A new lineage of *Cryptococcus gattii* (VGV) discovered in the Central Zambezi Miombo Woodlands

Rhys A. Farrer[1,2,3,4], Miwha Chang[5], Michael J. Davis[5], Lucy van Dorp[3], Dong-Hoon Yang[5], Terrance Shea[4], Thomas R. Sewell[6], Wieland Meyer[7], Francois Balloux[3], Hannah M. Edwards[6], Duncan Chanda[8], Geoffrey Kwenda[9], Mathieu Vanhove[6], Yun C. Chang[5], Christina A. Cuomo[4], Matthew C. Fisher[6], Kyung J. Kwon-Chung[5]

[1]Medical Research Council Centre for Medical Mycology, University of Exeter, Exeter, EX4 4PY, UK.

[2]Medical Research Council Centre for Medical Mycology, University of Aberdeen, Aberdeen, AB242TN, UK.

[3]UCL Genetics Institute, University College London, Gower Street, London, WC1E 6BT, UK.

[4]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America.

[5]Molecular Microbiology Section, Laboratory of Clinical Immunology and Microbiology, National Institute of Allergy and Infectious Diseases, National Institute of Health, Bethesda, Maryland, United States of America.

[6]MRC Centre for Global Infectious Disease Analysis, Imperial College London, London, United Kingdom

[7]Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Faculty of Medicine and Health, Sydney Medical School, Westmead Clinical School, Marie Bashir Institute for Infectious Diseases and Biosecurity, The University of Sydney, Westmead Hospital (Research and Education Network), Westmead Institute for Medical Research, Sydney, NSW, Australia

[8]Adult centre of Excellance, University Teaching Hospital, Lusaka, Zambia

[9]Department of Biomedical Sciences, School of Health Sciences, University of Zambia, Lusaka, Zambia

Corresponding author: Kyung J. Kwon-Chung (jkchung@niaid.nih.gov)

Abstract

We discovered a new lineage of the globally important fungal pathogen *Cryptococcus gattii*, based on six isolates collected from three locations spanning the Central Miombo Woodlands of Zambia, Africa. All isolates were from environments (middens and tree holes) that are associated with a small mammal, the African hyrax. Phylogenetic and population genetic analyses confirmed that these isolates form a distinct, deeply divergent lineage, which we name VGV. VGV comprises two subclades (A and B) that are capable of causing mild lung infection with negligible neurotropism in mice. Comparing the VGV genome to previously identified lineages of *C. gattii* revealed a unique suite of genes together with gene loss and inversion events. However, standard *URA5* RFLP analysis could not distinguish between VGV and VGIV isolates. We therefore developed a new *URA5* RFLP method that can reliably distinguish the newly described lineage. Our work highlights how sampling understudied ecological regions alongside genomic and functional characterisation can broaden our understanding of the evolution and ecology of major global pathogens*.*

Importance

Cryptococcus gattii is an environmental pathogen that causes severe systemic infection in immunocompetent individuals more often than in immunocompromised humans. Over the past two decades, researchers have shown *C. gattii* falls within four genetically distinct major lineages. By combining field work from an understudied ecological region (the Central Miombo Woodlands of Zambia, Africa), genome sequencing and assemblies, phylogenetic and population genetic analyses, and phenotypic characterization (morphology, histopathological, drug-sensitivity, survival experiments) we discovered a hither to unknown lineage which we name VGV (variety *gattii* five). The discovery of a new lineage from an under studied

ecological region has far-reaching implications for the study and understanding of fungal pathogens and diseases they cause.

Introduction

Cryptococcosis is a severe fungal infection responsible for high levels of mortality and morbidity worldwide(1). The etiological agents are two species complexes of the genus *Cryptococcus*: *C. neoformans* and *C. gattii.* Whilst, the first described cases of clinical cryptococcosis due to these two distinct species complexes were reported in the mid-1890s under the names *Saccharomyces hominis*(2) and *S. subcutaneous tumefacience*(3) respectively, clinical *Cryptococcus* isolates have been taxonomically treated as a single species (*C. neoformans*) for more than 100 years(4). Heterogeneity among cryptococcosis-causing yeast isolates became increasingly apparent from the middle of the 20th century onward, and led to the recognition of four serotypes (A, B, C, D) based on the antigenic determinant of capsular polysaccharide(5, 6). Subsequent discovery of two distinct sexual cycles produced by the isolates of A/D vs. B/C serotypes(7, 8) and phylogenetic analysis using various gene sequences(9–11) confirmed these complexes to be genetically divergent enough to be considered as separate species. Thus, in 2002, the isolates of serotype B/C were formally classified as *C. gattii*(12) while *C. neoformans* includes all serotype A/D strains(13).

Over the past two decades, population structure analysis of the two species using molecular typing methods such as PCR fingerprinting(14), AFLP analysis(15) and multi-locus sequencing(16) has demonstrated that both species contain genetically diverse lineages that qualify them to be considered as two species complexes, which have been further subdivided into numerous molecular types(14, 17). To date, four major lineages are recognised for *C. gattii*, which are denoted VGI/AFLP4, VGII/AFLP6, VGIII/AFLP5, and VGIV/AFLP7. Recently a fifth genotype was described on the basis of a single strain but with several different designations including Clade B (based on MLST), VGIIIc/VGIV, and *C. decagattii*(17, 18). It has been proposed to elevate these five lineages to separate species(17). However, such taxonomic treatment is currently controversial mainly due to the lack of clear biological differences between the lineages and no clear consensus on the limits and numbers of the putative species boundaries. As such, the various *C. gattii* lineages are collectively considered as the '*C. gattii* species complex' (18).

In this paper, we describe the discovery of a new lineage/molecular type within the *C. gattii* species complex, which we designate as VGV. The six VGV isolates were identified among 32 *C. gattii* isolates recovered from soil, animal dung and tree bark samples collected in Zambia by Vanhove *et al* in 2013 (19). In this paper, we characterize genomic and phenotypic features of the VGV molecular type. Additionally, we present a new improved genome assembly and gene-sets for *C. decagattii* (17) which we confirmed for the first time to be a separate lineage and therefore name as VGVI for consistency with the other lineages.

Results

<u>Comparative and population genomics for the six lineages of *C. gattii*</u>

We discovered a new lineage of *Cryptococcus gattii* from a panel of 32 (out of 55) genome sequenced isolates recovered from Southern tree hyrax (*Dendrohyrax arboreus*) middens, midden soil or tree holes from the Central Zambezian Miombo Woodland ecoregion, a densely forested ecoregion that covers much of Central and East Africa (**Fig. 1, Table 1, Table S1**). Isolates from the new lineage, which we have named VGV, were collected from a 430 km span of northern Zambia including the Mupata Hills (Luanshya, Copperbelt Province), Mutinondo wilderness area and Kapishya (Mpika, Northern Province), suggesting that the lineage has a broad regional distribution across this ecoregion (**Fig. 1a**). All VGV isolates were identified as serotype B, which also encompasses strains from VGI, VGII, the VGIIIa subgroup, and rare isolates among VGIV.

Phylogenetic analyses demonstrates that VGV, VGVI (*C. decagattii*), and the four previously described lineages, are genetically distinct from each other (**Fig. 2**). Indeed, none of the *C. gattii* lineages appear to be the result of hybridisation based on the distribution of private alleles (**Fig. 3a-b**), maximum likelihood phylogenetic reconstruction (**Fig. 2a**), F_{ST} (**Fig. S4**) or NeighborNet Network (**Fig. 2b**). Additional population genetics analyses confirmed low levels of genetic exchange between the six well resolved *C. gattii* lineages. For example, Principal Component Analysis (PCA) resolved distinct grouping for the lineages, with the first component (PC1) separating VGII from all other lineages, forming distinct clusters for VGIII and VGVI on PC2 (**Fig. 2c**). The projection of PC3 and PC4 further allows identication of distinct tight clusters for the VGI, VGIV and VGV lineages (**Fig. 2d**).

The new VGV lineage is represented by six isolates falling within two distinct subclades (A and B). Clade A comprises three VGV isolates (MF5, MF13, MF54) that were recovered from soil and animal dung sampled in hyrax middens, from which we also identified VGI and VGII isolates (**Fig. 1a**). Clade B comprises a further three VGV isolates: two that were recovered approximately 345 km away from Clade A (MF34 and MF51), and a third (MF56) that was recovered approximately 430 km away from the other Clade B isolates. Clade B isolates were recovered from both a tree hole and also hyrax middens, showing that the lineage can occupy both tree and dung, environments that are both associated with hyrax activity. The fact that Clade A and B were found in different geographic locations might reflect a degree of spatial genetic structure within VGV. All the VGV isolates were located in regions of granite and acidic kopjes/hills that are found occurring patchily across this ecoregion.

C. gattii VGV is highly diverged from all previously recognised *C. gattii* lineages. VGV isolates differ from VGII (reference isolate R265) by ~0.75 million SNPs on average (44 SNPs/Kb), and are thus similarly distant from VGII as the other lineages (**Table S1)**. The analysis of the relative proportion of shared private alleles for the *C. gattii* lineages (**Fig. 3a-b**), indicates VGII shared the fewest alleles with any of the other lineages, reflecting its greater divergence (<4.6 Kb total; <0.2 SNPs per Kb; **Fig. 2, Fig. 3a**). The newly discovered VGV shared fewer alleles with VGVI (0.24 per Kb) and VGIII (0.29 per Kb), than with VGI (1.18 per Kb) and VGIV (2.53 per Kb). The lineages that shared the most private alleles were VGVI and its closest relative VGIII (92 Kb total; 5.37 per Kb), which account for an average of 12% of all SNPs (based on alignments to VGII) found in isolates from those lineages.

