Cryptococcus gattii in North American Pacific Northwest: Whole-Population Genome Analysis Provides Insights into Species Evolution and Dispersal

David M. Engelthaler, a Nathan D. Hicks, a John D. Gillece, a Chandler C. Roe, a James M. Schupp, a Elizabeth M. Driebe, a Felix Gilgado, b Fabian Carricone, b,c Lucia Trilles, b,d Carolina Ficaricata, b,e Popchai Ngamskulrungroj, b,f Elizabeth Castañeda, b Marcia dos Santos Lazera, b Marcia S. C. Melhem, b, f Ása Pérez-Bercoff, b,h Gavin Huttley, b Tania C. Sorrell, b Kerstin Voelz, b,i Robin C. May, i,j Matthew C. Fisher, b George R. Thompson III, i,j Shawn R. Lockhart, m Paul Keim, a,n Wieland Meyer b

Translational Genomics Research Institute, Flagstaff, Arizona, USA; a Molecular Mycology Research Laboratory, Center for Infectious Diseases and Microbiology, Sydney Medical School, Westmead Hospital, Marie Bashir Institute for Infectious Diseases and Biosecurity, The University of Sydney, Westmead Millennium Institute, Sydney, Australia; b Institut Agronomique Néo-Calédonien (IAC), Noumea, New Caledonia; c Laboratório de Micologia, Instituto de Pesquisa Clínica Evandro Chagas, Fiocruz, Rio de Janeiro, Brazil; d Microbiology Group, Instituto Nacional de Salud, Bogotá, Colombia; e Department of Microbiology, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand; f Public Health Reference Laboratory, Institute Adolfo Lutz, São Paulo, SP, Brazil; g John Curtin School of Medical Research, Australian National University, Canberra, Australia; h Institute of Microbiology and Infection and School of Biosciences, University of Birmingham, Birmingham, United Kingdom; i National Institute for Health Research Surgical Reconstruction and Microbiology Research Centre, Queen Elizabeth Hospital, Birmingham, Birmingham, United Kingdom; j Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom; k University of California, Davis, Davis, California, USA; l Centers for Disease Control and Prevention, Atlanta, Georgia, USA; m, n Microbial Genetics and Genomics Center, Northern Arizona University, Flagstaff, Arizona, USA.

D.M.E. and N.D.H. contributed equally to this article.

ABSTRACT The emergence of distinct populations of Cryptococcus gattii in the temperate North American Pacific Northwest (PNW) was surprising, as this species was previously thought to be confined to tropical and semitropical regions. Beyond a new habitat niche, the dominant emergent population displayed increased virulence and caused primary pulmonary disease, as opposed to the predominantly neurologic disease seen previously elsewhere. Whole-genome sequencing was performed on 118 C. gattii isolates, including the PNW subtypes and the global diversity of molecular type VGII, to better ascertain the native source and genomic adaptations leading to the emergence of infection in the PNW. Overall, the VGII population was highly diverse, demonstrating large numbers of mutational and recombinational events; however, the three dominant subtypes from the PNW were of low diversity and were completely clonal. Although strains of VGII were found on at least five continents, all genetic subpopulations were represented or were most closely related to strains from South America. The phylogenetic data are consistent with multiple dispersal events from South America to North America and elsewhere. Numerous gene content differences were identified between the emergent clones and other VGII lineages, including genes potentially related to habitat adaptation, virulence, and pathology. Evidence was also found for possible gene introgression from Cryptococcus neoformans var. grubii that is rarely seen in global C. gattii but that was present in all PNW populations. These findings provide greater understanding of C. gattii evolution in North America and support extensive evolution in, and dispersal from, South America.

IMPORTANCE Cryptococcus gattii emerged in the temperate North American Pacific Northwest (PNW) in the late 1990s. Beyond a new environmental niche, these emergent populations displayed increased virulence and resulted in a different pattern of clinical disease. In particular, severe pulmonary infections predominated in contrast to presentation with neurologic disease as seen previously elsewhere. We employed population-level whole-genome sequencing and analysis to explore the genetic relationships and gene content of the PNW C. gattii populations. We provide evidence that the PNW strains originated from South America and identified numerous genes potentially related to habitat adaptation, virulence expression, and clinical presentation. Characterization of these genetic features may lead to improved diagnostics and therapies for such fungal infections. The data indicate that there were multiple recent introductions of C. gattii into the PNW. Public health vigilance is warranted for emergence in regions where C. gattii is not thought to be endemic.
Cryptococcus gattii, an environmental basidiomycetous yeast and the primary cause of cryptococcosis in immunocompetent patients, has been previously primarily associated with central nervous system disease in tropical and subtropical regions of the world (1). This distribution and disease phenotype are in contrast with those of another causative agent of cryptococcosis, Cryptococcus neoformans, which is primarily an opportunistic pathogen of immunosuppressed patients and is distributed worldwide. In addition, C. neoformans is classically found in bird excreta and decaying wood from a wide variety of trees (2–4), whereas C. gattii is primarily found growing in tree bark, tree detritus, and decaying wood. C. gattii has been closely associated with a number of trees, notably Eucalyptus species, which were thought to be the dominant environmental source allowing for transportation of C. gattii out of Australia where it is endemic to locales, such as California, Greece, and elsewhere (5, 6). It has since been shown to be endemic in South America where it is found in close association with a large number of tropical trees (2, 7–9). C. gattii is currently classified into four distinct populations, referred to previously as “major molecular types”: VGI (AFLP4), VGII (AFLP6A), VGIII (AFLP5), and VGV (AFLP7) (10), based on multilocus sequence typing (MLST), amplified fragment length polymorphism (AFLP), and other subgenomic genotyping methodologies. Each of these populations or molecular types can be divided further into distinct lineages or subtypes, as defined by AFLP profiling or MLST sequence typing, which identify individual clonal lineages within each molecular type.

In the late 1990s, infections caused by C. gattii molecular type VGII were reported on Vancouver Island in British Columbia, Canada (11–13); until this time, reports of C. gattii infection were uncommon in North America (1). Two clonal subtypes within the VGII molecular type known as VGIIa (AFLP6A) and VGIIb (AFLP6B) were identified as the cause of the outbreak, which subsequently spread to the Canadian mainland (14) and other regions of the Pacific Northwest (PNW) (15, 16). Within a decade, an additional novel VGII subtype, VGIIc, appeared in the PNW, primarily in the states of Oregon and Washington (16). Given the genomic similarity of isolates within each subtype and finding only a single mating type locus, all three subtypes were thought to be distinct clones (17). While all three PNW subtypes of C. gattii display a shared novel biome preference (presence in a temperate versus tropical climate), only VGIIb maintains the classical clinical features of neurologic dominance (18). The VGIIa and VGIIc subtypes, on the other hand, share distinct novel phenotypes (dominantly pulmonary disease and increased virulence, both in vitro and in vivo) (18).

Given their geographic proximity, time frame of emergence, and genotypic and phenotypic similarity, it was proposed that VGIIa and VGIIc originated relatively recently from a common ancestor (15, 19). Several studies were conducted to better understand the genotypic and phylogenetic nature of the C. gattii strains associated with these outbreaks (11, 13, 15, 17). VGIIa and VGIIb strains have now been found in other non-North American locales, and genotyping studies have provided evidence that these strains emerged independently from a South American source (9). The VGIIc subtype has not been identified elsewhere thus far.

A number of studies have described in vitro and in vivo clinical phenotypic variances (i.e., virulence in animal model, clinical presentation, etc.) between the emergent North American strains (15, 18, 20). In animal models of infection, VGIIa and VGIIc strains have been shown to be more virulent than VGIIb and other C. gattii molecular types (9, 21). It is also clear that the virulence profile within subtype VGIIa does not remain consistent globally with isolates of lesser virulence being described (13, 21; C. Firacative and W. Meyer, unpublished data).

Niche adaptation and phenotypic changes in pathogenic fungi are known to happen through a series of evolutionary mechanisms, including gene duplication, sexual recombination, base mutation, and to some degree, horizontal gene transfer (22). The potential genetic causes for the phenotypic and niche differences in the PNW C. gattii strains are not well understood, although potential contributing factors have been deduced from targeted expression analyses (20, 23). Whole-genome analyses provide a more comprehensive view of phylogenetic relationships and genomic evolution. Here we apply population-level whole-genome approach with the following aims: (i) to more thoroughly understand the population genomics of the C. gattii VGII molecular type; (ii) to explore the emergence of its subtypes in North America; and (iii) to identify possible genetic causes for its changing phenotype.

RESULTS
Whole-genome sequencing and SNP analysis. Whole genomes from 118 Cryptococcus gattii isolates, representing both the Pacific Northwest (PNW) subtypes and the known global VGII molecular type diversity, along with isolates of each of the other major C. gattii molecular types (VGI, VGIII, and VGIV), were sequenced and assembled de novo, using data from multiple next-generation sequencers (all Illumina technology). We obtained at least 10× coverage across an average of 98.27% of the genome (mode, 98.64%; range, 90.63% to 99.43%) compared to the assembled R265 reference (VGIIa), with an average depth of 111× (range, 24× to 532×). The average assembly size was ~17.55 Mb (range, 17.36 to 17.83 Mb).

