Genomic Analysis of the Population Structure and Virulence of the Human Pathogen, Cryptococcus Neoformans Var. Grubii

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:33813402">http://nrs.harvard.edu/urn-3:HUL.InstRepos:33813402</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>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 <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Genomic analysis of the population structure and virulence of the human pathogen, 

_Cryptococcus neoformans_ var. _grubii_

Sean M. Sykes

A Thesis in the Field of Biotechnology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

March 2017
Abstract

This study examined a large cohort of clinical and environmental strains of *Cryptococcus neoformans* var. *grubii* to better understand its population structure and identify potential markers significantly associated with clinical isolates or increased virulence capacity. *C. neoformans* var. *grubii* is an opportunistic fungal pathogen infecting nearly 1 million patients a year, mostly immunocompromised individuals. Previous work using multi-locus sequence typing analysis identified three sublineages: VNI, VNII, and VNB. Unlike the other two sublineages, the VNB sublineage is constrained to sub-Saharan Africa, exhibits high genetic variability, and encompasses a mix of environmental and clinical isolates. Nearly 400 diverse isolates were selected for sequencing and SNP genotype data were generated and provided as input to the STRUCTURE and smartPCA software to further clarify the population structure. This population analysis revealed a split in the VNB sublineage along with the presence of hybrid and introgressed strains. These genotype data and population information were used with the GEMMA software to identify potential markers significantly associated with clinical strains. GWAS identified twelve markers significantly associated with clinical strains, several within known virulence genes. Additional GWAS analysis utilized phenotypic characterization data for oxidative stress, fluconazole drug resistance, and melanization along with the genotype data. This revealed that three clinical strains harbored significantly associated mutations in *BZP4* and exhibited decreased melanin production similar to the melanin-deficient *Lac1* deletion strain. These data highlight the
utility of GWAS in the analysis of complex traits of pathogenic fungi and provide new markers for further virulence studies.
Dedication

To my family, Kristie and Clara
Acknowledgements

Foremost, I would like to thank my thesis director, Dr. Christina Cuomo. Her willingness to advise me during this process truly means so much to me as her scientific knowledge, guidance, and support has been invaluable to me throughout my time at the Broad Institute and during this thesis process. I would also like to thank my colleague on this project, Dr. Christopher Desjardins, for his patience and eagerness to share his knowledge of comparative genomics with me.

More personally, I’d like to thank my family for their patience during this whole process. My wife, Kristie, has been tireless in her support and willingness to step in when I could not, and my daughter, Clara, has been the best homework buddy anyone could ask for. I hope that my pursuit of this degree inspires her to work hard to pursue her goals.

Finally, this thesis work supported a project that was funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Grant Number U19AI110818 to the Broad Institute.
Table of Contents

Dedication ........................................................................................................................................ v
Acknowledgements ............................................................................................................................... vi
List of Tables ......................................................................................................................................... ix
List of Figures ................................................................................................................................. x
I. Introduction ........................................................................................................................................ 1
  Cryptococcus neoformans Species ........................................................................................................ 1
  Cryptococcus neoformans var. grubii Population Structure ............................................................... 3
  C. neoformans var. grubii Incidence ...................................................................................................... 4
  C. neoformans var. grubii Virulence ..................................................................................................... 6
  Genome Wide Association Studies ....................................................................................................... 8
  Discovering Markers Associated with C. neoformans var. grubii .................................................. 10
II. Materials and Methods ................................................................................................................ 12
  Variant Detection ............................................................................................................................ 12
  Investigation of Population Structure ............................................................................................... 13
  Identification of Hybrid Strains ........................................................................................................ 15
  Analysis of Genomic Introgressions ............................................................................................... 16
  Classification of Clinically Associated Markers .............................................................................. 17
III. Results ........................................................................................................................................... 19
  Investigation of Population Structure ............................................................................................... 19
  Identification of Hybrid Strains ........................................................................................................ 22
List of Tables

Table 1. *Cryptococcus neoformans* var. *grubii* Strain Collection .................................................. 13
Table 2. Fraction ancestry of Hybrid Strains ........................................................................... 24
Table 3. Clinical versus environmental markers ....................................................................... 31
Table 4. Top ten L-DOPA markers .......................................................................................... 31
Table 5. Melanization Scores for Loss of Function *BZP4* Markers ......................................... 32
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>STRUCTURE and smartPCA Subpopulation Results (387 strains)</td>
<td>21</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Evanno $\Delta K$ Results</td>
<td>21</td>
</tr>
<tr>
<td>Figure 3</td>
<td>STRUCTURE and smartPCA Results of VNI Strains</td>
<td>22</td>
</tr>
<tr>
<td>Figure 4</td>
<td>STRUCTURE and smartPCA Results of 392 Strains Showing Hybrid Strains</td>
<td>24</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Introgression Counts Per Strain</td>
<td>26</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Clinical vs. Environmental Strain GWAS Q-Q Plot</td>
<td>27</td>
</tr>
<tr>
<td>Figure 7</td>
<td>L-DOPA GWAS Q-Q Plot</td>
<td>30</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Phylogenetic Tree of <em>Cryptococcus neoformans</em> var. <em>grubii</em></td>
<td>38</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Top three principal components from smartPCA</td>
<td>39</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Chromosome Ancestry of Hybrid Strains</td>
<td>40</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Genomic Introgressions Plot</td>
<td>41</td>
</tr>
</tbody>
</table>
Chapter I

Introduction

_Cryptococcus neoformans_ var. _grubii_, an opportunistic fungal pathogen, causes an estimated 1 million new cases of cryptococcosis per year with a 60% mortality rate, particularly in immunocompromised individuals (Park et al., 2009). Though _C. neoformans_ var. _grubii_ is found in environmental reservoirs, it is suited for pathogenicity due to the presence of a polysaccharide capsule, an ability to grow at human body temperature, and the capacity to produce melanin (Liu et al., 2008; Kronstad, Jung, & Hu, 2008). Despite the identification of numerous genes related to virulence, there is not a clear understanding on how genotypes affect virulence phenotypes. The goal of this study was to identify potential markers associated with _C. neoformans_ var. _grubii_ phenotype.

_Cryptococcus neoformans_ Species

The basidiomycete, _Cryptococcus neoformans_, is a saprobic fungus in the Order Tremalleles (Findley et al., 2009) that is a major cause of opportunistic infection. Many fungal species within this order thrive in environmental reservoirs such as soil, grass, and wood (Findley et al., 2009). _Cryptococcus neoformans_ occupies an interesting subpopulation among Tremalleles species because of its ability to inhabit environmental niches and infect animal hosts. In the environment, _C. neoformans_ inhabits rotting wood
(Chayakulkeeree & Perfect, 2006; Hiremath et al., 2008) and pigeon guano (Litvintseva et al., 2005). In human hosts, there is a large impact of systemic infections by \textit{C. neoformans} causing meningitis and death (Park et al., 2009), despite antifungal treatments. Hence, \textit{C. neoformans} remains a worldwide health burden.