Nearly one in ten nucleotides in the *C. gattii* genome has an alternative allele across the six lineages (1.55 x 10^6 sites; 9.01% of the *C. gattii* genome). Indeed, >180 Kb of these unique/private alleles were identified for each lineage, including for VGV which had 220 Kb private alleles (12.75 per Kb) (**Fig. 3b**). VGI is the most distinct in terms of both the highest count of private alleles (378 kb / 21.93 SNPs per Kb) and its nucleotide diversity (π) (**Fig. 3c**), which is reflected in the three distinct subclades of VGI isolates in the whole genome phylogeny (**Fig. 2a-b**). Conversely, the three VGVI isolates are thought to be derived from a single isolate recovered from a patient in Mexico and subsequently distributed to different labs where they have been renamed and sequenced(14, 17, 20, 21). Its few clonal differences are illustrated by its low nucleotide diversity (π) (**Fig. 3c**).

Unsupervised model-based clustering identified highly structured ancestry components enriched in each of the lineages. The clustering solution with the lowest cross-validation error (K=9) grouped the VGV isolates into a single genetically homogenous group (**Fig3d-e, Fig. S1**) whilst identifying four unique components within the VGII lineage. Of these, subclades VGIIx and VGIIb share small proportions of ancestry with other defined VGII subclades. For example, VGIIb is inferred to share ancestry with other VGII subclades (isolates Ram5 and B8554) and other lineages (B7394 has alleles from VGIV, and B7735 has alleles from VGV). Conversely, none of the isolates in VGIIa and VGIIc have demonstrable admixture with other subclades or lineages, both being formed by single unique ancestry components. VGIII isolate B8212 (a clinical isolate from Oregon, USA in 2007(22)) is also modelled as sharing ancestry with VGVI.

Finer-scale clustering was performed by considering patterns of genome-wide haplotype sharing in fineSTRUCTURE(23). Here, VGV isolates forming a separate cluster with greater haplotype similarity to isolates from VGI, VGIII and VGIV than VGII (**Fig. S2**). While haplotype sharing patterns were overwhelmingly in accordance with each lineage being genetically distinct, a notable exception was VGIII isolate B8212 that shares haplotypes with VGIV and VGVI (also in accordance with model-based clustering), perhaps owing to a small amount of genetic exchange with one or both of those lineages. As also observed using ADMIXTURE based clustering (**Fig. 3e**), two isolates from VGII, B7394 and B7735, were also genetically distinct and were assigned to their own cluster which was most closely related to isolates from subclade VGIIb.

All six VGV isolates were haploid with no evidence for aneuploidy based on allele-frequencies and depth of coverage (**Fig. S3**). However, we did observe copy number variation (CNV) between the three VGVI isolates derived from a single clinical isolate from Mexico(14, 17, 20, 21). Specifically, isolate CBS11687 acquired a ~200Kb duplication of supercontig (sc) 5 (position 1,040,000 through the end of the supercontig). Separately, isolate WM1804 had a smaller 40kb duplication on sc21 (positions 150,000 – 190,000). Isolate WM1802 had neither CNV. In terms of base changes, the three VGVI isolates (WM1802, WM1804 and CBS11687) differed by only 419 SNPs, with the fewest found between WM1804 and CBS11687 (n=126) and the most found between WM1802 and CBS11687 (n=315). These genetic differences may have occurred as a result of micro-evolution during or following passaging or cryo-preserving, although large CNVs are common in *C. gattii*(24, 25). All of the newly isolated VGI (n=7) and VGVA (n=3) samples from Zambia had a small <10 kb duplication within supercontig 6 of the R265 genome (position 400 kb to 410 kb). This genomic region encodes a single 87aa protein that is conserved in *C. neoformans* and *C. gattii*, but has no functional annotation (PFAM, GO-terms, KEGG-EC, TMHMM or SigP4).

The results from our phylogenetic and population genetic analyses are in line with previous work(26), indicating that lineages within the *C. gatti* species complex have remained largely genetically isolated since their divergence. Pairwise-lineage calculations of θ, Weir's formulation of Wright's fixation index (F_{ST}) suggest very low levels of genetic exchange between each of the lineages (**Fig. S4**) which is also reflected in analyses of genetic structure (**Fig. 2-3, Fig. S1-S2**). Both depth of coverage plots and F_{ST} non-overlapping sliding 10 Kb window plots across the mating type locus (*MAT*) at the start of supercontig 18 demonstrate that all VGIV and VGV isolates included in this study are *MAT*α (the reference genome of R265 is *MAT*α; high depth of coverage and θ > 0.98 across the *MAT* loci). In contrast, for VGI, VGII, VGIII and VGVI *MAT***a** isolates were included in our panel.

Genome assembly and analysis of VGV reveals *C. gattii* lineage-specific differences

We assembled and annotated a near complete genome for the newly discovered lineage *C. gattii* VGV (isolate MF34) using both Oxford Nanopore and Illumina sequencing reads. The resultant assembly consisted of 15 contigs corresponding to the 14 chromosomes; the single break in one chromosome corresponds to the ribosomal (rDNA) region. Other than under-representing rDNA genes, this assembly provides a complete representation of the genome, with telomeric repeats (TTAGGG) present at 28 contigs ends. Gene annotation revealed 6,322 predicted protein coding genes, which is similar to the seven other representative *C. gattii* isolates with publicly-available complete genomes(26, 27) representing the four previously known major lineages (ranging from 6,092 to 6,763), as well as *C. neoformans* H99(28) (*n*=6,962) (**Fig. S5-S6**).

To establish the evolution of protein coding genes in *C. gattii*, we compared the gene content for two representative annotated genomes per lineage where possible (no second annotated reference genomes were available for VGIV, VGV and VGVI), identifying 4,565 single copy core orthologs that are shared amongst the five lineages of *C. gattii* and *C. neoformans* (~74% of *Cryptococcus* genes) (**Table 2**). For VGVI, we sequenced and assembled the WM1802 isolate obtaining a similar genome length (17.42 Mb) and protein coding gene count (*n*=6,092). For VGII, we included the updated VGII R265 PacBio assembly in our panel of genomes(29) (**Table 2**). Orthology detection between just the two R265 assemblies identified 91% of genes in 1:1 orthology (*n*=5,642), ~4% of genes unique to the new assembly (*n*=252) and ~6% of genes in paralogous clusters (*n*=364). The previous VGII R265 assembly had 635 genes that were not called in the new assembly, likely a difference in the annotation protocol. Analysis of Core Eukaryotic Genes (CEGs) and BUSCO revealed a high completeness of gene-sets, and an increased completeness in the new annotation (**Fig. S6**). Furthermore, all assemblies generated using long read sequencing technology assemble into 14 scaffolds/supercontigs, suggesting all *Cryptococcus* lineages/species have conserved numbers of chromosomes.

Ortholog amino acid differences within and between lineages were consistent with results from our phylogenetic and population genetic analyses. VGV MF34 had the highest amino acid sequence similarity to VGIV IND107 (53,000 amino-acid differences = 97.55% similarity), which is observed in both alignment-based and ortholog-based phylogenies (**Fig. 2, Fig. 4**). The most similar inter-lineage orthologs were between VGIII and VGVI (49,500 predicted amino-acid differences = 97.71% similarity) (**Table S1**), while the most distinct pairwise comparisons were between *C. gattii* and *C. neoformans* (between 205,000 and 218,000 amino acid changes; ~90% protein similarity).

Overall, synteny is conserved within *C. gattii*(26), though with notable differences between some lineages. For example, VGV has a single 171 Kb inversion on supercontig 7 (positions 544,906-716,249) compared with the middle of VGIV IND107 supercontig 7 and the middle of VGIII CA1280 supercontig 5 (**Fig. 4**). VGVI also has some syntenic differences between its closest relative VGIII (**Fig. 2, Fig. 3, Fig. 4**). For example, approximately half of VGVI supercontig 5 is syntenic for the start of VGIII (CA1873) supercontig 16, while the second half of VGVI supercontig 5 is syntenic for a middle region of VGIII supercontig 1, indicative of a chromosomal translocation. Further improvements and additional genome assemblies should establish the full number and genetic impact of lineage-specific genomic rearrangements.

Lineage specific genes and multi-lineage specific genes (found in two or more lineages) were identified in each of the lineages (**Fig. S7, Fig. 3f**). Many of these lineage-specific genes (223/605; 37%) were previously identified from a panel of genome assemblies without the addition of VGV and VGVI(26). A further 53/605 (9%) of newly detected lineage-specific genes were previously categorised as multi-lineage-specific genes. Lineage specific genes in newly sequenced lineages (VGV and VGVI) include 74 genes that were unique to VGV and 49 genes that were uniquely absent in VGV. Genes unique to VGV include two sugar transporters (D1P53_002216, D1P53_002944) an alcohol dehydrogenase (D1P53_004471), and an aldehyde dehydrogenase (D1P53_006242). Conversely, eight transmembrane proteins and a single uncharacterised secreted protein were uniquely missing in VGV. All of the genes involved in the ergosterol biosynthesis pathway were present in single-copy in VGV. The VGVI WM1802 genome encodes 80 genes that are unique and 127 genes that are uniquely absent. Among the unique genes in WM1802, 14 are predicted to be involved in transport and include three monosaccharide transporters, one hexose transporter, one cadmium ion transporter and one monocarboxylic acid transporter.

Predictably, *C. neoformans* VNI H99 had the greatest number of lineage-specific/absent genes, with 578 unique genes and 28 absent genes. These included 47 genes predicted to be transmembrane proteins (including four sugar transporters, five MFS transporters, and a caffeine resistance transporter), and 34 secreted proteins. Fewer genes were uniquely absent in *C. neoformans*, which included the eukaryotic translation initiation factor 3 subunit B, an ACC oxidase, a copper amine, an allantoate permease of the major facilitator, and a 3-hydroxyacyl-dehydrogenase with oxidoreductase activity. Full details of all lineage-specific genes are provided in **Table S1**.

