doi:10.5598/imafungus.2018.09.01.07 IMA FUNGUS · 9(1): 91–105 (2018)

ARTICLE

NGS barcode sequencing in taxonomy and diagnostics, an application in "*Candida*" pathogenic yeasts with a metagenomic perspective

Claudia Colabella[1], Laura Corte[1], Luca Roscini[1], Matteo Bassetti[2], Carlo Tascini[3], Joseph C. Mellor[4], Wieland Meyer[5], Vincent Robert[6], Duong Vu[6], and Gianluigi Cardinali[1,7]

[1]Microbiology Section, Department of Pharmaceutical Sciences, University of Perugia, 06121, Italy
[2]Infectious Diseases Division, Santa Maria Misericordia University Hospital, Udine, 33100, Italy
[3]Infectious Diseases Division, Cotugno Hospital Napoli, 80131, Italy
[4]seqWell Inc., 376 Hale Street, Beverly, MA 01915, USA
[5]Molecular Mycology Research Laboratory, Centre for Infectious Diseases and Microbiology, Sydney Medical School, Westmead Hospital, Marie Bashir Institute for Infectious Diseases and Biosecurity, The University of Sydney, Westmead Institute for Medical Research, Sydney, NSW 2006, Australia
[6]Bioinformatics Unit, Westerdijk Fungal Biodiversity Institute, 3508 CT, Utrecht, Netherlands
[7]CEMIN Research Centre of Excellence, University of Perugia, Borgo 20 Giugno 74, 06121, Italy; corresponding author e-mail: gianluigi.cardinali@unipg.it

Abstract: Species identification of yeasts and other Fungi is currently carried out with Sanger sequences of selected molecular markers, mainly from the ribosomal DNA operon, characterized by hundreds of tandem repeats of the 18S, ITS1, 5.8S, ITS2 and LSU *loci*. The ITS region has been recently proposed as a primary barcode marker making this region the most used one in taxonomy, phylogeny and diagnostics. The introduction of NGS is providing tools of high efficacy and relatively low cost to amplify two or more markers simultaneously with great sequencing depth. However, the presence of intra-genomic variability between the repeats requires specific analytical procedures and pipelines. In this study, 286 strains belonging to 11 pathogenic yeasts species were analysed with NGS of the region spanning from ITS1 to the D1/D2 domain of the LSU encoding ribosomal DNA. Results showed that relatively high heterogeneity can hamper the use of these sequences for the identification of single strains and even more of complex microbial mixtures. These observations point out that the metagenomics studies could be affected by species inflection at levels higher than currently expected.

Key words:
identification
ITS
LSU
Next generation sequencing
Sanger

Article info: Submitted: 21 July 2017; Accepted: 10 May 2018; Published: 22 May 2018.

INTRODUCTION

The regions encoding for the ribosomal DNA in yeasts are organized in an array ranging from 10 to 20 kb according to the species, including the small subunit (18S), the internal transcribed spacers (ITS1 and ITS2) and the large subunit (5S, 5.8S and 26S) (Dujon *et al.* 2004). These arrays vary in number from a few dozen to hundreds (Maleszka & Clark-Walker 1993, Amend *et al.* 2010). In *Candida albicans*, the operon is 12756 bp long and the diploid genomes can contains 110 repeats in a single locus (Jones *et al.* 2004), whereas in *Candida glabrata* these sequences are dispersed in two subtelomeric regions (Maleszka & Clark-Walker 1993, Dujon *et al.* 2004). This difference reflects the taxonomic distance (*C. albicans* belongs to *Debaromycetacee* whereas *C. glabrata* belongs to *Saccharomycetaceae*) and the evolution, the former being a pre-genome-duplication species and the latter a post-genome-duplication species (Dujon 1996). The sequences of these genes have been largely used in the last decades in taxonomy and phylogenetic studies thanks to their high conservation (Kurtzman & Robnett 1998, Groenewald *et al.* 2011), by means of Sanger sequencing that produces a single sequence of each gene. However, secondary peaks have been observed suggesting some level of heterogeneity among the various copies of the tandem repeats in Fungi (Korabecna 2007, Woo *et al.* 2010, Vydryakova *et al.* 2012, Li *et al.* 2014), yeasts (Alper *et al.* 2011), ciliates (Gong *et al.* 2013) and in some plants (Wang *et al.* 2015). The extent of this variability is critical for the exact understanding governing the homogenization of multigene families, which is typically attributed to concerted evolution, i.e. the internal evolution of the members of a multigene family that mutate over the time more or less in the same way. An effect is that the paralogous genes within a species are much closer to one another than to the same genes in other species. This phenomenon can be explained by gene conversion or asymmetric crossing-over (Liao 1999, Nei & Rooney 2005,

Ganley & Kobayashi 2007, Naidoo *et al.* 2013). Another model to explain the copy homogenization is the birth-and-death model, according which new versions of the genes appear with the "birth" transition and all the variants, but one (or very few) are purged during the "death" transition (Nei & Rooney 2005). However, the possibility of explaining this homogenization of the rDNA loci with the birth-and-death model is still a matter of debate, and some authors have claimed that the variation observed fits this model better than that of concerted evolution (Nei & Rooney 2005). In general, it is possible that different mechanisms have been active in different taxa, and maybe even in different regions (Vydryakova *et al.* 2012). A mixed model of evolution involving both models simultaneously was considered, although not for this gene family (Nei & Rooney 2005). The major differences between these models are that concerted evolution is expected to produce scarce heterogeneity, whereas the birth-and-death mechanism should yield more Intra-Genomic Polymorphisms (IGP) (Ganley & Kobayashi 2007). From the above observations it appears that concerted evolution is not a satisfactory model when IGPs frequency is particularly high (Simon & Weiß 2008).

Beyond the model governing the homogenization of the repeat units, the internal variability within the rDNA is a source of additional information useful in phylogenetic, environmental, and clinical microbiology, to trace the origin of strains (West *et al.* 2014). As long as Sanger sequencing was the sole or predominant technology, the sequence reported the most frequent nucleotides hiding the least frequent, occasionally visible as secondary peaks (Woo *et al.* 2010).

Since ITS has been agreed as the universal barcode for *Fungi* (Schoch *et al.* 2012), the possibility of applying Next Generation Sequencing (NGS) to this locus offers several advantages, such as the possibility of studying microbial communities, independently of their viability and capacity of growing on existing media (Bokulich & Mills 2012, Hajibabaei 2012). Problems still exist in the exact quantification of taxa on the basis of NGS read abundance (Amend *et al.* 2010), and care should be taken in data analysis because the internal heterogeneity could cause an inflation of the species richness (Lindner & Banik 2011, Lindner *et al.* 2013). Since database quality and completeness are mandatory for NGS analyses (Bokulich & Mills 2012), the presence of a few alternative barcode markers proposed and under evaluation (Stielow *et al.* 2015) leads to the use of the rDNA genes for identification, although a much deeper understanding of the problems and effective analytical pipelines are necessary (Medinger *et al.* 2010).

NGS technology is now sufficiently mature to move from specialized research centres to environmental and clinical laboratories. However, the massive amount of data and the internal heterogeneity can be a serious problem to get sound and rapid high-throughput identifications (Ahmed 2016), especially when a more complex metagenomics approach is taken to describe microbial communities in patients and healthy controls (Imabayashi *et al.* 2016). Among the various approaches described, a BLAST search followed by an assembly that is further tested with BLAST has been recently described and proposed (Ahmed 2016, Imabayashi *et al.* 2016).

In this paper we describe an innovative system of yeast strain identification using next generation sequencing of amplicons covering the region spanning the ITS1 to the D1/D2 domain of the LSU. The former region is that accepted as the universal barcode in Fungi (Schoch *et al.* 2012) and is now included in highly curated databases (Schoch *et al.* 2014, Irinyi *et al.* 2015); the latter was the first locus introduced in yeast identification (Kurtzman & Robnett 2013, Susca *et al.* 2013, Yurkov *et al.* 2015). These regions were amplified simultaneously in single amplicons in order test two hypotheses: (1) the possibilities offered by NGS for amplicon based single strain identification; and (2) the accuracy of this technique in a metagenomic study. The latter concept is based on the multi-copy nature of these markers, and their intrinsic endogenous variability, which could lead to the erroneous attribution of the reads to the right species and also to other ones. With our settings, it is possible to evaluate the rate of erroneous identifications caused by the above issues.

The analysis was performed with 286 strains of pathogenic yeasts isolated from two Italian hospitals, which had previously been studied to show that the success of these strains in the hospital environment is strictly related to their ability to form biofilms (Corte *et al.* 2016). This set of strains is large enough to represent the identification routine occurring in a clinical setting, as well as in other environmental studies, and to pave the way to further studies on the composition of repeats present in the rDNA region.

