Copy number variation contributes to cryptic genetic variation in outbreak lineages of Cryptococcus gattii from the North American Pacific Northwest

Jacob L. Steenwyk, John S. Soghigian, John R. Perfect and John G. Gibbons

Abstract

Background: Copy number variants (CNVs) are a class of structural variants (SVs) and are defined as fragments of DNA that are present at variable copy number in comparison with a reference genome. Recent advances in bioinformatics methodologies and sequencing technologies have enabled the high-resolution quantification of genome-wide CNVs. In pathogenic fungi SVs have been shown to alter gene expression, influence host specificity, and drive fungicide resistance, but little attention has focused specifically on CNVs. Using publicly available sequencing data, we identified 90 isolates across 212 Cryptococcus gattii genomes that belong to the VGII subgroups responsible for the recent deadly outbreaks in the North American Pacific Northwest. We generated CNV profiles for each sample to investigate the prevalence and function of CNV in C. gattii.

Results: We identified eight genetic clusters among publicly available Illumina whole genome sequence data from 212 C. gattii isolates through population structure analysis. Three clusters represent the VGIIa, VGIIb, and VGIIc subgroups from the North American Pacific Northwest. CNV was bioinformatically predicted and affected ~300–400 Kilobases (Kb) of the C. gattii VGII subgroup genomes. Sixty-seven loci, encompassing 58 genes, showed highly divergent patterns of copy number variation between VGII subgroups. Analysis of Pfam domains within divergent CN variable genes revealed enrichment of protein domains associated with transport, cell wall organization and external encapsulating structure.

Conclusions: CNVs may contribute to pathological and phenotypic differences observed between the C. gattii VGIIa, VGIIb, and VGIIc subpopulations. Genes overlapping with population differentiated CNVs were enriched for several virulence related functional terms. These results uncover novel candidate genes to examine the genetic and functional underpinnings of C. gattii pathogenicity.

Keywords: Cryptococcus gattii, Copy number variation, Pathogen, Pathogenicity, Population genomics

Abbreviations: BIC, Bayesian information criterion; bp, Base pairs; CN, Copy number; CNV, Copy number variation; CV, Structural variation; DAPC, Discriminant analysis of principal components; GO, Gene ontology; Kb, Kilobases; LASSO, Least absolute shrinkage and selection operator; MCMC, Markov Chain Monte Carlo; MFS, Major facilitator superfamily; NMRV, Normalized mapped read value; nt, Nucleotides; PNW, Pacific Northwest; RBBH, Reciprocal best BLAST hit; SNPs, Single nucleotide polymorphisms

© 2016 The Author(s). Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
Background

Copy number variants (CNVs) are a class of structural variants (SVs) and are defined as fragments of DNA, typically larger than 1 Kilobase (Kb), that are present at variable copy number in comparison with a reference genome [1]. Mutation rates of CNVs are typically higher than those of single nucleotide polymorphisms (SNPs) [2, 3] and several mechanisms have been proposed to explain the formation of CNVs, including non-allelic homologous recombination, non-homologous end-joining, retrotransposition, and fork stalling and template switching [4–7]. Recent developments in sequencing technologies and bioinformatics tools have made it possible to more accurately identify and quantify CNVs, revealing their prevalence in the genome [8–12]. For instance, estimates suggest CNVs encompass up to 10% of the human genome, account for more base pair differences between individuals than single nucleotide polymorphisms (SNPs), and overlap hundreds of genes [13].

Copy number variation is an important source of both genotypic and phenotypic variation, and commonly exerts its functional consequences through the modulation of gene expression [14]. Copy number variation can influence gene expression and gene function through a number of molecular mechanisms including gene dosage, regulatory element dosage, gene interruption, gene fusion, and position effect [15]. In humans, CNVs are associated with numerous genomic disorders including cancers, and autism-related disorders [16–19]. Conversely, CNVs can also lead to adaptive phenotypes such as the diet driven evolution of salivary amylase (AMY1) gene copy number (CN) in human populations, increased drug resistance of Plasmodium falciparum through elevated gch1 CN, and nematode resistance in soybean through increased CN of the multi-gene Rhgl locus [20–22].

In recent years, a number of studies have utilized high-throughput sequencing technologies to characterize the prevalence and function of CNVs in vertebrate populations [11, 23–26]. However, few studies have examined population-level patterns of CNVs in fungi and oomycetes despite the observation that SVs can play a role in host specificity, fungicide resistance, and degree of pathogenicity [27–30]. For example, multiple copies and modulated expression of the effector gene Avr1a in the oomycete Phytophthora sojae increases virulence and enables evasion of soybean immunity [27]. CNVs have also been demonstrated as playing a role in Cryptococcus neoformans virulence. Chromosomal CN variation in C. neoformans can lead to reduced fungicidal susceptibility by increasing copy number of ERG11, a target of fluconazole, and ARF1, an ABC transporter linked to azole susceptibility [29].

More than 1 million individuals die annually as a result of invasive mycoses and the vast majority of these life threatening infections occur in immuno-compromised individuals [31]. However, Cryptococcus gattii recently emerged as a primary human and animal pathogen, capable of infecting immuno-competent individuals [32]. A series of prominent and deadly C. gattii outbreaks occurred in Vancouver Island in 1999 and subsequently spread to mainland British Columbia, Oregon, and Washington [32, 33]. The epidemiology of these outbreaks was surprising considering C. gattii was believed to be endemic to subtropical and tropical regions [34]. The isolation of C. gattii from soil and trees revealed the global, but patchy, presence of C. gattii and the existence of four major phylogenetic lineages (denoted VGI, VGII, VGIII and VGIV) [33, 35, 36].

