NGS-based rDNA barcoding in fungal species identification and delimitation: limits, opportunities and relation to phenotypic HT FT-IR spectroscopy

NGS-baserte rDNA barkoding i identifisering og avgrensning av gjærarter: rammer, muligheter og sammenligning med fenotypisk HT FT-IR spektroskopi

Philosophiae Doctor (PhD) Thesis

Claudia Colabella

University of Perugia - Department of Pharmaceutical Sciences
Norwegian University of Life Sciences - Faculty of Science and Technology

Perugia - Ås 2017

Thesis number: 2017:28
ISSN: 1894-6402
ISBN: 978-82-575-1431-0

TABLE OF CONTENTS

ABSTRACT 1
NORSK SAMMENDRAG 4
LIST OF PAPERS 7
1. AIM OF THE THESIS 8
2. INTRODUCTION 9
2.1 SPECIES CONCEPT AND DELIMITATION IN FUNGI 9
2.2 DNA BARCODING 10
2.2.1 Molecular markers in fungi 12
2.2.2 rDNA ribosomal genes 13
2.2.3 D1/D2 domains of the LSU (26S) rDNA genes 14
2.2.4 Internal Transcribed Spacer (ITS) as universal barcode for fungi 15
2.2.4.1 Heterogeneity and limits of ITS region 16
2.2.5 Molecular evolution of the tandem repeats rDNA genes 17
2.3 PHENOTIPYC APPROACH AS A POTENTIAL TOOL FOR FUNGAL IDENTIFICATION 18
2.3.1 MALDI-TOF 18
2.3.2 Fourier Transform Infrared spectroscopy (FT-IR) 19
3. METHODOLOGIES 22
3.1 PCR-BASED METHODS FOR YEAST IDENTIFICATION 22
3.1.1 Amplification of the rDNA genes 22
3.2 EARLY DNA SEQUENCING 23
3.3 DNA SEQUENCING - THE NEXT GENERATION 24
3.4 DATABASE AND BIOINFORMATIC TOOLS 27
3.5 RAPID IDENTIFICATION OF FUNGAL RIBOSOMAL PROTEINS 29
3.6 FT-IR SPECTROSCOPY - A HIGH-THROUGHPUT PHENOTYPIC APPROACH 31
3.6.1 Absorption of infrared light 31
3.6.2. The Fourier Transform Infrared (FT-IR) spectra of microorganisms 33
3.6.3. Pre-processing of FT-IR spectra 34
3.6.4. Multivariate data analysis 35
4. RESULTS AND DISCUSSION 42
4.1 IDENTIFICATION OF PATHOGENIC BIOFILM-FORMING STRAIN USING ITS BARCODE 42
4.2 LIMIT OF ITS BARCODE IN THE DIAGNOSE OF FILAMENTOUS FUNGI 42
4.3 DELIMITATION OF YEASTS FOOD/CLINIC RELATED STRAINS USING PHENOTYPIC AND MOLECULAR APPROACHES 43
4.4 EXPLOITATION OF THE INTERNAL VARIABILITY OF THE rDNA OPERON: NGS-LIKE APPROACH 43
4.5 BRINGING THE ITS BARCODE IN THE NGS ERA 44
4.6 HT-NGS TECHNOLOGY AS A POTENTIAL TOOL FOR SNPs DETECTION 45
4.7 IDENTIFICATION OF PATHOGENIC YEASTS USING NGS BARCODING AND FT-IR JOINT-POSSIBILITIES 45
5. CONCLUSIONS AND FUTURE PROSPECTS 47
REFERENCES 48

ABSTRACT

The abundance of ribosomal DNA (rDNA) in the yeast and fungal genomes derives from their multigene nature. During the last decade of the XX century, this DNA region has become very popular for the molecular characterization of fungi. Unfortunately, the multigene nature of rDNA cannot be completely identified by the Sanger sequencing that records only the most prevalent nucleotide at each position. Conversely, Next Generation Sequencing (NGS) has unveiled the internal heterogeneity of rDNA, due to its mechanism of reporting individual reads. For these reasons, rDNA sequencing and particularly the Internal Transcribed Spacer (ITS) marker, have huge advantages in taxonomy, barcoding, ecological microbiology and diagnostics.

The aim of this thesis was to achieve a closer understanding of the rDNA organization and to link molecular and phenotypical analysis in order to obtain a stable and meaningful phenetic taxonomy, which accounts for the phylogeny.

The first part of the introduction of this thesis is a critical review of the literature on rDNA and its taxonomic variability. In the second part, the thesis illustrates how the application of new strategies to detect the variability within the rDNA, allows the identification and classification of species by analysing species derived from different environments that are relevant for white, green and red biotechnologies. Limitations of the significance of markers in the application of DNA-based molecular taxonomy of microorganisms are discussed. Therefore, to avoid a sterile taxonomic approach leading to a pure nomenclatural exercise, phenotypic characterization was associated to the genotyping of selected microorganisms. For this reason, as example, results obtained in studies on the ability of selected microorganisms to form biofilm in addition to their metabolomic characterization are presented. The biofilm forming ability of more than two hundred pathogenic strains belonging to *Candida* genus identified using ITS marker are presented. The relation between different variables was tested and results showed that species and biofilm forming ability appeared to be distributed almost randomly whereas the relation between biofilm formation and species isolation frequency was highly significant (R^2 around 0.98).

The identification of saprophytic filamentous fungi, which cause invasive infections, is also presented. In this case the current molecular diagnostic tools, based on the barcode marker ITS, failed in discriminating this fungi between the complex *Trichoderma*

longibrachiatum/*Hypocrea orientalis,* even using different tools. The definitive identification was carried out combining molecular approach and microbiological test.

A combined approach in the delimitation of ninety-six food-related strains of the complex *Meyerozyma*/*Candida guilliermondii* is presented. Results of both approaches (ITS and FT-IR spectroscopy) showed that the possibility to discriminate among strains with molecular and metabolomic analyses represents an additional tool to empower food and industrial monitoring and to gain further knowledge on the genetic variations of this species.

In order to study the variability of the rDNA an NGS-like approach on a new species *Ogataea uvarum* sp.nov. was carried out. Results showed that the ITS marker was more variable than the LSU gene, especially in the ITS2 region. In order to test the origin of this heterogeneity the whole region was introduced in a mini library and several clones were sequenced separately. The cloning of a sample of single copy sequences showed that indeed an internal heterogeneity is present and that the process of generating a consensus using Sanger sequencing hides a large part of it.

For instance, the introduction of NGS leads to a deeper knowledge of the individual sequences and of the variants between the same DNA sequences located in different tandem repeats. With this purpose, more than two hundred strains belonging to *Candida* genus were sequenced with NGS and a pipeline for the identification using different bioinformatics tools was carried out. The NGS also offers the possibility to evaluate this heterogeneity by analysing the Single Nucleotide Polymorphisms (SNPs) within the reads of an rDNA region amplified from a single strain DNA. Results performed on the four prevalent *Candida* species *(C. albicans, C. glabrata*, *C. parapsilosis* and *C. tropicalis*) indicated the presence of high variability among the strains and between the species, especially in the ITS2 region.

Moreover, a combined approach on these four *Candida* species using NGS and FT-IR spectroscopy was applied in order to improve the identification of pathogenic strains. Multivariate data analysis (MVA) by Consensus Principal Component Analysis (CPCA) was carried out. Partial Least Squares Regression (PLSR) was applied to build a classification model based on most relevant IR variables. The model was then cross-validated with the a success rate of 94.2%. Identification was also performed considering both the distance to the type strain and the central strain resulting in 97.4% correct classification.

In conclusion, in this thesis an identification method for the diagnose of pathogenic yeasts was developed on the basis of NGS. The internal variability of the rDNA was exploited and the relative limitations of the current methodologies presented. The comparison of results from totally different characters (molecular vs. phenotypic) and expressed with different data types (categorical vs. continuous) is one of the challenges necessary to try a reconciliation between the molecular DNA-based taxonomy, characterized by stable and "potential" characters, and the phenotypic data describing "actual" traits of the cells.

NORSK SAMMENDRAG

Overflod av ribosomalt DNA (rDNA) i gjær og sopp genomer stammer fra deres multigen natur. I løpet av det siste tiåret av XX århundre, har denne DNA regionen blitt svært populært for molekylær karakterisering av sopp. Dessverre kan den multigene naturen av rDNA ikke fullstendig identifiseres ved Sanger-sekvensering, som registrerer bare de mest utbredte nukleotider ved hver posisjon. I motsetning har Next Generation Sequencing (NGS) avduket den interne heterogeniteten av rDNA, på grunn av sin mekanisme for rapportering av enkelte 'reads'. Derfor har rDNA sekvensering og spesielt Internal transkribert Spacer (ITS) markører store fordeler i taksonomi, barcoding, økologisk mikrobiologi og diagnose.

Målet med denne avhandlingen var å oppnå en bedre forståelse av rDNA organiseringen og å lage en forbindelse mellom molekylær og fenotypisk analyse for å oppnå en stabil og meningsfull fenetisk taksonomi, som uttrykker fylogenien.

Den første delen av innledningen av denne avhandlingen er en kritisk gjennomgang av litteraturen om rDNA og dens taksonomisk variabilitet. I den andre delen, viser avhandlingen hvordan anvendelsen av nye strategier for å oppdage variasjonen innenfor rDNA, tillater identifisering og klassifisering av arter ved å analysere arter som stammer fra ulike miljøer som er relevante for hvite, grønne og røde bioteknologi. Begrensninger i betydningen av markører i anvendelsen av DNA-baserte molekylære taksonomi av mikroorganismer diskuteres.