MATERIALS AND METHODS

Strains and growth conditions

The 286 strains of opportunistic *Candida* species analysed (Table S1) were isolated from patient blood cultures obtained in two Italian hospitals in Pisa and Udine; these had been incorporated into the Cemin Microbial Collection of the Microbial Genetics and Phylogenesis Laboratory (Cemin, Centre of Excellence on Nanostructured Innovative Materials for Chemicals, Physical and Biomedical Applications, University of Perugia). *Candida albicans*, *C. parapsilosis*, *C. tropicalis,* and *C. glabrata* represented the majority of the isolates. The first three species are phylogenetically related among themselves, where *C. glabrata* is more closely related to *S. cerevisiae*. Strains were stored at –80 °C in 17 % glycerol immediately after isolation. Short-term storage was carried out on YEPDA (YEPD supplemented with 1.7 % agarose) at 4°C. Strains were grown in YEPD (yeast extract 1 %, peptone 1 %, and dextrose 1 %); all products were obtained from Biolife (http://www.biolifeitaliana.it/) and kept at 37 °C with 150 rpm shaking.

DNA extraction and molecular techniques

Genomic DNA was extracted as indicated by Cardinali *et al.* (2001). The ITS1, 5.8S, ITS2 rDNA gene cluster regions and the D1/D2 domain of the LSU gene were amplified with FIREPole® Taq DNA Polymerase (Solis BioDyne, Estonia), using the ITS1 (5'-TCCGTAGGTGAACCTGCGG) - NL4 (GGTCCGTGTTTCAAGACGG) primers pair. The amplification protocol was carried out as follows: initial

denaturation at 94 °C for 3 min, 30 amplification cycles (94 °C for 1 min, 54 °C for 1 min and 72 °C for 1 min) and final extension at 72 °C for 5 min. Amplicons were subjected to electrophoresis a 1.5 % agarose gel (Gellyphor, EuroClone, Italy). Amplicons were sequenced with NGS plexWell™ technologies (http://www.seqwell.com/) with the same primers used for the generation of the amplicons. The reads of each strain, contained in FASTQ file, were analysed with Geneious R9 (Kearse *et al.* 2012).

Bioinformatics analysis

Mapping against a reference vs de novo *assembling*

Mapping and assembling analyses were carried out on 12 strains (Table 1), representative of six species showing different numbers of reads, ranging from 19 388 to 53 497. Each strain was mapped against the ITS-LSU concatenate rDNA sequences of the respective ex-type strain, used as reference and provided by the fungal collection database of the Westerdijk Fungal Biodiversity Institute (CBS): CBS 562 *C. albicans*, CBS 138 *C. glabrata*, CBS 604 *C. parapsilosis*, CBS 10906 *C. orthopsilosis*, CBS 94 *C. tropicalis,* and CBS 2030 *Meyerozyma guilliermondii*.

The 12 FASTQ files were filtered to remove reads shorter than 140 bp. For the mapping analysis, contigs were obtained using two algorithms: Bowtie2 with setting "local" (hereinafter referred to as BTL) searching only for the best match (Langmead & Salzberg 2012) and BBMap (BBm) set to mapping as multiple best matching in random mode (Bushnell 2014). The two algorithms can align reads from all major NGS platforms; BTL is typically used with short reads and large reference sequences while BBm is particularly used with highly mutated sequences or reads with long indels; a comparison between the two algorithms has been reported (Berrocal *et al.* 2016, Lubock *et al.* 2017). Mappings were performed using High Sensitivity mode and no trimming of the sequences. *De novo* assembling was performed using Geneious assembler after testing other algorithms, with High and Low sensitivity settings without trimming. Contig identification was carried out with BLAST search using a local library containing 15 ITS-LSU concatenate rDNA sequences of the *Candida* ex-type strains (Table 2).

The *Saccharomyces cerevisiae* CBS 1171 ex-type strain was used as outgroup. The highest similarity matches were searched for using Megablast.

Mapping against a reference using large libraries

Mapping procedures were performed using strains CMC 1793 and CMC 1818 with 21 238 and 58 263 reads respectively and three selected and accurate libraries: Westerdijk Fungal Biodiversity Institute collection database (CBS) containing ITS sequences and ITS-LSU sequences of ex-type strains, and the ISHAM-ITS database (http://its.mycologylab.org/) containing ITS sequences of medical related strains. The two FASTQ files were mapped against the different libraries using the two algorithms BTL setting "local" and BBm with High Sensitivity mode and no trimming of the sequences.

Mapping against a selected library (M1)

All the 286 FASTQ files were filtered to remove reads shorter than 140 bp and were mapped against the local library of 16 ITS-LSU concatenate rDNA sequences of ex-type strains using two algorithms (BTL and BBm) with High Sensitivity mode and no trimming. All the 16 ITS-LSU rDNA sequences were provided by the fungal collection database of the Westerdijk Fungal Biodiversity Institute (CBS). Results were exported from Geneious R9 software in Microsoft Excel®. Using a built-in macro, six indices were calculated for both algorithms and data were assembled in order to give an unambiguous taxonomic interpretation of the results.

- *Iread*: This index is the ratio of the number of reads attributed to each member of the reference library, and therefore indicates the share of reads (R_i) attributable to the species represented by the ex-type strain (R).

- *Inuc*: This index is the ratio of the nucleotides mapped with the ex-type strain (N_i) on the total number of nucleotides present on the reads of the FASTQ file (N).

- *Icov*: This index of coverage describes the average number of reads that align to, or "cover", known reference bases. At higher levels of coverage, each base is covered by a greater number of aligned sequence reads, so base calls can be made with increased confidence. The index is represented by the ratio of coverage value of the strain (C_i) on the total coverage of the reads (C).

- *Iref*: Refseq percentage.

- *Isim*: Similarity percentage.

- *Isyn*: This index represents the ratio of the sum of all the five indices of one species (Isp) to the sum of the indices of all the species (Itot).

Mapping against the ex-type strain of the presumptive species and "unused reads" (M2 followed by the final M3 mapping)

All the 286 FASTQ files were mapped against the ITS-LSU concatenated sequences of the ex-type strains resulting from the first mapping (M1). The second mapping procedure was performed using the two algorithms (BTL and BBm) with High Sensitivity mode and no trimming. Results were exported from Geneious R9 software in Microsoft Excel®. The six indices applied in the first mapping were used to facilitate a taxonomic interpretation of the results. The reads that did not match with sequences of the local library were filtered in order to remove reads shorter than 140 bp and re-mapped against the local library using the same settings and indices of the first and second mapping (M1 and M2, respectively).

Table 1. Output parameters of *de novo* assembly and mapping.

		Mapping		De novo assembling	
		BTL	**BBM**	**High sensitivity**	**Low sensitivity**
C. albicans CMC 1853 23,810 reads	n° reads assembled	23,389	22,755	23,721	23,674
	n° reads not assembled	421	1,055	89	136
	Assembly duration	7.19 seconds	21.58 seconds	38 minutes-10 seconds	12 minutes-29 seconds
	CPU time	8.54 seconds	12.49 seconds	2h-11 minutes	41 minutes-15 seconds
	Contigs	1	1	119	219
C. albicans CMC 1856 53,497 reads	n° reads assembled	52,270	51,599	52,187	53,022
	n° reads not assembled	1,227	1,898	1,310	475
	Assembly duration	11.47 seconds	33.06 seconds	47 minutes+34 seconds	10 minutes-46 seconds
	CPU time	13.37 seconds	17.31 seconds	2h-49 minutes	34 minutes-1 second
	Contigs	1	1	242	357
C. glabrata CMC 1837 19,762 reads	n° reads assembled	18,883	18,257	19,454	19,334
	n° reads not assembled	879	1,505	308	428
	Assembly duration	5.10 seconds	21.01 seconds	26 minutes-10 seconds	10 minutes-8 seconds
	CPU time	6.68 seconds	10.52 seconds	1h-33 minutes	34 minutes-2 seconds
	Contigs	1	1	121	209
C. glabrata CMC 1934 49,337 reads	n° reads assembled	47,611	46,352	48,897	48,479
	n° reads not assembled	1,726	2,985	440	858
	Assembly duration	10.48 seconds	34.93 seconds	1h-40 minutes	21 minutes-11 seconds
	CPU time	11.48 seconds	21.11 seconds	5h-20 minutes	1h-11 minutes
	Contigs	1	1	243	470
C. orthopsilosis CMC 1880 20,089 reads	n° reads assembled	19,533	19,472	19,960	19,923
	n° reads not assembled	556	617	129	166
	Assembly duration	7.69 seconds	27.36 seconds	34 minutes-27 seconds	11 minutes-5 seconds
	CPU time	10.98 seconds	22.71 seconds	1h-58 minutes	37 minutes-37 seconds
	Contigs	1	1	119	236
C. orthopsilosis CMC 2011 36,955 reads	n° reads assembled	36,136	36,019	36,720	36,591
	n° reads not assembled	819	936	235	364
	Assembly duration	11.45 seconds	38.25 seconds	59 minutes-5 seconds	15 minutes-47 seconds
	CPU time	7.91 seconds	14.50 seconds	3h-20 minutes	51 minutes-58 seconds
	Contigs	1	1	219	477
C. parapsilosis CMC 1838 37,683 reads	n° reads assembled	36,717	36,584	37,683	37,432
	n° reads not assembled	966	1,099	1,264	251
	Assembly duration	12.19 seconds	27.66 seconds	44 minutes-3 seconds	15 minutes-11 seconds
	CPU time	17.55 seconds	29.27 seconds	2h-33 minutes	51 minutes-48 seconds
	Contigs	1	1	190	335
C. parapsilosis CMC 2039 23,817 reads	n° reads assembled	21,549	21,502	22,569	22,336
	n° reads not assembled	2,268	2,315	1,248	1,481
	Assembly duration	7.24 seconds	19.88 seconds	41 minutes-34 seconds	32 minutes-57 seconds
	CPU time	5.73 seconds	11.48 seconds	2h-24 minutes	1h-1 minute
	Contigs	1	1	281	632
C. tropicalis CMC 2003 44,744 reads	n° reads assembled	41,841	41,361	43,253	44,140
	n° reads not assembled	2,903	3,383	1,491	604
	Assembly duration	12.23 seconds	31.09 seconds	43 minutes-4 seconds	15 minutes-9 seconds
	CPU time	9.70 seconds	14.13 seconds	2h-28 minutes	45 minutes-59 seconds
C. tropicalis CMC 2017 20,125 reads	n° reads assembled	18,007	17,593	19,671	19,550
	n° reads not assembled	2,118	2,532	454	575
	Assembly duration	6.07 seconds	22.98 seconds	33 minutes-32 seconds	11 minutes-47 seconds
	CPU time	5.45 seconds	8.12 seconds	1h-56 minutes	39 minutes-35 seconds
	Contigs	1	1	369	527