Within the Pacific Northwest (PNW), infections have been predominantly caused by isolates belonging to the VGIIa, VGIIb, and VGIIc subgroups [33]. Genomic and phylogenetic analyses suggest the PNW VGII subgroups likely originated from South America and Australia [33, 34, 37, 38]. The PNW VGII subgroups exhibit low levels of genetic variation indicative of a recent clonal expansion [30, 33, 37]. Despite the high degree of genetic similarity between PNW isolates, VGII subgroups display pathological diversity. For instance, VGIIa and VGIIc isolates exhibit heightened virulence and more commonly present as pulmonary infections, compared with VGIIb isolates which are generally less virulent and maintain the classical clinical pathology of neurological dominance [32, 33, 37]. Better understanding the genetic basis underlying VGII range expansion, niche adaptation, and pathological differences between subgroups could aid in the prevention, control, and treatment of future C. gattii outbreaks.

We hypothesize that CN variation could rapidly generate genotypic diversity in C. gattii subgroups and this variation might contribute to phenotypic and pathologic differences between VGII subgroups. To address this hypothesis, we identified 90 VGII isolates from more than 200 publicly available C. gattii whole-genome Illumina sequencing datasets and generated high-resolution genome-wide CNV profiles for each isolate (Additional files 1 and 2) [30, 33, 35, 39]. Within predominantly PNW isolates (one isolate from Caribbean, and one isolate from Australia), we (i) broadly characterized genomic patterns of CNVs, (ii) identified divergent CNVs between the VGII subgroups, and (iii) examined the functional associations of genes overlapping with CNVs.

Methods

Data mining and sequence read processing

Whole-genome Illumina data for C. gattii isolates were retrieved from the NCBI Sequence Read Archive using the SRA Toolkit (Additional file 1) [30, 33, 35, 39]. For all samples, trim_galore (http://www.bioinformatics.babraham.ac.uk/
projects/trim_galore/) was used to remove residual adapter sequences and trim reads at bases with quality scores below 30. Trimmed reads shorter than 100 nucleotides (nt) were discarded. Quality and adapter trimmed read sets were then mapped against the C. gattii R265 reference genome (https://www.broadinstitute.org/annotation/genome/cryptococcus_neoformans_b/MultiHome.html) using the ‘sensitive’ preset parameters in Bowtie2 [40]. Mapping outputs were converted to sorted bam format using samtools ‘view’ and ‘sort’ parameters [41]. We first calculated average genome-wide coverage for each sample by dividing the sum of the per-base nucleotide depth values, using the samtools depth function, by the C. gattii R265 genome size [41]. Because differences in coverage between samples can bias CNV estimates, we used seqtk (https://github.com/lh3/seqtk) to randomly subsample reads such that each sample was normalized to 10× coverage. We chose a 10× coverage in order to retain the majority of publicly available data, and to ensure reliable CN estimation [42]. Samples with average pre-normalized coverage values less than 10× were removed from analysis, resulting in a total of 212 isolates. These 10× coverage subsampled datasets were mapped against the reference genome, and sorted bam files were generated as described above. The 10× coverage datasets were used for SNP calling and CNV estimation.

Inferring population structure

We performed population structure analysis to identify isolates belonging to the VGIIa, VGIIb, and VGIIc subgroups. SNPs were conservatively called for each sample using VarScan.v2.3.9, requiring a minimum coverage of 8X and a SNP frequency of 100 % [43]. SNP sites were removed from a consensus whole population matrix when >10 % of the population harbored an ambiguously called base, or when minor allele frequency was <10 %. We subsampled the resulting matrix to ensure SNP sites were separated by >10 Kb (average distance markers ~13 Kb). The final matrix used for evolutionary analysis consisted of 1223 SNPs.

We employed several methods to infer the population structure of the 212 isolates. First, we used the Bayesian model-based software Structure 2.3.4, using the “admixture” ancestry model and the “allele frequencies are correlated among populations” frequency model [44]. We ran a burn-in period of 100,000 and a Markov Chain Monte Carlo (MCMC) of 200,000 generations for K = 1 – 13, where K represents the number of genetic clusters or populations. The optimal number of clusters was predicted by calculating the average log probability (LnP(D)) of each K value and by calculating the rate of change in the LnP(D) between successive Structure runs with different K values (ΔK) [44, 45]. Both statistics were calculated using Structure Harvester [46]. Because Structure assumes Hardy-Weinberg equilibrium and linkage equilibrium and natural populations often violate these assumptions, we also employed Discriminant Analysis of Principal Components (DAPC) in the adegenet v1.3-1 package. DAPC is a non-model-based multivariate approach, to infer population structure [47]. The number of distinct populations was predicted using the “find.clusters” k-means clustering algorithm and by calculating the Bayesian Information Criterion (BIC) value for each K = 1–16. Predicting the optimal population number is often unclear and complex in panmictic natural populations. An alternative approach is to define the number of populations that are useful in describing the genetic data [47]. Using this approach, BIC values smaller than 1000 represent useful summaries of population structure [47].

When assigning individuals to subpopulations, we assigned isolates to clusters when cluster assignment between Structure and DAPC were in agreement. Isolates with membership coefficients <90 % were removed. Clusters were assigned a VGIIa, VGIIb, or VGIIc classification according to predominant clustering of subgroups previously identified by other studies [30, 33, 35, 39]. This classification scheme resulted in 35 VGIIa, 22 VGIIb, and 33 VGIIc isolates.