Derfor, for å unngå en steril taksonomisk tilnærming som fører til en ren taksonomi øvelse, ble fenotypisk karakterisering knyttet til genotypingen av utvalgte mikroorganismer. Derfor presenteres, for eksempel, det resultater som er oppnådd i biofilmstudier, hvor evnen av utvalgte mikroorganismer for dannelse av biofilm i tillegg til deres metabolomisk karakterisering undersøkes.

Evnen til å danne biofilm ble presentert for mer enn to hundre patogene stammer tilhørende slekten *Candida* og som er identifisert ved hjelp av markeringen ITS. Forholdet mellom evnen til å danne biofilm og artene ble undersøkt. Resultatene viste at det ikke er noe korrelasjon mellom arten og biofilmformingsevne, mens korrelasjonen mellom biofilmdannelse og isolasjonsfrekvensen for arten.

Identifiseringen av saprophytic trådformede sopp, som forårsaker invasive infeksjoner, blir også presentert. I dette tilfellet, mislykkes dagens molekylære diagnostiske verktøy basert på strekkode markør ITS i å diskriminere denne sopparten i komplekset

Trichoderma longibrachiatum/*Hypocrea orientalis*, selv ved hjelp av ulike verktøy. Den endelige identifikasjonen ble utført ved å kombinere molekylær tilnærming og mikrobiologiske test.

En kombinert tilnærming i avgrensningen av nitti-seks matrelaterte stammer av komplekset *Meyerozyma*/*Candida guilliermondii* er presentert. Resultater av begge tilnærminger (ITS og FT-IR spektroskopi) viste at muligheten til å diskriminere mellom stammer med molekylære og metabolomiske analyser representerer et tilleggsverktøy som kan styrke mikrobiell kontroll i matindustri og for å få mer kunnskap om de genetiske varianter av denne arten.

For å studere variasjonen av rDNA ble en NGS-lignende metode testet for en ny art *Ogataea uvarum sp.nov*. Resultatene viste at ITS markøren var mer variabel enn LSU genet, spesielt i ITS2 regionen. For å teste opprinnelsen av denne heterogeniteten, ble hele regionen innført i en mini-bibliotek og flere kloner ble sekvensert separat. Kloning av et utvalg på enkelkopi sekvenser viste at faktisk en intern heterogenitet er til stede, og at prosessen med å generere en konsensus ved hjelp av Sanger-sekvensering skjuler en stor del av denne heterogeniteten.

Innføringen av NGS fører til en dypere forståelse av de individuelle sekvensene og av variantene mellom de samme DNA-sekvensene som ligger i forskjellige tandemrepetisjoner. Med dette formålet, ble mer enn to hundre stammer tilhørende *Candida* slekten sekvensert med NGS og en rutine for identifisering ved hjelp av ulike bioinformatiske analyser ble satt opp. NGS tilbyr også muligheten for å evaluere heterogeniteten ved å analysere enkelt-nukleotider (SNPs) i lesninger av en rDNA region amplifisert fra DNAen til en enkelt stamme. Resultatene utført på de fire viktigste *Candida*-arter (*C. albicans*, *C. glabrata*, *C. parapsilosis* og *C. tropicalis*) indikerte tilstedeværelse av høy variabilitet blant stammene og mellom artene, spesielt i ITS2 regionen.

Videre ble en studie gjennomført, hvor en kombinasjon av NGS og FT-IR-spektroskopi ble utført for de fire *Candida*-artene for å forbedre den identifikasjon av patogene stammer. Multivariat dataanalyse (MVA) ved Konsensus Principal Component Analyse (CPCA) ble utført. Partial Least Squares Regression (PLSR) ble brukt til å bygge en klassifiseringsmodell basert på de mest relevante IR variablene. Modellen ble deretter kryss-validert med en suksessrate på 94,2%. Identifikasjon ble også utført med tanke på både avstanden til typestammen og den sentrale artsstammen og resulterte i 97,4% korrekt identifisering.

I denne avhandlingen ble en identifikasjonmetode for diagnose av patogene gjærsopper utviklet på basis av NGS. Den nye metoden utnytter den indre variasjon av rDNA. De relative begrensningene ved eksisterende metoder blir diskutert. Sammenligningen av identifikasjonsresultater som stammer fra data med helt forskjellige karakterer (molekyl vs. fenotypiske) og som er uttrykt med ulike datatyper (kategoriske vs. kontinuerlig) er nødvendig hvis man vil komme fram til en avstemming mellom en molekylær DNA-baserte taksonomi, preget av stabil og "potensielle" tegn, og en fenotypiske taksonomi som beskriver egenskapene til cellene.

LIST OF PAPERS

The thesis is based on the following papers:

I. Corte, L., Roscini, L., Colabella, C., Tascini, C., Leonildi, A., Sozio, E., ... & Cardinali, G. (2016). **Exploring ecological modelling to investigate factors governing the colonization success in nosocomial environment of *Candida albicans* and other pathogenic yeasts**. *Nature Publishing Group. Scientific Reports*, *6*, 26860.

II. Tascini, C., Cardinali, G., Colabella, C., Barletta, V., Di Paolo, A., Leonildi, A., Zucchelli, G., ... & Pasticci, M. B. (2016). **First Case of *Trichoderma longibrachiatum* CIED (Cardiac Implantable Electronic Device) - Associated Endocarditis in a Non-immunocompromised Host: Biofilm Removal and Diagnostic Problems in the Light of the Current Literature**. *Mycopathologia*, *181*(3-4), 297-303.

III. Corte, L., di Cagno, R., Groenewald, M., Roscini, L., Colabella, C., Gobbetti, M., & Cardinali, G. (2015). **Phenotypic and molecular diversity of *Meyerozyma guilliermondii* strains isolated from food and other environmental niches, hints for an incipient speciation**. *Food microbiology*, *48*, 206-215.

IV. Colabella, C., Roscini, L., Tristezza, M., Corte, L., Perrotta, C., Rampino, P., Cardinali, G., Grieco, F. **Travel Into the Internal Variability of Cloned rDNA Operon.** *In progress*.

V. Colabella, C., Corte, L., Roscini, L., Bassetti, M., Tascini, C., Mellor, J., Meyer, W., Cardinali, G. **Moving to NGS barcode sequencing for identification and diagnostics, an application in "*Candida*" pathogenic yeasts.** *Studies in Mycology*. (*Submitted*).

VI. Colabella, C., Corte, L., Roscini, L., Casagrande P, D., Bassetti, M., Tascini, C., Cardinali, G. **High Depth Next Generation Sequencing of single colony DNA reveals large variation levels of the Ribosomal DNA region ITS-LSU D1/D2 in the four prevalent pathogenic species of the genus *Candida.*** *In progress*.

VII. Colabella, C., Corte, L., Roscini, L., Kohler, A., Shapaval, V., Tafintseva, V., Cardinali, G. **Approaches and tools for species delimitation with FTIR and NGS in the four prevalent species of *Candida* pathogenic yeasts.** *PlosOne*. (*To be submitted in its current form*).

1. AIM OF THE THESIS

This thesis is aimed at exploring the rDNA organization in fungi and at demonstrating that the association of molecular with phenotypic analysis can lead to a more stable and phenetic taxonomy that takes into consideration also the evidences of the phylogeny.
The sub-goals were:

1. To study the limit of the DNA barcoding in the diagnose of species;
2. To describe yeast delimitation using both phenotypic and molecular approaches;
3. To develop an identification method for the diagnose of pathogenic yeasts using NGS;
4. To explore the internal variability of the rDNA using standard procedures and High-throughput Next Generation Sequencing technology;
5. To connect DNA barcoding and HT FT-IR spectroscopy.

2. INTRODUCTION

2.1 SPECIES CONCEPT AND DELIMITATION IN FUNGI

Species are one of the fundamental units of biology, comparable in importance to genes, cells and organisms [1, 2]. During the past half century, the issue of species delimitation has been confused by a problem involving the concept of species itself [3]. Among higher eukaryotes, it is possible to discriminate species according to biological discontinuities, such as the reproductive barrier at the basis of the Biological Species Concept [4]. This is impossible in lower eukaryotes since most of the Fungi are known to have solely an asexual cycle. Fungi displaying both asexual and sexual cycles can reproduce in both ways, with the consequence that the lack of a partner for the sexual reproduction is not a survival limitation. This implies that sexuality is an accessory mean of reproduction and cannot therefore be used as a general criterion of discontinuity and limitation in all fungal species [5]. The lack of effective barriers based on sexual reproduction in most fungal species, suggests that a continuous distribution of species could occur as probably happens in bacteria [6, 7] which can be described with some basic species concepts such as "*a species is a category that circumscribes (preferably) genomically coherent group of individual isolates/strains sharing a high degree of similarity in (many) independent features, comparatively tested under highly standardized conditions*" [8]. The fungal species is ruled by the Botanical and Mycological Code of Nomenclature, which defines different kind of "types". Basically, a type is defined as for the art 7.1 of the Vienna code: "*The application of names of taxa of the rank of family or below is determined by means of nomenclatural types (types of names of taxa)*" enforcing the Principle II: "*The application of names of taxonomic groups is determined by means of nomenclatural types*." The type is not necessarily the best representative of the *taxon* (*7.2 A nomenclatural type (typus) is that element to which the name of a taxon is permanently attached, whether as the correct name or as a synonym. The nomenclatural type is not necessarily the most typical or representative element of a taxon*.) [9]. The taxonomic practice has led, however to compare when possible, the unknown strain with the type strain, transforming the type in a sort of reference for the whole species. In the DNA sequencing era the comparison with the type strain of a presumptive species is a good practice, sometimes without any other alternative, because only the type strain marker sequences are available for many species.

