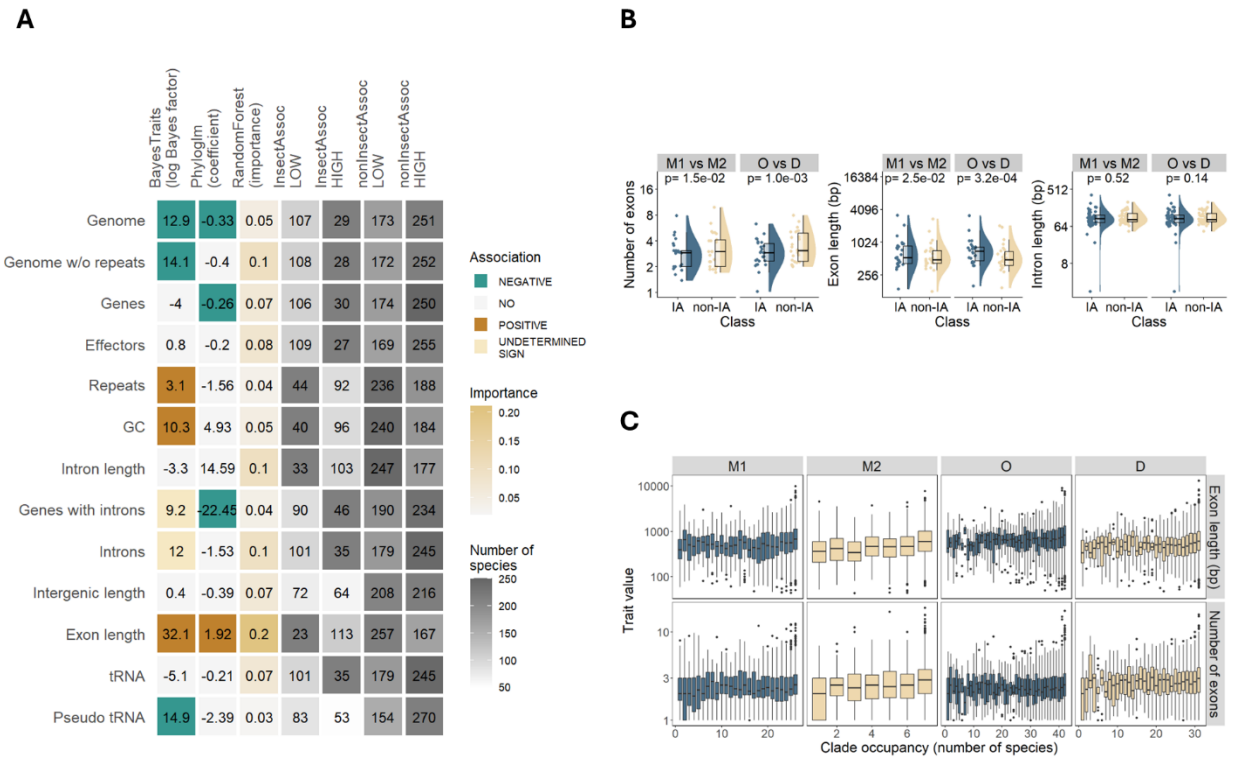


Figure 3. Genomic traits associated with insect association (IA). A. Association of 13 genomic traits with IA estimated with BayesTraits, phylogenetic logistic regression (Phyloglm), and random forest classifier. In BayesTraits and Phyloglm, colors correspond to statistically significant associations, either positive (brown) or negative (green). In Phyloglm these are based on coefficient sign, in BayesTraits they were inferred based on transition rates between binary traits estimated from the model. "YES" indicates that dependent co-evolution was detected with BayesTraits but a uniform direction of association could not be deduced from transition rates. Grey color scale depicts prevalence of species classified as IA and non-IA with traits size below median (LOW) or above median (HIGH). **B.** Comparison of exon and intron metrics in 38 one-to-one orthologs between two clades, Microascales (M), Ophiostomatales (O) and Diaporthales (D). In the comparisons, one taxon is IA (M1, O) and another one non-IA (M2, D). Exon/intron metrics were averaged within each clade and compared using paired Mann Whitney U test. **C.** Correlation between prevalence of gene families within clades and two exon metrics, in the same four clades. In all clades, more common gene families have longer and more exons (negative binomial generalized linear model, P-values < 0.05). Raw data underlying figures are in Figure3-Source Data 1-3.

Pathogens gene evolution depends on lifestyle

Given that our results indicated genome size and gene content vary depending on fungal lifestyle, we investigated how IA pathogens evolve compared to non-IA ones. Using phylogenetic logistic regression, we modeled the IA trait, incorporating key genomic traits one at a time as predictors and including pathogenicity as an interaction

term. This analysis was performed separately in three distinct fungal clades (H, M and a O/Ma/D/S comprising four subclades, Figure 2A). Models incorporating genome size and exon length could only be fitted for clade H, revealing a negative association between IA species and genome size, and a positive association with exon length in both pathogens and non-pathogens (Figure 4A). In clades O/Ma/D/S and M, genes showed a stronger negative association with IA in pathogens compared to non-pathogens. In clade H, tRNAs exhibited a stronger negative association with IA in pathogens than in non-pathogens. Additionally, effectors were negatively associated with IA in non-pathogens across all three clades, but not in pathogens.

Host-pathogen interactions often result in rapid gene turnover due to ongoing adaptation to host defense mechanisms, leading to gene family expansions and contractions (Gómez Luciano et al. 2019; Pendleton et al. 2014; Hartmann and Croll 2017; Stalder et al. 2023). To explore this, we analyzed a subset of ~500 gene families to test whether the rate of gene evolution (λ, the rate of gene loss and duplication (Mendes et al. 2020)) differs across species with different lifestyles. Species were categorized into four groups based on their pathogenicity and IA status, and we fitted models with one, two, or four rate parameters. The model with four distinct rates was significantly better than the simpler models (likelihood ratio tests (LRT) 4 vs. 1 rates: χ^2=7.8, P-value=0; LRT 4 vs. 2 rates: χ^2=6, P-value=0.002). Gene evolutionary rates were higher in non-IA species and lower in IA species, regardless of pathogenicity (Figure 4B). Non-IA pathogens showed the highest evolutionary rates, while IA pathogens exhibited the lowest.

These results indicate that IA species are associated with slower gene evolutionary rates overall, and some IA clades also with fewer genes, particularly in pathogenic species. However, some genes important for pathogens, such as effectors, are not depleted to the same extent as in non-pathogens.

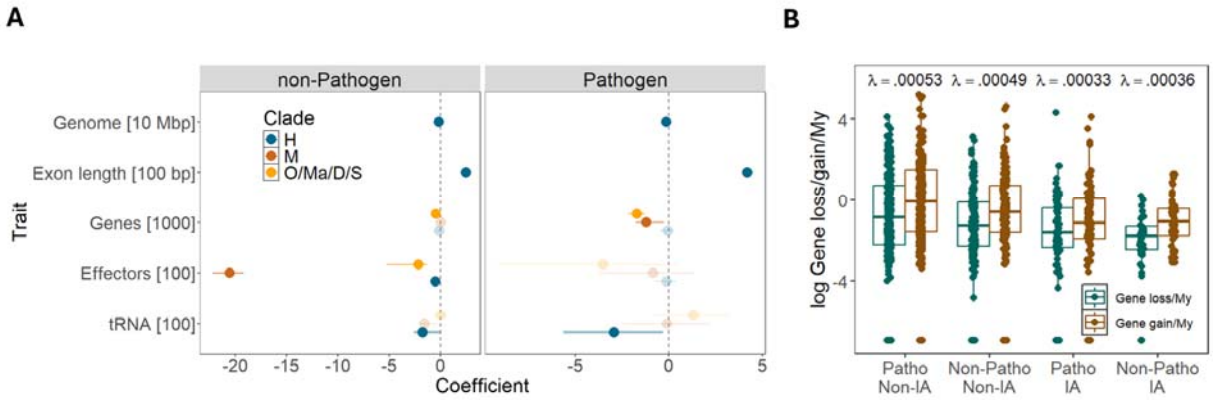


Figure 4. Evolution of pathogens with different lifestyles. **A.** Models of IA trait using phylogenetic logistic regression for each of the five genomic traits and pathogenicity as predictors. Coefficients of genomic traits for pathogenic and non-pathogenic species with 95% credible intervals are shown. **B.** Rates of gene loss and gain in four groups of species with different pathogenicity and IA status, estimated based on 527 small gene families using birth and death model in the program CAFE v5. Estimates of gene evolutionary rates (λ) are shown above the boxplots. Raw data underlying figures are in Figure4-Source Data 1-2.

IA plant pathogens preferentially lose genes important for host colonization

To investigate the functional gene classes gained or lost in fungi with different lifestyles, we calculated the fold change in the number of various gene classes across 19 clades (Figure 2A) since their last common ancestor with sister clades. The clades include plant pathogenic fungi (H1, Ma, G, D), entomopathogens (H2.3, H2.4, H2.6), , insect-vectored species (M1, O, H2.8, H2.2), saprotrophs (S), plant symbionts (H2.1, H2.2), and mixed lifestyle groups (Figure 5A), which display variation in the number of genes and genome size (Figure 5B). We utilized a range of databases to annotate genome assemblies or *ab initio* gene models (as detailed in the Methods).

Nearly all IA clades (M1, H2.8, O, H2.4, H2.2), along with a non-IA clade derived from an IA ancestor (H2.1, IA status of the MCRA of H2.1, H2.2, H2.3 and H2.4 estimated with BayesTraits equal to 0.97 [0.96-0.99]), a saprotrophic clade (S), and one mixed clade (H2.9), experienced the largest gene losses (Figure 5C). Entomopathogens exhibited a range of changes, with H2.4 showing mostly gene losses, H2.3 mostly gene gains, and H2.6 showing a mix of both. Conversely, the three clades that experienced the most gene gains were plant pathogenic clades (D, H1, G) and one IA clade (H2.3).

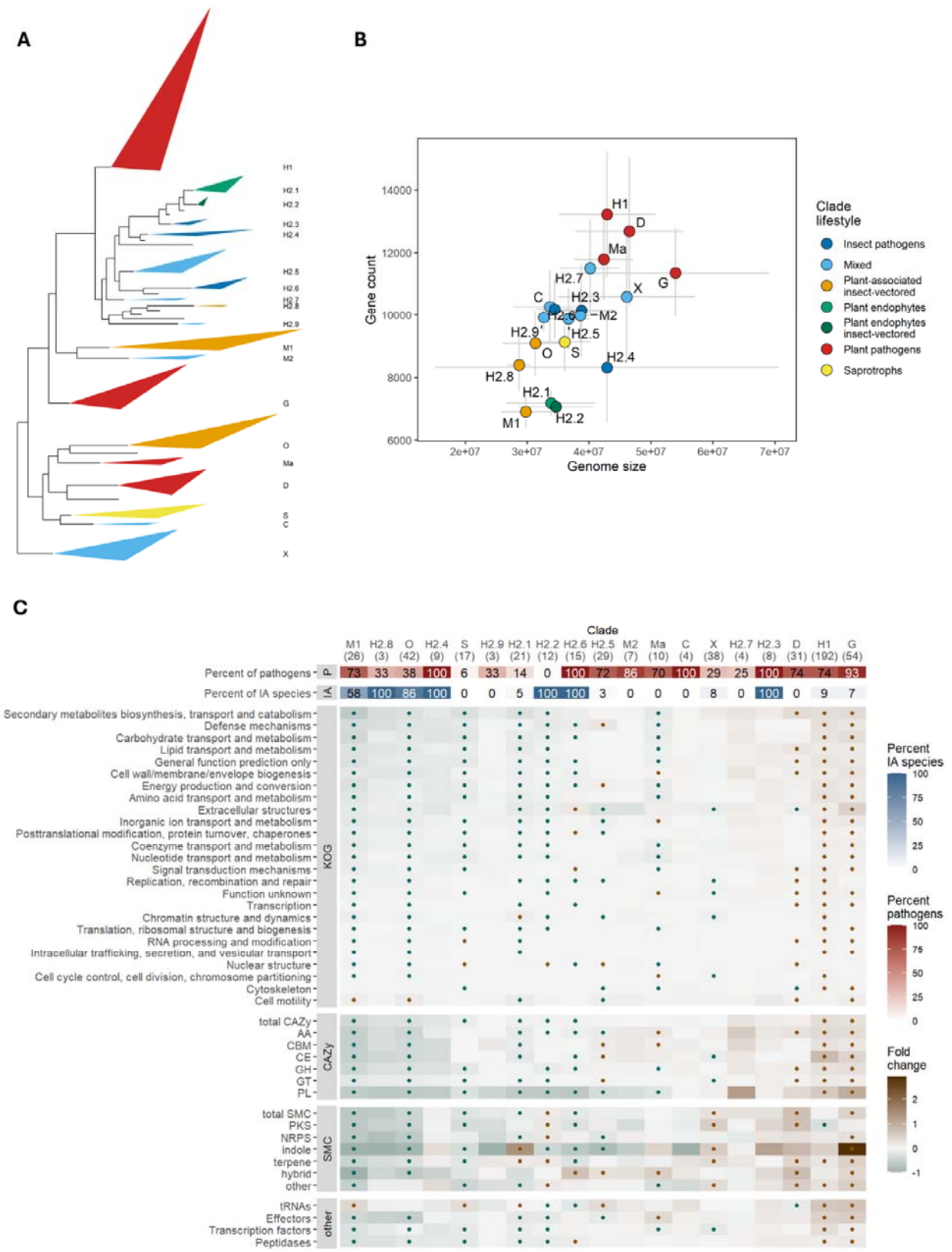

A
H1
H2.1
H2.2
H2.3
H2.4
H2.5
H2.6
H2.7
H2.8
H2.9
M1
M2
G
O
Ma
D
S
C
X
B
Gene count
Genome size
Clade lifestyle
Insect pathogens
Mixed
Plant-associated insect-vectored
Plant endophytes
Plant endophytes insect-vectored
Plant pathogens
Saprotrophs
C
Clade
M1 (26) H2.8 (3) O (42) H2.4 (9) S (17) H2.9 (3) H2.1 (21) H2.2 (12) H2.6 (15) H2.5 (29) M2 (7) Ma (10) C (4) X (38) H2.7 (4) H2.3 (8) D (31) H1 (192) G (54)
Percent of pathogens
Percent of IA species
KOG
CAZy
SMC
other
Percent IA species
Percent pathogens
Fold change

