DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Ecological Population Genomics in the Emerging Amanita System

Citation
Elmore, Holly. 2020. Ecological Population Genomics in the Emerging Amanita System. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Permanent link
https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37365849

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

Ecological population genomics in the emerging *Amanita* system

A dissertation presented

by

Holly Elmore

to

The Department of Organismic & Evolutionary Biology

In partial fulfilment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Organismic & Evolutionary Biology

Harvard University
Cambridge, Massachusetts

April 6, 2020

© 2020 Holly Elmore
All rights reserved.

Dissertation Advisor: Professor David Haig Author: Holly Elmore

Ecological population genomics in the emerging *Amanita* system

Abstract

The genus *Amanita* (Agaricomycetes) is an emerging non-model system for ecological population genomics. *Amanita* is a charismatic genus of beautiful and sometimes deadly poisonous mushrooms. Several *Amanitae* are invasive species, including the ectomycorrhizal *A. phalloides* and saprotrophic *A. thiersii.* Ecological population genomics combines population genomics, the study of differences within and between populations using genomic data, with ecological perspectives on the contexts of populations and natural history of each specimen.

In Chapter 1, I develop *Amanita*BASE as a resource and foundation for ecological population genomics in the *Amanita* system. *AmanitaBASE* consists of hundreds of physical specimens of *Amanita*, mushrooms and cultures; associated sequences, including 93 whole genomes, 52 of which are from the same two populations over three timepoints: 2004, 2014, and 2015; metadata about specimens including GPS coordinates, dates of collection, exact positions of collected specimens, photos of specimens, and descriptions of the surrounding area; and protocols and best practices developed alongside and as part of *AmanitaBASE*. The data in *Amanita*BASE serves as the basis for Chapter 2 and parts of Chapter 3.

In Chapter 2, I survey a natural population of highly variable mating compatibility genes, *HD1* and *HD2,* by sequencing directly from specimens collected from the field. Compatible mates must have different *HD1* and *HD2* alleles, therefore diversity at *HD1* and *HD2* has a large effect on the mating dynamics of the population. Yet, because of difficulties with sequencing the region, it has been difficult to study *HD1* and *HD2* directly from natural populations. Without population-level data, it has been difficult to determine the source of the high multiallelism typically found in *HD1* and *HD2* across Basidiomycetes. This study pioneers the use of next generation sequencing to study the HD locus in its natural context and methods for obtaining the sequences of *HD1* and *HD2*. Methods for phasing haplotypes are discussed. I conclude that the diversity of *HD1* and *HD2* alleles is ancient and maintained by balancing selection rather than continuously generated by ongoing negative frequency-dependent selection on new variation.

Chapter 3: Contributions to the *Amanita* system is a collection of smaller projects and work to which I contributed. **Section I** suggests a new lens for viewing “individuality” in filamentous fungi, in which the mycelium is less the individual in itself and more a shared resource for its constituent nuclei. The degree of conflict or integration in the interests of the nuclei determines how well-integrated of an individual a mycelium. **Section II** describes a population genetic analysis of the *Amanita*BASE data to determine the size of genets in *A. phalloides.* **Section III** describes my investigation into the apparent loss of the *HD1* mating gene in *A. thiersii* and its implications for *A. thiersii*’s rapid invasion of the Eastern US and low genetic diversity across its range. **Section IV** details the sequencing and assembly of *A. phalloides* genome Dr4M1 in advance of the massively parallel *Amanita*BASE project and the early exploration of MSDIN toxins in the *A. phalloides* genome. Relevant history of the *Amanita* system is included throughout.

Together, these chapters explore the ecological population genomics of *Amanita* in breadth and depth, with a focus on discovery science in the development of resources in the *Amanita* system.

Dedicated to

My parents, Jenifer and Charles Elmore, for their love, support, and inspiration throughout graduate school and my entire life.

And my grandfather, Albert Elmore, for his unwavering belief in me.

“The feeling of awed wonder that science can give us is one of the highest experiences of which the human psyche is capable. It is a deep aesthetic passion to rank with the finest that music and poetry can deliver. It is truly one of the things that make life worth living.”

–Richard Dawkins*, Unweaving the Rainbow*

Acknowledgments

First, I must thank my Committee members. My advisor, David Haig, has been the highlight of my time at Harvard, and I thank him for sharing his brilliance and gentle guidance. Dan Hartl was of particular help during rough times in the writing of this dissertation, and I will never forget how he gave me a desk when I found myself without a place to sit. Jim Mallet welcomed me into his lab group and his dedication to teaching inspired me when I served as his Teaching Fellow. David Hibbett is the most impressive mycologist I know, and I marvel that he has been able to give me so much quality attention over the years despite the fact that he’s located at another university. And Anne Pringle I have to thank for my entire career at Harvard, the *Amanita* system, and for never abandoning me even as she moved to another institution.

I must also thank the Department of Organismic & Evolutionary Biology for offering me generous financial support as well as a nurturing community. Elena Kramer was a lifeline. In my brief time as her student, Kirsten Bomblies taught me much, and I thank her. I thank all my labmates past and present, especially Jacky Hess, Anne Kakouridis, Cat Adams, Jacob Golan, Denny Wang, Sam Harrow, and Jennifer Kotler. FAS Informatics not only provided the computational resources my work required, but provided excellent support and teaching. Brian Arnold, Allison Shultz, Michele Clamp, Adam Freedman, and Aaron Kitzmiller were each crucial to this research. Richard Rabideau Childers was not only an excellent second opinion and troubleshooter, but a top-notch friend. My HGWISE mentor, Neena Haider, and the rest of my mentoring group provided valuable perspective and support from outside my department.

I owe my undergraduate PI, Antonis Rokas, and the entire Rokas Lab at Vanderbilt, especially John Gibbons and Kris McGary, a great debt for all of the hands-on education and independent learning I was allowed in the lab. My now parents-in-law, Carter and Laurie Todd, supported my research over the summers by allowing me to stay with them.

Dora Farkas has my gratitude, for without her coaching this dissertation would not be where it is today. Harvard University Health Services and Counseling and Mental Health Services provided excellent care.

And, finally, I thank my husband, Hudson Todd, for his love, and my family for their stalwart moral support. I am proud of my accomplishment with this dissertation, but more than getting a PhD, I value the way they believed in me, cheered me on, and even when I couldn’t they helped me to see that I could achieve my dream of becoming “Dr. Elmore.”

Table of contents

Chapter 1
AmanitaBASE: a resource and model of ecological population genomic databases

Co-authors: Catharine Adams, Jacob Golan, Samantha Harrow, Jaqueline Hess, Natalia Vargas-Estupiñan, Susana C. Gonçalves, Anne Pringle

Abstract

This paper introduces *Amanita*BASE, the Pringle Lab's integrated collection of physical *Amanita* specimens, nucleic acids, sequences, variants, ecological data, and specimen metadata, both as a resource and as a model for building ecological population genomic databases. Included here are standard procedures for collecting, sequencing, and bioinformatic data processing in an ecologically-informed way as well as protocols that were newly developed for *Amanita*BASE. Databasing principles can be used to assist in collections and genomics at any level of complexity— indeed, the metadata portion of *Amanita*BASE passed through several incarnations as a relational database before settling on a version-controlled spreadsheet.

Introduction

AmanitaBASE is an ecological population genomic database consisting of hundreds of physical specimens of *Amanita*, mushrooms and cultures; associated sequences, including 93 whole genomes, 52 of which are from the same two populations over three timepoints: 2004, 2014, and 2015; metadata about specimens including GPS coordinates, dates of collection, exact positions of collected specimens, photos of specimens, and descriptions of the surrounding area including ground cover and potential host trees; and protocols and best practices developed alongside and as part of *AmanitaBASE*. *AmanitaBASE* is designed to preserve a piece of the field now so that many questions, anticipated and unanticipated, can be answered in the future.

Why *Amanita*? The Pringle Lab has studied ecological questions in *Amanita* from its inception (Pringle *et al.* 2009), and had moved into genomics to further that inquiry (Hess *et al.* 2014; Kohler *et al.* 2015). *Amanita phalloides* is famous for its deadly amatoxins and phallotoxins (Hallen *et al.* 2007), but we don't yet understand why the mushroom is loaded with these expensive compounds. *Amanita* is also a promising system for the study of invasion. *A. phalloides* is an ectomycorrhizal species whose invasion of the US West Coast was associated with a host switch. *A. thiersii* is a saprotrophic species that may be clonal which has spread rapidly from Texas to Illinois in only ~60 years (Wolfe *et al.* 2012). Having two species of *Amanita* exhibit such different invasion styles makes the genus a good choice for studying invasion biology in macrofungi. The study of invasion naturally leads to questions about life history (Golan *et al.* in review). *AmanitaBASE* was designed to study these longstanding questions as well as anticipating the data that would be needed to investigate unanticipated questions.

Genomes are often sequenced without particular hypotheses or applications in mind, as acts of "discovery science." Discovery science aims to simply describe the elements of a system rather than address any hypothesis about how the system works (Aebersold *et al.* 2000). Natural history could be considered a discovery science. The discovery science approach has been prevalent in genomics since the days of the Human Genome Project, when it was anticipated that the benefits of a complete human genome sequence would greatly exceed those of smaller groups doing only the hypothesis-driven sequencing that they needed for their own research.

The same spirit of discovery that leads us to sequence genomes should lead to the collection of accompanying ecological information, information which may be difficult or impossible to ascertain after genomes are sequenced, but which may explain possible findings.

Critically, for most organisms, if ecological context is not recorded at the time of collection, it cannot be reconstructed later. Even when a specimen is collected or a genome is sequenced with a particular application in mind, ecology often emerges as relevant to the original question, let alone the further questions that may be inspired by the data. For example, Elmore *et al.* (2015) found a gene cluster in *Fusarium oxysporum* species complex strains in the Broad Institute Fusarium Comparative genomes (Ma *et al.* 2010) sequenced from the USDA ARS NRRL collection that possibly conferred some resistance to cyanate fungicides. Phylogenetic evidence indicated the gene cluster was horizontally transferred among strains of the *Fusarium* species complex. Unfortunately, without required or reliable collection locations for the strains in the NRRL database, it was impossible to detect any kind of geographic pattern in the distribution of the gene cluster. Even if new *F. oxysporum* strains were collected and screened for the gene cluster, they would never replace the historical strains that could add the dimension of time. Those data were lost forever.

With a discovery science design that anticipates what data is most likely to be needed to address more than one particular hypothesis, *AmanitaBASE* enables us to ask a variety of questions that require irreplaceable data. There are many fascinating questions in the *Amanita* system, all of which potentially impact each other, so it is ideal to get answers to different questions using the same populations so that answers may be directly compared. For instance, under the “predator release” hypothesis, invasive *A. phalloides* should make less toxin over time, since it is presumably not encountering the same predators in its native range (Keane & Crawley 2002). Invasion meets toxicity. Under the “ecological release” hypothesis, invasive *A. phalloides* may be able to grow bigger or make more spores without its native predators (Wolfe *et al.* 2010). Invasion meets natural history and life history. Using the same *AmanitaBASE* populations to answer different questions minimizes uncontrolled variation when we attempt to understand the relationship between these different phenomena.

Ecological data is important context for population genomic data. Many population genomics methods are blind to the individuals that make up the population, usually caring more about allele frequencies across the entire population, but the unique ecological contexts of specimens may matter to interpreting the results of those analyses. An individual's local context may hold the key to explaining variation unaccounted for by genotype or shared environment. Phenotypic plasticity that causes variation within a population, for example, is generally due to non-shared environment. In many cases, capturing the differences among individuals within a single population requires describing the microhabitats of individuals in detail, as there is typically a "population" is defined from a restricted geographic range without obvious environmental macro-habitat variation or variability.

AmanitaBASE is an ecological genomic database. A database, in the most inclusive sense, is an organized collection of data. Ecological population genomics combines population genomics, the study of differences within and between populations using genomic data, with ecological perspectives on the contexts of populations and natural history of each specimen. Here, we describe the development of databases used to link disparate kinds of ecological and genomics data and metadata to specific individuals, enabling an approach to ecological population genomics that combines methods of natural history observation and population mapping with population genetic and genomic approaches to data analysis in what we term "ecological population genomic databases."

Anyone developing a study system for any reason should consider collecting ecological data for specimens a best practice. Collecting and saving more complete information about specimens at the time of collecting, rather than collecting targeted data aimed only a particular question, allows one to ask bigger questions and approach their investigation with less prejudice. But this is a daunting task, since there are potentially unlimited ecological data points to be collected. This chapter details the process by which we determined what biological and ecological data to collect for *AmanitaBASE*, which we hope will serve as both a resource and a guide.

Methods

All scripts may be found in this chapter's Github: https://github.com/elmoremh/Amanita-Population-Genomics

Collecting, processing, and accessioning specimens

The *Amanita*BASE protocol: California and Portugal, 2015

The core of the *AmanitaBASE* specimens are *A. phalloides* collected at several sites in Point Reyes National Seashore, California, USA in December 2015 following a detailed collection protocol. A smaller collection of *A. phalloides* was made by Susana Gonçalves in Coimbra, Vilarinho, and Agrária, Portugal during the same year.

Several sites from previous work in Pt. Reyes National Seashore were revisited (Drake 2, 3, and 4 and Heart's Desire 1, 2, and 3), located by GPS coordinates and positions refined by Anne Pringle's recorded landmarks. To find new sites, we visually scanned for mushrooms from a car window. Four new sites were established: Pet 1, Pet 2, Picnic 1, and Picnic 2. When several mushrooms were sighted together, we proceeded to define the area as a spatial population, generally ~10m x 10m. GPS coordinates were recorded at the center of the spatial population. Commercial GPS is only accurate to within ±5 meters, so the centerpoint is chosen based on both it and landmarks. There is therefore slight variation in the location of the centerpoint of some populations over the years. This spatial definition of a population is somewhat arbitrary, but we ensured that populations were separated by more than 50m or by a natural break such as a road or waterway. The environment of each population was photographed to help identify possible tree hosts, and trees and surrounding flora were identified and noted.

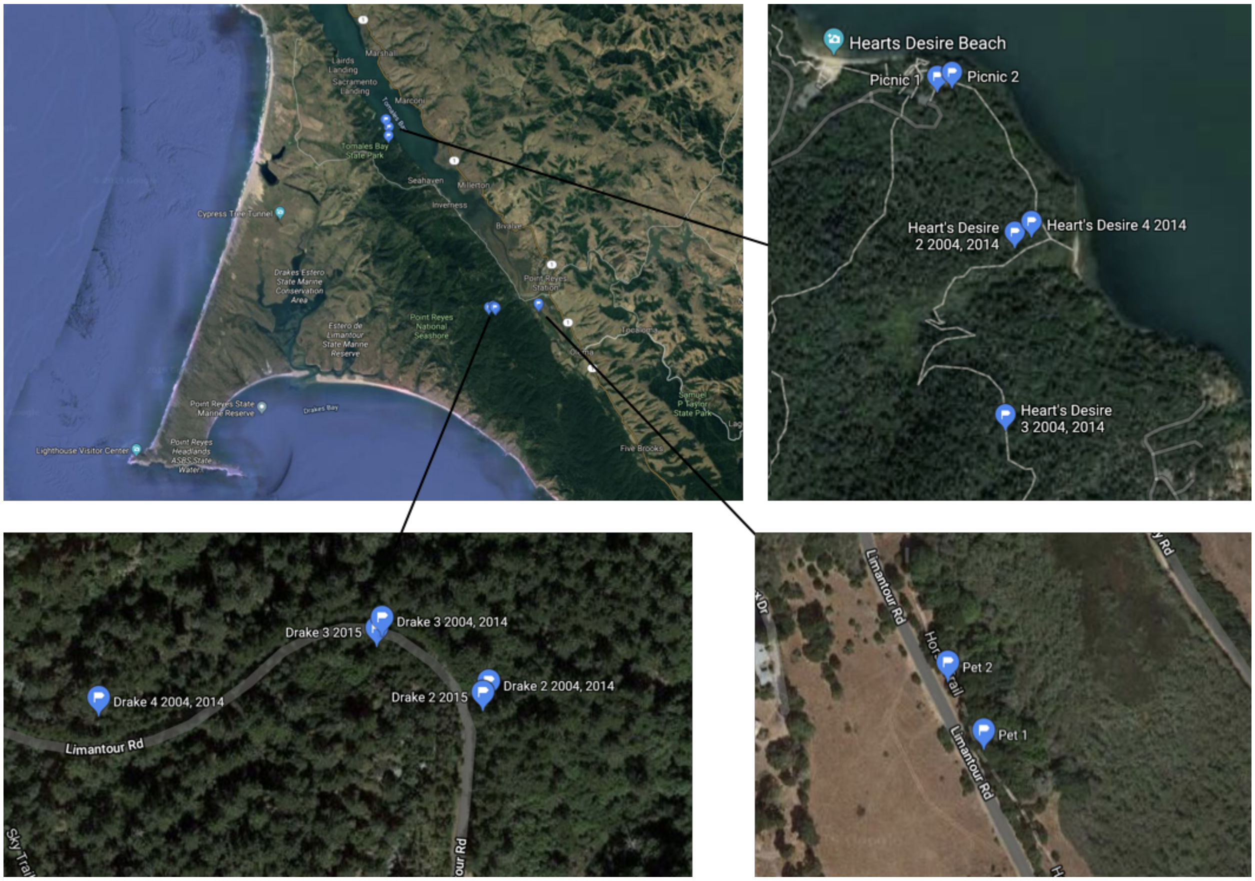


Figure 1.1: Google satellite images of the Pt. Reyes National Seashore populations. Each population is less than a mile from the Tomales Bay and usually directly adjacent to a road.

Once the limits of the population had been defined, we found every *A. phalloides* specimen within it and marked it with a flag. Next, we determined which ones we were going to collect. Overall, we wanted mostly mature and undamaged mushrooms, but we also collected a few older and younger specimens for microbiome studies. Mushroom age and condition were noted. There was a general goal of collecting 15-50 mushrooms per population in California and as many as possible in the Portuguese populations, as European *A. phalloides* populations (defined spatially as 10m x 10m) are frequently in the single digits and rarely exceed the teens. If there were three or more mushrooms to be collected, the position of each was recorded using a compass surveyor and field tape. If there were only two, the distance between them was measured and GPS coordinates collected at the midpoint.

After mapping, each mushroom was examined and given an entry at the bottom portion of the population worksheet (AmanitaBase Master Data Sheet_final.docx in Supplementary Materials). Each mushroom chosen was arrayed with collecting materials and labeled with a SpecimenID, a 5-digit number assigned to a single specimen and a unique identifier that would serve as its primary ID in the database. Batches of SpecimenIDs are given out by the team data manager before a collection trip so there is no possibility of accidentally assigning the same numbers during concurrent collection trips (Supplementary Table 1 shows past assigned number batches). During the collecting trip, the data manager assigned specimens their unique identifiers so that there was no confusion caused by other team members acting independently.

With the unique identifier visible, each specimen was photographed in detail. Only then would the specimen be disturbed. First, an inch-long isosceles triangle would be excised from the cap with a fresh, sterile scalpel blade and placed in a 15 mL Falcon tube of RNAlater. These samples will be useful for any extracting any nucleic acids but were intended particularly for transcriptome analysis. Second, the mushroom would be exhumed, making sure to include the entire volva, and placed in a wax bag. Third, an autoclaved spoon would be used to collect soil from beneath where the mushroom had emerged (and at some sites at further distances as well) and place it into a 50 mL Falcon Tube. At some sites, bags of soil were collected from around mushroom sites in the hopes of identifying root tips.

After days of collecting, we brought the specimens to the Bruns Lab at Berkeley for processing. Cap diameter was measured as well as length of the mushroom from the tip of the volva to the top of the cap. The entire mushroom was weighed in its wax bag. Then samples of the mushroom taken from the inside of the cap and the inside of the stipe were harvested for the purpose of microbiome analysis. For genomics and toxicity analysis, we cut large chunks of gill, cap, and interior of stipe tissue, placed them in 15 mL Falcon tubes and lyophilized them. Soil was preserved by placing it directly in a lyophilizer (no freezing required first).

We collected a total of 60 mushrooms from 7 populations in the December 2015 California trip (see Field Notes in Supplementary Materials).

Other specimens

*Amanita*BASE is about standardizing protocols and ensuring data completeness and integrity, maximizing general usefulness of the specimen for population genomics and ecology and minimizing loss of irretrievable data. But *Amanita*BASE must be flexible enough to accommodate valuable specimens that were collected by different collectors with different needs and priorities, particularly if those specimens are irreplaceable. Other specimens have been entered into *AmanitaBASE* that do not meet the above described protocol but merit inclusion due to their unique geography or age. The bulk of these specimens are *Amanita*e collected by the Pringle Lab and its affiliates with accompanying information that largely overlaps with *Amanita*BASE protocols. Some specimens were requested from other collections or Herbaria (particularly Kew Gardens), and these generally included information about location, date of collection, physical descriptions, etc. but rarely had GPS coordinates, for example. These specimens were generally much older than the rest of the *Amanita*BASE specimens or from a location that broadened the geographic reach of *Amanita*BASE.

Sequences generated for *Amanita*BASE

Sequences are deposited under NCBI Bioproject number PRJNA565149 and will be released in September 2020.

In 2015, I proposed creating *Amanita*BASE and using it to undertake a massive *A. phalloides* sequencing effort. The goal was to look in-depth at the invasive Californian Drake 2 and Drake 3 populations over time, compare them to native European populations, and include some greater variation through time and geography to help put the results into perspective. 89 specimens were ultimately selected, 86 *A. phalloides*, 1 *A. thiersii*, 1 *A. foetens,* and one isolate of uncertain species, either *A. thiersii* or *A. foetens.* Of the *A. phalloides*, 67 were from Drake 2 and Drake 3 in California over the years 2004, 2014, and 2015; 11 collected from Portugal in 2015; and 8 older specimens from across Europe. The *A. thiersii/A. foetens* specimens were included in the sequencing and variant-calling to help answer a species identification question and to see if there were any changes in the genome of 10802/*Ath* Skay 4041 as it has sat in culture for 5-6 years since it was originally sequenced (see Chapter 3: section III, subheading *HD1 and HD2 appear absent in other sequenced A. thiersii genomes*).

The 89 specimens were sequenced with Illumina HiSeq 2500 and two of those, 10511 and 10721, were sequenced with PacBio as well. The Illumina HiSeq sequencing was 250 bp paired-end reads with a 550 bp insert (with the exception of 10801, 10169, 10170, 10171, 10277, 10003, 10004, 10007, 10010, 10016, 10018, 10019, 10023, each of which had a 350-bp insert) prepared as IntegenX dual-index libraries. The mean sequencing depth of each of the samples ranged from 10.56 to 150.86 (and see Aggregated FastQC in Supplementary Materials). Two of the above 89 mushrooms, *A. phalloides* 10721 (USA, California, Drake 3, 2015) and *A. phalloides* 10511 (Portugal, São Jacinto, Dunas de Mira) were also sequenced using a PacBio Sequel platform at the University of Wisconsin-Madison Biotechnology Center, with a 20-kb single library per specimen and yielding average read sizes of 14,833 and 14,935 for specimens 10721 and 10511, respectively. 10511, which became the reference assembly, had raw PacBio coverage of 47x with N50 read length of 6,310 bp.

Trimming was performed on all 2015-sequenced specimens with the program Trim Galore v0.4.5 (Krueger, https://github.com/FelixKrueger/TrimGalore). Reads that became shorter than 100 bp after trimming and those with a quality score less than 30 were discarded. Adapter trimming was set to the highest stringency, 1, meaning a single nucleotide of overlap with the adapter sequence was trimmed from the read. Unpaired reads were retained, though they did not end up being used in either the assemblies or the alignments.

Alignments were run against the 10511 reference assembly using the Burrows-Wheeler Alignment software, BWA, mem algorithm with default parameters (Li & Durbin 2009), in the course of the GATK best practices pipeline (see **Variant Calling**).

Assemblies

Hybrid assemblies, incorporating both PacBio and Illumina HiSeq reads, of specimens 10721 (California) and 10511 (Europe) were completed by Jacky Hess. Extensive troubleshooting was performed in the course of generating the 10511 assembly, and then the workflow arrived at for 10511 (below) was applied to 10721. The workflow began with pre-processing: trimming and filtering with Trimmomatic v 0.35 (Bolger *et al.* 2014) with the following parameters: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 CROP:245 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:25 MINLEN:100, Illumina reads sequencing error correction with BFC (Li *et al.* 2015) and PacBio reads sequencing error correction with FMLRC (Wang *et al.* 2018). Dr. Hess proceeded to test several assemblers: CANU (Koren *et al.* 2016) and FALCON/HGAP4 (Gordon *et al.* 2016) are PacBio-only assemblers; ABySS (Simpson *et al.* 2009), Platanus (Kajitani *et al.* 2014), and Allpaths LG (Gnerre *et al.* 2014) are Illumina-only (although to meet the library prep criteria for Allpaths LG, Illumina libraries with the proper insert size were simulated using the program wgsim (https://github.com/lh3/wgsim) on PacBio reads); and SPAdes (Bankevich *et al.* 2012), DBG2OLC (Ye *et al.* 2016), and Cerulean (Deshpande *et al.* 2016) are hybrid assemblers. Based on metrics like contig number and lengths, SPAdes and Platanus were not considered further. Scaffolding was performed with LINKS (Warren *et al.* 2015) and each assembly was “polished” with the gap-filler PBJelly (English *et al.* 2012) and the base and indel correction program Pilon (Walker *et al.* 2014). The polished assemblies were analyzed with QUAST (Gurevich *et al.* 2013), BUSCO (Simão *et al.* 2019), and REAPR (Hunt *et* *al.* 2013) and evaluated on completeness of eukaryotic single copy complement, completeness of eukaryotic duplicated gene content, fragmentation, missing sequence, assembly size as percent of genome size, number of scaffolds, scaffold N50/NG50, and scaffold L50. Allpaths LG was chosen as the preferred assembler because the Allpaths LG assembly had the highest ranking in the greatest number of evaluation criteria (see Assembly Strategy in Supplementary Materials).

Variant calling

SNPs and Indels were called using the Genome Analysis Toolkit (GATK) v3.8-0-ge9d806836 software (Depristo *et al.* 2011) and following the GATK best practices (KateN 2016) as well as is possible for a non-model system, with help from Allison Shultz's Github page (github.com/ajshultz/whole-genome-reseq). I began by aligning the Illumina reads from each sample to the 10511 hybrid assembly using BWA (Li & Durbin 2009). Mapping rates ranged from 20.0% - 95.3%, with a median mapping percentage of 86.1%. Specimen 10003 had a low mapping rate and did not cooperate with the GATK workflow, so it was removed from the joint-calling cohort. The mapping rate of 10511's Illumina reads to the 10511 hybrid assembly was 93.8% (see Alignment and Deduplication metrics in Supplementary Materials). Following the steps, I marked duplicate reads, but I did not recalibrate base scores based on known variants because there were no known variants for *A. phalloides.* The GATK program Haplotypecaller makes variant calls jointly on all the samples, generating a GVCF file that contained a record of all sites of all the genomes, whether invariant or variant. The GATK program GenotypeGVCF creates the raw VCF files containing only the SNPs and Indels. Our samples contained 1831629 raw variants.

Again, due to the lack of known variants in *A. phalloides* to compare them to, the raw variants were hard-filtered according to the GATK default parameters of the VariantFiltration program. Because of the generally high coverage (~50x) of our sample set, hard filtering with default parameters was not expected to bias the filtered variants.

Database Design

Schema design

The *AmanitaBASE* schema— the layout of the data tables and connections between them— was first designed in MySQL Workbench following *SQL for Dummies* and the Linkedin Learning course *Programming Foundations: Databases* best practices. The work of designing a schema starts with determining the tables, which are organized around primary keys (marked with yellow diamonds in Figure 1.2). For example, in Figure 1.2 “Specimen” is a primary key. Then, the fields of each table are determined. Each field is an attribute of the primary key. Finally, the tables are linked by designating some fields as foreign keys (red diamonds in Results Figure 1.2), or primary keys in one table that appear as attributes in another table, reflecting the relationship between different tables. For example, in Results Figure 1.2, “Subspecimen” is a child table of “Specimen,” meaning Subspecimen is a foreign key of the Specimen table, because subspecimens, pieces or extracts of specimens, are properties of specimens. The ideal goal of designing a schema is to organize the data perfectly logically, so that there can be no conflict about where to put a piece of data or unanticipated ambiguity when the database management system enforces the rules of the schema. This process is called normalization. My goal was to achieve 3rd normal form, meaning that all non-key attributes depended only on keys, according to E. F. Codd’s classification (1970).

Database Management Systems (DBMS)

The original plan was to implement the schema in MySQL on a Harvard, University of Wisconsin-Madison, or Amazon Web Services server, though this never occurred. *Amanita*BASE v1 was implemented by Michele Clamp, then of Harvard Informatics, in her miniLIMS system, a DBMS designed for managing sequencing orders, on a Harvard Informatics server. miniLIMS is not based on SQL, but a “4-column semantic datastore” designed by Clamp. Though it is not a relational database, it can implement a relational schema.

*Amanita*BASE v2 was implemented in Filemaker 17 with guidance from the Linkedin Learning courses *Learning Filemaker* and *Filemaker: relational database design.* Filemaker is a GUI for relational database design that allows users to make attractive web-enabled databases without having to code. We chose Filemaker (despite its yearly $500+ price tag) because it was user-friendly, web-enabled, and would be easier for future database managers to pick up where the last manager left off. The hope was to make a user-friendly web-enabled database that was exhaustive enough for in-house use but intuitive enough to present to the public. As part of the Filemaker implementation process, the original schema (Figure 1.2) was slightly simplified (Figure 1.4).

*Amanita*BASE v3 is a version-controlled metadata google sheet with links to objects such as photos of specimens and site, sequences, and variants stored in my Google Drive and shared with members of the Pringle Lab. It is regularly exported to Excel and those workbooks saved in my Dropbox. It is a “flat” database consisting of one table, Specimen.

Metadata

*Amanita*BASE SpecimenIDs

SpecimenIDs are assigned by the database manager in batches so that no two people are at risk of using the same IDs for different mushrooms (see Supplementary Materials for a table of assigned SpecimenID batches). In accordance with databasing theory best practices (Taylor 2011), the SpecimenIDs are simply identifiers. There is no meaning conveyed through the name label rather than as a proper database field.

Specimen fields

The full *Amanita*BASE v3 collection protocol requires each specimen (mushroom or culture) to have the following accompanying metadata. Except for the colon following each field name, the names are written in precisely as they appear in the *Amanita*BASE Specimen Metadata v3 sheet. Consistent naming is key to data integrity, particularly if *Amanita*BASE becomes a relational database again in with a bulk import function.

Storage: Where is the specimen (or subspecimens) stored?
SpecimenID: Unique identifier, 5-digit number between 10000 and 99999. SpecimenIDs are given out in batches or on a case-by-case basis by the data manager.
Other Name(s): List any aliases used for this mushroom, whether by a previous owner or just as a shorthand in field notes.
Species: Entries should all be genus *Amanita*, but species must be specified
Date Collected: DD-MM-YY
Collected By: Who collected the specimen?
Method of Preservation: i.e. Lyophilized tissue, dried tissue, RNAlater
(Species) **Determiner:** Who determined the species of the specimen?
Previous Use of Specimen: Is this specimen associated with any notable earlier projects, e.g. AFLP from 2005 (Golan *et al.* in review)
Latitude Collected: GPS coordinate
Longitude Collected: GPS coordinate
Country
State/Province/Region
County
City
Site Name: Site name must be agreed upon and standardized at the time of collection, if not before
Day: Separating Day, Month, and Year makes it easier to search
Month: Separating Day, Month, and Year makes it easier to search. Particularly helpful when looking at seasonal patterns.
Year: Separating Day, Month, and Year makes it easier to search
Herbarium Source: if applicable
Host: note if known or suspected
Site Habitat: trees, terrain, any other notable features of the site
Site Picture: a picture or pictures of the site including ground cover, nearby trees, and landmarks
ITS: ITS sequence, if known
Work to be Done: any processing still to be done or purposes in mind for the specimen

Notes: Any observations, thoughts, or information about the specimen that doesn't fit elsewhere.
Photos & Field Notes: Link to photos of the specimen from the field and notes pertaining to it. Photos of the specimen must include their SpecimenID number (e.g. the label on the wax paper collecting bag) in the photo.
Sequences: link (or citation) to sequences from the that specimen; multiple links separated by commas
Variants: link to .vcf containing the specimen's variants; multiple links separated by commas.
Tag: For easier search, may refer to the collection expedition or sequencing run, etc. (Because of spreadsheet filter limitations, only one tag per specimen. Additional information can go in Notes.)

Data entry

Adding older specimens, from the Pringle Lab and from other collections around the world, to *Amanita*BASE required issuing an *Amanita*BASE SpecimenID number and some reformatting and relabeling. Any additional information accompanying specimens acquired outside the Pringle Lab is stored in the Notes column of the *Amanita*BASE v3 Specimen Metadata spreadsheet (Supplementary Materials). Specimens in Harvard's Farlow Herbarium were not given *Amanita*BASE SpecimenID numbers. The initial combined spreadsheet of old collections was compiled by Jacob Golan in 2015, and Sam Harrow prepared an updated version in February 2019 which served as the basis for the *Amanita*BASE v3 Specimen Metadata spreadsheet.

Entering specimens collected post-2015 is easier because the Collection Protocol matches the Specimen Metadata sheet, but it is still important to take care with typographical and spreadsheet errors such as erroneous autofill and copying over columns.

Results

Specimens

As of this writing, *Amanita*BASE contains 803 preserved mushrooms and 8 cultures. The majority of the specimens (505) are stored in the Pringle Lab at the University of Wisconsin-Madison, and the majority of the remainder are accessioned at the Farlow Herbarium at Harvard University. Specimens kept in Madison were given SpecimenID database numbers, while those accessioned to Herbaria were allowed to keep the labels associated with their accession there. The majority of physical specimens are either dried or lyophilized and stored in wax paper bags in cardboard boxes in a dry, cool place.

The vast majority of specimens are *A. phalloides*, with the next most represented species *A. thiersii,* followed by *A. muscaria.* 20 entries are *Amanita thiersii* specimens collected by Ben Wolfe and David Lewis in between 2007 and 2009 (Wolfe *et al.* 2012). A handful of *A. muscaria* were collected alongside the *A. phalloides* in 2015. 70 *A. phalloides* specimens from California and Portugal were collected following the *Amanita*BASE collection protocol in 2015. *Amanita*BASE also includes previous collections from the Drake and Heart's Desire sites in California that were revisited in 2015, 71 from 2004 and 149 from 2014. The culture entries include the cultures from which the *A. thiersii, A. brunescens, A. muscaria* var. *guessowii, A. polypyramis,* and *A. inopinata* genomes were sequenced. *Amanita*BASE also includes sequences and metadata (the specimens were consumed in sequencing) from 29 specimens borrowed from Kew Gardens that come from across Western Europe.

The age of the specimens stretches as far back as 1909, although there are only 36 specimens in *Amanita*BASE from before the year 2000. The specimens hale from 14 countries (Argentina, Canada, Czech Republic, Denmark, England, France, Italy, Northern Ireland, Portugal, Scotland, Serbia, Spain, Sweden, USA) across 163 unique collection sites and three continents (North America, South America, Europe), with the densest sampling in California, USA. See the *Amanita*BASE v3 Specimen Metadata spreadsheet in Supplementary Materials for complete information.

Protocols developed

Collection protocols

The *Amanita*BASE procedure described in Methods: Collecting, Processing, and Accessioning was developed to minimize errors in collecting that might lead to incorrect or incomplete data in *Amanita*BASE, as well as for the sake of efficiency. The version below (also included in Supplementary Materials as ***Amanita*BASE Master Collecting Protocol 01-01-2020**) is updated to reflect lessons learned and to be less specific to our experiments than the version actually used in the field during the 2015 California and Portugal expeditions (included in Supplementary Materials as **MasterProtocol_SamplingMapdPopls_FINAL**).

Before heading to the field

1. Specify data manager. This is ONE person who will have the final authority to make decisions about naming conventions, what data is collected, and how and where it is stored.
2. Ask data manager for a batch of unique SpecimenIDs. All labels must include the SpecimenID.
3. Bring a device with accurate GPS, compass surveyor or other tools to map mushrooms, field tape, waterproof notebook or data sheets printed on waterproof paper, camera, wax bags labeled according to the format described below, ziplock bags, tools to properly dig up *Amanita*, sterile implements for soil collection, permanent markers, etc.
4. Have a look at the laboratory protocols to make sure the lab has all necessary supplies (e.g. RNAlater, Falcon tubes).

In the field

5. *If revisiting a previously recorded population*, use latitude and longitude as well as landmarks to find the centerpoint of previous collections. (GPS coordinates are accurate ± 5 meters, so you will not find the exact same center.)

If defining a new population, find a centerpoint, record latitude and longitude, and define the population as an approximately 10 m x 10 m space around it (with some flexibility).

