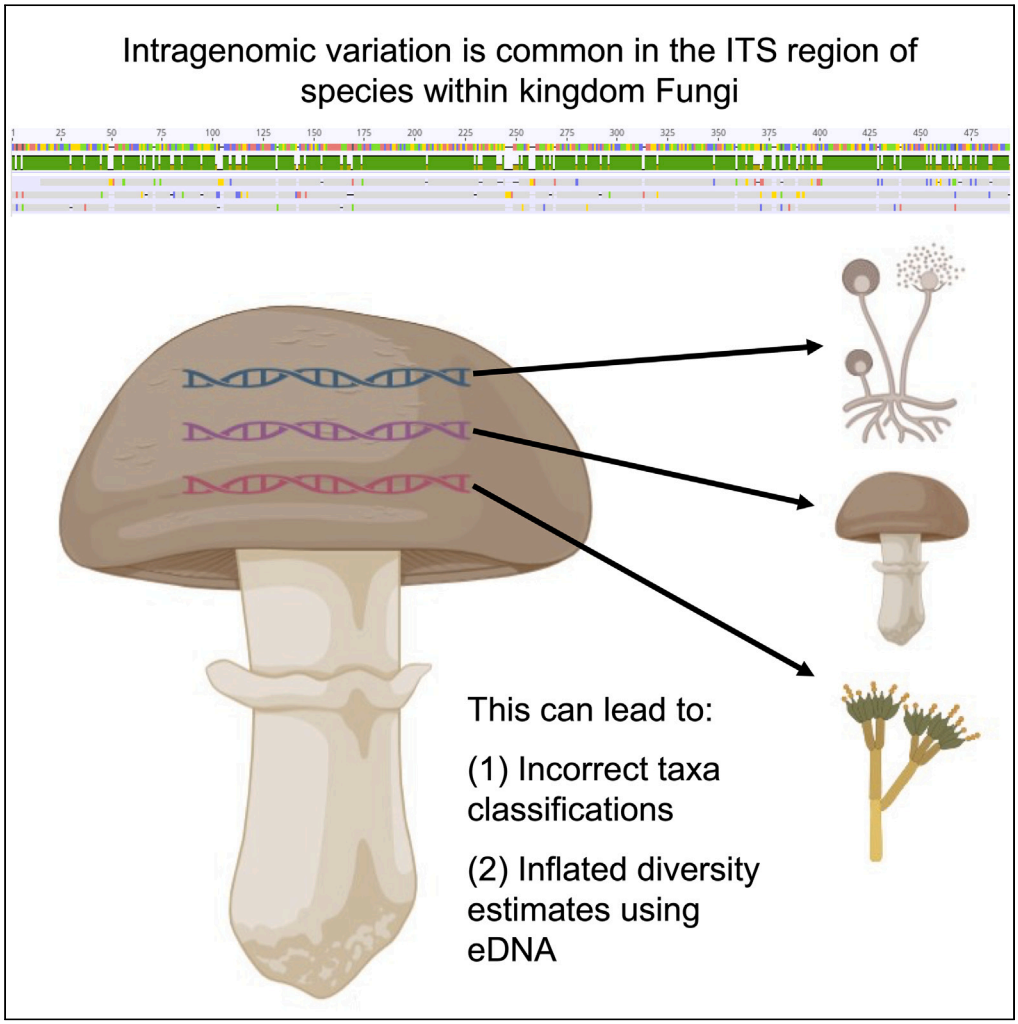
iScience

CellPress
OPEN ACCESS

Article

Extensive intragenomic variation in the internal transcribed spacer region of fungi



Michael J. Bradshaw, M. Catherine Aime, Antonis Rokas, ..., Binod Pandey, Yuanning Li, Donald H. Pfister

mbradshaw@fas.harvard.edu

Highlights

The ITS region is the most prevalent barcode used in studies of kingdom Fungi

There are multiple copies of the ITS region within the fungal genome

ITS intragenomic variation and contaminated genomes are common in fungi

Intragenomic variation can affect taxonomy, eDNA studies, and diversity estimates

Bradshaw et al., iScience 26, 107317
August 18, 2023
https://doi.org/10.1016/j.isci.2023.107317

iScience

CellPress
OPEN ACCESS

Article

Extensive intragenomic variation in the internal transcribed spacer region of fungi

Michael J. Bradshaw,[1,8,*] M. Catherine Aime,[2] Antonis Rokas,[3] Autumn Maust,[4] Swarnalatha Moparthi,[5] Keila Jellings,[2] Alexander M. Pane,[4] Dylan Hendricks,[4] Binod Pandey,[6] Yuanning Li,[7] and Donald H. Pfister[1]

SUMMARY

Fungi are among the most biodiverse organisms in the world. Accurate species identification is imperative for studies on fungal ecology and evolution. The internal transcribed spacer (ITS) rDNA region has been widely accepted as the universal barcode for fungi. However, several recent studies have uncovered intragenomic sequence variation within the ITS in multiple fungal species. Here, we mined the genome of 2414 fungal species to determine the prevalence of intragenomic variation and found that the genomes of 641 species, about one-quarter of the 2414 species examined, contained multiple ITS copies. Of those 641 species, 419 (~65%) contained variation among copies revealing that intragenomic variation is common in fungi. We proceeded to show how these copies could result in the erroneous description of hundreds of fungal species and skew studies evaluating environmental DNA (eDNA) especially when making diversity estimates. Additionally, many genomes were found to be contaminated, especially those of unculturable fungi.