Figure 5. Insect-vectored clades lose genes involved in breaking plant host barriers. A. Phylogenetic position of the selected clades. Names correspond to the ones in Figure 2A and colors correspond to general lifestyles explained in the legend in plot B. These broad categories indicate clades dominated by plant pathogenic fungi (H1, G, D), entomopathogens (H2.4, H2.6, H2.3), insect-vectored species (M1, O, H2.8, H2.2), saprotrophs (S), plant symbionts (H2.1, H2.2) or mixed lifestyle groups. The basis of triangles indicates span between minimum and maximum branch length for given clades. The height of the triangles was scaled with a factor of 0.2. Empty branches correspond to smaller clades whose members are not shown. **B.** Means (dots) and standard deviations (error bars) of number of genes and genome size for selected clades. C. Heatmap shows the fold change of genes/clusters relative to the ancestral state ((observed - ancestral)/ ancestral state). Clades are shown in columns with the number of clade members in parentheses; functional classes are shown in rows. The dots indicate significant gain (brown) or loss (green) of genes/clusters across clade members estimated from 100 rounds of bootstrapping of 10 species in clades with >= 10 members. SMC: secondary metabolite clusters, CAZy: carbohydrate-active enzymes. Same heatmap but for pathogenic species only is shown in Supplementary Figure 4. Raw data underlying figures are in Figure5-Source Data 1-4.

Insect-vectored species (M1, O, H2.8) demonstrated the highest gene losses in genes associated with secondary metabolite pathways, defense mechanisms, and carbohydrate and lipid transport and metabolism (Figure 5C). This was also evident from the gene losses among different categories of CAZymes (carbohydrate-active enzymes), and SMCs (secondary metabolite clusters), as well as peptidases, transcription factors, and effectors. These patterns remained consistent when focusing solely on plant pathogens (Supplementary Figure 4). The only gene classes that did not exhibit a reduction or showed gains in these three clades were tRNAs, genes involved in cytoskeleton functions, and cell motility (Figure 5C). Other IA species displayed more variability in changes to their gene repertoires, typically showing some gains in SMC genes (H2.2, H2.6). Two clades enriched with non-IA plant pathogens (H1, G) demonstrated the highest gains in host-related genes. Clade H1 experienced the most gene gains in genes related to secondary metabolite pathways, defense mechanisms, carbohydrate and amino acid transport and metabolism, wheres clade G experienced most gene gains in genes related to extracellular structures, secondary metabolite pathways, carbohydrate transport and metabolism, and the cell wall/membrane/envelope biogenesis.

Overall, we observed common gene gains, particularly in CAZy and SMC classes, among many crop (clades H1, G) and some tree pathogens (clade D, Figure 5C, Supplementary Figure 4). On the other hand, insect-vectored species (clade H2.2), including plant pathogens (clades M1, H2.8, O), and specialized entomopathogens (H2.4, H2.6) exhibited variable levels of gene losses, sometimes in genes critical for host interactions (Figure 5C, Supplementary Figure 4). An exception was the IA clade

H2.3, which includes the genera *Metarhizium* and *Pochonia*, both known for their ability to adapt to various lifestyles or hosts.

Discussion

Fungal pathogens are commonly linked to notably large genomes (Raffaele and Kamoun 2012; Wacker et al. 2023) yet they exhibit considerable diversity in both genome size and structure (Möller and Stukenbrock 2017). We focused on the class of Sordariomycetes, which includes fungi with diverse lifestyles, to investigate three questions: first, which genomic traits best explain genome size; second, whether genome size can be attributed to pathogenic lifestyle; and third, whether pathogens share common genomic traits. Our analysis revealed that gene number is the strongest predictor of genome size, followed by non-coding regions. We found no correlation between genome size and pathogenicity. Instead, pathogenicity was on average best explained by the number of effectors and tRNAs, but the overall genome architecture of pathogens was strongly influenced by the type of interacting host and the presence of potential vector species.

Genome size evolution in fungi

We found that the number of genes and their length—specifically, exon length and number—are the primary drivers of genome size in Sordariomycetes. While non-coding regions of the genome also contribute to this, they are less significant and are more frequently found in gene-poor species. These findings suggest that genes and non-coding regions do not evolve in concert or at the same rate.

To determine whether neutral processes can explain genome size in Sordariomycetes, we tested two hypotheses. The first posits that a decrease in effective population size (Ne) reduces the effectiveness of selection, including purifying selection, which in eukaryotes typically leads to a greater accumulation of deleterious non-coding elements and larger genome sizes (Lynch and Conery 2003, Lynch 2007). The second hypothesis suggests that, due to a higher rate of deletions compared to insertions, smaller Ne results in more streamlined genomes, similar to observations in bacteria (Kuo, Moran, and Ochman 2009; Mira, Ochman, and Moran 2001). Using the genome-wide ratio of non-synonymous to synonymous substitution rates (dN/dS) as a proxy for Ne (Lefébure et al. 2017; Kelkar and Ochman 2012; Marino et al. 2024), we did not find the link between Ne and the non-coding portion of the genome. Although we observed a weak negative correlation between the gene deletion rate and genome size, there was no correlation between the gene deletion rate and dN/dS.

The lack of clear association of dN/dS with the non-coding content could arise from several factors. First, dN/dS may not be an appropriate indicator of Ne, or species analyzed may not be in mutation-selection balance. Second, fungi typically have much more gene-rich genomes than most of other eukaryotes (Wells and Feschotte 2020; Merényi et al. 2023), and the species in this study generally have modest genome sizes, which may weaken the relationship between Ne and the non-coding regions. Lastly, the large non-coding content is often driven by the proliferation of transposable elements (TEs), which are known to be very dynamic. The absence of significant association between dN/dS and TE content on the long evolutionary timescales has been noted (Marino et al. 2024), suggesting possible variations in selective effects of TEs across different taxonomic groups.

Support for the neutral hypothesis of genome size evolution in our study is indirect. We found that species associated with insects—such as endoparasites and symbionts, therefore species that likely undergo transmission bottlenecks and in effect have reduced Ne (Mira and Moran 2002)—have smaller genomes and in some cases fewer genes. Additionally, insect-associated species have genes with fewer but longer exons, suggesting the evolution of a more streamlined gene architecture. While these findings align with neutral theories of genome evolution, further research is necessary to identify the primary processes, including any potential selective factors, that drive the dynamics of gene content in fungi.

Genomic predictors of fungal pathogens

Large genome size driven by repeats has often been implicated as one of the hallmarks of fungal pathogens, especially fungal plant pathogens (Raffaele and Kamoun 2012). However, in our analysis of over 500 genomes of Sordariomycetes, we found no evidence supporting a positive association between genome size or repeat content and pathogenic lifestyle. While a few species with the largest genomes are indeed plant pathogens and possess a high proportion of repeats, such instances appear to be exceptions rather than the rule on a broader scale. Although the high activity of transposable elements can contribute to genome expansion and facilitate the evolution of pathogenicity-related traits (Hartmann et al. 2017; Wacker et al. 2023), our findings suggest that this is not the predominant pathway for fungal pathogen evolution.

We found that the overall number of genes, and genome size without repeats showed positive associations with pathogenicity on average. However, we can only speculate whether this pattern arises from drift or selection. Fungal pathogens comprise a diverse group of species, with specialists or generalists' adaptations and varying demographic

histories. Pathogens vary from obligatory endoparasites, free-living saprotrophs who become opportunistic pathogens, to pathogens of common crops or plants with fluctuating or large Ne. Unfortunately, the specifics of the population dynamics and life histories of many pathogens remain largely unknown. Following the idea that smaller effective population sizes lead to smaller, more streamlined genomes in Sordariomycetes, the higher gene counts in pathogens may imply that most pathogens possess larger Ne compared to non-pathogens. Alternatively, the greater number of genes in pathogens might reflect selective pressures favoring pathogen-specific gene expansions. An arms race between hosts and pathogens drives the continual evolution of new effector gene variants in pathogens, enabling them to evade the host immune system (Boller and He 2009; Stalder et al. 2023). Our findings indicate the association of genes important for pathogenicity, such as effectors, with pathogens which can contribute to this overall pattern. In addition, we observe that clades with different lifestyles undergo distinct gene loss and gain dynamics, further suggesting a selective advantage for certain gene duplications.

Effectors and tRNAs were the only specific gene classes we examined in our analysis, and they proved to be more closely associated with pathogenicity than the overall gene count. The importance of effectors in pathogens is well established (Stergiopoulos and de Wit 2009; Wilson and McDowell 2022). For instance, mutations in effector genes can provide pathogens with an advantage by enabling them to evade host immune defenses (Sánchez-Vallet et al. 2018). The role of tRNA modifications in pathogenicity has also been recognized (Hinsch, Galuszka, and Tudzynski 2016; G. Li et al. 2023), although the overall count of tRNAs remains less clear. In large eukaryotic genomes, short interspersed nuclear elements can lead to the misannotation of tRNAs (Gibbs et al. 2004). Therefore, the role of tRNAs in fungal pathogens warrants further investigation. Other gene families could be important predictors of pathogenicity. Carbohydrate-active enzymes profiles are reported to be different in pathogenic and non-pathogenic fungi (Dort et al. 2023; Haridas et al. 2020; Hane et al. 2019). Given that overall gene repertoires between pathogens can vary drastically, small-scale evolution of specialized genes or gene families can better reflect fungal transmissions to pathogenic lifestyles. Because of the uniform yet simplified gene annotation approach, the total number of genes may be underestimated in some assemblies in our dataset, as observed when comparing the same species in JGI Mycocosm. Although this pattern is not biased toward any particular group of species, access to high-quality, well-annotated genomes could provide a clearer picture of the relative contributions of specific gene families.

Different genomic evolutionary pathways for fungal pathogens

Our study showed that the insect-association (IA) trait was negatively correlated with genome size and positively with exon length. Additionally, the total number of genes was also significantly reduced, particularly in insect-transmitted clades. Previous research showed that specialized fungi tend to have smaller genomes than generalist fungi (Loos et al. 2024), and the associations we observed between genomic traits and IA species may partially reflect the degree of specialization towards their hosts. In our dataset, IA species include several species of entomopathogens, many of which are host-specific. Plant pathogens vectored by insects are also largely host specific. Unfortunately, it is challenging to obtain comprehensive data on the full range of host species represented in our dataset, which limits our ability to compare genome sizes between specialist and generalist species.

We found significant differences in gene evolution between IA and non-IA pathogens. Non-IA pathogens exhibited the highest gene evolution rates, whereas IA pathogens on average had the lowest. However, among the three clades of IA species included in our study, non-pathogenic species displayed a more pronounced reduction in effector repertoires compared to pathogenic species, highlighting the critical role of these genes across a broad range of fungal pathogens. Two IA clades, Ophiostomatales and Microascales, exhibited losses in gene classes important for host colonization and wood decomposition, such as CAZymes, or secondary metabolite clusters (Scharf, Heinekamp, and Brakhage 2014; Cantarel et al. 2009). This is likely because the insect vector often facilitates host entry in these species, rendering many genes redundant and prone to pseudogenization and loss. In the third clade, Hypocreales, patterns of gene losses and gains varied across species. Hypocreales is our dataset's largest order and includes both specialized and generalist species with diverse host and insect associations. Furthermore, many species within this order have lifestyles that remain poorly understood. Distinguishing between pathogenic and non-pathogenic lifestyles can be challenging, particularly for species found on dead host tissues, which may function as opportunistic pathogens or secondary saprotrophs that benefit from decaying matter. These factors may partially contribute to the variation that we see across species.

Conclusions

Our study shows that genome size of Sordariomycetes is primarily driven by gene number, with non-coding regions contributing less. We found no association between pathogenicity and larger genome size or repeat content. Although many pathogenic species have a high number of genes, especially effectors and tRNAs, this pattern does not hold true for all species. While gene gains are common to some plant pathogens, others, such as insect-vectored pathogens exhibit gene losses, including in genes

critical for host interactions. Insect-associated species also tend to have smaller genomes and longer exons, likely reflecting their level of specialization toward hosts and vectors—factors that may influence their effective population size. Overall, our findings suggest that the genome architecture of fungal pathogens is mainly shaped by the dynamics of pathogenicity-related genes and the nature of interactions with hosts and other species.