We determined the presence of single nucleotide polymorphisms (SNPs) across and between subtypes. A total of 1,282,876 SNPs were identified from all genomes, including the VGI, VGII, VGIII, and VGIV molecular types; 544,881 SNPs were parsimony informative, that is they distinguished between more than one strain and were not autapomorphic (Fig. 1). Within the 115 VGII genomes, 310,969 SNPs were found, with 221,248 being parsimony informative (Fig. 2). Consistent with recently emerged and low-diversity clones, only 717, 246, and 528 SNPs were identified within the VGIIa, VGIIc, and VGIIb populations, respectively (Fig. 3, Fig. 4, and Fig. 5, respectively). A combined analysis of the three PNW clones identified little to no homoplasy (data not shown), as previously described (17).

Genome sequence data analysis. Population-level genome analyses were initiated with an SNP-based assessment of phylogenetics and population structure, followed by gene presence/absence comparisons to understand the genetic relationships between and among the VGII subtypes in the PNW and other geographically diverse subtypes. A detailed description of the data analysis flowchart is shown in Fig. S1 in the supplemental material.

Phylogenetics. Isolates chosen for phylogenetic analysis were representative of the global diversity of C. gattii VGII based on their multilocus sequence type (MLST), including a number of isolates from each of the PNW lineages (VGIIa, VGIIb, and VGIIc). Additionally, isolates of the other three C. gattii popula-
tions were analyzed. Based on whole-genome SNP typing (WGST), the VGII strains were clearly separated from other molecular types (i.e., VGI, VGIII, and VGIV), with greater than 500,000 SNPs between VGII and any other major molecular type (Fig. 1).

Using maximum-parsimony analysis for WGST, the VGII isolates separated mostly along the lines of their MLST (Table 1; see Fig. S2 and S3 in the supplemental material). A number of isolates with distinct sequence types (STs) were clearly more related to each other and likely represent members of the same clone with single or limited nucleotide variation within the MLST markers (e.g., VGIIc includes ST6 and ST49). Multiple clonal lineages are seen in the whole VGII phylogeny, including the PNW lineages (Fig. 2 and Fig. S2), and the long terminal branches indicate substantial genetic distance between most lineages. The maximum-parsimony tree (Fig. S2) had a relatively low consistency index of 0.4, indicating either a high degree of homoplasy from multiple mutations at the same sites or that VGII is not evolving in a strictly tree-like manner. To better visualize the relationships between lineages, we developed a neighbor-joining phylogenetic split tree network, or “neighbor net” (Fig. 2). The network contained many phylogenetic splits, which could not be condensed into a single tree phylogeny. This finding may indicate that sexual recombination events between different subtypes are contributing to the genetic diversity of VGII, as these events will cause additional splits due to different regions of the genome being best described by different trees.

When the phi test for recombination was performed using the SNP data, the test indicated that recombination was present with a very high level of significance ($P = 0.0$). We then examined the phi value across the genome to see whether recombination could be localized to a particular subset of the genome. Using nonoverlapping windows of 500 concatenated SNP positions, we tested for recombination separately in 673 “windows” across the genome (encompassing variable genome window sizes due to the nonuniform density of SNPs in the genomes). Recombination was detected at a significant level in 631 of these windows when the $P$ value cutoff was adjusted for making 673 comparisons ($P < 7.2^{-12}$), indicating that the recombination signal was common across most of the genome and not localized to a single region.

Due to the evidence for recombination among isolates, a model-based Bayesian approach (24) for identifying population structure and individual assignment to populations was added to our phylogenetic analyses. fineStructure analysis placed the 32 distinct VGII subtypes, as identified by WGS and MLST (see Fig. S2 and S3 in the supplemental material), into 11 “populations” or groups related by shared genomic regions, with some of these containing only a single representative (Fig. 6). For several of
these groups, members shared many genome regions with members from other groups, indicating incomplete or recent separation between groups. This finding is also supported by the presence of isolates, such as B9764 and WM 05-525, which apparently share genetic material with several populations and may represent admixed individuals. Importantly, all three of the PNW subtypes were placed into separate fineStructure "populations," all of which include members from South America. The neighbor net network (Fig. 2), maximum parsimony-based WGST (Fig. S2), and MLST (Fig. S3) analyses all demonstrate broad dispersal of subtypes from South America throughout the VGII population.

The members of the PNW subtypes (i.e., VGIIa, VGIIc, and VGIIb) showed strict clonality within their respective subtype, with consistency indices (CI) of 1.0 for each subtype tree (Fig. 3 to 5), and significant separation from each other, as described previously (17). Bioinformatic attempts to determine a date for the most recent common ancestor of each subtype did not yield useful estimates (analysis not shown), which may be due to the small amount of SNP data available or potentially the nonclocklike evolution of C. gattii.

The VGIIa subtype contained two distinct clades, a North American-dominated group and a South American group (Fig. 3). The North American VGIIa clade showed little topological structure with a 15-branch polytomy consistent with a radiation emanating from a recent common ancestor. One subclade, joined by 13 shared SNPs, contained a set of three isolates previously described as having lower virulence (13, 20), one of which was isolated in Brazil in 1981 and was genomically separated from the

*FIG 2*  Phylogenetic network of *C. gattii* VGII. A neighbor net representation of the relationships among all VGII subtypes based on WGST SNP data, using the uncorrected p-distance transformation. Each band of parallel edges indicates a split. For example, the two splits that group ST33 with two distinct groups of other VGII subtypes are labeled with red A and B letters. The split joining VGIIc with three additional STs is labeled with a red C letter.
others by 261 additional SNPs. The other two members of this lower-virulence subclade were isolated in Washington State in 1972 and California in 1990 and had SNP branch lengths of 0 and 52, respectively. The VGIIc population was also strictly clonal (CI = 1.0) with a total of 55 parsimony informative SNPs (Fig. 4). This subtype phylogeny contained a 5-branched polytomy but with several distinct subclades. The branch lengths were short (~16 SNPs); notably, all of these isolates were collected between 2008 and 2012.

Unlike VGIIa and VGIIc, the VGIIb subtype is dispersed beyond the Western Hemisphere (6, 13, 15, 25). This subtype contained 528 SNPs (147 parsimony informative) and multiple distinct clades (Fig. 5). North American isolates were found in multiple locations across the tree, including in one large derived clade where they shared membership with isolates that were collected in Asia and Australia. The Asian-Australian/North American subclade shows polytomy, which is consistent with the presence of a genetic bottleneck and subsequent radiation, or may be a reflection of limited sampling. A single isolate from Florida shared membership in a separate basal clade with an isolate from Australia.

**Gene content.** *Ab initio* gene prediction using the pretrained Augustus software identified a median of 6,276 genes per genome (range, 6,213 to 6,359). The gene content analysis initially focused on differences and similarities between the PNW subtypes. The use of BLAT score ratio (BSR) for this analysis allowed us to make gene predictions from the genome sequences, where exon-coding DNA sequences were concatenated into single contiguous sequences to create the *in silico* coding DNA (is-cDNA), the presence or absence of which could then be scored for the PNW genomes. The gene sequences from the three PNW subtypes clustered into 6,302 is-cDNAs. The gene sequences from the remaining 29 non-PNW subtypes clustered into 6,541 is-cDNAs. The individual and shared gene differences in the PNW subtypes were scored against representative genomes from the other known global VGII subtypes. Additionally, we surveyed for the presence or absence of specific genes from any identified PNW clone (e.g., PNW-VGIIa) within the representative genomes from the other *C. gattii* major molecular types (VGI, VGIII, and VGIIV), as well as within the *C. neoformans* var. grubii and *C. neoformans* var. neoformans reference genomes (Table 2). A heat map was developed (Fig. 7) to visualize the gene presence, absence, partial presence, and/or presence of genes with low sequence similarity. We identified several gene content similarities and differences in all of the comparisons (Table 2). Many of the identified putative genes, or is-cDNAs, were located in contiguous blocks of assembled genome sequence, labeled clusters 1 to 7 in Fig. 7. When the PNW clones were compared with each other, VGIIa had the largest number of PNW clone-specific gene content differences. Additionally, the PNW VGIIa-specific genes were typically absent in...
the other VGII subtypes and non-VGII genomes. Conversely, the PNW VGIIb-specific and VGIIc-specific genes (i.e., those genes found only in their respective subtype in the PNW) were commonly found in the other VGII subtypes and non-VGII C. gattii genomes. A large number of genes that were present in both VGIIa and VGIIc were absent from VGIIb. These genes, most of which were located adjacent to each other in clusters 1 and 2, were found in nearly all other C. gattii genomes, indicating a likely gene loss event in the PNW VGIIb population. An analysis of genes common to all three PNW populations but absent in other VGII subtypes identified a three-gene cluster (“cluster 6”) that was found to be absent in most C. gattii genomes but was present, with a high degree of similarity, in the C. neoformans var. grubii genome (Fig. 7).

A comparison of phylogenies based on linked SNPs and gene content showed dissimilar relationships among VGII subtypes (data not shown). This suggests that differential gene content has not become fixed within the VGII lineages. As such, placement of an isolate in a phylogeny or within a subtype using SNP data is not necessarily a good indicator of its gene content, and conversely, gene content is not a good indicator of phylogenetic placement.