Previous research identified four serotypes of \textit{C. neoformans} (A, B, C, and D) (Bennett et al., 1977). Serotypes B and C were rare in the environmental and were infrequent causes of human infection (Bennett et al., 1977); the organisms with these serotypes were different enough from the organisms in the other serotypes to warrant their classification as a new species, \textit{Cryptococcus gattii} (Kwon-Chung & Varma, 2006). Two variety designations are used to describe serotype A and D strains: \textit{C. neoformans} var. \textit{grubii} constitutes serotype A strains while \textit{C. neoformans} var. \textit{neoformans} constitutes serotype D strains (Bennett et al., 1977). Recent phylogenetic analysis suggests further delineation of these two serotypes into separate species (Hagen et al., 2015). Of these species, \textit{C. neoformans} var. \textit{grubii} (serotype A) is crucial for further study owing to the fact that it causes over 90% of clinical cases worldwide, particularly in sub-Saharan Africa among HIV+ individuals (Chayakulkeeree & Perfect, 2006). This study focused exclusively on the \textit{C. neoformans} var. \textit{grubii} population and featured one of the largest strain sets of \textit{Cryptococcus} investigated up to this point. To examine population structure and differences between properties including clinical or environmental origin, we examined the whole genome sequence from a large cohort of strains isolated from both environmental and clinical sources.
Cryptococcus neoformans var. grubii Population Structure

*C. neoformans* var. *grubii* is a haploid organism with a genome size of 19Mb organized into 14 chromosomes (Janbon et al., 2014). This species undergoes both an asexual and sexual life cycle (Chayakulkeeree & Perfect, 2006). Unique among basidiomycetes, *C. neoformans* var. *grubii* contains a bipolar mating type (*MAT*) system (Nielsen & Heitman, 2007). Each strain contains one of two mating types, *MAT*a or *MAT*a, and, during conjugation, can facilitate recombination between strains (Nielsen & Heitman, 2007). However, there is a strong bias towards a single mating type, *MAT*a, in both environmental and clinical strains (Nielsen & Heitman, 2007). Though limited availability of both mating types in close proximity could reduce the potential for sexual recombination, studies have shown that strains exhibiting the same mating type locus mate and recombine (Sun et al., 2014).

Phylogenetic analysis of *C. neoformans* var. *grubii* strains using multi-locus sequence typing (MLST) revealed 3 subpopulations within the larger population: VNI, VNII, and VNB (Litvintseva et al., 2006). An analysis of 102 isolates revealed that the VNI and VNII lineages were distributed worldwide while the VNB lineage was confined to southern Africa (Litvintseva et al., 2006). Further, the VNI and VNII strains were found to be highly clonal as evidenced by linkage disequilibrium within subclades and lack of genotypic variability across strains isolated from disparate worldwide sites, and, notably, these strains appear to contain only one of the two mating types, *MAT*a (Litvintseva et al., 2006). Conversely, VNB strains were mostly confined to southern
Africa and showed greater genotypic than VNI and VNII strains (Litvintseva et al., 2006). Further, researchers identified VNB strains from southern Africa containing the rare \textit{MATa} mating type (Litvintseva et al., 2003). This finding coupled with the higher diversity exhibited in VNB strains suggests an African origin for \textit{C. neoformans} var. \textit{grubii} (Litvintseva et al., 2011). Though the population structure of \textit{C. neoformans} var. \textit{grubii} has been described previously, the analysis focused on a limited set of typically seven marker genes with limited resolution. Through the examination of the full variation within the population using whole genome sequence, this thesis project further clarified the population structure of \textit{C. neoformans} var. \textit{grubii}, by identifying stratifications within the population and aiding in our understanding \textit{C. neoformans} var. \textit{grubii}.

\textit{C. neoformans} var. \textit{grubii} Incidence

\textit{C. neoformans} var. \textit{grubii} is typically found in two environmental reservoirs: decayed wood and avian excrement (Chayakulkeeree & Perfect, 2006). It is believed that pigeons may harbor the organism without infection due to their high body temperature; however, pigeons are essential for the dispersal of spores (Litvintseva et al., 2011). Because of the intermingling of pigeons with human habitats, the opportunity exists for airborne spores of \textit{C. neoformans} var. \textit{grubii} to enter the human body through the lungs (Chayakulkeeree & Perfect, 2006). In individuals without suppressed immunity, the inhalation of spores would not typically cause disease (Chayakulkeeree & Perfect, 2006). Importantly, in the immunocompromised, cryptococcal infection is of great concern. If
cryptococcosis occurs, it may be confined to the lungs or become systemic, ultimately causing cryptococcal meningitis (CM) and death (Chuck & Sande, 1989; Chayakulkeeree & Perfect, 2006). Therefore, understanding how an organism that normally inhabits environmental niches is also adapted to infect humans is of utmost importance. Some studies have suggested that interactions with other eukaryotes in the environment such as nematodes and amoebae may select for traits that also enable pathogenesis in humans (Derengowski et al., 2013).

In areas with high incidence of HIV/AIDS, such as sub-Saharan Africa, cryptococcosis mortality rates for the 1 million people newly infected per year are near 60% (Park et al., 2009). In fact, the number of deaths in southern Africa due to Cryptococcus exceeds those from tuberculosis (Park et al., 2009). Therefore, controlling and understanding the mechanisms cryptococcal infection is a high priority. In the U.S., during the height of the AIDS epidemic, the median survival for patients with AIDS was estimated to be about 6 months (Chuck & Sande, 1989). Fortunately, the mortality rate appears to declining for those AIDS patients with access to good healthcare and anti-retroviral therapies (Mirza et al., 2003) as well as the highest standard of antifungal treatment. With the increase in immunosuppressive therapies associated with organ transplantation or other treatments, cryptococcosis remains a concern in the U.S. For example, one alarming study found an increased mortality due to C. neoformans var. grubii in HIV negative individuals and no recent history organ transplantation (Brizendine, Baddley, & Pappas, 2013), highlighting the continued need to understand C. neoformans var. grubii virulence. Working with our clinical collaborator, Dr. John Perfect, we included clinical and phenotypic information in this study that helped identify
markers that track with clinical isolates and may add to our knowledge of possible virulence factors.

*C. neoformans* var. *grubii* Virulence

*C. neoformans* var. *grubii* virulence is greatly enhanced by its ability to inhibit phagocytosis via a polysaccharide coat or capsule, produce melanin to reduce environmental stresses, and grow at human body temperature (Liu et al., 2008; Kronstad, Jung, & Hu, 2008). Due to its role in confounding host defenses, the main focus of virulence studies has been the polysaccharide capsule. A total of 35 *Cryptococcus* genes may be involved in the synthesis of this capsule (O’Meara & Alspaugh, 2012). Capsule formation begins within the cell through carbohydrate metabolism turning sugar into carbohydrate backbones (O’Meara & Alspaugh, 2012). One important component of the capsule, hyaluronic acid, is thought to allow *C. neoformans* var. *grubii* to enter the cerebrospinal fluid, causing meningitis (O’Meara & Alspaugh, 2012). Once assembled, the capsule is excreted and attached to the cell wall, requiring an estimated additional 40 genes for this process alone (O’Meara & Alspaugh, 2012). Owing to the role of the capsule in virulence and the fact that many genes have been identified as crucial to its synthesis and assembly, these genes are an important first line of study for virulence markers.