6. Take a picture of the site.
7. In a fresh ***Amanita*BASE Population Data Sheet** (Supplementary Materials), describe the site, especially biotic features and landmarks. Note potential tree hosts and their approximate percentage cover. Name the site and, if it is similar to other site names, specify an abbreviation on the population data sheet.
8. Decide which mushrooms to collect. Target young mushrooms, unless you have some reason to prefer old, as old or rotting mushrooms preserve poorly and are harder to handle. If there are more than a few mushrooms at the site, plant flags by them labeled with their SpecimenID.
   In North America, collect 15-50 mushrooms evenly from across the population.
   In Europe, collect 5-as many as you can, but even singletons are often valuable (if they are from a novel location, for example).
9. If there are three or more mushrooms, map them using a compass surveyor or other mapping technique. If there are two, take GPS coordinates at the midpoint between them.
10. Note the SpecimenIDs, maturity, and any notes on the ***Amanita*BASE Population Data Sheet** (Supplementary Materials)**.**
11. Lay down your collecting materials (wax bag, falcon tubes, scalpel blade, etc.) beside each mushroom. Label every bag and tube that will store a mushroom with the SpecimenID. Before disturbing a mushroom, take several photos of the mushroom with the SpecimenID labels visible.
12. Exhume the entire mushroom, including the volva (you will probably have to dig around the mushroom a bit).
13. *RNAlater samples:* If you wish to extract RNA from your samples, have Falcon tubes of RNAlater prepared. Using a fresh scalpel blade and sterile technique, cut a chunk of tissue from the edge of the pileus with gills included and immediately place it into the SpecimenID-labeled tube of RNAlater.
14. Place the entire mushroom into its SpecimenID-labelled wax bag.
15. *Soil samples:* If you wish to examine the soil around the specimen, bring autoclaved spoons wrapped in tinfoil (sterility is important if, for example, you are interested in soil microbes) and Falcon tubes or sufficiently large plastic bags for your desired sample size, always labelled with the associated SpecimenID.
16. Store all samples in refrigerator or cold room until you are able to process them in the lab.

Remember to sterilize boots and tools if moving into an uninvaded area.

In the lab

17. Brush away loose dirt on mushrooms before processing.
18. *Size and biomass*:
    Measure the cap diameter and length from the tip of the volva to the top of the cap.

Weigh the entire mushroom.

19. *Archiving:*
    Put about a half of the mushroom in a labeled paper bag in a drier. This will become the herbarium specimen. Please try to include all parts of the mushroom: cap, gills, stipe, annulus, volva.
20. *Microbiomes:*
    Using sterile technique, cut 3 sections and place each in an individual Falcon tube filled with RNAlater. (Take these sections from the same side of the mushroom so that the other half of the mushroom can go to the herbarium.)
    Take one section of gill tissue, one section of cap tissue, and one section of tissue from the interior of the mushroom, being careful to break apart the mushroom without contaminating it with microbes from the pileus or other exterior surfaces (as if you were collecting a chunk of tissue for culturing).
21. *Soil:*
    Soil can go directly into the freezer.
22. *Genomics and toxin analysis:*
    With remaining tissue, cut large chunks of gill tissue, cap tissue, and tissue from the interior of the mushroom. Place in labeled Falcon tubes and lyophilize. Lyophilization of mushrooms involves 1) flash freezing in liquid nitrogen 2) placement on a lyophilizer. If there is no lyophilizer, dry tissues on a mushroom dryer at low setting (e.g. 50° C).

Data integrity protocols

Data integrity protocols have the goal of keeping specimens tied to their accurate metadata, avoiding confusion, data loss, and typographical errors.

Data Integrity General Practices

All notes and collections are tied to a SpecimenID. SpecimenIDs are assigned before any collecting or note-taking takes place, and specimens are always explicitly referred to by their SpecimenIDs. Allowing multiple labels and labeling systems for the same specimens from stage to stage of research introduces unnecessary cognitive overhead and invites preventable errors. A single numerical ID that only conveys the specimen's designation and no inherent information about the specimen preempts the tendency to shoehorn information into the name haphazardly rather than creating a proper field for necessary data. When names are allowed to become grab bags of attributes, they often either become ambiguous or grow longer and more confusing over time to distinguish specimens. There is also a tendency to give relative names without specifying an objective point of reference, e.g. “the darker one” or “next to the other one,” which are unnecessarily cognitively demanding at the least and can lead to confusing specimens at worst.

Photos of specimens contain the wax bag with their SpecimenID written on it, and notes about specimens always use the SpecimenID instead of relative or circumlocutory designations, even in quick field notes.

Designate a Data Manager (1 person!) before collecting. The Data Manager hands out unique SpecimenIDs in batches to ensure that no two collections could accidentally be using the same “next up” numbers. The Data Manager also controls naming conventions and data entry practices such as capitalization, use of whitespace, abbreviation, punctuation, etc. to ensure ease of data search and retrieval as well as to minimize errors. The Data Manager must be a single person so that they can make executive decisions. When multiple people share authority over data organization, there is a diffusion of responsibility and systems tend to fall into disarray and disuse more quickly.

Minimize cognitive load in the field. Cognitive load refers to the burden on working memory. Keeping track of too many details and making too many decisions while collecting can lead to errors, and there is plenty of cognitive load in the field with even the best-laid plans. When designing your collection protocol, modularize and parallelize tasks as much as possible, so that multiple people can contribute without creating confusion. Choreograph the procedure as much as possible beforehand, and practice that choreography-- is it going to work? Is there too much, for example, getting up and down that could be reorganized to be less tiring and allow for greater focus on precision tasks? At the beginning of the collection, make a video of the proper protocol that all may refer to if they become confused and perhaps include in publications along with collecting methods. Carry laminated workflow diagrams or multiple print copies of checklists of your collection protocol rather than trying to remember every step (or worse, not having collection protocol at all). Do as much of the work of you can beforehand as possible, such as pre-bundling the supplies needed to collect an individual. Take photos (including SpecimenID numbers) as much as possible when it can take the place of writing notes. This is especially helpful if it is cold or wet where you are collecting.

Everyone involved does a "mind dump" within one week. Many details and environmental variables seem obvious to us at the time that we collect specimens, but when those memories fade, we can be left with notes that are unintelligible, or forgetting important observations that seemed so important that we were sure we would remember. Each member of the collecting expedition must follow the procedure below within one week of returning to get crucial information out of their heads and into the database materials.

Post-field Reflection and Mind Dump

Bolded *words contained links the appropriate Dropbox folder.*

1. Scan all your field notes and deposit them **here** in the Dropbox.
2. If you can, transcribe all the notes you took into typeface and put them **here**. Feel free to offer clarifications or additional thoughts in [brackets] within the transcribed text. Only transcribe notes you took personally-- the whole point of this is to avoid puzzling over other people's handwriting.
3. Upload all specimen and habitat photos to their appropriate SpecimenID# folder in **Specimens**.
4. Upload all other photos to **Misc_Photos**.
5. Reflect: Is there anything we did that was unexpected? Are there any general impressions you had of the fieldwork, of the weather, the soil, etc.? Write down as many of these thoughts as you can and deposit them **here**. You can keep putting thoughts here whenever they occur to you-- just be sure to note the date!

6. Re-read the master protocol and note what changes we all made and what tweaks you might have personally made as comment on **this copy of the collecting protocol**.

7. [For the Data Manager] Collect everything produced by yourself and your labmates, scan all hard documents, upload all photos to appropriate storage place, and back up all collected materials (i.e. in Dropbox or Google Drive).

Sequences included in *Amanita*BASE

As of this writing, 93 *Amanita* specimens have fully sequenced whole genomes: 6 assemblies and 87 alignments. (One, *A. thiersii* 10801/Skay 4041, has had its whole genome sequenced twice.) The *A. muscaria* Koide v1.0 (Kohler *et al.* 2015) and *A. thiersii* Skay 4041 v1.0 (Hess *et al.* 2014; Chaib de Mares *et al.* 2015) are available through the Joint Genome Institute of the Department of Energy's Mycocosm portal (Nordberg *et al.* 2014). *A. brunescens, A. inopinata,* and *A. polypyramis* were sequenced and assembled in-house by the Pringle Lab (Hess *et al.* 2014).

A reference assembly for *A. phalloides,* 10511 from the Mira site in Portugal, was completed by Jacky Hess. *A. phalloides* 10511 is a hybrid assembly of PacBio and Illumina reads with combined 84x coverage. Using the pipeline identified with 10511, *A. phalloides* 10721, from California population Drake 3, was also hybrid assembled, although it was not used to align any of the other genomes described below. Based on kmer analysis, the size of *A. phalloides* 10511's genome is 45.5 MB, close but not matching the estimate of 43 MB for 2014 *A. phalloides* assembly Dr4M1 (Chapter 3: section IV).

A set of 89 genomes (86 *A. phalloides* and three *A. thiersii*) were processed in parallel using the GATK variant discovery best practice pipeline (KateN 2016), which includes creating alignments with the BWA software (Li & Durbin 2009), and *A. phalloides* reference assembly 10511. The majority of the 89 genomes that were sequenced in parallel are from California populations Drake 2 and Drake 3, because these sites were collected from in 2004, 2014, and

2015, allowing us to track changes over time and space. A complete list of the sequenced specimens can be accessed by filtering the Tag column of the AmanitaBASE v3 Specimen Metadata google sheet (Supplementary Materials) for the tag “2016 Sequencing,” but note that one of the 89 specimens sequenced that year (10801) is under the Cultures tab. The 86 *A. phalloides* genomes associated with Golan *et al.* (in review) are deposited under NCBI BioProject PRJNA565149, which will be released in September 2020.

Called variants

After filtering, our 89 mushroom samples contained 212119 indels and 1580133 SNPs. Both raw and filtered variant sets are available in Supplementary Materials.

Database

Schema

The first complete draft of the *AmanitaBASE* schema was in 3rd normal form, meaning that all non-key attributes depended only on keys (Codd 1970), but, on the advice of Aaron Kitzmiller, for the sake of efficiency some tables were combined with larger tables to give the final schema in Figure 1.2.

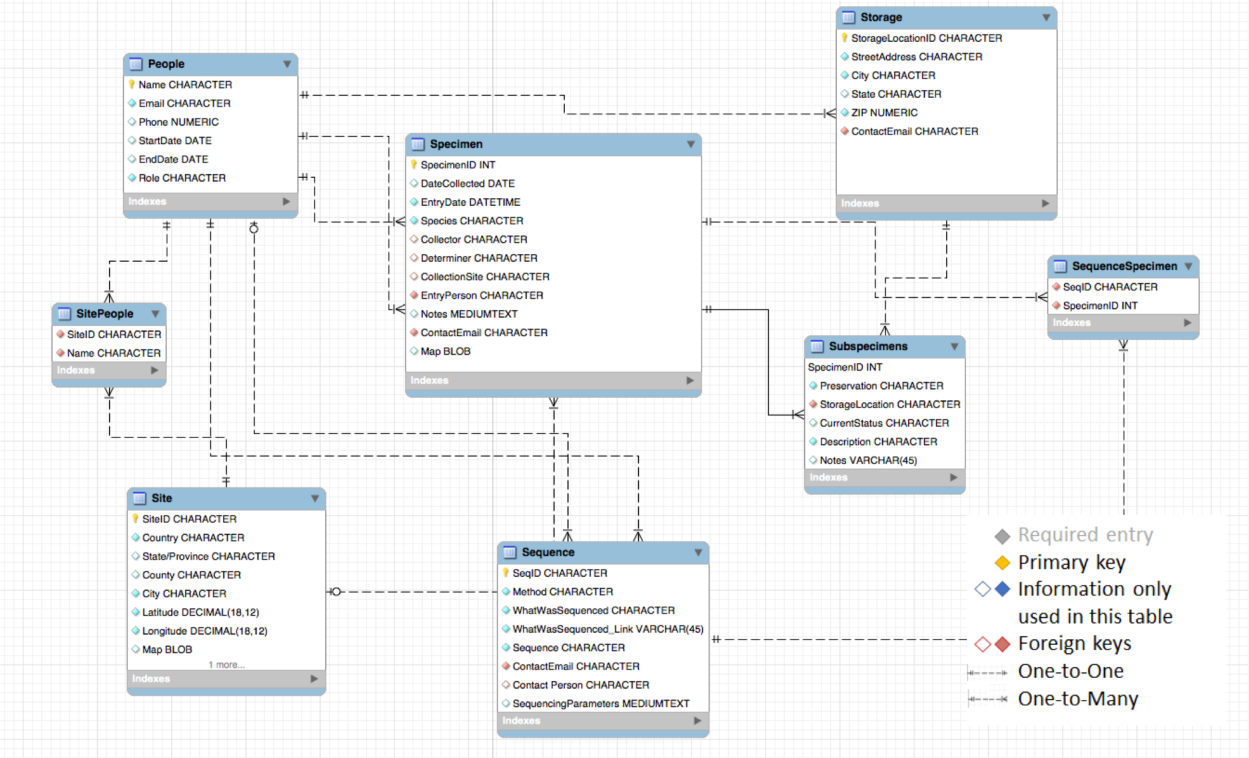
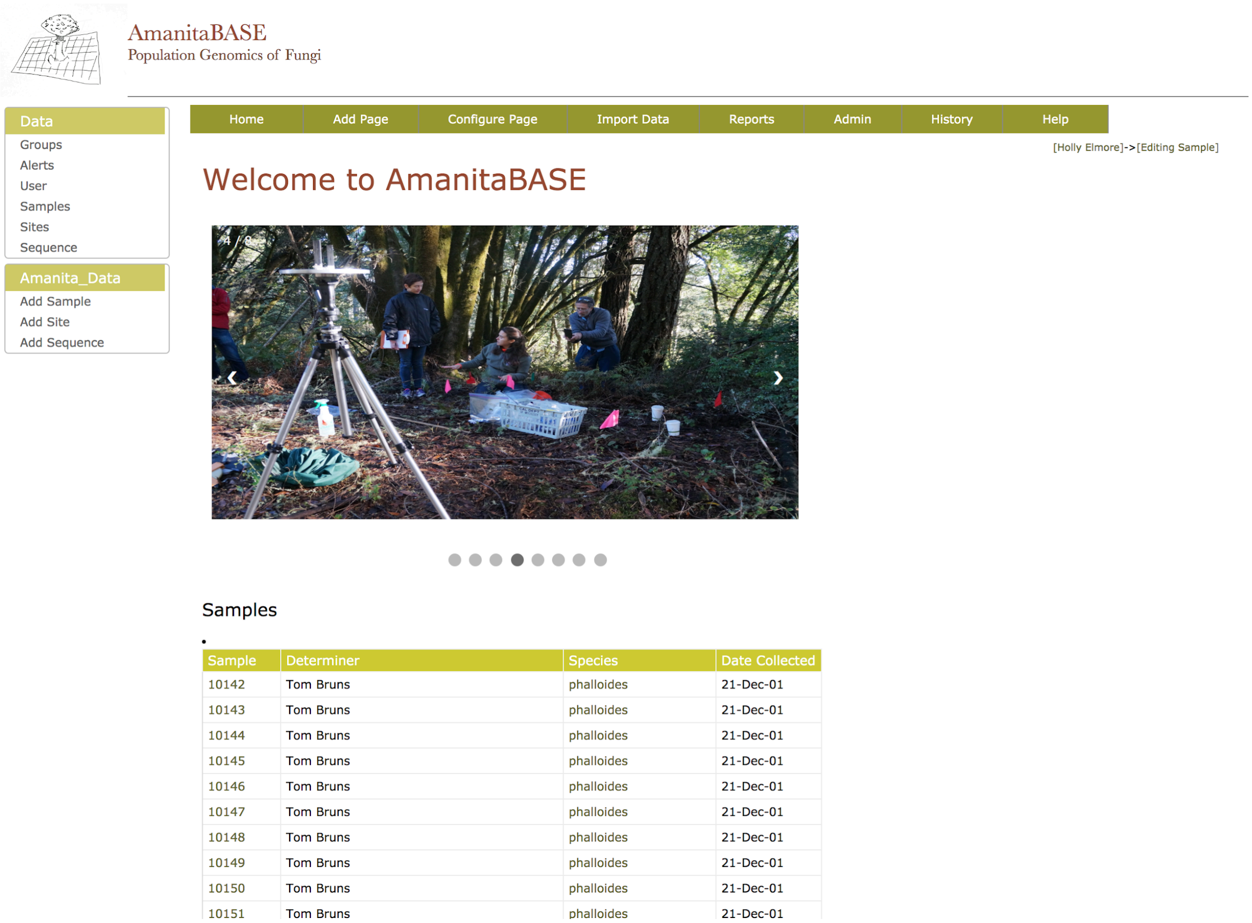


Figure 1.2: The *Amanita*BASE v1 schema in MySQL. Here, the yellow diamond indicates a primary key and the red diamond indicates a foreign key. Lines connecting the tables represent one-to-many relationships between the primary keys of parent tables and the foreign keys of child tables. For example, the “Subspecimen” table is a child of parent table “Specimen.”

*Amanita*BASE v1: miniLIMS



Welcome to AmanitaBASE

Samples

Sample	Determiner	Species	Date Collected
10142	Tom Bruns	phalloides	21-Dec-01
10143	Tom Bruns	phalloides	21-Dec-01
10144	Tom Bruns	phalloides	21-Dec-01
10145	Tom Bruns	phalloides	21-Dec-01
10146	Tom Bruns	phalloides	21-Dec-01
10147	Tom Bruns	phalloides	21-Dec-01
10148	Tom Bruns	phalloides	21-Dec-01
10149	Tom Bruns	phalloides	21-Dec-01
10150	Tom Bruns	phalloides	21-Dec-01
10151	Tom Bruns	phalloides	21-Dec-01

Figure 1.3: The frontpage of *Amanita*BASE v1 on Michele Clamp's miniLIMS system running on one of Harvard's FAS Informatics servers.

The schema in Figure 1.2 was implemented by Michele Clamp, then of Harvard FAS Informatics, in the miniLIMS system she designed for Harvard in conjunction with sequencing orders. Michele Clamp left Harvard suddenly, before the initial creation of the miniLIMS *Amanita*BASE was completed, and I did not have sufficient access to the back-end of miniLIMS to make all the necessary changes. It became clear at this point that, even if someone else helped me to complete the initial *Amanita*BASE, it would not do to have *Amanita*BASE so tied to a particular school, let alone a particular group within a school, using an idiosyncratic system. *Amanita*BASE v1 now lives at https://amanitabase.rc.fas.harvard.edu/amanitaBASE//plugins/Amanita/home.php and requires login credentials, which you may request by registering as a new user. It is no longer maintained and could be erased at any time.

*Amanita*BASE v2: Filemaker

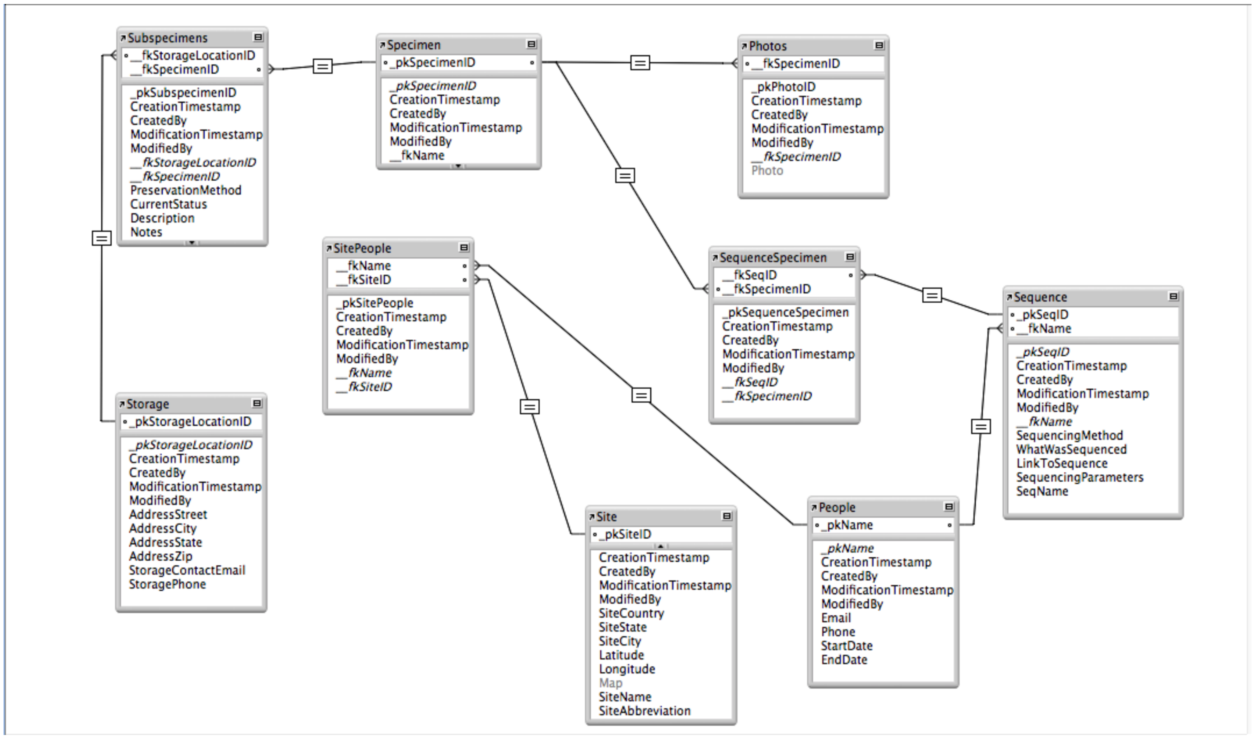


Figure 1.4: The schema of the Filemaker implementation of *Amanita*BASE. Here, the _pk prefix indicates a primary key, __fk indicates a foreign key. Lines connecting the tables represent one-to-many relationships between the primary keys of parent tables and the foreign keys of child tables. For example, the “Photos” table is a child of parent table “Specimen.”

Even user-friendly Filemaker proved to be too complicated. The Filemaker *Amanita*BASE presented a high enough barrier to entry that the revealed preference of *Amanita*BASE’s users (including its designer) was to access the same data through spreadsheets. There were enough idiosyncratic decisions I had to make in designing the Filemaker database that it seemed likely I would have to remain involved in maintaining it for many years to come. Reliance on a single person who is no longer directly involved in working with the database’s data to maintain the database is not a solid continuity plan. Indeed, this is one of the most common reasons that databases fall into obsolescence. Queries, one of the major motivations for setting up a relational database, did not turn out to be as easy to use, let alone teach others to use, as I hoped compared to filtering a spreadsheet. Maintaining an annual subscription was a further disincentive.

*Amanita*BASE v3: version-controlled metadata spreadsheet

Though not a relational database, a “flat database” spreadsheet is often good enough for storing, accessing, and searching collections data. A spreadsheet does not offer sufficient data security, however, especially if there are multiple people accessing it, because of the ease of erasing or copying over columns. If these errors aren’t quickly discovered, they are often irremediable. To mitigate this danger, the metadata component of *Amanita*BASE v3 is now a version-controlled Google sheet with a read-only archival Excel sheet saved in a Dropbox as a further backup. With a Google sheet, the user that makes each change is recorded as well, one of the major advantages of a full-blown DBMS over spreadsheets up until this point. Sequences, variants, and specimen photos are included in *Amanita*BASE v3 Specimen Metadata as links, rather than local copies as they would have been in v1 or v2.

Some fields were dropped in order to compress a seven (main) table schema into one flat spreadsheet. Essentially, all fields are properties of the Specimen table now. This was not a difficult translation, because it reflects the organization of the original collection record spreadsheet. The Photos table has been replaced by a Dropbox/Drive link to a folder containing all that specimen’s photos and field notes. The People table does not exist anymore. Not recording the contact details of every person involved in collecting and which collections they were connected to will mean that information about the specimens gets lost, but without a convenient way for people to enter their information once (and restrict access to that information to users with the proper credentials), repeating this information for each specimen entry becomes tedious, most likely more effort than it is worth. Future users will just have to do a little investigation with the names provided in the Collector and Determiner fields. The Storage table was jettisoned because the collection is very consolidated-- there are currently only two Storage locations, and that is denoted by different tabs in the spreadsheet. For the Specimens stored in the Pringle Lab at the University of Wisconsin-Madison, there is a field, “Box/Drawer (at UW-Madison)” specifying which box the specimen is in. The boxes are labeled with the letters of the alphabet. The loss of the Site table means it’s less convenient to check the details of a site, but many properties of the site are easily converted to properties of the specimens (Latitude Collected, Longitude Collected, Country, State/Province/Region, County, City, Site Name, Site Habitat, Photos & Field Notes). It’s less elegant from a databasing perspective to repeat data in this way, but there’s nothing wrong with it. In fact, my original vision for the Site table was too prescriptive, because it didn’t allow for slight changes in the information about a site from year to year due to GPS resolution or field conditions.

The most lamentable lost table is Subspecimen. Trying to include subspecimens, such as multiple pieces of the same mushroom, extracted DNA or RNA, associated soil samples, or soil samples possibly containing root tips, on a flat Specimen table is difficult. Users might be tempted to break important database rules, such as creating “sub” rows (where information about the relationship between two rows is conveyed visually instead of as a relationship specified within the database as a field) or creating invalid SpecimenIDs (such as “10511a” or “10220.1”) that undermine the integrity and searchability of *Amanita*BASE. For now, it is preferable not to include information about subspecimens in *Amanita*BASE v3. It may be too demanding to ask users to keep track of every time they create a subspecimen, and the locations of large collections of subspecimens (such as DNA extracted for sequencing) are few and known to the Pringle Lab.

Lessons learned: when are ecological genomic databases worth it?

A relational database version of *AmanitaBASE* is unlikely to be released to the public or even replace Excel/Google sheets in the Pringle Lab. For all the advantages of a fully functioning relational database when it comes to data integrity and queries, building and maintaining a database that functions takes a lot of time, effort, coordination, and even money both upfront and on an ongoing basis. A defunct relational database that no one uses can be worse than no organizational structure at all, because people will turn to undisciplined and incompatible means of recordkeeping. It is important to be realistic and to ensure that you do not go to the effort of designing a database only to have it backfire, making data organization worse! Even if the database works fine, it’s possible that it was not better enough than a flat database to justify the investment and ongoing costs to maintain.

How to decide whether to pursue an ecological genomic database? In the case of *Amanita*BASE, after years of developing different aspects of the database and trying different platforms, I realized that it was simply too much work for me as the data manager for the level of anticipated benefit. I wanted to see the database finished regardless, but I was not confident that I could build a database that was polished enough for another data manager to take over. Despite my best efforts to keep schemata simple and note and explain my decisions while building the database, I suspected that, if I wanted the database to survive past my PhD, unless I simplified it I would have to continue to be personally involved in running it.

Meanwhile other lab members continued to prefer the Excel sheet to prototypes of *Amanita*BASE on miniLIMS. Despite pleasing layouts and pictures, the user experience of the miniLIMS or Filemaker databases never came close to the ease and straightforwardness of a spreadsheet. I myself went to the Excel sheet to perform complexes search, the kind that database queries were supposed to be so good for. Query functionality was a major reason I was interested in an SQL-based database, but unfortunately it was much more difficult to specify queries than I had anticipated, let alone to teach other users how to specify queries. Filtering in Excel was much faster and easier. Filtering doesn't scale with number of records as easily as queries, but it is unlikely that *Amanita*BASE will ever contain enough records to make filtering and manual search unworkable.

I began to think that, by making a relational database, I was "fixing" all the features of spreadsheets that weren't broken. The bugs I had intended to fix were Excel auto-formatting errors, pulldown copy-over errors, lack of standardization that impedes searchability, entering grossly incomplete records and never returning to complete them, and disagreements in editing with multiple users. I learned that with git version control, Google sheets version control, or just better practices with accessing Excel sheets as read-only unless you are authorized to make changes, the Excel errors that were part of the impetus for the database are less of a problem than they seemed when I embarked on the project.

Although the relational database itself was aborted, the lab benefited a great deal from the procedures that were put in place to develop *Amanita*BASE*.* Planning the database required us to step up our critical thinking about what data to collect and how to choreograph collection. Databasing theory and best practices gave us a deeper way of thinking about the relationships between datapoints and how data is accessed. The reform of the specimen naming system alone was hugely beneficial-- it transformed not only collecting and the integrity and reliability of records, but it turned fragmented previous collections into one dataset. And, finally, it was beneficial for me as the data manager to learn so much about databasing firsthand.

Other reasons that ecological genomic databases may not be worth the effort:

- It may not be worth it to preserve a piece of the field if you don't have that many questions whose answers you want to cross-compare on the same specimens.
- You may frequently require fresh material, so being able to reuse the same material isn't worth so much to you.

- If you don't need to update the same records over and over again, so a one-time entry list is fine.
- If your material is consumed by sequencing and so there is no physical specimen left to store or whose information you will have to update.
- If making the database useful requires a manager with considerable knowledge and skill or investing too much in user experience— perhaps it is better for people to just manage their own specimens for their own projects in that case.
- If the DBMS software you use costs a lot of money and you would be on the hook for an indefinite number of years.
- If the database designer doesn't want to be on the hook for updates and answering questions for an indefinite number of years.

Situations when ecological genomic databases may be worth the effort:

- When you're already fully invested in population genomics or spatially-explicit ecological collecting.
- When spatial distribution is related to dispersal, allelopathy, pollution or any process which could strongly relate to genotype.
- When you find you already have massive duplication of effort in your lab group which could, with some upfront effort, be more standardized so sequences or specimens could be reused.
- When you're looking in detail at changes in a population over time (climate change datasets, microevolution), you can't predict the exact nature of the changes, and you need to preserve maximum information.
- When your current recordkeeping system is overwhelmed, you need an overhaul anyway, and you could use more functionality.

- When you have a very valuable dataset that you would like others to use and cite you for.
- When you have frequent, time-consuming requests for data that could be stream-lined.

Discussion

Here I describe the creation of *Amanita*BASE*,* an ecological population genomic database and collection, and the theoretical motivations that drove it. I describe the development of the database itself and the protocols for collecting specimens to be entered into the database. Making *Amanita*BASE was a sort of methodological experiment-- not only have we developed an ecological population genomic resource, but we've experimented with the process of collecting non-model organisms in order to populate a database that can be used, along with the physical specimens, to do ecological population genomics. 70 specimens were newly collected in 2015 and gathered together with 741 older specimens to populate the database. 89 of these specimens had their genomes fully sequenced and variants called.

*Amanita*BASE can serve as a "model" non-model system platform. *Amanita*BASE was conceived to allow us to collect and use more complete, high quality data and seamlessly share that data by putting all the organized relevant information in one place. *Amanita*BASE improves data integrity, minimizes duplication of effort, and improves recordkeeping. *Amanita*BASE is not the all-in-one relational database platform I originally envisioned, but it has succeeded in its goals. This chapter demonstrates the need for flexible design when building an ecological genomic database and the reality of trade-offs, thus can provide guidance to researchers collecting and storing data in a range of different circumstances.

Ecological population genomics databases like *Amanita*BASE may help to bridge the methodological divide between the highly conceptually interconnected fields of ecology and population genomics. Population genomics doesn't perfectly map onto ecology, owing in part to the population-level gene pool focus in population genetics and exacerbated by the computational resource limitations of early bioinformatics which encouraged population genomics techniques to focus on alleles in a population without keeping track of which individuals the alleles were in. Therefore, it is difficult to take differences in microhabitat between individuals into account when analyzing a population. *Amanita*BASE is an experiment in using population genomics methods, like joint variant-calling, with an eye to the ecology of the individual. But because *Amanita*BASE collects a wide array of general purpose information and tissues, the potential applications of *Amanita*BASE and its contents to any number of fields are vast.

*Amanita*BASE contains not only specimens and thorough, general purpose ecological metadata, but also RNAlater-preserved tissues, nucleic acids, and sequences. This thorough collection of ordered data will be around for all manner of unanticipated purposes. It's preserving a slice of the field in a particular time and place for future scientists, with their future methods and questions, to probe the past. It's important for researchers to preserve the actual physical collections they worked on, if possible, along with digitized data and results. First, so that their own results can be verified or replicated. Second, so that, if possible, future work, ideally even by other researchers, can be carried out on the same specimens, thereby minimizing uncontrolled variation. And, third, there may be unanticipated uses for a thorough data set collected at a particular time in the past. Today's natural history collections are proving useful in investigating the past effects of climate change, in most cases because they were collected systematically and with exact dates and locations (Li *et al.* 2013; MacLean *et al.* 2016; Robbirt *et al.* 2014). In fact, *A. thiersii* specimens previously collected by the Pringle Lab have already been used to measure carbon isotope ratios as an indicator of increased C3 vs C4 photosynthesis due to climate change (Hobbie *et al.* 2017). By the time a need for material from the past arises, it is too late to collect it. Though we cannot necessarily predict what material will be needed, we can at least collect what we do collect thoroughly and with “general purpose” metadata as laid out above.

In developing *Amanita*BASE*,* we made many observations about the organizational psychology of collecting and databasing. We found, for example, that we better organized the process of collecting our data because we were collecting with ease of later access to the data in mind. The 2015 *Amanita*BASE collection trips were somewhat unique in that they began with the question “How would we like to access this data when we’re analyzing it?” Working backwards from the endpoint revealed some interesting assumptions we normally make when working forwards. Knowing the format and desired end state of our data allowed us to bypass a lot of the customary confusion in multi-person collection efforts. Because the database was going to be based on a 5-digit unique SpecimenID as the primary key, we had no confusion about naming conventions and no need to change names down the line in order to enter the specimens into a spreadsheet or database. The need to know exactly what data we would collect ahead of time led us to work out a detailed choreography of our collection protocol ahead of time. Working the protocol out in advance in granular detail allowed us to preempt many of the usual fieldwork confusions about, for example, what units we were using or what step we were supposed to do first.

Databasing theory made the process of planning what specimens and data we would collect a bit more rigorous. Designing a schema (Figure 1.2, Figure 1.3) helped to determine what pieces of data were the most pivotally connected to other pieces of data and which were extraneous or reconstructable from already included data. It would have been easy for researchers with our motives of making a general purpose ecological population genomics database to collect *too much* data of a certain type for the sake of completeness, without considering how it will be used by us or by anyone. For example, we decided not to include Color as a field because we realized ahead of time that qualitative, subjective descriptions of color were likely to be useless for analysis. Not only would qualitative descriptions have subjective error and low resolution, but by imagining how we would try to analyze color data, we realized we would need to control for local lighting conditions. (At that time, we devised a color standard which can be included in the specimen photos that are already part of the *Amanita*BASE protocol, and anyone studying color can use a photo editing program to control for lighting conditions.) Had we not gone through this “backwards thinking” exercise, it seems likely that we would had jotted a useless note like “light olive green” when collecting each specimen. It is desirable to collect only as much data as is useful in order to minimize cognitive load in the field and minimize data entry labor, which also minimizes the opportunity to make data entry errors. The amount of required data for each specimen decreased with subsequent schema versions, and with the loss of tables when the metadata database became flat in *Amanita*BASE v3, Each time the decision of what to cut was guided by the logical relationships of the schema and the building and use of the database itself.

*Amanita*BASE has already borne scientific fruit. It is the direct foundation for Chapter 2 of this dissertation, Golan *et al.* (in review), an analysis of genet size in *A. phalloides*, and Harrow *et al.* (in prep), a population survey of *A. phalloides*’s characteristic MSDIN toxins (see Chapter 3). The results of these studies and all future studies based on the *Amanita*BASE whole genome dataset can be directly compared, because they include the same individuals. Our knowledge of populations Drake 2 and Drake 3, sampled in 2004, 2014, and 2015 and densely sequenced, will become very deep and thorough with time, allowing us to draw deep connections between the many fascinating questions in the genus *Amanita*. Thanks to Golan *et al.* (in review), we are confident that each of fully sequenced mushrooms in *Amanita*BASE is a genetic individual. When surveying the *Amanita*BASE genomes for their MSDIN toxin genes, we needn’t worry about each genome might not be unique. We needn’t even extrapolate from Golan *et al.*’s findings that all *A. phalloides* mushrooms they examined were individuals that the *A. phalloides* we are screening are all individuals, because we are working on the exact same proven individuals.

Unfortunately, investing in a database management system (DBMS) is not a trivial undertaking. Labs must weigh the costs of developing a database along with the benefits, and the process will require trial and error. *Amanita*BASE's metadata database is a perfect example of this. After the schema was designed, *Amanita*BASE's platform went from SQL to Michele Clamp's miniLIMS to Filemaker to Google sheets. Fortunately, however, all that work in developing a relational database is not lost just because, in the end, building and maintaining such a database was not worth the increase in data integrity. Relational database structure (schema) can be transferred across platforms. If a Filemaker or SQL *Amanita*BASE metadata database were ever attempted in the future, the schema is there and the data are organized accordingly. Today, *Amanita*BASE is not an all-in-one platform as originally envisioned, but it is an organized collection of data, and it is a far more complete collection of data than it would have been were it not for the way of thinking that relational database structure imposes.

Acknowledgments

I thank Allison Schultz, Tom Bruns, Jeffrey Grass, Zachary Muscavitch, Rod Tulloss, and Todd Mitchell for technical expertise and assistance. The computations in this study were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. Thanks to Harvard FAS Informatics, especially Allison Schultz, Michele Clamp, Aaron Kitzmiller, and Brian Arnold. Thanks to the National Science Foundation for funding grant 1021606 to Anne Pringle and NSF Graduate Research Fellowship 2013170866 to Holly Elmore. We collected at Tomales Bay State Park and Point Reyes under permits granted to Tom Bruns, Sydney Glassman, and Anne Pringle.