Materials and Methods

Genome assemblies

We selected 14 strains (11 x *Ophiostoma* and 3 x *Leptographium*, Supplementary Table 2) 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. The fungal isolates were handled in a facility that has received a Plant Pest Containment Level 1 certifcation by the Canadian Food Inspection Agency. 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 (Andrews 2010). Ten strains were used for *de novo* genome assembly (Supplementary Table 2). The depth of coverage of the generated data was between 32x and 555x. Reads were trimmed for adapters with Trimmomatic v0.33/0.36 (Bolger, Lohse, and Usadel 2014) using options 'ILLUMINACLIP:adapters.fa:6:20:10 MINLEN:21' and overlapping reads were merged with bbmerge from BBTools v36/v37 (Bushnell, Rood, and Singer 2017). *De novo* genome assemblies were generated with SPAdes v3.9.1 (Bankevich et al. 2012) with options '-k 21,33,55,77,99 --careful'. Mitochondrial DNA was searched in contigs with NOVOPlasty v3.8.3 (Dierckxsens, Mardulyn, and Smits 2017) 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 (H. Li and Durbin 2009), 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 (Supplementary Table 2). DNA extraction was done in the same way as for Illumina libraries, except that no vortexing and shaking were 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). The 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 (Chin et al. 2016). The range of parameters was tested and the final configuration files with the best-performing parameters for each species can be found on https://github.com/aniafijarczyk/Fijarczyk_et_al_2023. *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 (Darling et al. 2004). Assemblies were polished between two to four times by remapping long reads with minimap2 (pbmm v1 (H. Li 2018)) and correcting assemblies using arrow (pbgcpp v1). We also performed one round of polishing with pilon v1.23 (Walker et al. 2014) after mapping short-read Illumina reads (bwa v0.7.17 (H. Li and Durbin 2009)) to assemblies from this study (n=3) or from the previous study (Hessenauer et al. 2020) (n=7, Supplementary Table 2). The only exception was *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 (Waterhouse et al. 2018). Mitochondrial genomes were assembled using Illumina assemblies as baits (or those of related species). Long reads were mapped to bait assembly with pbmm2 (H. Li 2018), and a subsample of mapped reads was used for mtDNA assembly with mecat v2 (Xiao et al. 2017). 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 of a PacBio assembly, 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 (Flynn et al. 2020) using option -LTRStruct, and recovered repeat families together with fungal repeats from RepBase were used for masking assemblies with RepeatMasker v4.1.0 (Tarailo-Graovac and Chen 2009).

Genes were annotated with Braker v2.1.2 (Hoff et al. 2019) using Augustus v.3.3.2 (Stanke and Morgenstern 2005). 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 (Dobin et al. 2013). A

training set of genes was generated with GeneMark-ES-ET v4.33 (Lomsadze et al. 2005). 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

A total of 580 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 (Christendat 2013), *L. longiclavatum (Wong et al. 2020)*). The assemblies were obtained from NCBI rather than from other dedicated fungal genome resources such as JGI Mycocosm, primarily because this dataset provided better representation of various taxonomic clades of interest, such as insect-associated species from the order Microascales. The dataset comprised assemblies produced with different sequencing technologies, which may affect their overall quality, however, both short-read and long-read-based assemblies were distributed in similar proportions across different lifestyles that we investigated (pathogenic vs. non-pathogenic, and insect-associated vs. non-insect-associated) and were present across phylogeny (Supplementary Results). Indeed, long-read assemblies were on average longer and had a higher fraction of repeat content than short-read assemblies, but these differences did not affect average differences between lifestyles that we investigated (Supplementary Results).

Genes of all assemblies were subsequently annotated in a uniform manner. Assemblies were filtered to remove short contigs (< 1000 bp) and contaminants, as described above. The 1000 bp threshold was chosen because it is a common cutoff used by many assemblers when producing contigs. Gene completion was assessed with Busco v3 (Waterhouse et al. 2018) using an orthologous gene set from Sordariomycetes. *Ab initio* gene models were obtained with Augustus v.3.3.2 (Stanke and Morgenstern 2005) with different species-specific training files: Magnaporthe grisea, Fusarium, Neurospora, Verticillium longisporum1 or Botrytis cinerea (Supplementary Table 1). Different species training sets were selected to reduce the phylogenetic distance between the annotated species and the training species. Examination of the impact of choice of the training species on estimated traits did not reveal potential influence (Supplementary Results).Next, 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*). 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, $w = 2994$, $p = 0.57$). We also observed that the number of genes was lower than in JGI Mycocosm assemblies of

corresponding species. Again, this difference was consistent across studied lifestyles (Supplementary Results).

Phylogeny

The maximum likelihood tree was built from 1000 concatenated genes (retrieved with Busco v3 (Waterhouse et al. 2018)) with the highest species representation. Protein alignments were generated with mafft v7.453 (Katoh and Standley 2013) with E-INS-i method (option '--genafpair --ep 0 --maxiterate 1000'), trimmed with trimal v1.4.rev22 (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009) 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 (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017). The maximum likelihood tree was inferred with ultrafast bootstrap (Hoang et al. 2018), seed number 17629, and 1000 replicates implemented in IQ-TREE. Time-scaled phylogeny was inferred with program r8s v1.81 (Sanderson 2003), setting the calibration point of 201 My at the split of *Neurospora crassa* and *Diaporthe ampelina*, retrieved from TimeTree (Hedges and Kumar 2005).

Species and trait selection

New and downloaded genome assemblies after filtering together amounted to 573. Few species were represented by more than one strain, in particular, multiple *Ophiostoma* strains sequenced in this study. To avoid potential biases related to the overrepresentation of some species, only one strain per species was retained, leaving 563 species in total, including 11 outgroup species.

Pathogenicity and insect association traits were assigned based on a literature search. Pathogenicity was assigned if the species caused a well-recognized disease, or pathogenicity was experimentally documented in at least 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, number of effectors, 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). Repeat content was calculated as the proportion of masked bases in the raw assembly (after

removing contaminated contigs, but before filtering for contig length) compared to the raw 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 (Chan et al. 2021). Effectors were identified from predicted proteins, first by running SignalP v6 (Teufel et al. 2022) to identify proteins containing signal peptides, and then by running EffectorP v3 (Sperschneider and Dodds 2022) on them to identify effectors.

Correlation of genomic traits with genome size

The phylogenetic generalized least squares (pgls) model of genome size was built with gls function from the R package nlme. Maximum likelihood (ML) method was used, and the phylogenetic relationships were modeled with Pagel's lambda correlation structure. We considered 11 genomic traits (except for genome size, and effectors) as explanatory variables. First, the correlations among pairs of genomic traits were calculated after transforming variables to independent contrasts using pic function from the R package ape (Paradis, Claude, and Strimmer 2004). Pairwise correlations were tested with non-parametric Spearman’s rank correlation test. As it turned out, many variables were strongly correlated with each other. Therefore, instead of including all variables in the model, PCA was run on 11 genomic variables to retain three main components, which explained 33%, 19%, and 13% of the trait variance respectively. The loadings revealed that PC1 was mainly influenced by genic traits (exon length, introns, genome w/o repeats, genes), PC2 by non-coding traits (repeats, intergenic length), whereas PC3 by other traits (tRNA, genes with introns). We fitted the PGLS model with the three main PCs and interactions among them. The model was checked for homogeneity of variance, and normal distribution of residuals, and was refitted after removing 8 outliers. Model fit was assessed using anova() function. To compute the importance of each component in the model, R^2_{lik} was calculated with the R2 function from the rr2 R package after removing each component at a time and comparing it with the full model.

Correlation between dN/dS and non-coding regions

Nonsynnymous to synonymous substitution rate (dN/dS) was calculated as a proxy for effective population size (Ne). Species with a larger Ne are expected to exhibit a lower rate of nonsynonymous to synonymous substitution rate than species with smaller Ne due to stronger constraints on nonsynonymous deleterious mutations. dN/dS was calculated with codeml program in PAML v.4.10.6 (Yang 2007), on a codon-wise alignment of 300 random concatenated orthologs identified with Busco. Specifically, for each species, calculations were performed on a subset of ten most closely related species, with one species being the foreground branch and the rest serving as the background. Species, whose dN/dS estimate was based on a comparison with a

recently diverged species showed elevated dN/dS values, likely because of the stronger contribution of segregating polymorphism compared to fixed differences to divergence estimates. To account for this time-dependence, logarithm of species divergence times was fitted to a logarithm of dN/dS using generalized least square model, and model residues were taken to serve as dN/dS estimates in further correlations.

To test if species with smaller Ne tend to accumulate more non-coding deleterious material, correlation between phylogenetically independent contrasts of time-corrected dN/dS estimates and log transformed traits such as repeat content, intergenic length, intron length and number were calculated using Spearman's rank correlation.

Correlation between the gene loss rate and genome size

To test if differences in deletion rate can account for genome size changes, a correlation between the rate of gene loss and genome size was tested. Orthologous gene families for 563 species were determined with OrthoFinder v2.5.2 (Emms and Kelly 2019). Gene contractions were estimated by applying a birth-death model of gene evolution with CAFE v5 (Mendes et al. 2020). A single lambda for all branches was estimated, with a poisson distribution for gene family counts at the root. As the likelihood scores could not be computed for all gene families, the dataset was limited to 527 gene families present at the root and with a maximum change of 4 genes across species. This procedure has likely led to the removal of many fast evolving and big gene families and provides estimates on the rather conserved set of gene families. Lambda search was run three separate times to assure convergence. The rate of gene loss was calculated by summing gene family contractions and dividing by the calibrated branch length (in My). Correlation on phylogenetically independent contrasts of log transformed genome size and log transformed gene loss rate was calculated using Spearman's rank correlation.

Association of genomic traits with pathogenicity and IA

Association of pathogenicity and insect association (IA) with genomic traits was estimated using three approaches, i) a reversible jump MCMC discrete model of evolution implemented in BayesTraits v3 (Pagel and Meade 2006), ii) a phylogenetic logistic regression with phyloglm function in R package phylolm (Ho and Ané 2014), iii) random forest classification using scikit-learn python (v3) package. In tests of pathogenicity associations, analyses were also run separately within 10 subsets, where each subset comprised an equal number of pathogens and non-pathogens (n = 190,190) sampled from five genome size bins (between 20 and 70 Mbp, with a 10 Mbp step).

BayesTraits tests the coevolution of a pair of binary traits, by modeling eight possible transition rates between the two traits (transition of only one trait at a time). We modeled coevolution of each genomic trait with pathogenicity or IA. Genomic traits were converted into binary traits based on their median (0 if below median and 1 if above median), such that trait gain indicates increase of the size of that trait, and trait loss indicates decrease in the size of the trait. For each genomic trait, two models were compared, one with independent and one with dependent evolution of the two traits. In addition, to test if some transition rates are more frequent in one direction (gains vs losses of a trait), four models were run in which we set equal rates of change for the opposite transitions (for example the rate of gain of genome size in pathogens was set to equal the rate of loss of genome size in pathogens), and compared them to dependent models using log Bayes factor. Based on all significant transition changes (gains or losses of one of the traits in the presence of the other) we determined if the genomic trait is positively associated with the pathogenicity/IA (gain of the trait for pathogens/IA, or loss of the trait in non-pathogens/non-IA). Finally, we also compared dependent model with the covarion model in which coevolutionary transitions were allowed to vary across the phylogeny (Venditti, Meade, and Pagel 2011). Models were selected if the log Bayes factor surpassed 4. Each model was run three times to check for consistency, for 21 mln iterations, 1 mln burn-in, and thinning of 1000.

Phylogenetic logistic regression was used to fit each genomic trait to pathogenicity/IA 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. The scale of some genomic traits was adjusted. Assembly size with and without repeats was used in the unit of 10 Mbp, number of genes in unit of 1000 genes, length of introns and intergenic length in the unit of 1 kb, length of exons in the unit of 100 bp, number of tRNAs, pseudo tRNAs, and effectors in the unit of 100 genes.

To train a machine learning classifier for predicting pathogens and IA species, we tested the accuracy of five classifiers (KNeighbors, SVC, DecisionTree, RandomForest, and GradientBoosting). In each case, randomized grid search and five-fold cross validation was performed on the train data (80%), and balanced accuracy scores were obtained. Scores were averaged across 10 random splits into test and train datasets. Species with missing data were filtered out, and data were rescaled (between 0 and 1) for classification with SVC. As the random forest model showed one of the best scores, we used this model to train on the whole dataset and determine the most important features. Important features were determined using a mean decrease impurity method. In the final model, we used hyperparameters from the data split which gave the best balanced accuracy score. As the samples in the dataset are dependent through

phylogenetic relationships, we also determined the impact of phylogeny on the model, by including for each species distances to each node as additional features. We found that node distances never appeared among the top 5 most important features.

Ancestral reconstruction of genome size was performed using fastAnc function in R package phytools (Revell 2011) and visualized on the tree with ggtree R package (Yu 2020) using the continuous option. To estimate lifestyle traits in selected ancestral nodes, a discrete model of evolution implemented in BayesTraits v3 was used with the same parameters as above.

To test association of exon length and exon number between IA and non-IA clades, we obtained these features for 38 one-to-one orthologs found with OrthoFinder (described above). Mean values across IA species were compared to mean values in non-IA species for two clade pairs using paired Mann Whitney U test.

To determine whether long exon genes are more or less likely to be deleted compared to short exon genes, we examined the presence of gene families across different clade members. Gene families found in all clade members represent genes that are less likely to be deleted, whereas gene families present in only a few members of the clade represent genes that are more often deleted. We then compared features of the common and rare gene families in terms of exon length and number, by fitting a negative binomial generalized linear model using glm.nb() function from the MASS package in R.