Genome-wide copy number quantification

Whole supercontig CN variation can be detected through differences in the relative number of mapped reads. To estimate supercontig CN, a normalized mapped read value (NMRV) for each sample was determined by calculating the number of reads mapped per supercontig using the samtools ‘ixstats’ function [41]. The NMRV was calculated using only ten largest supercontigs which have an average length of 1,181,220 base pairs (bp). To estimate supercontig CN, the number of reads that mapped to each supercontig was calculated using the samtools ‘ixstats’ function and then divided by NMRV. Thus, relative read mapping serves as a proxy for CN events of larger sizes.

We used the Control-FREEC software for high-resolution CNV quantification [48]. Using the read depth approach, Control-FREEC additionally implements a sliding window and Least Absolute Shrinkage and Selection Operator (LASSO) approach for CNV detection [48–50]. Integer CN was estimated for each 100 bp non-overlapping window of the genome using the following parameters: window = 100, telocentric = 0, minExpectedGC = 0.33, and maxExpectedGC = 0.63. CN of high VST genes was calculated as the average copy number value of 100 bp windows that overlap with the gene of interest.
Identifying divergent CNVs between VGII subpopulations

The $V_{ST}$ measurement, derived from $F_{ST}$, is used to identify loci that differentiate by CN between populations [9]. $V_{ST}$ and $F_{ST}$ consider how genetic variation is partitioned at the individual, population, and global levels and range from 0 (no population differentiation) to 1 (complete population differentiation). $V_{ST}$ has proven useful in identifying CNVs under positive selection in genetically distinct populations [8, 25, 51]. We calculated $V_{ST}$ between VGIIa, VGIIb, and VGIIc groups for each non-overlapping 100 bp bin in the genome using the following formula:

$$V_{ST} = \frac{\{V_{VGIIa} + N_{VGIIb} + V_{VGIIb} + V_{VGIIc} + N_{VGIIc}\}}{V_{total} \cdot N_{total}}$$

$V_{total}$ is total variance, $V_{VGIIx}$ is the CN variance for each respective subpopulation, $N_{VGIIx}$ is the sample size for each respective subpopulation, and $N_{total}$ is the total sample size. We considered $V_{ST}$ values in the upper 99th percentile, corresponding to values greater than 0.3397, as significant.

Genomic statistics

To assess differences in gene gain or loss between subgroups, we used a non-parametric Fisher-Pitman permutation test implemented in the R package coin, using 1,000,000 permutations [52]. The R package boot was utilized for bootstrapping statistics to provide confidence intervals with 1,000,000 bootstrap replicates [53, 54]. If statistical significance was observed between subgroups, permutation-based pairwise $T$-tests (100,000 permutations) with Bonferroni-corrected $p$-values, as implemented in the R package RVAideMemoire, were used to assess statistical differences between subgroups [55].

Enrichment analysis

We used Gene Ontology (GO) annotation from the C. neoformans H99 genome to assign functional classification to C. gattii orthologs [56]. Orthologs were predicted using the reciprocal best-BLAST hit approach with an e-value cut-off of $1e^{-6}$ [57, 58]. PFam domains in high $V_{ST}$ genes were predicted using interproscan-5.8-49.0 [59]. Enrichment analysis was conducted using default settings in GOEAST (GO terms) and dcGO (PFam domains) [60, 61]. Results from both web-based programs were visualized in REVIGO [62].

Results

Population structure of C. gattii isolates

We were primarily interested in identifying loci with distinct CN profiles between the VGII subpopulations. Thus, we first performed population structure analysis in order to identify VGIIa, VGIIb, and VGIIc isolates using Structure and DAPC [44, 47]. We generated a SNP matrix of 1223 informative loci for 212 individuals (see Methods), and determined that a $K$ of eight was the optimal number of populations by the $AK$ method (Fig. 1a) [45]. With the same SNP matrix, DAPC did not determine a single optimal $K$ but rather an informative range of $K$ from 4 to 16 (Fig. 1b) [47]. We chose a $K$ value of 8 as it was the optimal value in Structure and fell within DAPC’s suggested range. Membership coefficients were independently calculated using Structure and DAPC (Fig. 1c and d, respectively) [44, 47]. When assessing membership coefficients across $K$ 5–13, VGIIa, VGIIb and VGIIc individuals clustered together consistently (Fig. 1e). Using this approach, a total of 35, 22, and 33 VGIIa, VGIIb, and VGIIc isolates were identified, respectively (Fig. 1b, d, and e). Subsequent CN variation analysis was performed on this set of 90 VGII isolates.

Copy Number Variant (CNV) analysis

We generated high-resolution CNV profiles for the 90 VGIIa, VGIIb, and VGIIc isolates to better understand the contribution of CNVs to genome variation. On average, CNV regions encompassed 1.63 % (284,860 bp) of the VGIIa genomes, 2.32 % (404,935 bp) of the VGIIb genomes, and 1.96 % (341,803 bp) of the VGIIc genomes, in comparison to the reference R265 genome. The average number of CNV regions per sample for VGIIa, VGIIb, and VGIIc was 156 ± 20.97, 171 ± 20.34, and 194 ± 21.29, respectively. We performed several analyses to (i) identify large-scale chromosomal and segmental duplication and deletion events, (ii) broadly characterize genomic patterns of CNVs, and (iii) identify CNVs that are differentiated between VGII subgroups.

