Fungal Genetics and Biology 75 (2015) 64–71

Contents lists available at ScienceDirect

Fungal Genetics and Biology

journal homepage: www.elsevier.com/locate/yfgbi

Tools and Techniques

Next generation multilocus sequence typing (NGMLST) and the analytical software program MLSTEZ enable efficient, cost-effective, high-throughput, multilocus sequencing typing

Yuan Chen [a,b,*], Aubrey E. Frazzitta [a,1], Anastasia P. Litvintseva [b,2], Charles Fang [a], Thomas G. Mitchell [b], Deborah J. Springer [b], Yun Ding [c], George Yuan [d], John R. Perfect [a,*]

[a] Division of Infectious Diseases, Department of Medicine, Duke University Medical Center, Durham, NC, USA
[b] Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, NC, USA
[c] Janelia Research Campus, HHMI, Ashburn, VA, USA
[d] Pacific Sciences, Menlo Park, CA, USA

ARTICLE INFO

Article history:
Received 28 October 2014
Accepted 17 January 2015
Available online 24 January 2015

Keywords:
MLST
Next generation sequencing (NGS)
Multiplex PCR
PacBio CCS sequencing
Software
Cryptococcus neoformans

ABSTRACT

Multilocus sequence typing (MLST) has become the preferred method for genotyping many biological species, and it is especially useful for analyzing haploid eukaryotes. MLST is rigorous, reproducible, and informative, and MLST genotyping has been shown to identify major phylogenetic clades, molecular groups, or subpopulations of a species, as well as individual strains or clones. MLST molecular types often correlate with important phenotypes. Conventional MLST involves the extraction of genomic DNA and the amplification by PCR of several conserved, unlinked gene sequences from a sample of isolates of the taxon under investigation. In some cases, as few as three loci are sufficient to yield definitive results. The amplicons are sequenced, aligned, and compared by phylogenetic methods to distinguish statistically significant differences among individuals and clades. Although MLST is simpler, faster, and less expensive than whole genome sequencing, it is more costly and time-consuming than less reliable genotyping methods (e.g. amplified fragment length polymorphisms). Here, we describe a new MLST method that uses next-generation sequencing, a multiplexing protocol, and appropriate analytical software to provide accurate, rapid, and economical MLST genotyping of 96 or more isolates in single assay. We demonstrate this methodology by genotyping isolates of the well-characterized, human pathogenic yeast *Cryptococcus neoformans*.

1. Introduction

Efficient methods for estimating the genetic diversity among microorganisms are essential for understanding their evolutionary history, geographic distribution, and pathogenicity. In the past decades, numerous methods have been developed for typing bacteria and fungi (Li et al., 2009; Vanhee et al., 2010). Some of these methods can characterize a large number of isolates at low cost, such as pulsed-field gel electrophoresis (PFGE) (Schwartz and Cantor, 1984) and amplified fragment length polymorphism (AFLP) (Vos et al., 1995). However, the results of these methods are laboratory specific and usually are not comparable among laboratories. Conversely, DNA sequencing results can be archived and shared among laboratories, and therefore, these methods are widely used in microbial studies today (Janbon et al., 2014; Li et al., 2009; Litvintseva et al., 2006; Tavanti et al., 2005; Taylor and Fisher, 2003; Vanhee et al., 2010). Multilocus sequence typing (MLST) targets multiple genomic loci and is considered one of the most reliable and informative methods for molecular genotyping (Maiden et al., 1998; Schwartz and Cantor, 1984). MLST has been applied to many pathogenic microorganisms, and there is increasing interest in the variation among isolates and within microbial populations, especially in studies of microbial evolution, pathogenesis, ecology, and microbiomes (Byrnes et al., 2009; Chen et al., 2013; Litvintseva and Mitchell, 2012; Meyer et al., 2009). Moreover,

* Corresponding authors at: Division of Infectious Diseases, Department of Medicine, Duke University Medical Center, Durham, NC, USA (Y. Chen and J.R. Perfect). Fax: +1 919 6848902.
E-mail addresses: ychenbioinfo@gmail.com (Y. Chen), perfe001@mc.duke.edu (J.R. Perfect).
[1] Present address: Oregon Health and Science University, School of Medicine, Portland, OR, USA.
[2] Present address: Mycotic Diseases Branch, Centers for Disease Control and Prevention, Atlanta, GA, USA.

http://dx.doi.org/10.1016/j.fgb.2015.01.005
1087-1845/

online MLST databases have been constructed for several bacterial and fungal species to facilitate molecular epidemiological studies and surveillance (Chan et al., 2001). MLST genotyping is a superb approach to delineate species and strains, but the current methodology is costly, time-consuming, and laborious.

To accelerate automation and expand the versatility of the current MLST method, we developed a high-throughput next-generation sequencing approach, NGMLST, and an automated software program for data analyses, MLSTEZ. We adapted multiplex PCR, which may save more than 75% of the PCR work (calculated based on using seven MLST loci). For next-generation sequencing, we employed the Pacific Biosciences (PacBio) circular consensus sequencing (CCS) technology, which is capable of generating relatively inexpensive, single-molecule consensus reads of 1–2 kbp in length. Unlike the usual PacBio read, a CCS read is an error-corrected consensus read generated from the consensus alignment of single-molecule circular sequencing (Eid et al., 2009). Therefore, the accuracy of a CCS read is correlated with the number of sequencing passes of the template molecule (Travers et al., 2010). With the benefit of these higher quality reads, our software, MLSTEZ, can automatically identify the barcodes and primers used in the PCR, correct sequencing errors, generate the MLST profile for each isolate, and predict potentially heterozygous loci.

Cryptococcus neoformans is a well-characterized, opportunistic human fungal pathogen, and it is responsible for approximately 600,000 annual deaths worldwide (Park et al., 2009). In this study, we targeted the nine MLST loci that are commonly used to genotype isolates of the *C. neoformans*/*Cryptococcus gattii* species complex. As controls, we selected 28 clinical and environmental haploid strains with known MLST genotypes that represented each major subpopulation or molecular type of the species complex, as well as six previously described diploid hybrid strains (Litvintseva et al., 2006; Simwami et al., 2011; Stephen et al., 2002; Sun et al., 2012; Xu et al., 2009). We pooled the amplicons of these 34 isolates with those of another 62 wild type *C. neoformans* isolates and sequenced them in one PacBio SMRT Cell. The NGMLST method and MLSTEZ software produced high quality, unambiguous MLST profiles of all 96 isolates, and the sequences of the reference strains were identical to their genotypes, which were previously determined by the conventional MLST method. The MLSTEZ successfully detected heterozygous loci in the hybrid strains and identified the sequences of each allele.