Evolution of pathogens with different lifestyles

The insect association (IA) trait was modeled using phylogenetic logistic regression in the phyloglm model, as described earlier, with each of the five genomic traits (genome size, exon length, number of genes, effectors, and tRNAs) tested individually, including pathogenicity as an interaction term. Models were fitted separately for 3 clades comprising species with a combination of pathogenicity and IA traits (H, M and O/Ma/D/S).

To test if the rate of gene evolution varies between pathogenic species that are IA or non-IA, we used the birth-death model implemented in CAFE v5, to fit one, two or four parameters of lambda on a subset of 527 small gene families. Pathogenicity and IA trait states were determined for each node in the phylogeny using function ace in the R package phytools. A model with equal rates of transition was used for IA (ER model) and a model with non-equal transition rates was used for pathogenicity (ADR model). Models were compared using likelihood ratio tests.

Functional gene classes

Functional gene annotations were determined using several databases. Protein sequences were searched against KOG (release 2003 (Tatusov et al. 2003)) and MEROPS Scan Sequences (release 12.1 (Rawlings et al. 2018)) using diamond v2.0.9 (Buchfink, Xie, and Huson 2015) with options '--more-sensitive -e 1e-10' and against CAZymes (download 24/09/2021 (Cantarel et al. 2009)) with options '--more-sensitive -e 1e-102'. Pfam domains were searched with pfam_scan.pl (Mistry et al. 2021) script with hmmer v3.2.1 (Mistry et al. 2013) 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 (Blin et al. 2019).

For each functional group of genes, and for each one of 19 selected clades, the ancestral number of genes/clusters was inferred in the most recent ancestor of the clade and its sister clade using the fastAnc function from R package phytools (Revell 2011). Fold change was calculated as a difference in observed and ancestral state divided by ancestral state and averaged across all clade members. Gains or losses of genes per clade were tested with a bootstrap of 10 species with 100 replicates for clades with >= 10 members.

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_2023).

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.

Figure Supplements

Supplementary Figure 1. Loadings of the main PCs. Principal component analysis of genomic traits. Histogram shows proportion of variance explained by each PC. Dark gray PCs together explain 66% of the variance. Heatmap shows loadings for each PC, with brown colors depicting positive, and green colors depicting negative contributions to the PC.

Supplementary Figure 2. Distributions of genomic traits across lifestyles. Boxplots indicate 25th and 75th percentiles of distributions, and dots indicate mean values. IA stands for insect-associated.

Supplementary Figure 3. Numbers of gains and losses among four transition types obtained in BayesTraits run. Opaque colors show transitions significantly shifted towards gains or losses. The x axis lists all four possible transitions for the two traits (genomic trait and pathogenicity). Two states (before change, after change) are separated by an underscore. 1 stands for either a value of a trait above median or presence of pathogenicity, 0 stands for a value of a trait below median or absence of pathogenicity. For example transition 01_11, indicates a gain of a trait (eg. transition from a small to a large genome size, 0x_1x), and no change in pathogenicity (x1_x1).

Supplementary Figure 4. Insect-associated pathogens lose genes involved in breaking host barriers. Heatmap shows the fold change in genes/clusters number relative to the ancestral state for pathogenic species only. Clades are shown in columns (clade names correspond to the ones in Figure 2A) with the number of clade members in parentheses; functional classes are shown in rows. The dots indicate significant gain (red) or loss (green) of genes/clusters across clade members estimated from 100 rounds of bootstrapping of 10 species in clades with >= 10 members. SMC: secondary metabolite clusters, CAZy: carbohydrate-active enzymes.

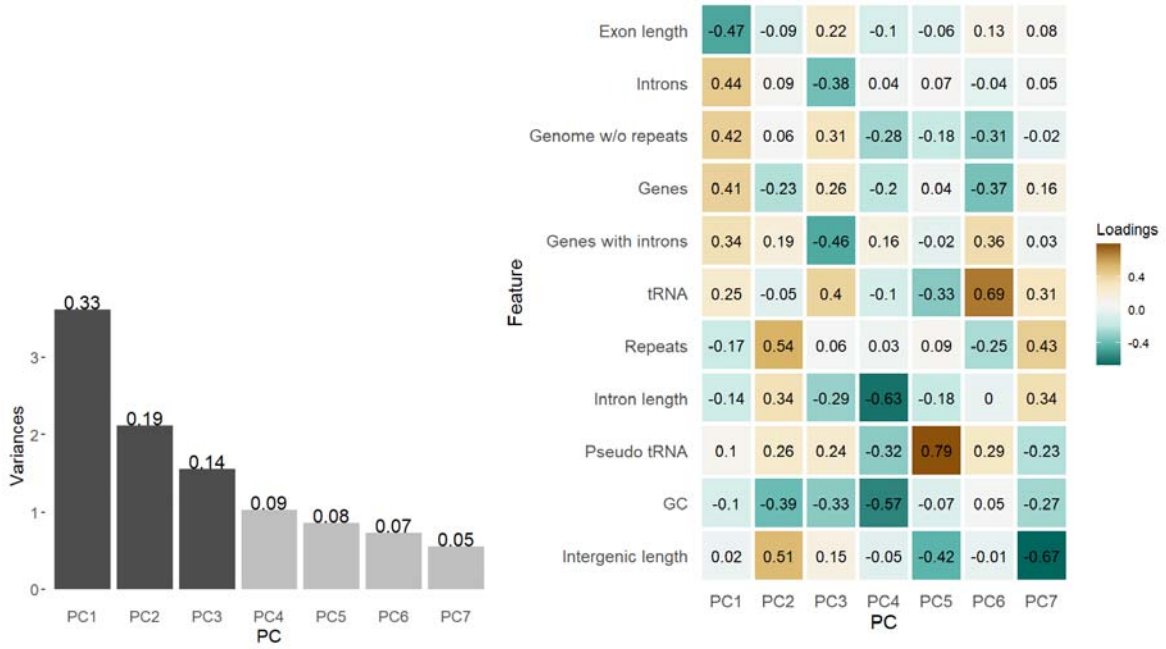
Supplementary Results. An overview of genome assemblies and their annotations, in relation to different sequencing technologies and JGI Mycocosm database.

Supplementary Files

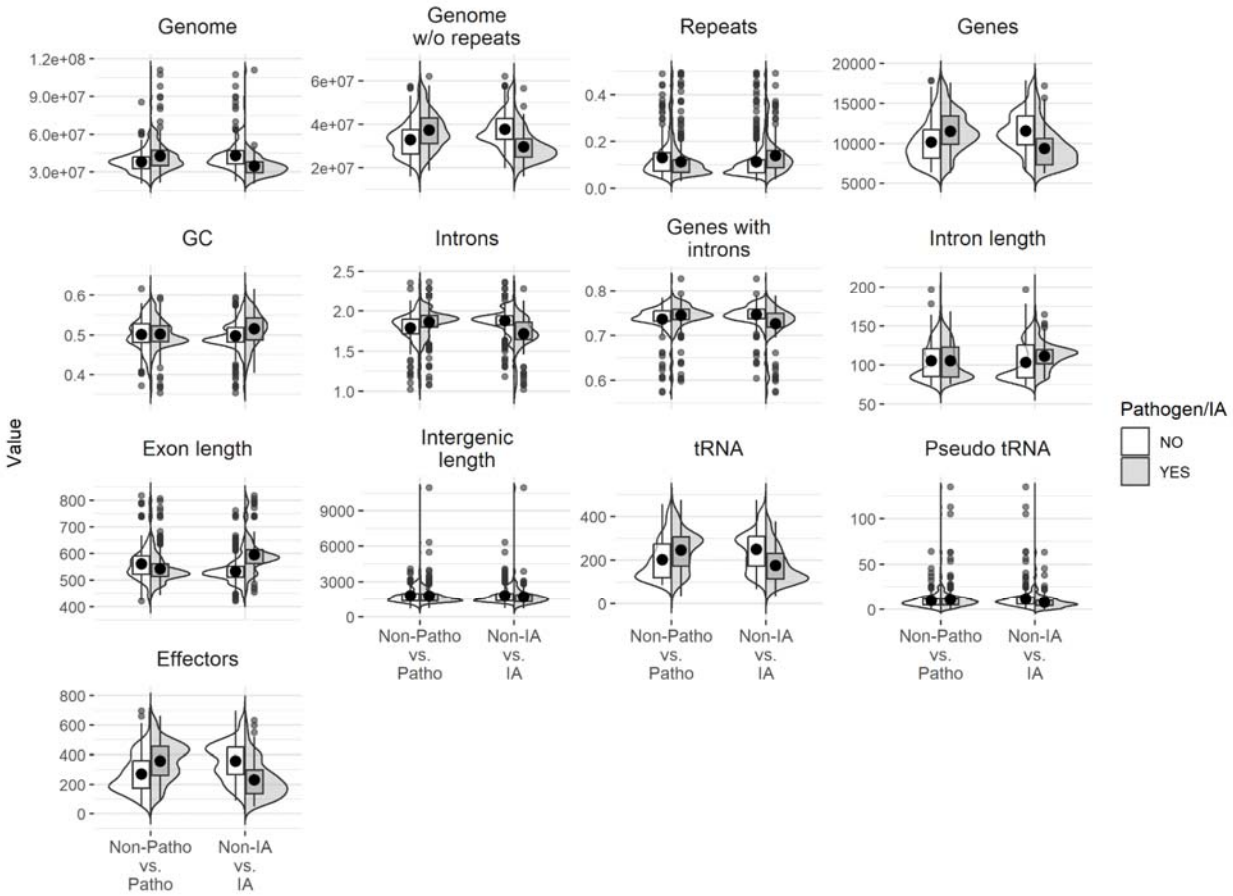
Supplementary Table 1. Information on Sordariomycetes genome assemblies including assembly identifiers, quality statistics and estimated trait values.
Supplementary Table 2. Sequencing information for species sequenced in this study.
Supplementary Table 3. Log marginal likelihoods and bayes factors from likelihood ratio tests comparing different models of coevolution of pathogenicity with genomic traits.
Supplementary Table 4. Log marginal likelihoods and bayes factors from likelihood ratio tests comparing different models of coevolution of insect association with genomic traits.

Source Data

Figure 1-Source Data 1. Spearman's rho correlation coefficients for pairs of genomic traits.
Figure 1-Source Data 2. Principal components and loadings of all genomic traits.
Figure 1-Source Data 3. Contrasts of genomic traits.
Figure 2-Source Data 1. Results of three tests of association of genomic traits with pathogenicity.
Figure 2-Source Data 2. Results of three tests of association of genomic traits with pathogenicity conducted in 10 random balanced subsets.
Figure 2-Source Data 3. Numbers of gains and losses among four transition types obtained in BayesTraits run.
Figure 2-Source Data 4. Pathogenicity, raw and binary genomic traits for 10 subsampled datasets.
Figure 3-Source Data 1. Results of three tests of association of genomic traits with insect association.
Figure 3-Source Data 2. Exon and intron metrics in 38 one-to-one orthologs.
Figure 3-Source Data 3. Prevalence of gene families within clades and two exon metrics, number and lengt of exons.
Figure 4-Source Data 1. Coefficients of genomic traits for pathogenic and non-pathogenic species with 95% credible intervals, estimated with phylogenetic logistic regression.
Figure 4-Source Data 2. Rates of gene loss and gain in four groups of species with different pathogenicity and IA status.
Figure 5-Source Data 1. Fold change of genes/clusters relative to the ancestral state ((observed - ancestral)/ ancestral state).
Figure 5-Source Data 2. Mean and high and low confidence intervals of fold change from bootstrapping species within clades with >= 10 members.
Figure 5-Source Data 3. Fold change of genes/clusters relative to the ancestral state ((observed - ancestral)/ ancestral state), in pathogenic species only.
Figure 5-Source Data 4. Mean and high and low confidence intervals of fold change from bootstrapping species within clades with >= 10 pathogenic members.



Supplementary Figure 1. Loadings of the main PCs. Principal component analysis of genomic traits. Histogram shows proportion of variance explained by each PC. Dark gray PCs together explain 66% of the variance. Heatmap shows loadings for each PC, with brown colors depicting positive, and green colors depicting negative contributions to the PC.



Supplementary Figure 2. Distributions of genomic traits across lifestyles. Boxplots indicate 25th and 75th percentiles of distributions, and dots indicate mean values. IA stands for insect-associated.

Supplementary Figure 3. Numbers of gains and losses among four transition types obtained in BayesTraits run. Opaque colors show transitions significantly shifted towards gains or losses. The x axis lists all four possible transitions for the two traits (genomic trait and pathogenicity). Two states (before change, after change) are separated by an underscore. 1 stands for either a value of a trait above median or presence of pathogenicity, 0 stands for a value of a trait below median or absence of pathogenicity. For example transition 01_11, indicates a gain of a trait (eg. transition from a small to a large genome size, 0x_1x), and no change in pathogenicity (x1_x1).

Supplementary Figure 4. Insect-associated pathogens lose genes involved in breaking host barriers. Heatmap shows the fold change in genes/clusters number relative to the ancestral state for pathogenic species only. Clades are shown in columns (clade names correspond to the ones in Figure 2A) with the number of clade members in parentheses; functional classes are shown in rows. The dots indicate significant gain (red) or loss (green) of genes/clusters across clade members estimated from 100 rounds of bootstrapping of 10 species in clades with >= 10 members. SMC: secondary metabolite clusters, CAZy: carbohydrate-active enzymes.

