Fungal Genetics and Biology xxx (2015) xxx–xxx

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 Farm 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 xxxx

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,

* Corresponding authors at: Division of Infectious Diseases, Department of Medicine, Duke University Medical Center, Durham, NC, USA (Y. Chen). 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/

Please cite this article in press as: Chen, Y., et al. Next generation multilocus sequence typing (NGMLST) and the analytical software program MLSTEZ enable efficient, cost-effective, high-throughput, multilocus sequencing typing. Fungal Genet. Biol. (2015), http://dx.doi.org/10.1016/j.fgb.2015.01.005

ecology, and microbiomes (Byrnes et al., 2009; Chen et al., 2013; Litvintseva and Mitchell, 2012; Meyer et al., 2009). Moreover, 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′-GCTGTCAACGATACGCTACG) 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.

Please cite this article in press as: Chen, Y., et al. Next generation multilocus sequence typing (NGMLST) and the analytical software program MLSTEZ enable efficient, cost-effective, high-throughput, multilocus sequencing typing. Fungal Genet. Biol. (2015), http://dx.doi.org/10.1016/j.fgb.2015.01.005