2. Materials and methods

2.1. Strains of C. neoformans

As reference controls, we selected conventionally MLST-genotyped strains of *C. neoformans* var. *grubii* (*Cng*), *C. neoformans* var. *neoformans* (molecular type VNIV), and *C. gattii*. Distinct genetic subpopulations of these recognized species and varieties were also considered when we selected control strains. For example, we included all three molecular types of *Cng* (VNI, VNB and VNII) (Litvintseva et al., 2006) and the four molecular types of *C. gattii* (VGI, VGII, VGIII, and VGIV). The number of strains for each molecular type are as follows (Table S1): 11 strains of *C. neoformans* var. *grubii* (five VNI strains, three VNB strains, three VNII strains); three strains of *C. neoformans* var. *neoformans* (VNIV); 14 strains of the sibling species, *C. gattii* (four VGI strains, three VGII strains, five VGIII strains, two VGIV strains); and six hybrid strains (three VNIII, two VGII/VGIII, one VNB/VNII). The other 62 isolates were wild type clinical and environmental isolates of *C. neoformans* collected from Brazil and Botswana.

2.2. MLST target loci and primer design

As routinely employed for genotyping strains of *C. neoformans* and *C. gattii*, the following nine MLST loci were used to analyze the genetic diversity of the strains: *CAP59*, *GPD1*, IGS1, *LAC1*, *PLB1*, *SOD1*, *URA5*, *TEF1* and *MPD1* (Colom et al., 2012; Litvintseva et al., 2006, 2011; MacDougall et al., 2007; Meyer et al., 2009). The locus-specific primers are listed in Table 1. A 20-bp universal primer (5′-CTGGAGCACGAGGACACTGA) was added at the 5′ end of each locus-specific primer (Fig. 1). Each barcode primer included a 5-bp padding sequence (GGTAG) at the 5′ end, followed by the 16-bp barcode sequence as suggested by PacBio (http://www.smrtcommunity.com/servlet/servlet.FileDownload?file=00P7000000W067VEAR), and a 20-bp universal primer was added to the 3′ end. The sequences of the 96 barcode primers used in our study are listed in Table S2.

2.3. NGMLST library preparation

Genomic DNA was isolated from each yeast strain using a MasterPure yeast DNA purification kit (Epicentre Biotechnologies, Madison, WI) according to the manufacturer's instructions. MLST loci of interest were amplified by two rounds of PCRs to prepare the library. The first PCR was used to amplify the target loci and then the unique barcodes for labeling the amplicons from each isolate were added in the second PCR.

For the first round, each multiplex PCR mixture contained 12.5 μL 2× Master Mix (QIAGEN Multiplex PCR *Plus* Kit, cat # 206152), approximately 2.5 ng genomic DNA, and nine primer pairs at the optimized concentration for each pair (Table 1). The PCR was conducted with the following thermocycling conditions: initial denaturation at 95 °C for 5 min, followed by 35 cycles of 30 s at 95 °C, 1.5 min at 58 °C, and 1.5 min at 72 °C, and finally, 10 min at 68 °C for extension.

These multiplexed products were then diluted 1:50 and used as templates for the second round of PCR, which were carried out in volumes of 25 μL that contained LongAmp *Taq* DNA Polymerase (New England BioLabs Inc., catalog # M0323L), 1 μL of diluted

Table 1
Nine pairs of MLST locus specific primer sequences and corresponding primer concentrations and product lengths.

Locus	Upper primer	Lower primer	Concentration (μM)	Product length[a]
SOD1	5′-GGCACAACTCCACCGATCA	5′-CTTACATGACACCGCAGGCA	0.3	668
LAC1	5′-AACATGTTCCCTGGACCTGTG	5′-ACGTGGATCTCGGGAGGA	0.3	816
MPD1	5′-TGCCCTGGATCCTAATGCTCT	5′-ACCCAGACTGCCGCTGTCGTC	0.8	1008
TEF1	5′-AATCGTCAAGGAGACCAACG	5′-CGTCACCAGACTTGACGAAC	0.4	811
CAP59	5′-CTCTACGTCGAGCAAGTCAAG	5′-TCCGCTGCACAAGTGATACCC	0.3	564
PLB1	5′-CTTCAGGCGGAGAGAGGTTT	5′-GATTTGGCGTTGGTTTCAGT	0.3	635
GPD1	5′-ATGGTCGTCAAGGTTGGAAT	5′-GTATTCGGCACCAGCCTCA	0.4	561
IGS1	5′-GGGACCAGTGCATTGCATGA	5′-ATCCTTTGCAGACGACTTGA	0.1	845
URA5	5′-ATGTCTTCCCAAGCCCTCGAC	5′-TTAAGACCTCTGAACACCGTACTC	0.4	733

[a] The production lengths are based on the H99 genome, and the primer lengths are not counted into products.

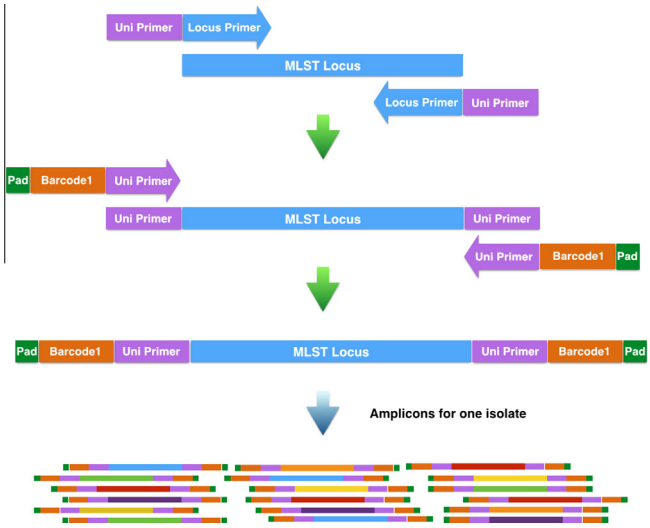

Fig. 1. Two rounds of PCRs are employed in NGMLST. In the first PCR round, each primer consists of a locus-specific sequence (blue, see Table 1) and a 20-bp universal primer sequence (purple, 5′-GCTGTCAACGATACGCTACG). The diluted PCR product is used as template for the second PCR round. The barcode primers consist of three parts: (i) a 20-bp universal sequence (purple), which amplifies the template; (ii) a 16-bp barcode sequence (orange) that identifies the amplicons from each different isolate; (iii) and a 5-bp padding sequence (green) to provide equivalent binding affinities for adding the PacBio sequencing adapters. Because multiplex PCRs were used in the first PCR round, primer pairs for each of the nine loci are added to the PCR mix at the same time. In the second PCR round, the various barcode primers are used to identify each isolates. The final products of each isolate would have the same sequence structure on both ends, flanking different target locus sequences in the middle, which are shown with different colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

multiplex PCR product, and 2 μL 10 μM barcode primer. The PCR was performed with the following cycling conditions: initial denaturation at 94 °C for 30 s followed by 35 cycles of 30 s at 94 °C, 30 s at 50 °C, and 60 s at 65 °C, and lastly, 10 min at 65 °C for extension.