Table 1. (Continued).

		Mapping		De novo assembling	
		BTL	**BBM**	**High sensitivity**	**Low sensitivity**
M. guilliermondii CMC 1825 40,659 reads	n° reads assembled	36,085	35,598	40,094	39,914
	n° reads not assembled	4,574	5,061	565	745
	Assembly duration	12.25 seconds	27.05 seconds	51 minutes-46 seconds	19 minutes-55 seconds
	CPU time	15.88 seconds	15.43 seconds	2h-58 minutes	1h-5 minutes
	Contigs	1	1	507	711
M. guilliermondii CMC 1924 19,388 reads	n° reads assembled	17,593	17,335	19,195	19,152
	n° reads not assembled	1,795	2,053	193	236
	Assembly duration	6.33 seconds	19.18 seconds	22 minutes-39 seconds	9 minutes-18 seconds
	CPU time	7.02 seconds	9.02 seconds	1h-20 minutes	31 minutes-10 seconds
	Contigs	1	1	129	248

Table 2. Type strains employed for contigs identification.

Species	Strain
Candida albicans	CBS 562
Candida dubliniensis	CBS 7987
Candida famata	CBS 1795
Candida glabrata	CBS 138
Issatchenkia orientalis	CBS 573
Candida metapsilosis	CBS 10907
Candida orthopsilosis	CBS 10906
Candida parapsilosis	CBS 604
Candida pararugosa	CBS 1010
Diutina rugosa	CBS 613
Candida sake	CBS 159
Candida tropicalis	CBS 94
Candida utilis	CBS 621
Candida lusitaniae	CBS 6936
Meyerozyma guilliermondii	CBS 2030
*Saccharomyces cerevisiae**	CBS 1171

*Outgroup type strain.

RESULTS

Mapping against a reference *vs de novo* assembly: different efficiency in terms of time

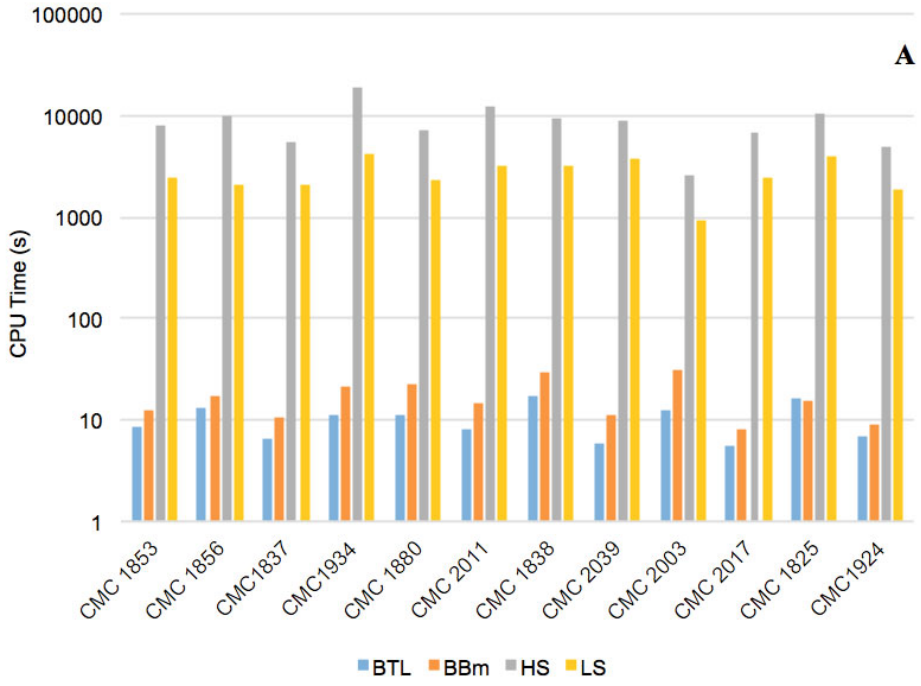
The reads contained in a FASTQ file can be analysed with the *de novo* assembly or by mapping to a reference, each with a variety of algorithms and settings. The former approach does not theoretically require any *a priori* knowledge. Moreover, it could produce contigs of the reads without the bias due to reference sequences non including the actual species to identify. On the other hand, assembling thousands of reads deriving from hundreds of repeats without a reference could produce inaccurate assemblies. We tested these two approaches by comparing the accuracy obtained and the computation time requested. The mapping was carried out using an *ad hoc* library containing the ITS-LSU region of the ex-type strains of 16 yeast species, most of which are known pathogens. These two analyses were carried out on 12 strains, representative of six species and exhibiting a different number of reads, ranging from 19 388 to 53 497. These analyses showed that the *de novo* assembly takes much more than the mapping in terms of the total time necessary to carry out the operation using an i7 Intel ^{R} processor with 8Gb RAM and the Geneious 9 interface (Table 1). The CPU time necessary for the two types of analysis diverges by two or three orders of magnitude (Fig. 1A). Within the two different approaches, BTL showed lower processing times than BBm; the former carried out the operation with an average of 0.329 milliseconds per read, *vs* 0.544 milliseconds for the latter. More interestingly, the standard deviations of the BTL and BBm analyses were 0.120 and 0.226 milliseconds respectively (corresponding to 0.31 and 0.42 variation coefficients).

Using the *de novo* approach with the Geneious algorithm, the CPU time required was 282.38 and 91.27 milliseconds per read, with the high and low sensitivity settings (HS and LS), respectively. Even in this case, a large difference was observed between settings; the HS had a standard deviation of 91.9 milliseconds (variation coefficient 0.33) and the LS 34.9 standard deviation, corresponding to a 0.38 variation coefficient.

Altogether, these data indicate that the *de novo* assembly takes much more computational time than mapping against a reference sequence. The variability observed between the performances of four algorithms recognized in our study posed the question of the influence of the number of reads on the operational time required. Surprisingly there was a low correlation between the CPU time and the number of reads: 0.67 BTL, 0.56 BBm, 0.52 HS, and 0.19 LS. Finally, the correlation analysis of the computational time required by the four algorithms showed relatively high correlation values between BTL and BBm (0.741) and between the two *de novo* procedures (0.827) (Fig. 1B). These data indicate that the time required by mapping and assembly are independent, whereas a weak relation exists between the algorithms employed within the same type of approach.

Contigs quality obtained with "mapping against a reference" and "*de novo* assembly"

The contigs obtained with the two methods were analysed with a local blasting using an *ad hoc* library containing sequences of the ex-type strains of 16 yeast species. Typical results (Fig. 2A–B) showed that high levels of



B

	BTL	BBm	HS	LS
BTL				
BBm	**0.741**			
HS	0.242	0.038		
LS	0.174	-0.124	**0.827**	

Fig. 1. Evaluation of the computational time requested for the two different approaches (Mapping against a Reference and *De novo* Assembly). **A.** CPU time needed by the two different types of procedures with the four settings. **B.** Correlation between the four different algorithms.

nucleotide similarity were obtained between the single contig of mapping with most of the library members. These nucleotide similarities spanned from approximately 80 % to 99.6 % of the correctly identified species (Fig. 2A). In this analysis no major differences were observed between the BTL and the BBm algorithms. On the contrary, blasting the several contigs derived from the *de novo* assembly produced different putative identifications with homologies spanning from approximately almost 0 % to 13 %, whereas the correct species displayed 69.4 % and 74.6 % homology with the high and low sensitivity algorithms. These types of results deriving from the two approaches were confirmed for all the strains analysed. For the mapping approach, the similarities with the correct species were in the range between 97.88 % and 99.8 %, whereas the homology shown with the second more similar species varied between 89.9% and 98.8% using BTL and BBm algorithms (Fig. 2C). The blasting of the contigs of the 12 strains obtained with *de novo* assembly gave homologies with the correct species ranging from 52.4–87.5 % and from 48.6 –90.5 % with the high and the low sensitivity algorithms, respectively (Fig. 2D). The second most similar species showed homologies in the ranges 3.8–38 % and 3.8–34 %, respectively with the HS and LS algorithms.