Phenotypic characteristics of VGV

All six isolates that belonged to the new VGV lineage based on whole genome sequencing (**Table 1**) were first incorrectly identified as VGIV based on the *URA5* restriction fragment length polymorphism (RFLP) banding pattern. Unlike most of VGIV, all six VGV isolates were serotype B. These were further characterised as *MAT*α, melanin and urease positive (**Fig. S8**) and grew well at 37^{o}C (**Fig. 5A**). However, VGV strains grew slightly more slowly on CGB agar (**Fig. S8**) than the serotype C VGIV strain (MF46) isolated from the same environment in the Central Miombo Woodland (**Table 1**). Hence, the positive blue/green colour development on CGB agar took longer in VGV than the control strains (**Fig. S8**). Since VGV is genetically closest to VGIV (**Fig. 2a-b**, **Fig. S2**), two Serotype C VGIV strains were used as control isolates for further phenotypic comparisons (WM779 isolated from a cheetah in South Africa(16), and MF46 isolated from Miombo tree bark in Zambia near Hyrax middens) (**Table 1**).

The size and morphology of VGV yeast cells were typical for *Cryptococcus* and indistinguishable from the control strains (**Fig. 5B**). Two distinct patterns of capsule formation were found among the six VGV isolates grown in YEPD broth (**Fig. 5B**). The isolates recovered from soil, Clade A (MF5, MF13, MF54), produced thinner capsule (≤ 1μm) compared to those recovered from tree bark, Clade B (MF34, MF51, MF56), which produced thick (2-4 μm) capsules similar to the VGIV control strains.

The VGV isolates and the control strains of VGIV manifested unusually high resistance toward fluconazole (FLC), particularly given they were sampled from an environmental niche. The three isolates of VGV clade A were more resistant to FLC, with minimal inhibitory concentration (MIC) ≥ 128 μg/ml than the three isolates in clade B which showed MICs of 24-64 μg/ml. All six VGV isolates showed MIC of 0.0625 μg/ml for 5-fluorocytocine (5-FC) similar to WM779. The MIC of MF46 for 5-FC was unusually high, 4μg/ml. The VGV MIC of amphotericin B ranged between 0.5 to 1μg/ml, higher than the control strains which had MIC below 0.5 μg/ml (**Fig. 5C**).

To explore the relative pathogenicity amongst VGV subclades we selected two isolates from Clade A (MF5, MF13) and two isolates from Clade B (MF34, MF51) for inoculation in mice models. Mice infected by all four isolates survived for 70 days while all the mice infected by WM779, a virulent serotype C control isolate, succumbed to infection within 30 days post infection (**Fig. 5D**). The VGIV environmental isolate MF46 (serotype C) caused no death in the mouse model. Fungal loads in the lungs of VGV infected mice were substantially lower than that of WM779 and slightly lower than those infected by MF46. Brain fungal loads of mice infected with the VGV strains were also low to negligible. The control isolates of VGIV showed little neurotropism (**Fig. 5E**). Histopathological analysis of the lungs demonstrated significant pathology in WM779 infected mice, with yeast found throughout the lung together with notable disruption of lung tissue. In many locations extensive leukocyte recruitment was evident in areas of concentrated infection (**Fig. 6**).

Histopathological analysis of the VGV isolates displayed substantially lower pulmonary yeast levels. That said, mice infected by MF46, MF34 and MF51 had higher levels of yeast than those infected by MF5 and MF13 in both CFU (**Fig. 5E**) and histopathological analysis (**Fig. 6**) in which MF51 was shown to represent Clade

B. The lung histopathology of the mice infected by MF34 was similar to that of MF51 (data not shown). Notable for its absence, leukocyte infiltration was mostly low or absent from sites of VGV infection. MF13 showed some leukocyte infiltration to a subset of infectious foci (**Fig. 6**).

Identification of VGV by *URA5* RFLP

The patterns of *URA5* RFLP, obtained by double digestion with *Sau*961 and *Hha*1, has been widely used to identify the lineage/molecular type in both *C. neoformans* and *C. gattii* species complexes(14). The *URA5* RFLP of Clade A isolates obtained by *Sau*96I/*Hha*I digest showed identical pattern with that of VGIV (**Fig. S9A**). Those of Clade B, however, produced an additional 1.3 kb amplicon which was absent from Clade A or any other VG molecular type reference isolates. This 1.3 kb amplicon was present even in uncut DNA of Clade B isolates (**Fig. S9A-B**), but its nature is not known at this juncture. Since the *URA5* RFLP patterns of both VGIV and VGV were not conclusively different, we compared the *URA5* gene sequence of the VGV isolate MF34 to that of the VGIV reference strain WM779 to identify possible restriction enzymes that can clearly distinguish the two lineages. This led to us identifying two highly discriminatory restriction enzymes; *Stu*I and *Ear*I. The expected sizes (bp) of the *URA5* gene fragment resulting from *Stu*I digestion are: 221 bp, 237 bp and 322 bp in VGIV and 237 bp and 543 bp in VGV. The *Ear*I digestions produced 247 and 501bp fragments in VGIV and 247 and 300 bp fragments in VGV. We compared the *URA5* RFLP of 17 VGIV isolates (**Table S1**) with the 6 VGV isolates by *Stu1* or *Ear*I digestion and the results are shown in **Fig. S9C-D**.

Type strain of VGV

We have designated MF34 (Clade B isolate) as the type strain of VGV which was isolated from a tree hole located in Mutinondo (latitude -12.45, longitude 31.29), Central Zambezian Miombo Woodlands (**Table 1**). Its genome has been assembled and annotated to near completion (15 scaffolds with N50=1.3Mb and telomeric repeats at 28 contigs ends). MF34 is serotype B and *MAT*α and causes mild pneumonia in C57BL/6 mice with neglible neurotropism. The genome assembly has been submitted to NCBI under the project accession PRJNA487802 and the culture has been deposited at the American Type Culture Collection (accession number pending).

Discussion

Over the past decade, increased sampling world-wide alongside whole-genome sequencing (WGS) methods have uncovered a greater genetic diversity of important pathogens including the *C. neoformans* and *C. gattii* species complexes. For example, sampling from Botswana revealed the existence of the *C. neoformans* VNB lineage (30), which itself has recently been shown to be deeply split into two genetically isolated lineages, VNBI and VNBII (31). Thus far, VGVI is the only lineage that exists as a single genotype since the three isolates previously designated as *C. decagattii* appear to have been isolated from the same patient (21). Each of the previously identified lineages of *Cryptococcus* have recently been designated as separate taxonomic species based on phylogenetic species recognition criteria (17). While we agree that *Cryptococcus* contains a number of genetically diverse and monophylectic clades that may be viewed as species under an Evolutionary Species Concept (32), we have previously argued that it is premature to give each clade a separate taxonomic name at this juncture (18, 33). One notable concern raised by Kwon-Chung *et al.* (18) was that the proposed seven-species taxonomy (33) was likely to be unstable due to incomplete knowledge of the true extent of *Cryptococcus* diversity worldwide. Our discovery of *C. gattii* VGV from the Miombo woodlands of Zambia clearly shows that we have not yet achieved a full understanding of the global biodiversity of *Cryptococcus*, and that further exploration will likely yield more phylogenetic species. Until we have a more accurate consensus on the true numbers of *Cryptococcus* lineages, we propose that the names 'VN' and 'VG' serve as a practical 'zip-code' within *C. neoformans* and *C. gattii*, offering a convenient way to describe newly discovered lineages or recombinants without introducing unwanted nomenclatural instability and confusion.

Our discovery of *C. gattii* VGV from hyrax-associated environments suggests an association with these mammals. Hyrax are small herbivores that are most closely related to elephants (Proboscidea) and sea cows (Sirenia), and are characterised by the behaviour of defecating in communal latrines, usually located in crevices in rocky kopjes, over many generations (34). These locations are often sheltered in rocky caves and droppings are likely to accumulate for upwards of 50,000 years, in some cases forming a stable paleoenvironmental hotspot of urea-rich nitrogenous material (35). *Cryptococcus* has a pronounced trophism for urea as a nutritive substrate, and pigeon guano is known to support prolific growth of *C. neoformans* and (to a lesser extent) *C. gattii* (36). Our finding that hyrax middens are hotspots of *Cryptococcus* diversity suggests that their ecological stability in landscapes that are low in nitrogen availability may lead to them being important arenas for the evolution of *Cryptococcus*, and will likely be fertile ground for further discovery of diversity within this genus.

Fungal association with small mammals may suggest adaptations that confer pathogenicity, known as the ‘endozoan, small-mammal reservoir hypothesis’ (37), and deserves to be explored further following our findings of an association of *Cryptococcus* with hyrax. Accordingly, alongside further study of potential mammalian reservoirs, the search for VGV clinical isolates is also needed in order to understand the true virulence potential of VGV and whether it can spillover into humans. Murine models have shown that environmental isolates are less virulent than clinical isolates of the same molecular type in both the *C. gattii* and *C. neoformans* species complexes suggesting that polymorphic virulence factors exist(38, 39). However, despite its large genetic distance from all other lineages, the new VGV lineage is not clearly distinguishable from others by existing methods such as serotyping or the routinely used *Sau*961 and *Hha*1 digested *URA5* RFLP analysis(14). Thus, it is possible that previous isolates belonging to VGV may have been misidentified using non-WGS methods. The most likely candidates for the search of clinical VGV are VGIV serotype B isolates recovered from patients. Geographically, the most likely place to find the VGV clinical isolates appear to be in sub-Saharan Africa since the current panel of isolates were found in the Zambian environment within an ecoregion that includes Tanzania, Burundi, Democratic Republic of the Congo, Angola and Malawi.