References

Allardice, S. (2015) *Programming Foundations: Databases (2015).* Video Course Series, LinkedIn Learning, first accessed 2016 <https://www.linkedin.com/learning/programming-foundations-databases-2015>

Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).

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

Chaib De Mares, M. *et al.* Horizontal transfer of carbohydrate metabolism genes into ectomycorrhizal Amanita. *New Phytol.* **205**, 1552–1564 (2015).

Codd, E. F. A Relational Model Data for Large Shared Data Banks. *Commun. ACM* 377–387 (1970)

Depristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–501 (2011).

Deshpande, V., Fung, E. D. K., Pham, S. & Bafna, V. Cerulean: a hybrid assembly using high thoughput short and long reads. *Algorithms Bioinform Lect Notes Com Sci.* **8126,** 349–350 (2013).

Elmore, M. H. *et al.* Clustering of two genes putatively involved in cyanate detoxification evolved recently and independently in multiple fungal lineages. *Genome Biol. Evol.* **7**, 789–800 (2015).

English, A. C. *et al.* Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS One* **7**, 1–12 (2012).

Golan J., Adams C. A., Cross H., Elmore M. H., Gardes M., Glassman S., Gonçalves S., Hess J., Richard F., Wang Y., Wolfe B., Pringle A. *Native and invasive populations of the ectomycorrhizal death cap Amanita phalloides are highly sexual but dispersal limited.* (in review) preprint: https://www.biorxiv.org/content/10.1101/799254v1

Gordon, D. *et al.* Long-read sequence assembly of the Gorilla Genome. *Science (80-.).* **0344**, 1–21 (2016).

Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci.* **108**, 1513–1518 (2011).

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

Hallen, H. E., Luo, H., Scott-Craig, J. S. & Walton, J. D. Gene family encoding the major toxins of lethal Amanita mushrooms. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 19097–101 (2007).

Hess, J. *et al.* Transposable element dynamics among asymbiotic and ectomycorrhizal amanita fungi. *Genome Biol. Evol.* **6**, 1564–1578 (2014).

Hobbie, E. A., Schubert, B. A., Craine, J. M., Linder, E. & Pringle, A. Increased C3 productivity in Midwestern lawns since 1982 revealed by carbon isotopes in Amanita thiersii. *J. Geophys. Res. Biogeosciences* **122**, 280–288 (2017).

Ippolite, C. (2018) *Filemaker: relational database design.* Video Course Series, Linkedin Learning, first accessed 2018 <https://www.linkedin.com/learning/filemaker-relational-database-design/>

Ippolite, C. (2018) *Learning Filemaker 17.* Video Course Series, Linkedin Learning, first accessed 2018 <https://www.linkedin.com/learning/learning-filemaker-17>

KateN. (howto) Discover variants with GATK - A GATK Workshop Tutorial. *GATK 3 User Guide*. (2016). <https://software.broadinstitute.org/gatk/documentation/article?id=7869>

Kajitani, R. *et al.* Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).

Keane, R. M. & Crawley, M. J. Exotic plant invasions and the enemy release hypothesis. *Trends Ecol. Evol.* **17**, 164–170 (2002).

Kohler, A. *et al.* Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. *Nat. Genet.* **47**, 410–415 (2015).

Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. 1–35 (2016). doi:10.1101/gr.215087.116.Freely

Krueger, F. Trim Galore: https://github.com/FelixKrueger/TrimGalore

Li, H. BFC: Correcting Illumina sequencing errors. *Bioinformatics* **31,** 2885–2887 (2015).

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

Li, Z., Wu, N., Gao, X., Wu, Y. & Oli, K. P. Species-level phenological responses to ‘global warming’ as evidenced by herbarium collections in the Tibetan Autonomous Region. *Biodivers. Conserv.* **22**, 141–152 (2013).

Ma, L. J. *et al.* Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. *Nature* **464**, 367–373 (2010).

MacLean, H. J., Kingsolver, J. G. & Buckley, L. B. Historical changes in thermoregulatory traits of alpine butterflies reveal complex ecological and evolutionary responses to recent climate change. *Clim. Chang. Responses* **3**, 1–10 (2016).

Nordberg, H. *et al.* The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.* **42**, 26–31 (2014).

Perez-Riverol, Y. *et al.* Ten Simple Rules for Taking Advantage of Git and GitHub. *PLoS Comput. Biol.* **12,** 1–11 (2016).

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E. & Jones, S. J. M. ABySS : A parallel assembler for short read sequence data ABySS : A parallel assembler for short read sequence data. 1117–1123 (2009). doi:10.1101/gr.089532.108

Pringle, A., Adams, R. I., Cross, H. B. & Bruns, T. D. The ectomycorrhizal fungus Amanita phalloides was introduced and is expanding its range on the west coast of North America. *Mol. Ecol.* **18**, 817–833 (2009).

Robbirt, K. M., Roberts, D. L., Hutchings, M. J. & Davy, A. J. Potential disruption of pollination in a sexually deceptive orchid by climatic change. *Curr. Biol.* **24**, 2845–2849 (2014).

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E. & Jones, S. J. M. ABySS : A parallel assembler for short read sequence data ABySS : A parallel assembler for short read sequence data. 1117–1123 (2009). doi:10.1101/gr.089532.108

Taylor, A. G. *SQL for Dummies*. Hoboken, New Jersey, Wiley & Sons, 2011.

Torvalds, L. Initial revision of "git", the information manager from hell. (2005).

Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, (2014).

Wang, J. R., Holt, J., McMillan, L. & Jones, C. D. FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinformatics* **19,** 1–11 (2018).

Warren, R. L. *et al.* LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience* **4**, (2015).

Wgsim: https://github.com/lh3/wgsim

Wolfe, B. E., Richard, F., Cross, H. B. & Pringle, A. Distribution and abundance of the introduced ectomycorrhizal fungus Amanita phalloides in North America. *New Phytol.* **185**, 803–816 (2010).

Wolfe, B. E., Kuo, M. & Pringle, A. Amanita thiersii is a saprotrophic fungus expanding its range in the United States. *Mycologia* **104**, 22–33 (2012).

Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* **6,** 1–9 (2016).

Chapter 2

The natural history of *HD1* and *HD2*, the heart of the Basidiomycetes HD mating locus, in collections of *Amanita phalloides*

Co-authors: Yen-Wen Wang, Jacqueline Hess, Anne Pringle

Abstract

This study is the first natural population survey of the Basidiomycete HD mating locus, famous for its high multiallelism (Raper 1966, Brown & Casselton 2001). There is high diversity at the *HD1* and *HD2* genes: 101 *HD1* and 80 *HD2* alleles in 86 dikaryotic *Amanita phalloides* individuals. 67/86 of the samples are two densely sampled and mapped invasive populations, Drake 2 and Drake 3, from Point Reyes National Seashore, California, USA. There are individuals from both sites that were collected from in 2004, 2014, and 2015. 65 distinct *HD1* and 46 distinct *HD2* alleles were found in 49 individuals in Drake 2, and 9 distinct *HD1* and 17 distinct *HD2* alleles found in 17 individuals in Drake 3. Some alleles occur at both sites, but most alleles are restricted to a single site. Similarly, some alleles persist over time, but the alleles present at each site in a given year are largely unique alleles that were not sampled again. There is no indication of new alleles emerging between 2004 and 2015. We conclude that the diversity of *HD1* and *HD2* alleles is ancient rather than continuously generated by ongoing negative frequency-dependent selection. The “ancient diversity” interpretation of the *HD1/HD2* multiallelism is supported by Yen-Wen Wang’s identification of de Bruijn assembly graph bubbles, areas of the assembly graph that were significantly diverged in sequence, indicating suppressed recombination. Methods for obtaining the sequences of *HD1* and *HD2*, particularly methods for phasing haplotypes, are discussed.

Introduction

The tetrapolar mating systems of Basidiomycete fungi are famously complex, and despite decades of high quality work, the population genomics of the mating loci have remained largely unexplored. This study is the first to sequence the *HD1* and *HD2* alleles of multiple natural populations over time. The *Amanita*BASE infrastructure (Chapter 1) provides spatially explicit population maps at multiple timepoints. It is possible to see which alleles occur together in the same specimen, and to track trends in the movement of alleles over space and through time. As a first step to investigating tetrapolar mating systems at a population level, I use genomics and phylogenetics to interrogate diversity across the sample and across time and space. I analyze the sequences of *HD1* and *HD2* from 86 sporocarps of *A. phalloides*, an ectomycorrhizal Agaricomycete. The sample includes breadth, with mushrooms collected between 1978 and 2015 and across Europe and North America, and depth, with a large subset coming from two invasive California populations within 100 meters of one another and collected from in 2004, 2014, and 2015.

There has long been interest and excellent work on mating in Basidiomycetes. Notably, John Raper developed *Schizophyllum commune* as a model system for classical genetics of mating in higher fungi and Lorna Casselton developed *Coprinopsis cinereus* for genetics and molecular biology of mating. There has even been some demographic work on mating compatibility in natural (Nieuwenhaus *et al.* 2013) and lab populations (Raper 1966). All of these systems require copious crosses and genetic manipulations. Experimental challenges aside, the genetic and molecular details of tetrapolar mating systems themselves can be hellishly complex. John Raper (1966) commented that, "many of the details [of mycosyngamology are] well-nigh ludicrous by any standards other than those imposed by the evolutionary history of the higher fungi." Therefore, I am forced to give a far from complete introduction to the topic of tetrapolarity. For a more thorough but still concise review, see Brown & Casselton (2001), extensively cited below.

Phylum Basidiomycota has an ancestrally tetrapolar mating system (Coelho *et al.* 2017), meaning that individuals have two mating loci and that compatible mates must have different alleles at both. The system is “tetrapolar” because there are four possible combinations of mating alleles when potential mates meet: same-same, same-different, different-same, different-different. Only different-different combinations are fully sexually compatible. The evolutionary logic behind this is thought to be to promote outbreeding. The parent mycelium is heterozygous at both loci. Therefore, ¼ of the spores in a tetrad will be sexually compatible-- half the chance of sib-mating allowed under bipolarity. High multiallelism at each locus drives up the chances that almost any unrelated germinating spores or monokaryons can mate while still reducing the chance of inbreeding.

The HD (*A*) locus encodes homeodomain transcription factor proteins and PR (*B*) locus encodes pheromones and receptors. The HD and PR loci work together in interleaving steps to allow the dikaryon life stage, in which two nuclear types live together in the same mycelium without fusing. The PR locus controls nuclear migration to the hyphal tips, where the HD locus controls conjugate nuclear division via clamp connections with cell division. The focus of this study is *HD1* and *HD2*, the genes encoding homeodomain proteins at the heart of the HD locus. HD1 is homologous to *Saccharomyces cerevisiae* MATα2 and HD2 is homologous to MATa1, the genes involved in fusion of haploid yeasts into a diploid. Just as MATα2 and MATa1 form a transcription factor that governs the diploid life stage, HD1 and HD2 form a heterodimer transcription factor that governs the dikaryon stage (Figure 2.1).

On the chromosome, *HD1* and *HD2* alleles appear as genetic “cassettes” (Brown & Casselton 2001) or modules, with the genes divergently oriented and recombination suppressed within the cassette (Figure 2.2). Genes within a cassette are not sufficient to initiate a dikaryon. Recombination within cassettes is prevented by the divergence of the sequences flanking the alleles-- like *MATα2* and *MATa1*, the allele cassettes have become idiomorphs. Some Basidiomycete species, such as *Schizophyllum commune* and *Coprinopsis cinereus*, have multiple *HD1-HD2* cassettes within the same locus or even duplicate HD loci.

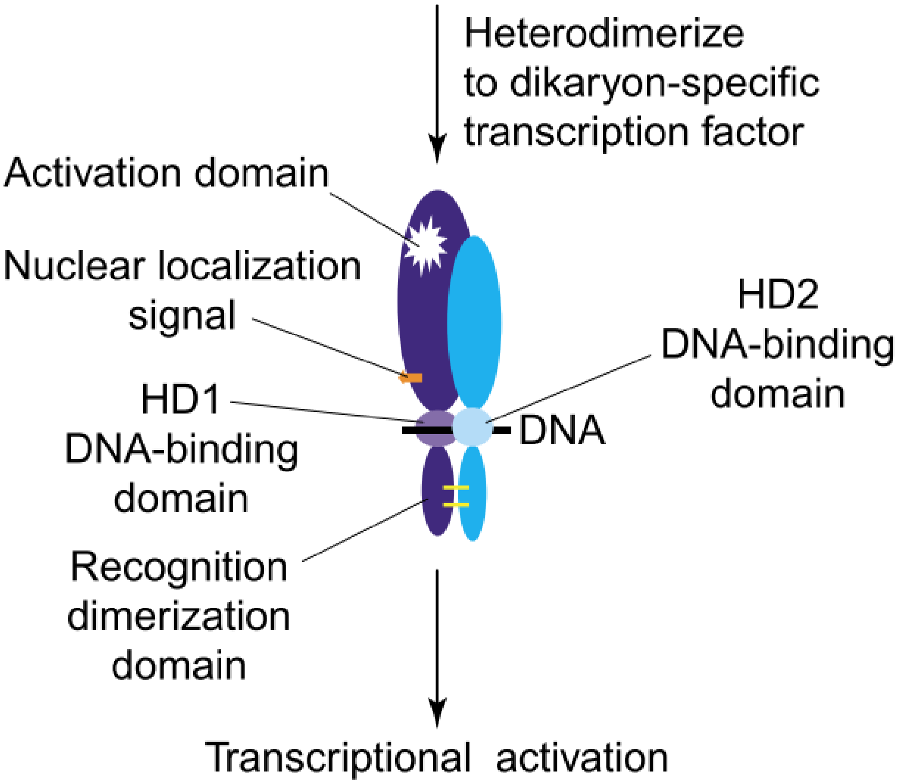


Figure 2.1: The heterodimerization of HD1 and HD2. HD1 has more domains, but HD2 has the stronger DNA-binding domain. Adapted from Brown and Casselton (2001).

As shown in Figure 2.1, HD1 has multiple domains: a C-terminal activation domain, a nuclear localization signal, the HD1-DNA-binding domain, and an N-terminal variable domain, consisting of a coiled-coil that dimerizes with HD2. HD2 has only the HD2-DNA-binding domain and the variable domain, although HD2's DNA-binding domain is the stronger of the two proteins. The high multiallelism at the HD locus comes from variation in the variable domains. Experiments show that a single amino acid substitution at the variable domain of either HD1 or HD2 within a cassette is sufficient to turn self (won't dimerize) into non-self (will dimerize)

(Kämper *et al.* 1995). This potential for variation has been exploited. Factoring in duplications of the *HD1-HD2* cassette, *C. cinereus* has 120 possible HD locus alleles (Brown & Casselton 2001) and *S. commune* has 400 (Raper 1966).

Therefore, an implication of our current picture is that diversity at *HD1* and *HD2* largely determines the mating dynamics of the population. In a large population with a high number of HD alleles, mating should barely be restricted at all outside of siblings, but in a recently bottlenecked population with lower HD allele diversity, mating would be expected to be more restricted. Raper and Casselton were able to estimate HD allele counts through experimental crosses, but they did not know the sequences and evolutionary histories of their alleles due to the technical limitations of the time, nor did they know the dynamics of pairing in natural populations. This study pioneers the use of next generation sequencing to study the HD locus in its natural context.

The question of how *HD1* and *HD2* alleles evolved, and possibly are still evolving, has also been waiting for the technology to study it. John Raper (1966) postulated that new alleles emerged by negative frequency-dependent selection similar to many angiosperm self-incompatibility systems. Either this is happening constantly, and the space of possible alleles is continually "rediscovered" by evolution or the majority of the evolution of *HD1* and *HD2* alleles occurred at some distant time in the past and is maintained by balancing selection, similar to alleles at the human major histocompatibility complex (Aguilar *et al.* 2004). This study is the first to have enough sequences from the same population of *HD1* and *HD2* to begin to distinguish between the continuous negative frequency-dependent selection hypothesis and the ancient diversity hypothesis on a demographic basis. Answering this question has implications for the manner in which tetrapolarity reduces inbreeding. If there's constant short-term evolution of new alleles because of the selective advantage to novel alleles, this could "cover up" inbreeding as well as preventing it, because the selection pressure for new alleles would be strongest when genome-wide diversity across the population was lowest. But if *HD1* and *HD2* alleles are ancient diversity, then having many alleles may still allow inbreeding but without “bending the rules” when diversity is low.

This is the first study to catalogue the diversity of *HD1* and *HD2* alleles and investigate their evolutionary history at a population level. This study boasts massively parallel high quality sequence data, sequences taken from invasive, spatially-explicit populations of *Amanita phalloides* over three timepoints (2004, 2014, 2015) and experiments with new techniques for sequencing and assembling or aligning the *HD1* and *HD2* alleles. Having surveyed the diversity of *HD1* and *HD2*, I ask, what drives that diversity, and what are its dynamics?

Methods

All scripts are available in this chapter’s Github (https://github.com/elmoremh/HD1-HD2-natural-history) and that of Amanita*BASE (github.com/elmoremh/Amanita-Population-Genomics).*

Specimens

The sequences in this study came from 86 *A. phalloides* sporocarps. 66 of these were from the populations Drake 2 and Drake 3 in Tomales Bay State Park, Point Reyes National Seashore, California, USA, where *A. phalloides* is invasive. These 66 mushrooms came from three different collecting trips in 2004, 2014, and 2015, and each was mapped using GPS coordinates and a field compass (Supplementary Figures S2.1-7 are population maps and a satellite image of the sites). 11 mushrooms were collected in Portugal in 2015 and followed the same mapping procedure. 8 were older specimens, going as far back as 1978, from across the native range Europe.

Sequencing

86 *A. phalloides* specimens were sequenced with Illumina HiSeq 2500 and two of those, 10511 and 10721, were sequenced with PacBio as well. The Illumina HiSeq sequencing was

250 bp paired-end reads with a 550 bp insert (with the exception of 10801, 10169, 10170, 10171, 10277, 10003, 10004, 10007, 10010, 10016, 10018, 10019, 10023, each of which had a 350-bp insert) prepared as IntegenX dual-index libraries. The mean sequencing depth of each of the samples ranged from 10.56 to 150.86 (and see Aggregated FastQC in Supplementary Materials). Two of the above 89 mushrooms, *A. phalloides* 10721 (USA, California, Drake 3, 2015) and *A. phalloides* 10511 (Portugal, São Jacinto, Dunas de Mira) were also sequenced using a PacBio Sequel platform at the University of Wisconsin-Madison Biotechnology Center, with a 20-kb single library per specimen and yielding average read sizes of 14,833 and 14,935 for specimens 10721 and 10511, respectively. 10511, which became the reference assembly, had raw PacBio coverage of 47x with N50 read length of 6,310 bp.

Quality control, assembly, & alignment

Trimming was performed on all 2015-sequenced specimens with the program Trim Galore v0.4.5 (Kruger, https://github.com/FelixKrueger/TrimGalore). Reads that became shorter than 100 bp after trimming and those with a quality score less than 30 were discarded. Adapter trimming was set to the highest stringency, 1, meaning a single nucleotide of overlap with the adapter sequence was trimmed from the read. Unpaired reads were retained, though they did not end up being used in either the assemblies or the alignments. Raw and trimmed reads will be available in NCBI BioProject PRJNA565149 in September 2020.

Alignments were run against the 10511 reference assembly using the Burrows-Wheeler Alignment software, BWA, mem algorithm with default parameters (Li & Durbin 2009), in the course of the GATK best practices pipeline (see **Variant Calling**).

Assemblies

Hybrid assemblies, incorporating both PacBio and Illumina HiSeq reads, of specimens 10721 (California) and 10511 (Europe) were completed by Jacky Hess. Extensive troubleshooting was performed in the course of generating the 10511 assembly, and then the workflow arrived at for 10511 (below) was applied to 10721. The workflow began with pre-processing: trimming and filtering with Trimmomatic v 0.35 (Bolger *et al.* 2014) (with the following parameters: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 CROP:245 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:25 MINLEN:100), Illumina reads sequencing error correction with BFC (Li *et al.* 2015) and PacBio reads sequencing error correction with FMLRC (Wang *et al.* 2018). Dr. Hess proceeded to test several assemblers: CANU (Koren *et al.* 2016) and FALCON/HGAP4 (Gordon *et al.* 2016) are PacBio-only assemblers; ABySS (Simpson *et al.* 2009), Platanus (Kajitani *et al.* 2014), and AllpathsLG (Gnerre *et al.* 2014) are Illumina-only (although to meet the library prep criteria for Allpaths LG, Illumina libraries with the proper insert size were simulated using the program wgsim (https://github.com/lh3/wgsim) on PacBio reads); and SPAdes (Bankevich *et al.* 2012), DBG2OLC (Ye *et al.* 2016), and Cerulean (Deshpande *et al.* 2016) are hybrid assemblers. Based on metrics like contig number and lengths, SPAdes and Platanus were not considered further. Scaffolding was performed with LINKS (Warren *et al.* 2015) and each assembly was "polished" with the gap-filler PBJelly (English *et al.* 2012) and the base and indel correction program Pilon (Walker *et al.* 2014). The polished assemblies were analyzed with QUAST (Gurevich *et al.* 2013), BUSCO (Simão *et al.* 2019), and REAPR (Hunt *et al.* 2013) and evaluated on completeness of eukaryotic single copy complement, completeness of eukaryotic duplicated gene content, fragmentation, missing sequence, assembly size as percent of genome size, number of scaffolds, scaffold N50/NG50, and scaffold L50. Allpaths LG was chosen as the preferred assembler because the Allpaths LG assembly had the highest ranking in the greatest number of evaluation criteria (see Assembly Strategy in Supplementary Materials).

Sequences are deposited under NCBI Bioproject number PRJNA565149 and will be released September 2020.

Variant calling

SNPs and Indels were called using the Genome Analysis Toolkit (GATK) v3.8-0-ge9d806836 software (Depristo *et al.* 2011) and following the GATK best practices (KateN 2016) as well as is possible for a non-model system, with help from Allison Shultz's Github page (github.com/ajshultz/whole-genome-reseq). I began by aligning the Illumina reads from each sample to the 10511 hybrid assembly using BWA (Li & Durbin 2009). Mapping rates ranged from 20.0% - 95.3%, with a median mapping percentage of 86.1%. Specimen 10003 had a low mapping rate and did not cooperate with the GATK workflow, so it was removed from the joint-calling cohort. The mapping rate of 10511's Illumina reads to the 10511 hybrid assembly was 93.8% (see Alignment and Deduplication metrics in Supplementary Materials). Following the steps, I marked duplicate reads, but I did not recalibrate base scores based on known variants because there were no known variants for *A. phalloides.* The GATK program Haplotypecaller makes variant calls jointly on all the samples, generating a GVCF file that contains a record of all sites of all the genomes, whether invariant or variant. The program GenotypeGVCF creates the raw VCF files containing only the SNPs and Indels. Our samples contained 1831629 raw variants.

Again, due to the lack of known variants in *A. phalloides* for comparison, the raw variants were hard-filtered according to the default parameters of the GATK VariantFiltration program. Because of the generally high coverage (~50x) of our sample set, hard filtering with default parameters was not expected to bias the filtered variants.

Phasing

The mushroom cap tissue from which we derived our sequencing material is dikaryotic (functionally diploid). In order to distinguish the two alleles of *HD1* and *HD2* in each mushroom, it was necessary to phase the GATK-called variants into haplotypes. WhatsHap (Martin *et al.*

2016) took the GATK all samples vcf and individual sample read alignments to the reference assembly as input for read-backed phasing.

Extracting alleles of *HD1* and *HD2* as alternate references

HD1 and *HD2* were identified in the 10511 (Portugal) assembly by the tblastn (Altschul *et al.* 1990) protein-to-nucleotide search using queries from the *Amanita muscaria* Koide v1.0 assembly hosted by the Joint Genome Institute of the Department of Defense (Kohler *et al.* 2015). Protein ID 65150 was used for mitochondrial intermediate peptidase (MIP) (best hit E value = 0), protein ID 19973 for HD1 (best hit E value = 9.00E -37), and protein ID 56764 and transcript ID 1084176 for HD2 (best hit E value = 3.00E -15) (tblastn results table in Supplementary Materials). The low expect values of the BLAST hits indicated these were the correct genes and that they were single copy, as multiple hits were simply slightly different ways of aligning the query over the same coordinates of the subject sequence. The positions of the genes also confirmed their identity-- *HD1* and *HD2* were divergently oriented and *MIP* was directly adjacent to them (Figure 2.2).

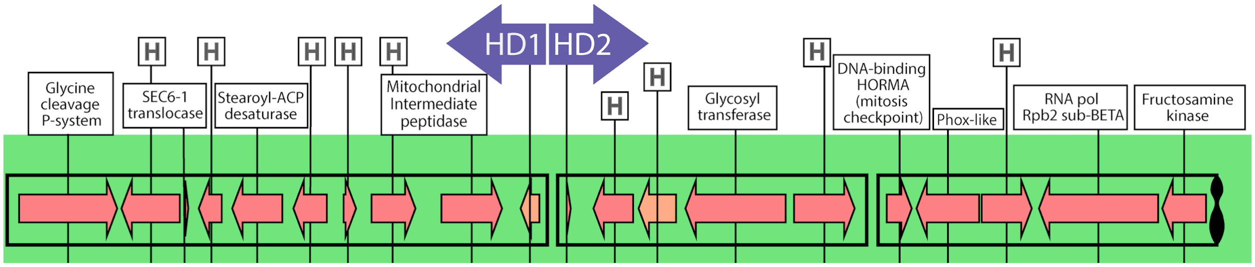


Figure 2.2: The core *Amanita* HD locus is centered on *HD1* and *HD2*, which are divergently oriented and always directly adjacent to MIP. *HD1* and *HD2* are surrounded by genes involved in mating and dikaryosis. “H” indicates a hypothetical protein. Adapted from Chapter 3: section III.

I did not attempt *de novo* whole genome annotations. I attempted to use MAKER (Cantarel *et al.* 2008) web version to annotate only several kb on either side of the area identified by BLAST, but found it impossible to use existing annotation profiles (other fungi were far too distant). I attempted to refine the boundaries of the genes BLAST had identified by making a STAR (Dobin *et al.* 2013) splice-aware alignment of our transcriptome from mushroom 10721 (California) to our reference genome 10511, but no reads mapped to the BLAST area even in a maximally permissive alignment (STAR_map_10721_RNA_to_10511_5-1-19.sh in https://github.com/elmoremh/HD1-HD2-natural-history). *HD1, HD2*, and *MIP* were apparently not expressed in that tissue (pileus) at the time we harvested it in the field. The distribution of called variants did not visually suggest the boundaries of the genes.

In the absence of other evidence and because the BLAST results created a plausible picture, the coordinates identified for BLAST, Contig87:398637-399359 and Contig87:398113-397802 (rev), were used as the start and end points of *HD1* and *HD2*, respectively. Variants within these coordinates were extracted from the master vcf of all positions in 89 individuals (to create HD1-HD2_vcf.xlsx in Supplementary Materials) using vcftools position filtering (Danecek *et al.* 2011).

The variants from each individual were imposed onto the reference (10511) assembly *HD1* and *HD2* sequence to obtain facsimiles of that individual's sequences, known as "alternate references," at those sites using bcftools-consenus in the bcftools suite (Danecek *et al.* 2011). This method was chosen because it takes advantage of the superior positional information of the reference assembly and superior accuracy of the variant calls by GATK compared to Illumina-only *de novo* assemblies of the reads from each individual separately.

Because the sequenced individuals are functionally diploid (dikaryotic), there are two alleles of each gene, *HD1* and *HD2*. However, because phasing of the haplotypes across this region was imperfect, some sites could not be assigned to one or the other haplotype in the alternate references. After corresponding with me about this issue, Petr Danecek, creator of bcftools, added the option `--haplotype 1pIu` and `--haplotype 2pIu` to bcftools development version 1.9-197-g4e51a29. The new option assigns phased alleles to the proper haplotype and gives an IUPAC ambiguity character at unphased heterozygous sites. Using bcftools with this new option, a list of alleles was generated with IUPAC ambiguity codes at unphased heterozygous sites. Because the ambiguity codes can distort tip weights in phylogenies, and because each allele contained either one variant or the other rather than some probability of each, for downstream analyses, all IUPAC codes were converted to “N” via Find and Replace (“alignments” folder in Supplementary Materials).

Generating phylogenies

Because the sequences were all alternate references, they were already aligned during the GATK pipeline. The only change necessary to use the aligned *HD1* and *HD2* sequenced for phylogenetics was to convert all the IUPAC ambiguity codes (W, Y, etc.) to Ns to avoid biasing the phylogeny, which was done via Find and Replace in TextWrangler. Maximum likelihood phylogenies were obtained using RAxML v8.2.11 (Stamatakis *et al.* 2014) in raxmlHPC mode with the GTRGAMMA model of rate heterogeneity and default parameters (script on Github).

Searching de Bruijn assembly graphs for *HD1* and *HD2* alleles

Yen-Wen Wang assembled each of the 86 genomes *de novo* with SPADES v3.13.1 (Bankevitch *et al.* 2012). He identified *HD1* and *HD2* in the de Bruijn graph of each assembly by nucleotide BLAST using the sequences of *HD1, HD2*, and their intergenic region in the 10511 reference assembly as the query within the assembly graph visualization tool Bandage v0.8.1 (Wick *et al.* 2015). If the sequence of two alleles in a diploid (dikaryotic) assembly are distinct enough they will not collapse into one contig, but be assembled into two different “unitigs” which are contiguous with at least one unitig at either end, forming a “bubble” (Iqbal *et al.* 2012). Wang hypothesized that, under the recent and ongoing negative frequency-dependent selection hypothesis, alleles in the same population should be fairly closely related, and so should not form a de Bruijn bubble. However, under the ancient diversity hypothesis, the alleles diverged long ago, and so should form a bubble.

The unitigs containing HD hits were annotated for genes with webAUGUSTUS (Hoff & Stanke 2013) and the predicted genes searched using HD1 and HD2 annotations from *Coprinopsis cinerea* (Genbank accession numbers XP_001829154.1 and XP_001829153.1) as BLASTp v2.8.1 queries. Protein sequences of the genes from the 11 Portuguese mushrooms and 66 Californian mushrooms (Drake 2 and Drake 3) were aligned with MAFFT v7.407 (Katoh *et al.* 2013). Phylogenetic trees were built using the best evolution model determined by ProtTest v 3.4.2 (Darriba *et al.* 2017) with RAxML ver. 8.2.9 (Stamatakis *et al.* 2014).

Determining patterns over space and time

The analysis of diversity over space and time is restricted to the populations, Drake 2 and Drake 3, that were sampled in 2004, 2014, and 2015. The two sites are approximately 94 meters apart and on opposite sides of a two lane road. Commercially available GPS is only accurate to within 5 meters, so we use a combination of GPS and landmarks to determine the center point where the field compass is placed. Because of this, the exact center point of the populations shifts slightly each year. Maps of mushrooms collected from both populations during all three years and a satellite image of the sites are in Supplementary Materials.

Using the “exactly identical sequence” warnings given by the RAxML in the creation of the phylogenies (see identical_alleles_Ns.txt worksheet in Supplementary Materials), I worked out which alleles were identical to each other and which were unique. Alleles with exactly the same sequence were placed into numbered “allele identity groups.” Allele identity groups were determined strictly. Only alleles with exactly the same nucleotide sequence were considered to be the same. The list of allele identity groups and unique alleles was made into an Excel spreadsheet (identical_HD1_HD2_alleles.xlxs in Supplementary Materials) and filtered to determine how many alleles were present in specimens collected from Drake 2 or Drake 3 and during which years.

The phylogenies were labelled with the SpecimenIDs, locations, and years collected of each tip and examined visually for patterns.

Results

A. phalloides HD1-HD2 cassette appears to be single-copy

In both the 10511 reference assembly and when collaborator Yen-Wen Wang (Pringle Lab, University of Wisconsin-Madison) searched for *HD1* and *HD2* in *de novo* assembly graphs, *HD1* and *HD2* appeared to be present as a single cassette (in contrast to, for example, *C. cinereus*, which has three (Brown & Casselton 2001)). Wang occasionally got three hits rather than two, but believes this was due to low coverage or poor quality sequencing rather than duplication of the *HD1-HD2* cassette on the chromosome.

Phasing (and caveats)

Phasing was not fully successful. According to the GATK joint-calling on the entire cohort of individuals, 250 sites out of *HD1*'s 734 nucleotides are variable. *HD1* has an average of 15.1 unphased heterozygous sites per 734 bp allele (6% of variable sites and 2% of sites overall). According to the GATK joint-calling on the entire cohort of individuals, 96 sites out of *HD2*'s 316 nucleotides are variable. *HD2* has an average of 4.1 unphased heterozygous sites per 316 bp allele (1.3% of sites). *HD2* has an average of 2.1 '-' gap characters indicating indels per 316 bp allele (0.7% of sites). (See also HD1-HD2_vcf.xlsx in Supplementary Materials.)

Because some heterozygous sites were unable to be phased, in the consensus sequences they appear as IUPAC ambiguity codes (see Supplementary Materials IUPAC alignment) or as ‘N’ unknown nucleotide characters (see Supplementary Materials Ns alignment).

de Bruijn graph of *HD1* and *HD2* alleles shows unlinked unitigs and de Bruijn bubbles

Among the 86 samples, 11 samples have one hit, 74 samples have two hits, and 2 samples have three hits for *HD1* and *HD2.* In three hit samples, at least one unitig is truncated and not linked to other unitigs, implying these samples don't have three alleles but the sequencing quality of these samples are not good enough to assemble the whole alleles. However, duplication of the *HD1-HD2* cassette is well-known in other Agaricomycetes such as *Coprinopsis cinerea* (Brown & Casselton 2001), so truncated duplications or perhaps even higher *HD1-HD2* copy number are not out of the realm of possibility. Some of the samples with two hits formed the "bubbles" while others might have one opening or are not linked at all (see Supplementary Figure S2.8 for examples of linked and unlinked unitigs). At least nine of the 11 single hit samples appear to be haploid (monokaryotic) or homothallic based on minor allele frequency graphs. The Pringle Lab is investigating this intriguing finding, but for now it does not appear that single hit samples have defied tetrapolarity by mating with a mycelium with the same *HD1* and *HD2* alleles. The presence of unitigs and de Bruijn bubbles weighs in favor of the ancient diversity hypothesis.

Among the Drake 2, Drake 3, and 2015 Portuguese samples, a total of 144 *HD1* and 147 *HD2* genes were predicted. Three *HD1* genes were not annotated due to truncated unitigs. In addition, four *HD1* genes and four *HD2* genes are not predicted at full length due to truncated unitigs as well. There are 28 *HD1* alleles and 27 *HD2* alleles in the assemblies, much less than the 48 *HD1* and 31 *HD2* that are estimated in the same samples by the phased variant-alternate reference approach.

Allele diversity

Following the stated methodology of obtaining sequences and considering only identical sequences to be the same allele, the sequenced sample contained 101 distinct *HD1* alleles and

80 distinct *HD2* alleles (out of a total of 172 sequences each). *HD1* has an average of 180/86 = 2.1 '-' gap characters indicating indels per 734 bp allele (.3% of all sites). *HD2* has an average of 180/86= 2.1 '-' gap characters indicating indels per 316 bp allele (0.7% of sites). A list of alleles is available in Supplementary Materials (“alignments” folder) along with a spreadsheet giving the groups of identical alleles (“Identical” alleles spreadsheet).

The evolutionary relationships between the alleles can be seen on the RAxML phylogenies (Figure 2.3 and Figure 2.4). The “ancient diversity” interpretation of the *HD1/HD2* multiallelism is supported by the fact that there are Portuguese alleles nestled within the large California allele identity groups (Figure 2.3).