Aneuploidy has been previously linked to variance in fungal pathogenicity, thus, we estimated CN for each supercontig (Fig. 2) [29, 30, 63, 64]. The vast majority of supercontigs were present at a CN of ~1; however, some supercontigs were present at higher or lower CN between subgroups. We calculated $V_{ST}$ between CN estimates of entire supercontigs to identify those that differed between VGII subgroups. The average $V_{ST}$ value for supercontigs was 0.097 suggesting most supercontigs were equally represented between subgroups and that variation is similarly partitioned between VGII subpopulations rather than within individual subpopulations. However, supercontig 26 was found at greater CN in VGIIa (1.14 ± 0.111 CN) compared to VGIIb (0.75 ± 0.068 CN) and VGIIc (0.67 ± 0.066 CN) ($V_{ST} = 0.861$) (Fig. 2). Supercontig 26 is one of the smallest supercontigs (15,500 bp) in the R265 reference genome and contains only one short uncharacterized gene (CNBG_9678) with no annotated protein domains. We also identified three individuals with atypical supercontig CNs. Two
VGIIb isolates had an estimated CN of 1.87 and 1.84 for supercontig 13, and another VGIIb isolate had an estimated copy number of 1.75 for supercontig 9. A large scale duplication on supercontig 13 has been previously reported in a clinically isolated VGII sample from France [37].

For high-resolution CNV prediction in each VGII isolate, we calculated integer CN for each 100 bp non-overlapping window in the genome using Control-FREEC [48]. To examine the potential functional effects of CNVs, we compared the frequency of genes for which...
CN variable regions partially or fully overlapped relative to the R265 reference genome (Fig. 3). VGIIa had the fewest partially deleted genes, VGIIb had the greatest partially deleted genes, and VGIIc had the fewest partially duplicated genes and the greatest partially deleted genes (Fig. 3a). The number of partially duplicated genes significantly differed between subgroups (Fisher-Pitman permutation test; \( p = 8.88 \times 10^{-3} \)) and post-hoc pairwise comparisons revealed that VGIIc had fewer partially duplicated genes than VGIIa (\( p = 0.034 \)) but not VGIIb (\( p > 0.05 \)) (Fig. 3c). Further, we found that the frequency of partially deleted genes was significantly different between VGII subgroups (Fisher-Pitman permutation test; \( p < 2.2 \times 10^{-16} \)) with post-hoc pairwise comparisons revealing that VGIIb had higher levels of partially deleted genes than VGIIa (\( p = 0.006 \)) and VGIIb (\( p = 0.012 \)) (Fig. 3d). We also found significant differences in the number of whole duplicated (\( p = 0.0051 \)) and deleted (\( p < 2 \times 10^{-6} \)) genes between VGII subpopulations. VGIIb had significantly more duplicated genes than VGIIc (\( p = 0.039 \)), while VGIIa had significantly fewer whole deleted genes compared with VGIIb (\( p < 6 \times 10^{-5} \)) and VGIIc (\( p < 6 \times 10^{-5} \)) (Fig. 3e and f). Reference genome bias might account for the lower frequency of CNVs in VGIIa since R265 is a member of the VGIIa subgroup.

Identifying divergent CNVs between VGII subgroups

We hypothesized that CN profiles that segregated by VGII subgroups might contribute to the pathological differences between VGII subgroups. To identify these divergent CNV regions, we calculated \( V_{ST} \) for each 100 bp bin across the genome between VGIIa, VGIIb, and VGIIc subpopulations [9]. Values near 0 represent no CN differentiation between subpopulations while values of 1 are indicative of fixed copy number differences between subpopulations. The average \( V_{ST} \) value for genomic bins where at least one isolate was copy number variable was 0.041, suggesting that the majority of CNV loci are not differentiated between VGII subpopulations. To identify highly differentiated CNVs, we examined loci in the upper 99 % percentile of \( V_{ST} \) values (\( V_{ST} > 0.3397 \)). Using this approach, we identified 67 loci harboring a total of 58 genes (Fig. 4a, Additional file 3). One of the highest \( V_{ST} \) regions (0.885) is present on supercontig 8, encompasses 10,900 bp, and contains four genes. Each gene in this locus is present near one copy in VGIIa, two copies in VGIIb, and zero copies in VGIIc. Two of the high \( V_{ST} \) loci are separated by 100 bp on supercontig 11 (Fig. 4b). The total length of these two neighboring high \( V_{ST} \) loci is 20.5 Kb and the region contains eight genes. All genes
within this region are present as a single copy in VGIIa but absent in VGIIb and VGIIc. This region has been reported previously using a different computational approach and reinforces the efficacy of our methodology [33].

Lastly, we compiled a list of candidate genes associated with *Cryptococcus* and fungal pathogenicity from the literature and examined their CNV profiles [65–89]. A comprehensive suite of genes empirically related to virulence in *C. gattii* or *C. neoformans* (53 total genes) and chitin synthase genes (10 total genes) revealed very few deletion or duplication events. However, several potential virulence associated genes were found in high V_{ST} regions and are discussed later. We did identify CNVs in several genes related to transport activity. The majority of high V_{ST} transporters show the same pattern in which VGIIa isolates retain a single copy, while VGIIb and VGIIc isolates have experienced deletions (CNBG_9639, CNBG_2482, CNBG_4491, CNBG_2944) (Additional file 3).