**DISCUSSION**

The emergence of C. gattii in temperate British Columbia in Canada and the U.S. PNW was unexpected in that C. gattii was previously thought to be prevalent only in tropical and semi-tropical regions. There has since been much speculation on the origins of the outbreak strains (9, 12, 15, 16). In this study, whole-genome sequencing was performed on 118 C. gattii isolates, primarily of the molecular type VGII, to better ascertain the source and genomic adaptations behind its emergence in the PNW. Our primary strategy was to analyze SNP mutations and gene content and to compare the PNW subtype genomes with a genomically diverse global collection of isolates. This study shows the following: (i) There were most likely multiple distinct North American introductions of at least two of the three PNW VGII subtypes. (ii) Gene content varies between the VGII subtypes. (iii) C. gattii has developed a number of evolutionary strategies that allow for continual niche adaptation. This study also explores the relationship of the North American VGII subtypes to a global C. gattii collection and the potential role of unique genes in virulence and niche adaption.
Out of South America? (phylogenetic analyses). Analysis of the population genome structure of the PNW VGII clones provides additional significant evidence of multiple and continued emergence of apparently regionally fit lineages from within VGII originating in South America, as has been proposed previously (9, 26). As the apparent level of recombination across VGII negates the use of a single SNP-based tree to understand the true phylogenetic relationships among all VGII members, we conducted a neighbor net analysis, providing a better understanding of relationships within the phylogenetic network. For example, the Australian ST33 subtype, which appears most genetically distant on the WGST SNP tree (see Fig. S2 in the supplemental material), still appears most genetically distant in the network; however, there are several distinct splits uniting this subtype with other subtypes (see splits A and B on Fig. 2). The ST33 lineage, therefore, may have continued to exchange genetic material with these taxa after its derivation from the bulk of currently sampled VGII taxa.

From within the genomic diversity of VGII, we can see the emergence of the three low-diversity and highly fit clones (i.e., VGIIa, VGIIb, and VGIIc) that have become prevalent in North America. The emergence of these three clonal subtypes clearly represents at least three distinct events. Our phylogenetic analyses suggest that the phenotypically differentiated VGIIa and VGIIc lineages evolved in South America and spread to the PNW. Additionally, the widely disseminated VGIIb lineage and the VGIIa subtype were likely exported, possibly out of South America, on multiple occasions, and the VGIIb subtype has emerged in multiple geographic locations.

(i) VGIIa. The likely most recent translocation of the VGIIa lineage, as inferred from the very short branch lengths within its respective clade (Fig. 3), resulted in its distinct emergence on Vancouver Island, spreading to the Canadian mainland and into Northwest United States. There is a distinct separation of the PNW dominant VGIIa clade and a clade of South American origin (i.e., Brazil, Argentina, and Japan [in a Brazilian emigrant]) strains. The two VGIIa clades likely have a recent common ancestor, and the lack of its presence elsewhere suggests that this subtype originated in South America. Additionally, WGST allowed us to identify genetic separation between the strains identified during the recent emergence on Vancouver Island and the subclade composed of the older, but less virulent, North American VGIIa strains: NIH-444 (1972; Washington) and CBS7750 (1990; Cali-
| Genotype | Strain  | ST | Country  | Source      | Yr of isolation | Figure inclusion |
|----------|---------|----|----------|-------------|-----------------|-----------------|
| VGI      | WM 179  | 51 | Australia| Clinical    | 1993            | 1, 6, 7         |
| VGIa     | B6864   | 20 | USA      | Clinical    | 2004            | 2, 3, S2, S3    |
|          | B7395   | 20 | USA      | Veterinary  | 2008            | 2, 3, S2, S3    |
|          | B7422   | 20 | USA      | Veterinary  | 2009            | 2, 3, S2, S3    |
|          | B7436   | 20 | USA      | Veterinary  | 2009            | 2, 3, S2, S3    |
|          | B7467   | 20 | USA      | Veterinary  | 2009            | 2, 3, S2, S3    |
|          | B8555   | 20 | USA      | Clinical    | 2006            | 2, 3, S2, S3    |
|          | B8377   | 20 | Canada   | Veterinary  | 2009            | 2, 3, S2, S3    |
|          | B8793   | 20 | USA      | Veterinary  | 2010            | 2, 3, S2, S3    |
|          | B8849   | 20 | USA      | Veterinary  | 2010            | 2, 3, S2, S3    |
|          | B9457   | 20 | USA      | Clinical    | 2011            | 2, 3, S2, S3    |
|          | B9458   | 20 | USA      | Clinical    | 2011            | 2, 3, S2, S3    |
|          | B9757   | 20 | USA      | Veterinary  | 2003            | 2, 3, S2, S3    |
|          | B9759   | 20 | USA      | Veterinary  | ?               | 2, 3, S2, S3    |
|          | CA-1014 | 20 | USA      | Environmental | 1990         | 2, 3, S2, S3    |
|          | CBS 7750| 20 | USA      | Clinical    | 2001            | 1, 2, 3, 6, 7, S2, S3 |
|          | R265    | 20 | Canada   | Clinical    | 2002            | 2, 3, S2, S3    |
|          | HL_A1   | 20 | Canada   | Veterinary  | 2002            | 2, 3, S2, S3    |
|          | HL_A3   | 20 | Canada   | Environmental| 2002         | 2, 3, S2, S3    |
|          | HL_A11  | 20 | Canada   | Clinical    | 2003            | 2, 3, S2, S3    |
|          | HL_B5   | 20 | Canada   | Clinical    | 2002            | 2, 3, S2, S3    |
|          | HL_B6   | 20 | Denmark  | Clinical    | 2006            | 2, 3, S2, S3    |
|          | NIH 444 | 20 | USA      | Clinical    | 1975            | 2, 3, S2, S3    |
|          | WM 03.697| 20 | Canada    | Veterinary  | 2001            | 2, 3, S2, S3    |
|          | WM 05.432| 20 | Japan-Brazil | Clinical  | 2000            | 2, 3, S2, S3    |
|          | WM 05.554| 20 | Brazil    | Clinical    | 2002            | 2, 3, S2, S3    |
|          | WM 06.10 | 20 | Argentina | Clinical    | 2000            | 2, 3, S2, S3    |
|          | ICB-107 | 252| Brazil    | Clinical    | 1981            | 2, 3, S2, S3    |
| VGIb     | B7394   | 7  | USA      | Veterinary  | 2008            | 2, 5, S2, S3, S4 |
|          | B7735   | 7  | USA      | Clinical    | 2009            | 2, 5, S2, S3, S4 |
|          | B8554   | 7  | USA      | Veterinary  | 2008            | 2, 5, S2, S3, S4 |
|          | B8828   | 7  | USA      | Veterinary  | 2010            | 2, 5, S2, S3, S4 |
|          | B9157   | 7  | USA      | Veterinary  | 2011            | 2, 5, S2, S3, S4 |
|          | B9552   | 7  | USA      | Veterinary  | 2011            | 2, 5, S2, S3, S4 |
|          | B9563   | 7  | USA      | Veterinary  | 2011            | 2, 5, 6, 7, S2, S3, S4 |
|          | B9588   | 7  | USA      | Clinical    | 2012            | 2, 5, S2, S3, S4 |
|          | B9758   | 7  | USA      | Environmental| 2002         | 2, 5, S2, S3, S4 |
|          | CDC R272| 7  | Canada   | Clinical    | 2001            | 2, 5, S2, S3, S4 |
|          | WM 2552 | 7  | Malaysia | Clinical    | 1997            | 2, 5, S2, S3, S4 |
|          | WM 03.27| 7  | Australia| Environmental| 1992         | 2, 5, S2, S3, S4 |
|          | WM 04.71| 7  | Australia| Veterinary  | 1991            | 2, 5, S2, S3, S4 |
|          | WM 05.465| 7  | Brazil   | Clinical    | 1997            | 2, 5, S2, S3, S4 |
|          | WM 06.634| 7  | Thailand | Clinical    | 1994            | 2, 5, S2, S3, S4 |
|          | WM 06.636| 7  | Thailand | Clinical    | 1995            | 2, 5, S2, S3, S4 |
|          | WM 04.75 | 30 | Thailand | Clinical    | 1993            | 2, 5, S2, S3, S4 |
| VGIc     | B6863   | 6  | USA      | Clinical    | 2005            | 2, 4, S2, S3    |
|          | B7432   | 6  | USA      | Clinical    | 2009            | 2, 4, S2, S3    |
|          | B7434   | 6  | USA      | Clinical    | 2008            | 2, 4, S2, S3    |
|          | B7466   | 6  | USA      | Veterinary  | 2008            | 2, 4, S2, S3    |
|          | B7491   | 6  | USA      | Clinical    | 2009            | 2, 4, S2, S3    |
|          | B7493   | 6  | USA      | Veterinary  | 2009            | 2, 4, S2, S3    |
|          | B7641   | 6  | USA      | Veterinary  | 2008            | 2, 4, S2, S3    |
|          | B7737   | 6  | USA      | Clinical    | 2009            | 2, 4, S2, S3    |
|          | B7765   | 6  | USA      | Veterinary  | 2009            | 2, 4, S2, S3    |
|          | B8210   | 6  | USA      | Clinical    | 2008            | 2, 4, S2, S3    |
|          | B8214   | 6  | USA      | Clinical    | 2009            | 2, 4, S2, S3    |
|          | B8510   | 6  | USA      | Clinical    | 2009            | 2, 4, S2, S3    |
|          | B8571   | 6  | USA      | Clinical    | 2009            | 2, 4, S2, S3    |
|          | B8788   | 6  | USA      | Clinical    | 2010            | 2, 4, S2, S3    |
|          | B8798   | 6  | USA      | Clinical    | 2005            | 2, 4, S2, S3    |