Feasibility of "mapping against a reference" with large libraries

From the above results, it was clear that the use of *de novo* assembly is time consuming and produced relatively low homologies to the expected species. These two aspects suggested a more detailed analysis of the data to determine the possibilities offered by the mapping when the reference is represented by a large library of sequences. Ideally, these reference sequences should contain the sequences from the type strains of all known species, in order to avoid the presence of misidentified strains that would lead not only to an incorrect identification, but to an inflation of misidentifications. For this reason, *ad hoc,* highly curated libraries are produced and maintained, such as the Westerdijk Fungal Biodiversity Institute (http://www.westerdijkinstitute.nl/), a fungal reference library (RefSeq) within the NIH-GenBank (Schoch *et al.* 2014), or the dedicated ITS library of ISHAM for human and animal pathogenic fungal identifications (Irinyi *et al.* 2015).

In order to test the efficacy of mapping against a reference library using large collections of sequences, we used three libraries of different size: (1) a curated library of type strains from the Westerdijk Fungal Biodiversity Institute (CBS) containing only ITS sequences (15 565 sequences); (2) a curated library with both ITS and LSU D1/D2 sequences from Westerdijk Fungal Biodiversity Institute (CBS) (34 683 sequences); and (3) the ISHAM-ITS database (2727 sequences). The strains CMC 1793 and CMC 1818 with 21

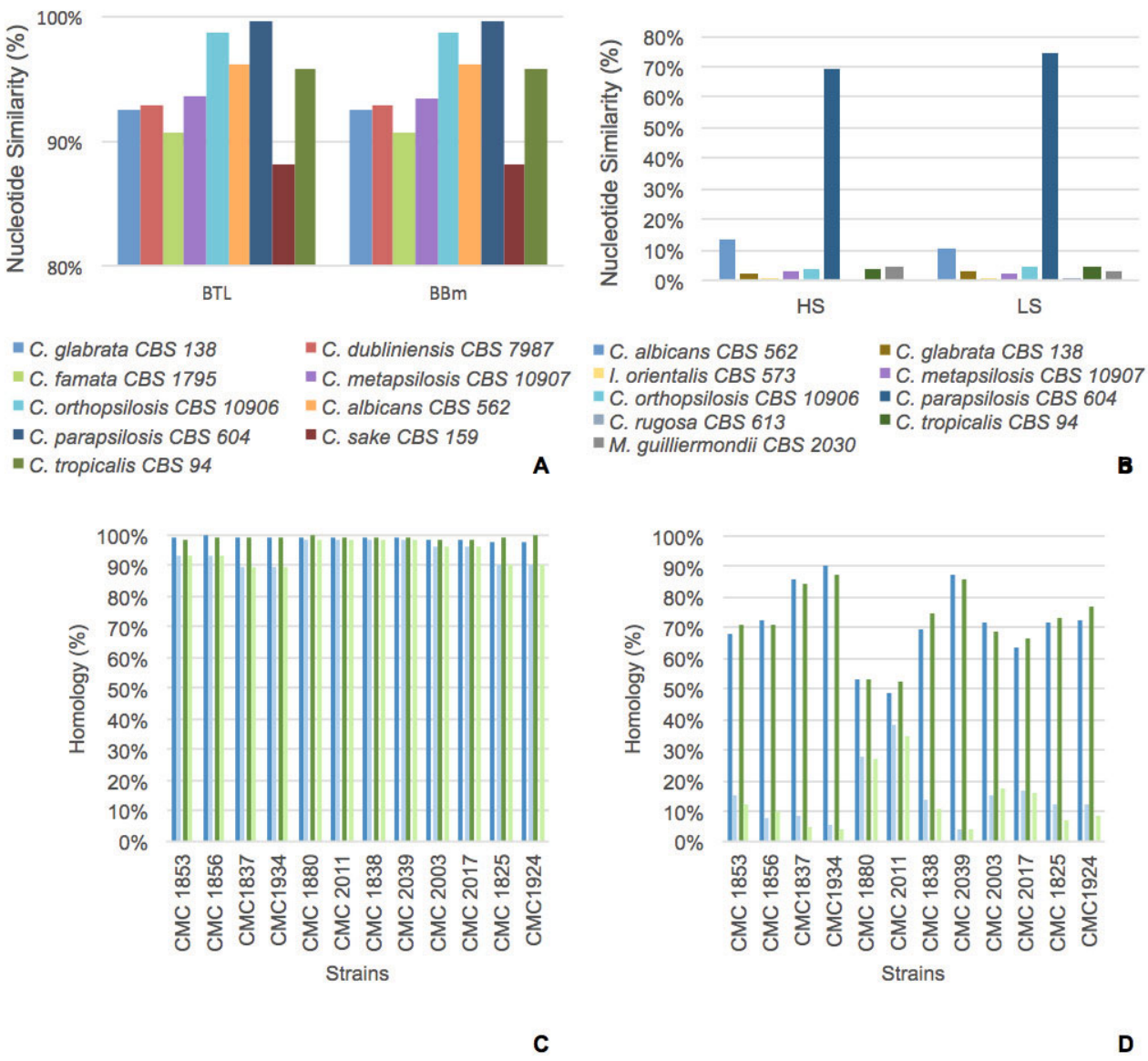


Fig. 2. Analysis of contigs quality. **A.** Similarity between a single contig of each of the two mapping algorithms and the members of the reference library. **B.** Variation of the similarity with the members of the library using High Sensitivity (HS) or Low Sensitivity (LS) algorithms. **C.** Homology of the contigs with the first and the second most similar species using BTL (dark-light blue) and BBm (dark-light green). **D.** Homology of the contigs analysed with High Sensitivity (dark-light blue) or Low Sensitivity (dark-light green) algorithms.

238 and 58 263 reads respectively were mapped against the three libraries using BTL and BBm algorithms. This scheme produced 12 mapping combinations that were tested in order to define the feasibility of using large libraries and these performance parameters.

Using an i7 Intel processor with 8Gb Ram and the Geneious 9 interface, the minimum CPU time was 13.59 s and the maximum 393 s. The average time for CMC 1793 mapping was 52.65 s, whereas CMC 1818 (with almost three times more reads) required an average time of 131.02 s. The time performances of the three libraries varied, as expected, according to their size, expressed as number of sequences. The relatively small ISHAM-ITS library required an average of 22.31 s, whereas the average CPU times of the two Westerdijk Fungal Biodiversity Institute (CBS) libraries were 61.43 and 191.75 s for the Westerdijk Fungal Biodiversity Institute CBS-ITS and Westerdijk Fungal Biodiversity Institute CBS ITS-LSU.

The time performance of the two tested algorithms changed according to the size of both the library and the FASTQ file (Fig. 3A). BBm was faster than BTL, especially when processing the large CMC 1818 file with over 58 000 reads, whereas it was slightly slower with the smaller FASTQ file from CMC 1793. The processing rate was obviously conditioned by the library size as well. Both algorithms showed the largest reads/second values with the ISHAM-ITS and the smallest with the very large Westerdijk Fungal Biodiversity Institute CBC ITS-LSU library. Taking together these observations, we tested the hypothesis that the reads/sec. processing rate may be function of the number of reads of the FASTQ file and the number of sequences of the library used. The regression analysis, carried out for BTL and BBm