**Functional analysis of CNV genes**

We performed enrichment analysis of genes overlapping CNVs to broadly characterize their functional associations. First, we identified orthologs between *C. neoformans* and *C. gattii* using the Reciprocal Best BLAST Hit (RBBH) approach [57, 58]. *C. neoformans* GO annotation was
assigned to *C. gattii* genes with one-to-one orthology [56]. We performed GO term enrichment analysis on all genes in the high *V*<sub>ST</sub> CNV regions, on the genes overlapping CNV regions within each VGII subpopulation, and on the genes overlapping CNV regions across all VGII subpopulations. We independently analyzed genes that were (i) deleted partially, (ii) deleted entirely, (iii) duplicated partially, and (iv) duplicated entirely because each type of mutation cannot be categorized as “functional” or “nonfunctional” without experimental validation. For instance, partial gene deletions and duplications may result in loss of function, but may also be adaptive [90–92]. No GO terms were significantly enriched in the combined VGII subpopulations when entire deleted genes, partially deleted genes, or entire duplicated gene sets were analyzed. However, in the gene set composed of CNV regions that partially overlap genes across the combined VGII samples, we identified several enriched GO terms (for a full list of enriched GO terms see Additional file 4). The top three enriched GO terms in the biological processes category were *transmembrane transport* (*p* = 1.7e<sup>−4</sup>), *methylation* (*p* = 2.4e<sup>−3</sup>), *transport* (*p* = 5.4e<sup>−3</sup>), the top three enriched GO terms in the molecular function category were *substrate-specific transmembrane transporter activity* (*p* = 3.41e<sup>−5</sup>), *transmembrane transporter activity* (*p* = 3.98e<sup>−5</sup>), *receptor activity* (*p* = 1.15e<sup>−2</sup>), and the top three GO terms in the cellular component category were *proton-transporting V-type ATPase, V1 domain* (*p* = 1.69e<sup>−4</sup>), *proton-transporting V-type ATPase complex* (*p* = 1.4.3e<sup>−3</sup>), and *proton-transporting two-sector ATPase complex, catalytic domain* (*p* = 2.2e<sup>−3</sup>) (Additional file 4).

We also examined the enrichment of Pfam domains using dcGO to provide an additional level of resolution for functional analysis of high *V*<sub>ST</sub> genes [61, 93]. Within the biological processes category, we observed enrichment of protein domains associated with primary metabolism.
In particular, several of these categories are associated with fundamental processes such as carbohydrate metabolic process \((p = 1.99 \times 10^{-5})\), telomere maintenance \((p = 2.04 \times 10^{-3})\), and mitochondrial proton-translocating ATP synthase, catalytic core \((p = 4.83 \times 10^{-7})\). Additionally, we identified several terms that are associated with environmental sensing, including regulation of response to biotic stimulus \((p = 4.86 \times 10^{-4})\) and response to xenobiotic stimulus \((p = 5.59 \times 10^{-4})\).

Analysis of terms related to the cellular component category revealed several terms associated with the cell surface, including external encapsulating structure \((p = 5.82 \times 10^{-4})\), microvillus \((p = 4.3 \times 10^{-4})\) and proton-translocating two-sector ATPase complex \((p = 6.72 \times 10^{-4})\) (Fig. 5b, Additional file 5). We also identified an enrichment of transport-related domains including active transmembrane transporter activity \((p = 8.3 \times 10^{-4})\), cation-transporting ATPase activity \((p = 9.56 \times 10^{-5})\), and proton-transporting ATPase activity, rotational mechanism \((p = 3.31 \times 10^{-6})\) (Fig. 5c, Additional file 5).

Discussion
We have conducted the first genome-wide population genomic analysis of \(C.\ gattii\) CN variation. Our high-resolution analysis revealed SVs ranging from 100 to 365,300 bp (Figs. 2, 3 and 4) encompassing ~1–2% of the genome. The proportion of the genome affected by CNVs are in agreement to levels found in \(Homo\ sapiens\) (~5%), \(Danio\ rerio\) (~4.5%), \(Sus\ scrofa\ domesticus\) (1.5%), \(Arabidopsis\ thaliana\) (0.75%), and the yeasts \(S.\ cerevisiae\) (1.2%) and \(S.\ paradoxus\) (3.5%) [13, 26, 31, 94, 95].

We were primarily interested in identifying differentiated CNVs between the VGIIa, VGIIb, and VGIIc subpopulations given their close phylogenetic relationship but distinct pathological phenotypes. We calculated the CN specific population statistic \(V_{ST}\) to identify CN profiles that differed between VGII subpopulations (Fig. 4). Selection can cause patterns of CNV divergence between populations, however, other population genetic process, such as genetic drift and gene flow can also contribute to CNV profiles that are stratified by population. Nonetheless, the \(V_{ST}\) approach has offered insights into phenotypic differences between populations [31, 51, 96–98]. For instance, high \(V_{ST}\) loci in South African Nguni cattle revealed CN variable genes involved in parasite resistance, body size, and fertility that may account for phenotypic differences between breeds [98].