The amplicons of the 96 strains were visualized on a 1.4% TAE agarose gel, and their concentrations were estimated. The amplicons were pooled into four groups of 24 strains based on having similar concentrations of DNA. Each pool of 24 amplicons was purified utilizing the QIAquick PCR Purification Kit (Qiagen, catalog # 28106), the DNA concentration of each pool was determined using a Nanodrop ND-1000 Spectrophotometer, and portions of the four purified pools containing equal concentration of DNA were combined.

2.4. PacBio sequencing

SMRT Cell sequencing libraries were prepared using Pacific Biosciences DNA Template Prep Kit 2.0 (catalog # 001-540-835) according to the 3-kb or 10-kb template preparation and sequencing protocol provided by Pacific Biosciences. Instead of using magnetic beads, the amplicons were loaded by diffusion at a concentration of 300 pM. The PacBio RS II platform was used for sequencing the amplicons. One SMRT Cell was used to sequence all 96 pooled isolates. The sequencing run used 1 × 180 min movie with P4-C2 chemistry.

2.5. Data analysis

Primary analysis was performed using the PacBio SMRT Analysis version 2.1 program, and the filtering parameters were as follows: minimum polymerase read quality of 0.75; minimum read length of 50 bp; and minimum subread length of 50 bp. Circular consensus

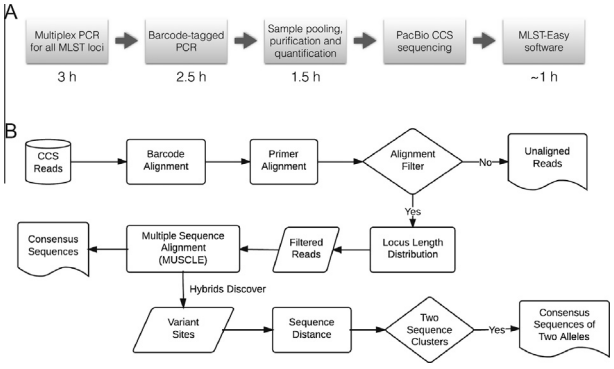

Fig. 2. The workflow for NGMLST with estimated time for each step (A) and flowchart of the analysis pipeline used in MLSTEZ (B).

sequencing (CCS) reads with less than four full passes were also filtered in further analysis. We used MLSTEZ to generate all the consensus sequences of each locus and searched for heterozygous loci. The analysis steps were outlined as flowchart in Fig. 2B. This software used the Smith–Waterman algorithm to identify each barcode and specific MLST locus in the reads. Then, the first quartile (Q1) and third quartile (Q3) of each MLST locus length among all sequenced isolates were calculated. The interquartile range (IQR) was calculated as Q3 − Q1. Reads with length less than Q1 − 1.5 * IQR or larger than Q3 + 1.5 * IQR of the specific locus were considered to be outliers and removed from the dataset. Then, all the reads were ranked by their sequencing scores. In this study, a minimum of three and a maximum of 10 reads of each locus were aligned using MUSCLE to generate the consensus sequence (Edgar, 2004). To detect heterozygosity, all the reads identified at each locus were aligned, and variation scores were calculated based on the number of variant sites among the sequences. A locus with two groups of reads that had significantly different variant scores ($p < 0.001$) was considered heterozygous. Consensus sequences of the two alleles were generated separately by different groups of reads. The following software parameters were used: barcode_length = 16; min_readnum = 3; max_readnum = 10; flanking_length = 5; match_score = 2; mismatch_score = −1; gap_score = −1; max_mismatch = 3. The entire analysis was performed on an iMac computer with 3.4G Intel Core i7, 16GB 1333MHz DDR3, and Mac OS X 10.9.2.

2.6. Software

The algorithm was written in Python, version 2.7.6. PyQt4 (http://www.riverbankcomputing.com/software/pyqt/download) and Qt Designer (http://qt-project.org/doc/qt-4.8/designer-manual.html) were used to create the graphic user interface (GUI) (Fig. 3). Mac and Windows versions of the GUI software were tested on computers with Mac OS X10.9 and Windows 7 operating systems, respectively.

3. Results

3.1. Development of multiplex PCR and resultant data production

To evaluate the multiplex PCR protocol for NGMLST, we selected the nine consensus, unlinked MLST loci adopted for genotyping isolates of *C. neoformans* and *C. gattii*: *CAP59*, *GPD1*, IGS1, *LAC1*, *PLB1*, *SOD1*, *URA5*, *TEF1* and *MPD1* (Colom et al., 2012; Litvintseva et al., 2011, 2006; MacDougall et al., 2007; Meyer et al., 2009). Of these loci, *MPD1* was used only for isolates of *C. gattii*. To enable simultaneous amplification of the other eight loci

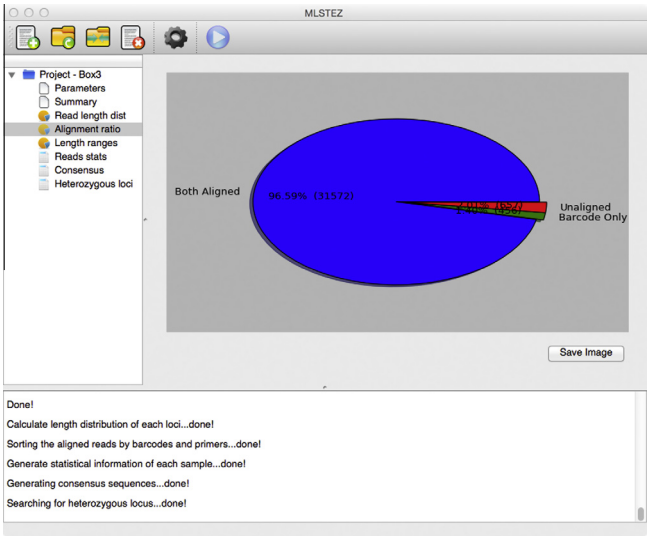

Fig. 3. Graphic user interface of MLSTEZ under Mac OS X system. The interface consists of four parts: toolbox bar (top), list of analyses panel (mid-left), analysis result panel (mid-right), and running status panel (bottom).

from most isolates of *C. neoformans* and *C. gattii*, we designed new pairs of primers that were specific for five loci (IGS1, *TEF1*, *LAC1*, *SOD1*, and *URA5*), which targeted the same regions used in previous studies, and we used previously designed primers for *CAP59*, *GPD1*, *PLB1*, and *MPD1* (Table 1). In addition, all nine MLST locus-specific primers were modified to include a universal primer sequence at the 5′ end (Fig. 1), which was needed to facilitate the addition of barcodes in the subsequent step (Fig. 2A). The nine pairs of locus-specific primers were admixed with the optimized concentrations (Table 1), and all the loci were amplified simultaneously. Although some strains and/or species differed in the efficiency with which they were amplified (Table 3), all the loci were successfully amplified in most tested isolates (Fig. 4).