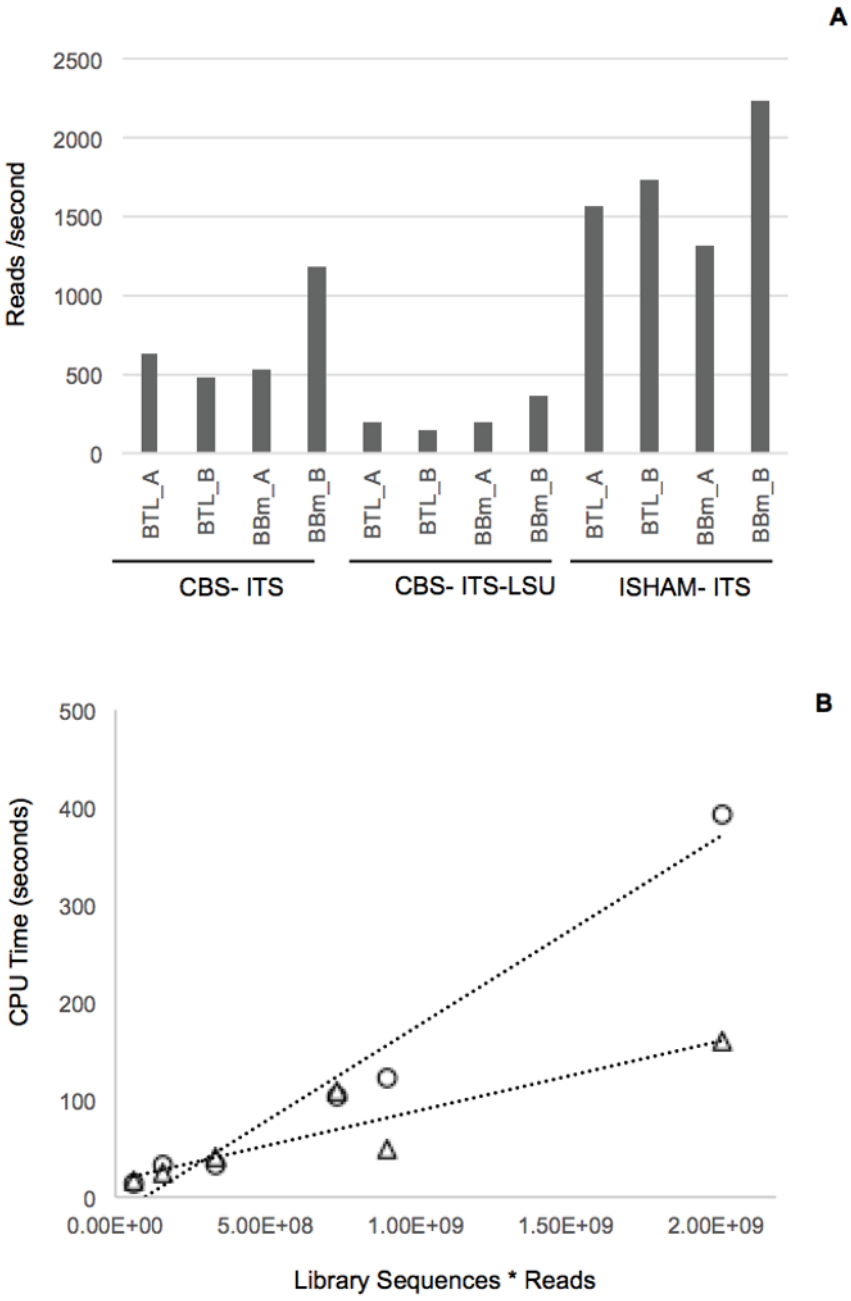


Fig. 3. Time performance of the mapping algorithms against three large libraries. **A.** Variation of the time performance using references and files of different size. **B.** Regression analysis between the time performance, the dimension of library and FASTQ files using the BTL (circle) and BBm (triangle) algorithms.

separately, yielded 0.9702 and 0.8354 R2 values for BTL and BBm, respectively (Fig. 3B). Taken together, these data indicated that mapping against a reference is feasible in terms of time and that it is more convenient than *de novo* assembly, even if large libraries and FASTQ files are employed. This analysis showed that the use of a large library as reference returns homology values ranging from 60–100 % (Table 3), indicating that a careful analytical protocol is necessary to discriminate the taxonomically positive identifications.

A pipeline to optimize the mapping against a reference

The tests described above showed that, even using large libraries, mapping is faster than *de novo* assembling. Furthermore, the levels of homology typically found with Sanger sequencing are closer to the data yielded by mapping than those produced by *de novo* assembling. These characteristics suggested the development of a pipeline to take into consideration all aspects emerging from our tests and accounting for the typical conditions in which identifications are carried out. The first step is mapping against a reference (mapping M1) to indicate the most likely species to which the strain belongs. Since mappings can be used with a wide range of parameters and their examination is quite difficult, especially when the libraries and the FASTQ files are large, we developed a series of indices, including a final synthetic index *Isyn* (Table 4A–B). These indices are reported as percentages, for easier reading. All the calculated indices are consistently higher with the expected species (e.g *C. glabrata* in Table 4A) as indicated by standard deviations not higher than 1.5 %. The reads of the conserved sequences are often very similar to more than one member of the library. This produces a biased decrease of all indices of the species to whom the unknown strain is expected to be attributed. This means that when the library increases in

Table 3. Performances of algorithms in mapping FASTQ files of different size with large reference libraries.

Library		Algorithm	FASTQ		Time Performances		Mapping parameters				
acronym	sequences		strain	reads	Mapping time (sec)	CPU-time (sec)	used reads	unused reads	Matches	min pairwise identities	max pairwise identities
	15,565	BTL	A	21,238	63	33,34	18,274	2,964	684	74,00%	100%
CBS			B	58,263	99	123	54,822	3,441	661	91,00%	100%
ITS		BBm	A	21,238	80	39,82	18,323	2,915	626	68,20%	100%
			B	58,263	114	49,57	55,675	2,588	708	62,30%	100%
	34,683	BTL	A	21,238	186	104	19,163	2,075	2,051	73,40%	100%
CBS			B	58,263	273	393	57,140	1,123	2,457	79,50%	100%
ITS-LSU		BBm	A	21,238	204	109	19,000	2,238	1,939	58,80%	100%
			B	58,263	311	161	56,930	1,333	2,445	63,60%	100%
	2,727	BTL	A	21,238	16,7	13,59	13,904	7,334	319	74,10%	100%
ISHAM			B	58,263	30,3	33,5	36,416	21,847	297	82,00%	100%
ITS		BBm	A	21,238	37,7	16,13	13,665	7,573	280	74,50%	100%
			B	58,263	45,2	26,03	35,653	22,610	305	85,70%	100%

Strain A: CMC 1793; strain B: CMC 1818.

size, and includes a large number of entries similar to the strain under identification, all indices and the *Isyn* will show relatively low values. The M1 mapping does not yield a definitive identification at the species level, but rather gives an indication of the most likely species and the ex-type strain that should be used in the next M2 mapping as reference. M2 produces the same parameters as M1 and normally sets aside a relatively number of "unused reads" that usually range from 5 % to over about 30 %. Some of these reads were highly homologous (i.e. > 98 % homology) to the rDNA of other species. These considerations led us to propose a third mapping (M3) similar to M1 in which all the unused reads are mapped against the same selected library. To ease the reading of the output parameters the indices were calculated on the outputs of M2 and M3 jointly (Table 4B). The major difference between the M1 and the M2-M3 mappings is that the conserved sequences are attributed to the most likely species in the latter, whereas they are distributed randomly and evenly in the former. The whole M2-M3 procedure led to homology values comparable to those usually observed with Sanger sequencing.

These findings suggest a pipeline, in which the preliminary attribution to species is carried out with M1. An alternative to the use of M1 is the case in which the microbiologist has a some evidence to restrict the identification to one or few presumptive specis. However, if even after M2-M3 mapping a high level of homology with the ex-type strain of a known species is not reached, the identification should be questioned and the possibility that the strain represents a new species should be considered. Residual unused reads after M2-M3 mapping, can be considered either as background noise or requiring further investigation.

Validation of the procedure with a large group of strains

Our procedure was tested with 286 strains, isolated as opportunistic pathogenic yeasts from patients at two Italian hospitals. Both BTL and BBm algorithms were used. Since the yeasts were supposedly part of the known pathogenic yeasts attributed to *Candida*, a restricted reference library of the 16 ex-type strains of species accepted in this genus was employed. The analyses produced output values with the members of the library that led to the calculation of the indices, including *Isyn* (Table 4). For simplicity, the identification characterized by the highest *Isyn* will be hereinafter referred to as a "correct identification", and the others as "incorrect identifications".

It must be noted that the *Isyn* value gives an overview of the homology of the reads in a FASTQ file with the members of the library. Even a high *Isyn* value does not preclude that some of the reads show a high homology with an ex-type strain not representing the species, which the unknown strain belongs to. This is, for instance, the case of the *C. glabrata* CMC 1912 strain, which included some 2.86 % of the reads with over 98 % of homology with *C. albicans*.

Results of the BTL mapping showed that the majority of the "correct identifications" ranged from 80–95 % and no strain showed an *Isyn* higher that 95 % in the M1 mapping. The majority of the M2-M3 mapping results were in the range 95–100 %, with only 10 % of the strains showing less than 95 % *Isyn* (Fig. 4A). Similar results were obtained with the BBm algorithm, although some 4 % of the strains had an *Isyn* higher than 95 % and only around 5 % of the strains had less than 95 % *Isyn* with M2-M3 mapping (Fig. 4B).

The distribution of the "incorrect identifications" was studied with both algorithms, showing a decrease of their maximum frequency from *Isyn* = 20 % to *Isyn* = 5 % respectively, with the M1 and M2-M3 mappings (Fig. 4C–D).

One of the major differences between the M1 and the M2-M3 mapping results is that in the former most of the reads display a very high homology (> 95 %) with the sequences contained in the library, whereas in the latter this frequency decreases to around 5 % (Table 4A–B). This was observed in all the analysed strains, and indicates that M1 alone can produce highly biased results, especially if

Table 4. Example of a M1 and M2-M3 mapping against an *ad hoc* library of 16 ITS-LSU concatenate rDNA sequences of type strains of pathogenic yeasts (CMC 1912 strain).