Supplementary Results

1. Impact of sequencing technology on the annotation of trait values
2. Comparison of assembly length and repeat content with JGI Mycocosm annotations
3. Comparison of gene annotations with JGI Mycocosm
4. Impact of the training set on gene annotations

1. Impact of sequencing technology on annotation of trait values

We investigated the potential impact of sequencing technology used to generate genome assemblies on annotation of genomic features. For each genome assembly, information on sequencing technology was extracted using NCBI datasets v18.0.2. Genome assembly sequencing technology was classified as long-read if it contained technology including Nanopore (ONT), PacBio or Sanger sequencing, otherwise it was classified as short-read assembly. 461 assemblies were classified as short-read, and 102 as long-read. Long-read assemblies were present across all of the large clades: Diaporthales (19%), Glomerellales (29%), Hypocreales (32%), Magnaporthales (10%), Microascales (12%), Ophiostomatales (29%), Sordariales (29%), Xylariales (61%). Pathogenic species comprised 63% of species in long-read vs 65% in short-read assemblies, and IA (insect-associated) species comprised 23% of species in long-read vs 25% in short-read assemblies (Fig. S5).

Reads	Non-pathogen	Pathogen
long	0.37	0.63
short	0.35	0.65

Reads	Non-IA	IA
long	0.77	0.23
short	0.75	0.25

Fig. S5. Proportion of species with different lifestyles among short and long-read assemblies.

The mean values of 13 traits were similar between long and short-read assemblies, with largely overlapping distributions (Fig. S6). Since the two sets consist of non-overlapping species, variation around the mean is expected. Median values in long-read assembly species were higher for total assembly length by 3.16 Mb, repeat content by 0.04, mean intron length by 6 bp, and mean intergenic length by 518 bp, and they were lower in long-read assembly species for the number of genes by 548, effectors by 34 and tRNA genes by 13.

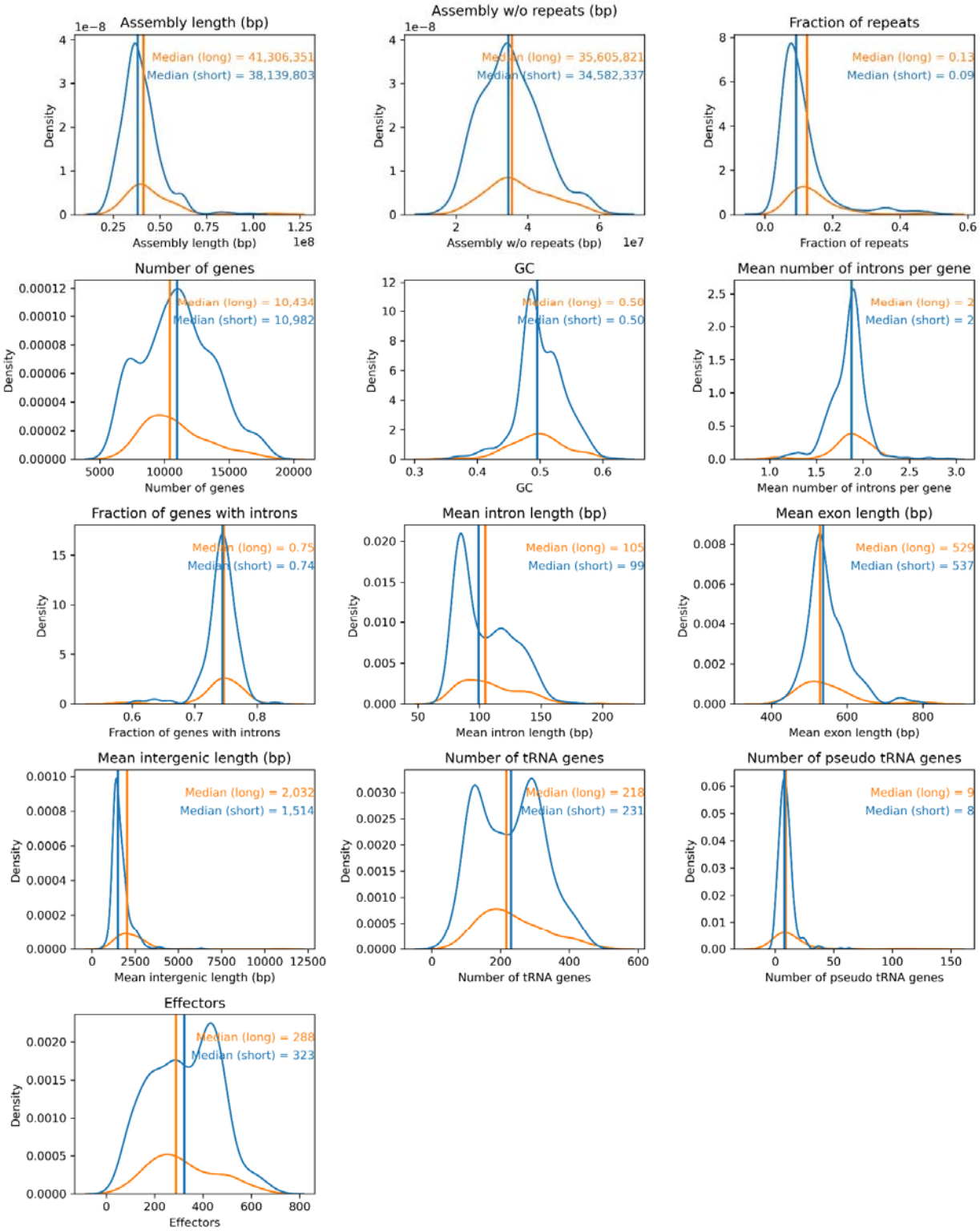


Fig. S6. Distribution of trait values calculated for long-read assembly (orange) and short-read assembly (blue) species. Vertical lines indicate the position of median values.

To investigate if genomic traits vary between lifestyles in the same way for long and short-read assemblies, we compared trait distributions for long and short-read assemblies between pathogens and non-pathogens (Fig. S7) and between IA and non-IA species (Fig. S8). Differences in trait values between lifestyles were similar and in the same direction for both short and long-read assemblies, except for the mean intron length, and to a lesser degree in GC content and the number of pseudo tRNAs in pathogens vs. non-pathogens (Fig. S7). Mean intron length was longer for pathogens than non-pathogens in long-read assemblies, but shorter for pathogens than non-pathogens in short-read assemblies.

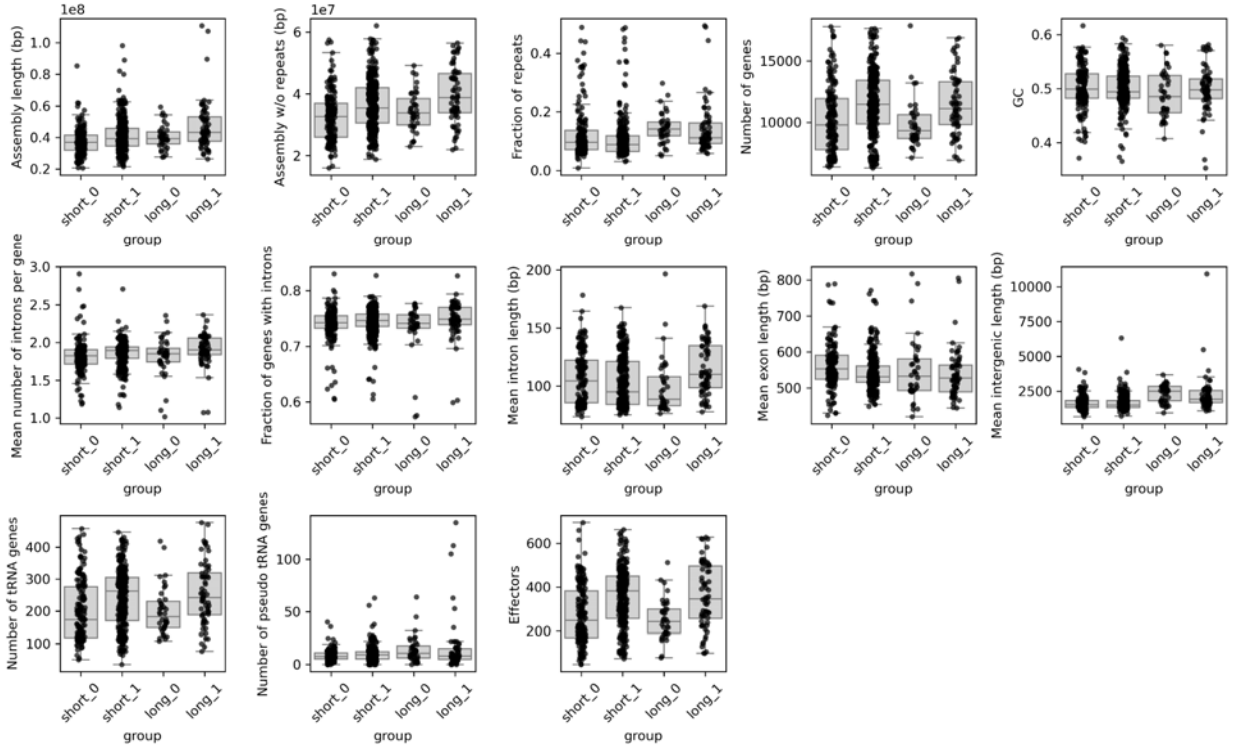


Fig. S7. Distribution of genomic trait values for short and long-read assemblies separately for pathogenic (1) and non-pathogenic (0) species.

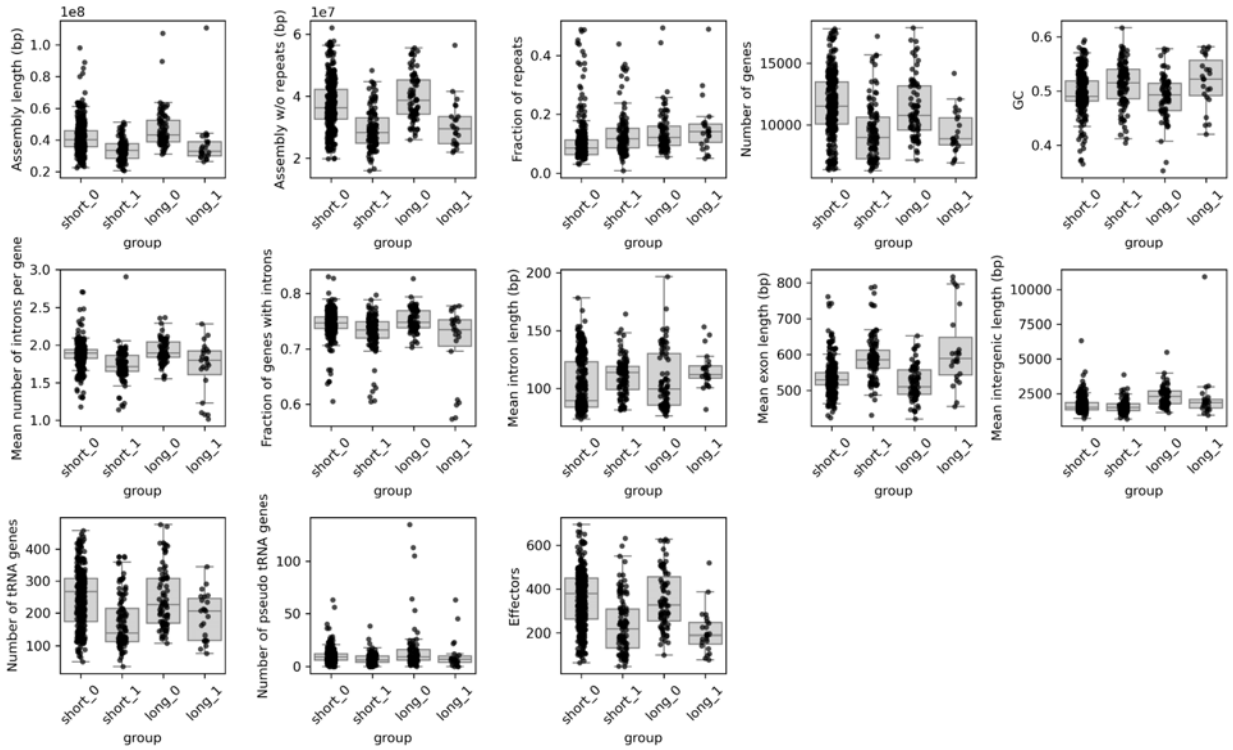


Fig. S8. Distribution of genomic trait values for short and long-read assemblies separately for IA (insect-associated) (1) and non-IA (0) species.

2. Comparison of assembly length and repeat content with JGI Mycocosm annotations

Genomic resources such as JGI Mycocosm provide another source of fungal genomes with standard pipelines for gene and repeat annotations. To investigate how well estimated genomic traits from our dataset (raw repeat annotation from NCBI and Augustus gene annotation) compare with those of JGI Mycocosm, we downloaded genome assemblies and gene annotations for published Sordariomycetes species from JGI Mycocosm (downloaded on 17/05/2025). Only 58 species were matching our dataset, and 19 of them were matching also at the strain level. The species set (n=58) included 33 Hypocreales, 11 Xylariales, 5 Sordariales, 4 Glomerellales and single representatives of three other clades. The strain set (n=19) included 14 Hypocreales, 2 Xylariales, 2 Diaporthales and 1 Coniochaetales species. Only three of these strains were sequenced with long-read sequencing technology (Sanger or PacBio). Genome assemblies were very similar in length between NCBI and Mycocosm datasets at the species level (median of 40.9 Mbp for NCBI vs 41.3 Mbp for Mycocosm), and at the strain level (median of 42.2 Mbp for NCBI vs 42.4 Mbp for Mycocosm, Fig. S9A). Differences between lifestyles were in the same direction for the NCBI and Mycocosm datasets (Fig. S9A). Repeat content (proportion of masked bases in the assembly) was higher for NCBI than Mycocosm assemblies both for the species (median of 0.091 for NCBI vs 0.053 for Mycocosm) and strain comparisons (median of 0.081 for NCBI vs 0.054 for Mycocosm, Fig. S9B). For same strain comparison,

repeat content in Mycocosm assemblies was lower for pathogens than non-pathogens (median of 0.054 in pathogens vs. 0.077 in non-pathogens) compared to NCBI assemblies which had similar repeat content for both lifestyles (median of 0.081 in pathogens vs. 0.076 in non-pathogens, Fig. S9B). Notably, this result is based on only four pathogenic species, which is too few to extrapolate to the entire dataset. Number of IA species in the dataset was generally very low (and absent from same strain dataset) to draw meaningful conclusions.