2.2 DNA BARCODING

The identification of biological entities, such as microbial species, is essential for fundamental biological research such as the assessment of biodiversity, conservation, taxonomy and evolutionary biology and for those applications in which humanity and biodiversity intersect (agriculture, ecology, bioremediation and pathology) [3, 10]. DNA molecule, which stores the biological information in the variable sequences of four bases (A, C, G, T), is a key to reveal biodiversity. DNA barcoding relies on the assumption that the genetic variation between species exceeds that within species. Therefore, the distributions of intra- and inter-specific variability separated by a distance called "DNA barcoding gap" can be determined combining molecular analysis with bioinformatics technique [11, 12]. Long before the term "DNA barcoding" assumed its present meaning, genetic information in different forms has been used for at least half a century for systematics research; the invention of Sanger sequencing marked a crucial point in the use of genetic data in the field of systematic [13]. DNA barcoding was proved to be a powerful tool to understand the biodiversity of fungi, their ecological roles as well as the geographical distribution of pathogenic species, with enormous potential also to resolve the so-called "cryptic" species. The DNA barcoding is a global initiative designed to provide rapid, accurate, and automated species identification by using short, standardized gene regions as internal species markers [11]. The critical issue underlying barcoding is accuracy, defined in taxonomic terms as the capability of unbiased and unequivocal identification at the species level. Accuracy depends especially on the extent of, and the separation between, intraspecific variation and interspecific divergence within the selected marker creating a significant barcoding ''gap'' [14]. Threshold values separate intraspecific variation and interspecific differences. In particular, the threshold is useful to compare the unknown species in existing samples with species that has been assumed to represent the characteristic sample species. The accuracy of a threshold-based approach critically depends upon the level of overlap between intra- and inter-specific variations across a phylogeny (**Fig. 1**). Sequences unique to single species make identification easier, but their lack of universality hampers their amplification and therefore the whole procedure.

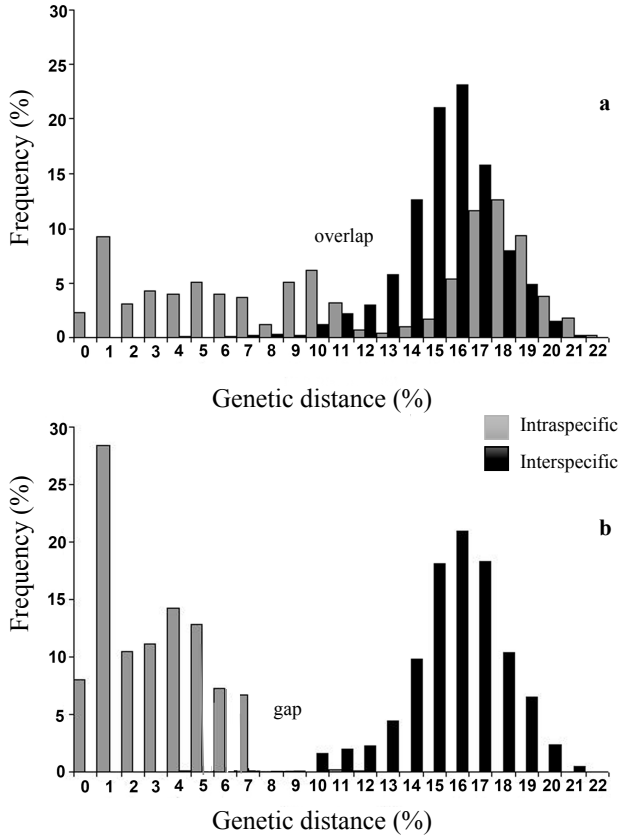


Fig. 1 Schematic distribution of intraspecific variation and interspecific divergence.
(**a**) Significant overlapping. (**b**) Ideal barcoding showing discrete distribution and no gap.

An ideal DNA barcode requires two fundamental characteristics: high taxonomic coverage and high resolution. Coverage, also mentioned as "universality", refers to the correct amplification of the genomic region chosen as DNA barcode in the broadest panel of *taxa*. On the other hand, a high resolution ensures the identification of different *taxa,* based on interspecific differences in DNA barcode sequences [12, 15]. A DNA barcode is not just any DNA sequences, it is a rigorously standardized sequence of a minimum length and quality from an agreed-upon gene, deposited in a major sequence database, and attached to a voucher specimen whose origins and current status are recorded. In this scenario, Arnot et al. [16] suggested the use of the hypervariable tandemly repeat DNA sequences as barcode to identify the strains of the parasite *Plasmodium*, while in 2002 Floyd et al. [17] proposed the use of nuclear small subunit ribosomal DNA (18S) sequences for defining Molecular Operational Taxonomic Units for the taxonomy of nematodes. In 2003 Hebert et al. [11] proposed that a DNA barcoding system for animal life could be based upon sequences diversity in cytochrome *c* oxidase

subunit 1 (*COX1*). They established that diversity in the amino acid sequences coded by the 5' section of this mitochondrial gene (mtDNA) was sufficient to reliably place species into higher taxonomic categories (from *phyla* to orders). However the use of the mtDNA in broad taxonomic analyses is constrained by the prevalence of insertions and deletions (indels) that greatly complicate sequence alignments [18].

2.2.1 Molecular markers in fungi

Many barcode markers have been described for fungi, such as *COX1* [19], nuclear large ribosomal subunit (LSU rDNA) [20], nuclear small ribosomal subunit (SSU rDNA) [21], β-tubulin (BenA) [22], partial translation elongation factor 1-α [23-25], protein-coding genes like RNA polymerase I and II [26-29] and internal transcribed spacer (ITS) [30-32]. Exploration of the animal barcode marker, cytochrome oxidase 1, has been fruitful for some fungi, but intron issues and lack of resolution in other *taxa* prevent its universal application. In fact, the length of fungal *COX1* varies from 1584 bp to 22 kb, with the barcode region that potentially ranges between 642 bp and 12.3 kb, the size range reflecting the number and length of introns. The problem is that introns can interfere with polymerase chain reaction (PCR), also the lack of conserved regions in existing sequences seemed to preclude universal primer design [19]. Protein-coding genes provided a good resolution for species delimitation giving greater levels of phylogenetic information under certain conditions. In fact, protein coding genes tend to be variable across the entire gene, often making primer design difficult [24, 26]. For yeasts, D1/D2 domain of the nuclear large ribosomal subunit (LSU) was adopted for the characterization of species long before the concept of DNA barcoding was promoted [20, 33, 34]. Within the region of the ribosomal operon, the internal transcribed spacer (ITS) showed the highest level of identification, displaying the most clearly defined barcoding gap between intra- and inter-specific variations for the most extended range of among fungi. Therefore, it has been adopted as the universal standard barcoding region for fungi [32]. In contrast, at higher taxonomic level the resolution ability of rDNA ITS barcode resulted lower than that of diverse protein-coding genes such as *RPB1* and *RPB2* [35, 36]. Nevertheless, the usefulness of ITS as a barcode was ascribed to its robust PCR amplification fidelity (>90% success rate), a Probability of Correct Identification (PCI) of about 70% and its applicability to a broad range of sample conditions [25].

2.2.2 rDNA ribosomal genes

Inspired by molecular bacterial taxonomy, and the need to work with easily isolated or amplified nucleic acids, the initial phylogenetic and molecular identification of fungi was based on the sequencing of the nuclear ribosomal genes. The ribosomal DNA is an essential genetic element connecting transcription to translation. The rRNA represents the main structural and catalytic component of the ribosome which is translated from a large tandem repeat found at one or more *loci* in each haploid genome [37]. Each repeat contains the 26S or 28S large subunit, the 18S small subunit and the 5.8S gene, which are transcribed as a single operon, two internal transcribed spacers (ITS1 and ITS2) and a large intergenic non-transcribed spacer [38] (**Fig. 2**). A significant advantage in the use of rDNA gene sequences is that ribosomes display highly conserved region, therefore suggesting a common evolutionary history, that can be used as a pan-specific primer attachment for PCR amplification [39].

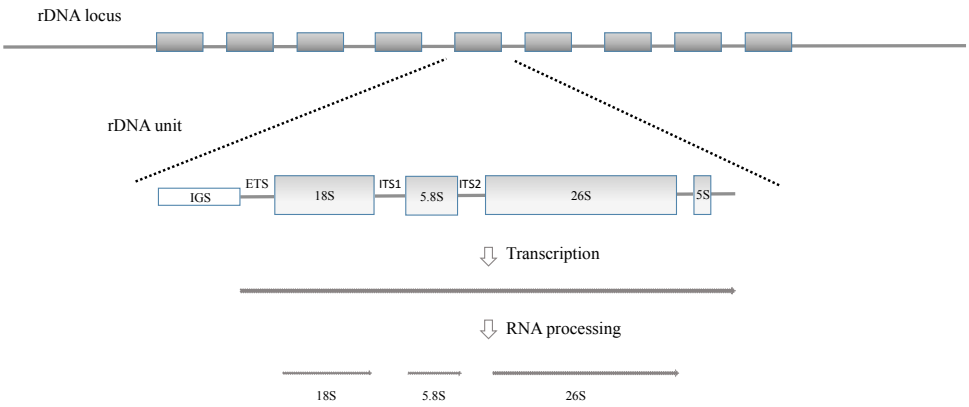


Fig. 2 Organization of the rDNA genes in eukaryotes.