INTRODUCTION

Fungi are one of the largest kingdoms among eukaryotes, containing an estimated 2.2–3.8 million species,[1] but only ~150,000 have been accepted so far. A key step in studying fungi, and for any downstream analyses, is species identification. Due to a dearth of diagnostic morphological characteristics and our inability to isolate or maintain many fungi in pure culture, accurate identification largely relies on the use of molecular markers. The nuclear ribosomal internal transcribed spacer (ITS) region of the rDNA array is the most prevalent marker used to identify fungi and is the primary fungal barcode to the Consortium for the Barcode of Life.[2] The internal transcribed spacer region is also the most frequently applied marker in studies evaluating environmental DNA (eDNA).[3] One of the main appeals of the ITS region is that each organism contains many paralogous copies[4] allowing successful amplification of samples (such as eDNA and unculturable fungi) where quantities of DNA may be scant. Additionally, the unique combination of conserved areas of DNA encompassing highly variable DNA allows the ability to design robust and nearly kingdom-wide PCR primers.[5] Considering that the ITS region was the first barcode commonly used for fungi, initial work using this region has led to many revelations regarding the generic and familiar classification of fungi. However, the multiple copies of ITS[4,6–8] can sometimes vary, raising concerns on its broad usage throughout the kingdom.[9,10]

The multiple copies of the ITS region tend to be in clusters within the genome and are often homogenized by concerted evolution.[11,12] Concerted evolution is not a definite phenomenon and mutations that are not homogenized can result in the formation of imperfect copies of functional genes and other intragenomic variation aberrations within the rDNA.[13–15] Intraspecific variation in the ITS region is widespread within fungi[16] and may cause ambiguous results when analyzing sequence data.[9,10,17,18] In some cases, analyses of sequences from divergent ITS copies can result in the description of species, thus highlighting how use of rDNA alone can mislead taxonomic inference.[9,10,17,19]

Thanks to systematic sequencing efforts,[20] fungi are one of the most densely genome-sequenced kingdoms of eukaryotes. Examination of fungal genomes offers new insight to study intragenomic variation.[10,21] Here, we evaluated a genome assembly for each fungal taxon available on GenBank to determine the prevalence of intragenomic variation in kingdom Fungi, and the consequences of this variation on taxonomic conclusions, eDNA studies, and fungal diversity estimates.

[1]Harvard University Herbaria and Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

[2]Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN 47907, USA

[3]Department of Biological Sciences and Evolutionary Studies Initiative, Vanderbilt University, Nashville, TN 37235, USA

[4]School of Environmental and Forest Sciences, University of Washington, Seattle, WA 98195, USA

[5]Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC 27695-7613, USA

[6]Department of Plant Pathology, North Dakota State University, Fargo, ND 58102, USA

[7]Institute of Marine Science and Technology, Shandong University, 72 Binhai Road, Qingdao 266237, China