The barcode primers for the second PCR round consisted of three parts (Fig. 1). The padding sequence was used to ensure that each product had equal efficiency to ligate to the sequencing adapter. The barcode sequence was unique to each isolate and was used to separate the amplicons from different isolates by MSLTEZ. The

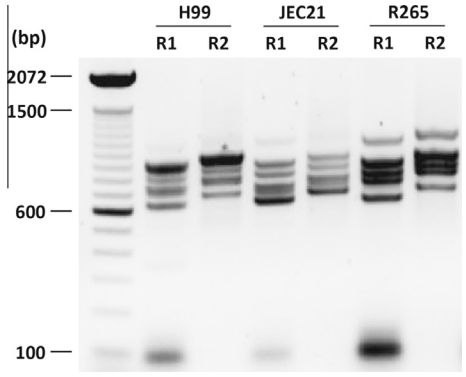

Fig. 4. Two rounds of PCR products of isolates H99 (VNI molecular type), R265 (VGII), and JEC21 (VNIV) are shown on 1.4% TAE agarose gel. R1 and R2 stand for the first and second PCR round, respectively. The expected PCR product sizes are shown in Table 1. The bands from top to bottom are PCR products of *MPD1*, IGS1, *LAC1*, *TEF1*, *URA5*, *SOD1*, *PLB1*, *CAP59* and *GPD1*. Some bands are overlapped because of similar product lengths. The gel image indicates that the *MPD1* (top band) locus was amplified with greater efficiency from R265 than H99 and JEC21. The primer pairs of other loci also reveal different amplification efficiencies among isolates from different molecular groups (Table 2).

amplicons of the first PCR round were amplified by the same universal primer that we had added into the barcode primer.

To test the accuracy of the PacBio sequencing platform for NGMLST, we selected 28 diverse reference strains that represented the eight major haploid molecular types of *C. neoformans* and *C. gattii* and six hybrids, which are very difficult to genotype using the conventional MLST protocol (Table S1). In addition, DNA from 62 wild type isolates of *C. neoformans* were also added to the test mixtures. We pooled all the barcoded amplicons of 96 isolates and sequenced them in one PacBio SMRT Cell. Four full passes yielded 37,906 CCS reads with an average CCS read length of 730 bp. As expected, more than 80% of the reads ranged between 600 and 1100 bp (Fig. 5A).

3.2. Data processing

The first step of the analysis pipeline to generate the MLST profile for each isolate is to identify the unique barcode sequence added during the second round of amplification. MLSTEZ successfully identified barcode sequences on 32,932 of 37,906 (86.9%) CCS reads. The average number of reads obtained for each isolate was 343.0 (Fig. 5B). Subsequently, the barcode-called amplicons were separated by the locus-specific primer sequences. Due to the low sequencing qualities of some reads, primer sequences

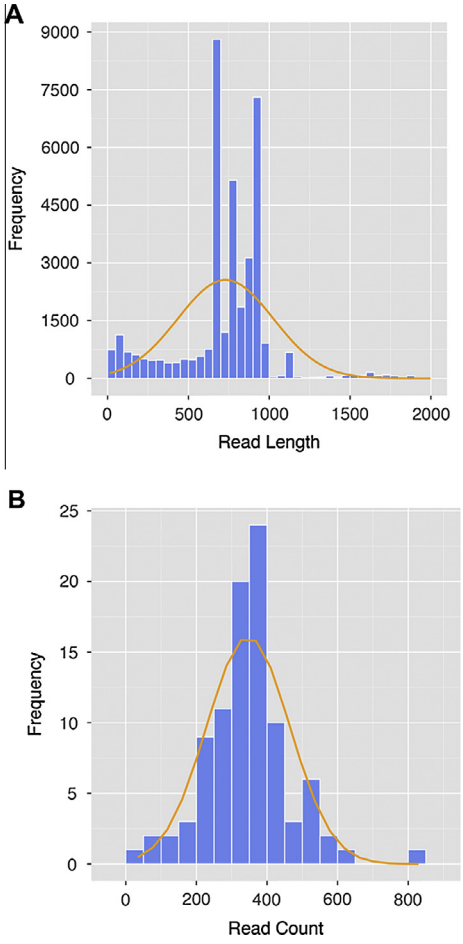

Fig. 5. Length distribution of CCS reads generated from 96 isolates (A). More than 80% of the reads have sequence lengths between 600 and 1100 bp. Normal distribution is shown for the read count of the 96 isolates (B). Distribution of read counts for isolates. The average read count of each isolate is 343, and minimal and maximal read counts are 34 and 829, respectively.

could not be identified on 1641 of 32,932 (5.0%) CCS reads. These reads were then removed from further analysis.

We obtained CCS reads from 818 of 864 (94.7%) alleles. The failure to obtain the sequences of certain loci in some isolates could probably be explained by the sequence diversity between the isolates and the primer sequences, which resulted in low amplification efficiencies of some primer in certain molecular type isolates (Table 2). This result was verified by electrophoretic gel images (data not shown), and the data from missing loci were then sequenced manually using Sanger technology.

3.3. Verification of NGMLST data

Both conventional MLST and NGMLST genotyping require sequence data of very high quality. Compared with other next generation sequencing platforms, such as Roche 454, Illumina HiSeq, or Ion Torrent, PacBio has the advantage providing reads of longer length, but the analysis of PacBio reads requires dealing with a relatively high error rate prior to consensus sequence determination (Koren et al., 2012; Quail et al., 2012). Therefore, we needed to verify that PacBio was able to generate high quality NGMLST profiles that were comparable to data obtained by conventional MLST. The 34 reference strains tested here included a total of 306 alleles, and 206 of these alleles were previously sequenced by Sanger method. Thus, the sequences of these alleles were compared with the corresponding sequences produced by NGMLST and MLSTEZ.