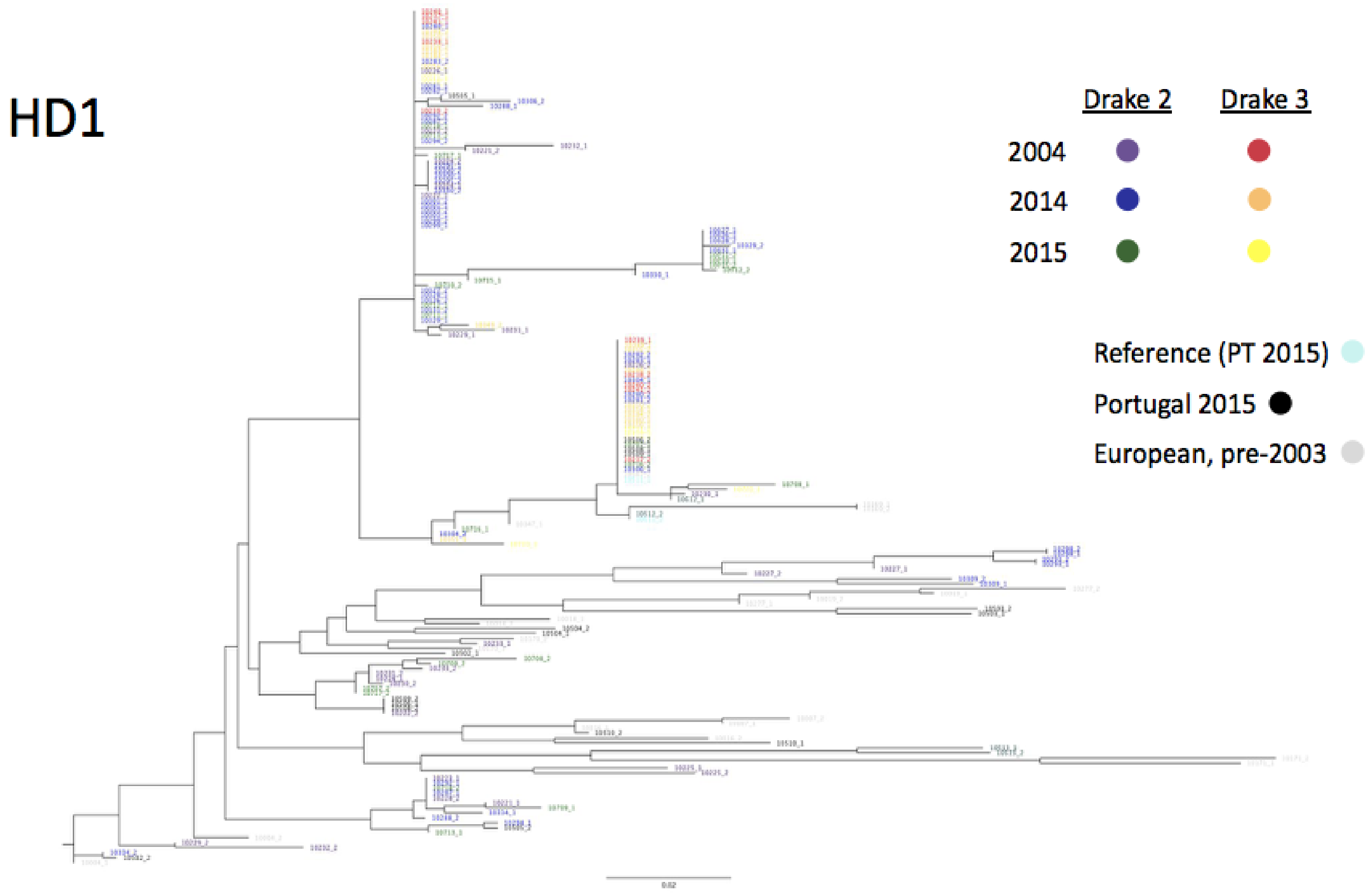


Figure 2.3: A maximum likelihood phylogeny of *HD1* alleles from all 86 sequenced *A. phalloides* individuals.

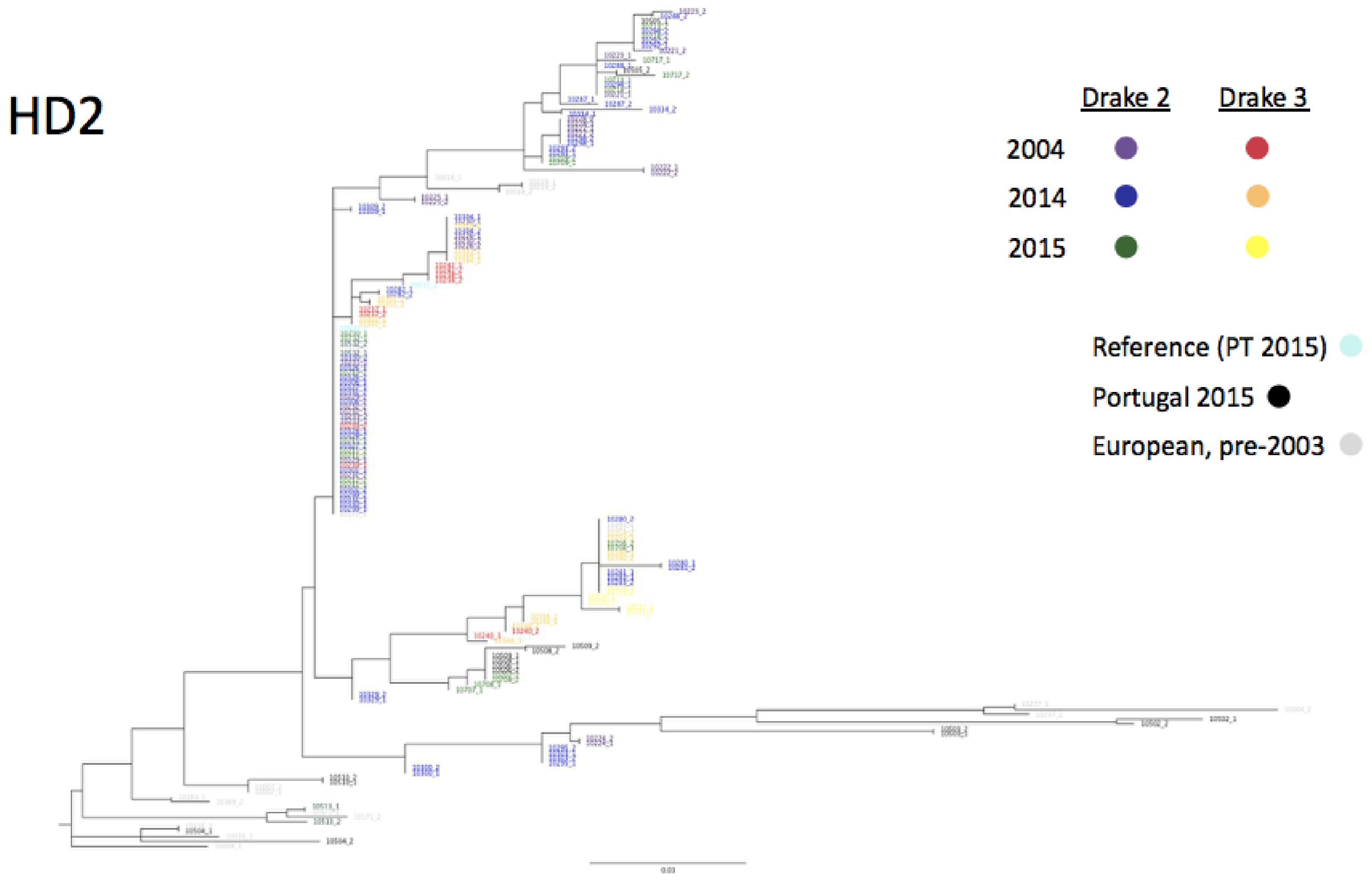
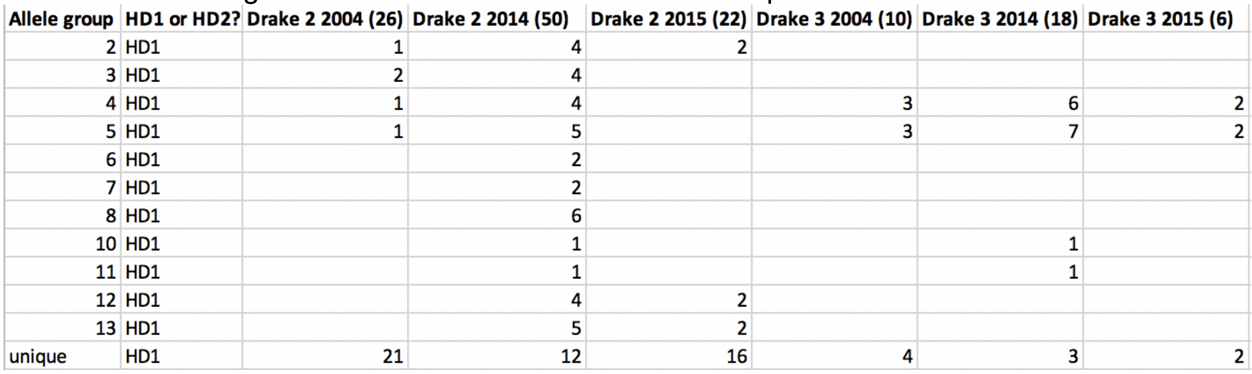


Figure 2.4: A maximum likelihood phylogeny of *HD2* alleles from all 86 sequenced *A. phalloides* individuals.

Diversity over space (Drake 2 and Drake 3)

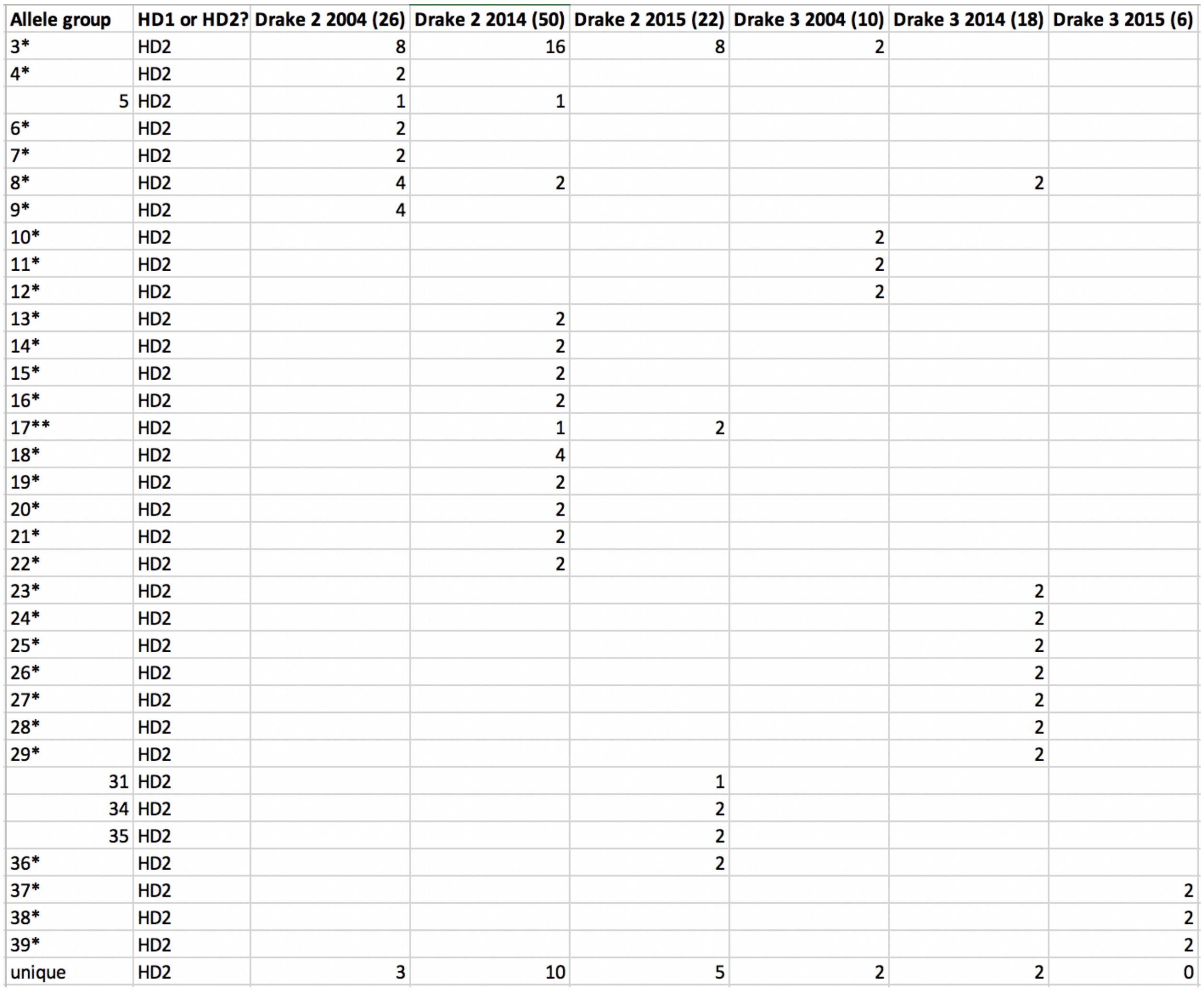
Over all three timepoints, 2004, 2014, and 2015, there were 65 distinct *HD1* and 46 distinct *HD2* alleles found in 49 individuals in Drake 2 and 9 distinct *HD1* and 17 distinct *HD2* alleles found in 17 individuals Drake 3. *HD1* allele groups 4, 5, 10, and 11 and *HD2* allele groups 3 and 8 are found in both Drake 2 and Drake 3.

Table 1: HD1 alleles of different allele groups present at Drake 2 and 3 in 2004, 2014, and 2015. Total number of *HD1* alleles present at a site each year is given in parentheses next to the column heading. Remember there are two *HD1* alleles per individual mushroom.



Allele group	HD1 or HD2?	Drake 2 2004 (26)	Drake 2 2014 (50)	Drake 2 2015 (22)	Drake 3 2004 (10)	Drake 3 2014 (18)	Drake 3 2015 (6)
2	HD1	1	4	2			
3	HD1	2	4				
4	HD1	1	4		3	6	2
5	HD1	1	5		3	7	2
6	HD1		2				
7	HD1		2				
8	HD1		6				
10	HD1		1			1	
11	HD1		1			1	
12	HD1		4	2			
13	HD1		5	2			
unique	HD1	21	12	16	4	3	2

Table 2: *HD2* alleles of different allele groups present at Drake 2 and 3 in 2004, 2014, and 2015. Total number of *HD2* alleles present at a site each year is given in parentheses next to the column heading. Remember there are two *HD2* alleles per individual mushroom. An asterisk (*) next to an allele group number indicates that all members of this group have identical co-alleles, which may indicate spurious “identity” due to phasing problems. A double-asterisk (**) indicates that some individuals in the group have identical co-alleles but other members do not.



Allele group	HD1 or HD2?	Drake 2 2004 (26)	Drake 2 2014 (50)	Drake 2 2015 (22)	Drake 3 2004 (10)	Drake 3 2014 (18)	Drake 3 2015 (6)
3*	HD2	8	16	8	2		
4*	HD2	2					
5	HD2	1	1				
6*	HD2	2					
7*	HD2	2					
8*	HD2	4	2			2	
9*	HD2	4					
10*	HD2				2		
11*	HD2				2		
12*	HD2				2		
13*	HD2		2				
14*	HD2		2				
15*	HD2		2				
16*	HD2		2				
17**	HD2		1	2			
18*	HD2		4				
19*	HD2		2				
20*	HD2		2				
21*	HD2		2				
22*	HD2		2				
23*	HD2					2	
24*	HD2					2	
25*	HD2					2	
26*	HD2					2	
27*	HD2					2	
28*	HD2					2	
29*	HD2					2	
31	HD2			1			
34	HD2			2			
35	HD2			2			
36*	HD2			2			
37*	HD2						2
38*	HD2						2
39*	HD2						2
unique	HD2	3	10	5	2	2	0

Diversity over time (Drake 2 and Drake 3)

Some large allele groups persist over time (Table 1, Table 2), but the alleles present at each site in a given year are largely unique alleles that were not sampled again. There are some large allele identity groups that are present at both sites and across time (*HD1* groups 4 and 5, *HD2* groups 3 and 8), but most of the identity groups in *HD2* are probably artifacts of sequencing and phasing, as they only contain 2 co-alleles or a few sets of co-alleles. *HD1* groups 10 and 11 are present at Drake 2 and 3 only in 2014, perhaps because 2014 saw many more fruiting bodies at both sites.

There is no indication of new alleles emerging between 2004 and 2015.

Differences between *HD1* and *HD2*

In space

For both *HD1* and *HD2*, no one or few alleles dominates the populations. Here I identify 90 unique alleles and 15 allele identity groups for *HD1* and 46 unique alleles and 39 unique allele identity groups for *HD2*, and the true number of unique alleles may be higher. Even the largest allele identity group was not present in both sites for all years in either *HD1* or *HD2.* The greatest pattern is most likely an artifact of sequencing and phasing: HD2's allele identity groups most often have 2 members, and they are co-alleles. I believe that those alleles are all, in fact, unique alleles. I suspect that *HD1* allele group 1, which consists of the two alleles found in 10169, is also an artifact.

In time

By visually inspecting the phylogenies, HD2 alleles in Drake 2 and 3 appear later in the tree and are better separated from the European samples. In *HD1*'s tree, alleles from Drake 2 are found in the earliest branches. This may reflect the fact that, unlike *HD2*, HD1 has conserved functional domains other than the variable domain and the DNA-binding domain.

Suppressed recombination?

There may be indications in this data of recombination between *HD1* and *HD2*, but it's difficult to say for sure because of 1) the phasing difficulties, which make it difficult to say which HD1 allele from a given individual is associated with which HD2 allele from that individual and 2) because, as is suggested by the phylogenies (Figure 2.3 and Figure 2.4), the two genes may have different evolutionary rates that account for part of the discrepancies in their phylogenies.

However, individuals that have “identical” (due to missing data) HD1 alleles have, in some cases, quite divergent *HD2* alleles (Supplementary Figure S2.9). This might indicate recombination, but if the alleles represent deep standing variation then a combination of different rates of evolution between the two genes and very rare recombination events in the past. This question could be resolved with better phasing and a molecular clock analysis.

By comparing phylogenies of *HD1* and *HD2* from *de novo* assembly graphs, Yen-Wen Wang identified nodes between which recombination events may have occurred (Figure 2.5).

A.

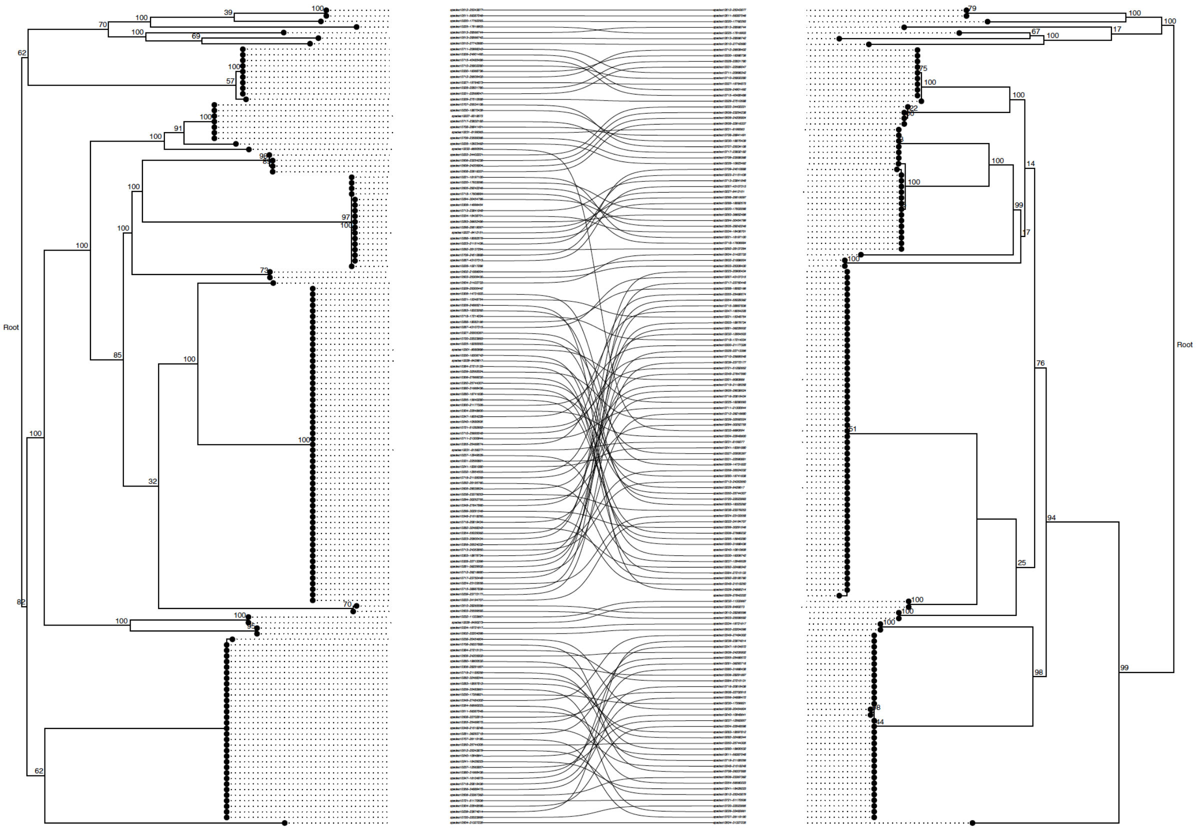


Figure 2.5: A) A co-phylogeny between *HD1* (left) and *HD2* (right) shows linkage over time. Note that the trees contain large polytomies which cause some crossed lines that are not due to differences in phylogeny.

B.

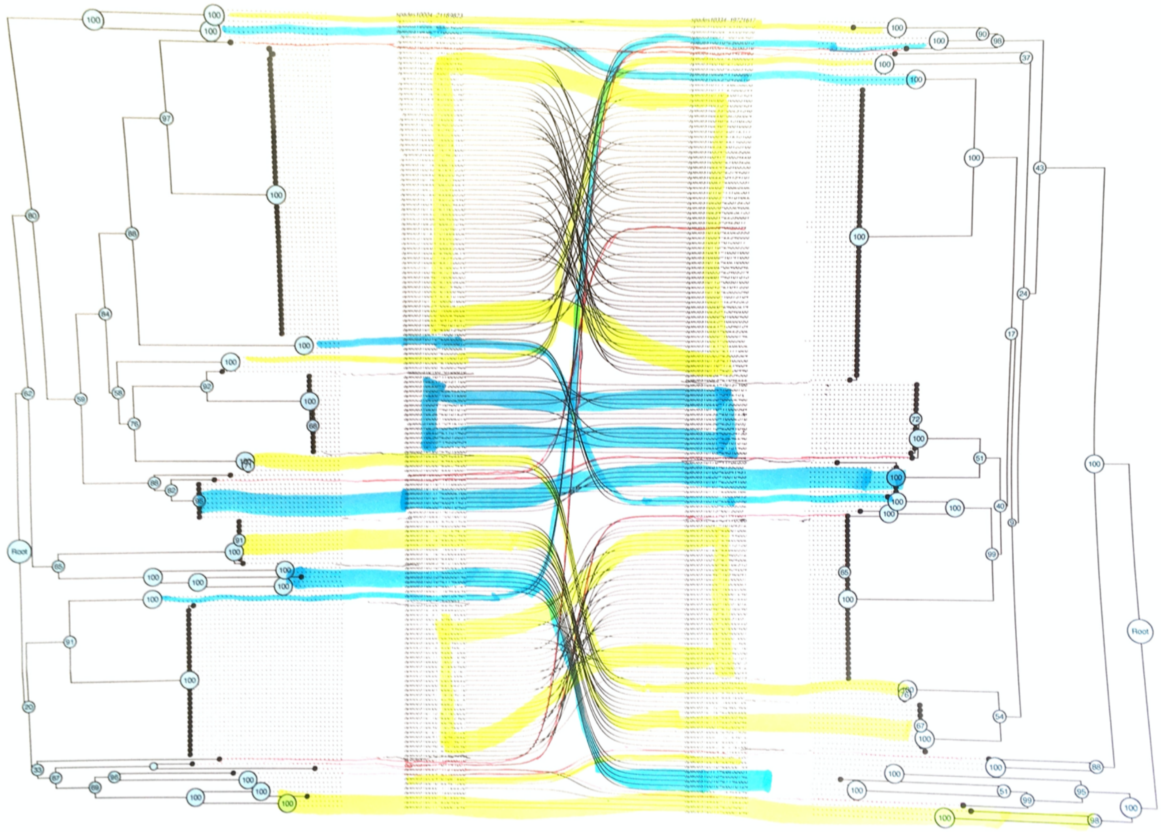


Figure 2.5 (Continued): B) Yellow and blue lines show how polytomies in the *HD1* and *HD2* trees line up, and red lines indicate possible recombination events. Note that the trees contain large polytomies which cause some crossed lines that are not due to differences in phylogeny.

Discussion

This study is a first-of-its-kind natural population survey of Basidiomycete tetrapolar mating loci. I found high diversity at the *HD1* and *HD2* genes: 101 *HD1* and 80 *HD2* alleles in 86 dikaryotic individuals, 65 *HD1* and 46 distinct *HD2* alleles found in 49 individuals in California population Drake 2, and 9 distinct *HD1* and 17 distinct *HD2* alleles found in 17 individuals in California population Drake 3. Some alleles occur at both sites but most alleles at each site are unique (Table 1, Table 2). Similarly, some large allele groups persist over time (Table 1, Table 2), but the alleles present at each site in a given year are largely unique alleles that were not sampled again. There is no indication of new alleles emerging between 2004 and 2015. The "ancient diversity" interpretation of the *HD1/HD2* multiallelism is supported by Yen-Wen Wang's identification of de Bruijn assembly graph bubbles at the site of *HD1, HD2*, and sometimes both alleles in *de novo* assemblies of individual specimens, and is further supported by the fact that there are Portuguese alleles nestled within the large California allele identity groups (Figure 2.3, Figure 2.4). There is also a possible indication that recombination is not as suppressed between *HD1* and *HD2* as has been the dogma (Figure 2.5)

Other Basidiomycetes that have had the number of their mating alleles determined by genetic crosses have found HD mating locus alleles in the hundreds. Raper (1966) determined that *Schizophyllum commune* had at least 400 alleles at the HD mating locus (there are multiple *HD1-HD2* cassettes on the chromosome and this number includes different combinations), the region containing *HD1* and *HD2*. If there were 101 distinct *HD1* and 80 distinct *HD2* alleles in a sample of 86 *A. phalloides* (172 alleles total of each of *HD1* and *HD2*), it stands to reason that *A. phalloides* populations throughout the world have hundreds of HD locus alleles as well. (Yen-Wen Wang found a smaller but still considerable number of *protein* alleles in the Drake 2, Drake 3, and 2015 Portuguese populations, 27 HD1 and 28 HD2 in 78 individuals compared to 84 HD1 and 65 HD2 alleles in those same individuals by the phased variant-alternate references method.) If *HD1* and *HD2*'s diversity is ancient are they are under balancing selection, they may be very well-dispersed and well-sampled even in recent founder populations, so this study may have sampled a fair bit. Some allele groups were recovered in both European individuals and invasive Californian populations (*HD2* groups 3 and 31) which indicates that there are not staggering numbers of HD mating locus alleles such that our sample size would not redraw the same allele from different populations. However, the small number of multiply-sampled alleles leads me to believe that our sample is not close to saturation.

My results are consistent with experimental findings that only a few substitutions are sufficient to make alleles compatible (Kämper *et al.*1995). Co-alleles, alleles that were recovered from the same mushroom, are frequently the most closely related on the tree (see coallele_flipbook.pptx in Supplementary Materials). Closely related co-alleles (with the caveat that there were phasing issues) support the idea that new alleles emerge by few mutations--and remember, these are nucleotide phylogenies.

In both cases, no one or few alleles dominates the populations. Here I identify 90 unique alleles and 15 allele identity groups for *HD1* and 46 unique alleles and 39 allele identity groups for *HD2*, but the true number of unique alleles may be higher or lower. Because of incomplete phasing, some variants that distinguished co-alleles were lost, which would lead to underestimating the number of all alleles, and the phased variant-alternate references method may have recombined variants from the same set of alleles into chimeric sequences that inflated the estimate of unique alleles. Even the largest allele identity group was not present in both sites for all years in the case of *HD1* and *HD2*. The greatest pattern in space is most likely an artifact of sequencing and phasing: *HD2*'s allele identity groups most often have 2 members, and they are co-alleles. I believe that those alleles are all, in fact, unique alleles. Even for *HD1*, I expect the phasing problems to disproportionately affect the variable regions and so turn closely related alleles into "identical" alleles.

The demographic results are consistent with both deep standing variation and ongoing evolution of alleles by negative frequency-dependent selection and ancient diversity maintained by balancing selection, but the *de novo* assembly de Bruijn bubbles in *HD1* and *HD2* corroborate the cassette model of suppressed recombination and tip our interpretation in favor of ancient diversity. Knowing the node ages, or putting a molecular clock on the phylogeny, would help to resolve the question, because the key issue here is not exactly how the alleles evolved but when that evolution occurred and how the alleles are maintained. The distribution of alleles from the same sites and years and the distribution of co-alleles across the phylogenies indicate that the entire sample of alleles is well-distributed among sites and years. In other words, there is no evidence of clustering of *HD1* and *HD2* alleles in space or time, or even by continent of origin. This is in stark contrast to whole genome SNPs in *A. phalloides*, which show strong genetic clustering by space and time, both on a global and local scale (Golan *et al.* in review). The *HD1* and *HD2* alleles appear overdispersed, which is consistent with maintenance by long-term balancing selection, and not locally clustered around a point of origin as would be expected for newly evolving alleles.

My findings call into question the claims that there is no recombination between *HD1* and *HD2*. Individuals that have “identical” (due to missing data) HD1 alleles have, in some cases, quite divergent HD2 alleles (Supplementary Figure S2.9). Interpretation of these results depends, again, on the age of the diversity among alleles-- recombination could be strongly suppressed, but if the alleles are old then rare combination events could explain what we see. This question, too, could be resolved with better phasing, local reassembly, or longer read sequencing for all individuals along with a molecular clock analysis. Yen-Wen Wang’s protein co-phylogeny also indicates possible recombination events (Figure 2.5B).

This study makes a new methodological contribution by trying two different ways of recovering *HD1* and *HD2* alleles from next gen sequencing data, the phased variant-alternate references method used in most of the text and the *de novo* assembly graph query method used by Yen-Wen Wang. When sequencing a natural population, it is important to sequence dikaryotic tissue to know which haplotypes are compatible and which, in fact, co-occur. Otherwise one is limited to collecting a species that can be cultured and culturing it to performing crosses. Genetic crosses in the lab will, with great effort, demonstrate which alleles are compatible, but not the actual mating dynamics (i.e. which alleles would have ended up together) in the natural population. Sequencing dikaryons allows for data collection directly from natural populations and at scale, but it creates the challenge of phasing haplotypes, a challenge made even more difficult by the high diversity and suppressed recombination at the HD locus. The variable domains of *HD1* and *HD2*, the areas that determine compatibility, phased especially poorly. The phased variant-alternate references method could be improved with longer reads. Fortunately, *HD1* and *HD2* are very short. In *A. phalloides*, *HD1, HD2*, and the intergenic region are each smaller than 1 kb, so long read sequencing would be expected to cover the entire area in a single read multiple times, allowing for much clearer inference of haplotypes. Deriving alleles from a *de novo* assembly suffers from similar problems to phasing, but querying the assembly graph allowed Wang to distinguish two *HD1-HD2* cassettes of suppressed recombination. Though not directly compared here, the assembly graph query method is likely to be superior to the variant phasing alternate references method because it preserves more of the haplotype information directly from the reads of the individual, rather than using reads from an individual to place variants that were called jointly with all the other samples on a shared reference backbone (even though in aggregate variants called by GATK are more accurate than those found on the raw reads).

In this study, we demonstrate, despite some uncertainty in the phasing of the sequences, that there is considerable diversity in *HD1* and *HD2* in *A. phalloides*, across the globe and within 10m x 10m populations*.* We tentatively conclude that that diversity is ancient, and maintained by balancing selection, rather than continuously maintained by negative frequency-dependent selection. The technical challenges of sequencing *HD1* and *HD2* were considerable, but the benefits of sampling directly from natural populations and carrying out the work bioinformatically rather than under a hood we also considerable, and we offer recommendations to smooth the way for the next attempt. This first natural population study of *HD1* and *HD2* has left many questions unanswered, but through the techniques and resources I've developed here, we've moved one step closer. John Raper (1966) once said, "...sexuality in the higher fungi is no more mystifying than elsewhere; the facts, examined in proper sequence, are simple enough, but it does seem there are quite few of them." I admit, I am still somewhat mystified by the tetrapolar mating system, but content to have added a few more significant facts to that number.

Acknowledgments

I thank Richard Childers for providing bioinformatic expertise and advice. The computations in this study were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University, with special thanks to Brian Arnold, Adam Freedman, and Allison Schultz. Thanks to the National Science Foundation for funding grant 1021606 to Anne Pringle and NSF Graduate Research Fellowship 2013170866 to Holly Elmore.

References

Aguilar, A. *et al.* High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 3490–3494 (2004).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).

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

Brown, A. J. & Casselton, L. A. Mating in mushrooms: Increasing the chances but prolonging the affair. *Trends Genet.* **17**, 393–400 (2001).

Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).

Coelho, M. A., Bakkeren, G., Sun, S., Hood, M. E. & Giraud, T. Fungal Sex: The Basidiomycota. *The Fungal Kingdom* 147–175 (2017). doi:10.1128/microbiolspec.funk-0046-2016

Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3 : fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2017).

Depristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–501 (2011).

Deshpande, V., Fung, E. D. K., Pham, S. & Bafna, V. Cerulean: a hybrid assembly using high thoughput short and long reads. *Algorithms Bioinform Lect Notes Com Sci.* **8126,** 349–350 (2013).

Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
English, A. C. *et al.* Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS One* **7**, 1–12 (2012).

Golan J., Adams C. A., Cross H., **Elmore M. H.**, Gardes M., Glassman S., Gonçalves S., Hess J., Richard F., Wang Y., Wolfe B., Pringle A. *Native and invasive populations of the ectomycorrhizal death cap Amanita phalloides are highly sexual but dispersal limited.* (in review) preprint: https://www.biorxiv.org/content/10.1101/799254v1

Gordon, D. *et al.* of the Gorilla Genome. **344,** 1–21 (2016).

Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci.* **108**, 1513–1518 (2011).

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

Hoff, K. J. & Stanke, M. WebAUGUSTUS--a web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic Acids Res.* **41**, 123–128 (2013).

Hunt, M. *et al.* REAPR: A universal tool for genome assembly evaluation. *Genome Biol.* **14**, (2013).

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226–232 (2012).

Kajitani, R. *et al.* Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).

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

Kämper, J., Reichmann, M., Romeis, T., Bölker, M. & Kahmann, R. Multiallelic recognition: Nonself-dependent dimerization of the bE and bW homeodomain proteins in Ustilago maydis. *Cell* **81**, 73–83 (1995).

KateN. (howto) Discover variants with GATK - A GATK Workshop Tutorial. *GATK 3 User Guide*. (2016). <https://software.broadinstitute.org/gatk/documentation/article?id=7869>

Kohler, A. *et al.* Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. *Nat. Genet.* **47**, 410–415 (2015).

Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. 1–35 (2016). doi:10.1101/gr.215087.116.Freely

Krueger, F. Trim Galore: https://github.com/FelixKrueger/TrimGalore

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

Li, H. BFC: Correcting Illumina sequencing errors. *Bioinformatics* **31,** 2885–2887 (2015).

Martin M., Patterson M., Garg S., Fischer S. O., Pisanti N., Klau G. W., Schoenhuth A., Marschall T. *WhatsHap: fast and accurate read-based phasing.* bioRxiv 085050. doi: 10.1101/085050

Nieuwenhuis, B. P. S., Nieuwhof, S. & Aanen, D. K. On the asymmetry of mating in natural populations of the mushroom fungus Schizophyllum commune. *Fungal Genet. Biol.* **56**, 25–32 (2013).

Raper, J. *The Genetics of Sexuality in Higher Fungi*. New York, USA, The Ronald Press Company, 1996.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E. & Jones, S. J. M. ABySS : A parallel assembler for short read sequence data ABySS : A parallel assembler for short read sequence data. 1117–1123 (2009). doi:10.1101/gr.089532.108

Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

Wang, J. R., Holt, J., McMillan, L. & Jones, C. D. FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinformatics* **19,** 1–11 (2018).

Wgsim: https://github.com/lh3/wgsim

Warren, R. L. *et al.* LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience* **4**, (2015).

Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).

Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* **6,** 1–9 (2016).

Chapter 3
Contributions to the *Amanita* system

Abstract

This chapter summarizes contributions I have made to the *Amanita* system. **Section I** suggests a new lens for viewing "individuality" in filamentous fungi, in which the mycelium is less the organism itself, but a shared resource for its constituent nuclei. The degree of conflict or integration in the interests of the nuclei determines how well-integrated of an individual a mycelium is, a paradigm which could explain seemingly Byzantine adaptations at the level of the mycelium, such as clamp connections. **Section II** describes a population genetic analysis of the *Amanita*BASE data to determine the size of genets in *Amanita phalloides.* **Section III** describes my investigation into the apparent loss of the *HD1* mating gene in *Amanita thiersii* and its implications for *A. thiersii*'s rapid invasion of the Eastern US and low genetic diversity across its range. **Section IV** details the sequencing and assembly of *A. phalloides* genome Dr4M1 in advance of the massively parallel *Amanita*BASE project and the early exploration of MSDIN toxins in the *A. phalloides* genome. Relevant history of the *Amanita* system is included throughout.

Introduction

This chapter summarizes several contributions I have made to the *Amanita* system. Some are otherwise unpublished ideas, and others resulted in manuscripts on which I am a middle author. My largest contribution to the system has been organizing a database of *Amanita* specimens, revamping collection protocols, sequencing 86 *A. phalloides* and three *A. thiersii/A. foetens* (one of each and one isolate that may be either species), and generating alignments and variants for each sequenced individual, a project I refer to as *Amanita*BASE (Ch. 1). These data have served as the basis for my own work in Chapter 2 and for many other projects, current and planned.