Because the polysaccharide capsule is crucial for virulence and as sugars are an important precursor for capsule formation, *Cryptococcus* must contain some mechanism for sugar intake. Recently, researchers began to uncover a role for a specific sugar
alcohol, inositol, in \textit{C. neoformans} var. \textit{grubii} virulence. For example, mutant \textit{C. neoformans} var. \textit{grubii} strains deficient in two inositol transporter genes (\textit{itr1a} \textit{itr3c}) exhibited a reduced ability to cross the blood-brain barriers in \textit{in vitro} cell and \textit{in vivo} animal models (Liu et al., 2013). Further, a microarray study of the response of wild type \textit{Cryptococcus} cells to inositol treatment revealed overexpression of genes for breaking down inositol (Liu et al., 2013). Transfer of the \textit{itr1a} \textit{itr3c} mutants into mouse models showed an increased immune response compared to infection with wild type \textit{Cryptococcus} (Liu et al., 2014). Moreover, inositol is important for capsule formation, since researchers observed a decrease in the polysaccharide capsule production due to a lack of the capsule building block, glucuronoxylomannan (Liu et al., 2014). These studies present further evidence for the role of the capsule in \textit{Cryptococcus} virulence and the importance of inositol during infection.

Despite its importance for \textit{Cryptococcus} virulence, the capsule is not the only factor conferring virulence. Screening of 1,201 targeted gene knockouts identified an additional 40 potential virulence factors in a mouse model, including replication factors, chromatin regulators, and immune response modulators (Liu et al., 2008). In addition, the ability to produce melanin is important for virulence. Melanin is thought to help protect the cells from stresses both in the host and in the environment (Nosanchuk and Casdevall 2003). Screening this knockout collection identified 38 mutants required for full melanin production; these knockouts include genes known to be required for melanin production. Both of these screens identified novel genes not previously linked to melanin production or virulence.
Although these knockout studies can test for the requirement of specific genes in *Cryptococcus* virulence, natural variation can modulate gene sequence at a finer scale, modulating function at more possible levels and with different impacts on virulence. For instance, MLST data identified environmental and clinical strains that appear to be similar genotypically, but, surprisingly, when these strains were inoculated into mice, the environmental strains exhibited a lack of infectivity (Litvintseva et al., 2009). This result suggests that finer scale variation is involved, and studies with limited marker sets lack the resolution necessary to detect such effects. It is also still unclear how variations in the gene sequences differ between clinical and environmental strains. Here, we utilized a more comprehensive marker set, and identified variations specific to clinical strains.

**Genome Wide Association Studies**

Because of the ease and falling cost of generating large genomic datasets, genome wide association studies (GWAS) have gained statistical power in identifying putative variations that are implicated in complex disease traits. The foundation of this method is the “common disease/common variant” (CD/CV) hypothesis wherein the genetic basis for common phenotypes is driven by common single nucleotide polymorphisms (SNPs), or nucleotide base changes in the genomic sequence, within a population of individuals (Bush & Moore, 2012). One principle behind GWAS is that the frequency of a SNP affecting uncommon phenotypes is low, and, for many phenotypes where a given SNP has a small effect, more than one SNP variant may be involved (Bush & Moore, 2012). Therefore, in GWAS, it is important to have high-quality SNP genotyping data across the
whole genome from a large cohort of the population representing both the presence (“case”) and absence (“control”) of a desired phenotype (Bush & Moore, 2012).

Typically, GWAS has been applied to human genomic data and common diseases, and as of 2014, the combined results of many studies identified roughly 12,000 SNPs with association to a disease (Welter et al., 2014).

GWAS is not limited to human genetics or looking for disease markers and has been applied to a few microbes. This method is applicable to any population with enough individuals exhibiting a heritable phenotype. Despite its previous success in human genetics and its translatability to other research areas, GWAS has rarely been applied to fungal genomics. The first published fungal GWAS involved clinical and non-clinical strains of the model organism, *Saccharomyces cerevisiae*. Researchers used a more general SNP genotyping array approach to examine a small number (n=88) of individual strains split equally between pathogenic and non-pathogenic phenotypes looking for markers distinguishing clinical from non-clinical isolates. Using this method, the researchers identified several loci implicated in pathogenicity (Muller et al., 2011); thus, demonstrating the utility of GWAS in fungal population analysis. Another GWAS examined virulence in the conifer pathogen, *Heterobasidion annosum s.s.*. Here, researchers used an even smaller set (n=23) of individual strains along with ~30,000 SNP genotype calls to find four variants associated with virulence (Dalman et al., 2013). In one of the larger fungal GWAS studies to date, researchers examined a set of wild type *Neurospora crassa* strains (n=112) to understand how asexual spores communicate and, ultimately, fuse. Here, ~1M SNP genotypes were generated and GWAS identified a single gene, *cse-1*, with significant association to increased communication (Palma-
Guerrero et al., 2013). This initial result along with analysis of deletion strains allowed the researchers to newly identify several cse-1 associated genes necessary for spore communication (Palma-Guerrero et al., 2013). These studies suggest that GWAS can be used to find markers associated with pathogenicity, virulence, or other complex traits in fungi. Contrary to these previous studies, we obtained a more extensive strain collection and a larger SNP genotype data set and had more power to identify potential virulence markers.

Discovering Markers Associated with *C. neoformans* var. *grubii*

Though previous research has advanced the understanding of *Cryptococcus neoformans* var. *grubii* and its role in human pathogenicity, studies of clinical and environmental isolates have not linked genotype to important phenotypes including virulence. Much of our understanding of the population structure of this organism is based on variations in a few genes. There are other underlying differences that these marker sets do not detect. This study aimed to overcome the challenge presented by limited marker sets by utilizing whole genome SNP genotype data, and the data used in this study is more comprehensive than previous studies using *C. neoformans* var. *grubii*.

Using this SNP genotype data, we re-evaluated major subdivisions in the *C. neoformans* var. *grubii* population and identified potential markers that distinguish clinical strains from environmental strains by performing a GWAS. Previous fungal GWAS focused on small strain datasets with less than 100 individuals and, at most, 30,000 SNP genotypes, yet still reported successful identification of markers for
virulence (Muller et al., 2011; Dalman et al., 2013). This study included a more comprehensive strain set with over 1 million markers, providing more resolution, and is the first use of GWAS for a major fungal pathogen.
Chapter II
Materials and Methods

The following section details the analysis employed in this study. Briefly, genotype data from a large cohort of environmental and clinical isolates of *Cryptococcus neoformans* var. *grubii* were used to evaluate population structure. Moreover, we used these data to identify and characterize hybrid and introgressed strains. Finally, phenotype and genotype data were combined in genome wide association analysis to select markers with significant associations between our clinical and environmental strains.