The classic paper by White et al. [40] describes universal primers that are still widely used for amplifying the three main components of the fungal ribosomal operon: the LSU (including the two variable domains called D1 and D2); the small subunit 18S, separated by the ITS that bracket the conserved 5.8S region. Because of the length limitations of manual sequencing, early studies of the fungal ITS often focused only on either the ITS1 or ITS2. The White et al. [40] primers are remarkably robust, working with the vast majority of fungi.

2.2.3 D1/D2 domains of the LSU (26S) rDNA genes

The genes encoding for the major and minor subunits of the ribosome (60S and 40S) are grouped into tandem repeat units, greatly conserved during the evolution. However, these repeats show variability with a different rate of nucleotide substitutions [41]. The variable domains D1 and D2, approximately 450-600 bp in length and located at the 5' end of the LSU (**Fig. 3**), are able to discriminate between closely related species, thus providing an invaluable tool for species identification and phylogenetic reconstruction [42].

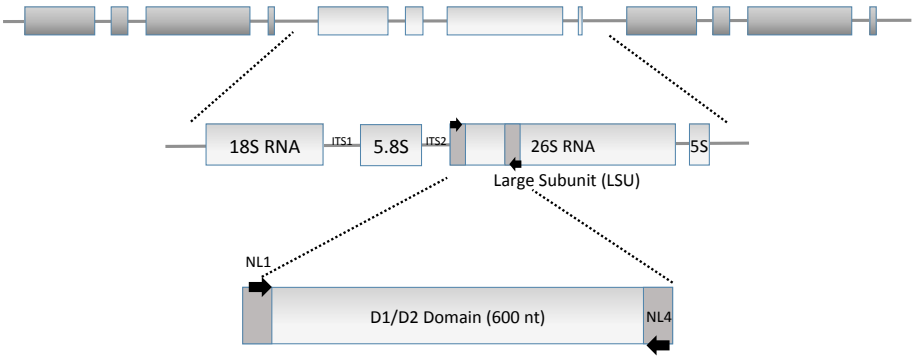


Fig. 3 rDNA ribosomal genes tandem repeats. D1/D2 domain of the LSU.

The LSU D1/D2 domain sequences are now available for the majority of the ascomycetous [20] and for a large set of basidiomycetous yeasts and yeast-like fungi [33]. Peterson and Kurtzman [43] described how various heterothallic sibling species of the genera *Issatchenkia*, *Pichia*, and *Saccharomyces* could be resolved by comparing the nucleotide sequences of their variable D2 region. They noted that conspecific strains generally had less than 1% nucleotide substitutions in region D2, whereas separate biological species had greater than this number, thus providing an empirical means for recognizing species. Further studies conducted by Meyer et al. [44] confirmed the effectiveness of D1/D2 region as a barcode. They established the degree of *taxon* separation by using LSU and actin gene. In association with the D1/D2 region of the LSU gene, the high variability of actin gene detected in sibling species permitted the best differentiation of closely related *taxa*. This demonstrated also the great advantage to use additional molecular markers. The LSU region has all the characteristics of the perfect barcode: (1) it is easy to amplify, (2) the procedures concerning sequencing and alignment do not constitute a problem, and (3) its high variability allows great discrimination ability among species. Although the LSU seems the most appropriate

locus for barcoding, the ITS region is most used as regards the kingdom of fungi, because it combines the highest resolution with the best results in terms of PCR for a wide range of species [32].

2.2.4 Internal Transcribed Spacer (ITS) as universal barcode for fungi

In the past 15-20 years, molecular identification through DNA barcoding has provided new insights into the biodiversity of many different groups of fungi thus becoming an integrated and essential part of ecological research. The entire ITS region, previously studied with traditional Sanger sequencing approaches, has been further characterized by the recently available high-throughput sequencing technologies leading to the identification and characterization in great detail of the ITS1 and ITS2 sub-regions [45, 46]. The ITS, typically 450-700 bp in length, can be further divided into three parts: ITS1 and ITS2 sub-regions with high mutation rate constitute the hyper variable portion of DNA and can be used as indicators of the evolutionary rate of the species. In addition, the conserved sequence 5.8S is comprised between ITS1 and ITS2 (**Fig. 4**) [47].

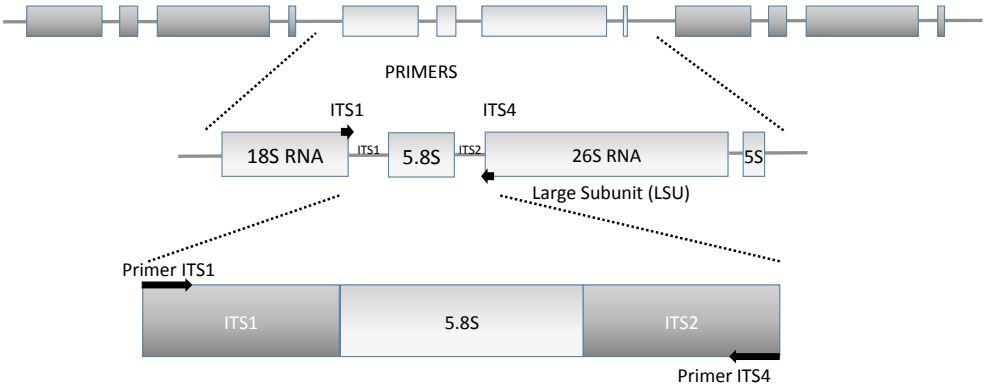


Fig. 4 Internal transcribed spacer regions.

Recently, the ITS region has been recognize as the official barcode for fungi by Schoch et al. [32]. The Fungal Barcode Consortium evaluated the potentiality of a number of fungal genes as barcode markers. Within a barcode database of 2,920 samples, a subset of 742 strains was selected and four markers, namely ITS, LSU, SSU, and RPB1 were further considered, respectively. This subset was separated into four taxonomically delimited datasets composed by 416 strains of *Pezizomycotina*, 81 strains of *Saccharomycotina*, 202 strains of *Basidiomycota*, and 43 strains from the collective lineages, respectively. Results analysis allowed to conclude that among the regions of the ribosomal cistron, the internal transcribed spacer (ITS) region has the highest

probability of successful identification for the broadest range of fungi, with the most clearly defined barcoding gap [32]. In addition, as a part of ribosomal operon this sequence is present in several copies, about 250, making the analysis possible even when the starting material is present in low amounts. For these reasons, the ITS is considered as the most attractive DNA region that can be used for the identification of organisms [47].

2.2.4.1 Heterogeneity and limits of ITS region

The rDNA is relatively conserved allowing the reconstruction of relationships of even distantly related *taxa*. Yet, there are rDNA regions variable enough to discriminate between species. The rDNA sequences may also exhibit variation within species. Different mechanisms can be responsible of this variability, for instance, a different length due to insertion or deletion (indels of single or several bases); Single Nucleotide Polymorphisms (SNPs) with no change in overall base pair numbers. The mutations that are observed with greater frequency in the ITS region are transversions, insertions and deletions, which have been recorded in a percentage higher than that expected based on the theory of concerted evolution [48]. Insertions and deletions can cause some problems during the alignment of sequences, sometimes hindering phylogenetic analysis [35]. The variability within the ITS sequences is most attributable to nucleotide polymorphisms (SNPs) which is particularly suitable for phylogenetic inference. For a long time the sequences heterogeneity within the rDNA unit has been a problem in conducting phylogenetic analyses of many species group [49-52]. A finely characterized rDNA sequence variation in multiple strains of *S. cerevisiae* for the first time [23] reported high levels of sequence variation among the individual rDNA units, ranging from 10 to 76 polymorphisms per strain across 227 variable sites. West et al. [53] used the term partial Single Nucleotide Polymorphism, or pSNP to indicate the impossibility to completely resolve polymorphisms detected across all units of tandem array. The pSNPs have been identified in species in which the hybridization events are very frequent. The same authors suggested that characterizing in fine detail the sequence variation present within the rDNA locus transforms a phylogenetic problem into a rich source of evolutionary information from which an accurate phylogenetic reconstruction can be achieved.

In fungi, the number of rDNA operon repeats ranges from a single copy to >200 copies [54, 55]. Different processes can occur within individual sequence heterogeneity in the ribosomal repeat that can, in some cases, complicate the analysis using ITS sequencing,

such as intra- and inter-*taxon* hybridization with the loss of the homogenization of the ribosomal repeat in a broad range of species. It is also demonstrated that the ITS region does not show the same degree of variability in all groups but there are differences that do not allow to determine a unique limit value through which an organism can be accurately assigned to a certain species. In fact, the inter- and intra-species distances measured through the analysis of the ITS region are often overlapping. Therefore, using only threshold value it is difficult to allocate an individual to a species within the kingdom of Fungi [56]. The ITS also showed insufficient variation in identifying some genera such as *Cladosporium* [57], *Penicillium* [58] and *Fusarium* [59]. These limitations in the use of ITS as a marker stimulated the exploration of robust primers for secondary barcodes in order to increase accuracy of species identification [25, 60]. Initially, the complete absence of reference data was a serious problem to find out additional barcodes. The standardization by the selection of one or more reference genes is crucial and stimulates large-scale phylogenetic analyses. For this reason, whether or not "one gene fits it all" is still an open debate [61]. However, the ITS barcode has been largely used in molecular identification and phylogenetic studies of a broad set of human pathogenic yeasts long before its selection as the universal fungal barcode [62-66]. The intra-species genetic analyses showed that the vast majority of medical related species had a low variability in the ITS regions. Additional analysis of alternative markers are required in order to reliably identify those species with high intra-species diversity in the ITS region [60].