[8]Lead contact

*Correspondence: mbradshaw@fas.harvard.edu

https://doi.org/10.1016/j.isci.2023.107317

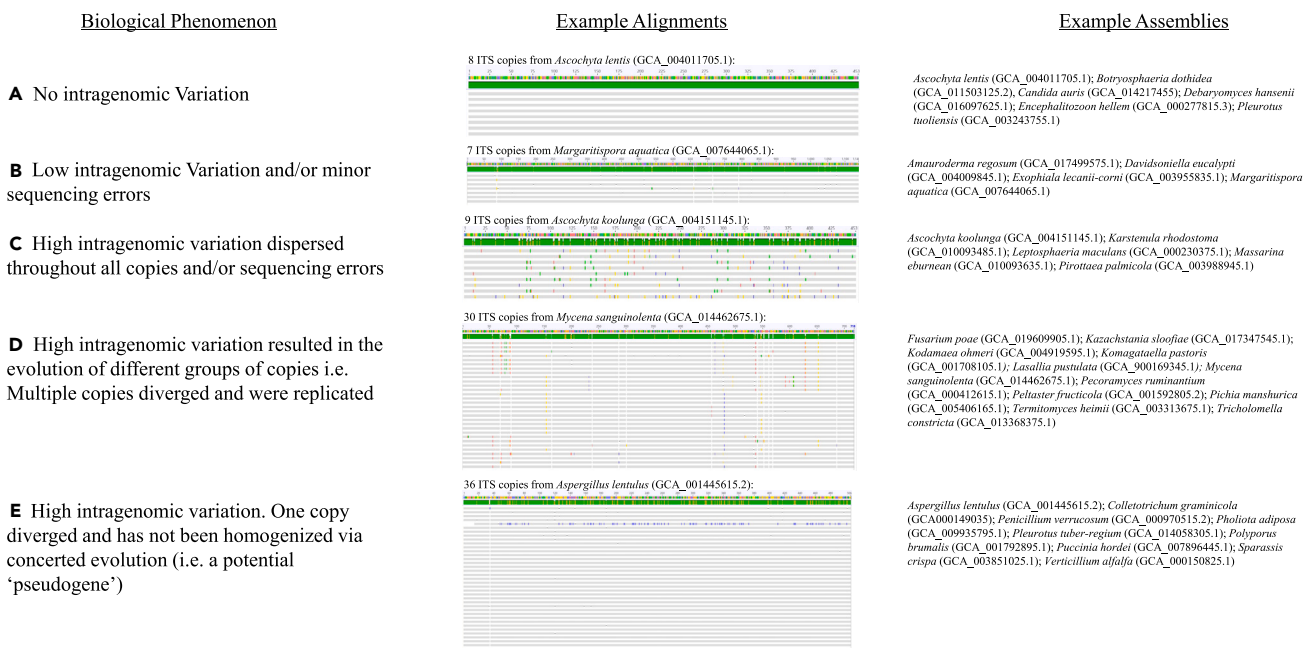


Figure 1. Common alignments observed when evaluating the ITS copies of fungi

Example assemblies are provided for the different commonly observed alignments as well as hypotheses on their biological significance (A–E). Different colors in the alignment represent intragenomic variation.

RESULTS

Data acquisition

In total, genome assemblies from 2414 taxa were mined and evaluated from GenBank (Data S1 and S2). The alignments generated are available at Dryad (https://doi.org/10.5061/dryad.g79cnp5t7). For approximately one-quarter of all the taxa evaluated (695/2414 assemblies), we were unable to locate an ITS sequence (ITS copies = 0). However, most of the taxa evaluated had one ITS copy (1080/2414). The remaining 641 taxa had two or more ITS copies.

Intragenomic variation

641/2414 genome assemblies (~27%) were found to have multiple ITS copies. The ITS copies were aligned, and five different phenomena were commonly observed (Figure 1). Of the 641 assemblies, 222 had 100% identical ITS copies (Figure 1A), 303 had low intragenomic variation (98–99.99% pairwise identity) (Figure 1B), and the remaining 116 assemblies had high variation (<98% pairwise identity) (Figures 1D and 1E). Highly divergent ITS copies, belonging potentially to "pseudogenes", were found in 46 assemblies (Figures 1D and 1E); these copies have less than 93% sequence similarity with the other copies in the alignment. In total, ~17% (419/2414) of the assemblies evaluated contained some level of intragenomic variation (pairwise identity <100%). The 17% intragenomic variation observed in the present study when analyzing all the assemblies is likely an underestimate as assemblies containing 0 or 1 ITS copy were coded as having no variation. In our analysis, when only multi-copy assemblies (ITS copies >1) were evaluated, ~65% of the assemblies contained intragenomic variation. The variation observed can affect phylogenies differently. For example, random point mutations that may arise from sequencing errors or biological variation (Figures 1B and 1C) usually will have no effect in multiple alignment-based analyses. However, pseudogenes (Figure 1E) will usually form a unique clade.

Although we cannot exclude sequencing errors, the high coverage and low sequencing error rate for most of the sequencing technologies used for the assemblies evaluated[22–24] suggest that the variation we observed among ITS copies was genuine. However, the sequencing technology, assembly methods, and technology read size can influence the data as discussed by Tedersoo et al.[25] and Appendix D of Paoli et al.[10] For example, assemblies constructed using long-read technology (PacBio, Oxford, etc.) tended to have largerer genomes (~46 mbps [million base pairs]), higher GC percentages (48.11), more

ITS copies (14.52) and less intragenomic variation (pairwise identity = 99.11) compared to short-read technology (Illumina, sanger, etc.) (genome size = ~37 mbps; GC percentage = 47.3; ITS copies = 1.49; pairwise identity = 96.61) (Data S2). The most accurate data undoubtedly occurs when both long and short-read technologies are used in tandem (genome = ~44 mbps; GC percentage = 46.01; ITS copies = 9.43; pairwise identity = 98.99) (Data S2).

Contamination

Multiple contaminated assemblies from the 2414 genomes were identified (Data S3). The contaminants were mostly identified from fragmented ITS copies that were GenBank nblasted and found to align with accessions of a different taxon than the assembly was designated. After noting that many of the contaminated genomes were from obligate parasitic fungi that cannot be cultured, we proceeded to evaluate intensely the assemblies of two commonly studied unculturable pathogens (Erysiphaceae, powdery mildews, and Pucciniales, rust fungi) to determine how common contamination is in unculturable full genome assemblies. In total, 12/35 of the taxa evaluated from Erysiphaceae and Pucciniales were contaminated. Another observation is that assemblies may be derived from multiple strains of the same species. For example, *Nosema bombycis* (GCA 000383075.1) has 9 ITS copies with a 95% pairwise identity among the copies. However, when GenBank blasted, the different copies aligned with different strains. A similar situation was found in *Suillus spraguei* (GCA016800925.1) where 4 copies aligned 100% with the ITS of *Suillus spraguei* voucher ACAD21063F (OL741513), whereas two other copies aligned 100% with *Suillus spraguei* strain EM44 (OL685247). Another possibility is that the authors of the accessions on GenBank sequenced the different ITS copies in the genome and reported these different "copies" as "strains." This phenomenon was also observed and discussed by Stadler et al.[15] who discovered 5 deviating ITS copies (one of which was a pseudogene) of *Hypomontagnella monticulosa*.