In our analysis, one of the highest \(V_{ST}\) values was obtained for a ~20 Kb region of supercontig 11 (VST-SC_1.11_1 and VST-SC_1.11_2) that contained 8 genes associated with sugar transport, alcohol dehydrogenase activity, and collagenase activity (Fig. 4b). In VGIIa the locus is present at a single copy, while deleted from the VGIIb and VGIIc isolates. This locus and pattern had been previously discovered through de novo genome assembly, \textit{in silico} gene prediction, and BLAT score ratio analysis between VGIIa, VGIIb, and VGIIc isolates [33]. Another high scoring \(V_{ST}\) region included a four gene, 10.9 Kb region on supercontig 8 (VST-SC_1.26_1) with a CN of 1, 2, and 0 in the VGIIa, VGIIb, and VGIIc subpopulations, respectively (Fig. 5b). This locus included a putative transcription factor with a Zinc cluster domain, a glycosyl hydrolase family gene, and a deoxyribose-phosphate aldolase. In \textit{Staphylococcus pneumoniae} a glycosyl hydrolase encoding gene (GHIP) is involved in

**Fig. 5** Functional enrichment of high \(V_{ST}\) loci. Gene Ontology (GO) enrichment based on PFam domains present in the genes overlapping high \(V_{ST}\) loci. Enrichment analysis was conducted via \textit{dcGO} and is independently reported for Biological Processes (a), Cellular Component (b), and Molecular Function (c) [66]. Terms with \(p\)-values <0.01 are depicted. The shading of red represents \(p\)-value, with dark red corresponding to the lowest \(p\)-value. The area of each circle is relative to the number of terms falling into the respective GO category. Grey lines connect related GO terms and the thickness of the line indicated the degree of relatedness.
host-cell invasion and knockout mutants exhibit reduced capacity to colonize mouse tissue [99]. In *Toxoplasma gondii*, a deoxyribose-phosphate aldolase plays an essential role in pathogenicity by mediating host-cell invasion, and providing energy during the invasion process [100].

Enrichment analysis of genes in high VST regions revealed cohesive networks of functional terms potentially linked to *C. gattii* pathogenicity, including many transport related GO terms (Fig. 5, Additional files 4 and 5). Transporter encoding genes are often associated with fungal virulence because of their ability to export fungi-cides from the cell [101]. Differences in fluconazole minimum inhibitory concentration within and between VGII subgroups have been observed, but have not been attributed to target gene (*ERG11*) expression or genetic variation [65, 67]. Rather, fluconazole resistance in *C. neoformans* and *C. gattii* appears to be driven by up-regulation of the ABC transporter encoding genes *ARF1*, *ARF2*, and *MDRI* [65, 102, 103]. Genes encoding a diversity of other transporters also contribute to *Cryptococcus* pathogenicity through such mechanisms as micronutrient scavenging and sugar transport for capsule formation [104–106]. Several Major Facilitator Superfamily (MFS) domains were predicted in the high VST genes including *CNBG_4491* (VGIIa: one copy, and VGIIb and VGIIc: zero copies), *CNBG_9477* (VGIIa: one copy, and VGIIb and VGIIc: zero copies), and *CNBG_9094* (VGIIa and VGIIb: two copies, and VGIIc: one copy).

The anti-phagocytic capsule is one of the defining virulence factors in *C. neoformans* and *C. gattii* and is primarily composed of a lattice of glucuronoxyloman-nan, glucuronoxylomannogalactan, and mannoproteins [34]. We observed the enrichment of several GO terms in the high VST genes that might be associated with the capsule, including polysaccharide catabolism and external encapsulating structure (Fig. 5). For instance, *CNBG_1084* contains a fibronectin type III-like domain. This domain can interact with the *ARFI* protein, which, along with its contribution to fluconazole resistance, is also involved in vesicle transport and influences capsule size [107, 108]. In VGIIc, the CNV found in *CNBG_1084* results in a ~600 bp deletion in the 3′ end of the gene and overlaps with the fibronectin type III-like domain and the stop codon. We also identified a highly differentiated 100 bp CNV (VST = 0.94) located in the 3′ end of a histone deacetylase (HDAC) encoding gene (*CNBG_5703*) (Additional file 3). This CNV did not overlap with the HDAC domain but did overlap with the stop codon (VGIIa: one copy, and VGIIb and VGIIc: zero copies). Through phenotypic screening of >1000 targeted gene deletion mutants in *C. neoformans*, Liu et al. [109] demonstrated that HDAC encoding genes contribute to capsule formation and influence infectivity.

CNVs can affect phenotype through changes in gene dosage, which most commonly correspond to alterations in gene expression [20, 22, 110–112]. Using a high resolution in silico approach, we have identified 67 CN variable regions in the *C. gattii* genome that contain 58 genes and are highly differentiated by VGII subgroup (Fig. 4, Additional file 3). These loci may account for some of the observed pathological differences between VGII subgroups. Moving forward, it will be essential to experimentally evaluate the function roles of these genes and to assess the regulatory impact of CNVs. Recent developments in molecular tools, such as RNA interference and CRISPR-Cas9 genome editing, as well as a murine model of *C. gattii* meningoencephalitis, will aid in assessing the function of *C. gattii* CNV gene candidates [113, 114].

**Conclusion**

We identified 90 *C. gattii* isolates corresponding to the subpopulations responsible for the deadly outbreaks in the North American Pacific North West. Within these isolates, we biinformatically predicted CNVs and uncovered loci that display subpopulation specific patterns of variation. Many of these CNV differentiated loci harbored genes that were enriched for metabolic, transport, and capsule composition associated functions. These genes represent novel candidates that may explain some of the pathological differences between VGII subgroups. Further functional investigation of candidate genes is needed to better understand the impact of copy number variation in *C. gattii* virulence.