The spores of *Amanita*BASE were germinated when Anne Pringle began working on the mystery of whether *A. phalloides* was native to California or invasive*.* In Pringle *et al.* (2009), using a combination of morphological identification, historical investigation, and multi-marker Sanger sequencing, Pringle showed that California *A. phalloides* were recently descended from European populations and showed evidence of genetic bottlenecking. The historical record revealed that *A. phalloides* had spread along the West Coast and into California's interior since 1938, meaning that it is both introduced and expanding its range, making *A. phalloides* invasive. Pringle's collections and the historical collections she amassed served as the basis for *Amanita*BASE.

Genomics on the *Amanita* system began with the *A. thiersii* Skay 4041 v1.0 (Hess *et al.* 2014; Chaib de Mares *et al.* 2015) and *A. muscaria* Koide v1.0 genomes (Kohler *et al.* 2015), sequenced and assembled by the Joint Genomics Institute of the Department of Energy. *A. thiersii* was chosen based on work by Wolfe *et al.* (2012) showing that it was invasive and nearly clonal across its large and rapidly expanding range, continuing the theme of invasion in *Amanita.* This paper was accompanied by substantial collecting. Next, the genomes of *A. brunnescens, A. polypyramis* and *A. inopinata* were sequenced and assembled in-house by Jacky Hess. These genomes set the stage for **Section III: *A. thiersii*, the loss of *HD1*, and Baker's Law.**

Section III sparked my interest in mating compatibility and opportunities for genetic conflict in a dikaryotic mycelium. Is a dikaryon really one organism when the two halves of its genomes move about in semi-autonomous nuclei that sometimes retain their own separate opportunities for reproduction? I attempt to answer this question in **Section I: Are there "individuals" in mycelial fungi?**

A. phalloides was not sequenced in these earlier efforts, despite the strong work that had already been invested in the species' ecology and natural history, because it does not culture, and so it was not possible to get enough high molecular weight DNA. However, due to the rapid pace of improvement in sequencing technology, a few years later when I wanted to know the synteny of the *A. phalloides'* HD mating locus, I was able to extract plenty of DNA from a lyophilized 2004 specimen to sequence the 48x Illumina-only genome in **Section IV: Assembling the genome of Dr4M1.** *A. phalloides* genomics expanded significantly when I proposed *Amanita*BASE as the basis of my dissertation and *A. phalloides* as the majority of *Amanita*BASE. The biochemical population genomics project in **Section IV** and **Section II: Genet size in *Amanita phalloides*** are based on the *Amanita*BASE dataset.

The *Amanita* system is rich in resources and only getting richer. There is a long, quality record of the natural history of *Amanitae*, large collections from around the world, a rapidly expanding set of sequenced genomes, and a growing community of researchers engaged in this work. I am honored to have contributed.

I. Are there "individuals" in mycelial fungi?

The concept of the individual organism is central to much of biological thought. In particular, individuals are often portrayed as the entities that possess adaptations and are assigned fitnesses in many formal models of evolutionary processes. Recent work on the concepts of individuality or organismality have viewed individuals as a "level of selection" that emerges from alignment of the fitnesses of formerly independent parts. Filamentous fungi provide a particularly challenging case for defining the biological individual because of hyphal fusions within mycelial networks. These fusions can result in physiological integration of formerly separate cytoplasms and, in some cases, produce heterokaryotic mycelia that contain genetically unrelated nuclei. The different nuclei within a heterokaryotic mycelium can possess a high degree of shared interests and yet retain the possibility of independent reproduction.

Genetic interests that do not entirely overlap leave open the possibility of conflict. In mycelial fungi, particularly in Basidiomycetes, there can be conflict between different nuclear types in mating interactions with third parties and in processes of nuclear divorce (Nieuwenhuis *et al.* 2013a) and conflict between nuclei and mitochondria (Aanen *et al.* 2004).

One approach has been to identify the mycelium as the individual, where territorial boundaries between mycelia are defined by reactions of somatic incompatibility that prevent cytoplasmic fusion (Lind *et al.* 2007). By this criterion, some mycelia are among the largest biological individuals on the planet (Anderson *et al.* 2018). Another approach has been to divide the mycelium into local 'reproductive units' defined as groups of nuclei that share a common fitness (Ma et al. 2016), to see the hypha as the fundamental unit (Falconer *et al.* 2005), or to see the nucleus as the operational selection unit (Johanneson & Stenlid 2004). A pluralistic model recognizes nuclei and mycelia as both fulfilling the criteria of 'Darwinian' individuality (Booth 2014).

There is value in all these approaches to individuality in fungi. This suggests that intuitions associated with biological individuals based on animals like ourselves where individuals are usually well-defined may not be the most appropriate model for thinking about evolutionary processes within the fungal mycelium. Instead, I propose thinking of the mycelium and its cytoplasm as the "constructed niche" of its genetic inhabitants—a vast mycelial metropolis providing the infrastructure for the support of its constituents. The genetic inhabitants of mycelia include nuclei, mitochondria, mycoviruses, and dispensable chromosomes. Fission of a mycelium can make multiple instantiations of the same genetic individual. Fusion of two genetically distinct mycelia by vegetative compatibility or mating can make one physiological individual.

Within a large city most residents do not directly interact but have shared interests in the provision of public goods from which all benefit while each also has their own local, private goods. The mycelium provides public goods including shelter, substrate digestion and nutrient absorption, and reproductive structures. Most residents contribute to public goods and collective wellbeing, but some may be more strictly parasitic. In this model, as applied to filamentous fungi, the provision of public goods is analogous to maintenance of the mycelium.

The more aligned the reproductive interests of all the genetic inhabitants of the mycelium are, the more “well-defined” an individual it becomes. In most animals, a gametic meiosis ensures that the physiological organism is composed of genetically identical cells, minimizing conflict. Basidiomycetes generally spend most of their life cycle as dikaryons, mycelia with two genetically distinct nuclear types. Each nuclear type has some autonomy and retains opportunities to profit at the expense of the other nuclear type, which can lead to conflict. Seemingly inefficient adaptations at the level of the mycelium, such as conjugate nuclear division with clamp connections, may be solutions to conflicts like these.

The relative autonomy of the different genetic inhabitants of the mycelium, particularly genetically distinct nuclei in a dikaryon, within the shared mycelium is reminiscent of eusocial insect colonies within a nest. Just as there is a continuum of sociality and eusociality, within Basidiomycetes there is a range of levels of adaptive integration (Queller & Strassman 2009), from heterokaryons of four or five nuclear types, not all of which are mated to all the others (Nieuwenhuis *et al.* 2013a) to the diploid “humongous fungus,” the largest single genetic individual by body mass on earth (Anderson *et al.* 2013).

II. Genet size in *A. phalloides*

This work resulted in authorship on the following manuscript:

Golan J., Adams C. A., Cross H., **Elmore M. H.**, Gardes M., Glassman S., Gonçalves S., Hess J., Richard F., Wang Y., Wolfe B., Pringle A. *Native and invasive populations of the ectomycorrhizal death cap Amanita phalloides are highly sexual but dispersal limited.* (in review) preprint: https://www.biorxiv.org/content/10.1101/799254v1

Introduction

This study sought to determine the size and lifespan of genets in *Amanita phalloides.* A genet is a genetic individual, even though it may be dispersed across many ramets, or apparent “individuals” which share the same DNA or even the same physiological processes (*c. f.* original usage in Harper 1977). We see mushrooms that appear separate above ground, but because the mycelium is embedded in the soil, we do not know whether each mushroom has its own mycelium, and is therefore each its own genet, or all the mushrooms come from the same mycelium and so are several ramets of one genet. Knowing the size and lifespan of genets in *A. phalloides* tells us several things about its life history. For example: Does the fungus propagate primarily by spores or mycelium? How prolific is an individual genet? What proportion of reproduction is vegetative (mycelial propagation) vs. sexual (sporic)?

Understanding the life history of *A. phalloides* is intricately intertwined with understanding its invasion, mating system (Chapter 2) and its individuality (following **Section I: Nuclear Individuality**).

Methods

All scripts may be found in github.com/elmoremh/Amanita-Population-Genomics

Anne Pringle took AFLP data from several populations in California in 2005, when she first attempted to determine genet size in *A. phalloides*. When I proposed AmanitaBASE as part of my qualifying exam, I proposed next generation sequencing of those 2005 specimens to give the genet size analysis greater resolution. The collection protocol for new specimens was designed in part to allow further investigation of the genet size question, and investigating genet size also guided the selection of new specimens for next generation sequencing.

Collecting, processing, and accessioning specimens

Find complete metadata for all Amanita*BASE specimens in* [Amanita*BASE v3 Specimen Metadata*] *(and find Amanita*BASE v3 Specimen Metadata 11-12-19.xlxs *in Chapter 1 Supplementary Materials).*

The *Amanita*BASE protocol: California and Portugal, 2015

The core of the *AmanitaBASE* specimens are *A. phalloides* collected at several sites in Point Reyes National Seashore, California, USA in December 2015 following a detailed collection protocol. A smaller collection of *A. phalloides* was made by Susana Gonçalves in Coimbra, Vilarinho, and Agrária, Portugal during the same year.

Several sites from previous work were revisited (Drake 2, 3, and 4 and Heart's Desire 1, 2, and 3), located by GPS coordinates and positions refined by Anne Pringle's recorded landmarks. To find new sites, we visually scanned for mushrooms from a car window. Four new sites were established: Pet 1, Pet 2, Picnic 1, and Picnic 2. When several mushrooms were sighted together, we proceeded to define the area as a spatial population, generally ~10 m x 10 m. GPS coordinates were recorded at the center of the spatial population. There is slight variation in the centerpoint of some populations over the years, and commercial GPS is only accurate to within ±5 meters. This spatial definition of a population is somewhat arbitrary, but we ensured that populations were separated by several hundred feet or by a natural break such as a road or waterway. The environment of each population was photographed to help identify possible tree hosts, and trees and surrounding flora were identified and noted.

Once the limits of the population had been defined, we found every *A. phalloides* specimen within it and marked it with a flag. Next, we determined which ones were to be collected. Overall, we wanted mostly mature and undamaged mushrooms, but also a few older and younger specimens for microbiome work. Mushroom age and condition were noted. There was a general goal of collecting 15-50 mushrooms per population in California and as many as possible in the Portuguese populations, as European *A. phalloides* populations are frequently in the single digits and rarely exceed the teens. If there were three or more mushrooms to be collected, the position of each was recorded using a compass surveyor and field tape. If there were only two, the distance between them was measured and GPS coordinates collected at the midpoint.

After mapping, each mushroom was examined and its location and other observations about it noted. Then, each chosen mushroom was arrayed with collecting materials and labeled with a unique identifier, the SpecimenID, that would serve as its primary ID in the database. The SpecimenID is a 5-digit number assigned to a single specimen. Batches of SpecimenID numbers are given out by the team data manager before a collection trip so there is no possibility of accidentally assigning the same numbers during concurrent collection trips. During the collecting, the data manager assigned specimens their SpecimenID so that there was no confusion caused by other team members acting independently.

With the unique identifier visible, each specimen was photographed in detail. Only then would the specimen be disturbed. First, an inch-long isosceles triangles would be excised from the cap with a fresh, sterile scalpel blade and placed in a 15 mL Falcon tube of RNAlater. These samples will be useful for any nucleic acids but particularly for potential transcriptome analysis. Second, the mushroom would be exhumed, making sure to include the entire volva, and placed in a wax bag. Third, an autoclaved spoon would be used to collect soil from beneath where the mushroom had emerged (and at some sites at further distances as well) and place it into a 50 mL Falcon Tube. At some sites, bags of soil were collected from around mushroom sites in the hopes of identifying root tips.

After days of collecting and chilling specimens in refrigerators overnight, the specimens were processed at the Bruns Lab at Berkeley for processing. Cap diameter was measured as well as length of the mushroom from the tip of the volva to the top of the cap. The entire mushroom was weighed in its wax bag. Then samples of the mushroom taken from the inside of the cap and the inside of the stipe were harvested for the purpose of microbiome analysis. For genomics and toxicity analysis, we cut large chunks of gill, cap, and interior of stipe tissue, placed them in 15 mL Falcon tubes and lyophilized them. Soil was preserved by placing it directly in a lyophilizer (no freezing required first).

We collected a total of 60 mushrooms from 7 populations in the December 2015 California trip.

Other specimens

*Amanita*BASE is about standardizing protocols and ensuring data completeness and integrity, maximizing general usefulness of each specimen for population genomics and ecology and minimizing loss of irretrievable data. But *Amanita*BASE must be flexible enough to accommodate valuable specimens that were collected by different collectors with different needs and priorities, particularly if those specimens are irreplaceable. Other specimens have been entered into *Amanita*BASE that do not meet the above described protocol but merit inclusion due to their unique geography or age. The bulk of these specimens are *Amanita*e collected by the Pringle Lab and its affiliates with accompanying information that largely overlaps with *Amanita*BASE protocols. Some specimens were requested from other collections or herbaria (particularly Kew Gardens), and these generally included information about location, date of collection, physical descriptions, etc. but rarely had GPS coordinates, for example. These specimens were generally much older than the rest of the *Amanita*BASE specimens or from a location that broadened the geographic reach of *Amanita*BASE.

Sequences generated for *Amanita*BASE

See Chapter 1: Methods *for more details and supplementary materials related to bioinformatics on the* Amanita*BASE specimens.*

In 2015, I proposed a massive *A. phalloides* sequencing effort as part of the creation of *Amanita*BASE. The goal was to look in-depth at the invasive Californian Drake 2 and Drake 3 populations over time, compare them to native European populations, and include some greater variation through time and geography to help put the results into perspective. 89 specimens were ultimately selected, 86 *A. phalloides* and 3 *A. thiersii/A. foetens.* Of the *A. phalloides*, 67 were from Drake 2 and Drake 3 in California over the years 2004, 2014, and 2015; 11 collected from Portugal in 2015; and 8 older specimens from across Europe. The *A. thiersii* specimens were included in the sequencing and variant-calling to help answer a species identification question and to see if there were any changes in the genome of 10802/*Ath* Skay 4041 as it has sat in culture for 5-6 years since it was originally sequenced (see **Section III**, subheading *HD1 and HD2 appear absent in other sequenced A.thiersii genomes*).

The 89 specimens were sequenced with Illumina HiSeq 2500 and two of those, 10511 and 10721, were sequenced with PacBio as well. The Illumina HiSeq sequencing was 250 bp paired-end reads with a 550 bp insert (with the exception of 10801, 10169, 10170, 10171, 10277, 10003, 10004, 10007, 10010, 10016, 10018, 10019, 10023, each of which had a 350-bp insert) prepared as IntegenX dual-index libraries. The mean sequencing depth of each of the samples ranged from 10.56 to 150.86. Two of the above 89 mushrooms, *A. phalloides* 10721 (USA, California, Drake 3, 2015) and *A. phalloides* 10511 (Portugal, São Jacinto, Dunas de Mira) were also sequenced using a PacBio Sequel platform at the University of Wisconsin-Madison Biotechnology Center, with a 20-kb single library per specimen and yielding average read sizes of 14,833 and 14,935 for specimens 10721 and 10511, respectively. 10511, which became the reference assembly, had raw PacBio coverage of 47x with N50 read length of 6,310 bp.

Trimming was performed on all 2015-sequenced specimens with the program Trim Galore v0.4.5 (Krueger, https://github.com/FelixKrueger/TrimGalore). Reads that became shorter than 100 bp after trimming and those with a quality score less than 30 were discarded. Adapter trimming was set to the highest stringency, 1, meaning a single nucleotide of overlap with the adapter sequence was trimmed from the read. Unpaired reads were retained, though they did not end up being used in either the assemblies or the alignments. Raw and trimmed reads are available in NCBI BioProject PRJNA565149.

Alignments were run against the 10511 reference assembly using the Burrows-Wheeler Alignment software, BWA, mem algorithm with default parameters (Li & Durbin 2009), in the course of the GATK best practices pipeline (see **Variant Calling**).

Assemblies

Hybrid assemblies, incorporating both PacBio and Illumina HiSeq reads, of specimens 10721 (California) and 10511 (Europe) were completed by Jacky Hess. Extensive troubleshooting was performed in the course of generating the 10511 assembly, and then the workflow arrived at for 10511 (below) was applied to 10721.

The workflow began with pre-processing: trimming and filtering with Trimmomatic v 0.35 (Bolger *et al.* 2014) (with the following parameters: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 CROP:245 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:25 MINLEN:100), Illumina reads sequencing error correction with BFC (Li *et al.* 2015) and PacBio reads sequencing error correction with FMLRC (Wang *et al.* 2018). Dr. Hess proceeded to test several assemblers: CANU (Koren *et al.* 2016) and FALCON/HGAP4 (Gordon *et al.* 2016) are PacBio-only assemblers; ABySS (Simpson *et al.* 2009), Platanus (Kajitani *et al.* 2014), and AllpathsLG (Gnerre *et al.* 2014) are Illumina-only (although to meet the library prep criteria for Allpaths LG, Illumina libraries with the proper insert size were simulated using the program wgsim (https://github.com/lh3/wgsim) on PacBio reads); and SPAdes (Bankevich *et al.* 2012), DBG2OLC (Ye *et al.* 2016), and Cerulean (Deshpande *et al.* 2016) are hybrid assemblers. Based on metrics like contig number and lengths, SPAdes and Platanus were not considered further. Scaffolding was performed with LINKS (Warren *et al.* 2015) and each assembly was “polished” with the gap-filler PBJelly (English *et al.* 2012) and the base and indel correction program Pilon (Walker *et al.* 2014). The polished assemblies were analyzed with QUAST (Gurevich *et al.* 2013), BUSCO (Simão *et al.* 2019), and REAPR (Hunt *et al.* 2013) and evaluated on completeness of eukaryotic single copy complement, completeness of eukaryotic duplicated gene content, fragmentation, missing sequence, assembly size as percent of genome size, number of scaffolds, scaffold N50/NG50, and scaffold L50. Allpaths LG was chosen as the preferred assembler because the Allpaths LG assembly had the highest ranking in the greatest number of evaluation criteria.

Variant calling

SNPs and Indels were called using the Genome Analysis Toolkit (GATK) v3.8-0-ge9d806836 software (Depristo *et al.* 2011) and following the GATK best practices (KateN 2016) as well as is possible for a non-model system, with help from Allison Shultz's Github page (github.com/ajshultz/whole-genome-reseq). I began by aligning the Illumina reads from each sample to the 10511 hybrid assembly using the BWA software, mem algorithm (Li & Durbin 2009). Mapping rates ranged from 20.0% - 95.3%, with a median mapping percentage of 86.1%. Specimen 10003 had a low mapping rate and did not cooperate with the GATK workflow, so it was removed from the joint-calling cohort. The mapping rate of 10511's Illumina reads to the 10511 hybrid assembly was 93.8% (see Alignment and Deduplication metrics in Supplementary Materials). Following the steps, I marked duplicate reads, but I did not recalibrate base scores based on known variants because there were no known variants for *A. phalloides.* The GATK program Haplotypecaller makes variant calls jointly on all the samples, generating a GVCF file that contained a record of all sites of all the genomes, whether invariant or variant. The program GenotypeGVCF creates the raw VCF files containing only the SNPs and Indels. Our samples contained 1831629 raw variants.

Again, due to the lack of known variants in *A. phalloides* for comparison, the raw variants were hard-filtered according to the default parameters of the GATK VariantFiltration program.

Because of the generally high coverage (~50x) of our sample set, hard filtering with default parameters was not expected to bias the filtered variants.

Ordination and cluster analysis

I handed raw and filtered SNPs over to Jacob Golan for ordination and cluster analysis.

Results

The analyses performed by Jacob Golan and the other co-authors found that:

1) Virtually every sporocarp is a single, short-lived genet. Only two genetically identical mushrooms were found, only in the AFLP dataset, and their “identity” may well be an artifact. In the multiply sampled populations, genets did not appear to persist between seasons, let alone grow larger over time. Because of the lack of large genets with multiple sporocarps, we conclude that *A. phalloides* spreads predominantly by sexual basidiospores as opposed to vegetative growth.

2) Genotypes cluster at continental scales. More interestingly, genotypes cluster at highly local scales as well. For example, the populations Drake 2 and Drake 3 are only 100m apart, but in all the years they were sampled (2004, 2014, 2015) their genotypes clustered separately.

3) Populations remain genetically differentiated across years. FST (Weir & Cockerham 1984) calculated using the 2004, 2014, and 2015 data of Drake 2 ranges from 0.0054–0.0138, and for Drake 3 ranges from 0.0003–0.0270. The FST statistic comparing the adjacent populations, Drake 2 and Drake 3, in 2004 is 0.0376 (0.021 with AFLP data), in 2014 is 0.0523, and in 2015 is 0.0398.

Discussion

Using the SNPs I generated from 86 whole *A. phalloides* genomes along with AFLP data from 2005 and 2007, Jacob Golan *et al.* found that virtually all of the mushrooms included in the analysis were unique genets (and the one instance of two mushrooms to a genet may well have been an AFLP artifact). There were no instances of the same genet being found in successive years nor of genets growing over multiple years in the populations that were sampled at several time points. The pattern of many singleton genets persists across our dataset. Thus, it appears that genets do not live long and reproduce sporically quickly after they are established. Genotypes cluster by continent as well as by local scales, on the order of 10m x 10m. Clustering lingers at local sites between years, leading to the interpretation that spores do not tend to disperse more than a few meters away from the parent sporocarp. FST values show that the California Drake 2 and Drake 3 populations remained differentiated over 2004, 2014, and 2015 despite being only 100m apart.

A. phalloides may be more "organismal" *a la* Queller & Strassman (2009) than other Basidiomycetes, because, unlike many other Basidiomycetes, there is frequent meiosis. A small mycelium that quickly produces a mushroom and dies off may indicate that the nuclei of the dikaryon have incentives which are well-aligned, leading to more cooperative, organismal behavior.

A. phalloides resembles H. G. Baker's (1965) classic paradigm of a weed. *A. phalloides* is small-bodied, ephemeral, annual, and may prefer disturbed areas (ruderal). Baker theorized that weed characteristics predisposed certain plants to invasion. Finding that *A. phalloides*, an invasive Basidiomycete, has a weed-like life history may help to explain its invasion of not only North America but Africa, Australia, and New Zealand.

A. phalloides's small genet size and implied ephemeral life history also have interesting implications for its mating system. A "weedy" lifestyle should favor self-compatibility under Baker's model (1965), and yet *A. phalloides* appears highly outbreeding (**Section II**; Chapter 2; Golan *et al.* in review). It is unlikely that undergoing meiosis and mating so frequently is merely a byproduct of weed syndrome selection for a small, ephemeral body and annual life cycle, because reversions to bipolarity and other means of reducing barriers to inbreeding are common in Basidiomycetes (Nieuwenhuis *et al.* 2013b). However, because of the high multiallelism in most Basidiomycetes, the tetrapolar system allows mating between close relatives (as well as ¼ of all within-tetrad matings) who happen to have different mating alleles. A great deal of diversity in mating locus alleles seems to have been sampled during the invasion of North America’s West Coast (Chapter 2). This may mean that, in practice, *A. phalloides* has a system that promotes outbreeding when possible but, like a weed, is able to mate with relatives or the ¼ of its own spore pairings that are compatible when necessary. *A. phalloides* may be getting the best of a few worlds in this interesting combination of life history strategies. Because fungal life history is understudied, further insights like these may be thick on the ground for future foragers.

III. *A. thiersii*, the loss of *HD1*, and Baker’s Law

Co-authors: Jacqueline Hess, Anne Pringle

Introduction

Amanita thiersii is both saprotrophic and invasive. *A. thiersii* was first noticed in the United States in Texas in 1952 and now found as far north as Illinois and as far east as Maryland (Figure 3.1), almost always in suburban lawns. The native range of *A. thiersii* is not known, but it is speculated to be native to Mexico or Central America.

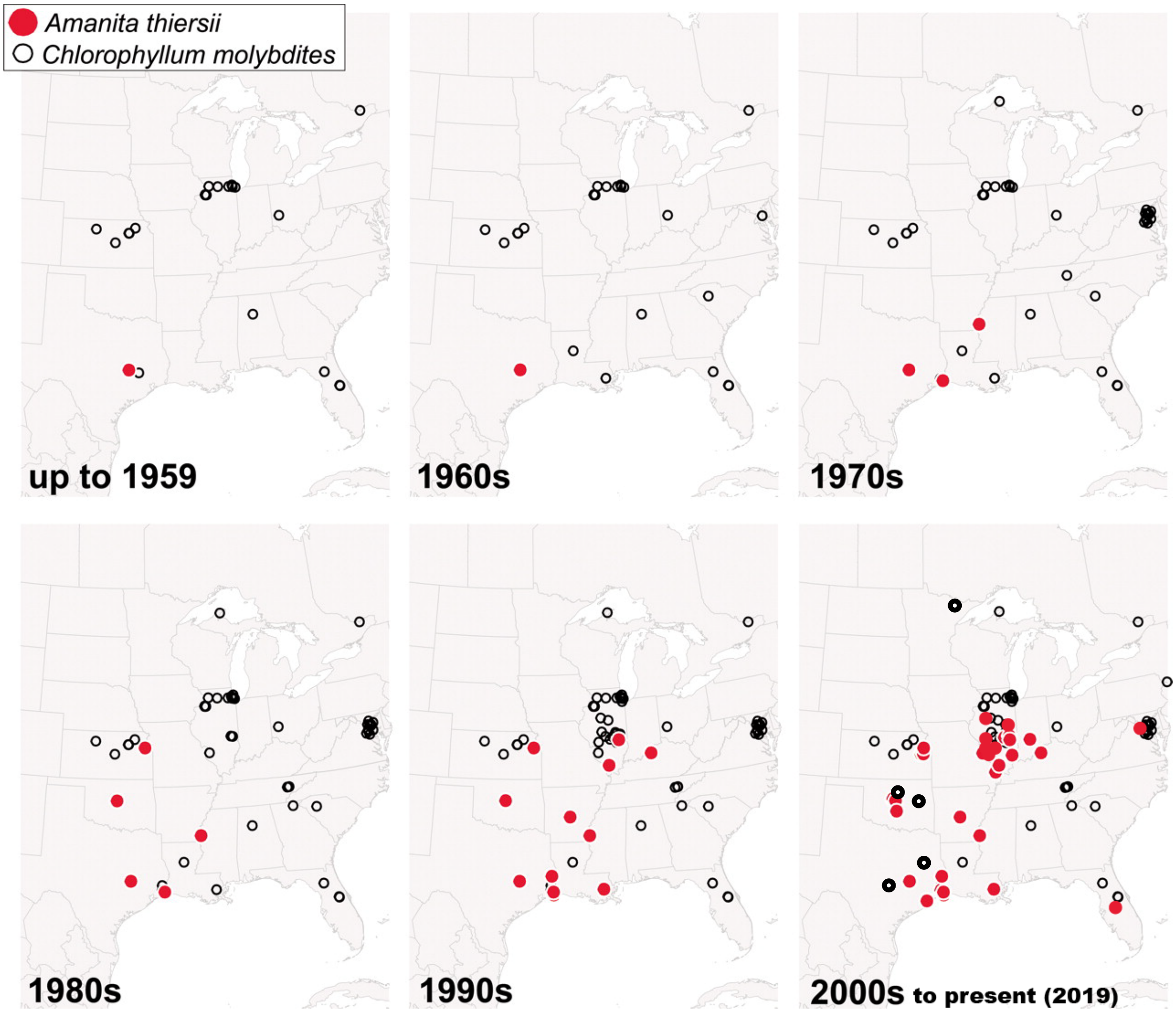


Figure 3.1: *A. thiersii*'s range has expanded rapidly across the US in only about 60 years. Red dots represent *A. thiersii* sightings and white dots represent sightings of *Chlorophyllum molybdites*, a mushroom of similar lifestyle, for comparison. *Adapted from Wolfe et al. (2012), updated with data from MushroomObserver.org.*

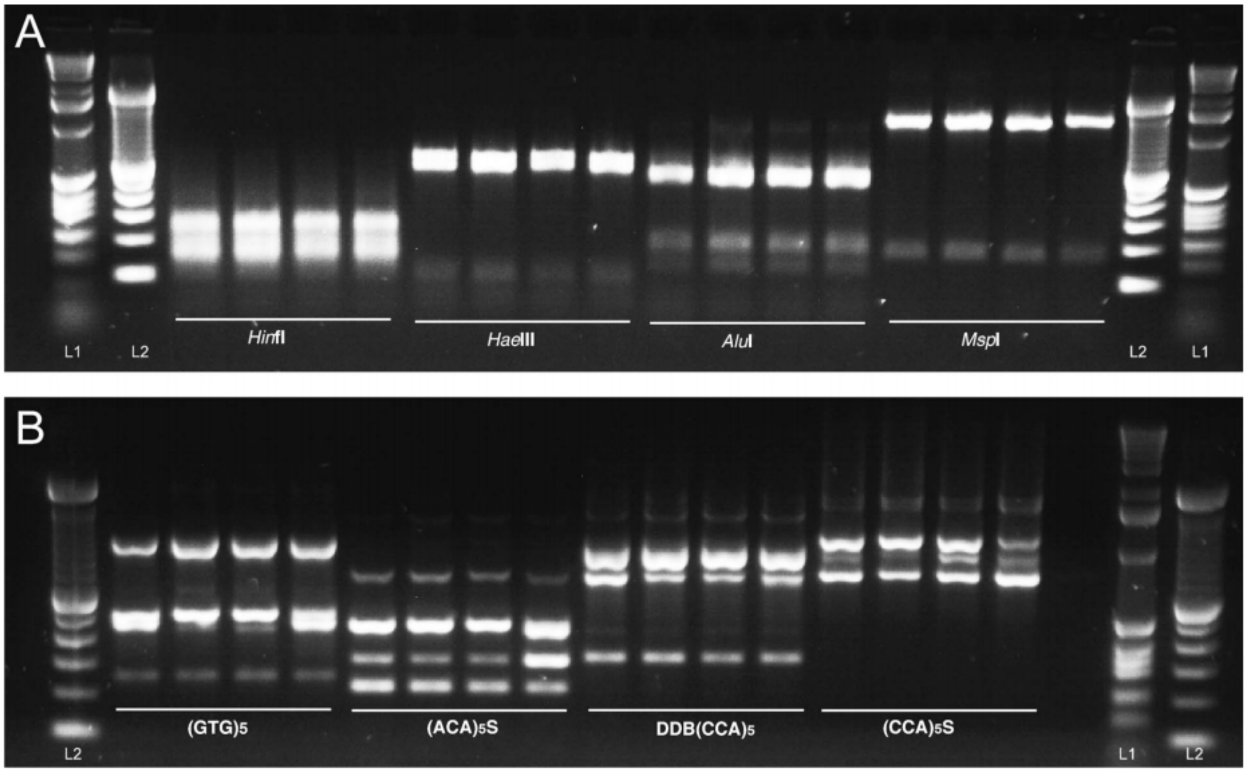


Figure 3.2: A. RFLP analysis of IGS-1 in four specimens of *A. thiersii.* B. Intersimple sequence repeat (ISSR) banding pattern of the same four specimens. Some variation is visible, but overall diversity appears low.

A. thiersii is highly genetically similar across its range (Wolfe *et al.* 2012). Figure 3.2 shows RFLP and ISSR genotyping of ribosomal intergenic spacer region 1, IGS-1, a locus that is generally highly variable within populations and exhibits band size variation on gels, in four *A. thiersii* individuals from across its range as of 2012. This is consistent with a rapid invasion that began with a small number of founding genotypes (Baker 1955; Stebbins 1967). However, although there are other invasive species in *Amanita*, and this type of rapid, low-diversity range expansion across such a large area is not common, nor is it common to see significantly depleted diversity in the invasive range (Golan *et al.* in review, and see **Section II**).

Intriguingly, the *A. thiersii* Skay reference genome (Hess *et al.* 2014; Chaib de Mares *et al.* 2015) indicates that the sequenced isolate was missing *HD1*, a key protein involved in Basidiomycete mating (Chapter 2). The disruption of the heart of the homeodomain (HD) mating locus would be expected to cause changes to the mating system or even the life cycle that might account for *A. thiersii*'s rapid range expansion and low diversity.

Amanitae share an ancestral tetrapolar mating system with other Basidiomycetes. Tetrapolar species have two mating loci, the pheromone/receptor (PR) locus and the homeodomain (HD) locus. The PR locus includes the genes responsible for making pheromones and receiving the pheromones of others. The HD locus contains, at its heart, the components of a transcription factor that unlocks downstream dikaryosis and mating functions (Casselton 1997; Casselton & Olesnicky 1998). For two mycelia to be sexually compatible, they must have different alleles at the PR and HD loci. The system is called tetrapolar because, when two monokaryons meet, there are four possible outcomes-- same HD-same PR (mating failure), same HD-different PR (mating failure), different HD-same PR (mating failure), different HD-different PR (mating success)-- and only one of them is a successful mating.

Tetrapolarity would be expected to put a limit on rapid, low-diversity expansion. Wild populations of mushrooms tend to have dozens or hundreds of PR and HD alleles (Raper 1966), so spores should mainly be prevented from sib-mating without finding it difficult to mate with non-relatives. However, in a low-diversity population, there may only be sibs, or at least individuals sharing the same mating alleles, left. *A. thiersii* must have avoided this trap somehow to expand as it has with such low diversity. If *HD1* really is missing from *Ath* Skay4041, it may be missing from some or all of the known *A. thiersii* population, and that must mean that *A. thiersii* has found a way around the requirements of tetrapolarity, maybe through inbreeding, selfing, or even cloning. But before we can speculate, we must know if *HD1* is actually missing.

Methods

Locating the HD locus in *Ath* Skay4041

The HD locus in the *A. thiersii* Skay4041 v1.0 assembly was located by the Joint Genome Institute (JGI) of the Department of Defense's tblastn search (Altschul *et al.* 1990), using the *A. muscaria* Koide v1.0 HD1, HD2, and mitochondrial intermediate peptidase (MIP) amino acid sequences as queries (Supplementary Materials). To one side of *MIP* were other known HD mating locus genes, and on the other side of MIP, where *HD1* would be expected, the scaffold abruptly ended. *HD2* was found as the only gene on scaffold_812. The *HD1* query returned nothing.

Investigating synteny of the core HD locus in sequenced *Amanita* genomes

The coordinates of the nine genes on either side of *HD1* and *HD2* in *A. thiersii* Skay4041 v1.0 and *A. muscaria* Koide v1.0 were recorded from their genome browsers on the Joint Genome Institute of the Department of Energy. Using the *A. muscaria* protein sequences as queries, matches were obtained from *A. brunescens, A. polypyramis, A. inopinata* (Hess *et al.* 2014) and *A. jacksonii* (van der Nest *et al.* 2014) assemblies in a local BLAST database (Supplementary Materials). The BLAST hits were used to construct a synteny diagram in Chromomapper (Niculita-Hirzel *et al.* 2008).

PCR to determine whether scaffold containing *HD2* is contiguous with rest of HD locus

PCR was carried out in order to determine whether the HD locus region was contiguous or disrupted, perhaps during whatever event made *HD1* appear missing. The objective was to amplify over the expected location of *HD1*. A PCR product that only accounted for the known sequence around the primers would indicate that *HD1* was truly absent.

To obtain primers that annealed in *HD2, MIP*, and the hypothetical protein predicted to be adjacent to *HD2* (see Fig 4 for schematic), their sequences were used as JGI tblastn queries the *Ath* Skay genome. Once unique areas of the genes were identified, primers were designed for them using Primer3 (Rozen & Skaletsky 2000). A list of the designed primers is in the Supplementary Materials.

Each PCR reaction contained the following reagents at the following concentrations: 0.05 U/μL Taq polymerase, 1 μM forward primer, 1 μM reverse primer, 0.75 mM dNTPs, 1x (2.5mM MgCl2) buffer. The PCR reaction used the following temperature profile: 94℃ for 30 seconds; 35 cycles of 94℃ for 15 seconds, 58℃ for 30 seconds, and 65℃ for 200 seconds; 65℃ for 10 minutes; 4℃ until product removed from the thermocycler.

Interpreting *Amanita*BASE called variants for presence or absence of *HD1* and *HD2*

While examining the vcf file that contained calls for 86 *A. phalloides* specimens and three *A. thiersii/A. foetens* specimens for variants in *A. phalloides*, it was observed that almost all data for variant sites in *HD1* and *HD2* were marked as missing, with '.' entries instead of values.