A

Mapping M1 Algorithm: BTL	**Mapping parameters**				**Indexes**					
	# Nucleotides	# Sequences	% Of Ref Seq	% Pairwise Identity	Mean Cover.	*Iread*	*Inuc*	*Icov*	*Iref*	*Isim*	***Isyn***
C. albicans	212,814	1,426	99,90%	99,00%	203,48	6,14%	6,14%	9,26%	6,14%	6,08%	6,80%
C. dubliniensis	12,362	80	53,20%	99,30%	10,01	0,34%	0,36%	0,46%	0,19%	0,35%	0,34%
C. famata	1,863	6	24,40%	98,80%	0,56	0,03%	0,05%	0,03%	0,01%	0,05%	0,03%
C. glabrata	3,004,133	20,223	100,00%	99,00%	1813,15	87,06%	86,74%	82,51%	86,74%	85,87%	**86,41%**
I. orientalis	2,689	12	42,20%	96,60%	1,3	0,05%	0,08%	0,06%	0,03%	0,07%	0,06%
C. metapsilosis	3,886	20	43,10%	99,00%	1,93	0,09%	0,11%	0,09%	0,05%	0,11%	0,09%
C. orthopsilosis	8,590	52	92,30%	99,10%	6,34	0,22%	0,25%	0,29%	0,23%	0,25%	0,25%
C. parapsilosis	70,672	468	99,10%	96,90%	61,27	2,01%	2,04%	2,79%	2,02%	1,98%	2,18%
C. pararugosa	0	0	0	0	0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
D. rugosa	1,034	2	4,50%	92,90%	0,05	0,01%	0,03%	0,00%	0,00%	0,03%	0,01%
C. sake	105,986	703	35,00%	99,00%	69,76	3,03%	3,06%	3,17%	1,07%	3,03%	2,69%
C. tropicalis	23,711	154	96,40%	99,20%	20,51	0,66%	0,68%	0,93%	0,66%	0,68%	0,73%
C. utilis	1,705	5	14,70%	99,70%	0,53	0,02%	0,05%	0,02%	0,01%	0,05%	0,03%
C. lusitaniae	970	2	12,70%	100,00%	0,13	0,01%	0,03%	0,01%	0,00%	0,03%	0,01%
M. guilliermondii	8,425	52	97,70%	98,90%	6,21	0,22%	0,24%	0,28%	0,24%	0,24%	0,25%
S. cerevisiae	4,642	25	19,00%	99,20%	2,32	0,11%	0,13%	0,11%	0,03%	0,13%	0,10%

B

Mapping M2-M3 Algorithm: BTL	**Mapping parameters**				**Indexes**					
	# Nucleotides	# Sequences	% Of Ref Seq	% Pairwise Identity	Mean Cover.	*Iread*	*Inuc*	*Icov*	*Iref*	*Isim*	***Isyn***
C. albicans	91,888	606	96,40%	98,20%	89	2,86%	2,90%	4,62%	2,33%	2,88%	3,13%
C. dubliniensis	0	0	0,00%	0,00%	0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
C. famata	0	0	0,00%	0,00%	0	0,01%	0,04%	0,01%	0,00%	0,03%	0,02%
C. glabrata	3,231,287	21,757	100,00%	87,70%	2413,9	96,13%	95,85%	94,02%	95,85%	94,32%	**95,74%**
I. orientalis	1,221	2	13,60%	77,10%	0,1	0,02%	0,04%	0,01%	0,01%	0,04%	0,02%
C. metapsilosis	0	0	0,00%	0,00%	0	0,02%	0,05%	0,01%	0,00%	0,05%	0,02%
C. orthopsilosis	2,355	9	44,50%	97,80%	1,1	0,06%	0,09%	0,08%	0,06%	0,09%	0,08%
C. parapsilosis	18,410	116	87,00%	99,00%	15,9	0,48%	0,51%	0,70%	0,39%	0,51%	0,52%
C. pararugosa	0	0	0,00%	0,00%	0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
D. rugosa	0	0	0,00%	0,00%	0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%

Table 4. (Continued).

B

Mapping M2-M3 Algorithm: BTL	Mapping		parameters		Indexes					
	# Nucleotides	# Sequences	% Of Ref Seq	% Pairwise Identity	Mean Cover.	*lread*	*lnuc*	*lcov*	*lref*	*lsim*	***lsyn***
C. sake	0	0	0,00%	0,00%	0	0,09%	0,12%	0,08%	0,01%	0,12%	0,09%
C. tropicalis	6,670	38	55,10%	99,40%	5,1	0,22%	0,25%	0,34%	0,18%	0,25%	0,25%
C. utilis	0	0	0,00%	0,00%	0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
C. lusitaniae	0	0	0,00%	0,00%	0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
M. guilliermondii	4,147	21	47,70%	98,80%	2,7	0,10%	0,13%	0,14%	0,09%	0,13%	0,12%
S. cerevisiae	0	0	0,00%	0,00%	0	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%

Legend. (**a**) Mapping of a FASTQ file against a selected library of 16 type strains of pathogenic yeasts; (**b**) Mapping of the FASTQ file against the type strains of the presumptive species and the resulting mapping of the residual unused reads.

the homology (e.g. pairwise identity) is used directly without any other correction. Altogether, it seems that the *lsyn* from the M2-M3 mapping produces data reliable and quite comparable to the Sanger homology levels commonly used by taxonomists for strain identification.

The reads with high homology (> 97 %) to one of the library's ex-type strains were 23.96 % with BTL Local and 18.31 % with BBm. These relatively high frequencies could be due to the similarity among the members of the library, because similar references would share most of the common conserved regions. In order to test this hypothesis, we analysed the behaviour of the strains belonging to the sister species *C. parapsilosis* and *C. orthopsilosis.* These strains had 23.97 % and 20.64 % reads with < 97 % of homology, using BTL and BBm, respectively; this indicates that the "incorrect identifications" would be increased by libraries containing highly related ex-type strains, when using the M1 mapping.

Different estimates of the internal variability among rDNA copies

Our results show that M1 mapping of NGS reads detects the variability intrinsic among the rDNA copies, generating possible misclassifications. Analysis of the data from all M1 mappings showed that 28.6 % and 24.1 % average internal heterogeneity was detected by BTL and BBm (Fig. 5). This heterogeneity ranged from 14.7–45.7 % with BTL, and from 24.1– 42.8 % with BBm. These data indicate that the M1 mapping can only indicate which is the most probable species of the unknown isolate, but cannot determine its precise similarity to the ex-type strain. The subsequent M2 mapping allows the determination of this similarity at a good level, and produces a marked reduction in heterogeneity, averaging 3.4 % and 2.9 % for BTL and BBm, respectively. However, the application of M2 still detects up to 6.1 % (BTL) and 7 % heterogeneity (BBm).

These analyses showed that the sole application of M1 mapping produced a high rate of "incorrect identifications", generating over 20 % of false positive identifications, that can be effectively corrected by M2-M3 mapping.

DISCUSSION

The increasing expansion of NGS and the unparalleled wealth of output reads occurring with rapidly decreasing prices, necessitated a consideration of this technology in the identification of fungal strains. There are currently two different scenarios in which identification to species level is required, the classical identification of an isolate, and the massive identification of the members of a population by direct sequencing. The latter approach, metagenomics, is technically bound to NGS, whereas the identification of an isolate can be carried out with both Sanger sequencing and NGS. The cost and the throughput possibilities of NGS are already more competitive than the traditional Sanger sequencing in many analytical settings, bringing the possibility of employing NGS in both scenarios. At this time, the molecular identification of yeasts and filamentous fungi is carried out with markers included in the chromosomal region encoding the ribosomal RNA (Schoch *et al.* 2012). The advantage of these markers is the ease of manipulation due to the amount of DNA present