We obtained on average 37.8 CCS reads for each allele of the 34 isolates. However, due to the low efficiency of several primers in isolates of certain molecular types, 22 of 206 alleles did not have more than three reads, which was our minimal requirement to generate a consensus. The newly generated NGMLST profiles were compared with 184 MLST alleles previously obtained by Sanger sequencing. The result demonstrated that 172 alleles were 100% identical between the two protocols, and the other 12 alleles only had very limited mismatches ($\leqslant$3 SNP per sequence). Thus, the sequencing accuracy has surpassed 99.98%. Using the phylogenetic analysis, the other 62 isolates were identified as 1 VNI, 3 VNB, 18 VGI, 15 VGII, 1 VGIII, 22 VGIV and 2 VN/VG hybrids. This result clearly confirmed the high quality of MLST profiles generated by NGMLST, which is also a more rapid and less expensive alternative to the conventional method.

3.4. Identification of hybrids and allelic sequences

We assessed utility of NGMLST for simultaneous sequencing and differentiating alleles in the diploid hybrid strains by including six hybrid *C. neoformans* strains: three VNIII (VNI + VNIV) hybrids, two VGII/VGIII hybrids, and one VNII/VNB hybrid. The heterozygous locus discovery function of MLSTEZ was used to analyze the sequencing data. A minimal of five reads were required for analysis for the heterozygous locus analysis. As expected, multiple heterozygous loci were reported by the software for each hybrid (Table 3). Phylogenetic analysis of the recovered alleles showed that the compositions of most heterozygous loci of the hybrids were consistent with previous studies (Fig. 6). A few loci from some haploid isolates were erroneously reported as having a heterozygous locus. Additional analysis revealed that these false positive results were caused by reads of low quality and quantity.

Table 2
Primer efficiencies in multiplex PCR of different molecular type isolates. Increased number of "+" stands for higher efficiency of the primers. Primers with "+++" have very high efficiency in all test isolates. Primers "++" work well in most isolates, and enough read coverage ($\geqslant$3) for loci to be obtained. Primers with "+" work inconsistently among the isolates, and they may occasionally not be able to yield sufficient reads. The primers labeled "–" rarely worked with the corresponding molecular types among the isolates tested.

	IGS1	*TEF1*	*GPD1*	*LAC1*	*PLB1*	*MPD1*	*CAP59*	*SOD1*	*URA5*
VNI	+++	+++	+++	+++	+++	+	+	+++	+++
VNB	+++	++	+++	++	+++	–	+	+++	+++
VNII	+++	+++	+++	++	+++	+	+	+++	+++
VNIV	++	+++	+++	–	+++	+	+++	+++	+++
VGI	++	+++	+++	+++	++	–	+++	++	+++
VGII	–	+++	–	+++	+	+++	+++	+++	+++
VGIII	++	+++	+++	+++	–	+++	+++	+++	+++
VGIV	++	+++	+++	+++	+	+	+++	+++	+++

4. Discussion

In studies of molecular epidemiology, pathogenicity, and phylogenetics, MLST has become the standard method of genotyping many fungi, including strains of the *C. neoformans*/*C. gattii* species complex. Furthermore, it also widely used in genotyping other fungal species such as *Candida* (Jackson et al., 2009), *Aspergillus* (Bain et al., 2007), and *Pseudallescheria* (Bernhardt et al., 2013). Although whole genome sequence typing (WGST) is becoming more accessible, especially for organisms with small genomes, such as bacteria and viruses, it is not yet practical for genotyping numerous isolates of eukaryotic species. MLST will not soon become obsolete because it provides an economical and efficient method of screening wild type isolates, assigning them to established clades, subpopulations, or phenotypic groups, and determining whether they warrant more extensive analysis or WGST. However, compared to less reproducible and more subjective methods of rapid genotyping, such as generating amplified fragment length polymorphisms, MLST has the disadvantages of being more time consuming as well as costly due to the use of Sanger sequencing. To resolve these issues, we have developed a new high-throughput method of MLST genotyping that generates CCS PacBio next-generation sequencing reads, NGMLST, and a novel multifunctional software program, MLSTEZ, which provides simplified and automated analysis of NGMLST data. The average time required for processing DNA from 96 isolates was 7 h. Previously, 2–4 weeks were required to obtain sequence data from the same number of strains using conventional MLST.

To interface with NGMLST, we developed the multifaceted software program, MLSTEZ, which is available on the Internet at no cost (https://sourceforge.net/projects/mlstez/). The program is fully automated and requires a general sequencing format file (FASTQ and FASTA) as input, which means that NGMLST will support all sequencing platforms that can generate full-length barcoded amplicons. Because a sequence assembly feature is not included in the program, fragmented amplicons must first be assembled before analysis by MLSTEZ. MLSTEZ can perform barcode and primer identification, recognize consensus sequences, and predict heterozygous loci. All the results that are generated by the software can be easily exported as sequence files, graphs, or tables. In addition, the MLSTEZ output sequence files can be used directly for phylogenetic analyses, which significantly reduces the time required for many follow-up studies. With the multiprocessing features of MLSTEZ, the analyses of data from

Table 3
Heterozygous loci of the hybrids predicted by MLSTEZ. 'Yes' indicates the identification of two alleles, 'No' indicates that only one allele was identified, and the loci without insufficient reads (<5) for analysis are labeled 'NA'.

Strain	*CAP59*	*GPD1*	*IGS1*	*LAC1*	*MPD1*	*PLB1*	*SOD1*	*TEF1*	*URA5*
Cng9	NA	Yes	No	No	No	No	Yes	No	Yes
ZG287	No	Yes	No	Yes	NA	Yes	Yes	Yes	Yes
CBS132	No	Yes	No	NA	No	No	Yes	No	Yes
NC34-21	No	Yes	No	Yes	NA	Yes	Yes	Yes	Yes
EJB34	No	No	NA	Yes	Yes	NA	Yes	Yes	Yes
EJB41	No	No	No	Yes	No	NA	Yes	No	No