Previous work has shown that most isolates of the *C. gattii* species complex generally cause pulmonary infection in a murine model with low neurotropism(20, 40, 41). The four VGV isolates tested here were less neurotropic than the VGIV isolate MF46 that was collected from the same Zambian environment, and all the examined Zambian environmental isolates were significantly less virulent than a VGIV control strain, WM779. It remains to be shown if the differences in neurotropism are due to lineage-specific genes, or alleles in VGV. As previous work has shown in *C. neoformans* (42), capsule size difference manifested by Clade A and B *in vitro* was unrelated to virulence in mice

Although serotypes have not yet been conclusively linked to virulence in *Cryptococcus*, they remain important for strain identification. The majority of *C. gattii* tested to date are serotype B, except for a subset of VGIII and the majority of VGIV isolates which are serotype C. The six VGV isolates are also all serotype B - but due to the slower growth rate on CGB agar, the CGB reaction was weaker than other isolates of serotype B or serotype C. It took 24 hours longer for VGV compared to other VG isolates (VGI-VGIV) to turn the medium dark blue. As the six VGV isolates are all serotype B whilst the majority of VGIV isolates (their most closely related lineage) reported thus far have been serotype C, it is possible that VGV may also occur in serotype C. Additional environmental sampling of VGV is therefore necessary to establish the dominant serotype, since the current sample size of six is insufficient to make definitive conclusions.

Surprisingly, five of the six VGV isolates and the two control VGIV isolates were highly resistant to fluconazole (MIC of >64μg/ml), a commonly used anti-fungal drug. The three isolates of the VGV Clade A were more resistant to FLC than those of the VGV Clade B. Although the *C. gattii* species complex was previously known to be on average more resistant to FLC than *C. neoformans*(43), such high resistance to FLC in environmental isolates is notable and has not yet been reported(44). All of the VGV isolates had identical nucleotide sequences for *ERG11* and *AFR1*, demonstrating the resistant isolates are not a result of genetic differences in the target or transporter of FLC. However, innate fungal resistance to FLC can be due to multiple factors besides the *ERG11* gene or efflux pumps and the mechanism(s) of FLC resistance in VGV remain a subject for future investigation. Why environmental VGV isolates should have such high resistance to azoles is unclear as it is unlikely that they have come into contact with agrichemicals owing to the relatively pristine environments from which they were recovered. More likely, fluconazole resistance is a pleiotrophic effect that has evolved as a consequence of exposure by xenobiotics other than azoles. Further investigations into the evolution of FLC resistance in VGV may take on additional importance as clinical cases due to VGV are a distinct possibility in the Sub-Saharan regions where 12% of the Zambian population are living with the HIV virus(45).

In this paper, we present a near complete genome assembly for the VGV type strain, MF34. The MF34 genome allowed us to conclusively establish that VGV is a separate and distinct lineage of *C. gattii* from any previously identified, and not the result of hybridisation, as has been seen for other divergent isolates(31). Indeed, while both *Cryptococcus* species complexes appear to have a conserved chromosome number of 14 based on the current panels of assembled and annotated genomes available, intra- and inter-chromosomal rearrangements as well as large CNV's appear to be common. This chromosomal variation may provide the genetic basis for phenotypic variation and may act as a genetic barrier to recombination between more divergent isolates such as those from separate lineages. At the within-lineage level, there are also a number of unique and uniquely lost "lineage-specific genes', which may contribute to phenotypic differences between lineages. However, it should be noted that many of the main phenotypes routinely measured, including virulence in animal models, growth rates, and ability to cause pulmonary versus CNS infections, appear to vary as much within as between lineages.

One line of future inquiry towards explaining this phenotypic diversity may come from the characterisation of further transcriptional differences. For example, VGII upregulates many of the ergosterol genes during co-incubation with bone-marrow derived macrophages(46) and it will be important to determine whether other traits exist which differentiate the lineages of *Cryptococcus*. Further, it will be important to examine whether similar lineage-specific differences underpin VGV's increased FLC resistance, and whether clinically-relevant traits such as drug resistance are linked to the environment within which these isolates have evolved. Ultimately, our study testifies to the deep reservoir of diversity that exists within *Cryptococcus* which, despite decades of research into this genus, still harbours abundant surprises.

Acknowledgments

We thank Elizabeth Geigner and Scott Steelman for assistance with Oxford Nanopore sequencing of MF34 and José Muñoz for advice running SplitsTree. Michael Fisher, Mark and Mel Harvey, Jo and Charlie Harvey, Ian Bruce-Miller and Matthew Jones assisted with fieldwork. Shannon Moyer assisted with experimental animal work. MCF was supported by the UK MRC (MR/R015600/1) and a UK NERC studentship to MV. CAC was supported by NIAID Grant U19AI110818 to the Broad Institute. MCF and CAC are CIFAR fellow in the 'Fungal Kingdom' programme. MC, MJD, DHY, YCC and JK-C were supported by the Division of Intramural Research (DIR), NIAID, NIH. LvD and FB acknowledge financial support from the Newton Fund UK-China NSFC initiative (grants MR/P007597/1 and 8166113800).

Data access

The raw sequence and genome assembly of VGV MF34 is available in NCBI under BioProject PRJNA487802.

Author contributions

RAF, LvD, and TRS performed the genomic analyses.
RAF, MC, CAC, MCF, JK-C, HE, LvD, and FB wrote the manuscript.
MC, MJD, and DHY performed the phenotypic assays.
JK-C, YCC, WM and CAC sequenced the isolates.
TS and CAC assembled the VGV genome
MV, DC, GK, and MCF conducted the field work.

Figure and Table Legends

Fig. 1. Environmental sampling of *C. gattii* VGV in Zambia. A) Location of *C. gattii* VGV isolates across Central Zambian Miombo Woodlands. Isolates MF5, MF13 and 54 (Mupata Hills, Copperbelt province) and MF56 (Kapishya, Northern Province) were found in or near to Hyrax middens created by Southern tree hyrax

(*Dendrohyrax arboreus*). MF34 and MF51 were found from sampling Miombo tree holes in the Mutinondo wilderness area, Northern Province. **B)** A tree hyrax feeding on leaves. **C)** Sampling from hyrax middens at Kapishya from which MF56 was isolated.

Fig. 2. Whole genome analysis supports VGV as a distinct lineage. a) Maximum likelihood (RAxML) phylogeny of 101 *C. gattii* genomes generated over all non-ambiguous sites with at least a SNP in ≥1 isolate (1,518,323 sites, or 8.7% of the total genome). Isolate names are coloured according to lineage (dark blue = VGI, green = VGII, purple = VGIII, light blue = VGIV, orange = VGV, and red = VGVI). Asterisks indicate 100% bootstrap support at each node after 1,000 tree-building replicates. **b)** SplitsTree NeighborNet Network**. c-d)** Principal Component Analysis (PCA) of genomic variant sites, showing separation of isolates into lineages (isolates plotted with 2% random noise (jitter) for clarity of individual points). The first four Principal Components (PCs) account for 71.26% of the total genetic variation.

Fig. 3. Population genetic analyses for each of the *C. gattii* lineages (based on 101 genomes). a) Shared alleles (SNPs per Kb) between each of the *C. gattii* lineages. **b)** Private alleles (SNPs per Kb) between each of the *C. gattii* lineages. **c)** Nucleotide diversity (π) within each lineage against the number of isolates representing each lineage. **d)** Admixture K optimisation based on Cross Validation Error. **e)** Unsupervised ADMIXTURE clustering analysis of all isolates at K=9. **f)** Lineage specific gene and lineage specific gene-loss counts. The tree topology is based on the core-ortholog RAxML tree setting equal branch lengths, and the number of multi-lineage-specific gene gains and losses are shown above internal nodes.

Fig. 4. A phylogenetic tree for ten *Cryptococcus* genomes belonging to the six *C. gattii* lineages and one *C. neoformans* lineage outgroup alongside their genome synteny. The phylogenetic tree was constructed in RAxML with branch lengths indicating the mean number of nucleotide substitutions per site. To the right is a synteny plot, visualizing regions that span three or more orthologs between any two species as a connected grey line. Supercontig numbers are shown above each

genome axis if longer than 400 kb, where + represents the forward orientation and – represents the negative orientation.

Fig.5. Phenotypic characteristics. a) Growth of six VGV isolates and two VGIV control isolates at 30^{o}C and 37^{o}C on YEPD agar. **b)** India ink staining of VGV cells grown in YEPD broth for 24 hours at 30^{o}C. The isolates of Clade A (MF5, 13, 53) produce thinner capsules than the isolates of Clade B cells which produce a similar size capsule compared to the two VGIV control isolates. Bar = 5μm. **c)** MIC of VGV isolates for FLC, 5-FC and amphotericin B. All tested isolates had high MIC for FLC ranging from 20 to 256 µg/ml. The MIC for 5-FC was low except for MF46, an environmental isolate of VGIV which showed average 4μg/ml. All VGV isolates showed higher MIC for amphotericin B than VGIV controls. **d)** Survival curve of mice infected by four VGV isolates (intrapharyngeal aspiration of 5,000 cell/mouse) and two VGIV control isolates. Only the mice infected with VGIV isolate WM779 succumbed to infection. **e)** Lung and brain fungal loads. VGV isolates grew moderately in lungs but the CFU in the brains were negligible.