**Additional files**

Additional file 1: NCBI Sequence Read Archive run accession numbers for whole-genome Illumina data of the 212 *Cryptococcus gattii* isolates. VGII subgroup classification based on *Structure* and DAPC analysis (where K = 8) is provided where applicable. (XLSX 13 kb)

Additional file 2: Individual copy number estimates and VST for each 100 bp bin relative to the *Cryptococcus gattii* R265 genome. (XLSX 49905 kb)

Additional file 3: High VST regions with genome coordinates and gene identifiers relative to *Cryptococcus gattii* R265. (XLSX 15 kb)

Additional file 4: Gene Enrichment Analysis of GO terms for CNV regions that partially overlap genes. Enrichment analysis is based on *Cryptococcus neoformans* annotation for genes with 1:1 orthology between *C. neoformans* and *C. gattii*. (XLSX 12 kb)

Additional file 5: Functional enrichment analysis GO terms for high VST genes based on Pfam domains. (XLSX 12 kb)

**Acknowledgements**

We acknowledge financial support from Clark University (JGG). This work was conducted in part using the Clark University supercomputing cluster.

**Availability of data and materials**

Accession numbers for all whole-genome Illumina data used in this study are provided in Additional file 1.
Cavendish B, Alspaugh JA. Ras1 and Ras2 contribute shared and unique roles in physiology and virulence of the primary pathogenic yeast Cryptococcus neoformans. Eukaryot Cell. 2006;5(7):1065–9.

Bass et al. Jr LR, Gast CE, Bruzal J, Wong B. Identification and properties of plasma membrane azole efflux pumps from the pathogenic fungi Cryptococcus gattii and Cryptococcus neoformans. J Antimicrob Chemother. 2015;70(3):396–407.

Schneider Rde O, Fogaca Nde S, Knettitz L, Schrank A, Vainstein MH, Staats CC. Zap1 regulates zinc homeostasis and modulates virulence in Cryptococcus gattii. PLoS One. 2012;7(8):e43773.

Gast CE, Basso Jr LR, Bruzal J, Wong B. Azole resistance in Cryptococcus gattii from the Pacific Northwest: investigation of the role of ERG11. Antimicrob Agents Chemother. 2013;57(11):5478–85.

Feder V, Knettitz L, Staats CC, Vidal-Figuereido N, Ligabue-Braun R, Carlini CR, Vainstein MH. Cryptococcus gattii urease as a virulence factor and the relevance of enzymatic activity in cryptococcosis pathogenesis. FEBS J. 2015;282(8):1406–18.

Chen YL, Lehman VN, Lewit Y, Averette AF, Heitman J. Calcineurin governs thermotolerance and virulence of Cryptococcus gattii. G3 (Bethesda). 2013;3(3):527–39.

Schneider Rde O, Diehl C, dos Santos FM, Pffer AC, Garcia AW, Kullmni M, Schrank A, Knettitz L, Vainstein MH, Staats CC. Effects of zinc transporters on Cryptococcus gattii virulence. Sci Rep. 2015;5:10104.

Shimizu K, Imanishi Y, Toh-e A, Uno J, Chibana H, Hull CM, Kawamoto S. Functional characterization of Pmt2, encoding a protein-O-mannosyltransferase, in the human pathogen Cryptococcus neoformans. Fungal Genet Biol. 2014;69:13–22.

Ngamkulthung P, Himmelreich U, Breger JA, Wilson JK, Chayakulkeeree M, Krooienberge MB, Malik R, Daniel HM, Toffaletti D, Djordjevic JT, et al. The trehalose synthesis pathway is an integral part of the virulence composite for Cryptococcus gattii. Infect Immun. 2009;77(10):4584–96.

Ren P, Springer DJ, Behr MJ, Samsonoff WA, Chaturovedi V, Chaturovedi V. Transcription factor STE11alpha has distinct roles in morphogenesis, virulence, and ecological fitness of the primary pathogenic yeast Cryptococcus gattii. Eukaryot Cell. 2006;5(7):1065–9.

Narasinga Sudha SD, Ault IG, Behr MJ, Chaturovedi V, Chaturovedi V. Characterization of Cu, Zn superoxide dismutase (SOD1) gene knock-out mutant of Cryptococcus neoformans var. gattii: role in biology and virulence. Mol Microbiol. 2003;47(6):1681–94.

Chen SC, Muller M, Zhou JZ, Wright LC, Sorrell TC. Phospholipase B activity in Cryptococcus neoformans: a new virulence factor? J Infect Dis. 1997;175(2):144–20.

Ganendren R, Carter E, Sorrell T, Widmer F, Wright L. Phospholipase B activity enhances adhesion of Cryptococcus neoformans to a human lung epithelial cell line. Microbes Infect. 2006;8(4):1006–15.

D’Souza CA, Alspaugh JA, Yue C, Harashima T, Cox GM, Perfect JR, Heitman J. Cyclic AMP-dependent protein kinase controls virulence of the fungal pathogen Cryptococcus neoformans. Mol Cell Biol. 2001;21(9):3717–9.

Bahn YS, Kojima K, Cox GM, Heitman J. A unique fungal two-component system regulates stress responses, drug sensitivity, sexual development, and virulence of Cryptococcus neoformans. Mol Biol Cell. 2006;17(7):3122–35.

Kronstad JW, Attarian R, Cadenius B, Choi J, D’Souza CA, Griffiths EJ, Geddes JM, Hu G, Jung WH, Kretscher M, et al. Expanding fungal pathogenesis: Cryptococcus breaks out of the opportunistic box. Nat Rev Microbiol. 2011;9(3):193–203.