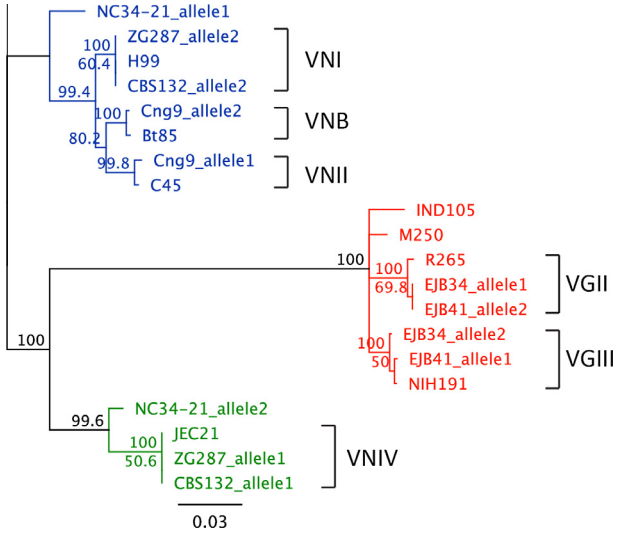

Fig. 6. Phylogeny of the *SOD1* locus among isolates with different molecular types and both alleles of six hybrids visualized by the neighbor-joining dendrogram. Different species and molecular groups of the isolates are color-coded (blue, *C. neoformans* var. *grubii*; red, *C. gatii*; green, *C. neoformans* var. *neoformans*). All the sequences were generated using MLSTEZ based on NGMLST sequencing result. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

one PacBio SMRT Cell can be completed within an hour on a modern desktop computer. Thus, the rate-determining step of this protocol is the time required for PacBio sequencing.

Recently, a NGS genotyping method (HiMLST) was proposed by Boers et al. (2012) for typing four different bacterial species using 454 pyrophosphate sequencing. The comparisons among conventional MLST, HiMLST and NGMLST are shown in Table 4. The major advantages in our MGMLST approach are: (i) the employment of multiplex PCR greatly reduces the amount of labor; (ii) the cost of PacBio CCS sequencing is only about 20% of Roche 454 sequencing; (iii) PacBio greatly extends the maximum read length of target loci or genes from 500-bp to 2-kb without requiring fragmentation into shorter sequences; (iv) the NGMLST workflow was optimized to reduce unnecessary steps; (v) MLSTEZ can be easily implemented and does not require technical expertise or a background in bioinformatics; and (vi) for analysis of hybrid isolates, unlike most programs, MLSTEZ can detect heterozygous loci and sequence their alleles.

PacBio CCS reads have an error rate of 2.5% with ~1.5 kb insertion size (Jiao et al., 2013), which is considerably higher than other platforms. However, because these errors occur randomly and are not biased toward homopolymeric regions (Carneiro et al., 2012), accuracy approaching 100% can be achieved by increasing the level of coverage or number of reads. Our software routinely employs multiple PacBio CCS reads to generate consensus sequences, and accuracy can surpass 99.98%, which is sufficient for genotyping. In preliminary experiments, we determined that more than three reads were required to generate an accurate consensus sequence. However, our tests showed that including more than 10 reads per locus did not significantly improve the quality. On the contrary, exceeding 10 reads per allele tended to overfill the program with low quality reads, which sometimes reduced the accuracy. Therefore, to generate optimal consensus sequences, only the top 10 scored reads were used. We also observed that a small proportion of the reads were longer or shorter than expected. Most of the shorter reads were leftover adaptor and incomplete PCR products, and the longer reads represent concatemers generated by ligation during preparation of the PacBio library. To resolve this issue, we added a length filter in our analysis pipeline (Fig. 2B) to ensure that only sequencing reads within the correct length range would be used for generating the consensus sequences.

For this evaluation of NGMLST and MLSTEZ, we targeted the nine unlinked loci that are routinely used to genotype isolates of the *C. neoformans*/*C. gattii* species complex. Five of the primer pairs were identical to those used in previous studies but with different annealing PCR temperatures (Colom et al., 2012; Litvintseva et al., 2011, 2006; MacDougall et al., 2007; Meyer et al., 2009). After adjusting and standardizing the PCR conditions, these primers worked well in the thermocycling parameters for multiplexing. The primer pairs of the other four loci were developed specifically for this study, but they targeted the same regions used in previous reports. Under these optimized conditions, the primer pairs amplified the previously established cryptococcal MLST loci. Preliminary results with reference strains confirmed that the primers used here (Table S1) accurately genotyped both species and the molecular types of *C. neoformans* and *C. gattii* in addition to the hybrid strains. The use of species-specific primers could further improve the results. For example, we used the same protocol to genotype 96 isolates of *C. neoformans* var. *grubii* with eight pairs of primers (without *MPD1*), and 762 of 768 (99.2%) alleles had more than three filtered reads to generate consensus (data not shown).

Although we have only demonstrated the application of NGMLST to *C. neoformans* and *C. gattii*, this approach can be used for any MLST investigation. Most MLST analyses performed by conventional MLST could be readily adapted to this method. In our study, we found that the primers previously used under different PCR conditions (Litvintseva et al., 2006) worked reasonably well in a single multiplex PCR system using the same conditions. Several caveats are suggested for successfully replacing conventional MLST with NGMLST: (i) NGMLST can accommodate amplicoms up to 2 kbp in length; however, the maximal difference in length among the amplicons cannot exceed 500 bp to avoid affecting the yield of sequencing reads; (ii) the concentration of locus-specific primers needs to be optimized to obtain equal amounts of each product; and (iii) considering the quality and amount of data that are generated with the current protocol, the numbers of target MLST loci and tested isolates need to be balanced. We suggest analyzing no more than 11 loci for 96 isolates at one time. Multiple groups of multiplex PCRs could then be employed to accommodate different PCR conditions and/or the need for a large number of loci required by species with low amounts of genetic variation.

Table 4
Comparisons between conventional MLST, HiMLST and NGMLST based on 96 isolates with 8 target loci.

	Conventional MLST	HiMLST	NGMLST
Number of PCRs	768	864 (768 + 96)	192 (96 + 96)
Number of PCR product purifications	768	More than 96	4
Sanger sequencing reaction	1536 (768 × 2)	None	None
Automate data analysis tool	None	None	MLSTEZ
Estimated experimental time	>1 week	>30 h	7 h
Estimated data analysis time	>10 h (manually)	>10 h (manually)	≈1 h (automatically)[a]
Estimated cost per isolate	$70.56 (bi-directional)	$42.23 (1/4 plate)	$8.83 (1 SMRT Cell)

[a] Tested with 8 threads on iMac (Mac OS X 10.9.2) on 3.4G Intel Core i7, 16GB 1333MHz DDR3.