Results

The *Ath* Skay4041 assembly does not contain *HD1*

The *A. thiersii* Skay 4041 reference genome indicates that the sequenced isolate was missing *HD1*, the key homeodomain protein that contains the nuclear localization domain. It was not identified by BLAST or by synteny. This may be because the region did not align well to *A. phalloides.*

Synteny analysis reveals core HD locus genes present and syntenic, except *HD1*

Figure 3.3 shows that, in all of the *Amanita* genomes searched, synteny of the core HD locus was maintained, except for the loss of *HD1* in *A. thiersii.*

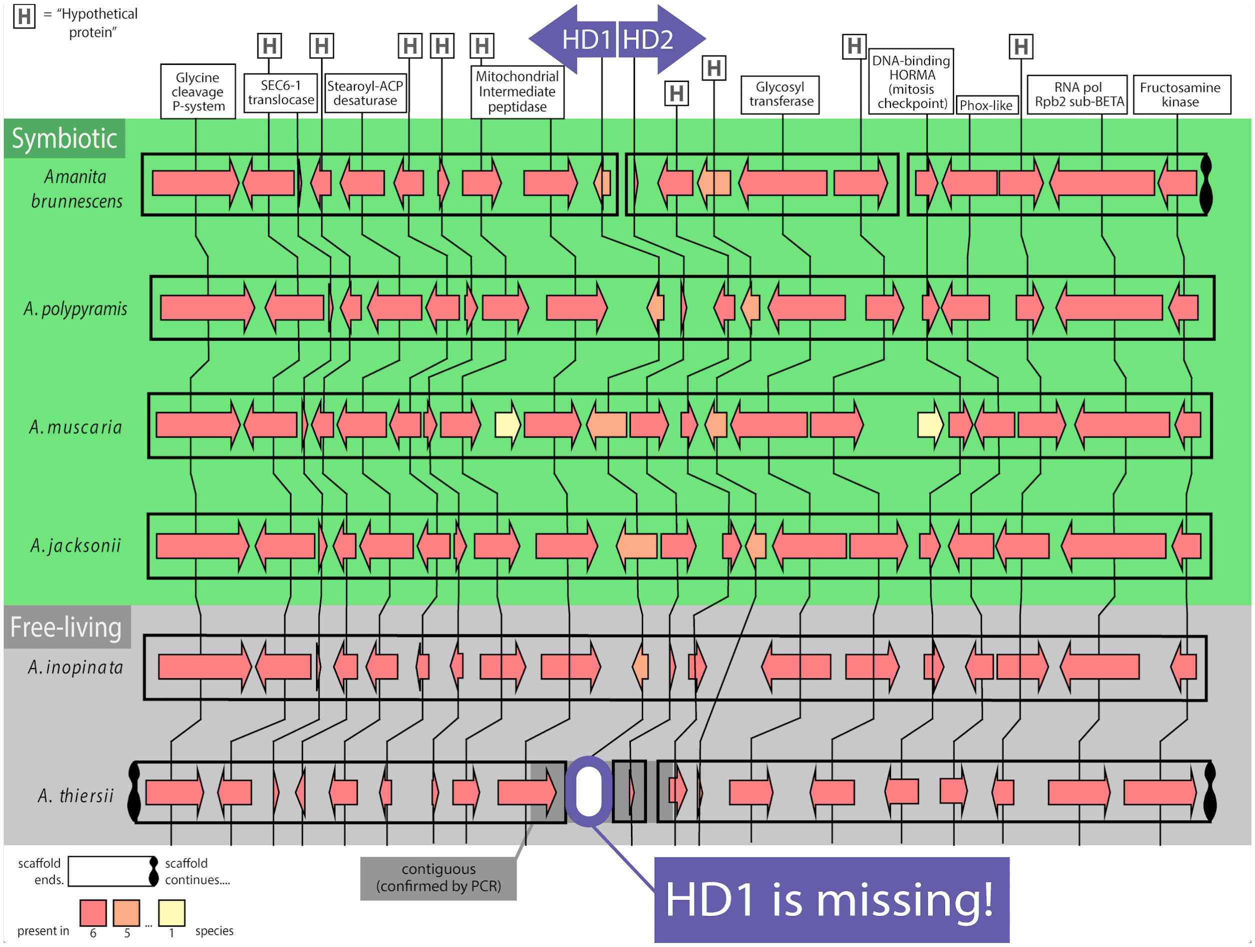


Figure 3.3: A synteny diagram of the core HD locus, centered on *HD1* and *HD2*. Despite small scaffolds in some of the assemblies, the overwhelming message here is of conserved synteny within *Amanita.*

PCR confirms *Ath* Skay4041 scaffold containing *HD2* is contiguous with rest of HD locus

PCR showed that the regions normally flanking *HD1* were contiguous (Figure 3.4). Strangely, there does not seem to be a difference between primer 1 and primer 2 even though the -2 -> 2 fragment should be longer. This result was never repeated. Figure 3.4 shows the only time amplification across the scaffold break was achieved. The region seems to resist PCR and eludes sequencing, possibly due to repetitive sequences surrounding the assembly breakpoints. It's possible the result in Figure 3.4 was some sort of accident, but the fragment size is within the predicted range and primer 2 may simply bind closer to primer 1 in *Ath* Skay4041 than was predicted based on the well-assembled HD loci of the other *Amanita* genomes.

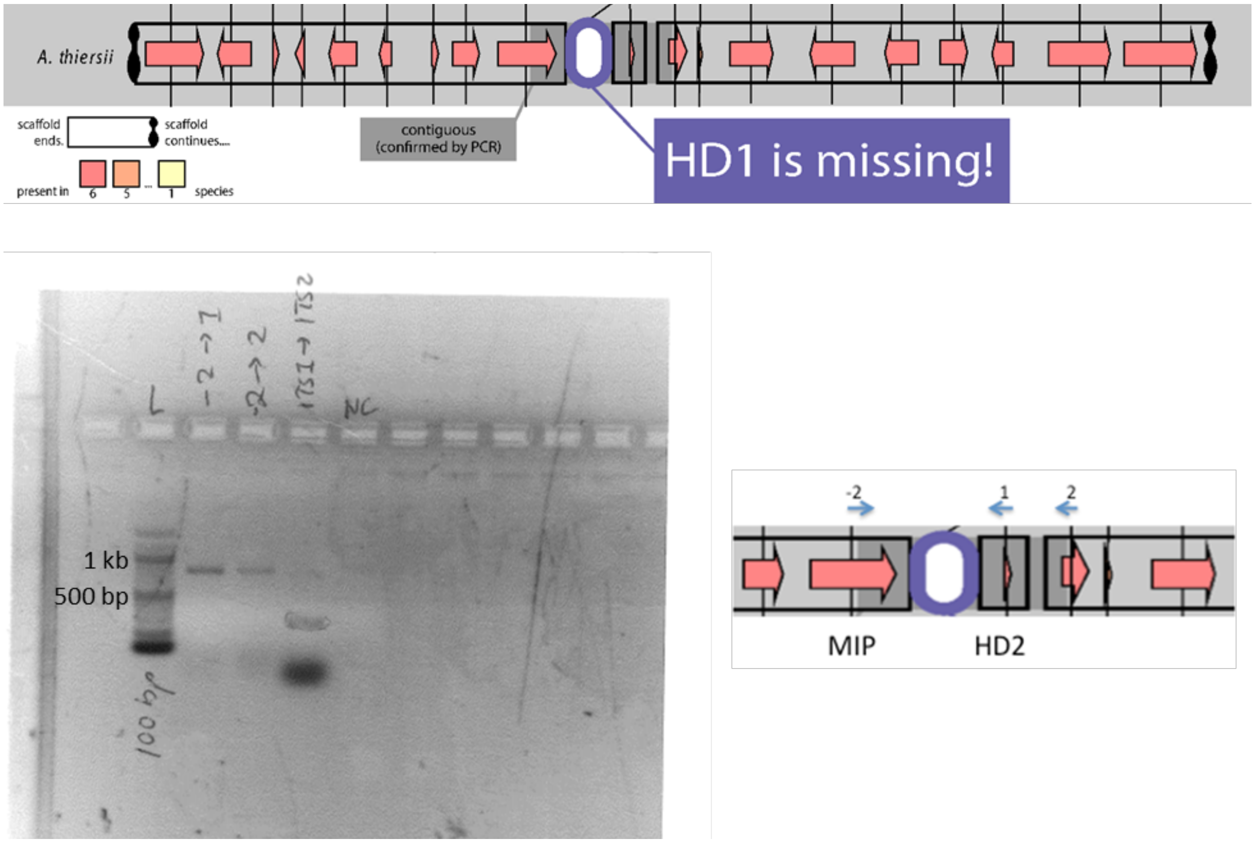


Figure 3.4: PCR confirms that the scaffold that contains *HD2* is contiguous with the rest of the mating locus. The PCR product is ~800 bp and appears too short to contain *HD1* (expected to be ~800-1200 bp plus intergenic space). Strangely, there does not seem to be a difference between primer 1 and primer 2 even though the -2 -> 2 fragment should be longer. *Note that the synteny diagram is not to scale.*

HD1 and *HD2* appear absent in other sequenced *A.thiersii* genomes

*Amanita*BASE's fully sequenced *A. thiersii* genomes also appear to be missing *HD1* judging by GATK-produced variant call file (.vcf) (Supplementary Materials) (DePristo *et al.* 2011). Out of 250 variants identified by GATK across 722-bp *HD1*, 10175 has only 5 variants assigned, all contiguous over a 9 nucleotide space (Contig87:399248-399256), although each appears to be high quality, with 14-15 reads supporting each of the called alleles, and each were able to be assigned to phase groups. *HD1* in 10801 has 4 variants assigned, all contiguous over an 8 nucleotide space (Contig87:398740-398747) and with only a single read supporting each call. 10801's variants were also phased. *HD1* in 10802 had no variants called and consisted entirely of missing data.

Interestingly, most of *HD2* is also missing from these *A. thiersii* specimens in the vcf. Out of 96 variants identified by GATK across 311-bp *HD2*, 10175 has 10 variants assigned, contiguous except for one site with only missing data (Contig87:397807) across Contig87:397805-397841. Each called site has three reads supporting the call. *HD2* in 10801 has only three called sites, the three variants directly following those that could be called in 10175, Contig87:397857-397863, with 2 reads supporting each call. As in *HD1*, 10802 did not have any variants mapped to *HD2*.

All of the *A. thiersii* specimens had low mapping rates generally because they were the minority (3/89) included on a majority *A. phalloides* GATK joint-calling run (Van der Auwera 2018). Including the *A. thiersii* individuals did not negatively affect the accuracy of the calls for the *A. phalloides* samples (*pers. comm.* Allison Schultz), but the calls aren't necessarily highly accurate for *A. thiersii* variant sites. Additionally, and consistent with the suggestion in Chapter 2 that *HD2* evolves more rapidly than *HD1*, variant-calling was more difficult at *HD2* than *HD1* across the board, with a much higher proportion of variants called based on user-defined filtering parameters than by clearly passing GATK quality checks. However, there are reads mapped nearby to *HD1* and *HD2* on contig 87, so it's not the case that these specimens' reads completely failed to map to the HD locus of *A. phalloides* 10511 reference genome.

*Amanita*BASE specimen 10802 is a culture derived from the *Ath* Skay4041, the same culture that was used in the JGI assembly (Grigoriev *et al.* 2014). In that assembly, *HD1* is nowhere to be found but *HD2* is present on scaffold_812. From this vcf, there is no evidence that either *HD1* or *HD2* is present in 10802. Again, this may simply be an artifact of imperfect alignment of *A. thiersii* reads to an *A. phalloides* assembly. 10802 is a monokaryotic culture and so does not require *HD1* or *HD2* to stay alive. It is possible that even more of the HD locus has been lost during the 5-6 years 10802 has spent in culture since it was originally sequenced.

Discussion

As I have shown above, *HD1* appears to be absent from the *Ath* Skay 4041 assembly and the three *AmanitaBASE A. thiersii/A. foetens* genomes. *HD1* is missing from the *Ath* Skay 4041 assembly. Synteny analysis shows that the core HD locus genes are present in *Ath* Skay 4041, all except for *HD1*. PCR confirms that *Ath* Skay 4041 scaffold_812 which contains *HD2* is contiguous with the rest of the HD locus. (Although the PCR was finicky and sequencing never worked, the fragment in Figure 3.4 was within the predicted size range and repetitive sequences indicated in the assembly may have caused the PCR and the sequencing problems.) *HD1 and HD2* appear absent from newly sequenced *Amanita*BASE *A. thiersii/A. foetens* genomes, even though *HD2* is found in *Ath* Skay 4041, which was sequenced from the same individual as 10802 5-6 years earlier. These results are interesting; however, they cannot be taken at face value. It is not clear what we would *expect* to see when looking at a minority species (3/89) in a GATK joint variant calling run and at a locus that may harbor difficult-to-map haplotypes due to ancient diversity.

Without *HD1*, tetrapolarity doesn't work. If *HD1* really is missing, it implies that *A. thiersii* is reproducing in an alternate way. The *HD1-HD2* heterodimer is the transcription factor that initiates the dikaryon, the sexual life stage. It has too many targets to simply be replaced or subverted. We know *A. thiersii* from its plentiful mushrooms, which are made only in the dikaryon stage, a stage that is supposed to require *HD1*. I don't know how *A. thiersii* can continue to enter the dikaryon stage and make mushrooms while the *Ath* Skay 4041 genome apparently lacks *HD1*, but have identified several possibilities:

1. The *Ath* Skay 4041 assembly is flawed, that individual *did* possess *HD1*, and *A. thiersii* is tetrapolar like other Basidiomycetes. That *HD1* is missing in the *AmanitaBASE A. thiersii* variant calls is not strong evidence it is actually missing in those individuals, either. There may simply be difficulties with amplifying or sequencing *HD1*. (Those difficulties may still be related to changes, *cis* or *trans*, that also affect life history.)
2. *HD1* (and possibly *HD2*) are apt to get lost over time in culture. The wild organism was tetrapolar.
3. *HD1* is lost, but the HD1-HD2 heterodimer is being approximated by a modified HD2-HD2 homodimer or a chimeric protein. There is an example of *Coprinus cinereus* in the lab with a chimeric HD1-HD2 protein, in which the C-terminal domains of *HD1* that are required for localization to the nucleus and activation were translocated to *HD2*, leading to constitutively active HD loci (Kues *et al.* 1994).
4. A fragment of HD1 or another protein is binding HD2's variable domain to make a sufficient transcription factor, or it has mutated to have enough DNA binding affinity that HD1 is not required. The HD2 homeodomain is more important for DNA-binding, and protein-protein interaction at the C-terminal variable domain appears to induce conformational changes that make the binding sufficiently strong without the contribution of HD1's homeodomain (Asante-Owusu *et al.* 1996). This transcription factor would still be without HD1's N-terminal domain and a nuclear localization signal.
5. *HD1* is lost and *A. thiersii* is undergoing some kind of selfing or apomixis.
6. A *thiersii* is spreading clonally as a dikaryon.

If *HD1* is missing and the affected *A. thiersii* have some means of bypassing its role in mating, then there may be a change in their life history. A change in mating system has been known to accompany invasiveness (Baker 1955). Inbreeding, selfing, and cloning all save resources that would otherwise be spent finding and screening a mate as well as preserving combinations of genotypes that are suited to invasion (Stebbins 1957). Baker's Law states that invading species often have low genetic diversity at the margin of expansion, but it doesn't stipulate that either low diversity causes invasion or vice versa. Is low genetic diversity a cause of invasion, for example, allowing successful genotypes to remain unbroken by recombination (Stebbins 1957), or simply the result of rapid expansion of a few colonizing genotypes into a new niche? If the loss of HD1 marks a transition to a new mating system, perhaps with more study we will be able to pinpoint which came first: invasion or low diversity?

Future directions

- Is *A. thiersii* polymorphic for the presence/absence of *HD1*, or do all invasive isolates lack it? Collecting and testing more specimens will shed light.
- What are the sequences of *HD1* and *HD2* in *A. thiersii* specimens? Is *HD2* chimeric with portions of *HD1*?
- Following suggestion of Taylor *et al.* 2015, we can use the sequenced *A. thiersii* genomes to probe the degree of recombination restriction via linkage equilibrium decay distances. *A. thiersii*'s degree of recombination restriction can be compared to *A. phalloides* and other multiply sequenced fungal species, invasive and noninvasive, to probe the degree of inbreeding or clonality and investigate the relationship between low diversity and invasion.

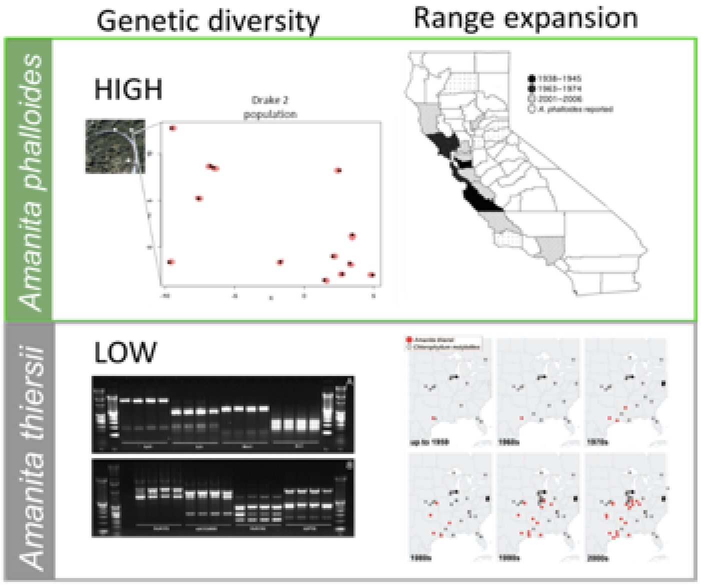


Figure 3.5: *A. thiersii* and *A. phalloides* are invasive in unique ways. *A. thiersii* follows Baker's Law while *A. phalloides* appears not to. Clockwise from top left: image provided by Anne Pringle, adapted from Pringle *et al.* 2009, adapted from Wolfe *et al.* 2012, from Wolfe *et al.* 2012.

- Within *Amanita*, both *A. phalloides* and *A. thiersii* are invasive, but in unique ways (Figure 3.5). *A. phalloides* is moving slowly and has retained high genetic diversity despite what must have been a population bottleneck when it arrived in California from its native range in Europe (Pringle *et al.* 2009). *A. thiersii* appears to be clonal over large stretches of its rapidly expanding range. *A. thiersii* follows Baker's Law, but *A. phalloides* appears not to. Is this difference due to their differences in sexuality, with *A. phalloides* being highly sexual and *A. thiersii* being nearly clonal? Perhaps because *A. phalloides* is ectomycorrhizal and *A. thiersii* is saprotrophic? *AmanitaBASE* was developed in part to be able to study the data relevant to understanding what distinguishes these two styles of invasion.

IV. Assembling the genome of Dr4M1: a test run for the sequencing and assembly of *Amanita* herbarium specimens

Co-authors: Jacqueline Hess, Inger Skrede, Anne Pringle

Introduction

Dr4M1 is a mushroom collected in 2004 from the Drake's Landing trail in Tomales Bay State Park in Point Reyes National Seashore, California, USA. In 2014, we sequenced and assembled Dr4M1's genome in what became something of a pilot experiment for *AmanitaBASE.* Given that *A. phalloides* does not culture, experimenting with sequencing directly from dikaryotic mushrooms was a key step in developing *Amanita* genomics. This was the first test of how well herbarium specimens of *A. phalloides* would sequence and assemble. Contamination was very low, which boded well for all the other *A. phalloides* specimens that had been stored in the same manner. Because of the Dr4M1 assembly's success, using 10+ year old specimens became a major feature of the *AmanitaBASE* 2016 sequencing scheme and added the valuable dimension of time. Even though most of its contigs remained unscaffolded after the first assembly pass, Dr4M1 exceeded our expectations and assured us that 10-year-old and probably older specimens of *A. phalloides* were suitable for further work.

Methods

All scripts may be found in https://github.com/elmoremh/Contributions-to-Amanita, also linked in Supplementary Materials.

Collection of the specimen

The sequenced specimen, "Drake 4, Mushroom 1" or Dr4M1 (*Amanita*BASE SpecimenID 10243), was collected along the Drake trail from a population marked with the coordinates N

38.054675, W 122.836545 in Pt. Reyes National Seashore, CA., in the fall of 2004. Fresh gill tissue was lyophilized for long-term storage using an ACME Mark V Lyophilizer.

DNA extraction

Approximately 50 mg of the lyophilized gill tissue from each mushroom was placed in a 2.0 ml microcentrifuge tube with 4 to 5 3mm glass beads and macerated using a MiniBeadbeater-8 (BioSpec Products Inc., Bartlesville, OK) set at 3/4 speed for one minute. Genomic DNA was extracted from lyophilized gill tissue using a Qiagen DNeasy Tissue kit, according to manufacturer's specifications with the following modifications: Incubation times with lysis buffer (Buffer ATL) were extended to one hour; after incubation, the tubes were spun in a microcentrifuge briefly to collect cellular debris at the bottom and the liquid was removed to a new tube before proceeding; incubation with buffer AE was extended to five minutes, and the final volume eluted was 200 ul. To avoid cross-contamination between samples all equipment was washed with dilute bleach solution, dedicated equipment and filtered tips were used for all extraction procedures, and extraction blanks were extracted along with samples.

Quantity and purity of genomic DNA was determined using a Nanodrop Spectrophotometer (Nanodrop Technologies, Wilmington, DE). Dr4M1 had a 260:280 ratio of 1.96 and 260:230 ratio of 1.72. Specimen Dr4M1 was chosen because its extracted DNA appeared relatively less fragmented than the others when run out on a gel (Figure 3.6)

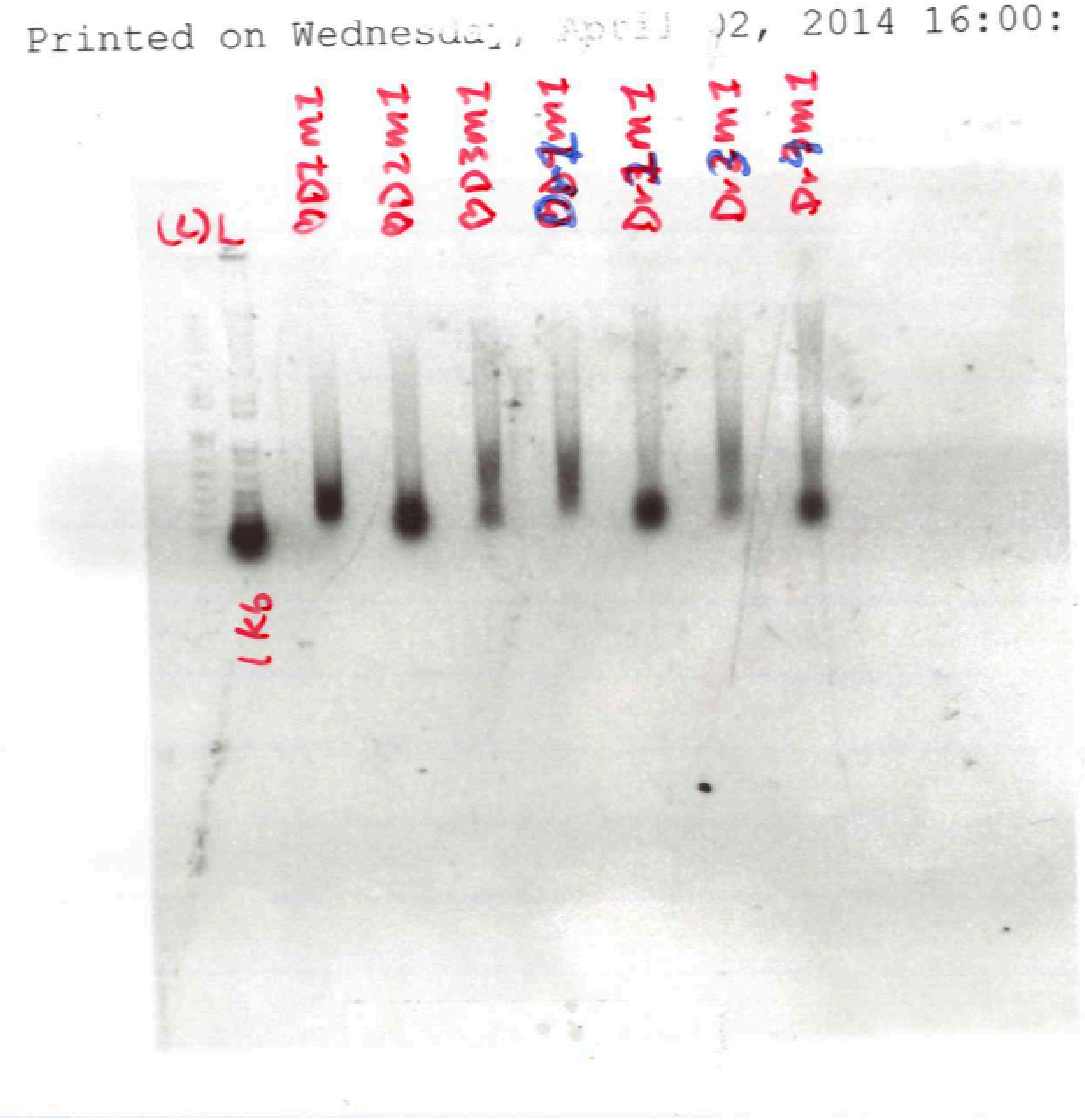


Figure 3.6: 1% agarose gel showing genomic DNA of several 10+ year old specimens of *A. phalloides*. Dr4M1 (furthest right) was chosen because it appeared to have more high molecular weight DNA fragments than the others.

Sequencing

1 µg of Dr4M1 genomic DNA was sequenced on a single lane of Illumina MiSeq with 500 bp insert to yield 100 bp paired end reads. Bioanalyzer sample quality control, library

preparation, and sequencing was completed at the Biopolymer Facility at Harvard Medical School.

kmer Analysis

The script ErrorCorrectReads.pl from ALLPATHS-LG assembler software package was used to obtain kmer spectra. Plots were made in Excel (Supplementary Materials).

de novo Assembly

The reads were trimmed with trimmomatic using the following parameters: LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:50.

We attempted to align the *A. phalloides* reads to the *A. brunescens* assembly (Hess *et al.* 2014) with bowtie2 so that the mapped reads could form the basis for assembling the genome. *A. brunescens* was the closest *Amanita* to *A. phalloides* that had already been sequenced, but mapping percentage was an extremely poor 0.45%. The resulting alignment was not enough to be useful going forward in the assembly process.

After experimenting with several assemblers and kmer hashes, the current assembly of Dr4M1 was made with SPAdes 3.1.1 (Bankevich *et al.* 2012) using default parameters. SPAdes was chosen in large part because of its tolerance for highly heterozygous “diploids.”

BLAST contamination screen

After the genome had been assembled into contigs, the assembly was given as input to Jacqueline Hess’s script BLAST_contamination_multi-thread.py, which divides the assembly into 10 kb chunks and submits them as queries to the NCBI-BLAST nr database. The results were then assessed manually using Excel filters. Any taxid within Kingdom Fungi was considered not to be contamination. Contigs that were contained genuine-appearing hits were removed.

Identifying MSDIN genes in *A. phalloides* assembly Dr4M1

I queried the NCBI-BLAST (Altschul *et al.* 1990) non-redundant protein sequences (nr) database with the following MSDIN amino acid sequence from Hallen *et al.* (2007)'s *Amanita bisporigera* genome. The header preserves the NCBI output.

```
>gi|159033014|gb|ABW87778.1| putative MSDIN-like protein 6 [Amanita
bisporigera]
MSDINGTRLPIPGLIPLGIPCVSDDVNPTLTRGER
```

This search returned a cloned sequence from an *A. phalloides* individual (Li *et al.* 2014):

```
>gi|559795661|gb|AHB18711.1| MSDIN-like protein [Amanita phalloides]
MSDINATRLPIFWFIYFPCVGDNVDNTLTRGER
```

The *A. phalloides* amino acid sequence was used to as a tblastn query a nucleotide blast database of the Dr4M1 assembly (BLAST script in Supplementary Materials).

Results

kmer analysis

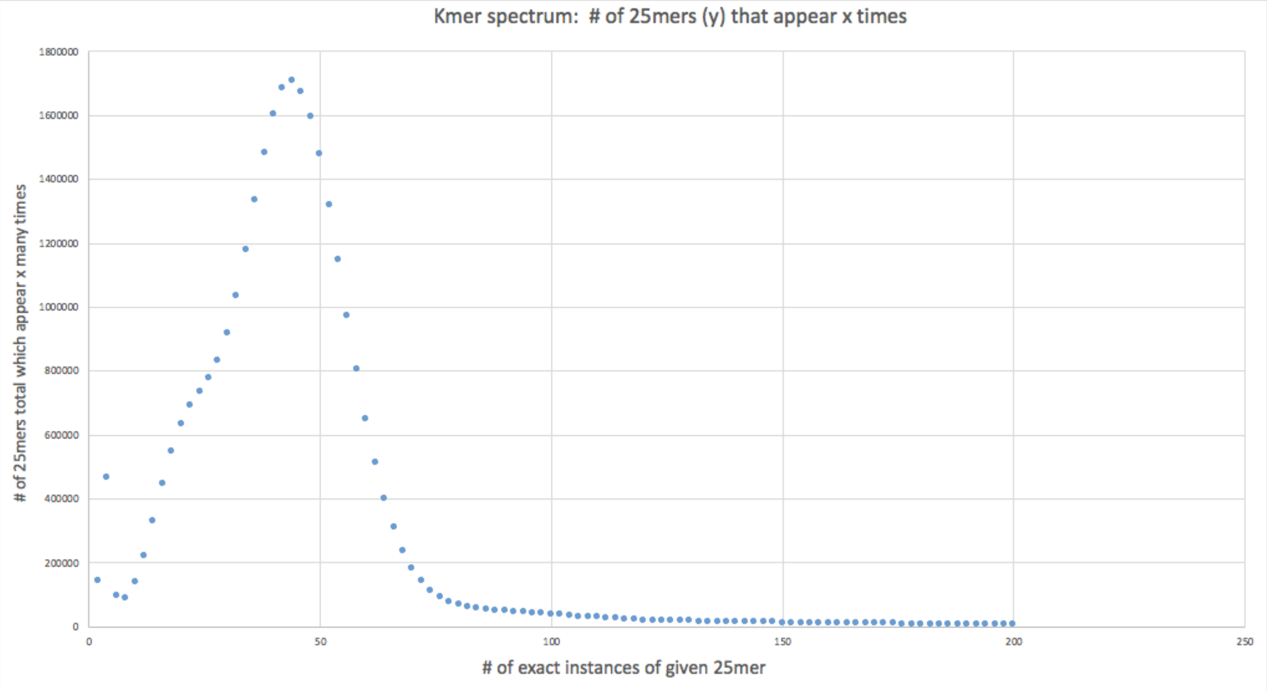


Fig 7: This kmer spectrum plot shows the distribution of 25mers (y axis) that appear x number of times (x axis) in the Illumina reads of Dr4M1.

The largest peak shows homozygous coverage, because the vast majority of 25mers in a genome are unique and should therefore exact matches should be due to sequencing depth. Qualitative analysis of the kmer spectra (Figure 3.7) indicated that homozygous coverage was ~48x. Heterozygosity, indicated by the height of the heterozygous coverage peak to the left of the main peak, was high. The size of the genome was estimated at 43 MB. SNP rate was estimated at 1/40,000 bp. By analyzing the tail of the kmer distribution-- 25mers that occur many, many times-- repeat content was estimated at about 40%.

The assembly was surprisingly successful

We implemented the assemblathon_stats.pl script (Bradnam, last updated 10-13-2011) to obtain common statistics and metrics used to describe the quality of genome assemblies (Table 1).

Table 1: Statistics describing the Dr4M1 SPAdes assembly using the assemblathon_stats.pl script. Statistics referring to sequences are in bp.

```
                     Number of scaffolds       95863
                  Total size of scaffolds   46741056
                         Longest scaffold    1255653
                        Shortest scaffold         56
              Number of scaffolds > 1K nt       4365   4.6%
             Number of scaffolds > 10K nt        417   0.4%
            Number of scaffolds > 100K nt         28   0.0%
              Number of scaffolds > 1M nt          1   0.0%
             Number of scaffolds > 10M nt          0   0.0%
                       Mean scaffold size        488
                     Median scaffold size        147
                      N50 scaffold length       2342
                       L50 scaffold count       1303
                              scaffold %A      27.04
                              scaffold %C      22.93
                              scaffold %G      22.95
                              scaffold %T      27.02
                              scaffold %N       0.06
                      scaffold %non-ACGTN       0.00
          Number of scaffold non-ACGTN nt          0

 Percentage of assembly in scaffolded contigs      14.3%
Percentage of assembly in unscaffolded contigs      85.7%
```

```
                      Average number of contigs per scaffold        1.0
Average length of break (>25 Ns) between contigs in scaffold        211

                                           Number of contigs      95987
                              Number of contigs in scaffolds        230
                          Number of contigs not in scaffolds      95757
                                       Total size of contigs   46714815
                                              Longest contig    1255653
                                             Shortest contig         56
                                   Number of contigs > 1K nt       4489   4.7%
                                  Number of contigs > 10K nt        457   0.5%
                                 Number of contigs > 100K nt         31   0.0%
                                   Number of contigs > 1M nt          1   0.0%
                                  Number of contigs > 10M nt          0   0.0%
                                            Mean contig size        487
                                          Median contig size        148
                                           N50 contig length       2290
                                            L50 contig count       1424
                                                   contig %A      27.06
                                                   contig %C      22.95
                                                   contig %G      22.96
                                                   contig %T      27.04
                                                   contig %N       0.00
                                           contig %non-ACGTN       0.00
                               Number of contig non-ACGTN nt          0
```

Though these numbers are not impressive in and of themselves, this was only the result of one round of assembly and without any polishing. We considered even these numbers a success because of that fact that Dr4M1 was a 10-year-old herbarium specimen that wasn't stored particularly carefully or its DNA prepared in any special way. This level of assembly is usually sufficient to find proteins, which I did next.

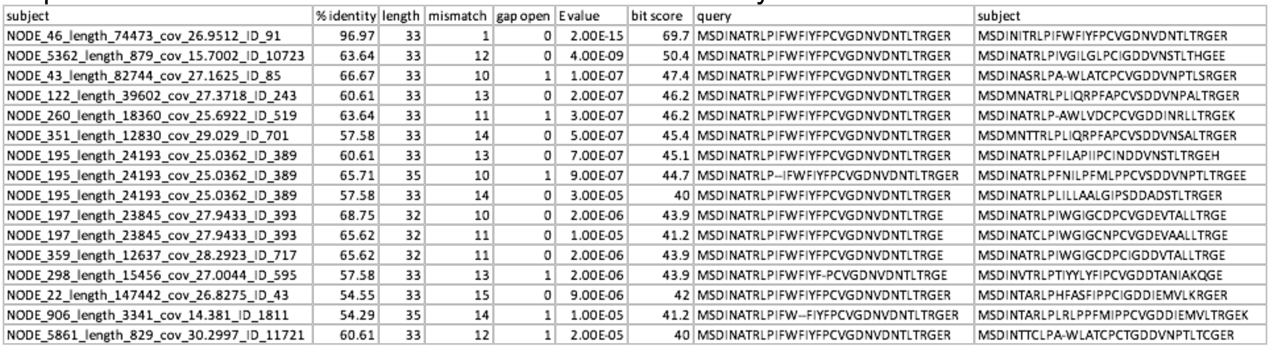
Very little contamination found

Though much contamination by microbes might have been expected in a specimen kept in a box at room temperature for 10 years, remarkably little was found. Most taxids outside of Fungi were the result of contamination or mislabeling in GenBank rather than in Dr4M1. For example, the sample whose BLAST hit was labeled “Camel” almost certainly had fungal contamination of its own. The most commonly repeated taxids of hits were “Enterobacter” and “Enterobacter phage.” Contigs made up of Enterobacter hits were removed. The BLAST screen reinforced what the kmer spectra had indicated, that Dr4M1 was not significantly contaminated with other organisms' DNA.

Novel MSDIN genes identified in Dr4M1

BLAST tblastn search identified 16 MSDIN genes in Dr4M1 (Table 2), several of which were not in the BLAST nr database at the time.