Variant Detection

A large cohort of isolates (n=392), both environmental and clinical was collected (Table 1), and high quality genotype data were produced. Briefly, the Broad Institute Genomics Platform generated 100-bp short insert mate pair data for each strain on the Illumina HiSeq sequencing platform. Using the BWA-MEM algorithm v0.7.12 (Li, 2013), the sequencing read data were aligned to the *C. neoformans* var. *grubii* H99 assembly (Janbon et al., 2014; Genbank accession GCA_000149245.2). Alignments around insertion/deletion (indel) events were corrected using the Genome Analysis Toolkit (GATK) v3.4 software, indelRealigner (McKenna et al., 2010), and genotype calls were generated from these alignments using GATK HaplotypeCaller in GVCF
Finally, each strain’s GVCF output were merged using GATK GenotypeGVCFs to produce variant calls for each strain.

To ensure high-quality genotype calls, variants were filtered using GATK VariantFiltration software to remove variants with low-quality mapping scores (MQ < 40.0), significant strand bias (FS > 60.0), and low sequence quality (QD < 2.0). Genotype calls were further filtered for a minimum genotype quality of 50, read coverage depth of 10, and fraction alternate allele of 0.8. These filtered variant calls were annotated for functional effects using SnpEff v4.2 (Cingolani et al., 2012). In total, over 1 million SNP variant sites were generated for analysis.

<table>
<thead>
<tr>
<th>Subpopulation</th>
<th>Total Strains</th>
<th>Clinical Strains</th>
<th>% Clinical Strains</th>
</tr>
</thead>
<tbody>
<tr>
<td>VNI</td>
<td>185</td>
<td>151</td>
<td>81.62%</td>
</tr>
<tr>
<td>VNII</td>
<td>16</td>
<td>14</td>
<td>87.50%</td>
</tr>
<tr>
<td>VNB</td>
<td>191</td>
<td>108</td>
<td>56.54%</td>
</tr>
</tbody>
</table>

Investigation of Population Structure

A phylogeny was inferred from the variant positions across all strains. A concatenated sequence was generated for each strain from variant sites containing at least one unambiguous site and no more than 10% ambiguous sites. In these strain consensus sequences, indel events were ignored to maintain variant call alignment. RAxML v8.2.4 (Stamatakis, 2014) using the GTRCAT model in rapid bootstrapping mode with 1,000 replicates was used to estimate phylogenetic trees.
To further define the population structure in the strains, we ran the STRUCTURE software v2.3.4 in site-by-site mode (Pritchard et al., 2000) using our full variant dataset. Known bugs in the software when running in this mode required us to update the code before evaluating our variant data. However, due to STRUCTURE’s inability to scale with extremely large datasets and its lack of multithreaded support, we were unable to run this full dataset; thus, requiring us to create a subset of our variant data. In our subsampling approach, we first filtered out sites with less than two SNPs total and greater than 5% missing data across strains. Next, we randomly subsampled the remaining sites at known frequencies. For our full strain dataset, we used a 50% subsampling frequency that allowed STRUCTURE to complete in roughly a month. Once this analysis was complete, we removed potential hybrid strains variant data from our full variant dataset and subsampled the remaining sites at a 15% frequency. This lower subsampling threshold was necessary to allow the STRUCTURE analysis to complete in a more reasonable runtime. We employed these same methods to investigate finer-scale population structure within two of the larger cohort of strains, VNI and VNB.

Evanno (2005) suggested a statistic, $\Delta K$, for determining the number of populations, $K$, to use as input to STRUCTURE. Calculation of this statistic requires replicate STRUCTURE runs, so we ran STRUCTURE across 10 replicates with $K$ values ranging from 2 to 9. Each replicate was run using a different random subset of approximately 10,000 SNPs. We generated summary results and $\Delta K$ statistics using structureHarvester (Earl, 2012) to help understand the optimal $K$ for STRUCTURE.

This $K$ value was also compared to the number of clusters detected by PCA of the SNP data and major clades present in phylogenetic trees. To compare to our
STRUCTURE results, we ran principal components analysis on the full genotype set across all strains using the software, smartPCA v9102 (Patterson et al., 2006). In addition to the full strain set, we ran smartPCA on each of the larger cohort of strains, VNI and VNB, and compared the population structure within these subpopulations to those found in our STRUCTURE runs. Upon examination of the PCA results, we found that many strains had standard deviation values for the top principal components greater than the default “outliersigmathresh” value of 6.0. In these cases, we ran the smartPCA analysis to evaluate the maximum standard deviation observed in the dataset, and then we adjusted the “outliersigmathresh” argument values to prevent strains being removed from our analysis.

Identification of Hybrid Strains

For a small number of strains, STRUCTURE revealed large contributions from multiple ancestry groups. To examine these more closely, for each strain in our STRUCTURE analysis, we extracted the individual marker population ancestries from the site-by-site output. Strains with less than 90% of a single ancestry were selected as potential hybrid strains. To examine the chromosomal distribution of sites of each ancestry in these strains, we combined the ancestries of single markers into larger genomic segments. First, we assigned a marker to a population if the fraction of population ancestry was $\geq 0.9$. We combined consecutive markers with shared population into larger chromosome segments and visualized the results using Python’s matplotlib package to identify strains exhibiting mixed ancestry.
Analysis of Genomic Introgressions

To more precisely map small regions of secondary ancestry, corresponding to introgressions between the ancestry groups, we partitioned our variant marker data into 500kb segments along each chromosome and then ran STRUCTURE in site-by-site mode. For each segment, we used STRUCTURE output coupled with the strain population information to assign each sublineage (VNI, VNII, VNBI/II) to a primary ancestry. Briefly, strains were assigned to one of four ancestry populations based on the maximum fraction ancestry value output by STRUCTURE. We mapped each ancestry population to a sublineage based on the maximum number of sublineages found in that population and discarded segments for which we found incomplete separation of the sublineages. In total, we evaluated 35 genomic segments covering 15.7Mb (83%) of the genome.

For each 5kb window within these 35 chromosome segments, we identified markers that had a fraction ancestry $\geq 0.9$. We labeled markers as an alternate population if the ancestry population did not match the determined sublineage population. Next, we assigned each window to a sublineage based on the presence of alternate population ancestry. Finally, we combined these segments into full-length chromosomes and visualized the results using Python’s matplotlib package to identify strains exhibiting introgressed windows.
Classification of Clinically Associated Markers

For our genome wide association study (GWAS), we limited our analysis to the VNB subpopulation due to the diverse population of clinical and environmental strains. First, we combined the annotated variants from SnpEff with our genotype data and removed marker sites that were classified by SnpEff as “low impact” effects, e.g. synonymous mutations. Next, we filtered out marker sites with ≥ 10% ambiguous variant calls. Those individual marker sites with ≤ 5% variant calls across the strains were aggregated into their respective gene features. We converted these data into Plink formatted ped and map files (Purcell et al., 2007). To look for loss of function associations, we filtered the marker set based on annotated frameshift or gained stop variants. Because these variants are rare and may occur at different locations along a gene, these variants were aggregated across their associated gene regions.