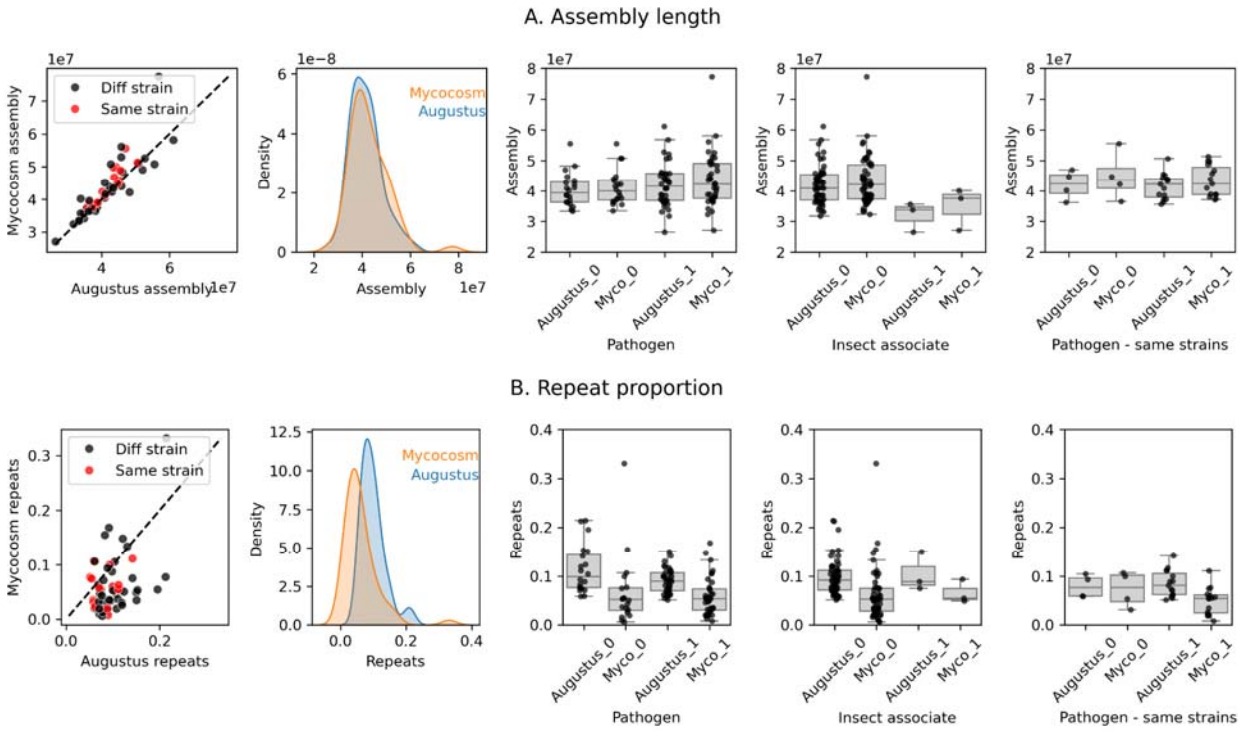


Fig. S9. Comparison of assembly length and repeat annotations between matching species from our dataset (NCBI) and JGI Mycocosm. In boxplots categories “1” corresponds to pathogens or IA (insect-associated) species, and “0” to non-pathogens or non-IA species.

3. Comparison of gene annotations with JGI Mycocosm

Similarly we compared five gene traits (number of genes, number of introns, intron length, exon length, and fraction of genes with introns) between our dataset annotated with Augustus, and JGI Mycocosm annotations. For the JGI Mycocosm dataset, gene traits were calculated from gff files with the same scripts as those from Augustus. Number of genes was systematically higher for Mycocosm than Augustus datasets for almost all 58 matching species, and it was higher for both pathogenic and non-pathogenic species (Fig. S10A). Other gene traits were on average slightly lower for Mycocosm than Augustus datasets, in particular for the same strain comparisons in both pathogen and non-pathogenic species (Fig. S10).

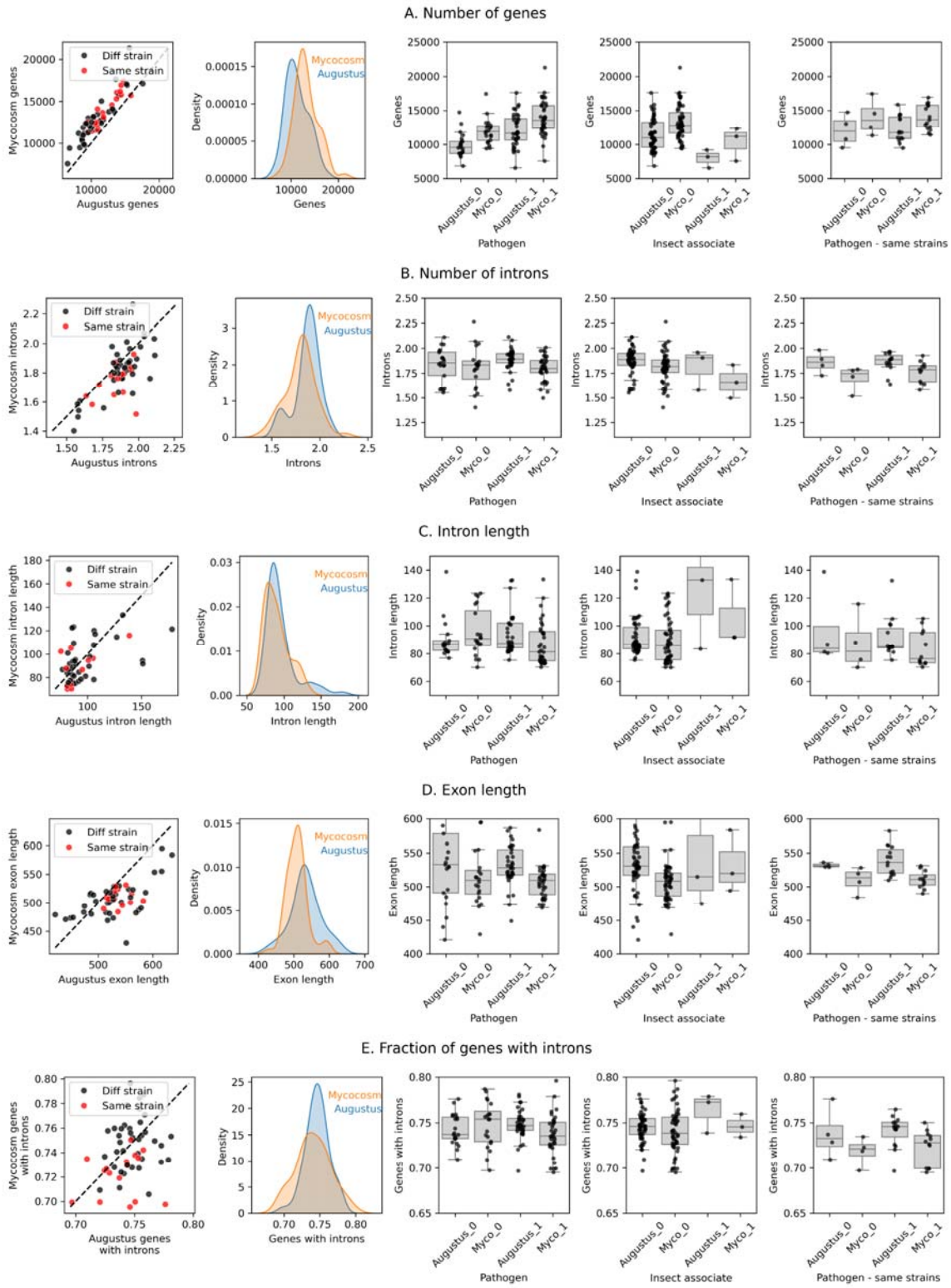


Fig. S10. Comparison of gene annotations between matching species from our dataset (NCBI) and JGI Mycocosm. In boxplots categories “1” corresponds to pathogens or IA (insect-associated) species, and “0” to non-pathogens or non-IA species.

4. Impact of the training set on gene annotations

To investigate whether the choice of the species as a training set for Augustus gene annotation had an impact on gene traits, we compared five gene traits as a function of distance from the species used in a training set. Distance was calculated as the cumulative branch length distance estimated from the phylogenetic tree using dist.nodes() function from R package ape v5.8. The scatterplots are shown in Fig. S11. Overall there were no consistent patterns across training sets and within each training set differences between species belonging to different orders were usually more pronounced than those within orders. In the case of the fusarium training set, there was a clear decrease in the number of genes with increasing distance, and an increase in intron and exon length with increasing distance (Fig. S11). To verify if these trends could be impacted by a species selected for training, we compared gene traits between species annotated with Augustus (n=353) and species from the JGI Mycocosm dataset (n=146 including species not overlapping with Augustus) in three groups of species with increasing distance from Fusarium: species from the order Hypocreales belonging to Fusarium genus (corresponding to clade H1 in this study), other species from Hypocreales order (corresponding to clade H2 in this study) and species from Microascales order. Even though representation for the Microascales was very small (n=2), and the mean values differed between Augustus and Mycocosm (as demonstrated above), the differences between three groups of species were of the same magnitude and direction (Fig. S12), confirming that these trends are not driven predominantly by Augustus annotation.

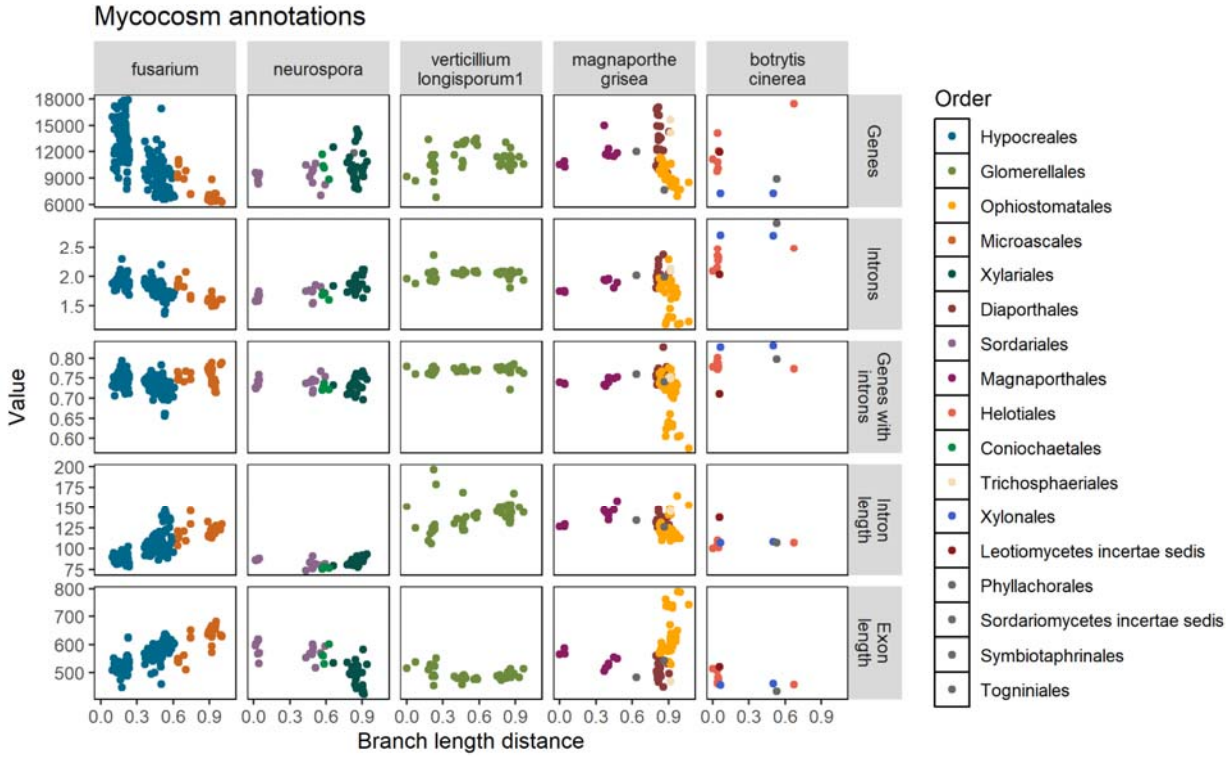


Fig. S11. Gene trait values as a function of distance from the five species or genera used in Augustus training.