2.2.5 Molecular evolution of the tandem repeats rDNA genes

The most conserved and most utilized genes in fungi, as well as in all eukaryotes, are those encoding ribosomal RNA (rRNA). Because of the massive numbers of ribosomes needed during periods of rapid growth, eukaryotes typically encode hundreds of copies of this transcription unit. Those units, organized in tandem arrays, show a uniform sequence, which is different among species. This homogeneity may occur by homologous recombination or unequal crossing over between tandem repeats, and other mechanism extensively described in concerted evolution [67]. In the concerted evolution all the members of a gene family are assumed to evolve in a concerted manner rather than independently. Concerted evolution occurs when sequence differences among reiterated copies in the genome, which are accumulating their own distinct mutations, show uniformity within the same sequence type. The role of crossing over on the

patterns of genetic diversity and genome evolution is well known [37]. A second mechanism is non crossing over gene conversion, NCGC, which occurs at the site of a double-strand DNA break without crossing over. Both crossing over and NCGC shuffle combinations of alleles across *loci* lead to degradation of linkage disequilibrium [68]. More recently, Nei et al. [69] reported that occasional duplication/deletion can occur also within the birth-and-death model of evolution where the repeats are probably maintained as a coherent family by selection and not homogenization. In this model, new genes created by gene duplication stay in the genome for a long time, whereas others are inactivated, deleted from the genome or become non functional through deleterious mutations. However, the controversy over the two models is still debated because it is difficult to distinguish between the two mechanisms when there are only a few sequence differences [67, 69, 70].

2.3 PHENOTIPYC APPROACH AS A POTENTIAL TOOL FOR FUNGAL IDENTIFICATION

The acceptance of rDNA sequence diversity as a criterion for phylogenetic discrimination heralds the transition from microbiological identification methods mainly based on the morphological features and biochemical properties of microorganisms, to molecular assays techniques. Robust amplification assays and sensitive direct detection methods are rapidly becoming standard protocols in microbiological laboratories. As mentioned above, in species discrimination the existence of some limitations in the use of the ITS marker stimulated the identification, validation and development of alternative and/or complementary tools to apply in order of increasing the accuracy of species identification. Phenotyping techniques such as time of flight mass spectrometry (MALDI-TOF) and Fourier transform infrared spectroscopy (FT-IR) represents two useful approaches that can be applied to perform high-throughput analysis and obtain rapid identification of fungal species in samples.

2.3.1 MALDI-TOF

Matrix-assisted laser desorption/ionization (MALDI) time-of-flight (TOF) mass spectrometry (MS) enables the analysis of biomolecules such as DNA, proteins, peptides and sugars in sample. It has proven to be a reproducible, accurate, fast and cost effective approach for the identification and classification of microorganisms. Actually, it represents a relevant tool for the rapid identification of pathogenic species [71-73].

MALDI was first developed in the 80's, and it represents a "soft" ionization method compared to other ionization techniques. The MALDI-TOF MS technique is versatile: it can be performed directly on intact cells [74] and even on biological samples, such as blood [71]. It has been applied to characterize molecular profiles [75] typical of yeasts and bacteria species and can be assimilated to fingerprints for the identification of microbial cells [76]. Spectral profile detected by MALDI-TOF and identified by bioinformatics procedures can be compared to those stored in reference libraries allowing the rapid identification and classification of microorganisms within a few minutes. Spectral libraries are composed of dynamic databases which, growing with the number of classified species, contribute to increase the sensibility and specificity of the classification procedure [77]. Many papers in the recent literature report that as compared to other techniques of phenotyping, MALDI-TOF MS shows superior capability to identify and classify microorganisms. In a work by Dhiman et al. [78] MALDI-TOF mass spectrometry yielded 96.3% and 84.5% accurate species level identifications, respectively. MALDI-TOF MS has been successfully applied to identify and classify with high reproducibility *Escherichia coli*, *Staphylococcus aureus,* bacteria of the HACCEK Group *(Haemophilus, Actinobacillus, Cardiobacterium, Capnocytophaga, Eikenella,* and *Kingella)*, and many others [79-81]. Also the MALDI-TOF MS identification of yeasts with clinical relevance has been reported [72, 73, 82, 83]. Yeasts infections represent a relevant problem in hospitals and in general in nosocomial environment where patients benefit either of a fast identification of pathogens and of an appropriate antimycotic therapy. The ability in classifying microorganisms together with easy sample preparation and rapid data analysis is promoting MALDI-TOF MS as an invaluable tool for clinical microbiology. It is expected in forthcoming years that self-learning procedures applied to expand MALDI-TOF MS profiles in dynamic databases will further increase the classification accuracy of available libraries.

2.3.2 Fourier Transform Infrared spectroscopy (FT-IR)

Fourier transform (FT) infrared (IR) spectroscopy (FT-IR) is a very specific and sensitive analytical technique applied to identify and quantify all detectable molecular components within the spectrum of a sample. It is also indicated as vibrational spectroscopy, which comprises both FT-IR, and Raman spectroscopy. FT-IR absorbance spectroscopy measures the loss of IR radiation transmitted through a sample across an interval of frequencies of electromagnetic spectrum. Mid-IR spectroscopy

plots the recorded intensity of absorption bands versus an interval of energies, which corresponds to changes of vibrational energy levels measuring the corresponding quantic transition from the ground level to the first energy level in molecules. This first definition summarizes the nature of this analytical technique in analysing molecules within complex biological matrices. FT-IR spectroscopy in the mid-infrared has non-destructive effects in the sample. Its limited spatial (lateral) resolution could be greatly increased when the FT-IR interferometer was coupled to an IR microscope thus enabling FT-IR microspectroscopy [84]. Since the 90s, FT-IR spectroscopy has been applied to characterize the biochemical profiles of microorganisms [85]. In FT-IR spectroscopy, absorption signatures of chemical absorption bands are obtained by transmitting mid-infrared radiation through the whole microbial cell. Different FT-IR spectroscopic techniques have been extensively used to characterize and identify fungi in many different fields like food microbiology, medical diagnostics and microbial ecology [86-89]. For example, FT-IR spectroscopy has been applied for the identification of fungal genera such as *Penicillum* and *Fusarium* spp [87], fungal phyto-pathogenes [86] and for the differentiation of *Aspergillus* and *Penicillium* at species and strain levels [88]. During the last decade infrared spectroscopy has been also employed in the identification and characterization of yeast food-related strains [90, 91] and of pathogenic strains belonging to *Candida* genus [92-96]. The advantages of using FT-IR spectroscopy are its high sensitivity, rapidity, low running cost and the applicability to all microorganisms. Currently, FT-IR spectroscopy represents the most advantageous technique to obtain complete chemical, structural and dynamical analyses of biomolecules within the spectrum of a representative population of microorganisms starting from a few biomass [84, 85, 93, 97, 98]. Recent advances in the development of high-throughput sample preparation techniques, allow the measurement of a high number of samples in short time [99]. In this approach, fungi are cultivated in 96-microwell plates for one day for yeasts and 2 days for filamentous fungi, and representative samples, subsequently deposited and dried on microwell plates (96- up to 384-microwell plates for FT-IR), are measured by high-throughput FT-IR spectroscopy setting. Also the interfering growth medium can be measured and eventually subtracted from the average spectrum of microorganisms. Otherwise, significant variations induced by microorganisms in selected media can be used to study and/or classify microorganisms [100,101], an approach which has been already applied for genome-wide phenotyping via growth parameters [102]. Identification of microorganisms via FT-IR fingerprints can be

accomplished by the use of validated spectral databases. Comprehensive databases composed of several reference strains covering a large range of species and genera are now available [103]. When suitable databases are established, spectra of unknown strains can be compared with database spectra and rapidly identified on genus, species and sometimes even at strain levels.

3. METHODOLOGIES

3.1 PCR-BASED METHODS FOR YEAST IDENTIFICATION

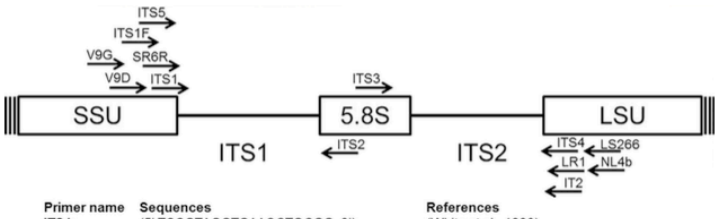
Yeast have traditionally been classified on the basis of their morphological, phenotypic and biochemical properties performing different physiological and cultural tests including: colony, cell and sporulation morphology; sugar fermentation; carbon and nitrogen assimilation, growth at different temperatures and growth in the presence of various concentration of sugars and salt [42, 104]. However, these procedures are complex and time-consuming. The progress in molecular biology has provided a large number of DNA-based approaches for the identification and characterization of yeasts including DNA-DNA hybridization [105, 106], PCR-RFLP (restriction-enzyme fragment length polymorphism) [107-109], random amplified polymorphic DNA (RAPD) analysis [110], amplified fragment length polymorphisms (AFLP) [111], microsatellite PCR fingerprinting [112] and ribosomal DNA sequencing [20]. Within these molecular techniques PCR-based methods had permit both intra-species differentiation and species identification of yeast isolates [113].