(Continued on following page)
| Genotype | Strain | ST | Country | Source | Yr of isolation | Figure inclusion |
|----------|--------|----|---------|--------|----------------|-----------------|
| B8825    | 6      | USA| Clinical| 2010   | 1, 2, 4, S2, S3 |
| B8833    | 6      | USA| Clinical| 2010   | 2, 4, S2, S3   |
| B8838    | 6      | USA| Clinical| 2010   | 2, 4, S2, S3   |
| B8843    | 6      | USA| Clinical| 2010   | 2, 4, S2, S3   |
| B9816    | 6      | USA| Veterinary| 2012 | 2, 4, S2, S3 |
| B9933    | 6      | USA| Clinical| 2012   | 2, 4, S2, S3   |
| GT-11-7650 | 6 | USA| Veterinary| ? | 2, 4, S2, S3 |
| HL_B3    | 6      | USA| Clinical| 2005   | 2, 4, S2, S3   |
| HL_B4    | 6      | USA| Clinical| 2008   | 2, 4, S2, S3   |
| B7390    | 49     | USA| Clinical| 2008   | 2, 4, S2, S3   |
| VGII     | B8973  | 50 | USA      | Clinical | 2010 | 1, 2, 6, 7, S2, S3 |
| B9764    | 21     | USA| Veterinary| 2012 | 1, 2, 6, 7, S2, S3 |
| IAL-3225 | 264    | Brazil| Veterinary| 1994 | 1, 2, 6, 7, S2, S3 |
| IAL-3234 | 137    | Brazil| Veterinary| 1998 | 1, 2, 6, 7, S2, S3 |
| IAL-3243 | 283    | Brazil| Veterinary| 2000 | 1, 2, 6, 7, S2, S3 |
| HL_A2    | 28     | Brazil| Clinical| Pre-2008 | 2, S2, S3 |
| HL_A4    | 3      | Uruguay| Environmental| 1996 | 1, 2, 6, 7, S2, S3 |
| HL_A5    | 124    | Brazil| Environmental| Pre-2008 | 1, 2, 6, 7, S2, S3 |
| HL_A6    | 25     | Aruba| Veterinary| 1953 | 2, S2, S3 |
| HL_A8    | 185    | Brazil| Environmental| Pre-2008 | 2, S2, S3 |
| HL_B2    | 248    | Brazil| Veterinary| 2000 | 1, 2, 6, 7, S2, S3 |
| HL_B11   | 35     | Greece| Clinical| 1996 | 1, 2, 6, 7, S2, S3 |
| HL_B12   | 18     | Greece| Clinical| 1998 | 2, S2, S3 |
| WM 178   | 21     | Australia| Clinical| 1991 | 2, S2, S3 |
| WM 1850  | 25     | Venezuela| Clinical| 1999 | 2, S2, S3 |
| WM 1851  | 45     | Venezuela| Clinical| 1999 | 1, 2, 6, 7, S2, S3 |
| WM 3032  | 33     | Australia| Clinical| 1983 | 1, 2, 6, 7, S2, S3 |
| WM 04.78 | 31     | Colombia| Clinical| 1998 | 1, 2, 6, 7, S2, S3 |
| WM 04.84 | 34     | Brazil| Clinical| 1986 | 1, 2, 6, 7, S2, S3 |
| WM 04.274 | 29 | Colombia| Clinical| 2002 | 1, 2, 6, 7, S2, S3 |
| WM 05.275 | 25 | Colombia| Clinical| 2001 | 1, 2, 6, 7, S2, S3 |
| WM 05.339 | 43 | Colombia| Environmental| 2005 | 2, S2, S3 |
| WM 05.342 | 25 | Colombia| Environmental| 2005 | 2, S2, S3 |
| WM 05.419 | 39 | Brazil| Clinical| 1988 | 1, 2, 6, 7, S2, S3 |
| WM 05.452 | 16 | Brazil| Clinical| 1995 | 1, 2, 6, 7, S2, S3 |
| WM 05.456 | 17 | Brazil| Environmental| 1994 | 1, 2, 6, 7, S2, S3 |
| WM 05.457 | 19 | Brazil| Clinical| 1995 | 2, S2, S3 |
| WM 05.461 | 14 | Brazil| Clinical| 1997 | 2, S2, S3 |
| WM 05.462 | 15 | Brazil| Clinical| 1997 | 1, 2, 6, 7, S2, S3 |
| WM 05.525 | 41 | Brazil| Clinical| 1997 | 1, 2, 6, 7, S2, S3 |
| WM 05.527 | 28 | Brazil| Clinical| 1997 | 1, 2, 6, 7, S2, S3 |
| WM 05.528 | 40 | Brazil| Clinical| 2001 | 1, 2, 6, 7, S2, S3 |
| WM 05.529 | 27 | Brazil| Clinical| 1997 | 1, 2, 6, 7, S2, S3 |
| WM 05.530 | 13 | Brazil| Clinical| 1999 | 1, 2, 6, 7, S2, S3 |
| WM 05.533 | 11 | Brazil| Clinical| 1997 | 1, 2, 6, 7, S2, S3 |
| WM 05.536 | 9 | Brazil| Clinical| 1997 | 2, S2, S3 |
| WM 05.545 | 24 | Brazil| Clinical| 2001 | 2, S2, S3 |
| WM 05.546 | 23 | Brazil| Clinical| 2001 | 2, S2, S3 |
| WM 05.547 | 22 | Brazil| Clinical| 2001 | 1, 2, 6, 7, S2, S3 |
| WM 06.12 | 37 | Venezuela| Clinical| 1996 | 2, S2, S3 |
| WM 08.309 | 48 | Australia| Veterinary| 1997 | 1, 2, 6, 7, S2, S3 |
| WM 08.311 | 5 | Australia| Veterinary| 1996 | 1, 2, 6, 7, S2, S3 |
| WM 09.94 | 38 | Australia| Veterinary| 2001 | 1, 2, 6, 7, S2, S3 |
| WM 09.152 | 48 | Australia| Environmental| 2009 | 2, S2, S3 |
| WM 11.65 | 33 | Australia| Veterinary| 2011 | 2, S2, S3 |
| VGII     | WM 11.139 | 143 | USA      | Veterinary| 2011 | 1, 7 |
| VGIV     | WM 779  | 70 | South Africa| Veterinary| 1994 | 1, 7 |

* Cryptococcus neoformans

**VNI H99** 2 USA Clinical 1978 7
**VNIV JEC21** 126 USA Clinical 1997 7

* The strain designation, major molecular type, MLST sequence type, and figures that the strains are shown on are indicated.
This lower-virulence subclade also contained a 1981 Brazilian isolate (ICB-107). It is possible that the latter isolate represents the originating lineage in Brazil and that the two North American isolates represent two distinct translocation events of that same lineage to North America prior to the eventual PNW emergence. Strain ICB-107 was isolated from a patient in Recife, a major port city in Brazil, representing a possible point of export. A lack of epidemiological information associated with that particular isolate prevents further understanding at this time.

(ii) VGIIc. The VGIIc subtype has been found only in the PNW. Its most recent splits (i.e., closest genetic relatives) in the phylogenetic network were found exclusively in South America (see split A in Fig. 2). The very short branch lengths and lack of distinct separate clades within this subtype (Fig. 4) are indicative of an individual emergence of this subtype, likely from South America, where all of the most closely related subtypes were identified.