Fig. 6. Histopathology of the lung infected by VGV isolates. Sections of the mouse lungs infected by three different VGV isolates and two VGIV control strains stained by Alcian blue, Periodic acid-Schiff stain and counterstained with haematoxylin (left and middle columns) or with standard haematoxylin (right column). Note that Alcian blue stains cryptococcal cells blue. Images in the left columns were acquired using a 2.5X objective. Images in the middle and right columns are higher magnification (10X) images of the area indicated by the yellow boxes in the 2.5X images.

Table 1. Environmental isolates of VGIV and VGV from the Central Zambezian Miombo Woodlands.

Table 2. The genome assemblies used for phylogenetic analysis and orthology detection. *indicates newly described genome assemblies for this paper. All others have been described previously(26).

Supplemental Figure and Table Legends

Fig S1 Admixture analysis of 101 isolates based on K=2 through to K=15. The lowest Cross Validation Error was found at K=9.

Fig. S2 Chromopainter's inferred proportion of genome-wide DNA that each strain shares with every other based on pairwise haplotype matching profiles. The tree at the top provides fineSTRUCTURE's inferred hierarchical merging of clusters based on these profiles. Tick marks at the bottom are coloured according to lineage. VGV form their own cluster, tending to match more haplotypes genome-wide with isolates from VGI, VGIII and VGIV compared to VGII.

Fig. S3 Normalised read depth across 10 kb sliding windows along each supercontig relative to the R265 reference sequence. Aneuploid regions were only identified in two of the three *C. decagattii* (VGVI) isolates, which are highlighted by red circles. The lower depth of coverage across the start of supercontig 18 indicates a *MAT***a** isolate, compared with the R265 *MAT*α.

Fig. S4. Genome-wide variation in θ, Weir's formulation of Wright's fixation index (FST), on pairwise comparisons in each lineage. For comparison of isolates between each VG group, θ was calculated across window lengths of 10 kb. The lower F_{ST} at the start of supercontig 18 shows the location of the *MAT*α locus. Below the non-overlapping windows, mean pairwise F_{ST} values from all nuclear supercontigs are shown.

Fig. S5. The numbers of protein coding genes, rRNA, tRNA, genes with PFAMs, KEGG-ECs, GO-terms, predicted secreted genes (SignalP4) and transmembrane genes (TMHMM) for each assembly described in this paper.

Fig. S6. Coverage of the 248 Core Eukaryotic Genes (CEGs) by the *C. gattii* and *C. neoformans* gene-sets described in this paper.

Fig. S7. Synima/Orthofinder identified orthogroups (a) All orthogroups (representing every gene in each lineage) were grouped into a variety of categories including Orthologs 1:1 and Orthologs 1:>1 – which are orthogroups with 1:1 orthology in all isolates except for within a single isolate of the lineage that includes paralogs. "Orthologs divergent" and "Paralogs divergent" are genes that were previously lineage or isolate specific, but BLASTn revealed them to be unique orthogroups that, when joined to the database search genes, would make a 1:1 ortholog or Paralog conserved, respectively. Paralogs Lineage-specific (L.S.) and Paralogs Strain-specific (S.S.) are those genes that are represented by a single gene in all the other isolates in the category Orthologs 1:1 divergent. Paralogs miscellaneous (misc.) are all other Orthogroups including paralogous clusters. Genes absent in one lineage and lineage specific genes are broken down into further categories in **panel b**. Absent in one strain and present in one strain are self-explanatory, while Miscellaneous (Misc.) contains all remaining Orthogroups (such as genes found in multiple gene-sets but not lineage-specific or an ortholog. **b)** Details the number of genes in each lineage that are either absent in another lineage or specific to just that lineage.

Fig. S8. CGB reaction, melanin and urease production by VGV isolates. The melanin and urease production by VGV isolates were similar to the other VG type isolates but the CGB reaction took a longer time due to the slower growth on CGB agar.

Fig. S9. Patterns of *URA5* RFLP of VGIV and VGV isolates. **A)** The banding patterns of *URA5* uncut or *Sau*96I/*Hha*I digests are identical between VGIV and VGV Clade A isolates. However, Clade B isolates of VGV show 1.3 kb amplicon (red arrow) both in uncut as well as in *Sau*96I/*Hha*I digested DNA. This 1.3 kb amplicon is also absent in other molecular types. **B)** The uncut as well as *Stu*I and *Ear*I digests all show the 1.3 kb amplicon only in the Clade B isolates. **C)** The RFLP patterns of the 6 VGV isolates and 17 VGIV isolates digested by *Stu*I and **D)** *Ear*I showing clear difference between the two molecular types.

Table S1. (Tab 1) Details of all isolates used in this study. Isolates include those newly sequenced, and those presented in previous papers (with select citations included). Details of the alignments to R265 are given, along with the variants called.

(Tab 2) Pairwise comparisons of amino acid differences found among 1:1 core orthologs between all lineages of *C. gattii* and *C. neoformans* VNI H99. Each ortholog orthogroup has been aligned using MUSCLE and concatenated into a contiguous sequence used for phylogenetic reconstruction. **(Tab 3) Lineage specific genes for VGV, VGVI and VNI.** Details of all lineage-specific genes and genes uniquely absent in each of the lineages. Columns include unique orthogroup number, gene ID, GO-term annotation, length of gene, PFAMs, GO-terms, SigP4 predictions and TMHMM predictions. Genes that are uniquely absent in a lineage are represented by a separate lineage (VNI absent represented by VGV genes, VGVI (Cd) absent represented by VNI genes, VGV absent represented by VNI genes). **(Tab 4) Isolates of VGIV used to distinguish from VGV by *URA5* RFLP.**

Methods

Library preparation and sequencing of Zambian isolates

Environmental sampling took place in January and September of 2013. Samples were collected using “Transwab" Amies swabs (MWETM – MW170) and sterilized 30-mL screw- capped glass bottles. Amies liquid transport swabs were taken from tree bark (*n*=20), soil (*n*=19) and cracks in granite kopjes or droppings from rock Hyrax middens (*n*=16). Samples were collected and processed according to previously established protocols(47, 48), and the samples were kept at 4°C before being processed on niger seed agar. All samples were collected under license from the Zambian Wildlife Authority (ZAWA).

Single colonies purified from the original isolation media were maintained cryopreserved at -80°C at Imperial College in London since 2013. The isolates were revived on YPD agar (Yeast extract 1%, Peptone 1%, glucose 2%) and incubated at 30°C before use. Genomic DNA was isolated with CTAB extraction method as described previously with modification(49). Paired-end libraries (150 bp) were prepared and sequenced using the Illumina HiSeq 4000 platform by Novogene (Davis, CA). Two Oxford Nanopore libraries of isolate MF34 were constructed from genomic DNA using the 1D library construction kit (SQK-LSK109). A total of 243,660 reads with an N50 of 9,827 were generated on a FLO-MIN106 flow cell using a Minion. Reads were base called using Albacore v2.3.1. This resulted in 923,997,900 total bases (~46X coverage). Raw sequence data was submitted to the NCBI Sequence Read Archive under BioProject ID PRJNA476154 (all *C. gattii* non-VGV isolates) and PRJNA480403 (all *C. gattii* VGV isolates).

Genome assembly and annotation

For Isolate MF34, a hybrid assembly of Oxford Nanopore long-reads and Illumina short reads was generated. An initial assembly of the Oxford reads was generated using Canu v1.5(50) with parameter genomeSize=20,000,000. The assembly was inspected for the presence of telomeric repeat (TTAGGG) at contig ends; for two contig ends missing telomeric repeat, contigs were extended by aligning unassembled Canu contigs to these ends using NUCmer v3.1(51). Base called reads were then aligned to the contigs with BWA mem(52) with flag "-x ont2d", and the alignments used for polishing with Nanopolish(53). Two rounds of Pilon v1.13(54) correction were performed using Illumina BWA read alignments(52). Paired Illumina sequences of *C. decagattii* (VGVI) were assembled and scaffolded using SPAdes v3.1.1(55) with *k*-mer lengths (21, 33, 55 and 77). An assembly statistics summary for the assembly is provided in **Table 2**. Reads were aligned back to the assembly with BWA v0.7.4-r385 mem(52), and Pilon v1.12(54) was further used to improve the assembly. Scaffolds smaller than 1Kb were removed. The genome assembly has been submitted to NCBI under the project accession PRJNA487802.

The *C. gattii* VGV MF34 and VGVI WM1802 genomes were annotated using Genemark(56), BLASTx against SwissProt(57) and KEGG(58), and HMMER hmmscan(59) against PFAM(60). We ran tRNAscan(61) and RNAmmer(62) to identify non-protein coding genes. Gene predictions were checked for a variety of issues, including overlap with non-coding genes, overlap with coding genes, and the presence of in-frame stops. Genes were named according to evidence from BLASTx and HMMER following order of precedence: (1) SwissProt(57), (2) TIGRfam(63), and (3) KEGG(58), where BLASTx hits must meet the 70% identity and 70% overlap criteria to be considered a good hit and for the name to be applied. Otherwise, genes were named as hypothetical proteins.

Genes were functionally annotated by assigning PFAM domains(60), GO terms, KEGG assignment and ortholog mapping to genes of known function. HMMER3(59) was used to identify PFAM (release 27) domains, and BLASTx used against the KEGG v65 database(58) (e-value<1 × 10^{-10}). GO terms were assigned using Blast2GO version2.3.5(64), with a minimum e-value of 1 × 10^{-10}. SignalP 4.0(65) and TMHMM(66) were used to identify secreted proteins and trans-membrane proteins respectively (**Fig. S5**). Gene sets were aligned to the 248 Core Eukaryotic Genes (CEGs) and BUSCO basidiomycota_odb9 set to evaluate completeness (**Fig. S6**).