3.1.1 Amplification of the rDNA genes

Polymerase chain reaction (PCR) was developed in 1980s [114] and is based on the ability of DNA polymerase to synthesize new copies of DNA complementary to the original DNA template strand. Since DNA polymerase can add a nucleotide only onto a pre-existing 3'-OH group the presence of primers in the reaction mixture is essential to add the first nucleotides. The PCR reaction generates copies of the target sequence exponentially [115]. However, PCR reaction can be affected by some drawbacks such as sequence artefacts (PCR errors) and unequal amplification (PCR bias). PCR errors can take place with the formation of chimerical molecules, formation of heteroduplex molecules and error that can be ascribed to the lack of 3' to 5' exonuclease proofreading activity resulting in relatively low replication fidelity using *Thermus acquaticus* (*Taq*) thermo stable DNA polymerase. PCR biases can derive from the accumulation of phosphate molecules as well as from the self-annealing of the new-formed product in the last step of the amplification procedure. In this case, a "plateau effect" can occur in the PCR reaction which ceases the amplification of target DNA sequence at an exponential rate [116]. Strategies that can be adopted to prevent/reduce PCR reaction drawbacks are *i*) modify temperature setting, in particular when A/T-rich regions of

DNA are amplified [117] , *ii*) modify the number of PCR cycles [118] *iii*) modify mastermix composition including, for instance betaine, trehalose and dimethylsulfoxide (DMSO) [119] *iv*) use of new generation polymerases [120].

Primers selection represents a crucial step. The internal transcribed spacer (ITS) region contains two variable non-coding regions that are nested within the rDNA repeat between the highly conserved small subunit 5.8S and large rDNA subunit genes. The ITS region can be readily amplified with universal primers, complementary to sequences within the rDNA genes. Several primers have the ability of amplifying the entire or parts of the ITS region (**Fig. 5**).



Primer name	Sequences	References
ITS1:	(5' TCCGTAGGTGAACCTGCGG 3')	(White et al., 1990)
ITS2:	(5' GCTGCGTTCTTCATCGATGC 3')	(White et al., 1990)
ITS3:	(5' GCATCGATGAAGAACGCAGC 3')	(White et al., 1990)
ITS4:	(5' TCCTCCGCTTATTGATATGC 3')	(White et al., 1990)
ITS5:	(5' GGAAGTAAAAGTCGTAACAAGG 3')	(White et al., 1990)
ITS1F:	(5' CTTGGTCATTTAGAGGAAGTAA 3')	(Gardes and Bruns, 1993)
IT2:	(5' CCTCCGCTTATTGATATGCTTAGG 3')	(Beguin et al., 2012)
SR6R:	(5' AAGTATAAGTCGTAACAAGG 3')	(Vilgalys and Hester, 1990)
LR1:	(5' GGTTGGTTTCTTTCCT 3')	(Vilgalys and Hester, 1990)
V9D:	(5' TTAAGTCCCTGCCCTTTGTA 3')	(Gerrits van den Ende and de Hoog, 1999)
V9G:	(5' TACGTCCCTGCCCTTTGTA 3')	(Gerrits van den Ende and de Hoog, 1999)
LS266:	(5' GCATTCCCAAACAACTCGACTC 3')	(Gerrits van den Ende and de Hoog, 1999)
NL4b:	(5' GGATTCTCACCCTCTATGAC 3')	(O'Donnell, 1993)

Fig. 5 Schematic structure of the ITS region indicating universal and genus-specific primers [121].

They were recognized and utilized since early 90's [40, 122] when little was known about the variability of rDNA repeats in fungal kingdom [47]. Different PCR primers with the ability of amplifying ITS region have been described [123] but they are not greatly used as early primers.

3.2 EARLY DNA SEQUENCING

Yeasts species are now routinely identified by sequencing the internal transcribe spacer (ITS) of the ribosomal DNA repeat and sometimes in combination with the LSU rDNA genes. Previously, the traditional Sanger sequencing approach was applied to study the ITS region. In the Sanger sequencing DNA is replicated in the presence of chemically altered versions of the A, C, G, and T bases in four different tubes, each containing the appropriate amount of one of the four terminators. When incorporated into the growing strand, terminator stops the replication process, which generates a population of short

DNA fragments with variable lengths. All the generated fragments have the same 5'-end, whereas the residue at the 3'-end is determined by specific dideoxynucleotide used in the reaction. Electrophoresis on denaturing polyacrylamide gel orders these short DNA strands according to their lengths, from the shortest to the longest DNA fragments, allowing to reconstruct the whole sequence of original DNA [124]. The separation of the oligonucleotides is a difficult process but the progress of the technique has led to the development of new methods of electrophoresis, which offer the possibilities to differentiate fragments that differ in length by only one base. Technical variations of chain-termination sequencing include tagging with nucleotides containing radioactive phosphorus for radiolabelling, or using a primer labelled at the 5'-end with a fluorescent dye [125, 126]. For instance, capillary electrophoresis does not make the use of four different wells as in the sequencing by electrophoresis and separation occurs within a single column [127]. In addition, capillary electrophoresis combines high efficiency, sensitivity and resolving power allowing to separate longer DNA fragments (up to 1000 bp) with a velocity three times higher than other traditional methods [128]. Common challenges of DNA sequencing using Sanger method include poor quality in the first 20-40 bases of the sequence due to primer binding and deteriorating quality of sequencing traces after 700-900 bases. Finally, bioinformatics software can provide an estimate of quality achieved in sequences allowing to aid in the trimming of those with low-quality [129].

3.3 DNA SEQUENCING - THE NEXT GENERATION

The dideoxy method developed by Sanger marked a crucial point in the use of genetic data in the field of systematic. Advances in conventional sequencing methods led to large-scale, broad-scope biosystematics projects with a wide range of applications. The analysis of environmental DNA through the use of specific gene markers such as species-specific DNA barcodes has been a key application of next generation sequencing technologies to ecological, medical and environmental research [45]. Strategies adopted in newer sequencing technologies rely on a combination of template preparation, sequencing and imaging, and sequences alignment and assembly methods. One of the major advances offered by NGS is its ability to produce a huge amount of data cheaply, in some cases in excess of one billion short reads per instrument run [130]. Specific protocols distinguish one technology from another and determine the type of

data obtained by each platform (**Table 1**). These differences in data output present challenges when comparing platforms based on data quality and cost.

Tab. 1 Example of NGS platform families.

Platform	Clonal amplification	Chemistry	Average read lenght
454	Emulsion PCR	Pyrosequencing (seq-by-synthesis)	700bp
Illumina	Bridge amplification	Reversible dye terminator (seq-by-synthesis)	300bp
SOLiD	Emulsion PCR	Oligonucleotide chained ligation (seq-by-ligation)	75bp
Ion Torrent	Emulsion PCR	Proton detection (seq-by-synthesis)	400bp

Short-read sequencing approaches can be divided in two large categories: sequencing by ligation (SBL) and sequencing by synthesis (SBS). In most SBL and SBS configurations, DNA is clonally amplified on a solid surface [131]. In SBL approaches a probe sequence that is bound to a fluorophore hybridizes to a DNA fragment and is ligated to an adjacent oligonucleotide for imaging. The emission spectrum of the fluorophore indicates the identity of the base or bases complementary to specific positions within the probe. In SBS approaches a polymerase is used and a signal, such as a fluorophore or a change in ionic concentration, identifies the incorporation of a nucleotide into an elongating strand. This approach is defined by the use of terminator molecules that are similar to those used in Sanger sequencing, in which the ribose 3′-OH group is blocked, thus preventing elongation. To begin the process, a DNA template is primed by a sequence that is complementary to an adapter region, which will initiate polymerase binding to this double-stranded DNA (dsDNA) region. During each cycle, a mixture of all four individually labelled and 3′-blocked deoxynucleotides (dNTPs) are added. After the incorporation of a single dNTP to each elongating complementary strand, unbound dNTPs are removed and the surface is imaged to identify which dNTP was incorporated at each cluster. The fluorophore and blocking group can then be removed and a new cycle can begin [132] (**Fig. 7**).

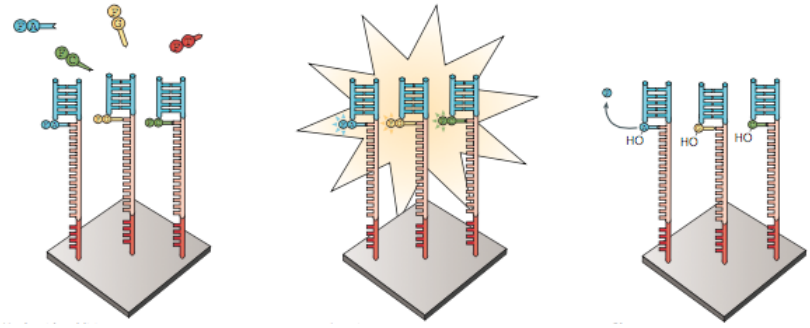


Fig. 7 Sequencing by synthesis: cyclic reversible termination approaches by Illumina system [132].