(iii) VGIIb. VGIIb is the most diverse of the VGII lineages found in the PNW and likely has been translocated to North America and other parts of the world on multiple occasions. The origin of PNW VGIIb strains is not clear. A VGIIb subclade defined by the putative loss of a 12-gene 10-kb contiguous genomic region (see Fig. S4 in the supplemental material; cluster 1 in Fig. 7) contains most, but not all, PNW VGIIb strains along with strains...
| Gene content similarity and is-cDNA | Family hit(s) | Pfam | E value(s) |
|-----------------------------------|--------------|------|------------|
| Present in PNW VGIIc              |              |      |            |
| is-cDNA 1                         | Transmembrane amino acid transporter protein family | PF01490.13 | 1.90E−33   |
| is-cDNA 2                         | Endoribonuclease L-PSP | PF01042.16 | 8.30E−39   |
| is-cDNA 3                         | Aminotransferase class I and II domain | PF00155.16 | 5.20E−47   |
| is-cDNA 4                         | No significant matches | |            |
| is-cDNA 5                         | Zinc finger double domain, fungus-specific transcription factor domain | PF13465.1, PF04082.13 | 4.4E−7, 2.3E−16 |
| is-cDNA 6                         | Major facilitator superfamily | PF07690.11 | 5.80E−11   |
| is-cDNA 7                         | Peptidase family M20/M25/M40, peptidase dimerization domain | PF01546.23, PF07687.9 | 3E−6, 4.2E−9 |
| is-cDNA 8                         | Major facilitator superfamily, fungus-specific transcription factor domain | PF07690.11, PF04082.13 | 5.5E−29, 9.6E−12 |
| is-cDNA 9                         | No significant matches | |            |
| is-cDNA 10                        | Arylsulfotransferase (ASST) family | PF14269.1 | 9.90E−14   |
| is-cDNA 11                        | Major facilitator superfamily | PF07690.11 | 7.20E−23   |
| is-cDNA 12                        | Phosphatidylethanolamine-binding protein domain | PF01161.15 | 5.70E−18   |
| is-cDNA 13                        | Fungus-specific transcription factor domain | PF04082.13 | 2.90E−05   |
| is-cDNA 14                        | Glycosyl hydrolase family 32 N-terminal domain, glycosyl hydrolase family 32 C-terminal domain | PF00251.15, PF08244.7 | 1.8E−51, 4.8E−10 |
| is-cDNA 15                        | No significant matches | |            |
| is-cDNA 16                        | Alcohol dehydrogenase GroES-like domain, zinc-binding dehydrogenase | PF08240.7, PF13602.1 | 8.5E−6, 2.8E−19 |
| is-cDNA 17                        | No significant matches | |            |
| Present in PNW VGIIb              |              |      |            |
| is-cDNA 18                        | E1-E2 ATPase, haloacid dehalogenase-like hydrolase | | 3.6E−47, 3.1E−21 |
| is-cDNA 19                        | No significant matches | |            |
| is-cDNA 20                        | No significant matches | |            |
| Present in PNW VGIIa              |              |      |            |
| is-cDNA 21                        | Sugar (and other) transporter | PF00083.19 | 1.20E−94   |
| is-cDNA 22                        | Alcohol dehydrogenase GroES-like domain, zinc-binding dehydrogenase | PF08240.7, PF00107.21 | 1.6E−28, 1.6E−16 |
| is-cDNA 23                        | Major facilitator superfamily | PF07690.11 | 1.20E−17   |
| is-cDNA 24                        | No significant matches | |            |
| is-cDNA 25                        | Aldo/keto reductase family | PF00248.16 | 1.30E−52   |
| is-cDNA 26                        | Protein of unknown function, putative collagen-binding domain of a collagenase | PF13204.1, PF12904.2 | 2.1E−54, 7.4E−19 |
| is-cDNA 27                        | Sugar (and other) transporter | PF00083.19 | 3.70E−93   |
| is-cDNA 28                        | Alcohol dehydrogenase GroES-like domain, zinc-binding dehydrogenase | PF08240.7, PF00107.21 | 3.5E−28, 1.2E−15 |
| is-cDNA 29                        | Fungus-specific transcription factor domain, fungal Zn(2)-Cys(6) binuclear cluster domain | PF04082.13, PF00172.13 | 2E−16, 1.8E−4 |
| is-cDNA 30                        | No significant matches | |            |
| is-cDNA 31                        | No significant matches | |            |
| is-cDNA 32                        | Aspartyl protease | PF09668.5 | 3.30E−05   |
| is-cDNA 33                        | Helicase conserved C-terminal domain, DEAD/DEAH box helicase | PF00271.26, PF00270.24 | 4.6E−7, 1E−4 |
| Present in PNW VGIIc              |              |      |            |
| is-cDNA 34                        | Fungus-specific transcription factor domain | PF04082.13 | 2.50E−11   |
| is-cDNA 35                        | Lyase, adenylosuccinate lyase C terminus | PF00206.15, PF10397.4 | 1.5E−47, 2.6E−18 |
| is-cDNA 36                        | Sugar (and other) transporter | PF00083.19 | 5.40E−80   |
| is-cDNA 37                        | No significant matches | |            |
| is-cDNA 38                        | No significant matches | |            |
| is-cDNA 39                        | Protein of unknown function | PF11175.3 | 1.40E−80   |
| is-cDNA 40                        | Sugar (and other) transporter | PF00083.19 | 8.30E−63   |
| is-cDNA 41                        | Protein of unknown function | PF11175.3 | 6.70E−83   |

(Continued on following page)
engthaler et al.

Table 2 (Continued)

| Gene content similarity and is-cDNA | Family hit(s) | Pfam | E value(s) |
|-------------------------------------|---------------|------|------------|
| is-cDNA 42                          | No significant matches | PFO03933, PFO14310 | 1.9E−70, 7.5E−20 |
| is-cDNA 43                          | No significant matches | | |

Present in PNW VGIIa/b/c and missing in half of VGII STs

| is-cDNA 44 | Glycosyl hydrolase family 3 N-terminal domain, fibronectin type III-like domain | PF00933, PF014310 | 1.9E−70, 7.5E−20 |
|------------|--------------------------------------------------------------------------------|-----------------|-----------------|
| is-cDNA 45 | Alpha- L-rhamnosidase N-terminal domain, bacterial alpha-L-rhamnosidase | PF08531, PF05592 | 0.7E−36, 9.5E−166 |
| is-cDNA 46 | Sugar (and other) transporter | PF00083 | 2.0E−57 |
| is-cDNA 47 | No significant matches | | |

Missing in PNW and present in half of VGII STs

| is-cDNA 48 | Fungus-specific transcription factor domain | PF04082.13 | 1.90E−23 |
| is-cDNA 49 | Proline dehydrogenase | PF01619.13 | 4.60E−31 |
| is-cDNA 50 | Transmembrane amino acid transporter protein | PF01490.13 | 2.8E−17, 1E−9 |

collected in the previous decade in Australia and Asia (6, 25). These strains are the most commonly found VGIIb strains in the PNW and they reportedly are less virulent than the other VGIIb strains found in the PNW (16). This 10-kb contig is absent from the derived clade yet occurs in all other, more-basal VGIIb strains, including two additional isolates found in PNW, which strongly suggests a single gene loss event in a parental strain. This event, again, provides evidence for multiple independent introductions of VGIIb to North America, which are discernible by both SNP states and genetic content. Further studies are necessary to assess the phenotypic impact of this loss of this genomic component.

The maximum possible consistency indices (CI = 1.0) for the VGIIa, VGIIb, and VGIIc phylogenetic trees demonstrate a lack of observable recombination occurring within or among the North America VGII isolates. These data, taken with the fact that all PNW strains contain only the mating type α locus, indicate that there is limited to no sexual reproduction occurring in these strains. Same-sex mating was previously hypothesized as an evolutionary driver of these subtypes (13). As no phylogenetic inconsistencies were identified among strains within each PNW subtype, there is no obvious evidence of same-sex mating in the region as well, although the progeny of such events may not be readily identified due to the genomic similarities of the parental strains. As sexual reproduction is typically necessary for the development of infectious propagules, it is possible that individual isolates undergo unisexual reproduction (selfing) or haploid fruiting, which have been previously described in C. neoformans (27, 28), to induce a yeast-to-hyphal transition.

Searching for North American adaptations (comparative genomics). In order to identify genetic content associated with niche adaptation to the PNW, we explored the population genome content for genes associated with each of the North American VGII subtypes compared to each other and to other C. gattii molecular types. We were able to identify a subtype-specific gene present (compared to other PNW subtypes) in each of the three subtypes, as well as shared genes in VGIIa and VGIIc compared to the VGIIb subtype. These genomic differences could be the basis for specific niche adaptations and possible differences in virulence and tissue tropism (18). None of the PNW VGIIa, VGIIb, and/or VGIIc clone-defining genes or gene regions was completely unique to North American populations; all were identified in at least once in other global VGII genotypes analyzed.

(i) Notable gene differences. The gene regions specific to VGIIa in the PNW were rarely seen in other global subtypes and may represent greater likelihood of novel genes. The majority of these VGIIa-specific regions were in a contiguous gene cluster (gene cluster 3, which contains is-cDNA 21 to 29, in Fig. 7) that was seen only in three other global subtypes, originating from Brazil, Colombia, and Australia. This gene cluster contains several genes associated with sugar transport and alcohol dehydrogenase and is possibly important for regional plant sugar metabolism (Table 2). This cluster also includes a collagen-binding domain, likely from a collagenase enzyme, which has been identified in other pathogenic fungi (29). Collagen is a component of all major lung cells (30), and collagen binding in the lung has been associated with colonization of Aspergillus in the asthmatic lung (31) and increased virulence in other pulmonary infections (32). The presence of a collagen-binding domain, therefore, may be relevant to binding of cryptococci to host lung cells and contribute to increased virulence or lung tissue tropism or both in the PNW VGIIa infections (18).