Read alignment and variant identification

The 36 newly sequenced isolates from this study were compared to an additional 65 isolates that were sequenced and described in previous studies(20, 26, 38, 67, 68). These additional isolates were obtained from the NCBI Sequence read archive (SRA) and converted from SRA format to FASTQ using SRAtoolkit version 2.3.3–4. Illumina reads were aligned to the *C. gattii* VGII R265 reference genome assembly using Burrows-Wheeler Aligner (BWA) v0.7.4-r385 mem(52) with default parameters and converted to sorted BAM format using SAMtools v0.1.9 (r783)(69).

Genome Analysis Toolkit (GATK) v2.7-4-g6f46d11(70) was used to call both variant and reference nucleotides from the 101 alignments (as previously described(24)). Briefly, the Picard tools AddOrReplaceReadGroups, MarkDuplicates, CreateSequenceDictionary, and ReorderSam were used to preprocess the alignments (http://broadinstitute.github.io/picard/). GATK RealignerTarget-Creator and IndelRealigner were then used to resolve misaligned reads close to indels. Next, GATK Unified Genotyper (with the haploid Genotyper ploidy setting) was run with both SNP and indel genotype likelihood models (GLM). We also ran Base Recalibrator and PrintReads for base quality score recalibration on those initial sites for GLM SNP. We then recalled variants with Unified Genotyper with the parameter "—output_mode EMIT_ALL_SITES." We merged and sorted all of the calls and then ran Variant Filtration with the parameters "QD < 2.0, FS > 60.0, MQ < 40.0." Next, we removed any base that had less than a minimum genotype quality of 50, a minimum percent alternate allele (AD) of 80%, or a minimum depth of 10. Finally, we removed any positions that were called by both GLMs (i.e., incompatible indels and

SNPs), any marked as “LowQual” by GATK, any nested indels, and any sites that did not include a PASS flag.

Phylogenetic and population genetic analysis

The variants identified from the 101 alignments were filtered for positions that were homozygous (reference or SNP) and polymorphic in one or more isolate (**Fig. 2**), resulting in an alignment of 1,517,353 nuclear sites and 970 mitochondrial sites. A FASTA file of these positions was created and converted into PHYLIP format, and a phylogenetic tree was generated using RAxML v7.7.8(71) with 1,000 bootstrap replications. RAxML was run with the generalized time-reversible (GTR) and category (CAT) rate approximation with final evaluation of the tree using GTR plus gamma-distributed rates. The same sites were analysed using the NeighborNet Network of SplitsTree v4.14.6(72).

A multi sample VCF of all 101 genomes was made with VCFtools(73) and converted to ped and map file formats for use in PLINK v1.90(74). Unsupervised ADMIXTURE(75) was run on a moderately Linkage Disequilibrium (LD) pruned alignment for values of K between 1-15. A value of K=9 provided the lowest cross-validation error (**Fig3d-e, Fig. S1**). To explore finer-patterns of population structure amongst our sampled lineages we applied a technique designed to characterise patterns of haplotype sharing between a panel of “donor” and “recipient” haplotypes within a recombining population. We ran Chromopainter v2(23) to infer, at each position in a recipient isolate’s genome, which donor they are most closely related to ancestrally relative to all others in the dataset. To do this, we assumed a uniform recombination rate of 1.5 morgans/megabase based on the genome wide recombination rate previously estimated in *C. neoformans*(76) and with Chromopainter’s switch and mutation rate parameters estimated using 10 runs of Expectation-Maximisation (-n 190.29, -M 0.0011). We then ran Chromopainter in linked mode using the haploid switch (-j) under an “all-versus-all” framework, painting all samples using all others to produce a pair-wise coancestry matrix describing the amount of DNA each isolate matches to every other under the copying model.

Haplotype based clustering was then implemented in fineSTRUCTURE(23) with an estimated normalization parameter of *c=0.51*, sampling cluster assignments every

10,000 iterations for 1x10^6 MCMC iterations after 1x10^6 initial burn-in steps. We then performed an additional 1x10^5 hill-climbing iterations beginning with the MCMC sample with the highest posterior probability. This classified our data into 34 clusters (**Fig. S2**).

For the *C. neoformans* VNI H99 rooted *C. gattii* tree, we identified 1:1 orthologs among each of the nine isolates with Orthofinder v2.1.2(77) using the Synima pipeline(78). We aligned orthologs with MUSCLE v3.8.31(79), extracted the CDS sequences in a codon context, and trimmed to the smallest contiguous sequence, and then concatenated alignments. In total, we aligned 2.16 Mb of transcripts for each genome. Prottest v3.4(80) was used to determine the best-fitting amino acid transition model (JTT) according to Bayesian information criterion. The final tree was produced using RAxML v7.7.8(71) using the CAT rate approximation and WAG amino acid replacement matrix with 1,000 bootstrap replicates. Synima(78) was used to visualise synteny between each of the genomes. The same pipeline was used to compare the previous and updated R265 genomes.

Phenotypic analysis

To determine the growth rate of *C. gattii* VGV, cells of all six VGV isolates were inoculated in YEPD broth and incubated at 30°C on a shaker (200rpm) for 18h. Cells were washed with sterile PBS and 2X10^5 cells/ml were resuspended in PBS. Three microliter aliquots of 10-fold serial dilutions were spotted onto YEPD agar and incubated at 30°C and 37°C. For biological confirmation of the species, isolates were inoculated on Canavanine glycine bromothymol blue (CGB) agar(81) for species specific CGB reaction and Christensen's urea agar (Sigma) and norepinephrine agar(82) for urease and melanin production respectively and incubated at 30^{o}C for 48 hours. India ink mount of the cells grown on YEPD broth for 24h at 30^{o}C were used for microscopic observation of the cell and polysaccharide capsule size. The reference strains used were WM148 or H99 (serotype A, VNI), WM626 (serotype A, VNII), WM179 (serotype B, VGI), WM178, R265 and R272 (serotype B, VGII), WM161 (serotype B, VGIII), and WM779 (serotype C, VGIV)(14). Mating type of each isolate was determined by PCR using primers specific to the *STE12*α and *STE20*a(83).

Determination of MIC for antifungal antibiotics

MICs for fluconazole (FLC), 5-fluorocytosine (5FC), and Amphotericin B were determined using Etest strips according to the Etest technical guide (AB Biodisk, Solna, Sweden), with slight modification. Fungal cells were grown in 5 ml of YEPD at 30°C for 18 hours. Harvested cells were diluted in sterile saline to an optical density of 0.05 at 600 nm (OD600) and plated on yeast nitrogen base (YNB) agar plates. Etest strips were placed at the center of the plates and incubated at 30°C for 72 hours. The susceptibility endpoint was read at the first growth inhibition ellipse. The concentration ranges tested were: FLC, 0.016 to 256 μg/ml; both 5-FC and Amphotericin B, 0.002 to 32 μg/ml.

URA5 gene RFLP

The *URA5* gene of each isolate was amplified from genomic DNA by PCR to identify the molecular type using 50 ng of two primers: URA5 (5′-ATGTCCTCCCAAGCCCTCGACTCCG-3′) and SJ01 (5′-TTAAG ACCTCTGAACACCGTACTC-3′). Reactions were carried out in a total volume of 50 μL as previously described(14). PCR was performed for 40 cycles at 94°C for 2 min initial denaturation, 30 s of denaturation at 94°C, 30 s annealing at 55°C, and 2 min extension at 72°C. The reactions were completed by a final extension step for 10 min at 72°C. PCR products were analysed by 1% agarose gel electrophoresis and 5 μL of PCR products were double digested with *Sau*96I (10 U/μL) and *Hha*I (20 U/μL) for 3 h at 37°C. Then, digested samples were separated by 3% agarose gel electrophoresis at 80V for 5 h. The RFLP patterns of *URA5* gene were analysed using well-defined bands in the gel images by comparing them with the patterns obtained from the standard reference strains.

Restriction enzyme analysis of the *URA5* gene to distinguish VGV from VGIV

We found the *URA5* RFLP banding patterns(14) of VGV and VGIV are not clearly distinguishable although *Sau*96I and *Hha*I (**Fig. S9A**). We compared the DNA sequences of the *URA5* gene from MF34 (VGV) and WM779 (VGIV) and found two restriction enzymes, *Stu*I and *Ear*I that can be used to distinguish the two molecular types based on *URA5* RFLP. Three microliters of *URA5* PCR products were digested with *Stu*I (10 U/μl) or *Ear*I (20U/μl) (New England BioLabs Inc) at 37^{o}C for

4h and restriction fragments were separated by electrophoresis in 3% agarose Tris-acetate-EDTA (TAE) gels at 80V for 5h. Standard reference strains for molecular typing were used as controls.

Virulence in mice

The virulence of four VGV isolates, two from Clade A and two from Clade B, was assessed using seven to eight weeks old female C57BL/6 mice (Taconic Farms). Isolates to be tested in mice were inoculated in YEPD broth and incubated overnight, washed twice and diluted to 2.5X10^5 cells/ml in PBS. Mice (14 mice per isolate) were inoculated with 20μl of cell suspension (5x10^3/mouse) by pharyngeal aspiration. Eight mice for each isolate were used for the survival rate and six mice each were used for the analysis of fungal burden and histopathology at the indicated time points. Mice were monitored twice per day and differences in survival were determined using GraphPad Prism, version 7 (GraphPad Software, San Diego, CA).

To assess the organ fungal burden, lungs and brains of four mice from each infected group were inspected. The mice infected with WM779 started to die on day 25 post infection and the lungs and brains were harvested immediately from the dead mice on day 25. Mice infected with other isolates were euthanized on day 60 and organs were harvested. Harvested lungs and brains were homogenized in 7ml and 2ml sterile water respectively and 5 μl aliquots of 10-fold serial dilutions were plated on YEPD agar and incubated at 30^{o}C for 48h. Fungal colonies were counted and the tissue fungal burden was analysed using GraphPad Prism, version 7 (GraphPad Software, San Diego, CA).