Having many thousands of identical copies of a DNA fragment, in a defined area, ensures that the signal can be distinguished from background noise. Massive parallelization is also facilitated by the creation of several millions of individual SBL or SBS reaction centres, each with its own clonal DNA template. A sequencing platform can collect information from millions of reaction centres simultaneously, thus sequencing millions of DNA molecules in parallel [130]. With this new technology it is now possible to process in parallel millions of oligonucleotides ensuring either high speed and accuracy [45]. Numerous NGS platforms have been implemented in a relatively short period of time worldwide, and the increasing demand of a number of potential users is further stimulating the market to develop new platforms. As technology progresses, a major goal will be to fill in the sequencing space with platforms that can produce higher numbers of sequences as well as longer reads per run [133]. However, some limitations of NGS platforms can negatively influence their optimal applicability and uptake in various applications. For example, time is needed to recognize and correct PCR-associated biases in a huge amount of generated sequences. Therefore, several bioinformatics methods have been developed in order to obtain optimal sequences screening and filtering of those reads that get low scores and short reads in length that may cause problems during the assembly procedures or mapping against a selected library [46]. Either the ITS1 or the ITS2 regions have been targeted in recent high-throughput sequencing studies [134-136]. Using high-throughput sequencing, thousands of sequences can be analysed from a single environmental sample, enabling in-depth analysis of the fungal diversity. When using next generation high-throughput methods, DNA barcoding is proved to be faster in species identification. This modern-automated method is now considered as accurate, economic, and less time-consuming when compared to the traditional methods [137].

3.4 DATABASE AND BIOINFORMATIC TOOLS

A correct species identification trough DNA-based method requires the continuous update of shared, public and well-annotated set of DNA sequences. Each of those sequences need to be associated with accurate specimen data and a current species name, which is strictly regulated by the International Code of Nomenclature for algae, fungi and plants (ICN) [36]. The ability to investigate the microbial complexity through DNA-based methods depends on the development of appropriate and reliable databases [138]. More than 100.000 fungal ITS sequences generated by conventional Sanger sequencing are deposited in the International Nucleotide Sequence Databases Collection (INSDC) and/or in other databases [139], providing a large reference material for identification of fungal *taxa*. The information included in INSDC comprises previous data stored in DNA Data Bank of Japan, the European Nucleotide Archive and GenBank, including the Sequence Read Archive [140]. However, these data are to some extent affected by misidentifications or technical errors such as mixing of DNA templates or sequencing errors. Nilsson et al. [141] showed that about 20% of the fungal DNA sequences from the public sequence databases leads to the incorrect identification of species, and that the majority of entries lack descriptive and up-to-date annotations. Additional databases storing highly accurate sequences, included ITS sequences, are now available: ITS Database III, UNITE, AFTOL, ITSoneDB, ISHAM database containing 2800 sequences from 421 species of pathogenic fungi for humans and animals [60] and MycoBank [142]. Although conventional sequencing has provided the most efficient method for the development of large DNA barcode reference libraries, a large amount of partial ITS sequences generated by NGS has recently been deposited in public sequence databases. All NGS sequencers produce observations of the target DNA molecule in the form of reads: sequences of single-letter base calls plus a numeric quality value (QV) for each base call [129]. Although QVs offer extra information, their use generally increases a program's CPU and RAM requirements. The reads that derive from NGS sequencing have an average length of 150bp or 300bp depending on the technology, and it is likely that finding similarity within several sequences stored in the reference database can cause ambiguous rather than correct results. Alignment of reads is one of the primary computational tasks in bioinformatics. Alignment is the process that describes how and where the reads are similar to the reference sequence. An alignment is a way of "lining up" some or all of the characters in the read with some

characters from the reference in a way that reveals how similar they are [143]. The optimal alignment of sequences with gigabases of data is quite expensive. In many cases, the alignment step could be very slow, because for each read the aligner must determine the read's likely point of origin with respect to a reference sequence [144]. Different algorithms have been developed for the alignment of the NGS reads; one of the most useful is the Bowtie (1 and 2) algorithm. The Bowtie sequence aligner was originally developed by Ben Langmead et al. [145]. The aligner is typically used with short reads and a large reference genome, or for whole genome analysis. Bowtie is promoted as "an ultrafast, memory-efficient short aligner for short DNA sequences." The speed increase of Bowtie is partly due to implementing the Burrows-Wheeler transform [146] for aligning, which reduces the memory footprint. In addition to the Burrows-Wheeler transform, Bowtie 2 also uses an FM-index [147] (similar to a suffix array) to keep its memory footprint small. Due to its implementation, Bowtie 2 is more suited to finding longer, gapped alignments in comparison with the original Bowtie method. In general, for reads longer than about 50 bp Bowtie 2 is generally faster, more sensitive, and uses less memory than Bowtie 1. Bowtie 2 supports gapped alignment with affine gap penalties and supports local alignment. Local alignments might be "trimmed" at one or both extremes in a way that optimizes alignment score. Bowtie 2 also supports end-to-end alignment, which, like Bowtie 1, requires that the read align entirely [144] (**Fig. 8**).

```
                                                  a
Read:        GACTGGGCGATCTCGACTTCG
Reference:   GACTGCGATCTCGACATCG

End-to-end alignment:

Read:        GACTG - - CGATCTCGACTTCG
             | | | | |     | | | | | | | | | | | | |
Reference:   GACTGGGCGATCTCGACATCG


                                                  b
Read:        ACGGTTGCGTTAATCCGCCACG
Reference:   TAACTTGCGTTAAATCCGCCTGG

Local alignment:

Read:        ACGGTTGCGTTAA - TCCGCCACG
                   | | | | | | | | | |  | | | | | |
Reference:   TAACTTGCGTTAAATCCGCCTGG
```

Fig. 8 Bowtie 2 algorithm: end-to-end alignment (**a**); local alignment (**b**). Dash symbols represent gaps and vertical bars show where aligned characters match.

An alignment score gives the similarity between the read and the reference sequence. The higher the score, the more similar they are. A score is calculated by subtracting penalties for each difference (e.g., mismatch and gap) and, in local alignment mode, adding bonuses for each match. The scores can be configured with the *--ma* (match bonus), *--mp* (mismatch penalty), *--np* (penalty for having an N in either the read or the reference), *--rdg* (affine read gap penalty) and *--rfg* (affine reference gap penalty) options [145]. The aligner cannot always assign a read to its point of origin with high confidence. For instance, a read that originated inside a repeat element might align equally well to many occurrences of the element throughout the reference sequence, leaving the aligner with no basis for preferring one over the others. Aligners characterize their degree of confidence in the point of origin by reporting a mapping quality such as a non-negative integer $Q = -10 \log_{10} p$, where p is an estimate of the probability that the alignment does not correspond to the read's true point of origin. Mapping quality is related to "uniqueness." An alignment is unique if it has a much higher alignment score than all the other possible alignments. The bigger the gap between the best alignment's score and the second-best alignment's score, the more unique the best alignment, and the higher its mapping quality should be [148].

Therefore, the NGS platforms have characteristic error profiles that change as the technologies improve. Error profiles can include enrichment of base call error toward the 3′ (terminal) ends of reads, compositional bias for or against high-GC sequence, and inaccurate determination of simple sequence repeats [149]. Index-aided alignment can be quite inefficient especially when alignments are permitted to contain gaps. The alignment gaps can result either from high-throughput sequencing errors or from true insertions and deletions of the sequences processed. For this reason, the analysis of data derived from NGS requires advanced computational tools that are able to align the amount of information obtained with those contained in well-annotated and possibly validated databases [47].

3.5 RAPID IDENTIFICATION OF FUNGAL RIBOSOMAL PROTEINS

In recent years matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) has emerged as a potential tool for fungal identification and diagnosis. During the MALDI-TOF MS process, fungi are identified using either intact cells or cell extracts such as ribosomal proteins. Mass spectrometry is

an analytical technique in which chemical compounds are ionized into charged molecules and ratio of their mass to charge (m/z) is measured [83]. The lasers most often used for MALDI-MS are pulsed ultraviolet (UV) lasers with wavelengths close to the maximum UV absorption of the matrix (e.g., 337 nm, nitrogen gas laser; and 355 nm, frequency- tripled Nd: YAG solid state laser) and pulse durations of 1–10 ns [150] . MALDI-MS can be performed in either the positive or the negative ion mode, measuring positive or negative ions, respectively. In most MALDI-MS applications and instruments, however, the positive ion mode usually provides higher sensitivity and spectral quality. Several chemical and physical pathways have been suggested including gas-phase photoionization, ion-molecule reactions, disproportionation, excited-state proton transfer, energy pooling, thermal ionization, and desorption of preformed ions [151]. The sample for analysis by MALDI MS is prepared by mixing or coating with solution of an energy-absorbent, organic compound called matrix. The choice of the matrix is crucial for the success of the experiment. Good matrices for proteins are derivatives of benzoic acid, cinnamic acid and other related aromatic compounds [152]. Desorption and ionization with the laser beam generates singly protonated ions from analytes in the sample. The protonated ions are then accelerated at a fixed potential, where these separate from each other on the basis of their mass to charge ratio (m/z). The ions desorption are mainly analysed using a time-of-flight mass spectrometer working in the linear or reflection mode that generate a peptide mass fingerprint (PMF). Identification of fungi by MALDI-TOF MS is done by either comparing the PMF of the sample with the PMFs contained in the database, or by matching the masses of biomarkers of unknown organism with the proteome database [83]. Several algorithms have been proposed to facilitate the matching of mass spectra from unknown sources with spectra from reference libraries. The multivariate linear least-squares regression algorithm is one method for finding the best match from a reference library for both the m/z and intensity values. A variety of software packages have been developed to process raw spectra and to perform the matching with the libraries; one feature of some of these software is the ability to create a super-spectrum for a particular strain by combining individual spectra obtained under different experimental conditions [153].