Interestingly, the VGIIa cluster 3 gene cluster is also absent from the other major C. gattii types (VGI, VGIII, and VGIV), except for variants of two of the sugar transport/metabolic domains that were present in the VGI genome, possibly an indication of a historical gene transfer event. Additionally, this gene cluster was absent from the C. neoformans var. grubii and C. neoformans var. neoformans reference genomes (Fig. 7).
An additional gene complex found in the VGIIa group, a DEAD box RNA helicase (is-cDNA 33), was rarely seen in VGII subtypes outside the PNW. Semihomologous genes or gene fragments were identified in VGIIc and four additional VGII subtypes and in the VGIII genome, as well as in the *C. neoformans* var. *grubii* genome. The DEAD box RNA helicase has been found previously in multiple eukaryotic and prokaryotic genomes and was recently established as an important virulence factor in *C. neoformans*, with reported impacts on environmental sensing, pathogenesis, and RNA metabolism (33). Additionally, RNA helicases have been associated with fungal cold response (34). As such, its presence/partial presence in VGIIa and VGIIc and relatively few other VGII lineages may suggest a linkage to differential virulence and/or temperate climate phenotype. Its absence from most other *C. gatt.*
An analysis of the genome content of the three North American VGII subtypes against all other VGII genomes in the current study revealed an ~10-kb region, containing three genes, that was consistently found in all VGIIa, VGIIb, and VGIIc genomes. This cluster (cluster 6, which contains is-cDNA 44 to 46, in Fig. 7) was present in only a limited number of other VGII subtypes and was absent from the representative VGI, VGIII, and VGIV strains. BSR analysis revealed that the region had high similarity with a homologous region of the same size in the C. neoformans var. grubii reference genome but was missing from C. neoformans var. neoformans. The three genes in this cluster were identified as encoding rhamnosidase, avenacinase, and an unspecified sugar transporter. Rhamnosidase has been previously identified in Cryptococcus albidos (35), Aspergillus spp. (36), and other basidiomycetes and ascomycetes (37). The presence of rhamnosidase may enable more efficient use of rhamnose as a carbohydrate source. Of particular note, rhamnose is a free sugar produced by Rhamnus (buckthorn), a plant with a large distribution in the PNW. Rhamnose is also found in members of the Uncaria family, a genus found throughout the tropical region. It is a component of many plant and bacterial polysaccharides, as well as plant glycodies. Additionally, rhamnosidase in soft-rot fungi has been shown to have lignocellulolytic activity, which is involved in the degradation of hardwoods (38). Its presence in the PNW population may also provide an adaptive advantage to old-growth forests of the region, as C. gattii has been recovered from rotting and decaying wood in other geographic areas (39, 40).

The genes found in VGIIc, but not VGIIa or VGIIb (is-cDNA 34 to 41), relate primarily to sugar transport, metabolism, and gene expression and are commonly seen in other VGII subtypes and non-VGII cryptococcal genomes, suggesting that the lack of homologous genes in VGIIa and VGIIb represents independent gene loss events.

Among the three gene regions found only in the North American VGIIb strains is a haloacid dehalogenase-like hydrolase (is-cDNA 18). This gene was rather infrequent in other VGII subtypes, although a highly homologous region was seen in VGI, VGIII, and VGIV. A haloacid dehalogenase was found to be a virulence factor in Paracoccidioides brasiliensis, necessary for lung epithelial cell adhesion (41). No function for this gene region has been established in Cryptococcus.

The majority of all other between-PNW subtype genetic differences include genes or domains that are seen relatively frequently in other VGII subtypes. The presence or absence of these genes will need to be further explored to understand any functional role associated with the PNW populations.

The 12-gene cluster that was absent in the derived clade of VGIIb (cluster 1, which contains is-cDNA 1 to 12, in Fig. 7; see Fig. S4 in the supplemental material) was present in all other VGIIb isolates, including the separate clade found in North America, and was found, wholly or partially, in all other Cryptococcus isolates. This gene complement is composed of structural, regulatory, transport, and repair genes. The significance of this deletion remains unknown, but it is a marker of the derived lower-virulence clade of VGIIb. Phenotypic studies are necessary to determine whether there are other differences between the derived clade and the rest of the VGIIb subtype.

(ii) Evidence of recombination. Beyond the clear differences of gene content between and among VGII subtypes, multiple lines of evidence demonstrate the likelihood of “genome sharing” throughout the VGII population. In addition to the “all VGII” tree consistency index of ~0.4 (meaning a high degree of homoplasy) (see Fig. S2 in the supplemental material), and a significant phi test for recombination (\(P = 0.0\)), it is evident that gene content does not dictate phylogeny. We assume that recombination throughout the VGII population accounts for the high degree of homoplasy, which may be accomplished through larger genome content sharing events (i.e., bisexual or same-sex mating [19]). While the neighbor net phylogenetic network recovers a mostly star-like shape, there are apparent relationships between lineages (Fig. 2). The fineStructure analysis conducted here assigned lineages to specific “populations” and assessed levels of relatedness between genomes using adjacent linked SNPs to identify the nearest neighbor relationships for portions of the genome bounded by recombination events. Evidence of recombination was identified in South American isolates. For example, isolates WM 05.275 (originating in Colombia) and WM 05.529 (originating in Brazil) share many of their linked loci, although they are distinct subtypes and are separated by WGST, and they have distinct gene content by BSR. The same pattern is seen with isolates WM 05.419 and WM 05.547, both Brazilian clinical isolates. The VGIIa (R265 [British Columbia, Canada]) and VGIIb (B9563 [Washington]) representatives showed clear evidence of limited genome sharing with most isolates in the analysis. Conversely, the VGIIc genome showed almost no sharing with the vast majority of other VGII subtypes, except its nearest phylogenetic relatives from Brazil (samples WM 05.462 and IAL-3234). Taken together, these results indicate a relatively weak population structure within VGII or a recent subdivision of populations that have not yet allowed gene absence and presence to become fixed within recombining groups. Again, the high consistency indices for whole-genome SNP phylogenies of each PNW subtype (Fig. 3 to 5) indicate a lack of observable recombination occurring within these clones.

The presence of the three-gene cluster with a high degree of homology to C. neoformans var. grubii (cluster 6) provides evidence of a possible previous hybridization event between C. gattii and C. neoformans var. grubii, resulting in the possible introgression of this cluster in either a limited number of C. gattii subtypes or in the C. neoformans var. grubii genome. More whole-genome sequencing of the latter will be needed to ascertain the direction of introgression. Natural C. gattii and C. neoformans var. grubii hybrids have been recently identified from clinical samples in Colombia and Brazil (42). Additionally, genetic remnants of previous hybridization events, also referred to as “identity islands,” have been reported in the case of a nonreciprocal exchange of a multigene cluster from C. neoformans var. grubii to C. neoformans var. neoformans (43). The existence of hybrids and subsequent introgression are an additional mechanism of genome evolution (beyond recombination, SNP mutation, and gene gain, loss, and duplication) and may allow continued successful emergence of novel phenotypes.

(iii) Conclusions. No clear progenitor strains have been identified for VGIIa, VGIIb, or VGIIc, although the first two have been found in South America. The phylogenetic evidence shown here is consistent with multiple translocations of VGIIb to the PNW and elsewhere, as well as possible multiple introductions of VGIIa into North America. The VGIIc subtype lacks representation outside...
the PNW; however, its consistent SNP and gene content relationship with subtypes from Brazil provides compelling evidence of a likely South American origin for VGIIc as well. The results of the analysis also confirm that South America is a major source of VGII diversity and, with numerous subtypes and extensive recombination, it appears to represent an evolutionary accelerator for the VGII subtype and clones emerging from it.

Our extensive whole-population genome analysis enabled an in silico search for genetic factors associated with differing phenotypes in the PNW populations; similar in silico genome-wide association studies (GWAS) have been useful for studying adaptations in other widely dispersed fungi (34). The initial SNP and gene analyses based on this population-level study provide an important perspective on C. gattii diversity and will serve as a baseline for future studies. The identification of variant sugar transport and metabolism genes, as well as previously unexplored virulence genes (e.g., collagen adhesion gene in VGIIa), provides excellent candidates to pursue in functional studies that will help explain the PNW outbreaks and other C. gattii emergence events. The comprehensive SNP analyses will also be useful for future mutation-based functional analyses. Given the obvious gene content differences seen within VGIIb, even demonstrably clonal subtypes are not uniform and may require intrasubtype comparative analyses.

We recognize that limitations of this study include sampling bias and hence possible overrepresentation of isolates from some geographic regions and underrepresentation from others. There could be further diversity of VGII in other regions that were not sampled or that were sparsely sampled. Additionally, given the finding of C. gattii VGII strains in undisturbed Amazon rainforest and the amount of diversity already seen in South America (7), there is doubtless much more unsampled diversity on that continent. Additionally, this study was designed to identify important genomic features from whole-genome sequence data and was not designed to identify important mutations impacting gene regulation or expression. Nevertheless, our findings provide a significant advance in our understanding of the emergence of C. gattii in the PNW though the importance of the gene differences is unknown and requires further functional genomic analyses. As has been seen in recent years, C. gattii continues to spread to new locales, and a comprehensive understanding of its population genomics may help researchers, public health officials, and clinicians to better respond to the impacts of its continued evolution and emergence.

**MATERIALS AND METHODS**

*Cryptococcus gattii* isolates and Illumina sequencing. One hundred eighteen C. gattii isolates were included in the analyses (Table 1). Information for clinical strains was collected with the approval and oversight of the institutional ethic committee. Additionally, the two near neighboring taxa, *C. neoformans* var. *neoformans* strain JEC21 and *C. neoformans* var. *grubii* strain H99, were obtained as whole-genome assemblies from NCBI (http://www.ncbi.nlm.nih.gov/Genbank/index.html) and the Broad Institute of Harvard and MIT (http://www.broadinstitute.org/annotation/genome/cryptococcus_neoformans_b/index.html) respectively. The 118 C. gattii isolates were selected based on previously established MLST sequence types (Table 1) to ensure global genomic diversity. The previously published VGIIa reference genome (R265) (see the Broad Institute link above) was resequenced for this analysis, due to the relatively high number of likely errors in the existing Sanger–based reference genome (17). The genomes of all 118 C. gattii isolates were sequenced using Illumina HiSeq, MiSeq, or GAIIx sequencing technology as previously described (17, 44). High-molecular-weight DNA was extracted using theZR Fungal/Bacterial DNA MiniPrep kit (catalog no. D6005; Zymo Research, Irvine, CA, USA). DNA samples were prepared for paired-end sequencing using the Kapa Biosystems library preparation kit (catalog no. KK8201; Woburn, MA, USA) protocol with an 8-bp index modification. Briefly, 2 μg of double-stranded DNA (dsDNA) sheared to an average size of 600 bp was used for the Kapa Illumina paired-end library preparation as described by the manufacturer. Modified oligonucleotides (Integrated DNA Technologies, Coralville, IA, USA) with 8-bp indexing capability (45), were substituted at the appropriate step. Prior to sequencing, the libraries were quantified using quantitative PCR (qPCR) on the ABI 7900HT (Life Technologies Corporation, Carlsbad, CA, USA) using the Kapa library quantification kit (catalog no. KK4835; Woburn, MA, USA). Libraries were sequenced to a read length of 100 bp or 150 bp on the Illumina HiSeq system, to 75 bp or 100 bp on the GAIIx system, or to 250 bp on the MiSeq system. All WGS data files have been deposited in the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/BIoproject; project PRJNA249927).