We ran our GWAS using five different phenotypes for association. First, we examined the variants in the context of clinical versus environmental isolates. Next, we examined associations for each of the strains based on growth measurements following exposure to the anti-fungal agent, fluconazole, and the oxidative stress agents, paraquat and hydrogen peroxide. Finally, we examined associations based on melanization measurements following exposure to laevodihydroxyphenylalanine (L-DOPA). Prior to each association test, we ran GEMMA (Zhou & Stephens, 2014) to build a relatedness matrix to reduce the effect of population structure on our analysis. We ran association tests using both variant datasets described above with each of the phenotype datasets and
evaluated the associations based on $p$-value scores. To prioritize markers for further investigation, we plotted the observed $p$-value scores versus the expected $p$-value scores and identified markers that fell above the diagonal. For each phenotype dataset, we identified the top 20 significant markers. We examined the genes affected and their possible roles in the observed phenotypes.
Chapter III

Results

This study aimed to further elucidate the population structure of the opportunistic pathogen, *Cryptococcus neoformans* var. *grubii*, and identify potential markers significantly associated with virulence. This goal required a large cohort of environmental and clinical strains across the previously identified sublineages: VNI, VNII, and VNB. High-quality variant data from these strains were used in standard population software to delineate population structure and identify genomic recombination events. The genotype data combined with phenotype data were then used in a genome-wide association study. Potential markers were identified based on observed $p$-value score for further evaluation.

Investigation of Population Structure

Phylogenetic analysis using all variant data recapitulated the previously described subpopulations of *Cryptococcus neoformans* var. *grubii*: VNI, VNII, and VNB. However, there was underlying support in the phylogeny for further separation of VNB into two distinct sublineages that we term VNBI and VNBII (Figure 8). When given an estimated number of populations replicating the previously characterized sublineages ($K=3$), STRUCTURE analysis placed both VNBI and VNBII sublineages together in a single population. However, using an estimated population of four ($K=4$), we found that
those strains contained within the proposed VNBI sublineage were assigned to a different population than the VNBII strains, further confirming our phylogenetic results. Moreover, the STRUCTURE output showed surprisingly little admixture between populations as evidenced by the absence of alternate ancestry within each strain across populations (Figure 1a).

The Evanno method of calculating $\Delta K$ produced multiple $K$ values that did not match our expected population number based on phylogeny. However, the values of $\Delta K$ were extremely low and may be attributable to limit of detection fluctuations (Figure 2). Despite this lack of clearly defined $K$ using the Evanno method, our STRUCTURE run with an estimated $K=4$ showed consistency with the four major sublineages identified by phylogenetic analysis and identified a complete separation of the sublineages based on the high fraction ancestry observed in each strain (Figure 1a).

PCA analysis on the full variant dataset also demonstrated these four major sublineages with strains from each sublineage clustered tightly together (Figure 9). Further, the phylogenetic analysis suggested a tripartite split in the VNI sublineage. STRUCTURE and PCA analysis of the VNI strains also demonstrated this split (Figure 3), though estimates of linkage disequilibrium suggest these VNI sublineages are not isolated but are recombining (Desjardins 2016, submitted). Taken together, this analysis suggests that the Cryptococcus neoformans var. grubii population be redefined to include the two VNB subpopulations, but not include the VNI subpopulations.
Figure 1. STRUCTURE and smartPCA Subpopulation Results (387 strains). STRUCTURE results (a) show lack of admixture of populations as denoted by lack of alternate populations within strains. SmartPCA results (b) show separation of the subpopulations; however, a tight cluster of VNB strains within first two principal components is observed.

Figure 2. Evanno $\Delta K$ Results. Delta $K$ values for 20 replicates at $K=(2 .. 8)$ show variable results at low values.
Identification of Hybrid Strains

Though our STRUCTURE analysis of the full strain dataset largely showed single ancestry of each strain within the subpopulations, there were seven strains exhibiting mixed ancestry. These hybrid strains were further verified by our PCA analysis, in which each hybrid strain did not cluster with any single subpopulation and were placed instead between two different clusters (Figure 4b). Based on STRUCTURE analysis, five of these strains were VNI-VNB hybrids while the remaining two were VNII-VNB (Figure 4a). Further, the VNI-VNB hybrids exhibited approximately 40% VNI ancestry and 60% VNB ancestry, but the VNII-VNB hybrids showed less mixed ancestry with 85% VNII ancestry and 15% VNB ancestry (Table 2). Of the five VNI-VNB hybrids, three strains
isolated from similar sources, HIV+ individuals in Botswana, exhibited high nucleotide identity (average of only 39 variant SNP positions) and, thus, close proximity on the phylogenetic tree. These similarities suggest that these hybrid strains represent the same recombination events. For the VNII-VNB hybrids, we found two such strains. Taken together, we found four unique hybrid strains: three VNI-VNB with roughly equal ancestry and one VNII-VNB with higher VNII ancestry.

Upon examination of the contribution of each ancestry across the genome, we found different patterns of ancestry between the VNI-VNB and VNII-VNB hybrids. Notably, the VNII-VNB hybrid strain, CCTP51, exhibited highly intermixed ancestries (Figure 10d) with a mosaic of VNB and VNI regions interspersed somewhat evenly across the entire genome. Conversely, the VNI-VNB strains contained less intermixing, attributable to large segments of single ancestry. Notably, the VNI-VNB hybrid, Bt131, appears to have two chromosomes, 6 and 9, that are nearly all derived from VNB ancestry (Figure 10b). The remaining VNI-VNB hybrids, Bt125 and Ftc260-1, also exhibited large segments representing VNB ancestry (Figure 10a,c), with unique patterns in each isolate. These differences in the fraction of mixed ancestry between VNI-VNB and VNII-VNB hybrids suggest differing timelines of lineage mixture. These VNI-VNB hybrids likely represent recent events while the VNII-VNB hybrids may be more ancient.
Figure 4. STRUCTURE and smartPCA Results of 392 Strains Showing Hybrid Strains. STRUCTURE results (a) show admixture of ancestry groups in VNB and VNII subpopulation strains. SmartPCA results (b) show hybrid strains situated between the subpopulations.