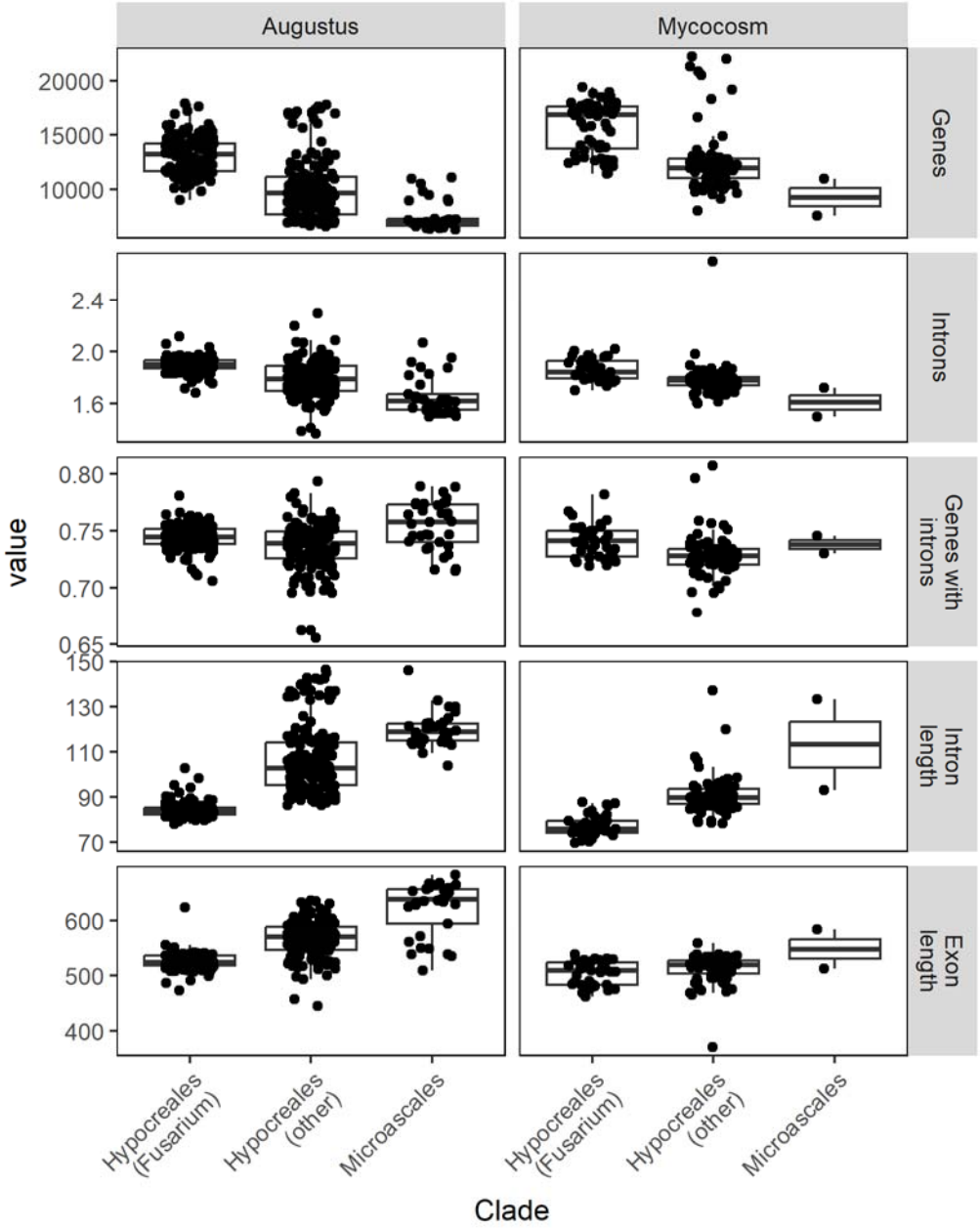


Fig. S12. Distributions and medians of gene traits compared between all species from Hypocreales and Microascales retrieved from JGI Mycocosm, and all species from the same two orders in our dataset annotated with Augustus using “fusarium” for training. Species were split into three groups with increasing distance from genus Fusarium, namely all Fusarium species in Hypocreales order, all other Hypocreales, and Microascales.

References

Andrews, Simon. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Badet, Thomas, and Daniel Croll. 2025. “A Systematic Screen for Genetic Factors Underpinning Transposon Defense Systems across the Fungal Kingdom.” bioRxiv. https://doi.org/10.1101/2025.01.10.632494.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.” Journal of Computational Biology: A Journal of Computational Molecular Cell Biology 19 (5): 455–77.

Bao, Jiandong, Meilian Chen, Zhenhui Zhong, Wei Tang, Lianyu Lin, Xingtan Zhang, Haolang Jiang, et al. 2017. “PacBio Sequencing Reveals Transposable Elements as a Key Contributor to Genomic Plasticity and Virulence Variation in Magnaporthe Oryzae.” Molecular Plant 10 (11): 1465–68.

Baroncelli, Riccardo, Daniel Buchvaldt Amby, Antonio Zapparata, Sabrina Sarrocco, Giovanni Vannacci, Gaétan Le Floch, Richard J. Harrison, et al. 2016. “Gene Family Expansions and Contractions Are Associated with Host Range in Plant Pathogens of the Genus Colletotrichum.” BMC Genomics 17 (August):555.

Bergin, Sean A., Fang Zhao, Adam P. Ryan, Carolin A. Müller, Conrad A. Nieduszynski, Bing Zhai, Thierry Rolling, et al. 2022. “Systematic Analysis of Copy Number Variations in the Pathogenic Yeast Candida Parapsilosis Identifies a Gene Amplification in RTA3 That Is Associated with Drug Resistance.” mBio 13 (5): e0177722.

Biderre, C., F. Duffieux, E. Peyretaillade, P. Glaser, P. Peyret, A. Danchin, M. Pagès, G. Méténier, and C. P. Vivarès. 1997. “Mapping of Repetitive and Non-Repetitive DNA Probes to Chromosomes of the Microsporidian Encephalitozoon Cuniculi.” Gene 191 (1): 39–45.

Blin, Kai, Victòria Pascal Andreu, Emmanuel L. C. de Los Santos, Francesco Del Carratore, Sang Yup Lee, Marnix H. Medema, and Tilmann Weber. 2019. “The antiSMASH Database Version 2: A Comprehensive Resource on Secondary Metabolite Biosynthetic Gene Clusters.” Nucleic Acids Research 47 (D1): D625–30.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” Bioinformatics 30 (15): 2114–20.

Boller, Thomas, and Sheng Yang He. 2009. “Innate Immunity in Plants: An Arms Race between Pattern Recognition Receptors in Plants and Effectors in Microbial Pathogens.” Science (New York, N.Y.) 324 (5928): 742–44.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. “Fast and Sensitive Protein Alignment Using DIAMOND.” Nature Methods 12 (1): 59–60.

Bushnell, Brian, Jonathan Rood, and Esther Singer. 2017. “BBMerge - Accurate Paired Shotgun Read Merging via Overlap.” PloS One 12 (10): e0185056.

Cantarel, Brandi L., Pedro M. Coutinho, Corinne Rancurel, Thomas Bernard, Vincent Lombard, and Bernard Henrissat. 2009. “The Carbohydrate-Active EnZymes Database (CAZy): An Expert Resource for Glycogenomics.” Nucleic Acids Research 37 (Database issue): D233–38.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. “trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” Bioinformatics 25 (15): 1972–73.

Chan, Patricia P., Brian Y. Lin, Allysia J. Mak, and Todd M. Lowe. 2021. “tRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes.” Nucleic Acids Research 49 (16): 9077–96.

Chin, Chen-Shan, Paul Peluso, Fritz J. Sedlazeck, Maria Nattestad, Gregory T. Concepcion,

Alicia Clum, Christopher Dunn, et al. 2016. “Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing.” Nature Methods 13 (12): 1050–54.

Christendat, Dinesh. 2013. “Ophiostoma Ulmi Resource Browser.” 2013. http://www.moseslab.csb.utoronto.ca/o.ulmi/.

Darling, Aaron C. E., Bob Mau, Frederick R. Blattner, and Nicole T. Perna. 2004. “Mauve: Multiple Alignment of Conserved Genomic Sequence with Rearrangements.” Genome Research 14 (7): 1394–1403.

Dierckxsens, Nicolas, Patrick Mardulyn, and Guillaume Smits. 2017. “NOVOPlasty: De Novo Assembly of Organelle Genomes from Whole Genome Data.” Nucleic Acids Research 45 (4): e18.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” Bioinformatics 29 (1): 15–21.

Dong, Suomeng, Sylvain Raffaele, and Sophien Kamoun. 2015. “The Two-Speed Genomes of Filamentous Pathogens: Waltz with Plants.” Current Opinion in Genetics & Development 35 (December):57–65.

Dort, E. N., E. Layne, N. Feau, A. Butyaev, B. Henrissat, F. M. Martin, S. Haridas, et al. 2023. “Large-Scale Genomic Analyses with Machine Learning Uncover Predictive Patterns Associated with Fungal Phytopathogenic Lifestyles and Traits.” Scientific Reports 13 (1): 17203.

Emms, David M., and Steven Kelly. 2019. “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics.” Genome Biology 20 (1): 238.

Fisher, Matthew C., Sarah J. Gurr, Christina A. Cuomo, David S. Blehert, Hailing Jin, Eva H. Stukenbrock, Jason E. Stajich, et al. 2020. “Threats Posed by the Fungal Kingdom to Humans, Wildlife, and Agriculture.” mBio 11 (3). https://doi.org/10.1128/mBio.00449-20.

Flynn, Jullien M., Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. 2020. “RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families.” Proceedings of the National Academy of Sciences of the United States of America 117 (17): 9451–57.

Galagan, James E., Sarah E. Calvo, Katherine A. Borkovich, Eric U. Selker, Nick D. Read, David Jaffe, William FitzHugh, et al. 2003. “The Genome Sequence of the Filamentous Fungus Neurospora Crassa.” *Nature* 422 (6934): 859–68.

Gibbs, Richard A., George M. Weinstock, Michael L. Metzker, Donna M. Muzny, Erica J. Sodergren, Steven Scherer, Graham Scott, et al. 2004. “Genome Sequence of the Brown Norway Rat Yields Insights into Mammalian Evolution.” Nature 428 (6982): 493–521.

Gómez Luciano, Luis B., Isheng Jason Tsai, Izumi Chuma, Yukio Tosa, Yi-Hua Chen, Jeng-Yi Li, Meng-Yun Li, Mei-Yeh Jade Lu, Hitoshi Nakayashiki, and Wen-Hsiung Li. 2019. “Blast Fungal Genomes Show Frequent Chromosomal Changes, Gene Gains and Losses, and Effector Gene Turnover.” Molecular Biology and Evolution 36 (6): 1148–61.

Hane, James K., Jonathan Paxman, Darcy A. B. Jones, Richard P. Oliver, and Pierre de Wit. 2019. “‘CATAStrophy,’ a Genome-Informed Trophic Classification of Filamentous Plant Pathogens - How Many Different Types of Filamentous Plant Pathogens Are There?” Frontiers in Microbiology 10:3088.

Haridas, S., R. Albert, M. Binder, J. Bloem, K. LaButti, A. Salamov, B. Andreopoulos, et al. 2020. “101 Dothideomycetes Genomes: A Test Case for Predicting Lifestyles and Emergence of Pathogens.” Studies in Mycology 96 (June):141–53.

Hartmann, Fanny E., and Daniel Croll. 2017. “Distinct Trajectories of Massive Recent Gene Gains and Losses in Populations of a Microbial Eukaryotic Pathogen.” Molecular Biology and Evolution 34 (11): 2808–22.

Hartmann, Fanny E., Andrea Sánchez-Vallet, Bruce A. McDonald, and Daniel Croll. 2017. “A Fungal Wheat Pathogen Evolved Host Specialization by Extensive Chromosomal