**SNP detection and phylogenetic comparison.** (i) Genome assembly. All sequenced genomes were de novo assembled using the SPAdes assembler v2.5.0 (46). Raw contigs were filtered by length and coverage, as SPAdes tends to produce many short and low-coverage contigs. For isolates assembled from 100-bp or 150-bp paired-end data, contigs shorter than 200 bp were filtered and removed, while contigs shorter than 500 bp were filtered/removed in genomes assembled from 250-bp paired-end MiSeq data. Additionally, for every genome, contigs with coverage below 20% of the average coverage of the 20 largest contigs were filtered out. (ii) MLST typing. The seven MLST loci, CAP59, GPD1, LAC1, SOD1, URA5, PLB1, and intergenic sequence (IGS), were aligned to the contigs according to the consensus MLST scheme for the *C. neoformans/C. gattii* species complex (10). Allele types (ATs) and sequence types (STs) were assigned according to the MLST database at http://mlst.mycologylab.org, and new ATs and STs were added to this database. A maximum likelihood tree was constructed based on ST locus scores using the program MEGA v. 5.05 (47). (iii) SNP variant detection. Illumina read data were aligned against the herein newly obtained R265 (high virulent VGIIa Vancouver Island outbreak strain) assembly using Novoalign 0.30.03 (Novocraft Technologies, Selangor, Malaysia), and then SNP variants were identified using the GATK Unified Genotyper v2.4 (48). SNP calls were then filtered (http://gennorth.github.io/NASP/) to remove positions with less than 10× coverage, with less than 80% variant allele calls, or identified by Nucmer as being within duplicated regions in the reference. In the cases of the VGIIb and VGIIc analyses, the read data were aligned by the same methods using the genome assemblies for B9563 and B8825, respectively. In total, five SNP matrices with different taxa were produced: (i) C. gattii molecular types VGII to VGIV; (ii) C. gattii molecular type VGII only with all previously established MLST sequence types; (iii) C. gattii VGIIa; (iv) C. gattii VGIIb; and (v) C. gattii VGIIc. (iv) Phylogenetic analysis. To understand relationships between isolates, we conducted whole-genome SNP typing (WGST) as previously described (17, 39, 49). Maximum- parsimony SNP trees, based on each of the above five SNP matrices, were constructed, using PAUP* v.4b10 (50) and visualized in FigTree v.1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/). The C. gattii VGII tree was rooted using the large C. gattii tree, while the VGIIa, VGIIb, and VGIIc trees were each rooted using the other VGII subtypes as outgroups. Roots are shown simply as the midpoint on the branch determined by outgroup rooting, because the exact number of SNPs assigned to each branch varies slightly with the inclusion of different taxa, due to the SNP filtering strategies employed. The phi test for recombination was conducted on the whole VGII SNP data set, as well as the same data set broken into 500-bp nonoverlapping windows, using the PhiPack software (51). Due to recombination signal strongly present in the VGII SNP data, a neighbor-joining split tree network or “neighbor net” (52) was drawn to visualize the relationship between isolates in the
VGII tree. The neighbor net was drawn using SplitsTree4 (53) with the uncorrected P-distance transformation. The GTR+G model of nucleotide evolution, selected by jModelTest2 (54), was also used to calculate distances for input to the neighbor net algorithm, which produced a network showing the same divisions among samples (data not shown).

(v) ChromoPainter and fineStructure population analysis. In addition to whole-genome SNP phylogenies, fineStructure analysis (25) was performed using SNP matrix 2 (see “SNP variant detection” above) to help identify the population structure within VGII. Using one representative from each of the 32 distinct VGII subtypes (i.e., MLST sequence types) (Table 1), visualized by placement in the SNP phylogeny, whole-genome SNP data were reduced to a pairwise similarity matrix using ChromoPainter, which estimates the number of regions in two genomes that coalesce prior to coalescing with another genome. ChromoPainter uses the linkage disequilibrium present between closely positioned SNP loci to help assign a given region of a genome to its best near neighbor, increasing the sensitivity of population detection over unlinked methods. ChromoPainter was run using the linkage model, assuming uniform rates of recombination per base pair of sequence. Next, the population structure is identified by fineStructure using this similarity matrix, where the underlying model assumes individuals within populations will share more regions of their genome with each other, and have a similar amount of admixture with individuals from different populations. After identifying populations, fineStructure then merges populations one at a time, at each step using the most likely population merge, to relate populations to each other through a tree.

Gene content comparison. (i) BSR analysis of gene content. Potential gene content comparison of taxa was performed using BLAT score ratio (BSR) analysis, as previously described for prokaryotic genomes (55), with modifications made to accommodate eukaryotic genes containing introns (described below). Gene sequences were predicted for each assembled genome using the ab initio eukaryotic gene prediction software Augustus v2.6.1 (56). For each gene prediction, the coding DNA region sequences were concatenated into a single contig, removing introns to create the putative in silico coding DNA (is-cDNA). is-cDNA sequences from all isolates were then clustered and dereplicated using USEARCH v5.0.144 (57) with a 90% identity cutoff, creating a nonredundant set of sequences from all isolates then were clustered and dereplicated using USEARCH v5.0.144 (57) with a 90% identity cutoff, creating a nonredundant set of consensus is-cDNA predictions encompassing the entire diversity of samples. Putative is-cDNAs were then aligned back to each genome assembly using BLAT v35x1 (38), which can align the cDNA sequence against a reference allowing for intervening intron sequences, using an 80% identity cutoff. BLAT alignments were converted to scores replicating the calculation used by the web-based BLAT server (https://genome.ucsc.edu/). For each putative is-cDNA, the raw score for an isolate was converted into a ratio by dividing the score of a particular isolate by the top score achieved by any isolate, as previously described (55). These BSRs were then imported into the MultiExperiment Viewer (MEV) to generate a heat map which could be manually inspected to identify genes present and absent across different sets of taxa (Fig. 7). Thus, a value of 1.0 (yellow on the heat map) represents the best hit, and a value of 0 (blue) represents no alignment by BLAT.

Using the BSR method described above, the potential gene content between VGIIa, VGIIb, and VGIIc sequence types was compared using representative isolates from each subtype present in the PNW (Table 2). Representative strains were selected based on assembly quality and being part of the PNW outbreak subtypes (i.e., isolates typed as VGIIa and VGIIb but not from the PNW were not included in these sets; all VGIIc isolates were from the PNW). This set included 9 VGIIa isolates, 6 VGIIb isolates, and 12 VGIIc isolates (Table 1). Putative consensus is-cDNAs showing differential content across the three PNW sequence types were identified using a cutoff BSR score of above 0.9 in the taxa of interest and below 0.9 in the excluded taxa. Multiple assemblies from each subtype were included to help reduce the number of genes that may appear to be absent due to variability of draft genome quality with multiple contigs per chromosome. Genes with a low BSR value in a single member of a subtype almost invariably were due to contig breaks in assembly. Unique putative genes were identified for each of the VGIIa, VGIIb, and VGIIc groups as well as for genes contained in both VGIIa and VGIIc but missing in VGIIb, as this comparison may help explain the previously described phenotype differences seen in the PNW.

Expanding the analysis beyond differences within the PNW sequence types, the putative is-cDNA sequences that define the PNW subtypes were scored in each of the additional 29 global VGII subtypes (Table 1 and Fig. 7), a representative sequence from each of the VGII, VGIII, and VGIV major molecular types, as well as C. neoformans var. grubii (H99) and C. neoformans var. neoformans (IEC21). This allowed for the gene differences found in the PNW subtypes to be examined across the diversity of the Cryptococcus neoformans/C. gattii species complex, as well as for the identification of genes present in all of the PNW isolates but not universally present in other VGII subtypes. In addition, the full complement of is-cDNAs was scored in all 29 global VGII subtypes.

To characterize genes absent in the PNW populations but present in many of the other global VGII subtypes, a second set of putative genes (is-cDNAs) was generated using the gene predictions in non-PNW VGII subtypes. While this would presumably contain many of the same genes found in the PNW gene set, it would also contain any genes that may be absent in all of the PNW subtypes. Using the same cutoff scheme as described above, putative genes were identified that were missing or truncated in all of the PNW subtypes but present in at least half of the other VGII subtypes. is-cDNA sequences identified as having a differential content pattern of interest were then located in the genome assemblies of strains R265 (VGIIa), B9563 (VGIIb), B8825 (VGIIc), depending on which taxa contained the is-cDNAs, to look for blocks of adjacent genes showing similar patterns of presence and absence.