DISCUSSION

Our study reveals that fungi exhibit intragenomic variation in the form of nucleotide substitutions, deletions/insertions, and likely "pseudogenes" which we show can be impacting taxonomic, eDNA, and diversity estimate studies. The high intragenomic variation in the ITS rDNA observed in the present study can be attributed to the divergence of multiple copies that have not been homogenized through concerted evolution or similar forces. Evaluating full genome assemblies can give an accurate representation of intragenomic variation[10,15]; as such the ITS intragenomic variation reported here is likely indictive of the true intragenomic variation in kingdom Fungi. In our analysis, when only assemblies with ITS copies >1 were evaluated, ~65% of the assemblies contained intragenomic variation. However, when only high-quality assemblies using both long- and short-read technology are considered, variation is still observed in a high percentage of the taxa evaluated (49.7% [90/181]), and in a similar proportion to that reported for all multicopy assemblies (~65% variation in assemblies containing >1 copy). The 49.7% is likely more indictive of the true proportion of ITS intragenomic variation of taxa in kingdom Fungi. Lindner et al.[16] estimated that rDNA intragenomic variation was widespread, yet rare in fungi, with polymorphisms to exist in ~3–5% of taxa. This is a considerable underestimate from the data shown in the current study. The intragenomic variation data presented are consistent with the data from Paloi et al.[10] and Stadler et al.[15] It should be noted that there were differences in intragenomic variation between all the taxonomic groupings evaluated (Data S2). Organisms within the earlier diverging taxa (Chytridiomycota, Mucoromycota, and Zoopagomycota) tended to have the most intragenomic variation.

Fungi were estimated to have between 14 and 1442 ITS copies in their genomes based on an in silico read depth approach.[4] Although mining assemblies can give an accurate representation of intragenomic variation, it does not give reliable data of ITS copy numbers.[10] The effect of the different sequencing technologies on ITS copy number is discussed thoroughly by Tedersoo et al.[25] and in Appendix D of Paoli et al.[10] Briefly, the impact of assembly programs on copy number can, in part, be due to the "stacking" function of the assembly pipelines. For example, similar data can be stacked on top of each other (under certain identity thresholds), essentially masking copy numbers. We hypothesize that this "stacking function" is likely the cause for the massive amount of single-copy ITS regions observed throughout the dataset, which is why they were not included in our main intragenomic variation analyses. When only assemblies with ITS copies >1 were evaluated, the number of copies ranged from 2 to 528 with an average of 10.9. However, when only long-read technology was considered, the average ITS copy number was 14.52, which is likely the most accurate calculation of copy number. Even so, this is a vast underestimate of the average of 133 copies presented by Lofgren et al.[4]

The evolution of extremely divergent ITS copies (i.e., likely "pseudogenes") is common throughout fungi (Figure 1). In our evaluation of the current data, we found instances in which it is possible that species were described based on a divergent ITS copy. For example, *Candida viswanathii* (GCA003327735) has 5 ITS copies that align 100% with *Candida viswanathii* accessions from GenBank and 4 copies that align 99.7% with the type material of *Candida pseudoviswanathii*. Unsurprisingly, Ren et al.[26] described *C. pseudoviswanathii* based primarily on ITS sequences. Taxonomic conclusions regarding these species should not be made until the type material of *C. pseudoviswanthii* and *C. viswanathii* can be further evaluated with additional genetic markers. Similar examples in yeasts with divergent ITS copies are discussed in Sipiczki[27] and Sipiczki.[28] Other examples of accessions/species that need to be evaluated further include *Mucor circinatus* (GCA_016758965.1), the ITS copies of which align with different *Mucor* spp. i.e., *Mucor plumbeus* and *M. mucedo*, and *Taphrina wiesneri* (GCA_005281515.1), the copies of which align with type material of *T. wisneri* and type material of *Taphrina confusa*. Similar cases likely exist in the dataset and we encourage researchers to further mine the data to locate doubtful species that were potentially described based on divergent ITS copies. Further critical analyses should also be conducted to see if any of these circumstances are examples of hybridization between different fungi. Only 2414 of the ~150,000 taxa accepted from kingdom fungi were analyzed here and as such intragenomic variation could have led to the description of hundreds of erroneous species. Interestingly, the "pseudogenes" analyzed from the present study often contained multiple mismatches with the common fungal primers (ITS1, ITS4, and/or ITS5) from White et al.,[5] and are often in lower proportion to the other ITS types (Figure 1E).[16] This likely explains why they are not commonly amplified in PCR. Additionally, when other primers are used, or non-specific binding occurs, it could lead to new species being described based on a divergent ITS copy as was reported by Harrington et al.[17] A similar phenomenon in *Fusarium* was noted where two highly divergent ITS 2 "types" were observed, of which, only the "major ITS2 type" was able to be sequenced with conserved primers. When the authors developed specific primers, they were able to anneal and amplify the other ITS type.[29] Different primers can anneal to different ITS copies and, as such, the primers used could be artificially skewing the phylogenetic relationships among certain fungal lineages.