Our results show that NGMLST and MLSTEZ not only work well on the haploid strains but also can be used to detect and analyze hybrid strains, which are difficult to MLST genotype using conventional Sanger sequencing. Among our six control hybrid strains, most were detected by more than three heterozygous loci. Unfortunately, some haploid strains were erroneously identified with heterozygous loci; these results were caused by low coverage or reads of poor quality. Therefore, we strongly recommend repeating the analysis on putative hybrids. In addition, MLST only targets a limited number of genomic loci, and aneuploid strains are very common in some fungal species (Kwon-Chung and Chang, 2012; Selmecki et al., 2009). It remains difficult to determine the ploidy of test strains even when multiple loci have been identified to be heterozygous. Other methods to determine aneuploidy, such as analysis of the cells by fluorescent-activated cell sorting (FACS) could help to confirm MLST data and ploidy.

This investigation evaluated a novel NGMLST method of genotyping, which has proven to be rapid and relatively inexpensive, as well as amenable to the high-throughput analyses of large samples. Coupled with the automated multifunctional software, MLSTEZ, high quality MLST profiles can be acquired with very simple operations in a short period of time. The approach demonstrated here was evaluated with the heterobasidiomycetous human pathogen, *Cryptococcus*, but it can be applied to many other fungal or other eukaryotic taxa, including haploid, diploid, and hybrid organisms.

Conflict of interest

None of the authors have a conflict of interest.

Data accessibility

The source code and pre-compiled GUI applications for Macintosh and Windows PC of MLSTEZ are freely available on https://sourceforge.net/projects/mlstez/files/. The GUI application is only available for Macintosh OS X 10.6+ and Windows XP and later version. The manual of the application is available on https://sourceforge.net/p/mlstez/wiki/Manual/. MLSTEZ is under active development. Bug reports and suggestions from users will be helpful for the improvements of the software in future version.

Acknowledgments

The authors thank to Sun Sheng (Duke University Medical Center), Matthew Fisher (Imperial College London), Tom Harrison (St George's, University of London), Vinicius Ponzio (Federal University of São Paulo), Arnaldo L. Colombo (Federal University of São Paulo) and Annemarie Brouwer (Radboud University Nijmegen) for providing isolates for the study. The authors also thank to Josh Granek (Duke University Medical Center), Olivier Fedrigo and Graham Alexander (Duke Center for Genomic and Computational Biology) for discussion and sample sequencing. This work was supported by Public Health Service Grants AI73896 and AI93257 (JRP).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fgb.2015.01.005.

References

Bain, J.M., Tavanti, A., Davidson, A.D., Jacobsen, M.D., Shaw, D., Gow, N.A.R., Odds, F.C., 2007. Multilocus sequence typing of the pathogenic fungus *Aspergillus fumigatus*. J. Clin. Microbiol. 45, 1469–1477. http://dx.doi.org/10.1128/JCM.00064-07.

Bernhardt, A., Sedlacek, L., Wagner, S., Schwarz, C., Würstl, B., Tintelnot, K., 2013. Multilocus sequence typing of *Scedosporium apiospermum* and *Pseudallescheria boydii* isolates from cystic fibrosis patients. J. Cyst. Fibros. 12, 592–598. http://dx.doi.org/10.1016/j.jcf.2013.05.007.

Boers, S.A., van der Reijden, W.A., Jansen, R., 2012. High-throughput multilocus sequence typing: bringing molecular typing to the next level. PLoS ONE 7, e39630. http://dx.doi.org/10.1371/journal.pone.0039630.

Byrnes, E.J., Bildfell, R.J., Frank, S.A., Mitchell, T.G., Marr, K.A., Heitman, J., 2009. Molecular evidence that the range of the Vancouver Island outbreak of *Cryptococcus gattii* infection has expanded into the Pacific Northwest in the United States. J. Infect. Dis. 199, 1081–1086. http://dx.doi.org/10.1086/597306.

Carneiro, M.O., Russ, C., Ross, M.G., Gabriel, S.B., Nusbaum, C., DePristo, M.A., 2012. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genom. 13, 375. http://dx.doi.org/10.1038/nbt.1754.

Chan, M.S., Maiden, M.C., Spratt, B.G., 2001. Database-driven multi locus sequence typing (MLST) of bacterial pathogens. Bioinformatics 17, 1077–1083.

Chen, Y., Toffaletti, D.L., Tenor, J.L., Litvintseva, A.P., Fang, C., Mitchell, T.G., McDonald, T.R., Nielsen, K., Boulware, D.R., Bicanic, T., Perfect, J.R., 2013. The *Cryptococcus neoformans* transcriptome at the site of human meningitis. MBio 5, e01087-13. http://dx.doi.org/10.1128/mBio.01087-13.

Colom, M.F., Hagen, F., Gonzalez, A., Mellado, A., Morera, N., Linares, C., García, D.F., Peñataro, J.S., Boekhout, T., Sánchez, M., 2012. *Ceratonia siliqua* (carob) trees as natural habitat and source of infection by *Cryptococcus gattii* in the Mediterranean environment. Med. Mycol. 50, 67–73. http://dx.doi.org/10.3109/13693786.2011.574239.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. http://dx.doi.org/10.1093/nar/gkh340.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S., 2009. Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138. http://dx.doi.org/10.1126/science.1162986.

Jackson, A.P., Gamble, J.A., Yeomans, T., Moran, G.P., Saunders, D., Harris, D., Aslett, M., Barrell, J.F., Butler, G., Citiulo, F., Coleman, D.C., de Groot, P.W.J., Goodwin, T.J., Quail, M.A., McQuillan, J., Munro, C.A., Pain, A., Poulter, R.T., Rajandream, M.-A., Renauld, H., Spiering, M.J., Tivey, A., Gow, N.A.R., Barrell, B., Sullivan, D.J., Berriman, M., 2009. Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. Genome Res. 19, 2231–2244. http://dx.doi.org/10.1101/gr.097501.109.

Janbon, G., Ormerod, K.L., Paulet, D., Byrnes, E.J., Yadav, V., Chatterjee, G., Mullapudi, N., Hon, C.-C., Billmyre, R.B., Brunel, F., Bahn, Y.-S., Chen, W., Chen, Y., Chow, E.W.L., Coppée, J.-Y., Floyd-Averette, A., Gaillardin, C., Gerik, K.J., Goldberg, J., Gonzalez-Hilarion, S., Gujja, S., Hamlin, J.L., Hsueh, Y.-P., Ianiri, G., Jones, S., Kodira, C.D., Kozubowski, L., Lam, W., Marra, M., Mesner, L.D., Mieczkowski, P.A., Moyrand, F., Nielsen, K., Proux, C., Rossignol, T., Schein, J.E., Sun, S., Wollschlaeger, C., Wood, I.A., Zeng, Q., Neuvéglise, C., Newlon, C.S., Perfect, J.R., Lodge, J.K., Idnurm, A., Stajich, J.E., Kronstad, J.W., Sanyal, K., Heitman, J., Fraser, J.A., Cuomo, C.A., Dietrich, F.S., 2014. Analysis of the genome and transcriptome of *Cryptococcus neoformans* var. *grubii* reveals complex RNA expression and microevolution leading to virulence attenuation. PLoS Genet. 10, e1004261. http://dx.doi.org/10.1371/journal.pgen.1004261.