Rearrangements." The ISME Journal 11 (5): 1189–1204.
Hedges, Blair, and Sudhir Kumar. 2005. "TIMETREE 5: The Timescale of Life." 2005. http://www.timetree.org/.
Hensen, Noah, Lucas Bonometti, Ivar Westerberg, Ioana Onut Brännström, Sonia Guillou, Sandrine Cros-Aarteil, Sara Calhoun, et al. 2023. "Genome-Scale Phylogeny and Comparative Genomics of the Fungal Order Sordariales." Molecular Phylogenetics and Evolution 189 (107938): 107938.
Hessenauer, Pauline, Anna Fijarczyk, Hélène Martin, Julien Prunier, Guillaume Charron, Jérôme Chapuis, Louis Bernier, Philippe Tanguay, Richard C. Hamelin, and Christian R. Landry. 2020. "Hybridization and Introgression Drive Genome Evolution of Dutch Elm Disease Pathogens." Nature Ecology & Evolution 4 (4): 626–38.
Hinsch, Janine, Petr Galuszka, and Paul Tudzynski. 2016. "Functional Characterization of the First Filamentous Fungal tRNA-Isopentenyltransferase and Its Role in the Virulence of Claviceps Purpurea." The New Phytologist 211 (3): 980–92.
Hoang, Diep Thi, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh. 2018. "UFBoot2: Improving the Ultrafast Bootstrap Approximation." Molecular Biology and Evolution 35 (2): 518–22.
Hoff, Katharina J., Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke. 2019. "Whole-Genome Annotation with BRAKER." In Gene Prediction: Methods and Protocols, edited by Martin Kollmar, 65–95. New York, NY: Springer New York.
Ho, Lam si Tung, and Cécile Ané. 2014. "A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models." Systematic Biology 63 (3): 397–408.
Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K. F. Wong, Arndt von Haeseler, and Lars S. Jermiin. 2017. "ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates." Nature Methods 14 (6): 587–89.
Katinka, M. D., S. Duprat, E. Cornillot, G. Méténier, F. Thomarat, G. Prensier, V. Barbe, et al. 2001. "Genome Sequence and Gene Compaction of the Eukaryote Parasite Encephalitozoon Cuniculi." Nature 414 (6862): 450–53.
Katoh, Kazutaka, and Daron M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." Molecular Biology and Evolution 30 (4): 772–80.
Kelkar, Yogeshwar D., and Howard Ochman. 2012. "Causes and Consequences of Genome Expansion in Fungi." Genome Biology and Evolution 4 (1): 13–23.
Kuo, Chih-Horng, Nancy A. Moran, and Howard Ochman. 2009. "The Consequences of Genetic Drift for Bacterial Genome Complexity." Genome Research 19 (8): 1450–54.
Lefébure, Tristan, Claire Morvan, Florian Malard, Clémentine François, Lara Konecny-Dupré, Laurent Guéguen, Michèle Weiss-Gayet, et al. 2017. "Less Effective Selection Leads to Larger Genomes." Genome Research 27 (6): 1016–28.
Li, Gang, Nawaraj Dulal, Ziwen Gong, and Richard A. Wilson. 2023. "Unconventional Secretion of Magnaporthe Oryzae Effectors in Rice Cells Is Regulated by tRNA Modification and Codon Usage Control." Nature Microbiology 8 (9): 1706–16.
Li, Heng. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." Bioinformatics 34 (18): 3094–3100.
Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." Bioinformatics 25 (14): 1754–60.
Lomsadze, Alexandre, Vardges Ter-Hovhannisyan, Yury O. Chernoff, and Mark Borodovsky. 2005. "Gene Identification in Novel Eukaryotic Genomes by Self-Training Algorithm." Nucleic Acids Research 33 (20): 6494–6506.
Loos, Daniel, Ailton Pereira da Costa Filho, Bas E. Dutilh, Amelia E. Barber, and Gianni Panagiotou. 2024. "A Global Survey of Host, Aquatic, and Soil Microbiomes Reveals Shared Abundance and Genomic Features between Bacterial and Fungal Generalists." Cell

Reports 43 (4): 114046.
Lynch, Michael. 2007. The Origins of Genome Architecture. Sinauer Associates.
Lynch, Michael, and John S. Conery. 2003. “The Origins of Genome Complexity.” Science 302 (5649): 1401–4.
Marino, Alba, Gautier Debaecker, Anna-Sophie Fiston-Lavier, Annabelle Haudry, and Benoit Nabholz. 2024. “Effective Population Size Does Not Explain Long-Term Variation in Genome Size and Transposable Element Content in Animals.” eLife. https://doi.org/10.7554/elife.100574.1.
McDonald, Megan C., Adam P. Taranto, Erin Hill, Benjamin Schwessinger, Zhaohui Liu, Steven Simpfendorfer, Andrew Milgate, and Peter S. Solomon. 2019. “Transposon-Mediated Horizontal Transfer of the Host-Specific Virulence Protein ToxA between Three Fungal Wheat Pathogens.” mBio 10 (5). https://doi.org/10.1128/mBio.01515-19.
Mendes, Fábio K., Dan Vanderpool, Ben Fulton, and Matthew W. Hahn. 2020. “CAFE 5 Models Variation in Evolutionary Rates among Gene Families.” Bioinformatics , December. https://doi.org/10.1093/bioinformatics/btaa1022.
Merényi, Zsolt, Krisztina Krizsán, Neha Sahu, Xiao-Bin Liu, Balázs Bálint, Jason E. Stajich, Joseph W. Spatafora, and László G. Nagy. 2023. “Genomes of Fungi and Relatives Reveal Delayed Loss of Ancestral Gene Families and Evolution of Key Fungal Traits.” Nature Ecology & Evolution 7 (8): 1221–31.
Mira, A., and N. A. Moran. 2002. “Estimating Population Size and Transmission Bottlenecks in Maternally Transmitted Endosymbiotic Bacteria.” Microbial Ecology 44 (2): 137–43.
Mira, A., H. Ochman, and N. A. Moran. 2001. “Deletional Bias and the Evolution of Bacterial Genomes.” Trends in Genetics: TIG 17 (10): 589–96.
Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, et al. 2021. “Pfam: The Protein Families Database in 2021.” Nucleic Acids Research 49 (D1): D412–19.
Mistry, Jaina, Robert D. Finn, Sean R. Eddy, Alex Bateman, and Marco Punta. 2013. “Challenges in Homology Search: HMMER3 and Convergent Evolution of Coiled-Coil Regions.” Nucleic Acids Research 41 (12): e121.
Mohanta, Tapan Kumar, and Hanhong Bae. 2015. “The Diversity of Fungal Genome.” Biological Procedures Online 17 (April):8.
Möller, Mareike, and Eva H. Stukenbrock. 2017. “Evolution and Genome Architecture in Fungal Plant Pathogens.” Nature Reviews. Microbiology 15 (12): 756–71.
Muszewska, Anna, John W. Taylor, Pawel Szczesny, and Marcin Grynberg. 2011. “Independent Subtilases Expansions in Fungi Associated with Animals.” Molecular Biology and Evolution 28 (12): 3395–3404.
Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. “IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.” Molecular Biology and Evolution 32 (1): 268–74.
Oggenfuss, Ursula, and Daniel Croll. 2023. “Recent Transposable Element Bursts Are Associated with the Proximity to Genes in a Fungal Plant Pathogen.” PLoS Pathogens 19 (2): e1011130.
Pagel, Mark, and Andrew Meade. 2006. “Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo.” The American Naturalist 167 (6): 808–25.
Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. “APE: Analyses of Phylogenetics and Evolution in R Language.” Bioinformatics 20 (2): 289–90.
Pendleton, Amanda L., Katherine E. Smith, Nicolas Feau, Francis M. Martin, Igor V. Grigoriev, Richard Hamelin, C. Dana Nelson, J. Gordon Burleigh, and John M. Davis. 2014. “Duplications and Losses in Gene Families of Rust Pathogens Highlight Putative Effectors.” Frontiers in Plant Science 5 (June):299.

Petrov, D. A. 2002. “Mutational Equilibrium Model of Genome Size Evolution.” Theoretical Population Biology 61 (4): 531–44.

Raffaele, Sylvain, and Sophien Kamoun. 2012. “Genome Evolution in Filamentous Plant Pathogens: Why Bigger Can Be Better.” Nature Reviews. Microbiology 10 (6): 417–30.

Raffa, Nicholas, and Nancy P. Keller. 2019. “A Call to Arms: Mustering Secondary Metabolites for Success and Survival of an Opportunistic Pathogen.” PLoS Pathogens 15 (4): e1007606.

Rawlings, Neil D., Alan J. Barrett, Paul D. Thomas, Xiaosong Huang, Alex Bateman, and Robert D. Finn. 2018. “The MEROPS Database of Proteolytic Enzymes, Their Substrates and Inhibitors in 2017 and a Comparison with Peptidases in the PANTHER Database.” Nucleic Acids Research 46 (D1): D624–32.

Revell, Liam J. 2011. “Phytools: An R Package for Phylogenetic Comparative Biology (and Other Things).” Methods in Ecology and Evolution 3 (2): 217–23.

Rokas, Antonis. 2022. “Evolution of the Human Pathogenic Lifestyle in Fungi.” Nature Microbiology 7 (5): 607–19.

Sahu, Neha, Boris Indic, Johanna Wong-Bajracharya, Zsolt Merényi, Huei-Mien Ke, Steven Ahrendt, Tori-Lee Monk, et al. 2023. “Vertical and Horizontal Gene Transfer Shaped Plant Colonization and Biomass Degradation in the Fungal Genus Armillaria.” Nature Microbiology 8 (9): 1668–81.

Sánchez-Vallet, Andrea, Simone Fouché, Isabelle Fudal, Fanny E. Hartmann, Jessica L. Soyer, Aurélien Tellier, and Daniel Croll. 2018. “The Genome Biology of Effector Gene Evolution in Filamentous Plant Pathogens.” Annual Review of Phytopathology 56 (August):21–40.

Sanderson, Michael J. 2003. “r8s: Inferring Absolute Rates of Molecular Evolution and Divergence Times in the Absence of a Molecular Clock.” Bioinformatics 19 (2): 301–2.

Scharf, Daniel H., Thorsten Heinekamp, and Axel A. Brakhage. 2014. “Human and Plant Fungal Pathogens: The Role of Secondary Metabolites.” PLoS Pathogens 10 (1): e1003859.

Shen, Xing-Xing, Jacob L. Steenwyk, Abigail L. LaBella, Dana A. Opulente, Xiaofan Zhou, Jacek Kominek, Yuanning Li, Marizeth Groenewald, Chris T. Hittinger, and Antonis Rokas. 2020. “Genome-Scale Phylogeny and Contrasting Modes of Genome Evolution in the Fungal Phylum Ascomycota.” *Science Advances* 6 (45): eabd0079.

Sipos, György, Arun N. Prasanna, Mathias C. Walter, Eoin O’Connor, Balázs Bálint, Krisztina Krizsán, Brigitta Kiss, et al. 2017. “Genome Expansion and Lineage-Specific Genetic Innovations in the Forest Pathogenic Fungi Armillaria.” Nature Ecology & Evolution 1 (12): 1931–41.

Sperschneider, Jana, and Peter N. Dodds. 2022. “EffectorP 3.0: Prediction of Apoplastic and Cytoplasmic Effectors in Fungi and Oomycetes.” Molecular Plant-Microbe Interactions: MPMI 35 (2): 146–56.

Stajich, Jason E. 2017. “Fungal Genomes and Insights into the Evolution of the Kingdom.” Microbiology Spectrum 5 (4). https://doi.org/10.1128/microbiolspec.FUNK-0055-2016.

Stalder, Luzia, Ursula Oggenfuss, Norfarhan Mohd-Assaad, and Daniel Croll. 2023. “The Population Genetics of Adaptation through Copy Number Variation in a Fungal Plant Pathogen.” Molecular Ecology 32 (10): 2443–60.

Stanke, Mario, and Burkhard Morgenstern. 2005. “AUGUSTUS: A Web Server for Gene Prediction in Eukaryotes That Allows User-Defined Constraints.” Nucleic Acids Research 33 (Web Server issue): W465–67.

Stergiopoulos, Ioannis, and Pierre J. G. M. de Wit. 2009. “Fungal Effector Proteins.” Annual Review of Phytopathology 47 (1): 233–63.

Tarailo-Graovac, Maja, and Nansheng Chen. 2009. “Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences.” Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.] Chapter 4 (March):Unit 4.10.

Tatusov, Roman L., Natalie D. Fedorova, John D. Jackson, Aviva R. Jacobs, Boris Kiryutin,

Eugene V. Koonin, Dmitri M. Krylov, et al. 2003. “The COG Database: An Updated Version Includes Eukaryotes.” BMC Bioinformatics 4 (September):41. Frontiers in Plant Science 5 (August):422.

Teufel, Felix, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Magnús Halldór Gíslason, Silas Irby Pihl, Konstantinos D. Tsirigos, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2022. “SignalP 6.0 Predicts All Five Types of Signal Peptides Using Protein Language Models.” Nature Biotechnology 40 (7): 1023–25.

Torres, David E., Ursula Oggenfuss, Daniel Croll, and Michael F. Seidl. 2020. “Genome Evolution in Fungal Plant Pathogens: Looking beyond the Two-Speed Genome Model.” Fungal Biology Reviews 34 (3): 136–43.

Venditti, Chris, Andrew Meade, and Mark Pagel. 2011. “Multiple Routes to Mammalian Diversity.” Nature 479 (7373): 393–96.

Wacker, Theresa, Nicolas Helmstetter, Duncan Wilson, Matthew C. Fisher, David J. Studholme, and Rhys A. Farrer. 2023. “Two-Speed Genome Evolution Drives Pathogenicity in Fungal Pathogens of Animals.” Proceedings of the National Academy of Sciences of the United States of America 120 (2): e2212633120.

Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, et al. 2014. “Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement.” PloS One 9 (11): e112963.

Waterhouse, Robert M., Mathieu Seppey, Felipe A. Simão, Mosè Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2018. “BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics.” Molecular Biology and Evolution 35 (3): 543–48.

Wells, Jonathan N., and Cédric Feschotte. 2020. “A Field Guide to Eukaryotic Transposable Elements.” Annual Review of Genetics 54 (1): 539–61.

Wilson, Richard A., and John M. McDowell. 2022. “Recent Advances in Understanding of Fungal and Oomycete Effectors.” Current Opinion in Plant Biology 68 (August):102228.

Wong, Barbara, Isabel Leal, Nicolas Feau, Angela Dale, Adnan Uzunovic, and Richard C. Hamelin. 2020. “Molecular Assays to Detect the Presence and Viability of Phytophthora Ramorum and Grosmannia Clavigera.” PloS One 15 (2): e0221742.

Xiao, Chuan-Le, Ying Chen, Shang-Qian Xie, Kai-Ning Chen, Yan Wang, Yue Han, Feng Luo, and Zhi Xie. 2017. “MECAT: Fast Mapping, Error Correction, and de Novo Assembly for Single-Molecule Sequencing Reads.” Nature Methods 14 (11): 1072–74.

Yang, Ziheng. 2007. “PAML 4: Phylogenetic Analysis by Maximum Likelihood.” Molecular Biology and Evolution 24 (8): 1586–91.

Yu, Guangchuang. 2020. “Using Ggtree to Visualize Data on Tree-Like Structures.” Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.] 69 (1): e96.