Table 2. Fraction ancestry of Hybrid Strains

<table>
<thead>
<tr>
<th></th>
<th>VNB</th>
<th>VNI</th>
<th>VNII</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bt125</td>
<td>0.5677</td>
<td>0.4323</td>
<td>0</td>
</tr>
<tr>
<td>Bt131</td>
<td>0.6018</td>
<td>0.3982</td>
<td>0</td>
</tr>
<tr>
<td>Bt162</td>
<td>0.6001</td>
<td>0.3999</td>
<td>0</td>
</tr>
<tr>
<td>Bt163</td>
<td>0.6004</td>
<td>0.3996</td>
<td>0</td>
</tr>
<tr>
<td>Ftc260-1</td>
<td>0.5361</td>
<td>0.4639</td>
<td>0</td>
</tr>
<tr>
<td>CCTP51</td>
<td>0.1430</td>
<td>0</td>
<td>0.8570</td>
</tr>
<tr>
<td>MW_RSA852</td>
<td>0.1403</td>
<td>0</td>
<td>0.8597</td>
</tr>
</tbody>
</table>

Note: Each row denotes a hybrid strain. Column values represent the fraction of genome derived from labeled ancestry.
Analysis of Genomic Introgressions

To examine if small regions not detected by STRUCTURE are exchanged between the lineages, the genome was partitioned into 500kb genomic segments along each chromosome, and we identified introgressions between lineages in our non-hybrid strain set. Though nearly all of the strains appeared to be of single ancestry using our full dataset, this finer scale analysis revealed the presence of smaller regions of alternate ancestry. In total, 118 strains contained at least one 5kb window with ≥ 75% alternate ancestry. Moreover, 40% of the introgressed strains contained ≥ 10 windows, corresponding to regions totaling 50 to 260kb or 0.02-1.4% total genome (Figure 5). Of these strains, 32 were from the VNBI subpopulation, 14 were from VNI, and 5 were from VNBII; we did not observe any VNII strains containing introgressions, though we had fewer VNII strains in our analysis than other sublineages (Figure 11). Interestingly, we found introgressed regions originating from each of the sublineages, including from VNII.

While we found no specific functional enrichment of genes within these introgressed regions, we did find introgressed regions shared across strains. This includes a shared 40kb VNI-derived introgression across 14 VNBI strains. This genomic region contains a glycolipid mannosyltransferase gene implicated in capsule formation (O’Meara 2012) and a proteolytic carboxypeptidase D secretome component (Geddes 2015). We observed a 60kb VNBI-derived introgression across 8 VNI strains. Along this region lies VCXI, a known virulence factor (Kmetzsch 2010), and ERG11, the azole
drug target. Contrasted with the whole genome STRUCTURE results where these strains appeared to be from a single ancestry, this chromosomal segment analysis suggests genomic intermingling at the gene level.

Figure 5. Introgression Counts Per Strain. Breakdown of sublineage strains and the total number of 5kb introgressions found.

Classification of Clinically Associated Markers

With our extensive collection of clinical and environmental strains within the VNB sublineage and our high quality genotype data, we were able to detect significant variants in a fungal GWAS analysis on our data. Our strain dataset included 112 clinical isolates, 73 environmental isolates, and 1 isolate of unknown origin while our genotype dataset containing high impact variants totaled 246,608 positions while our frameshift
and loss of function dataset totaled 2,463 positions distributed across the genome. GEMMA identified 11 sites from our high impact genotype set and 1 site from our frameshift/stop set with association \( p \)-value scores \( \leq 1 \times 10^{-5} \) (Table 3). These observed association \( p \)-value scores were slightly higher than the calculated expected scores (Figure 6). The variants significantly associated include 6 either in known virulence factors or in the intergenic region nearby. One of the top associations was downstream of a velvet family protein implicated in filamentation (Chacko et al., 2015), and lack of filamentation has previously been attributed to increased virulence (Wang et al., 2012; Zhai et al., 2015).

Fungal virulence is further increased by an ability to survive within the host. During infection, the host mounts an immune response including the release of oxidative
stress agents, so in this case, the ability of a strain to withstand these conditions increases its virulence capacity (Brown, Haynes, & Quinn, 2009). To identify potential markers associated with oxidative stress response, we measured the ability of each strain to grow in the presence of the oxidative stress agents, paraquat and hydrogen peroxide. Using the same genotype datasets used in our clinical versus environmental GWAS coupled with the normalized growth scores obtained from the oxidative stress measurement, we ran a GWAS for both the paraquat and hydrogen peroxide assays. Our paraquat GWAS did not find any sites with observed association \( p \)-value scores \( \leq 1e-5 \) nor any sites with \( p \)-values above the range of the expected scores. In our hydrogen peroxide GWAS, we found one high impact feature with an association \( p \)-value scores \( \leq 1e-5 \). However, this variant resides within a hypothetical protein (CNAG_00167) that has not been further characterized.

We extended our phenotypic characterization assay to include a screen for growth in the presence of fluconazole. Fluconazole inhibits fungal growth and is routinely used as part of a maintenance therapy in the treatment of cryptococcosis (Perfect & Bacanic, 2015). A recent study of HIV+ individuals in Uganda highlights the increased clinical strain resistance to fluconazole (Smith et al., 2015). Because of this increased resistance and the fact that fluconazole is a readily available therapy in sub-Saharan Africa, the opportunity for effective treatment will be reduced in those regions most affected (Smith et al., 2015). To identify potential markers contributing to fluconazole resistance, we measured strain growth in the presence and absence of fluconazole. We used these growth scores in conjunction with our SNP genotype data to run a GWAS. In this fluconazole GWAS, we found 6 high impact variant and one frameshift variant feature;
however, these variant sites were mostly isolated to hypothetical proteins or intergenic regions. We further examined our GWAS results for associations at any $p$-value score related to fluconazole resistance in the gene, $ERG11$ (Sionov et al., 2012). We then examined our genotype data for any variations within $ERG11$. In both analyses, we did not find any related sites.

Melanization is one of the three main virulence factors of $C. neoformans$ var. $grubii$, thought to protect the organism from environmental stress. Since melanin precursors are not produced naturally by $Cryptococcus$, a strain must acquire the necessary substrates from the environment (Kwon-Chung, Tom, & Costa, 1983). The laccase enzyme, $LAC1$, is essential to melanin production via the conversion of the substrate, L-DOPA, to melanin intermediates (Williamson 1994). To determine the melanization capacity of collected strains, we measured the pigmentation of each strain grown in the presence of L-DOPA. These melanin scores and SNP genotype data were used to perform a GWAS analysis. This L-DOPA melanization GWAS identified 66 high impact SNP sites with $p$-value scores $\leq 1e-5$ (Table 4), though because of a distinctive step pattern in the observed versus expected $p$-value scores, some of these associations may be symptomatic of co-segregating variants (Figure 7a). Though we accounted for population structure in our GWAS, these high impact SNP variants are adjacent in the genome and may all come from a single subpopulation of strains, like VNBII. However, there were frameshift genotypes that exhibited a noticeably higher observed score than expected one gene, $BZP4$ (Figure 7b). Prior work demonstrated that deletion of $BZP4$ reduced melanization capacity (Jung et al., 2015). There were 4 clinical strains exhibiting an associated variant within this gene, 2 from the VNBI and 2 from the
VNBI sublineage (Table 5). Of these 4 strains, 3 showed reduced melanization within range of the control strain, a deletion strain missing the $LAC1$ gene necessary for melanization (Salas et al., 1996).