Caution should be taken when describing species based predominantly on differences in the ITS region without corresponding secondary barcodes and/or morphological, ecological, and chemical data.[30] The effect of ITS copies likely has a large impact on eDNA studies that rely on ITS data. Kõljalg et al.[3] proposed the term "species hypothesis," for taxa discovered through ITS analyses that grouped together in different similarity thresholds ranging from 97 to 99%. The UNITE platform[31] variously delimited species hypothesis at 97–100% similarity based on intraspecific ITS variability. Additionally, the GlobalFungi database[32] classified ITS sequences according to the closest UNITE species hypothesis and a 98.5% similarity threshold. Evaluating our 641 assemblies that contained >1 ITS copy, at a 97% "species hypothesis" threshold, we could describe an additional 15% (93/641) species, at a 98% threshold, an additional 18% (116/641) species, and at a 99% threshold, an additional 27% (171/641) species. Similarly, Lindner and Banik[19] showed how the use of a 95% threshold for *Laetiporus* species descriptions based on the ITS region could artificially result in over twice the number of described species due to intragenomic variation of ITS copies. It is likely that the use of amplicon sequence variants evaluating the ITS region in eDNA studies is also being impacted by intragenomic variation. Lücking et al.[33] found that sequencing errors in some full genome technology can contribute to increased biological diversity estimates. As such, using eDNA data from the ITS region to establish diversity estimates[34] could be vastly overestimating the number of fungi.

During this study, a high number of assemblies were determined to be contaminated with non-target fungi (Data S3). This was especially the case with unculturable fungi (at least 82% of the taxa in the Erysiphaceae were found to contain contaminants and at least 12.5% of the taxa in the Pucciniales were contaminated). Similar to the current study, Vaghefi et al.[35] noticed the high contamination among Erysiphaceae genomes and recommended that assemblies of taxa within the order be treated as "eDNA," recognizing that the sequences from leaf tissue and surfaces were likely to include other organisms in the environment. Future research evaluating full genome phylogenies should consider removing unculturable fungi from their analyses or methodically checking them for contamination by blasting the assemblies with common contaminate DNA from multiple regions (blasting solely the ITS region is not sufficient as many assemblies do not contain an ITS region). Removing contaminated assemblies from datasets will undoubtedly improve phylogenetic inference.

The issue of misidentifications, contaminations, and assemblies that contain multiple strains cannot be discounted and is likely a common occurrence within fungi (even in pure cultures).[10,35,36] We hypothesize that the high amount of DNA required for full genome sequences may lead to these intra/inter species contamination issues through the accumulation of multiple individuals for processing. It is possible that sequencing multiple strains/taxa could be impacting other molecular statistics such as genome size. For example, some obligate, unculturable fungi have been reported to represent the largest, repeat rich, genomes.[37] In our dataset, a potential example of this phenomenon can be observed with the assembly from *Austropuccinia psidii* (GCA 902702905.1). This is the third-largest genome in our dataset (Data S1). It contains 10 ITS copies; one copy is likely a pseudogene and the remaining 9 fall into two genotypes. One genotype aligns 100% with *Puccinia psidii* isolate UY217 (EU348742) and the other aligns 100% with *Puccinia psidii* isolate SZ2 (EU071045). In this scenario, it is possible that the genome contains multiple strains that artificially increase the genome size and repetitive regions. Alternatively, in the case of this rust, we could be observing two parental genotypes.

The ITS region is widely accepted as the universal barcode for fungi.[2] Our analyses show that ITS intragenomic variation is common throughout kingdom Fungi (Data S1), a finding with wide implications for taxonomic assignments, eDNA analyses, and fungal diversity estimates. Future research evaluating the ITS region should consider the data generated (available at Dryad: https://doi.org/10.5061/dryad.g79cnp5t7) to ascertain whether intragenomic variation could be skewing research results. Internal transcribed spacer region data have been analyzed for fungi for over 30 years and these data should not be set aside, however, taxonomic conclusions using ITS data should be accompanied by secondary barcodes and/or morphological and ecological data.[38,39] Additionally, DNA-based typifications[30,40] should not be done solely with ITS. Future research should further deduce the role that intragenomic variation can play on eDNA studies, especially regarding diversity estimates. Other single-copy markers should be evaluated and compared to ITS data in eDNA studies to ascertain the effect of intragenomic variation. Additionally, the data presented could be mined further to answer a range of molecular biology questions including the substitution rate and most common intragenomic mutations occurring in the ITS rDNA region. Sequencing genomes is becoming easier and more affordable. We recommend future taxonomic research to consider taking a taxogenomic approach and eDNA studies to use other single-copy markers to circumvent the issues presented here.