Idnurm A, Walton FJ, Floyd A, Reedy JL, Heitman J. Identification of ENA1 as a virulence gene of the human pathogenic fungus Cryptococcus neoformans through signature-tagged insertional mutagenesis. Eukaryot Cell. 2009;8(3):315–26.

Idnurm A, Reedy JL, Nussbaum JC, Heitman J. Cryptococcus neoformans virulence gene discovery through insertional mutagenesis. Eukaryot Cell. 2004;3(2):420–9.

Wuah MS, Nichols CB, DeCesare CM, Cox GM, Heitman J, Alspaugh JA. Ras1 and Ras2 contribute shared and unique roles in physiology and virulence of Cryptococcus neoformans. Microbiology. 2002;148(Pt 1):191–201.

Kronstad JW, Hu G, Choi J. The CAMP/Protein Kinase A pathway and virulence in Cryptococcus neoformans. Mycobiology. 2011;39(3):143–50.

Giles SS, Stajich JE, Nichols C, Gerald QD, Alspaugh JA, Diethrich F, Perfect JR. The Cryptococcus neoformans catalase gene family and its role in antioxidant defense. Eukaryot Cell. 2005;4(9):1447–59.

Waterman SR, Hacham M, Hu G, Zhu X, Park YD, Shin S, Panepinto J, Vaty-Nagy T, Beam C, Husain S, et al. Role of a CUF1/CTR4 copper regulatory axis in the virulence of Cryptococcus neoformans. J Clin Invest. 2007;117(3):794–802.
86. Bahn YS, Hicks JK, Giles SS, Cox GM, Heitman J. Adenylyl cyclase-associated protein Aca1 regulates virulence and differentiation of Cryptococcus neoformans via the cyclic AMP-protein kinase A cascade. Eukaryot Cell. 2004;3(6):1476–91.

87. Jung WH, Saikia S, Hu G, Wang J, Fung CK, D’Souza C, White R, Kronstad JW. HapN positively and negatively regulates the transcriptional response to iron deprivation in Cryptococcus neoformans. PLoS Pathog. 2010;6(11):e1001209.

88. Okabayashi K, Hasegawa A, Watanabe T. Microreview: capsule-associated genes of Cryptococcus neoformans. Mycopathologia. 2007;163(1):1–8.

89. Mylonakis E, Idnurm A, Moreno R, El Khoury J, Rottman JB, Ausubel FM, Heitman J, Calderwood SB. Cryptococcus neoformans Km1 protein kinase homologue, identified through a Caenorhabditis elegans screen, promotes virulence in mammals. Mol Microbiol. 2004;54(2):407–19.

90. Fidalgo M, Barrales RR, Ibeas JL, Jimenez J. Adaptive evolution by mutations in the FLO11 gene. Proc Natl Acad Sci U S A. 2006;103(30):11228–33.

91. Katju V, Lynch M. The structure and early evolution of recently arisen gene duplicates in the Caenorhabditis elegans genome. Genetics. 2003;165(4):1793–803.

92. Qian W, Zhang J. Genomic evidence for adaptation by gene duplication. Genome Res. 2014;24(8):1356–62.

93. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Quek S, Rose DW, S Allan, Sonnhammer EL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44(D1):D279–285.

94. Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X, Chen W, Xie M, Wang W, Hammond S, et al. The genomic sequence of the Chinese hamster ovary (CHO-K1) cell line. Nat Biotechnol. 2011;29(8):735–41.

95. Bergstrom A, Simpson JT, Salinas F, Barre B, Lars P, Zia A, Nguyen BA, Moses AM, Louis EI, Mustonen V, et al. A high-definition view of functional genetic variation from natural yeast genomes. Mol Biol Evol. 2014;31(4):872–88.

96. Xu L, Hou Y, Bickhart DM, Zhou Y, He H, Song J, Sonstegard TS, Van Tassell CP, Liu G. Population-genetic properties of differentiated copy number variations in cattle. Sci Rep. 2016;6:23161.

97. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Stuve LL, Lander ES, Coluccio AE, Danilova TV, Kudrna D, et al. Global diversity, population stratification, and selection of human copy-number variation. Science. 2015;349(6253):aab3761.

98. Liu OW, Chun CD, Chow ED, Chen C, Madhani HD, Noble SM. Systematic genetic analysis of virulence in the human fungal pathogen Cryptococcus neoformans. Cell. 2008;135(1):174–88.

99. Bibbons JG, Salichos L, Slot JC, Rinker DC, McGary KL, King JG, Klich MA, Tabb DL, McDonald WH, Rokas A. The evolutionary imprint of domestication on genome variation and function of the filamentous fungus Aspergillus oryzae. Curr Biol. 2012;22(15):1403–9.

100. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA. Large multiallelic copy number variations in humans. Nat Genet. 2015;47(3):296–303.

101. Maron LG, Guimarães CT, Kirt M, Albert PS, Bircher JA, Bradbury PJ, Buckler ES, Coluccio AE, Danilova TV, Kudrna D, et al. Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc Natl Acad Sci U S A. 2013;110(13):5241–6.

102. Skowrya ML, Doering TL. RNA interference in Cryptococcus neoformans. Methods Mol Biol. 2012;895:165–86.

103. Thompson 3rd GR, Wiederhold NP, Najvar LK, Bocanegra R, Kirkpatrick WR, Graybill JR, Patterson TF. A murine model of Cryptococcus gattii meningoencephalitis. J Antimicrob Chemother. 2013;67(6):1432–8.