# Supplementary Figures and Figure Legends for:

# A robust phylogenomic timetree for biotechnologically and medically important fungi in the genera *Aspergillus* and *Penicillium*

Jacob L. Steenwyk1, Xing-Xing Shen1, Abigail L. Lind2,3, Gustavo H. Goldman4, and Antonis Rokas1,2,\*

1 Department of Biological Sciences, Vanderbilt University, Nashville, TN, 37235, USA

2 Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America

3 Gladstone Institute for Data Science and Biotechnology, San Francisco, California, USA

4 Departamento de Ciências Farmacêuticas, Faculdade de Ciências Farmacêuticas de Ribeirão Prêto, Bloco Q, Universidade de São Paulo, São Paulo, Brazil

\* Correspondence: [antonis.rokas@vanderbilt.edu](mailto:antonis.rokas@vanderbilt.edu)

**Supplementary Figure Legends**

### **Supplementary figure 1. Previously reported phylogenies of Aspergillaceae sections**

(top) The four previously reported phylogenies of Aspergillaceae sections based on genome-scale amounts of data. Note that the inferred relationships among sections differ between the four phylogenies and that the sampling of the sections is relatively sparse in all the phylogenies. (middle) The three previously reported phylogenies of Aspergillaceae sections based on multi-gene analyses. Note that here too the inferred relationships among sections differ among the phylogenies but that the sampling of sections in some of the studies is denser. (bottom) Comparisons between the phylogenies inferred from three representative previous studies. Lines denote sections that are shared between the pairs of phylogenies compared. Note the widespread lack of consensus in the inferred relationships among sections and the differences in sections sampled.

## **Supplementary figure 2. The 93 selected genomes exhibit high levels of genome completeness and are of high quality for phylogenomic studies**

81 fungal genomes belonging to the family Aspergillaceae and 12 outgroups were downloaded from NCBI and assessed for genome completeness based on their percentages of single-copy, duplicated, fragmented, and missing BUSCO genes. In the figure, the genomes are ordered from the genome showing the greatest percentage of single copy BUSCO genes (top) to that showing the lowest percentage of single copy BUSCO genes (bottom).

## **Supplementary figure 3. The phylogenomics pipeline used in this study**

For each genome assembly, BUSCO genes were identified resulting in a set of BUSCO genes per genome (93 total sets of BUSCO genes). Amino acid multi-fasta files (mfasta) were created for each BUSCO gene where at least 50% of each taxon was represented resulting in 3,138 BUSCO genes. Mfasta files were then aligned. To obtain nucleotide mfasta files, DNA sequences were threaded onto the protein alignment resulting in an additional set of 3,138 BUSCO mfasta files. Amino acid and nucleotide mfasta files were trimmed and BUSCOs were removed from subsequent analyses if the trimmed alignment was less than 50% the length of the original alignment for either amino acids or nucleotides resulting in 1,773 BUSCOs. Short BUSCO gene alignments defined as ≤ 167 amino acids or ≤ 501 nucleotides, respectively, were removed, resulting in a final set of 1,668 BUSCO gene alignments (BUSCO1668). These 1668 BUSCO gene alignments were used to reconstruct evolutionary relationships using maximum likelihood methods with gene-partitioned models using concatenation, without partitioning using concatenation, or with coalescence.

## **Supplementary figure 4. Construction of five different subsets of the BUSCO1668 phylogenomic matrix**

To explore the stability of the Aspergillaceae phylogeny inferred from analyses of the full BUSCO1668 data matrix, we constructed 5 different subsets of the BUSCO1668 matrix separately for nucleotides (a) and amino acids (b) where each subset contained 50% of the genes found in the BUSCO1668 matrix (834 genes). Each subset was created from genes scoring highest for various metrics of phylogenetic signal including alignment length, average bootstrap support, completeness, treeness / relative composition variability, and the number of variable sites.

## **Supplementary figure 5. Distribution of best-fitting substitution models for nucleotide and amino acid versions of gene sequence alignments of the BUSCO1668 data matrix**

(a) The overwhelming majority of nucleotide gene alignments showed the best fit for a general time reversible model with unequal rates and unequal base frequencies (GTR), empirical base frequencies (F), invariable site proportions (I), and gamma rate heterogeneity across sites with 4 rate categories (G4). (b) The JTT empirical substitution matrix with invariable site proportions (I) and gamma rate heterogeneity across sites with 4 rate categories (G4) was the most common best-fitting model (643 amino acid gene alignments), followed by the LG matrix with invariable (I) and gamma rate heterogeneity across sites with 4 rate categories (G4) (362 genes).

## **Supplementary figure 6. Distribution of internode certainty values among the 36 phylogenies**

Boxplots show the distribution of internode certainty (IC) values for each of the 36 phylogenies. Note that nucleotide-based data matrices show generally higher distributions of IC values than amino acid-based data matrices. Similarly, the distributions of IC values of phylogenies inferred from analyses of subset data matrices are generally higher than those from analyses based on the full data matrices. In contrast, there is little difference in the distributions of IC values by analytical scheme.