Jiao, X., Zheng, X., Ma, L., Kutty, G., Gogineni, E., Sun, Q., Sherman, B.T., Hu, X., Jones, K., Raley, C., Tran, B., Munroe, D.J., Stephens, R., Liang, D., Imamichi, T., Kovacs, J.A., Lempicki, R.A., Huang, D.W., 2013. A benchmark study on error assessment and quality control of CCS reads derived from the PacBio RS. J. Data Min. Genom. Proteom. 4. http://dx.doi.org/10.4172/2153-0602.1000136.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., Phillippy, A.M., 2012. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. Nat. Biotechnol. 30, 693–700. http://dx.doi.org/10.1038/nbt.2280.

Kwon-Chung, K.J., Chang, Y.C., 2012. Aneuploidy and drug resistance in pathogenic fungi. PLoS Pathog. 8, e1003022. http://dx.doi.org/10.1371/journal.ppat.1003022.

Li, W., Raoult, D., Fournier, P.-E., 2009. Bacterial strain typing in the genomic era. FEMS Microbiol. Rev. 33, 892–916. http://dx.doi.org/10.1111/j.1574-6976.2009.00182.x.

Litvintseva, A.P., Mitchell, T.G., 2012. Population genetic analyses reveal the African origin and strain variation of *Cryptococcus neoformans* var. *grubii*. PLoS Pathog. 8, e1002495. http://dx.doi.org/10.1371/journal.ppat.1002495.

Litvintseva, A.P., Thakur, R., Vilgalys, R., Mitchell, T.G., 2006. Multilocus sequence typing reveals three genetic subpopulations of *Cryptococcus neoformans* var. *grubii* (serotype A), including a unique population in Botswana. Genetics 172, 2223–2238. http://dx.doi.org/10.1534/genetics.105.046672.

Litvintseva, A.P., Carbone, I., Rossouw, J., Thakur, R., Govender, N.P., Mitchell, T.G., 2011. Evidence that the human pathogenic fungus *Cryptococcus neoformans* var. *grubii* may have evolved in Africa. PLoS ONE 6, e19688. http://dx.doi.org/10.1371/journal.pone.0019688.

MacDougall, L., Kidd, S.E., Galanis, E., Mak, S., Leslie, M.J., Cieslak, P.R., Kronstad, J.W., Morshed, M.G., Bartlett, K.H., 2007. Spread of *Cryptococcus gattii* in British Columbia, Canada, and detection in the Pacific Northwest, USA. Emerg. Infect. Dis. 13, 42–50. http://dx.doi.org/10.3201/eid1301.060827.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. U.S.A. 95, 3140–3145.

Meyer, W., Aanensen, D.M., Boekhout, T., Cogliati, M., Diaz, M.R., Esposto, M.C., Fisher, M., Gilgado, F., Hagen, F., Kaocharoen, S., Litvintseva, A.P., Mitchell, T.G., Simwami, S.P., Trilles, L., Viviani, M.A., 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/10.1080/13693780902953886.

Park, B.J., Wannemuehler, K.A., Marston, B.J., Govender, N., Pappas, P.G., Chiller, T.M., 2009. Estimation of the current global burden of cryptococcal meningitis among persons living with HIV/AIDS. AIDS 23, 525–530. http://dx.doi.org/10.1097/QAD.0b013e328322ffac.

Quail, M.A.M., Smith, M.M., Coupland, P.P., Otto, T.D.T., Harris, S.R.S., Connor, T.R.T., Bertoni, A.A., Swerdlow, H.P.H., Gu, Y.Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genom. 13, 341. http://dx.doi.org/10.1186/1471-2164-13-341.

Schwartz, D.C., Cantor, C.R., 1984. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell 37, 67–75.

Selmecki, A.M., Dulmage, K., Cowen, L.E., Anderson, J.B., Berman, J., 2009. Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance. PLoS Genet. 5, e1000705. http://dx.doi.org/10.1371/journal.pgen.1000705.

Simwami, S.P., Khayhan, K., Henk, D.A., Aanensen, D.M., Boekhout, T., Hagen, F., Brouwer, A.E., Harrison, T.S., Donnelly, C.A., Fisher, M.C., 2011. Low diversity *Cryptococcus neoformans* variety *grubii* multilocus sequence types from Thailand are consistent with an ancestral African origin. PLoS Pathog. 7, e1001343. http://dx.doi.org/10.1371/journal.ppat.1001343.

Stephen, C., Lester, S., Black, W., Fyfe, M., Raverty, S., 2002. Multispecies outbreak of cryptococcosis on southern Vancouver Island, British Columbia. Can. Vet. J. 43, 792–794.

Sun, S., Hsueh, Y.-P., Heitman, J., 2012. Gene conversion occurs within the mating-type locus of *Cryptococcus neoformans* during sexual reproduction. PLoS Genet. 8, e1002810. http://dx.doi.org/10.1371/journal.pgen.1002810.

Tavanti, A., Davidson, A.D., Johnson, E.M., Maiden, M.C.J., Shaw, D.J., Gow, N.A.R., Odds, F.C., 2005. Multilocus sequence typing for differentiation of strains of *Candida tropicalis*. J. Clin. Microbiol. 43, 5593–5600. http://dx.doi.org/10.1128/JCM.43.11.5593-5600.2005.

Taylor, J.W., Fisher, M.C., 2003. Fungal multilocus sequence typing – it's not just for bacteria. Curr. Opin. Microbiol. 6, 351–356. http://dx.doi.org/10.1016/S1369-5274(03)00088-2.

Travers, K.J., Chin, C.-S., Rank, D.R., Eid, J.S., Turner, S.W., 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. 38, e159. http://dx.doi.org/10.1093/nar/gkq543.

Vanhee, L.M.E., Nelis, H.J., Coenye, T., 2010. What can be learned from genotyping of fungi? Med. Mycol. 48 (Suppl. 1), S60–S69. http://dx.doi.org/10.3109/13693786.2010.484816.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 23, 4407–4414.

Xu, J.J., Yan, Z.Z., Guo, H.H., 2009. Divergence, hybridization, and recombination in the mitochondrial genome of the human pathogenic yeast *Cryptococcus gattii*. Mol. Ecol. 18, 2628–2642. http://dx.doi.org/10.1111/j.1365-294X.2009.04227.x.