Limitations of the study

A major limitation of the study is the effect of sequencing technology on the results, especially in regard to sequencing errors. Additionally mining genomes does not give reliable data of ITS copy numbers. As such, we were unable to determine the proportion of each of the different variant copies within a genome. Having said that, we believe the intragenomic variation data are genuine and a detailed discussion of the potential impact of the sequencing technology can be found in the results and discussion sections. Additionally, considering that the present study was accomplished bioinformatically, none of the results were verified in the lab.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107317.

ACKNOWLEDGMENTS

We would like to thank the Harvard University Herbaria Research Fellowship for funding this research.

AUTHOR CONTRIBUTIONS

M.J.B. designed the study, brought together the team of collaborators, helped mine the assemblies, organized the data, and wrote the first draft. M.C.A. helped design the study and bring together the team of collaborators. A.M., S.M., K.J., A.M.P., D.H., and B.P. helped mine the assemblies. A.R. helped design the study and formulate the manuscript. Y.L. contributed to data analysis. D.H.P. helped fund and organize the project. M.C.A., A.R., and D.H.P. considerably edited the first draft.

DECLARATION OF INTERESTS

The authors declare no conflict of interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: May 15, 2023
Revised: June 8, 2023
Accepted: July 4, 2023
Published: July 10, 2023

REFERENCES

1. Hawksworth, D.L., and Lücking, R. (2017). Fungal Diversity Revisited: 2.2 to 3.8 Million Species. Microbiol. Spectr. *5*, 4.

2. Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., and Chen, W.; Fungal Barcoding Consortium; Fungal Barcoding Consortium Author List (2012). Fungal Barcoding Consortium, nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. Proc. Natl. Acad. Sci. USA *109*, 6241–6246.

3. Köljalg, U., Nilsson, R.H., Abarenkov, K., Tedersoo, L., Taylor, A.F.S., Bahram, M., Bates, S.T., Bruns, T.D., Bengtsson-Palme, J., Callaghan, T.M., et al. (2013). Towards a unified paradigm for sequence-based identification of fungi. Mol. Ecol. *22*, 5271–5277.

4. Lofgren, L.A., Uehling, J.K., Branco, S., Bruns, T.D., Martin, F., and Kennedy, P.G. (2019). Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles. Mol. Ecol. *28*, 721–730.

5. White, T.J., Bruns, T., Lee, S., and Taylor, J. (1990). Amplification and Direct Sequencing of Fungal Ribosomal RNA Genes for Phylogenetics in PCR Protocols: A Guide to Methods and Applications (Academic Press), pp. 315–332.

6. Vilgalys, D., and Gonzalez, D. (1990). Organization of ribosomal DNA in the basidiomycete Thanatephorus praticola. Curr. Genet. *18*, 277–280.

7. Vydryakova, G.A., Van, D.T., Shoukouhi, P., Psurtseva, N.V., and Bissett, J. (2012). Intergenomic and intragenomic ITS sequence heterogeneity in *Neonothopanus nambi* (Agaricales) from Vietnam. Mycology *3*, 89–99.

8. Harder, C.B., Læssøe, T., Frøslev, T.G., Ekelund, F., Rosendahl, S., and Kjøller, R. (2013). A three-gene phylogeny of the *Mycena pura* complex reveals 11 phylogenetic species and shows ITS to be unreliable for species identification. Fungal Biol. *117*, 764–775.

9. Kiss, L. (2012). Limits of nuclear ribosomal DNA internal transcribed spacer (ITS) sequences as species barcodes for fungi. Proc. Natl. Acad. Sci. USA *109*, E1811.

10. Paloi, S., Luangsa-ard, J.J., Mhuantong, W., Stadler, M., and Kobmoo, N. (2022). Intragenomic variation in nuclear ribosomal markers and its implication in species delimitation, identification and barcoding in fungi. Fungal Biol. Rev. *42*, 1–33.

11. Hillis, D.M., and Dixon, M.T. (1991). Ribosomal DNA: Molecular evolution and phylogenetic inference. Q. Rev. Biol. *66*, 411–453.

12. Elder, J.F., and Turner, B.J. (1995). Concerted evolution of repetitive DNA sequences in eukaryotes. Q. Rev. Biol. *70*, 297–320.

13. Rooney, A.P., and Ward, T.J. (2005). Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. Proc. Natl. Acad. Sci. USA *102*, 5084–5089.

14. Sipiczki, M., Horvath, E., and Pfliegler, W.P. (2018). Birth-and-death evolution and reticulation of ITS segments of *Metschnikowia andauensis* and *Metschnikowia fructicola* rDNA repeats. Front. Microbiol. *9*, 1193.