## **Supplementary figure 7. Distributions of metrics used in the creation of subsets of the 1,668-gene matrix and their correlations between amino acid and nucleotide-based data sets**

(a, b) For completeness, we observed an average NT and AA taxon occupancy of 97.2 ± 4.6% with a 50th percentile cut-off of 99.0%. (d, e) For gene alignment length among the genes in the set included in the 1,668-gene matrix, we observed an average NT and AA alignment length of 1,896.4 ± 1,370.5 and 631.9 ± 456.0 sites with a 50th percentile cut-off of 1,489.5 and 497.5 sites, respectively. (g, h) For the number of variable sites, we observed an average number of variables sites for NT and AA gene alignments to be 1,446.5 ± 1,103.2 and 450.9 ± 349.5 sites with a 50th percentile cut-off of 1,120.0 and 349.5, respectively. (j, k) For average bootstrap value, we observed an average NT and AA average bootstrap value of 92.4 ± 2.7 and 87.9 ± 5.75 with a 50th percentile cut-off of 92.8 and 89.1, respectively. (m, n) For treeness / RCV, we observed an average NT and AA value of 8.2 ± 2.6 and 6.5 ± 2.8 with a 50th percentile cut-off of 8.2 and 6.0, respectively. Significant correlations were observed across NTs and AAs for each metric (c, f, i, l, and o). Each category is presented in order of the one with the strongest correlation between the NT and AA matrix to the weakest (i.e., completeness = alignment length = variable sites > average bootstrap support > treeness / RCV). For each distribution orange indicates genes that scored in the top 50% for each metric while green indicate genes in the lower 50%. For each correlation, the blue line indicates the line of best fit.

## **Supplementary figure 8. Intersections of genes present in the ten different nucleotide and amino acid subsets of the 1,668-gene matrix**

BUSCO gene intersections were highest between alignment length, average bootstrap support, treeness / RCV, and the number of variable sites for both (a) amino acids (b) and nucleotides. The next largest intersection was a set of genes unique to the matrix based off of completeness followed by a set of genes shared by all subsampled matrices for both (a) amino acids and (b) nucleotides. (c) The number of shared or unique BUSCO genes per subset between amino acid-based and nucleotide-based subsets shows that the number of shared BUSCO genes is very high for alignment length, completeness, and variable sites. In contrast, there are more unique genes for the average bootstrap support and treeness / RCV matrices.

## **Supplementary figure 9. The Aspergillaceae phylogeny shown in Figure 1 with each internode labelled**

To accompany supplementary tables containing specific information about different internodes of the reference Aspergillaceae phylogeny shown in Figure 1, all internodes of this phylogeny have been labelled with unique identifiers from I1 (internode #1) to I78 (internode #78).

## **Supplementary figure 10. Multiple correspondence analysis of matrix features and analytical schemes for each of the eight conflicting internodes**

(a-f) Each plot corresponds to a different internode (shown on the upper left), and each dot in a plot represents one of the 36 phylogenies and is colored according to whether the reference internode or the conflicting internode(s) is supported. Ellipses are drawn around dots that support the same topology if a sufficient number of data points are present. The percent of variance explained on each dimension or principal component is written in parentheses after each dimension label. (b) I74 and I78 reveal distinct clustering among the alternative top and alternative middle and alternative bottom topology (Figure 3b). (e) Similarly, distinct clustering is observed at I33 between trees that support the reference and alternative topologies. (g and h) Contribution of variables to dimension 2 reveals clustering is driving by the matrix sequence type. The red line represents the percent contribution expected by random chance, which is 9.1% or 1 / 11 where 11 is derived from the number of categories (2 sequence types + 6 matrix subsample types + 3 schemes).

## **Supplementary figure 11. Gene-wise log-likelihood scores for each of the eleven incongruent internodes**

(a-k) Each panel depicts the difference in gene-wise log-likelihood scores (ΔGLS) for the reference (blue) and alternative (red) topology for one of the twelve incongruent internodes. When evaluating support for each topology, two depictions are shown: (left) the distribution of support values were plotted and (right) the number of genes that favor either the reference or alternative topology. For each incongruent internode, the number of genes that favor the reference topology is greater than the number of genes that favor the alternative topology with the exception of I33 (b), however the absolute value of the sum of ΔGLS for genes that favor the reference topology (3,881.1) is greater than that of the alternative topology (3,585.0). It is noteworthy that a single gene, *EOG091N0434*, has a relatively high ΔGLS of 140 and favors the reference topology. To determine if this single gene skewed the results, we removed this gene and still found that the absolute value of the sum of ΔGLS for genes that favor the reference topology (3,741.1) was still greater than that of the alternative topology (3,585.0).

## **Supplementary figure 12. Measures of incongruence are largely robust to removal of potential hidden paralogs.**

Internode certainty values were highly correlated between the full 1,668-gene dataset and the 1,331-gene dataset, which was filtered for potential hidden paralogs (*rs* = 0.98, *p* < 0.01).

## **Supplementary figure 13. Distribution of the degree of violation of a molecular clock per gene and the correlation of estimated internode divergence times inferred using three different subsets of genes suggest stable divergence estimates.**

(a) The average degree of violation of a molecular clock (DVMC) value was 0.15 (min: 0.02 and max: 3.14) (File S5) and the 834 genes with the smallest DVMC values had a cut off of 0.09 and are depicted in orange while genes with higher DVMC values are depicted in green.