Figure 7. L-DOPA GWAS Q-Q Plot. Q-Q plot of GWAS marker $p$-values of L-DOPA high impact (a) and loss of function variants (b). $BZP4$ loss of function variant marked in red.

In our clinical versus environmental strain GWAS analysis, we identified markers proximal to known virulence genes significantly associated with clinical strains. The environmental stressors and drug resistance GWAS found potential new targets around uncharacterized hypothetical proteins. Finally, we detected significantly associated markers in a gene implicated in melanin production. Taken together, the genotype data coupled with the various phenotype data was able to identify numerous sites associated with clinical isolates.
### Table 3. Clinical versus environmental markers significantly associated with clinical strains identified by GWAS analysis

<table>
<thead>
<tr>
<th>Location</th>
<th>$p$-value</th>
<th>Gene(s)</th>
<th>Variant Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNAG_01348</td>
<td>2.02E-07</td>
<td>cyanate hydratase</td>
<td>Loss of function</td>
</tr>
<tr>
<td>downstream of CNAG_07989*</td>
<td>3.00E-07</td>
<td>velvet family protein</td>
<td>SNP</td>
</tr>
<tr>
<td>upstream of CNAG_07661</td>
<td></td>
<td>hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>downstream of CNAG_07972</td>
<td>3.22E-07</td>
<td>hypothetical protein, chromosome 3 centromere</td>
<td>SNP</td>
</tr>
<tr>
<td>upstream of CNAG_06879</td>
<td></td>
<td>hypothetical protein, chromosome 3 centromere</td>
<td></td>
</tr>
<tr>
<td>CNAG_05440*</td>
<td>1.19E-06</td>
<td>pyridoxamine 5'-phosphate oxidase PDX3</td>
<td>SNP</td>
</tr>
<tr>
<td>CNAG_03374</td>
<td>2.44E-06</td>
<td>DNA replication complex GINS protein PSFI</td>
<td>SNP</td>
</tr>
<tr>
<td>downstream of CNAG_06765*</td>
<td>2.91E-06</td>
<td>low mating performance LMP1</td>
<td>SNP</td>
</tr>
<tr>
<td>upstream of CNAG_06764*</td>
<td></td>
<td>short chain dehydrogenase FOX21</td>
<td></td>
</tr>
<tr>
<td>CNAG_06921</td>
<td>3.99E-06</td>
<td>hypothetical protein</td>
<td>SNP</td>
</tr>
<tr>
<td>CNAG_05037</td>
<td>5.07E-06</td>
<td>hypothetical protein</td>
<td>SNP</td>
</tr>
<tr>
<td>CNAG_07427*</td>
<td>5.62E-06</td>
<td>Casein Kinase 1CCK1</td>
<td>SNP</td>
</tr>
<tr>
<td>CNAG_01464*</td>
<td>6.00E-06</td>
<td>Flavohemoglobin YHBf</td>
<td>SNP</td>
</tr>
<tr>
<td>downstream of CNAG_07972</td>
<td></td>
<td>hypotheical protein, chromosome 3 centromere</td>
<td></td>
</tr>
<tr>
<td>upstream of CNAG_06879</td>
<td>6.84E-06</td>
<td>hypotheical protein, chromosome 3 centromere</td>
<td></td>
</tr>
<tr>
<td>CNAG_02007</td>
<td>9.28E-06</td>
<td>adenylate kinase 1</td>
<td>SNP</td>
</tr>
</tbody>
</table>

Note: Markers associated with virulence denoted with asterisk (*).

### Table 4. Top ten L-DOPA markers significantly associated with L-DOPA utilization

<table>
<thead>
<tr>
<th>Location</th>
<th>$p$-value</th>
<th>Gene(s)</th>
<th>Variant Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNAG_03346*</td>
<td>3.43E-09</td>
<td>hypothetical protein</td>
<td>Loss of function</td>
</tr>
<tr>
<td>CNAG_01996</td>
<td>4.39E-09</td>
<td>hypothetical protein</td>
<td>SNP</td>
</tr>
<tr>
<td>downstream of CNAG_06251</td>
<td></td>
<td>Ser/Thr protein phosphatase family protein</td>
<td></td>
</tr>
<tr>
<td>upstream of CNAG_06252</td>
<td>2.53E-08</td>
<td>hypothetical protein</td>
<td>SNP</td>
</tr>
<tr>
<td>downstream of CNAG_06251</td>
<td></td>
<td>Ser/Thr protein phosphatase family protein</td>
<td></td>
</tr>
<tr>
<td>upstream of CNAG_06252</td>
<td>2.53E-08</td>
<td>hypothetical protein</td>
<td>SNP</td>
</tr>
<tr>
<td>downstream of CNAG_06252</td>
<td></td>
<td>hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>upstream of CNAG_06253</td>
<td>2.53E-08</td>
<td>sugar transporter</td>
<td>SNP</td>
</tr>
<tr>
<td>CNAG_07009</td>
<td>2.53E-08</td>
<td>hypothetical protein</td>
<td>SNP</td>
</tr>
<tr>
<td>downstream of CNAG_06250</td>
<td></td>
<td>hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>upstream of CNAG_06251</td>
<td>2.85E-08</td>
<td>Ser/Thr protein phosphatase family protein</td>
<td>SNP</td>
</tr>
<tr>
<td>downstream of CNAG_06250</td>
<td></td>
<td>hypothetical protein</td>
<td></td>
</tr>
<tr>
<td>upstream of CNAG_06251</td>
<td>2.85E-08</td>
<td>Ser/Thr protein phosphatase family protein</td>
<td>SNP</td>
</tr>
</tbody>
</table>

Note: BZP4 marker denoted with asterisk (*).
Table 5. Melanization Scores for Loss of Function BZP4 Markers

<table>
<thead>
<tr>
<th>Strain Name</th>
<th>Melanization Score</th>
<th>Subpopulation</th>
<th>Strain Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bt147</td>
<td>0.22</td>
<td>VNBI</td>
<td>clinical</td>
</tr>
<tr>
<td>Bt102</td>
<td>0.72</td>
<td>VNBI</td>
<td>clinical</td>
</tr>
<tr>
<td>Bt75</td>
<td>0.70</td>
<td>VNBII</td>
<td>clinical</td>
</tr>
<tr>
<td>Bt103</td>
<td>0.73</td>
<td>VNBII</td>
<td>clinical</td>
</tr>
<tr>
<td>Lac1</td>
<td>0.69</td>
<td>N/A</td>
<td>control</td>
</tr>
</tbody>
</table>

Note: Strains exhibiting BZP4 loss of function variants exhibit similar melanization scores to the Lac1 reduced melanization mutant.
Chapter IV

Discussion

Using a large collection of *C. neoformans* var. *grubii* strains, this study helped to further characterize the population structure and variation of this important human pathogen. Though our goal was to investigate population structure to reduce our GWAS bias, the population analysis itself revealed a number of interesting discoveries including a proposed distinct sublineage of VNB, the identification of hybrid strains, and the presence of genomic introgressions in numerous strains. Further, by identifying and removing hybrid strains from the analysis, the GWAS was not affected by these unusual patterns of ancestry. Use of the population structure also enabled confirmation of whether potential markers were independently associated with different phenotypes or likely due to a single change and shared ancestry.