Table 2: *A. phalloides* MSDIN query of Dr4M1 assembly BLAST results table. Subject sequences "NODE…" are scaffolds of the Dr4M1 assembly.



subject	% identity	length	mismatch	gap open	E value	bit score	query	subject
NODE_46_length_74473_cov_26.9512_ID_91	96.97	33	1	0	2.00E-15	69.7	MSDINATRLPIFWFIYFPCVGDNVDNTLTRGER	MSDINITRLPIFWFIYFPCVGDNVDNTLTRGER
NODE_5362_length_879_cov_15.7002_ID_10723	63.64	33	12	0	4.00E-09	50.4	MSDINATRLPIFWFIYFPCVGDNVDNTLTRGER	MSDINATRLPIVGILGLPCIGDDVNSTLTHGEE
NODE_43_length_82744_cov_27.1625_ID_85	66.67	33	10	1	1.00E-07	47.4	MSDINATRLPIFWFIYFPCVGDNVDNTLTRGER	MSDINASRLPA-WLATCPCVGDDVNPTLSRGER
NODE_122_length_39602_cov_27.3718_ID_243	60.61	33	13	0	2.00E-07	46.2	MSDINATRLPIFWFIYFPCVGDNVDNTLTRGER	MSDMNATRLPLIQRPFAPCVSDDVNPALTRGER
NODE_260_length_18360_cov_25.6922_ID_519	63.64	33	11	1	3.00E-07	46.2	MSDINATRLPIFWFIYFPCVGDNVDNTLTRGER	MSDINATRLP-AWLVDCPCVGDDINRLLTRGEK
NODE_351_length_12830_cov_29.029_ID_701	57.58	33	14	0	5.00E-07	45.4	MSDINATRLPIFWFIYFPCVGDNVDNTLTRGER	MSDMNTTRLPLIQRPFAPCVSDDVNSALTRGER
NODE_195_length_24193_cov_25.0362_ID_389	60.61	33	13	0	7.00E-07	45.1	MSDINATRLPIFWFIYFPCVGDNVDNTLTRGER	MSDINATRLPFILAPIIPCINDDVNSTLTRGEH
NODE_195_length_24193_cov_25.0362_ID_389	65.71	35	10	1	9.00E-07	44.7	MSDINATRLP--IFWFIYFPCVGDNVDNTLTRGER	MSDINATRLPFNILPFMLPPCVSDDVNPTLTRGEE
NODE_195_length_24193_cov_25.0362_ID_389	57.58	33	14	0	3.00E-05	40	MSDINATRLPIFWFIYFPCVGDNVDNTLTRGER	MSDINATRLPLILLAALGIPSDDADSTLTRGER
NODE_197_length_23845_cov_27.9433_ID_393	68.75	32	10	0	2.00E-06	43.9	MSDINATRLPIFWFIYFPCVGDNVDNTLTRGE	MSDINATRLPIWGIGCDPCVGDEVTALLTRGE
NODE_197_length_23845_cov_27.9433_ID_393	65.62	32	11	0	1.00E-05	41.2	MSDINATRLPIFWFIYFPCVGDNVDNTLTRGE	MSDINATCLPIWGIGCNPCVGDEVAALLTRGE
NODE_359_length_12637_cov_28.2923_ID_717	65.62	32	11	0	2.00E-06	43.9	MSDINATRLPIFWFIYFPCVGDNVDNTLTRGE	MSDINATRLPIWGIGCDPCIGDDVTALLTRGE
NODE_298_length_15456_cov_27.0044_ID_595	57.58	33	13	1	2.00E-06	43.9	MSDINATRLPIFWFIYF-PCVGDNVDNTLTRGE	MSDINVTRLPTIYYLYFIPCVGDDTANIAKQGE
NODE_22_length_147442_cov_26.8275_ID_43	54.55	33	15	0	9.00E-06	42	MSDINATRLPIFWFIYFPCVGDNVDNTLTRGER	MSDINTARLPHFASFIPPCIGDDIEMVLKRGER
NODE_906_length_3341_cov_14.381_ID_1811	54.29	35	14	1	1.00E-05	41.2	MSDINATRLPIFW--FIYFPCVGDNVDNTLTRGER	MSDINTARLPLRLPPFMIPPCVGDDIEMVLTRGEK
NODE_5861_length_829_cov_30.2997_ID_11721	60.61	33	12	1	2.00E-05	40	MSDINATRLPIFWFIYFPCVGDNVDNTLTRGER	MSDINTTCLPA-WLATCPCTGDDVNPTLTCGER

Biochemical population genomics of *A. phalloides*

The above work, along with the AmanitaBASE *whole genome dataset and intellectual contributions, resulted in authorship on the following manuscript:*

Harrow S… **Elmore MH**… Pringle A. (manuscript) *Biochemical population genomics: diversity in the MSDIN genes of* Amanita phalloides.

Using the *AmanitaBASE* whole genome data set (86 *A. phalloides* genomes, 67 from California and 19 from Europe), Samantha Harrow, a graduate student in the Pringle Lab, built a bioinformatic pipeline to search for MSDIN protein sequences. Because Harrow was interested in synteny, copy number variation, and genes that varied by only a few amino acids, we agreed that using the reference assembly (10511) and creating alternate reference as I had in Chapter 2 would not do. Therefore, Harrow's pipeline starts with *de novo* assembly of each Illumina-only genome followed by BLAST search of an MSDIN protein query to a translated nucleotide genome (tblastn) to identify MSDIN candidate genes.

Harrow has found that although an individual *Amanita* genome may contain upwards of 30 MSDIN genes, there is generally little overlap in MSDIN between species. The *AmanitaBASE A. phalloides* dataset provides an excellent opportunity to study how segregated MSDIN complements are between populations and between continents. Harrow has already found that MSDIN complements in the invasive California *A. phalloides* are more similar to each other than those than those of the European specimens, which may owe to their different demographic history or reflect a response to different environments. Harrow plans HPLC-MS metabolomics to determine whether and how each of the MSDIN genes is expressed and its biological activity, as well as tests of selection, such as dN/dS, to determine whether selection is driving differences in MSDIN complements between populations.

Discussion

The sequencing and the assembly were much more successful than we had predicted. We feared contamination that overwhelmed the DNA of Dr4M1, or fragmentation and degradation so bad that assembly was thwarted. Had that happened, sequencing would still have provided novel information about genome size, repeat content, SNP rate, and heterozygosity via kmer analysis. Instead, we were able to assemble Dr4M1, and even though we did not continue to assemble or polish, we got enough to BLAST search for MSDIN proteins. Dr4M1 was the first step towards the 2016 mass *AmanitaBASE* sequencing, and both led to the way to Samantha Harrow's MSDIN gene-extraction pipeline and planned metabolomics of the MSDIN toxins.

Overall summary

In my work, patterns of diversity in *Amanita* spp., in their ecology, population structure, toxins, and genomes, shed light on life history which in turn gives insight into ecology and evolution.

Section I was a theoretical framework for all Basidiomycetes, but it ended up framing my ideas about the life history of *Amanita phalloides.* Small, short-lived dikarya may indicate lower nuclear autonomy and greater organismality (following Queller & Strassman 2009).

Section II was investigation of diversity in *A. phalloides* across the genome and the globe. Golan *et al.'*s findings indicate that *A. phalloides* has exclusively singleton sporocarp genets that do not persist between fruiting seasons. A small mycelium that quickly produces a mushroom and dies off may indicate that the nuclei of the dikaryon have incentives which are well-aligned, leading to more cooperative, organismal behavior. Altogether, *A. phalloides* is small-bodied, ephemeral, annual, and may prefer disturbed areas (ruderal)-- several of the characteristics of Baker (1965)'s classic description of a weed. Baker theorized that weed characteristics predisposed certain plants to invasion.

Section III investigated a potential disturbance in the *A. thiersii* HD mating locus, the loss of *HD1*, which may be related to its sudden emergence as an invasive. If *HD1* is missing and the affected *A. thiersii* have some means of bypassing its role in mating, then there may be a change in their life history. A change in mating system has been known to accompany invasiveness (Baker 1955). Inbreeding, selfing, and cloning all save resources that would otherwise be spent finding and screening a mate as well as preserving genotypes that are suited to invasion (Stebbins 1957). Again, there is a connection between life history evolution and invasiveness in *Amanita*, but in a very different way from *A. phalloides*. *A. thiersii* is a saprotrophic *Amanita* that has invaded very quickly, covering the Eastern US from Texas all the way up to Illinois since its discovery in Texas in 1952. *A. phalloides* has invaded the US West

Coast more slowly since its earlier introduction, some time before 1938, and it is ectomycorrhizal (symbiotic with trees). We can speculate that perhaps the ectomycorrhizal niche requires greater genetic diversity, perhaps related to the need to compete with other fungi to colonize the root tip. Saprotrophic *A. thiersii* may have had fewer constraints on sweeping through its new habitat with lower genetic diversity. Whatever the case, the *Amanita*BASE genomes and metadata are a rich source of comparative data for further study of the ecological correlates of invasion in *Amanita*, already arguably the best studied genus for invasive fungi (Pringle *et al.* 2009; Wolfe *et al.* 2012).

Section IV describes a test-run of sequencing a 10-year-old herbarium specimen, which made the later massively parallel 2016 *Amanita*BASE sequencing run possible, and which identified new MSDIN toxins. Further work by Samantha Harrow surveying the diversity of MSDIN toxins across *A. phalloides*' range will allow us to ask deeper questions about the still mysterious role of MSDIN toxins in the ecology of *A. phalloides. Amanita*BASE is a foundation for deeper genomic and ecological work on all of the questions raised here going forward.

Throughout this dissertation, I have contributed to knowledge about *Amanita,* its life history, and evolution and ecology. In this chapter, I describe projects that were not taken to publication or in which I was not the first author. In developing *Amanita* as an ecological genomics systems for the study of invasion, tetrapolar mating dynamics, amatoxins, and beyond, I have laid the groundwork for a large indirect contribution as others take charge of their own projects within the system. I am pleased that my dissertation includes examples of all of these different kinds of contributions to science-- as a first author, a collaborator, and a producer of data and resources that others will use-- because all are crucial to a healthy scientific community.

Acknowledgments

Deep thanks to the Universitetet i Oslo Mycology Group for hosting me while I worked with Inger Skrede and Jacqueline Hess in Oslo, Norway. The computations in this study were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. Thanks to the National Science Foundation for funding grant 1021606 to Anne Pringle and NSF Graduate Research Fellowship 2013170866 to Holly Elmore.

References

Aanen D. K., Kuyper, T. W., Debets, A. J. M. & Hoekstra, R. F. The evolution of non-reciprocal nuclear exchange in mushrooms as a consequence of genomic conflict. *Proc. R. Soc. B Biol. Sci.* **271**, 1235–1241 (2004).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

Anderson, J. B. *et al.* Clonal evolution and genome stability in a 2500-year-old fungal individual. *Proc. R. Soc. B Biol. Sci.* **285**, 1–6 (2018).

Asante-Owusu, R. N., Banham, A. H., Böhnert, H. U., Mellor, E. J. C. & Casselton, L. A. Heterodimerization between two classes of homeodomain proteins in the mushroom Coprinus cinereus brings together potential DNA-binding and activation domains. *Gene* **172**, 25–31 (1996).

Baker, H. G. Characteristics and modes of origin of weeds. in *The genetics of colonizing species: Proc. 1st Internat, Union biol Sci., Asilomar, California.* 147–172 (1965).

Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

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

Booth, A. Symbiosis, selection, and individuality. *Biol. Philos.* **29**, 657–673 (2014).

Bradnam, K. “assemblathon_stats.pl” last updated 10-13-2011 <https://github.com/elmoremh/Contributions-to-Amanita/blob/master/assemblathon_stats.pl>

Burt A. & Trivers R. *Genes in Conflict.* Cambridge, MA, Harvard University Press, 2006.

Casselton, L. A. Molecular recognition in fungal mating. *Endeavour* **21**, 159–163 (1997).

Casselton, L. A. & Olesnicky, N. S. Molecular genetics of mating recognition in basidiomycete fungi. *Microbiol. Mol. Biol. Rev.* **62**, 55–70 (1998).

Chaib De Mares, M. *et al.* Horizontal transfer of carbohydrate metabolism genes into ectomycorrhizal Amanita. *New Phytol.* **205**, 1552–1564 (2015).

Depristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–501 (2011).

Deshpande, V., Fung, E. D. K., Pham, S. & Bafna, V. Cerulean: a hybrid assembly using high thoughput short and long reads. *Algorithms Bioinform Lect Notes Com Sci.* **8126,** 349–350 (2013).

English, A. C. *et al.* Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS One* **7**, 1–12 (2012).

Falconer, R. E., Bown, J. L., White, N. A. & Crawford, J. W. Biomass recycling and the origin of phenotype in fungal mycelia. *Proc. R. Soc. B Biol. Sci.* **272**, 1727–1734 (2005).

Gordon, D. *et al.* of the Gorilla Genome. **344,** 1–21 (2016).

Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1513–1518 (2011).

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

Hallen, H. E., Luo, H., Scott-Craig, J. S. & Walton, J. D. Gene family encoding the major toxins of lethal Amanita mushrooms. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 19097–19101 (2007).
Harper, J. L. *Population biology of plants*. (1977).

Hess, J. *et al.* Transposable element dynamics among asymbiotic and ectomycorrhizal amanita fungi. *Genome Biol. Evol.* **6**, 1564–1578 (2014).

Johannesson, H. & Stenlid, J. Nuclear reassortment between vegetative mycelia in natural populations of the basidiomycete *Heterobasidion annosum*. *Fungal Genet. Biol.* **41**, 563–570 (2004).

Kajitani, R. *et al.* Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).

KateN. (howto) Discover variants with GATK - A GATK Workshop Tutorial. *GATK 3 User Guide*. (2016). <https://software.broadinstitute.org/gatk/documentation/article?id=7869>

Kohler, A. *et al.* Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. *Nat. Genet.* **47**, 410–415 (2015).

Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. 1–35 (2016). doi:10.1101/gr.215087.116.

Krueger, F. Trim Galore: https://github.com/FelixKrueger/TrimGalore

Kües, U. *et al.* A chimeric homeodomain protein causes self-compatibility and constitutive sexual development in the mushroom Coprinus cinereus. *EMBO J.* **13**, 4054–4059 (1994).
Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

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

Li, H. BFC: Correcting Illumina sequencing errors. *Bioinformatics* **31,** 2885–2887 (2015).

Li, P., Deng, W. & Li, T. The molecular diversity of toxin gene families in lethal Amanita mushrooms. *Toxicon* **83**, 59–68 (2014).
Lind, M., Stenlid, J. & Olson, Å. Genetics and QTL mapping of somatic incompatibility and intraspecific interactions in the basidiomycete Heterobasidion annosum s.l. *Fungal Genet. Biol.* **44**, 1242–1251 (2007).

Ma, L. *et al.* Defining individual size in the model filamentous fungus Neurospora crassa. *Proc. R. Soc. B Biol. Sci.* **283**, (2016).

Niculita-Hirzel, H. *et al.* Gene organization of the mating type regions in the ectomycorrhizal fungus Laccaria bicolor reveals distinct evolution between the two mating type loci. *New Phytol.* **180**, 329–342 (2008).

Nieuwenhuis, B. P., Nieuwhof, S. & Aanen, D. K. On the asymmetry of mating in natural populations of the mushroom fungus Schizophyllum commune. *Fungal Genetics and Biology* **56,** 25–32 (2013a).

Nieuwenhuis, B. P. S. *et al.* Evolution of uni- and bifactorial sexual compatibility systems in fungi. *Heredity (Edinb).* **111**, 445–455 (2013b).
Raper, J. *The Genetics of Sexuality in Higher Fungi*. New York, USA, The Ronald Press Company, 1996.

Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386 (2000).

Perez-Riverol, Y. *et al.* Ten Simple Rules for Taking Advantage of Git and GitHub. *PLoS Comput. Biol.* **12,** 1–11 (2016).

Pringle, A., Adams, R. I., Cross, H. B. & Bruns, T. D. The ectomycorrhizal fungus *Amanita phalloides* was introduced and is expanding its range on the west coast of North America. *Mol. Ecol.* **18**, 817–833 (2009).

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E. & Jones, S. J. M. ABySS : A parallel assembler for short read sequence data ABySS : A parallel assembler for short read sequence data. 1117–1123 (2009). doi:10.1101/gr.089532.108

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

Stebbins, G. L. Self Fertilization and Population Variability in the Higher Plants *The American Naturalist* **91**, 337–354 (1957).

Taylor, J. W., Hann-Soden, C., Branco, S., Sylvain, I. & Ellison, C. E. Clonal reproduction in fungi. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8901–8 (2015).

Torvalds, L. Initial revision of "git", the information manager from hell. (2005). <https://github.com/git/git/commit/e83c5163316f89bfbde7d9ab23ca2e25604af290>

Van der Auwera G. Germline variant discovery (SNPs + indels) (2018) <https://gatkforums.broadinstitute.org/gatk/discussion/11145/germline-short-variant-discovery-snps-indels>
van der Nest, M. A. *et al.* Draft genomes of Amanita jacksonii, Ceratocystis albifundus, Fusarium circinatum, Huntiella omanensis, Leptographium procerum, Rutstroemia sydowiana, and Sclerotinia echinophila. *IMA Fungus* **5**, 473–486 (2014).

Queller, D. C. & Strassmann, J. E. Beyond society: the evolution of organismality. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **364**, 3143–3155 (2009).

Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, (2014).

Wang, J. R., Holt, J., McMillan, L. & Jones, C. D. FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinformatics* **19,** 1–11 (2018).

Warren, R. L. *et al.* LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience* **4**, (2015).

Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution (N. Y).* **38**, 1358–1370 (2017).

Wgsim: https://github.com/lh3/wgsim

Wolfe, B. E., Kuo, M. & Pringle, A. *Amanita thiersii* is a saprotrophic fungus expanding its range in the United States. *Mycologia* **104**, 22–33 (2012).

Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* **6,** 1–9 (2016).

Appendix I
Supplementary materials

Link to electronic Word document (so you can follow links)

https://www.dropbox.com/s/c2lnndmfmaqzu8r/Elmore_dissertation_submitted_3-6-20.docx?dl=0

Chapter 1: *Amanita*BASE: a resource and model of ecological population genomic databases

All documents in this dedicated Chapter 1 Supplementary Materials Dropbox folder: https://www.dropbox.com/sh/kdpda16ipzsuniz/AADm2N8A8oarSK77wmQ79eOza?dl=0

Links:

- View-only G-Drive AmanitaBASE v3 Specimen Metadata sheet: https://docs.google.com/spreadsheets/d/1M43034L9533sQVBcFwZSG037HL8FOIkqSz_gowYzJno/edit?usp=sharing
- Scripts in github.com/elmoremh/Amanita-Population-Genomics

Supplementary Table 1: *Amanita*BASE SpecimenID number batches and their intended populations.

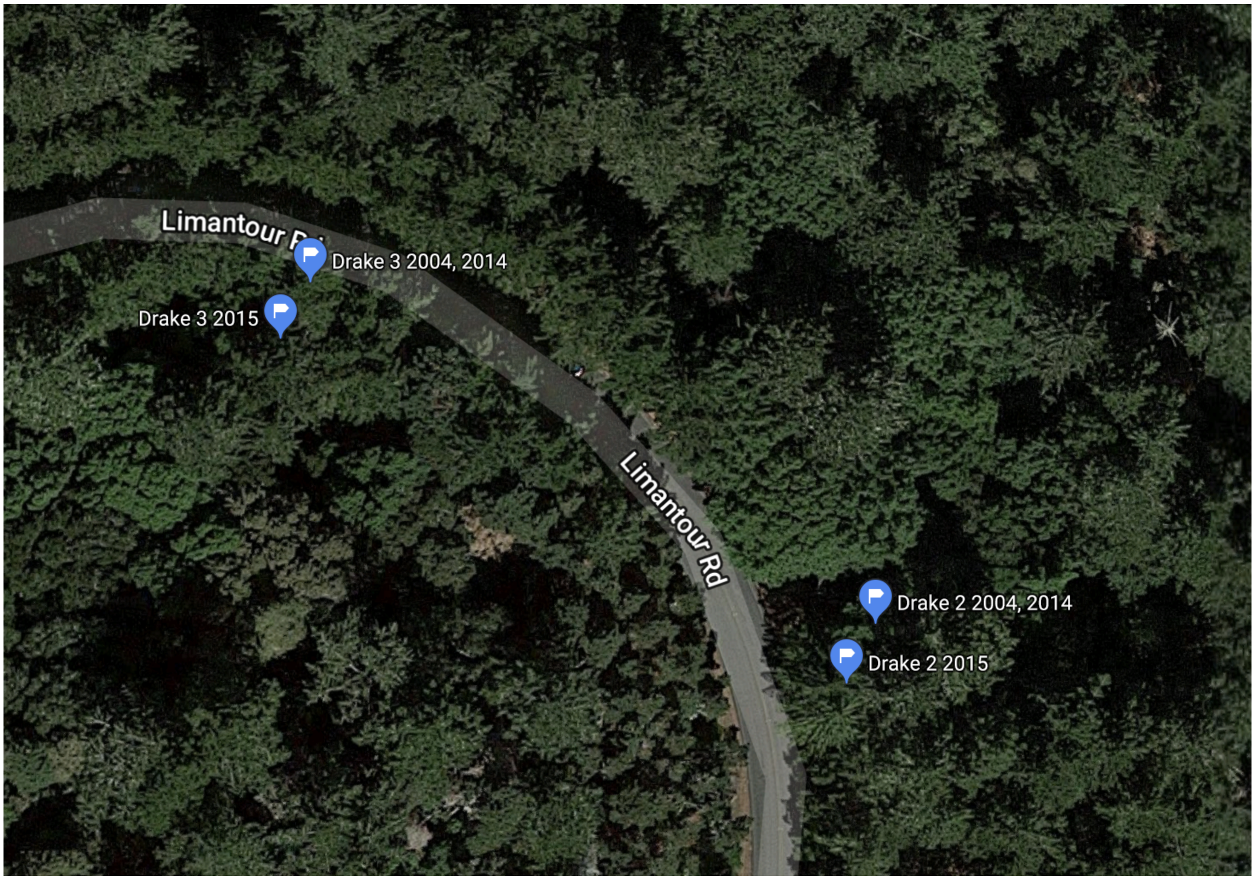
Numbers	Date assigned	Purpose
10001 - 10501	11/7/15	Accessioning of old (pre-2015) specimens

10502 - 10701	11/15/15	Susana Goncalves' Portuguese collecting expedition
10702 - 10800	12/2/15	Pt. Reyes 2015 collecting expedition
10801	2/14/16	Accessioning Michigan herbarium sample of *A. foetens*
10802	2/14/16	*A. thiersii* culture
10803 - 10807	3/16/16	Argentinian *A. thiersii* specimens
10817 - 10900	3/21/16	Reserved for cultures
10901 - 11499	12/11/17	Assigned to Cat Adams
11500 - 11650	10/3/18	Assigned to Jacob Golan for Jake's Landing population

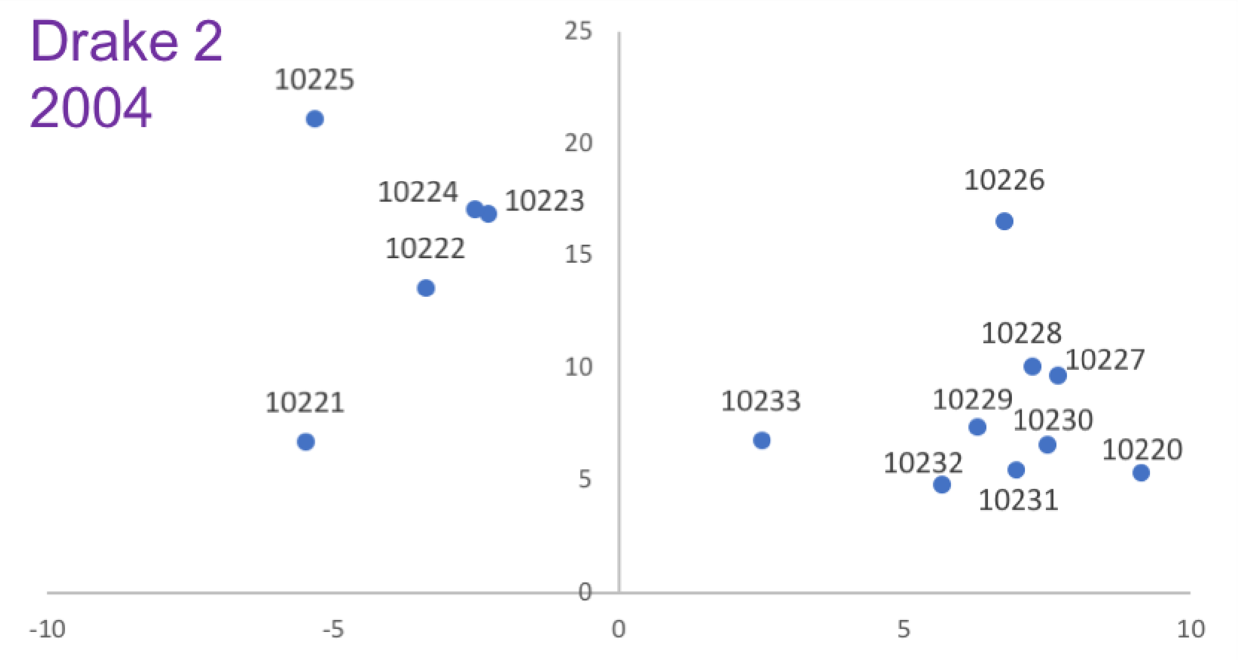
Chapter 2: The natural history of *HD1* and *HD2*, the heart of the Basidiomycetes HD mating locus, in collections of *Amanita phalloides*

All documents in this dedicated Chapter 2 Supplementary Materials Dropbox folder: https://www.dropbox.com/sh/7h367chlcl6fnut/AABMlyac3YYHkk-2VGt2N3FQa?dl=0

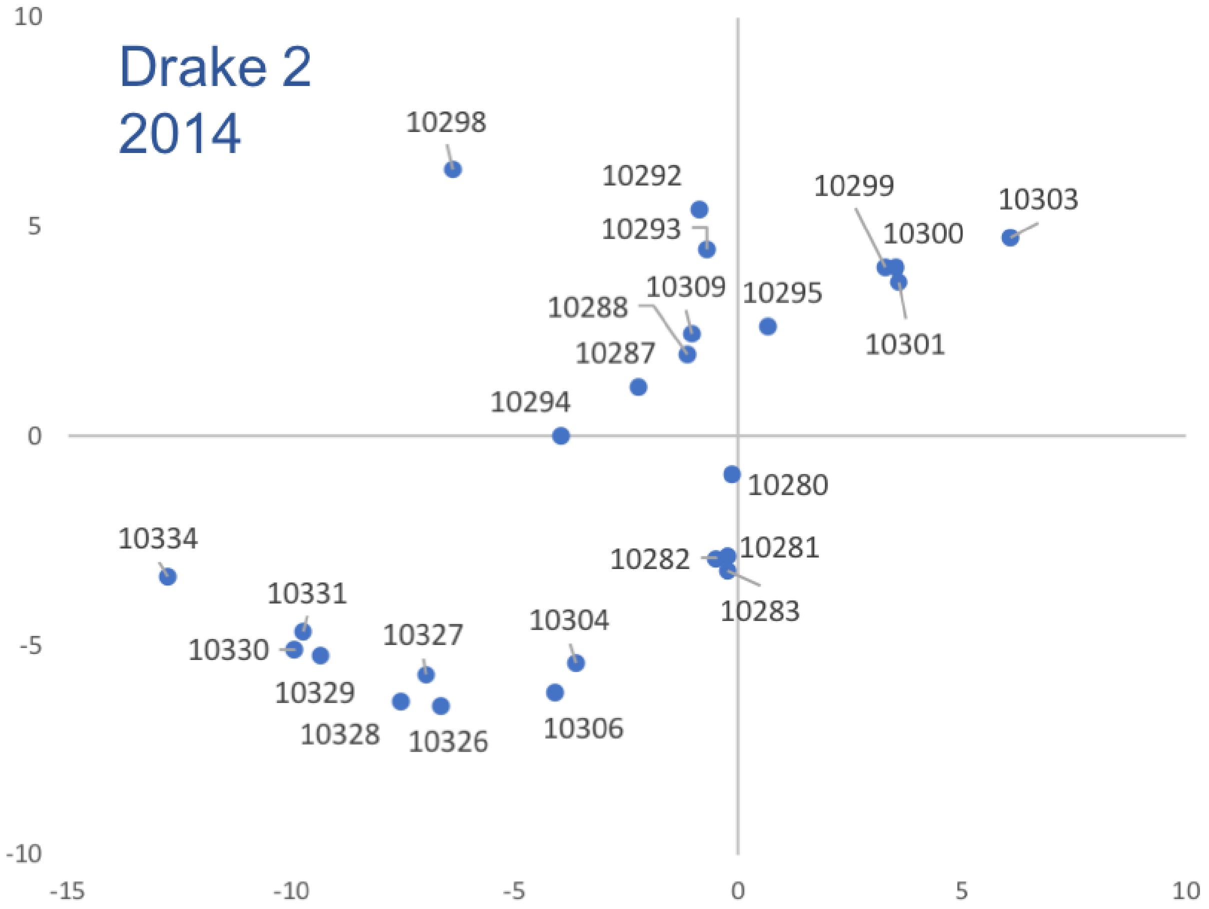
Scripts available in https://github.com/elmoremh/HD1-HD2-natural-history



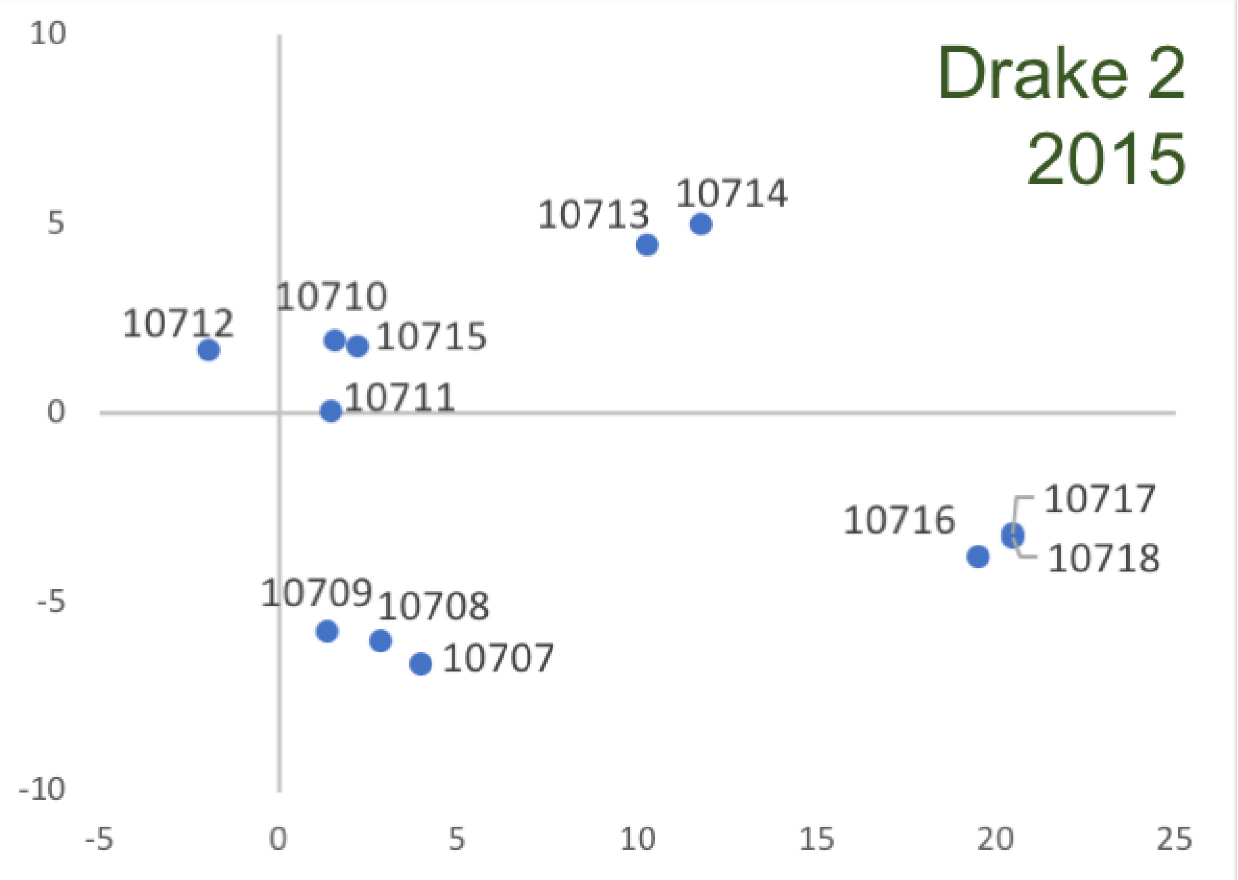
Supplementary Figure S2.1: Satellite view of the Drake 2 and Drake 3 sites, with centerpoints from 2004, 2014, and 2015 collections marked. The two sites are less than 100 meters apart and on either side of a two-lane road.



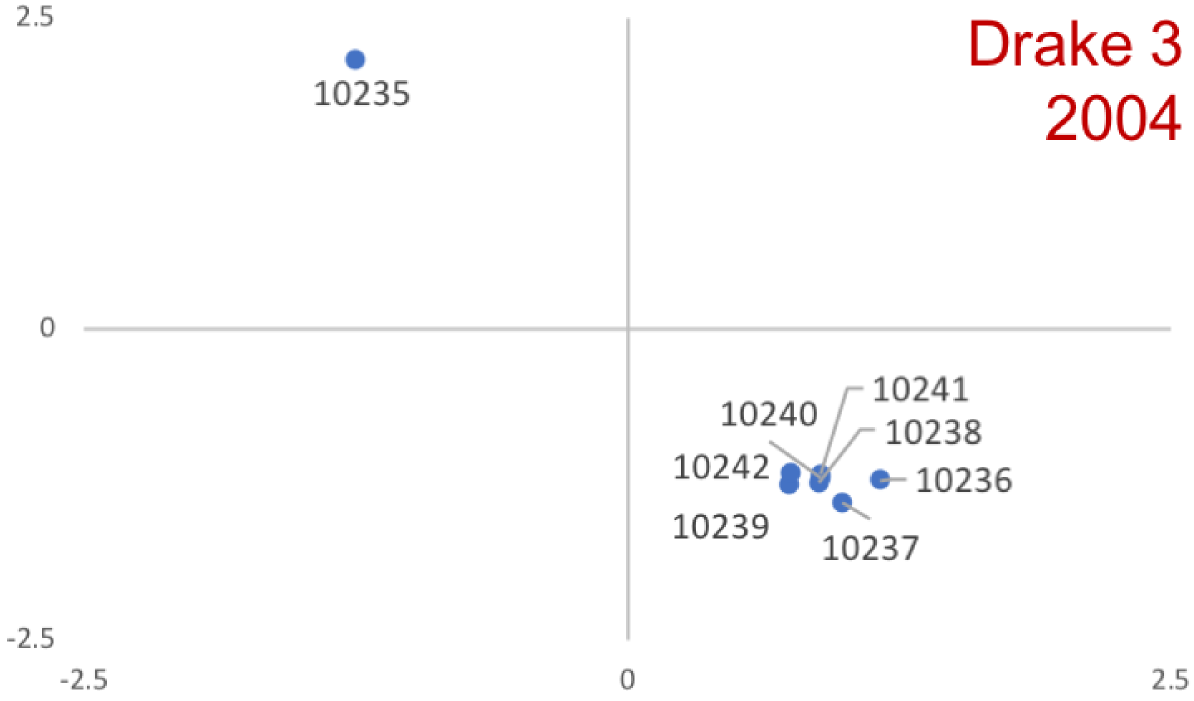
Supplementary Figure S2.2: A map of the *A. phalloides* population at Drake 2 in 2004. Each dot is a mushroom, labeled with its *AmanitaBASE* SpecimenID number. (0, 0) is the position of the field compass, and the location of each mushroom is given as x,y-coordinates relative to the field compass.



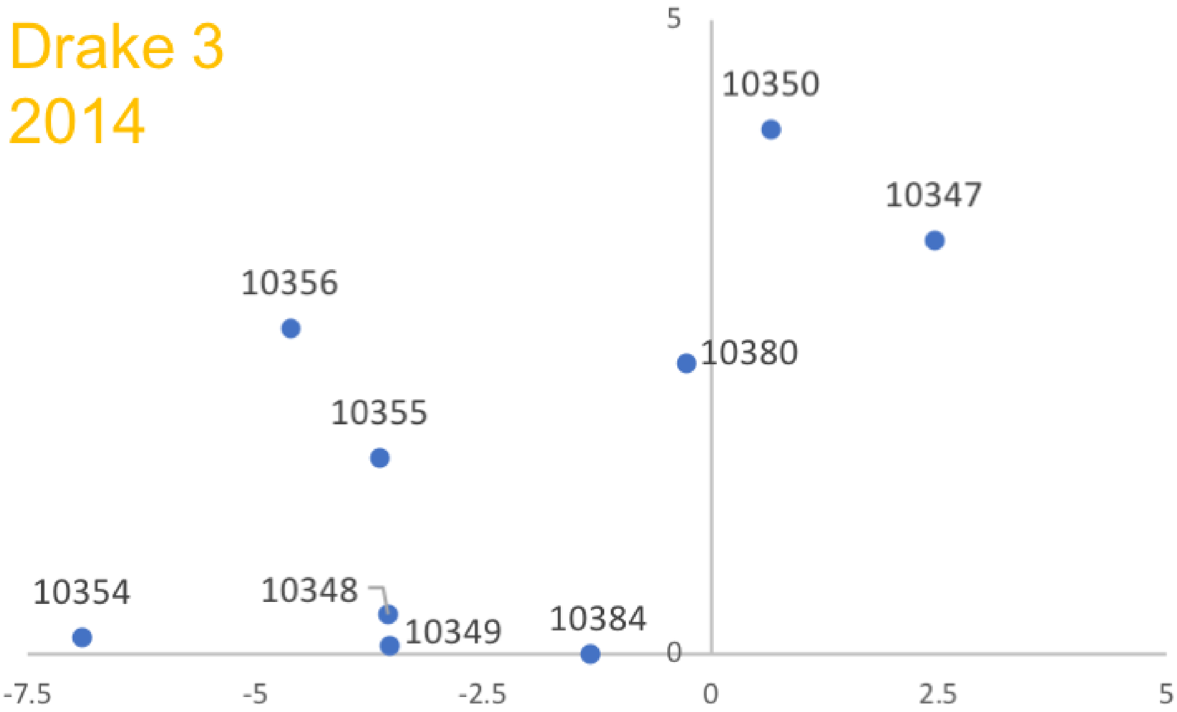
Supplementary Figure S2.3: A map of the *A. phalloides* population at Drake 2 in 2014. Each dot is a mushroom, labeled with its *AmanitaBASE* SpecimenID number. (0, 0) is the position of the field compass, and the location of each mushroom is given as x,y-coordinates relative to the field compass.