(ii) BSR confirmation with sequence read data alignment. Because a low BSR score may indicate gene loss or truncation as well as problems caused by assembly issues in draft genomes, putative gene sequences identified as differential across taxa of interest were subsequently assessed by raw read data alignment. This provides evidence for confirmation of the differential gene content and, in the case of scores between 0.0 and 0.9, helps to identify the type of gene structure differences present between isolates with different scores. To avoid potential problems in aligning genomic read data against putative is-cDNA sequences, which do not contain introns, full-length gene predictions containing introns were dereplicated with USEARCH to serve as read alignment references. Putative is-cDNAs of interest were aligned against the full-length gene sequences using BLAT to identify the corresponding full-length gene. Read data from each isolate in the PNW analysis was then aligned to the full-length gene. These alignments were inspected for each gene to ensure read coverage across the entire gene sequence for isolates with high BSR scores and to ensure a lack of read coverage for isolates with BSR scores indicating gene absence. Moderate BSR scores occurred in a few cases when limited read data successfully aligned to a putatively missing gene (e.g., is-cDNA 9 in VGIIb), or when read data aligned to truncated genes. In an effort to be conservative in identifying differential gene content using short-read data, gene differences that could not be confirmed with read data are not reported. This most often occurred in the case of putative genes that contained highly repetitive sequences, which cause assembly and definitive alignment issues in short-read data.

(iii) Putative gene characterization. is-cDNAs, which were confirmed as variable with read data (see above), were translated into amino acid sequences and then searched against the Pfam database (http://pfam.sanger.ac.uk/) to help identify potential protein functions. Protein families and domains achieving a significant score are reported in Table 2. Additional characterization for selected is-cDNAs was performed using a blastp search of the NCBI nonredundant protein database (http://www.ncbi.nlm.nih.gov/RefSeq/).

SUPPLEMENTAL MATERIAL
Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.01464-14/-/DCSupplemental.
ACKNOWLEDGMENTS

We thank all our colleagues from around the world, who have submitted isolates for this study.

This study was supported by a grant from the National Institutes of Health (R21AI098859 [P.K., D.M.E., G.R.T., S.R.L., and W.M.]) and an NHMRC grant (APP1031943 [W.M., T.C.S. and G.H.]).

The findings and conclusions of this article are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

REFERENCES

1. Kwon-Chung KJ, Bennett JE. 1984. High prevalence of Cryptococcus neoformans var. gattii in tropical and subtropical regions. Zentralbl. Bakteriol. Mikrobiol. Hyg. A 257:213–218.

2. Lanza MS, Salsomaggi CA, Londero AD, Trilles L, Nishikawa MM, Wanke B. 2000. Possible primary ecological niche of Cryptococcus neoformans. Med. Mycol. 38:379–383. http://dx.doi.org/10.1080/13693700075000014.

3. Escándon P, Sánchez A, Martínez M, Meyer W, Castañeda E. 2006. Molecular epidemiology of clinical and environmental isolates of the Cryptococcus neoformans species complex reveals a high genetic diversity and the presence of the molecular type VGI mating type in a from Colombia. FEMS Yeast Res 6:625–635. http://dx.doi.org/10.1111/j.1564-6646.2006.00055.x.

4. Firtaricite C, Torres G, Rodríguez MC, Escándon P. 2011. First environmental isolation of Cryptococcus gattii serotype B from Cúcuta, Colombia. Biomedica 31:118–123. http://dx.doi.org/10.1590/S0100-879X2011000100003.

5. Ellis DH, Pfeiffer TJ. 1990. Natural habitat of Cryptococcus neoformans var. gattii. J. Clin. Microbiol. 28:1642–1644.

6. Carriconde F, Gilgado F, Arthur I, Ellis D, Malik R, de Vrieze N, Robert V, Currie BJ, Meyer W. 2011. Clonality and alpha-a-recombination in the Australian Cryptococcus gattii VGI population—an emerging outbreak in Australia. PLoS One 6:1. http://dx.doi.org/10.1371/journal.pone.0016936.

7. Fortes ST, Lázera MS, Nishikawa MM, Macedo RC, Wanke B. 2001. First isolation of Cryptococcus neoformans var. gattii from a native jungle tree in the Brazilian Amazon rainforest. Mycoses 44:137–140. http://dx.doi.org/10.1111/j.1365-3179.2001.tb00651.x.

8. Baltazar LDM, Ribeiro MAA. 2008. First isolation of Cryptococcus gattii from the environment in the State of Espirito Santo, Rev. Soc. Bras. Med. Trop. 41:449–453. (In Portuguese.) http://dx.doi.org/10.1590/S0037-86822008000500003.

9. Hagen F, Ceresini PC, Polacheck I, Ma H, van Nieuwerburgh F, Gabaldón T, Kagan S, Pursell ER, Hoogveld HL, van Iersel LJ, Klaau GW, Kelk SM, Stougie L, Bartlett KH, Voelz K, Pryszcz LP, Castañeda GW, Klau GW, Kelk SM, Stougie L, Bartlett KH, Voelz K, Pryszcz LP, Castañeda GW.

10. Meyer W, Aanensen DM, Boekhout T, Cognati M, Diaz MR, Esposto MC, Fisher M, Gilgado F, Hagen F, Kaarsohoro S, Litvinseva 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. http://dx.doi.org/10.1080/1369378090293886.

11. Kidd SE, Hagen F, Tschark RL, Huynh N, Bartlett KH, Fyfe M, Macdougall L, Boekhout T, Kwon-Chung J, Meyer W. 2004. A rare genotype of Cryptococcus gattii caused the cryptococcosis outbreak on Vancouver Island (British Columbia, Canada). Proc. Natl. Acad. Sci. U. S. A. 101:17258–17263. http://dx.doi.org/10.1073/pnas.0402981101.

12. Fraser JA, Subaran RL, Nichols CB, Heitman J. 2003. Recapitulation of the sexual cycle of the primary fungal pathogen Cryptococcus neoformans var. gattii: implications for an outbreak on Vancouver Island, Canada. Eukaryot. Cell 2:1036–1045. http://dx.doi.org/10.1083/euc.2.5.1036-1045.2003.
Liu T, Yu H, Zhang C, Lu M, Piao Y, Ohba M, Tang M, Yuan X, Wei Y, Yadav V, Yadav PK, Yadav S, Yadav KDS, 37. Nghi DH, Bittner B, Kellner H, Jehmlich N, Ullrich R, Pecyna MJ, Kavanaugh LA, Fraser JA, Dietrich FS, 43. Lazéra MS, Cavalcanti MA, Trilles L, Nishikawa MM, Wanke B, 57. DePristo MA, Banks E, Poplin R, Garrison EL, Maguire JR, Hartl C, Kozarewa I, Turner DJ, 115:4893–4901. doi:10.1038/ng.806
Engelthaler et al. 2012. Whole genome sequencing for outbreak follow-up: a tornado in Joplin, Missouri, 2011. PLoS One 7:e49989. doi:10.1371/journal.pone.0049989
Kozarewa I, Turner DJ. 2011. 96-plex molecular bar coding for the Illumina Genome Analyzer. Methods Mol. Biol. 733:279–298. http://dx.doi.org/10.1007/978-1-61779-089-8_20
Bankевич A, Nurk S, Antipov D, Gurievich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prijibelski AD, Pshyk VN, Sarsakov AV, Vyashty N, Tesler G, Astakhov ME, Pavzer PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19:455–477. http://dx.doi.org/10.1089/cmb.2012.0021
Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using maximum likehood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28:2731–2739. http://dx.doi.org/10.1093/molbev/msr121
DePristo MA, Banks E, Poplin R, Garrison EL, Maguire JR, Hartl C, Philippakas AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernshty MS, Vavachenko AY, Gubulsky K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–498. http://dx.doi.org/10.1038/ng.806
Engelthaler DM, Kelley E, Driebe EM, Bowers J, Eberhard CF, Trujillo J, Garcia-Grauenea F, Schnip M, Moseong J, Keim P, Even J. 2013. Rapid and robust phylotyping of spa 003, a dominant MRSA clone in Luxembourg and other European countries. BMC Infect. Dis. 13:339. http://dx.doi.org/10.1186/1471-2431-13-339
Swofford DL. 2003. PAUP*: Phylogenetic Analysis Using Parsimony (and other methods), version 4. Sinauer Associates, Sunderland, MA.
Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics 172:2665–2668. http://dx.doi.org/10.1534/genetics.105.048975
Bryant D, Moulton V. 2004. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21:255–265. http://dx.doi.org/10.1093/molbev/msl018
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.
Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9:772. http://dx.doi.org/10.1038/nmeth.2111
Sahl JW, Gillece JD, Schupp JM, Waddell V, Driebe EM, Engelthaler DM, Keim P. 2013. Evolution of a pathogen: a comparative genomics analysis identifies a genetic pathway to pathogenesis in Acinetobacter. PLoS One 8:e54267. http://dx.doi.org/10.1371/journal.pone.0054287
Stanke M, Diekhans M, Baertsch R, Hausler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24:387–384. http://dx.doi.org/10.1093/bioinformatics/btn013
Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. http://dx.doi.org/10.1093/bioinformatics/btq461
Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12:656–664. http://dx.doi.org/10.1101/gr.229202