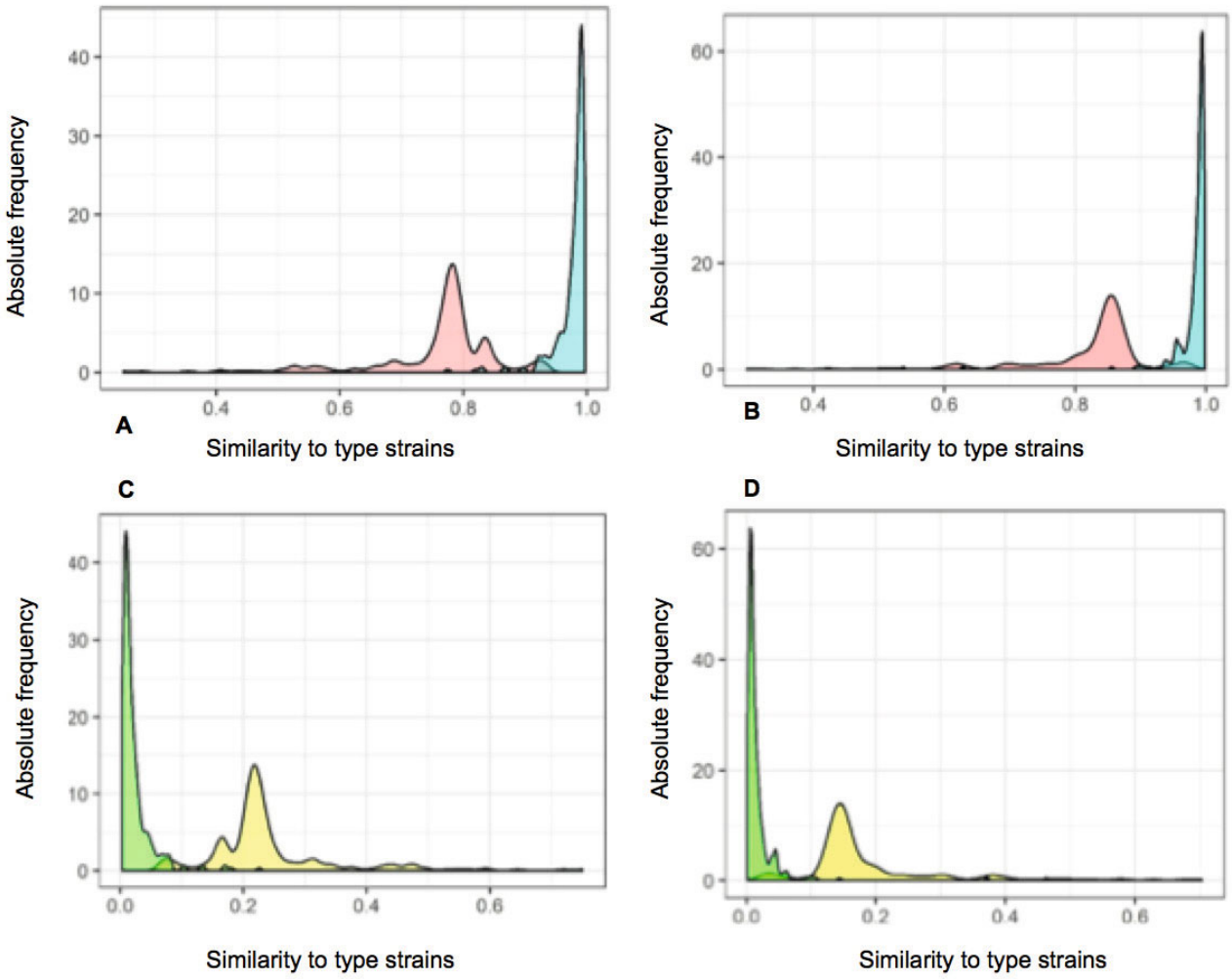


Fig. 4. Distribution of the similarity to ex-type strains with different analytical combinations. **A.** BTL Local - similarity to the correct species. **B.** BBm - similarity to the correct species. **C.** BTL Local - similarity to the incorrect species. **D.** BBm - similarity to the incorrect species. Light Red = mapping M1; Light Blue = mappings M2 and M3; Yellow = mapping M1; Green = Mappings M2 and M3.

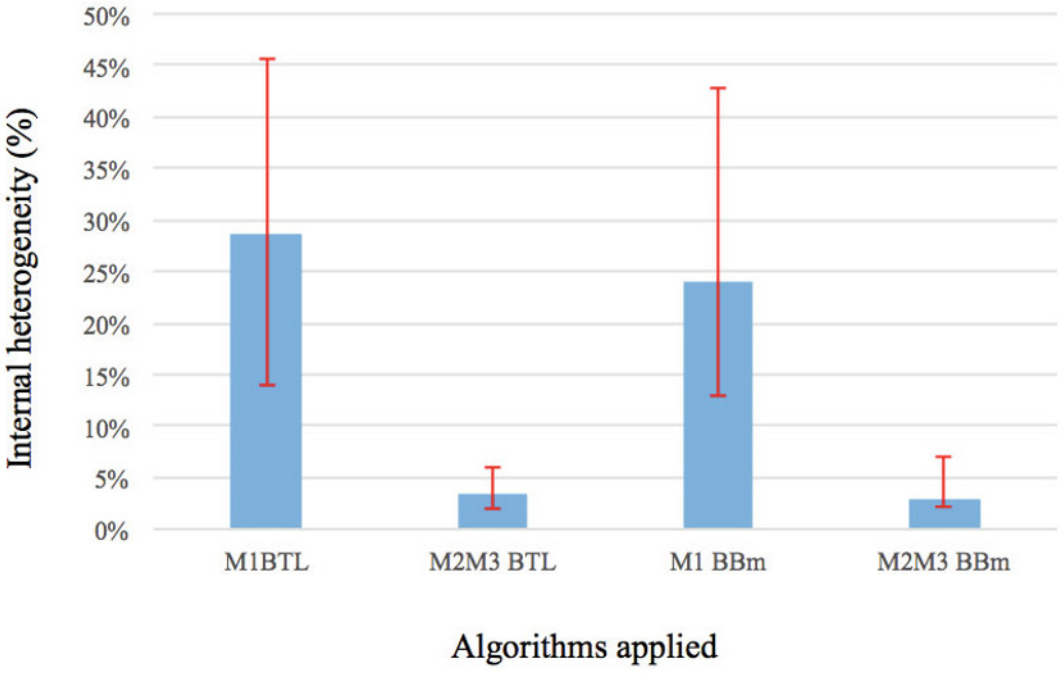


Fig. 5. Estimation of the internal variability among the rDNA copies.

in the cell, since this region represents a large fraction of the genome (Dujon *et al.* 2004). Not less important is that the tandem structure, with the alternation of conserved and variable regions, facilitates location of the anchoring places for primers and that both the LSU and the ITS regions are within the reach of a normal Sanger run. On the other hand, these markers belong to a large multigene family in which the presence of so many copies that are not necessarily identical poses some problems when using NGS both for strain identification and metagenomics. The dual aims of this paper were to find a rapid and reliable pipeline to employ NGS in strain identification, and to estimate the effects of internal heterogeneity among copies on the quality of the identification in a metagenomic scenario.

For single strain identification, the proposed pipeline was shown to have the rapidity to massively identify or re-identify thousands of strains in matter of days, with a very high level of accuracy arising from the high number of reads. This makes the final sequence several hundred times deeper than the Sanger ones that have a maximum depth of two in most routine applications. Of course, the high depth combined with the heterogeneity required the three mappings described here, which are not in contrast with the rapidity of the data processing. Considering the number of reads deriving from a single NGS analysis, and the relatively short length of the amplicons, this approach has the potential for multiplexing. The technique allowed the sequencing of two different marker loci, the LSU gene and ITS region, that represent a low level of multiplexing, although the depth obtained indicated that there is room for more markers (e.g. *TEF1-alpha, RPB1, RPB*2) (Kurtzman & Robnett 2013), especially if not represented by multi-copy genes, as was the case here. The use of Multi Locus Sequencing has been recommended for the superior ability to detect the species circumspection and is particularly important when the *taxa* are phylogenetically close and therefore require a high level of taxonomic resolution. Still, the markers proposed (Kurtzman & Robnett 2013) present some practical problems, such as the difficulty of finding universal anchoring positions for both the primary amplification and the sequencing, making their use not as straightforward as one would expect in routine procedures for identification or, even more, diagnostics.

Our initial hypothesis was that, in the case of multigene families, the sequencing depth obtainable with amplicon-based NGS could produce information on the real extent of the internal heterogeneity, possibly masked by the peculiarities of Sanger sequencing. This aspect is particularly relevant to predict the behaviour of heterogeneous multigene families, as the rRNA encoding genes, when subject to NGS for the direct exploration of environmental biodiversity, taking a metagenomic culture-independent approach, to explore the effects of the rDNA heterogeneity in metagenomic identifications. For this reason, we carried out identifications at the strain level with a technique amenable of working in metagenomics with DNA derived from an unknown number of taxa. The results indicated that the relatively high heterogeneity within a single genome must be taken into account in case NGS is used to sequence multigene markers for identification, both at the strain and metagenomics level. In the first case, the simple usage of the procedure described allows an accurate determination of the similarity of the strain with the ex-type of the species it belongs to. Moreover, the application of M2-M3 mappings yields similarity values comparable with those produced by Sanger sequencing, allowing to use the same threshold values proposed in studies carried out using Sanger sequencing data (Vu *et al.* 2016).

In general, our work point out that some sort of pipeline is necessary to ensure a “correct” species identification and that mapping is much more efficient than assembly. These observations are in contrast with those proposed by Ahmed (2016) using data from a different platform and experimental design, and indicate that some sort of pipeline is necessary to decrease the extent of misidentification in metagenomics settings.