Previous analysis using MLST of seven loci broadly defined the *C. neoformans* var. *grubii* subpopulations as VNI, VNII, and VNB (Litvintseva et al., 2006). Though that analysis utilized an extensive strain set, the MLST lacked the resolution to interrogate population thoroughly. In addition, the seven MLST loci include several virulence genes and may be more prone to selective pressure than phylogenies inferred from genome-wide variants. Here, we found that the population may be redefined to include a split within the VNB sublineage, denoted VNBI and VNBII, thus, further illustrating the utility of next-generation sequencing and whole genome SNP genotyping methods in fungal analysis. The significance of this split remains to be determined,
though, notably, the VNBII sublineage included 60 clinical isolates out of 70 strains total (Desjardins et al., 2016). For instance, the predominance of clinical strains within the VNBII sublineage may be attributed to an enhanced ability for growth in humans. However, we cannot discount the possibility that we have not identified the environmental niche or adequately sampled the geography of the VNBII sublineage. We found three partitions within the VNI sublineage, but, based on recombination between these three groups, the VNI sublineages do not represent distinct subpopulations (Desjardins et al., 2016). While the lack of frequent recombination between VNI, VNII, VNBI, and VNBII suggests that these lineages could be separate similar to those used for C. gattii (Hagen et al., 2015), further sequencing of additional diverse isolates would help confirm that the population has been well sampled making such subdivisions.

While most isolates fall into distinct sublineages, our analysis revealed the presence of four distinct hybrid strains with widespread mixed ancestry: 3 VNI-VNB and 1 VNII-VNB (Figure 10). The VNI-VNB hybrids contained large tracts of each sublineage suggesting more recent genomic intermixing within these strains. Although the percentage of each sublineage were similar across the VNI-VNB hybrids, phylogenetic analysis using sublineage specific markers revealed that the parental lineages of each hybrid differ (Rhodes et al., 2016). Two of these VNI-VNB hybrid strains and the VNII-VNB strain were from clinical sources while the remaining VNI-VNB hybrid strain was from the environment. Naturally occurring hybrid fungal strains in nature may confer a selective advantage (Morales & Dujon, 2012) or can contribute to the evolution of virulence (Friesen et al., 2006). Therefore, the need exists to determine if these clinical hybrid strains exhibit increased virulence compared to non-hybrid clinical
strains. For the environmental hybrid strain, the virulence profile needs to be examined. Potential differences in virulence between this strain and the clinical strains may be of interest for looking at different genomic ancestry regions to help understand virulence.

Genomic intermixing was not limited to the hybrid strains, as we found nearly one-third of our strains contained at least one 5kb region derived from an alternate ancestry. In the human pathogenic fungi, *Coccidioides immitis*, introgressions were observed to be derived from a sister species, *Coccidioides posadasii* (Neafsey et al., 2010). One of these regions contained a gene that showed increased expression in *Coccidioides* parasitic growth phase suggesting a role in virulence (Whiston et al., 2012). In *Neurospora crassa*, an introgressed region was found in strains collected from Louisiana (Ellison et al., 2012). Further, using gene deletion analysis, researchers determined that genes in these introgressed regions may be essential for growth in cooler climate (Ellison et al., 2012). Together, these studies suggest that introgressions are important for fungal adaptation. In our study, we did find shared regions of introgressions between strains encompassing potential virulence genes and the target of the frontline antifungal fluconazole. However, there remains a need to see if there are any particular phenotypes or gene expression differences exhibited by these introgressed strains versus non-introgressed strains.

To help understand genotypic associations to phenotype, we successfully employed GWAS to detect variants associated with clinical origin, oxidative stress resistance, antifungal resistance, and level of melanization. This provides one of the few examples of GWAS analysis of fungi (Muller et al., 2011; Dalman et al., 2013; Palma-Guerrero et al., 2013). Our identification of potential markers associated with clinical
isolates coupled with markers associated with melanization validates this approach in fungi. In our GWAS analysis between environmental and clinical strains, we found a significant association in the intergenic region downstream from a velvet family protein (CNAG_07989). This protein has been linked to filamentation (Chacko et al., 2015), and, importantly, the transition of Cryptococcus neoformans var. grubii from filamentous form to yeast form has previously been implicated in increased virulence (Wang et al., 2012; Zhai et al., 2015). Since this variant lies outside of the coding sequence and because of this significant association score, its close proximity to the gene may indicate a role in the regulation of gene expression (Macintyre et al., 2014). Further analysis is needed to determine if this intergenic region contains a regulatory element and how variations in this region affect filamentation and, ultimately, virulence. More interestingly, our melanization GWAS identified significant markers within BZP4 for four clinical strains. Melanin has been implicated in virulence due to its protection from oxidative stress (Liu et al., 2008; Kronstad, Jung, & Hu, 2008), and disruption of BZP4 has been linked to reduced melanization capacity (Jung et al., 2015). Three of the four clinical strains with BZP4 variants showed phenotypically lower melanin scores. Though the remaining clinical strain contained a significantly associated variant for melanization, it exhibited normal melanization capacity, suggesting the possible presence of an additional compensatory variant that restored melanin production.

Because of its worldwide health impact, understanding Cryptococcus neoformans var. grubii and its virulence is an utmost concern. In our genomic analysis, we discovered interesting population characteristics previously unresolved. The distinct subpopulation definitions will be useful for prospective comparative genomics and
additional phenotypic assays. The presence of hybrid strains and genomic introgressions suggests active exchange of material between strains, a possible driver for adaptation. How these events shape strain evolution will be an interesting future study. Using this phylogenetic background, we set out to identify natural variants associated with important disease related phenotypes; such variation in the pathogen may be important to consider in either clinical diagnostics or the design of therapeutic targets. This work provides a foundation for future GWAS to examine additional phenotypes to identify markers of virulence that may be informative in the clinic.
Appendix

Additional Figures

Figure 8. Phylogenetic Tree of *Cryptococcus neoformans* var. *grubii*. Phylogenetic tree of the 392 strains used in the study with putative split of VNB sublineage marked in red. Courtesy Chris Desjardins (Broad Institute).
Figure 9. Top three principal components from smartPCA. Plot of top 3 principal components using 387 strains highlights the proposed split with the VNB sublineage.
Figure 10. Chromosome Ancestry of Hybrid Strains. Fraction ancestry across genome of 3 VNB-VNI hybrids and 1 VNII-VNB hybrid.
Figure 11. Genomic Introgressions Plot. Introgressed regions within strains across the genome. Chromosomes are along the x-axis (not to scale). The y-axis contains strains that harbored $\geq 10$ introgressed windows. Strain sublineages are provided on the left.
References


http://doi.org/10.1038/nmeth.2848