Histopathological analysis

For histopathological analysis, organs of infected mice from each group were harvested at 10 and 20 days post inoculation and fixed in 3.7% buffered formalin and embedded in paraffin. Sections were stained with hematoxylin and eosine (H&E) or Alcian blue/periodic acid-Schiff (AB/PAS) at the Histoserv Inc.

Ethics statement

The Institutional Animal Care and Use Committee of the National Institute of Allergy and Infectious Diseases approved all animal studies (#LCIM-5E). Studies were performed in accordance with recommendations of the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health.

References

1. Rajasingham R, Smith RM, Park BJ, Jarvis JN, Govender NP, Chiller TM, Denning DW, Loyse A, Boulware DR. 2017. Global burden of disease of HIV-associated cryptococcal meningitis: an updated analysis. Lancet Infect Dis 17:873–881.

2. Busse O. 1894. Uber parasitare Zelleinschlusse und ihre Zuchtung. Cent Bakt Parasit 16.

3. Barnett JA. 2010. A history of research on yeasts 14: medical yeasts part 2, *Cryptococcus neoformans*. Yeast Chichester Engl 27:875–904.

4. Kwon-Chung KJ, Fraser JA, Doering TL, Wang Z, Janbon G, Idnurm A, Bahn Y-S. 2014. *Cryptococcus neoformans* and *Cryptococcus gattii*, the etiologic agents of cryptococcosis. Cold Spring Harb Perspect Med 4:a019760.

5. Evans EE. 1950. The Antigenic composition of *Cryptococcus neoformans*: I. A serologic classification by means of the capsular and agglutination reactions. J Immunol 64:423–430.

6. Wilson DE, Bennett JE, Bailey JW. 1968. Serologic grouping of *Cryptococcus neoformans*. Proc Soc Exp Biol Med 127:820–823.

7. Kwon-Chung KJ. 1975. A new genus, filobasidiella, the perfect state of *Cryptococcus neoformans*. Mycologia 67:1197–1200.

8. Kwon-Chung KJ. 1976. A new species of Filobasidiella, the sexual state of *Cryptococcus neoformans* B and C serotypes. Mycologia 68:943–946.

9. Franzot SP, Salkin IF, Casadevall A. 1999. *Cryptococcus neoformans var. grubii*: separate varietal status for *Cryptococcus neoformans* serotype A isolates. J Clin Microbiol 37:838–840.

10. Xu J, Vilgalys R, Mitchell TG. 2000. Multiple gene genealogies reveal recent dispersion and hybridization in the human pathogenic fungus *Cryptococcus neoformans*. Mol Ecol 9:1471–1481.

11. Diaz MR, Boekhout T, Kiesling T, Fell JW. 2005. Comparative analysis of the intergenic spacer regions and population structure of the species complex of the pathogenic yeast *Cryptococcus neoformans*. FEMS Yeast Res 5:1129–1140.

12. Kwon-Chung KJ, Boekhout T, Fell JW, Diaz M. 2002. (1557) Proposal to Conserve the Name *Cryptococcus gattii* against *C. hondurianus* and *C. bacillisporus* (Basidiomycota, Hymenomycetes, Tremellomycetidae). Taxon 51:804–806.

13. Kwon-Chung KJ, Varma A. 2006. Do major species concepts support one, two or more species within *Cryptococcus neoformans*? FEMS Yeast Res 6:574–587.

14. Meyer W, Castañeda A, Jackson S, Huynh M, Castañeda E, IberoAmerican Cryptococcal Study Group. 2003. Molecular typing of IberoAmerican *Cryptococcus neoformans* isolates. Emerg Infect Dis 9:189–195.

15. Boekhout T, Theelen B, Diaz M, Fell JW, Hop WC, Abeln EC, Dromer F, Meyer W. 2001. Hybrid genotypes in the pathogenic yeast *Cryptococcus neoformans*. Microbiol Read Engl 147:891–907.

16. Meyer W, Aanensen DM, Boekhout T, Cogliati M, Diaz MR, Esposto MC, Fisher M, Gilgado F, Hagen F, Kaocharoen S, Litvintseva AP, Mitchell TG, Simwami SP, Trilles L, Viviani MA, Kwon-Chung J. 2009. Consensus multi-locus sequence typing scheme for *Cryptococcus neoformans* and *Cryptococcus gattii*. Med Mycol 47:561–570.

17. Hagen F, Khayhan K, Theelen B, Kolecka A, Polacheck I, Sionov E, Falk R, Parnmen S, Lumbsch HT, Boekhout T. 2015. Recognition of seven species in the *Cryptococcus gattii/Cryptococcus neoformans* species complex. Fungal Genet Biol FG B 78:16–48.

18. Kwon-Chung KJ, Bennett JE, Wickes BL, Meyer W, Cuomo CA, Wollenburg KR, Bicanic TA, Castañeda E, Chang YC, Chen J, Cogliati M, Dromer F, Ellis D, Filler SG, Fisher MC, Harrison TS, Holland SM, Kohno S, Kronstad JW, Lazera M, Levitz SM, Lionakis MS, May RC, Ngamskulrongroj P, Pappas PG, Perfect JR, Rickerts V, Sorrell TC, Walsh TJ, Williamson PR, Xu J, Zelazny AM, Casadevall A. 2017. The case for adopting the “Species Complex” nomenclature for the etiologic agents of cryptococcosis. mSphere 2.

19. Vanhove M, Beale MA, Rhodes J, Chanda D, Lakhi S, Kwenda G, Molloy S, Karunaharan N, Stone N, Harrison TS, Bicanic T, Fisher MC. 2017. Genomic epidemiology of Cryptococcus yeasts identifies adaptation to environmental niches underpinning infection across an African HIV/AIDS cohort. Mol Ecol 26:1991–2005.

20. Firacative C, Roe CC, Malik R, Ferreira-Paim K, Escandón P, Sykes JE, Castañón-Olivares LR, Contreras-Peres C, Samayoa B, Sorrell TC, Castañeda E, Lockhart SR, Engelthaler DM, Meyer W. 2016. MLST and whole-genome-based population analysis of *Cryptococcus gattii* VGIII links clinical, veterinary and environmental strains, and reveals divergent serotype specific sub-populations and distant ancestors. PLoS Negl Trop Dis 10:e0004861.

21. Hagen F, Illnait-Zaragozí M-T, Meis JF, Chew WHM, Curfs-Breuker I, Mouton JW, Hoepelman AIM, Spanjaard L, Verweij PE, Kampinga GA, Kuijper EJ, Boekhout T, Klaassen CHW. 2012. Extensive genetic diversity within the dutch clinical *Cryptococcus neoformans* population. J Clin Microbiol 50:1918–1926.

22. Gillece JD, Schupp JM, Balajee SA, Harris J, Pearson T, Yan Y, Keim P, DeBess E, Marsden-Haug N, Wohrle R, Engelthaler DM, Lockhart SR. 2011. Whole genome sequence analysis of *Cryptococcus gattii* from the Pacific Northwest reveals unexpected diversity. PloS One 6:e28550.

23. Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. PLOS Genet 8:e1002453.

24. Chen Y, Farrer RA, Giamberardino C, Sakthikumar S, Jones A, Yang T, Tenor JL, Wagih O, Wyk MV, Govender NP, Mitchell TG, Litvintseva AP, Cuomo CA, Perfect JR. 2017. Microevolution of serial clinical isolates of *Cryptococcus neoformans var. grubii* and *C. gattii*. mBio 8:e00166-17.

25. Steenwyk JL, Soghigian JS, Perfect JR, Gibbons JG. 2016. Copy number variation contributes to cryptic genetic variation in outbreak lineages of *Cryptococcus gattii* from the North American Pacific Northwest. BMC Genomics 17:700.

26. Farrer RA, Desjardins CA, Sakthikumar S, Gujja S, Saif S, Zeng Q, Chen Y, Voelz K, Heitman J, May RC, Fisher MC, Cuomo CA. 2015. Genome evolution and innovation across the four major lineages of *Cryptococcus gattii*. mBio 6:e00868-00815.

27. D'Souza CA, Kronstad JW, Taylor G, Warren R, Yuen M, Hu G, Jung WH, Sham A, Kidd SE, Tangen K, Lee N, Zeilmaker T, Sawkins J, McVicker G, Shah S, Gnerre S, Griggs A, Zeng Q, Bartlett K, Li W, Wang X, Heitman J, Stajich JE, Fraser JA, Meyer

W, Carter D, Schein J, Krzywinski M, Kwon-Chung KJ, Varma A, Wang J, Brunham R, Fyfe M, Ouellette BFF, Siddiqui A, Marra M, Jones S, Holt R, Birren BW, Galagan JE, Cuomo CA. 2011. Genome variation in *Cryptococcus gattii*, an emerging pathogen of immunocompetent hosts. mBio 2:e00342-00310.

28. Janbon G, Ormerod KL, Paulet D, Iii EJB, Yadav V, Chatterjee G, Mullapudi N, Hon C-C, Billmyre RB, Brunel F, Bahn Y-S, Chen W, Chen Y, Chow EWL, Coppée J-Y, Floyd-Averette A, Gaillardin C, Gerik KJ, Goldberg J, Gonzalez-Hilarion S, Gujja S, Hamlin JL, Hsueh Y-P, Ianiri G, Jones S, Kodira CD, Kozubowski L, Lam W, Marra M, Mesner LD, Mieczkowski PA, Moyrand F, Nielsen K, Proux C, Rossignol T, Schein JE, Sun S, Wollschlaeger C, Wood IA, Zeng Q, Neuvéglise C, Newlon CS, Perfect JR, Lodge JK, Idnurm A, Stajich JE, Kronstad JW, Sanyal K, Heitman J, Fraser JA, Cuomo CA, Dietrich FS. 2014. Analysis of the genome and transcriptome of *Cryptococcus neoformans var. grubii* reveals complex RNA expression and microevolution leading to virulence attenuation. PLOS Genet 10:e1004261.