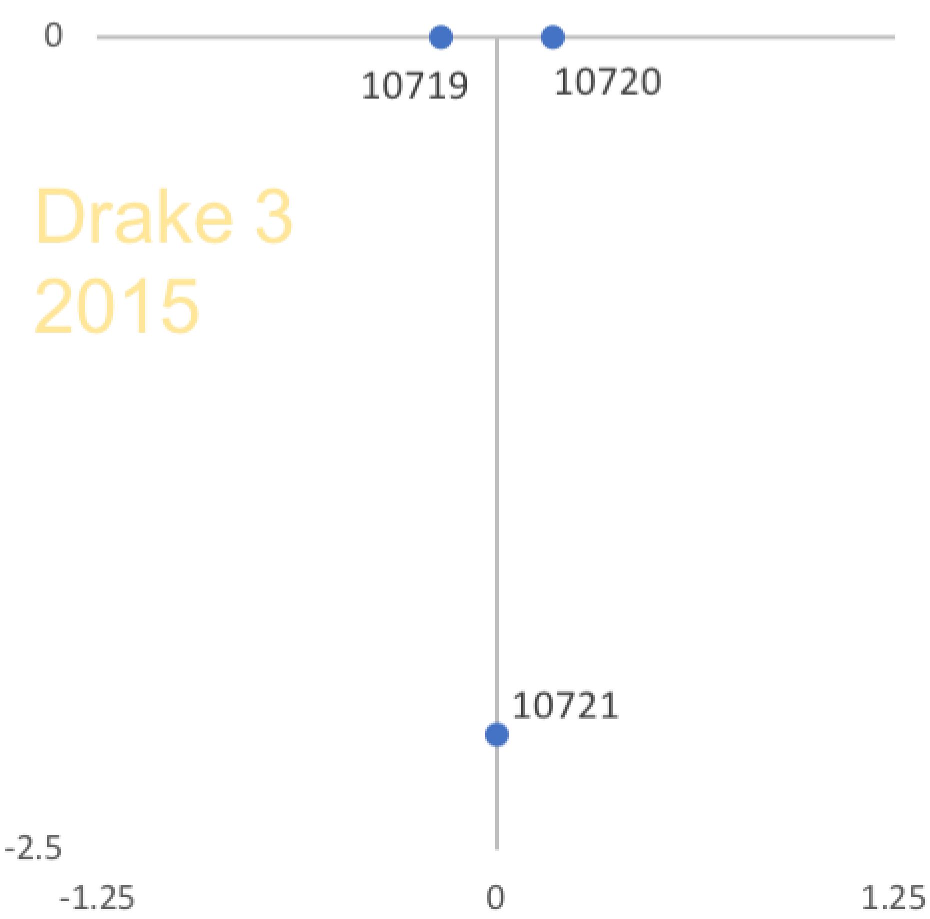
Supplementary Figure S2.4: A map of the *A. phalloides* population at Drake 2 in 2015. Each dot is a mushroom, labeled with its *AmanitaBASE* SpecimenID number. (0, 0) is the position of the field compass, and the location of each mushroom is given as x,y-coordinates relative to the field compass.



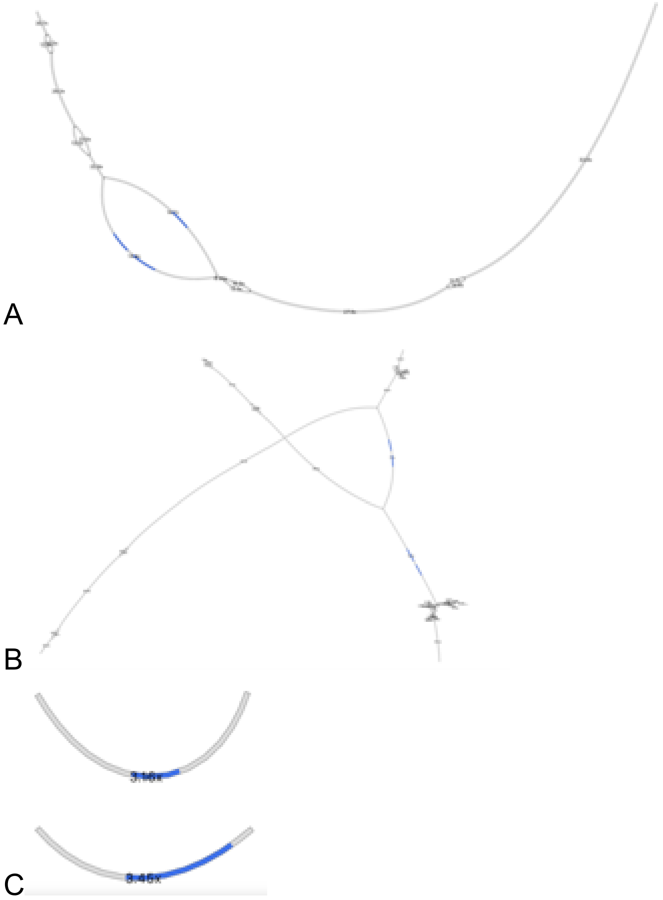
Supplementary Figure S2.5: A map of the *A. phalloides* population at Drake 3 in 2004. Each dot is a mushroom, labeled with its *AmanitaBASE* SpecimenID number. (0, 0) is the position of the field compass, and the location of each mushroom is given as x,y-coordinates relative to the field compass.



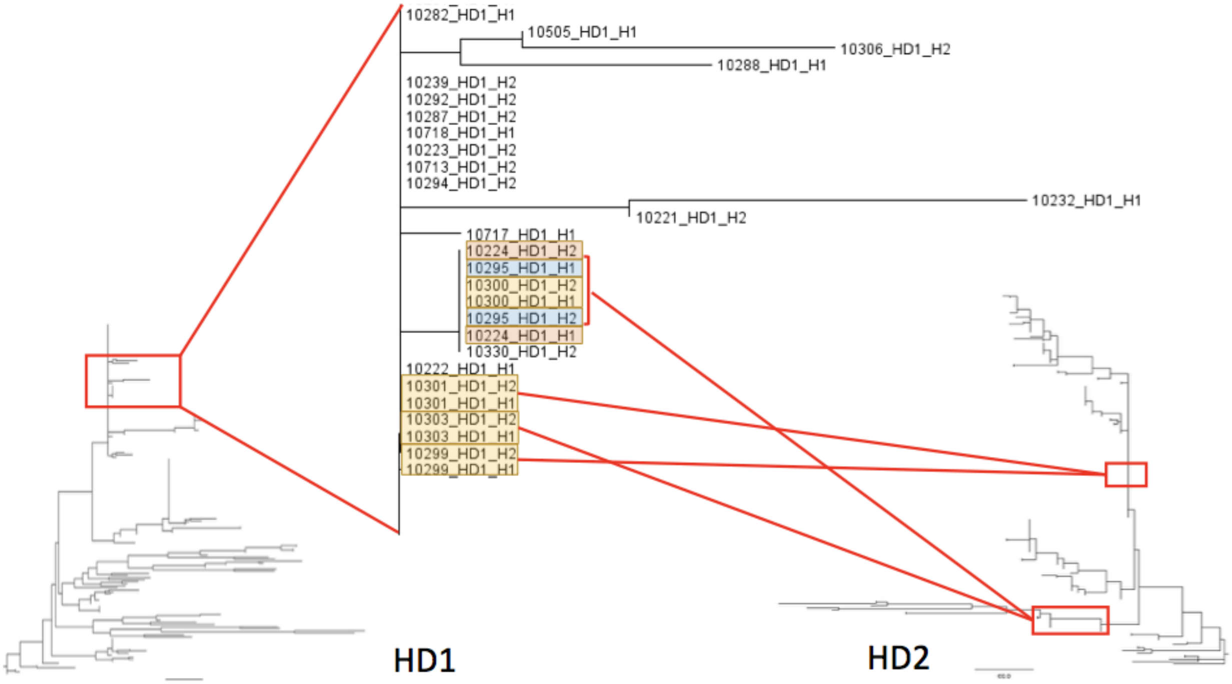
Supplementary Figure S2.6: A map of the *A. phalloides* population at Drake 3 in 2014. Each dot is a mushroom, labeled with its *AmanitaBASE* SpecimenID number. (0, 0) is the position of the field compass, and the location of each mushroom is given as x,y-coordinates relative to the field compass.



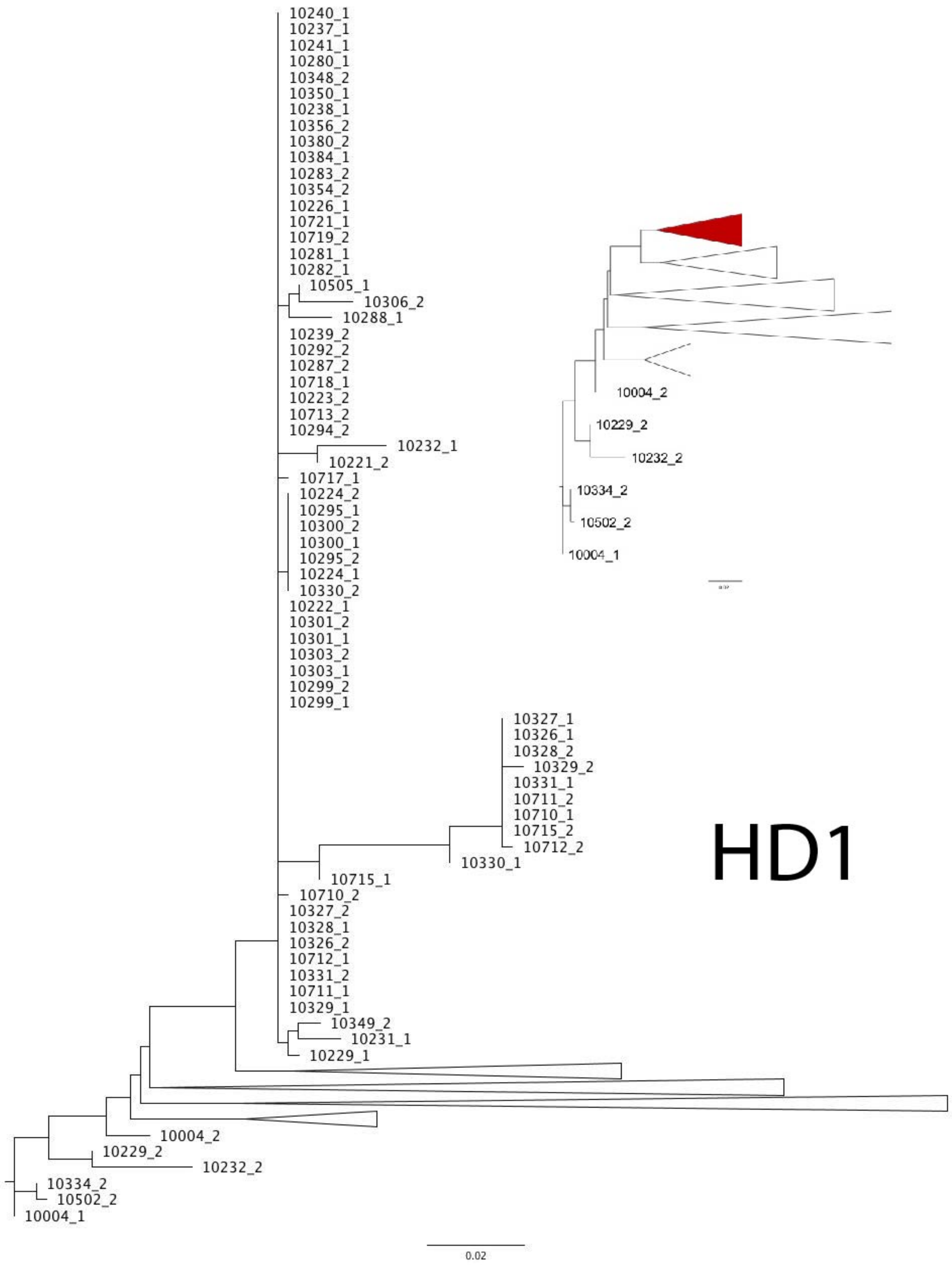
Supplementary Figure S2.7: A map of the *A. phalloides* population at Drake 3 in 2015. Each dot is a mushroom, labeled with its *AmanitaBASE* SpecimenID number. (0, 0) is the position of the field compass, and the location of each mushroom is given as x,y-coordinates relative to the field compass.



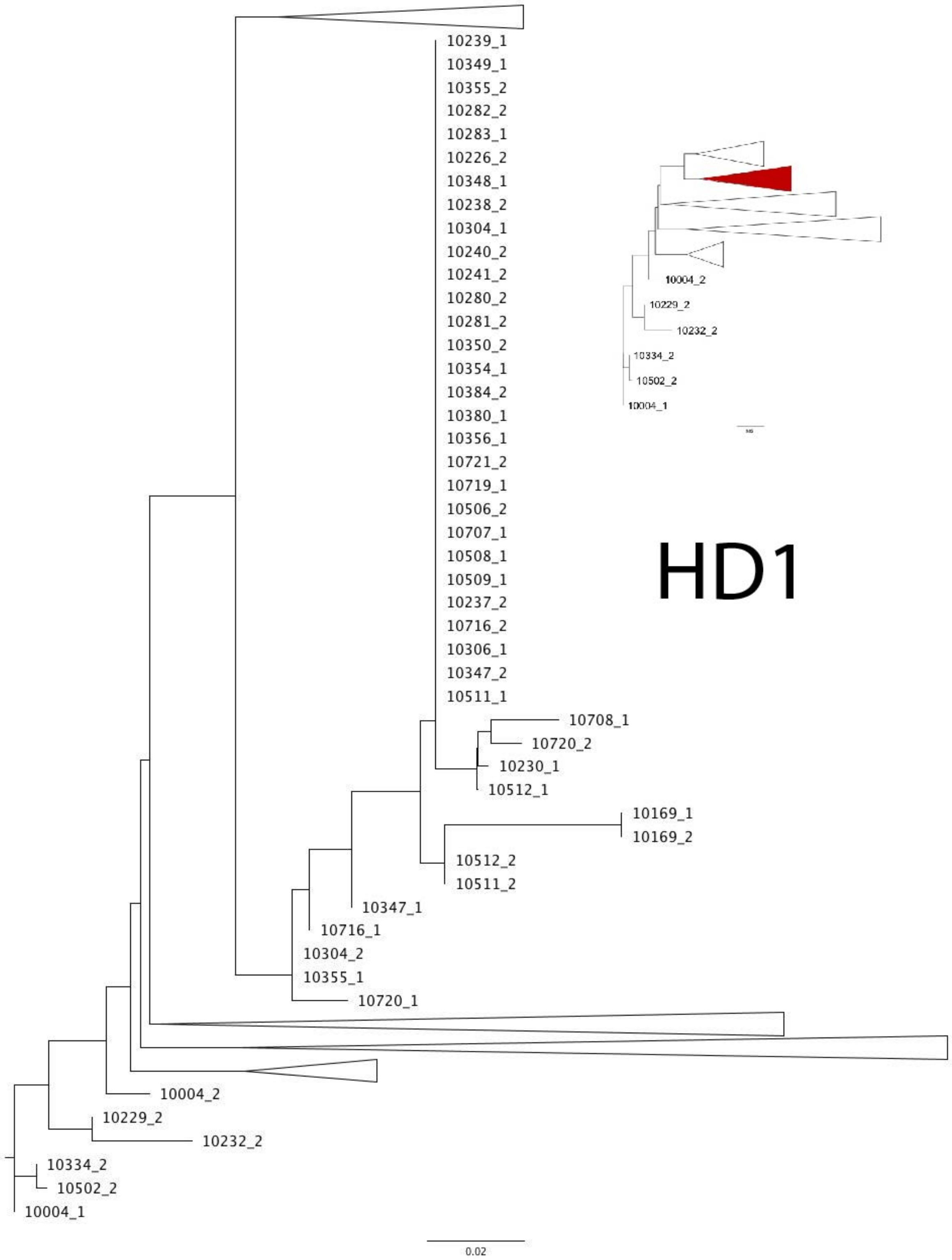
Supplementary Figure S2.8: Three types of de Bruijn graphs in the HD locus. Grey lines: unitigs; Black lines: links between unitigs; Blue lines: *HD1* and *HD2* hits; Numbers: depth of the unitigs. A. “bubble form”, two unitigs containing hits contiguous to same unitigs at both ends. B. “open bubble form”, two unitigs containing hits contiguous to same unitig(s) at only one end. C. “unlinked form”, two unitigs containing hits not contiguous to any same unitig at all.



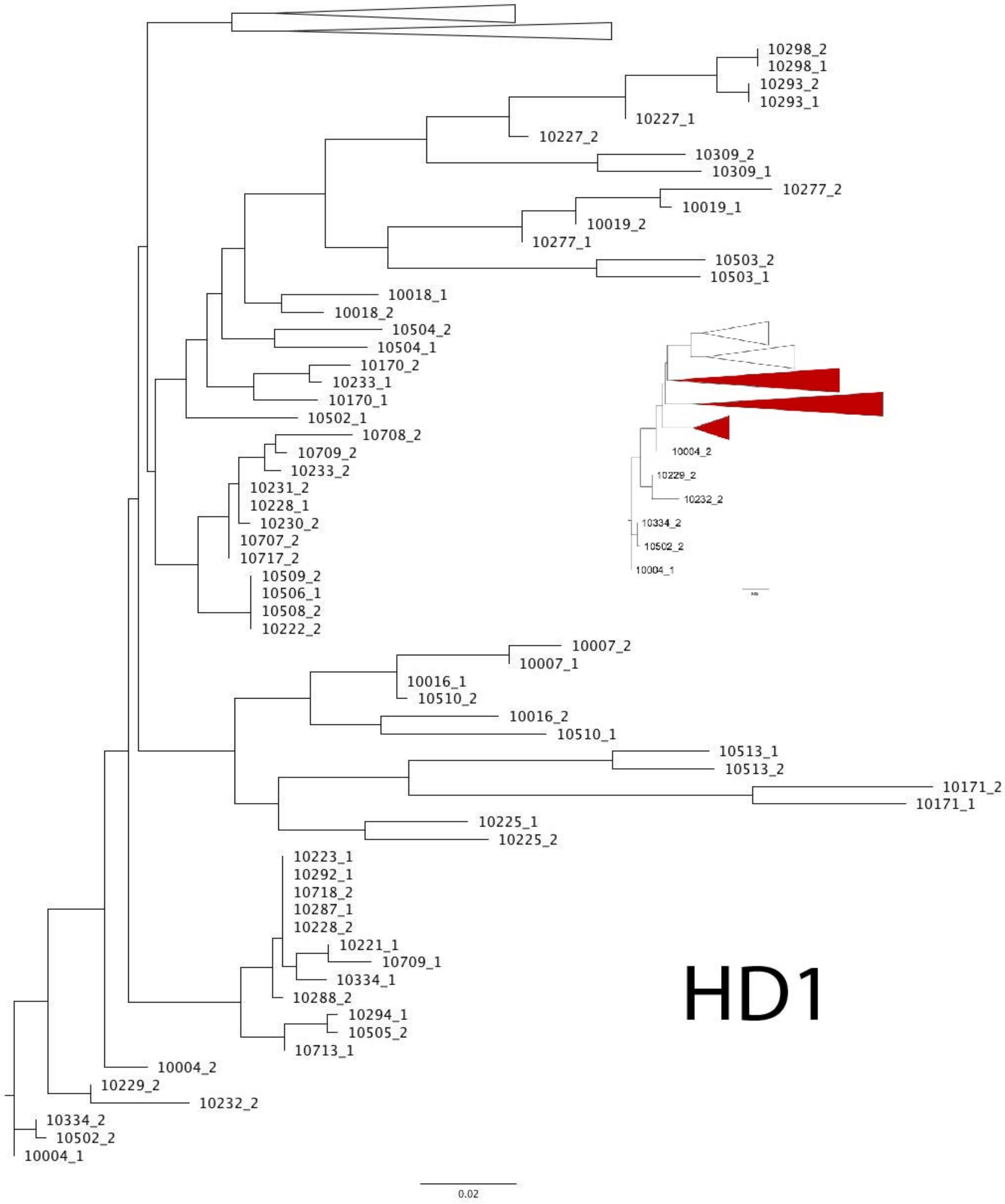
Supplementary Figure S2.9: Individuals that have “identical” (due to missing data) *HD1* alleles have, in some cases, quite divergent *HD2* alleles. Either there is convergent evolution on the same *HD2* alleles, or this calls into question the claims that there is no recombination between *HD1* and *HD2.*



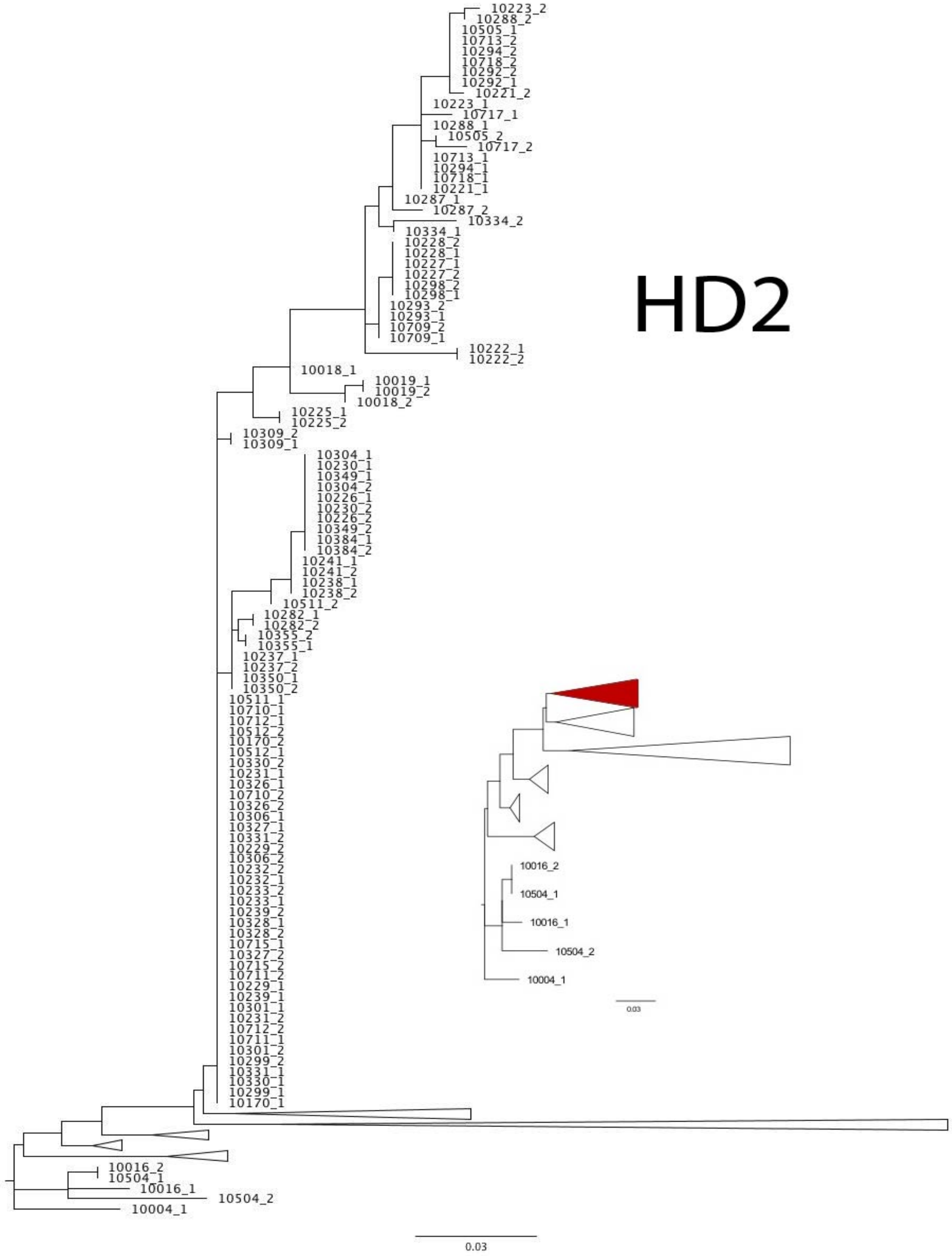
Supplementary Figure S2.10: Close-up *HD1* phylogeny. *Page 1/3.* A Maximum-Likelihood phylogeny of 172 *HD1* alleles, both alleles from each of 86 individual mushrooms, with nodes in increasing order. Here middle nodes are collapsed to allow greater visibility of the top of the tree. The open node is shown in red in the legend.



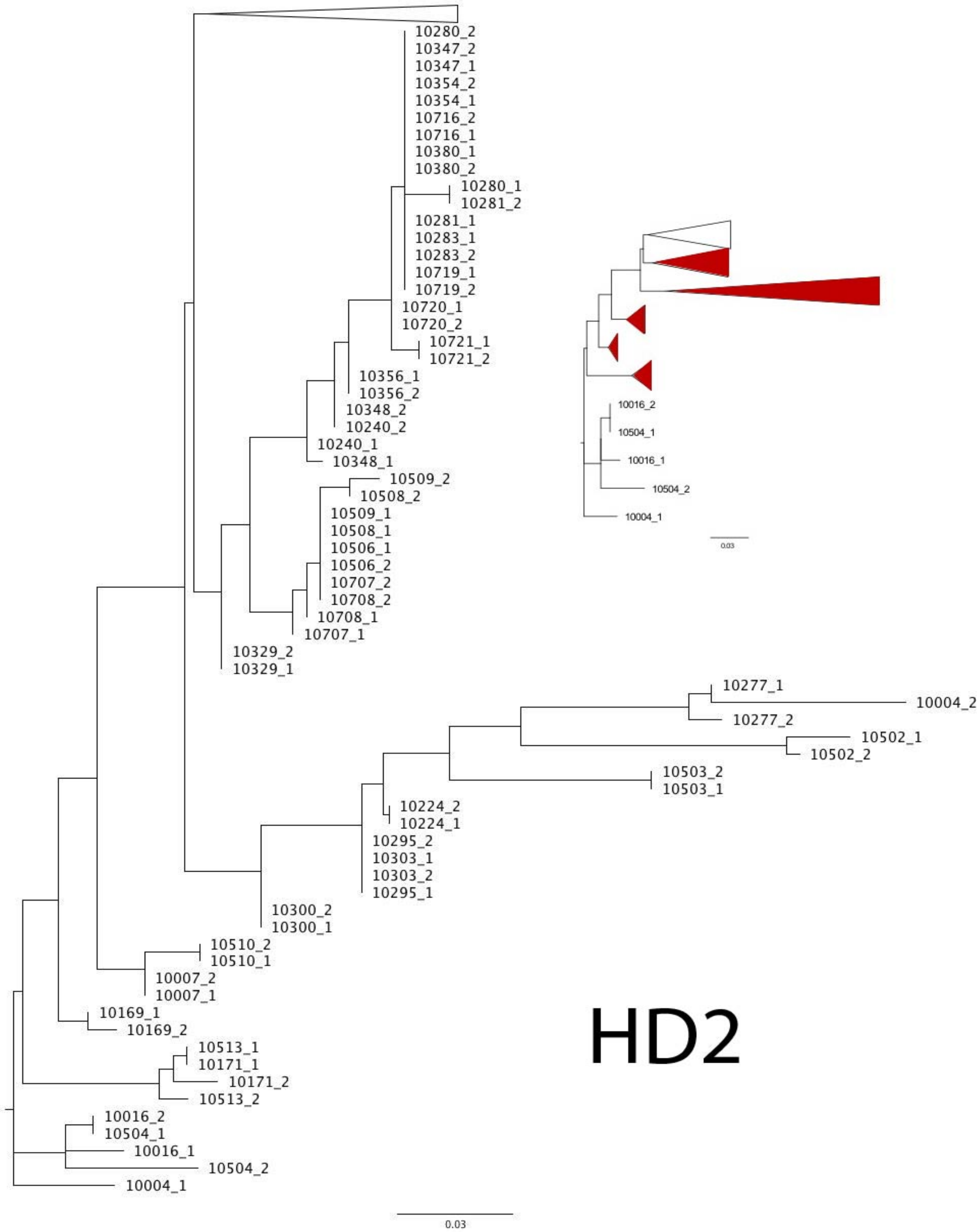
Supplementary Figure S2.10 (Continued): Close-up *HD1* phylogeny. *Page 2/3.* A Maximum-Likelihood phylogeny of 172 *HD1* alleles, both alleles from each of 86 individual mushrooms, with nodes in increasing order. The open node is shown in red in the legend.



Supplementary Figure S2.10 (Continued): Close-up *HD1* phylogeny. *Page 3/3.* A Maximum-Likelihood phylogeny of 172 *HD1* alleles, both alleles from each of 86 individual mushrooms, with nodes in increasing order. The open nodes are shown in red in the legend.



Supplementary Figure S2.11: Close-up *HD2* phylogeny. *Page 1/2.* A Maximum-Likelihood phylogeny of 172 *HD2* alleles, both alleles from each of 86 individual mushrooms, with nodes in increasing order. The open node is shown in red in the legend.



Supplementary Figure S2.11 (Continued): Close-up *HD2* phylogeny. *Page 2/2.* A Maximum-Likelihood phylogeny of 172 *HD2* alleles, both alleles from each of 86 individual mushrooms, with nodes in increasing order. The open nodes are shown in red in the legend.

Chapter 3: Contributions to the *Amanita* system

Scripts from section II are here: github.com/elmoremh/Amanita-Population-Genomics

Scripts from the rest of the chapter are found here: https://github.com/elmoremh/Contributions-to-Amanita

Each link below is to a file in my Chapter 3 Dropbox: https://www.dropbox.com/sh/pil9mbkbyzc8b69/AADXUcLKGaOmHmtHGN0kV-0ja?dl=0

Section III

A variant call file (vcf): https://www.dropbox.com/s/6zzn7ebw4eee2y6/HD1-HD2_vcf.xlsx?dl=0 of variants within the coordinates of *HD1* (Contig87:398638-399360) and *HD2* (Contig87:397803-398114), with *A. thiersii* specimens 10175, 10801, and 10802 and any reads mapped to those specimens highlighted in grey.

Amamu queries HD1, HD2, and MIP: https://www.dropbox.com/s/tnmzkuuji8qdrnj/A_locus_query_AAseqs?dl=0 for JGI BLAST of *Ath* Skay4041.

Amamu queries for synteny BLAST: https://www.dropbox.com/s/tnmzkuuji8qdrnj/A_locus_query_AAseqs?dl=0 and BLAST results: https://www.dropbox.com/s/xpgplxb7w1rr1du/A-locus_BLAST_results_synteny.xlsx?dl=0 organized by species and synteny.

A. thiersii HD1 check primer: https://www.dropbox.com/s/0fzmeybcc7htwjp/A_thiersii_HD1-HD2_primers.xlsx?dl=0 design and order form.

Section IV

SPAdes assembly scaffolds: https://www.dropbox.com/s/9oy3hrj1vhue2ou/scaffolds.fasta?dl=0

Scaffolds.stats: https://www.dropbox.com/s/mgn2ajjik2aqnxi/scaffolds.stats?dl=0 (statistics for the SPAdes assembly)

Kmer spectra and Excel plot-making sheet: https://www.dropbox.com/sh/kx1mcb27wdtb8m2/AACe3ncKyTrwz-ipX7UIae0Ka?dl=0

Full Aphal vs Aphal BLAST results table: https://www.dropbox.com/s/x5wkmv6nqqsdnpp/Aphal-MSDIN_v_Aphal-genome_2-5-15.xlsx?dl=0.

Appendix II

Contributions to the study of whole genome duplication in *Arabidopisis arenosa*

The following document was written in Spring 2015, as my then-PI Kirsten Bomblies was leaving Harvard, to help orient for the next person to work on that projects I had begun. It briefly covers my PhD work that wasn't about Amanita *or fungi. It references my lab notebook, which I gave to Dr. Bomblies before she left.*

Hello, heir to my project! I'm Holly Elmore, a former grad student in the Bomblies lab at Harvard. Here's how to contact me:

Cell: (561) 324 8787

Skype: m.holly.elmore (location listed as Wellington, FL [2019: Cambridge, MA])

Email: m.holly.elmore@gmail.com

Other people involved in my projects:

Kirsten Bomblies & Levi Yant

Kevin Wright: wright@fas.harvard.edu

Andrew Lloyd: andrewhmlloyd@gmail.com

Franchesco Molina: yhersonfranchesco@gmail.com

Jeremy O'Connell: theoconnell@gmail.com

Eli Swab: eli.swab@roxburylatin.org

I was only in the lab from June 2014 to February 2015, so I didn't get to do much.

I was interested in:

- The coevolution of the meiosis proteins, particularly ASY1 and ASY3. I wanted to investigate whether all of the polymorphisms associated with tetraploidy each contributed to polyploid meiosis or if only a few mutations helped in the adjustment to polyploidy and the rest occurred to accommodate them. This was why I raised tetraploid plants from populations that were known to harbor the diploid-like ASY1 and ASY3 alleles (described in pages 2-15 of my lab notebook).
- Why plants don't get cancer (in the sense of selfish cell lineages that harm the reproductive success of the organism in order to reproduce themselves).

Tolerance of aneuploidy and polyploidy is probably part of the equation. There are a lot of notes about this in my notebook, but I never got to do any work on it.

My major contributions to the lab were:

1. Screening tetraploid *A. arenosa* to look for diploid-like ASY1 and ASY3 alleles. I found a diploid-like ASY1 allele in STEmix84 (see notebook page 75).
    a. This group, the STEmix plants, were a mix of seeds from populations that were known to have the diploid-like alleles. The seeds were mixed before planting and so we don't know the parents of the individual plants.
        i. ASY1: TBG4, STE1, STE2, STE4, STE9, TBG1, TBG6
        ii. ASY3: STE3, Tz19, STE10, STE7
    b. Plate Holly 5 (key on page 75) has extracted DNA from these plants.
    c. Diploid-like and tetraploid-like ASY1 alleles can be distinguished by RFLP analysis with XmnI (see notebook page 65).
    d. ASY3 screening requires DCAPS primers, which need to be re-designed. I did some screens that were inconclusive because the difference between cut and uncut only 20 bp.

2. Caring for and extracting from DNA from a lot of diploids and colchicine-treated diploids. I did this mainly to help out Kevin and learn the process, but I was also hoping to eventually use any "colchi-ploids" that resulted in studying tolerance of polyploidy.
    a. Extracted DNA plates:
        i. Holly 1 (page 23)
        ii. Holly 2 (page 29)
        iii. Holly 3 (page 31)
        iv. Holly 4 was one row (from A to H): Kz32, Kz33, Kx59, Kz60, Kz61, Kz62, Kz63, Ø

3. Beginning the process of mutating (site-directed mutagenesis) the sequence of the *A. thaliana* ASY1 (diploid-like) allele at each of the four major selected SNPs to see how that would affect *A. thaliana* meiosis.
    a. See Kirsten's ASY1-SDM Dropbox folder for Kirsten's documents for the sequences of the different constructs and the placement of four key substitutions: K40E, F272S, R313H, and Q567R
    b. See my construct and primer spreadsheet: https://www.dropbox.com/s/arjuljbv0kb033v/Holly_legacy_ASY1-SDM.xlsx?dl=0
    c. See my Holly Cloning Box key (page 105) and cross-reference with spreadsheet
    d. Why so many constructs? We began by attempting to perform site-directed mutagenesis (NEB kit) on the entire 16 kb ASY1 construct, but got nothing. We thought it might be too big, so I tried to make two subclones, each with 2 of the substitution sites.

i. Subclone 1 would contain sites 40 and 272. I thought I had made this one successfully but it turned out just to look that way on the gel (see notebook page 84 for explanation).

- Protocol: Double digest 16 kb ASY1 construct with BamHI-HF and SacII. Take 3.9 kb fragment and ligate into pBlueScript vector. Save 12.1 kb fragment so that 3.9 kb fragment can go back in after site-directed mutagenesis.

ii. Subclone 2 would contain sites 313 and 567. I was going to make it by Gibson assembly (follow NEB kit instructions and see primer tab of construct and primer spreadsheet).

e. You'll also see a lot of notes about raising a bunch of asy1/+ *A. thaliana*. They were intended for transformation down the line, but, of course, I never got there. I got the seeds from Andrew.

All documents associated with the project are here, in the ASY1-SDM project legacy Dropbox folder.