The heterogeneity with M1 mapping was an average of 21.13 % throughout all the species considered. When M2-M3 mappings were sequentially applied, the number of reads not homologous to the species of appurtenance dropped to around 2.1 %. The latter estimate of internal heterogeneity is quite close to the *ca*. 3 % expected (Lindner *et al.* 2013). These figures mean that the actual heterogeneity is indeed around 3 %, which seems to support more the classical theory of gene conversion (Ganley & Kobayashi 2007). The problem in metagenomics is that the variability would be estimated at around 20 % because the M2-M3 mappings can only be applied to the scenario of isolated strains, and not in metagenomics. This is due to the number of species present in a metagenomics sample being unknown, whereas a single isolate can belong only to one species, allowing the redistribution of the various reads not attributed directly to the right taxon. A way around this problem would be to extend this type of study to other taxa in order to: (1) verify the extent of the internal variability of the rDNA region; and (2) define an algorithm that disregards all taxa below a given threshold to eliminate artefacts, i.e. reads that at the single strain level would not be attributed to the “correct” species.

Further works, similar to that presented here, will be needed to determine whether these thresholds are similar in different groups of fungi, and would in also allow for a correction of the alpha-diversity estimate that could be produced by an unsupervised usage of NGS, i.e. by automatic pipelines without the actual supervision of a microbiologist.

Taking these considerations together, we conclude that the possibilities offered by current NGS techniques, and their future developments, promise to shed more light on rDNA composition and to transform its internal variability into a powerful taxonomic tool.

ACKNOWLEDGEMENTS

CC was supported by a PhD program of the Biotechnology Doctorate at the University of Perugia paid by MIUR (Ministero dell’Istruzione, dell’Università e della Ricerca). LR was supported as temporary researcher with a grant from Fondazione Cassa di Risparmio di Perugia. Partially supported by the GILEAD (Gilead Sciences, Inc.,) Fellowship Programme: “Caratterizzazione del biofilm ed efficacia del trattamento con Amfotericina B liposomiale in ceppi filmogeni del

gruppo *Candida parapsilosis sensu lato*". WM and VR are supported by an Australian NH&MRC (National Health and Medical Research Council) grant #APP1121936. The authors acknowledge gratefully Angela Conti for friendly technical help.

REFERENCES

Ahmed A (2016) Analysis of metagenomics Next Generation Sequence data for fungal ITS barcoding: do you need advance bioinformatics experience? *Frontiers in Microbiology* **7**: 1061.

Alper I, Frenette M, Labrie S (2011) Ribosomal DNA polymorphisms in the yeast Geotrichum candidum. *Fungal Biology* **115**: 1259–1269.

Amend AS, Seifert KA, Bruns TD (2010) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology* **19**: 5555–5565.

Berrocal CA, Rivera-Vicens RE, Nadathur GS (2016) Draft genome sequence of the heavy-metal-tolerant marine yeast *Debaryomyces hansenii* J6. *Genome announcements* **4**: e00983–00916.

Bokulich NA, Mills DA (2012) Next-generation approaches to the microbial ecology of food fermentations. *BMB Reports* **45**: 377–389.

Bushnell B (2014) *BBMap: a fast, accurate, splice-aware aligner.* Berkeley: Ernest Orlando Lawrence Berkeley National Laboratory.

Corte L, Roscini L, Colabella C, *et al.* (2016) Exploring ecological modelling to investigate factors governing the colonization success in nosocomial environment of *Candida albicans* and other pathogenic yeasts. *Scientific Reports* **6**: 26860.

Dujon B (1996) The yeast genome project: what did we learn? *Trends in Genetics* **12**: 263–270.

Dujon B, Sherman D, Fischer G, *et al.* (2004) Genome evolution in yeasts. *Nature* **430**: 35–44.

Ganley AR, Kobayashi T (2007) Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Research* **17**: 184–191.

Gong J, Dong J, Liu X (2013) Extremely high copy numbers and polymorphisms of the rDNA operon estimated from single cell analysis of oligotrich and peritrich ciliates. *Protist* **164**: 369–379.

Groenewald M, Robert V, Smith MT (2011) The value of the D1/D2 and internal transcribed spacers (ITS) domains for the identification of yeast species belonging to the genus *Yamadazyma*. *Persoonia* **26**: 40–46.

Hajibabaei M (2012) The golden age of DNA metasystematics. *Trends in Genetics* **28**: 535–537.

Imabayashi Y, Moriyama M, Takeshita T, *et al.* (2016) Molecular analysis of fungal populations in patients with oral candidiasis using next-generation sequencing. *Scientific Reports* **6**: 28110.

Irinyi L, Serena C, Garcia-Hermoso D, *et al.* (2015) International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database—the quality controlled standard tool for routine identification of human and animal pathogenic fungi. *Medical Mycology*: myv008.

Jones T, Federspiel NA, Chibana H, *et al.* (2004) The diploid genome sequence of *Candida albicans*. *Proceedings of the National Academy of Sciences, USA* **101**: 7329–7334.

Kearse M, Moir R, Wilson A, *et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

Korabecna M (2007) The variability in the fungal ribosomal DNA (ITS1, ITS2, and 5.8 S rRNA gene): its biological meaning and application in medical mycology. *Communicating Current Research and Educational Topics and Trends in Applied Microbiology* **105**: 783–787.

Kurtzman CP, Robnett CJ (1998) Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. *Antonie van Leeuwenhoek* **73**: 331–371.

Kurtzman CP, Robnett CJ (2013) Relationships among genera of the *Saccharomycotina* (*Ascomycota*) from multigene phylogenetic analysis of type species. *FEMS Yeast Research* **13**: 23–33.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.

Li W, Sun H, Deng Y, Zhang A, Chen H (2014). The heterogeneity of the rDNA-ITS sequence and its phylogeny in *Rhizoctonia cerealis*, the cause of sharp eyespot in wheat. *Current Genetics* **60**: 1–9.

Liao D (1999) Concerted evolution: molecular mechanism and biological implications. *American Journal of Human Genetics* **64**: 24–30.

Lindner DL, Banik MT (2011) Intragenomic variation in the ITS rDNA region obscures phylogenetic relationships and inflates estimates of operational taxonomic units in genus *Laetiporus*. *Mycologia* **103**: 731–740.

Lindner DL, Carlsen T, Henrik Nilsson R, *et al.* (2013) Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi. *Ecology and Evolution* **3**: 1751–1764.

Lubock NB, Zhang D, Sidore AM, Church GM, Kosuri S (2017) A systematic comparison of error correction enzymes by next-generation sequencing. *Nucleic Acids Research* **45**: 9206–9217.

Maleszka R, Clark-Walker G (1993) Yeasts have a four-fold variation in ribosomal DNA copy number. *Yeast* **9**: 53–58.

Medinger R, Nolte V, Pandey RV, *et al.* (2010) Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Molecular Ecology* **19**: 32–40.

Naidoo K, Steenkamp ET, Coetzee MP, Wingfield MJ, Wingfield BD (2013) Concerted evolution in the ribosomal RNA cistron. *PLoS One* **8**: e59355.

Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics* **39**: 121.

Schoch CL, Robbertse B, Robert V, *et al.* (2014) Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. *Database- the Journal of Biological Database and Curation* **2014**.

Schoch CL, Seifert KA, Huhndorf S, *et al.* (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl Acad Sci U S A* **109**: 6241–6246.

Simon UK, Weiß M (2008) Intragenomic variation of fungal ribosomal genes is higher than previously thought. *Molecular Biology and Evolution* **25**: 2251–2254.

Stielow J, Lévesque C, Seifert K, *et al.* (2015) One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes. *Persoonia-Molecular Phylogeny and Evolution of Fungi* **35**: 242.

Susca A, Perrone G, Cozzi G, *et al.* (2013) Multilocus sequence analysis of Aspergillus Sect. Nigri in dried vine fruits of worldwide origin. *International Journal of Food Microbiology* **165**: 163–168.

Vu D, Groenewald M, Szöke S, *et al.* (2016) DNA barcoding analysis of more than 9000 yeast isolates contributes to quantitative thresholds for yeast species and genera delimitation. *Studies in Mycology* **85**: 91–105.

Vydryakova GA, Van DT, Shoukouhi P, Psurtseva NV, Bissett J (2012) Intergenomic and intragenomic ITS sequence heterogeneity in *Neonothopanus nambi* (Agaricales) from Vietnam. *Mycology* **3**: 89–99.

Wang W, Ma L, Becher H, *et al.* (2015) Astonishing 35S rDNA diversity in the gymnosperm species *Cycas revoluta* Thunb. *Chromosoma*: 1–17.

West C, James SA, Davey RP, Dicks J, Roberts IN (2014) Ribosomal DNA sequence heterogeneity reflects intraspecies phylogenies and predicts genome structure in two contrasting yeast species. *Systematic Biology* **63**: 543–554.

Woo PC, Leung S-Y, To KK, *et al.* (2010) Internal transcribed spacer region sequence heterogeneity in *Rhizopus microsporus*: implications for molecular diagnosis in clinical microbiology laboratories. *Journal of Clinical Microbiology* **48**: 208–214.

Yurkov A, Guerreiro MA, Sharma L, Carvalho C, Fonseca A (2015) Multigene assessment of the species boundaries and sexual status of the basidiomycetous yeasts *Cryptococcus flavescens* and *C. terrestris* (Tremellales). *PloS One* **10**: e0120400.