15. Stadler, M., Lambert, C., Wibberg, D., Kalinowski, J., Cox, R.J., Kolařík, M., and Kuhnert, E. (2020). Intragenomic polymorphisms in the ITS region of high-quality genomes of the Hypoxylaceae (Xylariales, Ascomycota). Mycol. Prog. *19*, 235–245.

16. Lindner, D.L., Carlsen, T., Henrik Nilsson, R., Davey, M., Schumacher, T., and Kauserud, H. (2013). Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi. Ecol. Evol. *3*, 1751–1764.

17. Harrington, T.C., Kazmi, M.R., Al-Sadi, A.M., and Ismail, S.I. (2014). Intraspecific and intragenomic variability of ITS rDNA sequences reveals taxonomic problems in *Ceratocystis fimbriata* sensu stricto. Mycologia *106*, 224–242.

18. Estensmo, E.L.F., Maurice, S., Morgado, L., Martin-Sanchez, P.M., Skrede, I., and Kauserud, H. (2021). The influence of intraspecific sequence variation during DNA metabarcoding: A case study of eleven fungal species. Mol. Ecol. Resour. *21*, 1141–1148.

19. Lindner, D.L., and Banik, M.T. (2011). Intragenomic variation in the ITS rDNA region obscures phylogenetic relationships and inflates estimates of operational taxonomic units in genus *Laetiporus*. Mycologia *103*, 731–740.

20. Grigoriev, I.V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., et al. (2014). MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Res. *42*, D699–D704.

21. Wibberg, D., Stadler, M., Lambert, C., Bunk, B., Spröer, C., Rückert, C., Kalinowski, J., Cox,

R.J., and Kuhnert, E. (2021). High quality genome sequences of thirteen Hypoxylaceae (Ascomycota) strengthen the phylogenetic family backbone and enable the discovery of new taxa. Fungal Divers. *106*, 7–28.

22. Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. Mol. Ecol. Resour. *11*, 759–769.

23. Chen, Y., Nie, F., Xie, S.-Q., Zheng, Y.-F., Dai, Q., Bray, T., Wang, Y.-X., Xing, J.-F., Huang, Z.-J., Wang, D.-P., et al. (2021). Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat. Commun. *12*, 60.

24. Stoler, N., and Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. NAR Genom. Bioinform. *3*, lqab019.

25. Tedersoo, L., Drenkhan, R., Anslan, S., Morales-Rodriguez, C., and Cleary, M. (2019). High-throughput identification and diagnostics of pathogens and pests: Overview and practical recommendations. Mol. Ecol. Resour. *19*, 47–76.

26. Ren, Y.C., Xu, L.L., Zhang, L., and Hui, F.L. (2015). *Candida baotianmanensis* sp. nov. and *Candida pseudoviswanathii* sp. nov., two ascosporic yeast species isolated from the gut of beetles. Int. J. Syst. Evol. Microbiol. *65*, 3580–3585.

27. Sipiczki, M. (2020). *Metschnikowia pulcherrima* and related pulcherrimin-producing yeasts: fuzzy species boundaries and complex antimicrobial antagonism. Microorganisms *8*, 1029.

28. Sipiczki, M. (2022). Taxonomic revision of the pulcherrima clade of *Metschnikowia* (fungi): Merger of Species. Taxonomy *2*, 107–123.

29. O'Donnell, K., and Cigelnik, E. (1997). Two divergent Intragenomic rDNA ITS2 types within a monophyletic lineage of the fungus *Fusarium* are nonorthologous. Mol. Phylogenet. Evol. *7*, 103–116.

30. Aime, M.C., Miller, A.N., Aoki, T., Bensch, K., Cai, L., Crous, P.W., Hawksworth, D.L., Hyde, K.D., Kirk, P.M., Lücking, R., et al. (2021). How to publish a new fungal species, or name, version 3.0. IMA Fungus *12*, 11. https://doi.org/10.1186/s43008-021-00063-1.

31. Nilsson, R.H., Larsson, K.-H., Taylor, A.F.S., Bengtsson-Palme, J., Jeppesen, T.S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F.O., Tedersoo, L., et al. (2019). The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Res. *47*, 259–264.

32. Vetrovsky, T., Morais, D., Kohout, P., Lepinay, C., Algora, C., Holla, S.A., Bahnmann, B.D., Bilohneda, K., Brabcova, V., D'Alo, F., et al. (2020). GlobalFungi, a global database of fungal occurrences from high-throughput-sequencing metabarcoding studies. Sci. Data *7*, 228.

33. Lücking, R., Lawrey, J.D., Gillevet, P.M., Sikaroodi, M., Dal-Forno, M., and Berger, S.A. (2014). Multiple ITS haplotypes in the genome of the lichenized basidiomycete cora inversa (Hygrophoraceae): fact or artifact? J. Mol. Evol. *78*, 148–162. https://doi.org/10.1007/s00239-013-9603-y.

34. Baldrian, P., Větrovský, T., Lepinay, C., and Kohout, P. (2021). High-throughput sequencing view on the magnitude of global fungal diversity. Fungal Divers. *114*, 539–547.