29. Yadav V, Sun S, Billmyre RB, Thimmappa BC, Shea T, Lintner R, Bakkeren G, Cuomo CA, Heitman J, Sanyal K. 2018. RNAi is a critical determinant of centromere evolution in closely related fungi. Proc Natl Acad Sci U S A 115:3108–3113.

30. Litvintseva AP, Thakur R, Vilgalys R, Mitchell TG. 2006. Multilocus sequence typing reveals three genetic subpopulations of *Cryptococcus neoformans var. grubii* (serotype A), including a unique population in Botswana. Genetics 172:2223–2238.

31. Desjardins CA, Giamberardino C, Sykes SM, Yu C-H, Tenor JL, Chen Y, Yang T, Jones AM, Sun S, Haverkamp MR, Heitman J, Litvintseva AP, Perfect JR, Cuomo CA. 2017. Population genomics and the evolution of virulence in the fungal pathogen *Cryptococcus neoformans*. Genome Res 27:1207–1219.

32. Wiley EO. 1978. The evolutionary species concept reconsidered. Syst Zool 27:17–26.

33. Hagen F, Lumbsch HT, Arsenijevic VA, Badali H, Bertout S, Billmyre RB, Bragulat MR, Cabañes FJ, Carbia M, Chakrabarti A, Chaturvedi S, Chaturvedi V, Chen M, Chowdhary A, Colom M-F, Cornely OA, Crous PW, Cuétara MS, Diaz MR, Espinel-Ingroff A, Fakhim H, Falk R, Fang W, Herkert PF, Rodríguez CF, Fraser JA, Gené J, Guarro J, Idnurm A, Illnait-Zaragozi M-T, Khan Z, Khayhan K, Kolecka A, Kurtzman CP, Lagrou K, Liao W, Linares C, Meis JF, Nielsen K, Nyazika TK, Pan W, Pekmezovic M, Polacheck I, Posteraro B, Telles F de Q, Romeo O, Sánchez M, Sampaio A, Sanguinetti M, Sriburee P, Sugita T, Taj-Aldeen SJ, Takashima M, Taylor JW, Theelen B, Tomazin R, Verweij PE, Wahyuningsih R, Wang P, Boekhout T. 2017. Importance of resolving fungal nomenclature: the case of multiple pathogenic species in the *Cryptococcus* genus. mSphere 2:e00238-17.

34. Scott L. 1990. Hyrax (Procaviidae) and dassie rat (Petromuridae) middens in palaeoenvironmental studies in Africa, p. 408–427. *In* Packrat Middens: The Last 40,000 Years of Biotic Change. University of Arizona Press, Tucson.

35. Chase BM, Scott L, Meadows ME, Gil-Romera G, Boom A, Carr AS, Reimer PJ, Truc L, Valsecchi V, Quick LJ. 2012. Rock hyrax middens: A palaeoenvironmental archive for southern African drylands. Quat Sci Rev 56:107–125.

36. Nielsen K, Obaldia ALD, Heitman J. 2007. *Cryptococcus neoformans* mates on Pigeon guano: Implications for the realized ecological niche and globalization. Eukaryot Cell 6:949–959.

37. Taylor JW, Barker BM. 2019. The endozoan, small-mammal reservoir hypothesis and the life cycle of Coccidioides species. Med Mycol 57:S16–S20.

38. Springer DJ, Billmyre RB, Filler EE, Voelz K, Pursall R, Mieczkowski PA, Larsen RA, Dietrich FS, May RC, Filler SG, Heitman J. 2014. *Cryptococcus gattii* VGIII isolates causing infections in HIV/AIDS patients in Southern California: identification of the local environmental source as Arboreal. PLoS Pathog 10:https://doi.org/10.1371/journal.ppat.1004285.

39. Litvintseva AP, Mitchell TG. 2009. Most environmental isolates of *Cryptococcus neoformans var. grubii* (serotype A) are not lethal for mice. Infect Immun 77:3188–3195.

40. Ngamskulrungroj P, Chang Y, Sionov E, Kwon-Chung KJ. 2012. The primary target organ of *Cryptococcus gattii* is different from that of *Cryptococcus neoformans* in a murine model. mBio 3:e00103-12.

41. Davis MJ, Moyer S, Hoke ES, Sionov E, Mayer-Barber KD, Barber DL, Cai H, Jenkins L, Walter PJ, Chang YC, Kwon-Chung KJ. 2019. Pulmonary iron limitation induced by exogenous type I IFN protects mice from *Cryptococcus gattii* independently of T Cells. mBio 10:e00799-19.

42. Dykstra MA, Friedman L, J W Murphy. 1977. Capsule size of *Cryptococcus neoformans*: control and relationship to virulence. Infect Immun 16:129–135.

43. Gomez-Lopez A, Zaragoza O, Dos Anjos Martins M, Melhem MC, Rodriguez-Tudela JL, Cuenca-Estrella M. 2008. *In vitro* susceptibility of *Cryptococcus gattii* clinical isolates. Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis 14:727–730.

44. Khan ZU, Randhawa HS, Kowshik T, Chowdhary A, Chandy R. 2007. Antifungal susceptibility of *Cryptococcus neoformans* and *Cryptococcus gattii* isolates from

decayed wood of trunk hollows of *Ficus religiosa* and *Syzygium cumini* trees in north-western India. J Antimicrob Chemother 60:312–316.

45. https://www.unaids.org/en/regionscountries/countries/zambia.

46. Farrer RA, Ford CB, Rhodes J, Delorey T, May RC, Fisher MC, Cloutman-Green E, Balloux F, Cuomo CA. 2018. Transcriptional heterogeneity of *Cryptococcus gattii* VGII compared with non-VGII lineages underpins key pathogenicity pathways. mSphere 3.

47. Litvintseva AP, Carbone I, Rossouw J, Thakur R, Govender NP, Mitchell TG. 2011. Evidence that the human pathogenic fungus *Cryptococcus neoformans var. grubii* may have evolved in Africa. PloS One 6:e19688.

48. Randhawa HS, Kowshik T, Khan ZU. 2005. Efficacy of swabbing versus a conventional technique for isolation of *Cryptococcus neoformans* from decayed wood in tree trunk hollows. Med Mycol 43:67–71.

49. Fujimura H, Sakuma Y. 1993. Simplified isolation of chromosomal and plasmid DNA from yeasts. BioTechniques 14:538–540.

50. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27:722–736.

51. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. 1999. Alignment of whole genomes. Nucleic Acids Res 27:2369–2376.

52. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio.

53. Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. Nat Methods 12:733–735.

54. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One 9:e112963.

55. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol J Comput Mol Cell Biol 19:455–477.

56. Lukashin AV, Borodovsky M. 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26:1107–1115.

57. Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45–48.

58. Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30.

59. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–W37.

60. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. 2014. Pfam: the protein families database. Nucleic Acids Res 42:D222–D230.

61. Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964.

62. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100–3108.

63. Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res 31:371–373.

64. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinforma Oxf Engl 21:3674–3676.

65. Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785–786.

66. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a Hidden Markov Model: application to complete genomes. J Mol Biol 305:567–580.

67. Farrer RA, Voelz K, Henk DA, Johnston SA, Fisher MC, May RC, Cuomo CA. 2016. Microevolutionary traits and comparative population genomics of the emerging pathogenic fungus *Cryptococcus gattii*. Phil Trans R Soc B 371:20160021.

68. Engelthaler DM, Hicks ND, Gillece JD, Roe CC, Schupp JM, Driebe EM, Gilgado F, Carriconde F, Trilles L, Firacative C, Ngamskulrungroj P, Castañeda E, Lazera M dos S, Melhem MSC, Pérez-Bercoff Å, Huttley G, Sorrell TC, Voelz K, May RC, Fisher MC, Thompson GR, Lockhart SR, Keim P, Meyer W. 2014. *Cryptococcus gattii* in

North American Pacific Northwest: whole-population genome analysis provides insights into species evolution and dispersal. mBio 5:e01464-14.

69. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinforma Oxf Engl 25:2078–2079.

70. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303.

71. Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

72. Huson DH. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinforma Oxf Engl 14:68–73.

73. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

74. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575.

75. Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664.

76. Roth C, Sun S, Billmyre RB, Heitman J, Magwene PM. 2018. A high-resolution map of meiotic recombination in *Cryptococcus deneoformans* demonstrates decreased recombination in unisexual reproduction. Genetics 209:567–578.

77. Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157.

78. Farrer RA. 2017. Synima: a synteny imaging tool for annotated genome assemblies. BMC Bioinformatics 18:507.

79. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

80. Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165.

81. Kwon-Chung KJ, Polacheck I, Bennett JE. 1982. Improved diagnostic medium for separation of *Cryptococcus neoformans var. neoformans* (serotypes A and D) and *Cryptococcus neoformans var. gattii* (serotypes B and C). J Clin Microbiol 15:535–537.

82. Kwon-Chung KJ, Polacheck I, Popkin TJ. 1982. Melanin-lacking mutants of *Cryptococcus neoformans* and their virulence for mice. J Bacteriol 150:1414–1421.

83. Halliday CL, Bui T, Krockenberger M, Malik R, Ellis DH, Carter DA. 1999. Presence of alpha and a mating types in environmental and clinical collections of *Cryptococcus neoformans var. gattii* strains from Australia. J Clin Microbiol 37:2920–2926.