35. Vaghefi, N., Kusch, S., Németh, M.Z., Seress, D., Braun, U., Takamatsu, S., Panstruga, R., and Kiss, L. (2022). Beyond nuclear ribosomal DNA sequences: Evolution, taxonomy, and closest known saprobic relatives of powdery mildew fungi (Erysiphaceae) inferred from their first comprehensive genome-scale phylogenetic analyses. Front. Microbiol. *13*, 903024.

36. Houbraken, J., Visagie, C.M., and Frisvad, J.C. (2021). Recommendations to prevent taxonomic misidentification of genome-sequenced fungal strains. Microbiol. Resour. Announc. *10*, e0107420.

37. Tavares, S., Ramos, A.P., Pires, A.S., Azinheira, H.G., Caldeirinha, P., Link, T., Abranches, R., Silva, M.d.C., Voegele, R.T., Loureiro, J., and Talhinhas, P. (2014). Genome size analyses of Pucciniales reveal the largest fungal genomes. Front. Plant Sci. *5*, 422.

38. Bradshaw, M., Guan, G.X., Nokes, L., Braun, U., Liu, S.Y., and Pfister, D. (2022). Secondary DNA barcodes (CAM, GAPDH, GS, and RPB2) to characterize species complexes and strengthen the powdery mildew phylogeny. Front. Ecol. and Evol. *10*, 918908. https://doi.org/10.3389/fevo.2022.918908.

39. Bradshaw, M., Braun, U., and Pfister, D. (2022). Phylogeny and taxonomy of the genera of the Erysiphaceae, part 1, *Golovinomyces*. Mycologia *114*, 964–993.

40. Nilsson, R.H., Ryberg, M., Wurzbacher, C., Tedersoo, L., Anslan, S., Põlme, S., Spirin, V., Mikryukov, V., Svantesson, S., Hartmann, M., et al. (2023). How, not if, is the question mycologists should be asking about DNA-based typification. MycoKeys *96*, 143–157.

41. Sayers, E.W., Cavanaugh, M., Clark, K., Pruitt, K.D., Schoch, C.L., Sherry, S., and Karsch-Mizrachi, I. (2022). Nucleic Acids Res. *50*, 161–164.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
ITS genome alignments	Dryad	https://doi.org/10.5061/dryad.g79cnp5t7
Raw Data	This paper	Data S1, S2, and S3
Software and algorithms		
GenBank	Sayers et al.[41]	GenBank Overview (nih.gov)
Geneious version 2021.2.2	Geneious	https://www.geneious.com
R (v. 3.31)	R Foundation for Statistical Computing	

RESOURCE AVAILABILITY

Lead contact

Further questions should be directed to Dr. Michael Bradshaw (mbradshaw@fas.harvard.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The data is available in Data S1, S2, and S3 as well as through Dryad. DOIs are listed in the key resources table.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This work has not involved the use of human subjects or samples, nor has it used experimental models that require reporting of experimental model and subject details.

METHOD DETAILS

Data were mined from at least one genome assembly of every fungal species from September-December of 2021 on GenBank.[41] Data mining was accomplished by extracting the multiple ITS copies from a given assembly and then aligning and analyzing the extracted copies for variation. Detailed methods are as follows.

(1) A list of all taxa with publicly available assemblies was compiled.

(2) For each taxon, GenBank's nucleotide database was searched for a fully annotated ITS region.

(3) The GenBank accession number determined from (2) was GenBank blasted (blastn) to ensure the taxon was identified correctly.

(4) The ITS region from (2) was trimmed to include only nucleotides present in the ITS1+5.8S+ITS2 region.

(5) A genome assembly was chosen for each fungal species on GenBank. If multiple assemblies for a given taxon were available, the assembly with the smallest number of scaffolds/contigs was evaluated first.

(6) A genome assembly was GenBank blasted (blastn) with the trimmed ITS region. For example, in Data S1, column A ('Assembly Reference') was GenBank blasted with column E ('GenBank Accession Number of ITS Region used to blast assembly'); if no ITS region was located or if it was very fragmented other assemblies were checked.

(7) The results of the assembly blast were downloaded into Geneious version 2021.2.2 and aligned.

(8) ITS copies from the genome assembly that were ~ >50 bases shorter than the length of the ITS region determined in step (4) were discarded to eliminate short contigs and to keep the data consistent.

(9) Alignments for these taxa are available on Dryad (https://doi.org/10.5061/dryad.g79cnp5t7) in both a .geneious and .fasta file format.

(10) The number of ITS copies in the assembly, identical site % and pairwise identity % among the different copies were calculated in Geneious and recorded.

(11) The ITS accessions used to blast the assemblies were downloaded into Geneious and their GC content was recorded.

(12) The remaining data from the assemblies were recorded from GenBank (Taxa ID, assembly method, sequencing technology used, genome coverage, contigs, scaffolds, assembly GC content (%), assembly release date, and genome size).

QUANTIFICATION AND STATISTICAL ANALYSIS

All data analyses were conducted in the software R v. 3.31.