3.6 FT-IR SPECTROSCOPY - A HIGH-THROUGHPUT PHENOTYPIC APPROACH

3.6.1 Absorption of infrared light

The infrared (IR) light is a spectral region of light that covers the range of wavelengths between 780 nm to approximately 100000 nm in the electromagnetic spectrum (**Fig. 9**). This wavelengths interval is further divided in three sub-regions: (1). near infrared region (from 780 nm to 2500 nm); (2). mid infrared region (2500 nm-25000 nm); (3). far infrared region (from 25000 to about 100000 nm). In the infrared spectrum the radiation is usually expressed as wavenumber $\bar{\upsilon}$ rather then wavelengths λ, which is the reciprocal of the wavelength and it is expressed in cm^{-1} units [154].

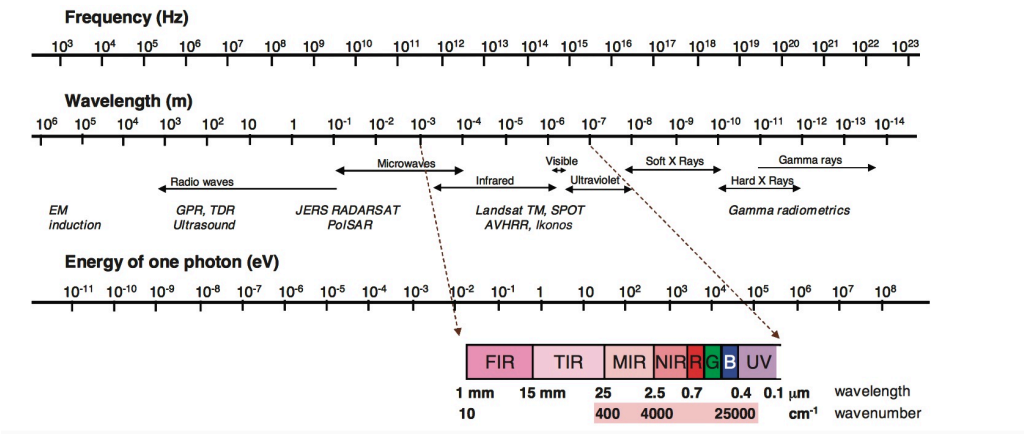


Fig. 9 Electromagnetic spectrum highlighting the visible and infrared portions [155].

There is a correlation between the wavelength and the energy of an electromagnetic wave:

$$\textbf{Eq. (1)}\ E = h\,c\,/\,\lambda = h\,f$$

where E is the energy, h is the Planck's constant, c is the speed of light and f is the frequency of the wave. Therefore, the energy of a wave is directly proportional to its frequency. The energy that relies in the molecular bonds provokes vibrations in their structures. These vibrations are associated to individual bonds and can be described as stretching, bending, rocking, twisting and wagging movements (**Fig. 10**). Therefore,

molecular vibrations result from many complex vibrational modes occurring among chemical groups and/or bonds characterizing specific molecules [156].

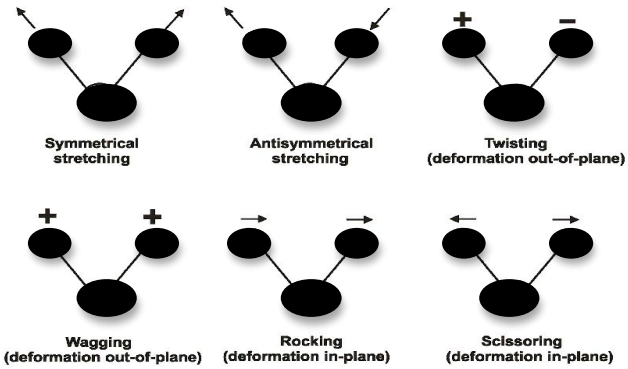


Fig. 10 Types of molecular vibrations.

When a pure molecule, is placed in the path of an IR beam light, the molecule will absorb only the frequencies of mid-IR that coincide with those of the vibrations in chemical groups and bonds allowing the molecule enters in a resonant vibration status. IR absorbance spectroscopy measures the loss of IR radiation transmitted through a sample across an interval of frequencies of electromagnetic spectrum. Depending on the selected interval of wavelengths, near-IR spectroscopy, mid-IR spectroscopy, and far-IR spectroscopy can be performed using near-IR radiation, mid-IR radiation, and far-IR radiation, respectively. Mid-IR absorbance spectroscopy plots the recorded intensity of absorption bands versus an interval of wavenumbers, $\bar{\nu}$, from 4000 to 400 cm^{-1}, which corresponds to changes of vibrational energy within the first energy level (E 0 →E 1) in molecules. Mid-IR photons have energy values comprised between 0.05 eV and 0.5 eV and fitting the quantized vibrational transitions of intra- and inter-molecular bonds of bonded atoms in molecules. Therefore, mid-IR photons can be absorbed by molecules that are in periodic (sinusoidal) motion (molecular vibrations) and the low energy associated with mid-IR photons do not induce irreversible modifications of bonds in sample molecules. This explains the non-destructive character of FT-IR spectroscopy [157,154]. According to the harmonic oscillator model, the frequency of a vibration, ν, is inversely proportional to the atomic masses and directly proportional to the bond strength and therefore the vibrational mode of a chemical bond and/or a chemical group may occur at a specific frequency. But, not all the functional groups, therefore not all the molecules, are sensitive to the mid-IR radiation. The second condition necessary for mid-IR radiation absorption is the net change of dipole moment between atoms (or

groups) connected by the bond. For instance, the planar CO_2 molecule has no permanent dipole moment, since the individual bond dipoles exactly cancel each other during symmetric stretching vibration occurring at ~1480 cm^{-1}. Nevertheless, the antisymmetric stretching, at ~2560 cm^{-1} and the bending vibrations at ~500 cm^{-1}, respectively, can be detected because there is a net change in the dipole moment of CO_2 molecule. Moreover, also the magnitude of the dipole moment change influences the intensity of absorption band. For instance, whereas the νC=O bands have strong absorbance values, more symmetric vibrations such as νC=C have weaker absorbance values or even they are not absorbing [157].

3.6.2. The Fourier Transform Infrared (FT-IR) spectra of microorganisms

The Infrared spectra of microorganisms are commonly divided in sub regions corresponding to the absorption of different macromolecules present in the cells such as lipids, proteins, nucleic acids and carbohydrates. Different molecules absorb mid-IR radiation within specific wavenumber intervals, as summarized in **Table 2**. The different spectral regions have different accuracy depending on the type of material analysed. FT-IR absorbance spectroscopy follows the Beer-Lambert Law. Therefore, the intensity of peaks in the spectrum will reflect the different samples composition. Assuming comparable thickness in homogeneous samples, the relative intensity value of the same peak in the spectra of two samples will indicate different quantities of the same molecule in the sample. The 1200-700 cm^{-1} is the most informative spectral region, often indicated as the "fingerprinting region", here not only carbohydrates but also nucleic acids DNA and RNA together with phospholipids, glycolipids, proteins, glycoproteins, phosphoproteins and many others molecular components absorb IR light. Therefore, within this spectral region we can observe a large number of bands many of which cannot be confidently assigned to vibrations of a particular molecular group [84]. The 1700-1500 cm^{-1} wavenumber interval has been associated with amide I and amide II absorbance modes. They usually give important information on proteins and in particular amide II reflects protein secondary structure [85, 158, 159]. The fact that more than one vibrational mode may concur in peaks requires the application of some type of "resolution enhancement" such as derivative spectroscopy (e.g. second derivative) together with some de-noising filter (e.g. smoothing).

Tab. 2 A summary of the vibrational frequencies of some functional groups in molecules within the mid-IR region of electromagnetic spectrum [84].

Wavenumber (cm^{-1})	Functional group	Vibrational mode	Commonly assigned biochemical component
3500 - 2500	X-H stretching vibrations (where X is C, O, or N)		
~3300	N–H	ν(N–H)	Amide A: peptide, protein
~3100	N–H	ν(N–H)	Amide B: peptide, protein
2957	$C–CH_3$	$\nu_{as}(CH_3)$	lipids
2920	$–(CH_2)_n–$	$\nu_{as}(CH_2)$	
2872	$C–CH_3$	$\nu_s(CH_3)$	
2851	$–(CH_2)_n–$	$\nu_s(CH_2)$	
2000 - 1500	fundamental stretching vibrations of double bonds (e.g., C=O, C=C, C=N)		
~1740	$–CH_2–COOR$	ν(C=O)	Phospholipid esters
~1655	O=C–N–H	80% ν(CO), 20% ν(CN)	Amide I peptide, protein
~1645	H–O–H	γ(HOH)	Water
~1545	O=C–N–H	60% γ(N–H), 30% ν(C–N), 10% ν(C–C)	Amide II peptide, protein
~1500 - 600	the "fingerprinting region": many overlapped vibrations		
~1450	$–(CH_3)_n–$ $–(CH_2)_n$	$\delta_{as}(CH_3)$ $\delta_{as}(CH_3)$	Lipid, protein
~1395	$–(CH_3)_n–$ $–(CH_2)_n$ $^-O–C=O$	$\delta_s(CH_3)$ $\delta_s(CH_3)$ ν(C=O)	Lipid, protein
~1380	$C–CH_3$	$\gamma_s(CH_3)$	Phospholipid, fatty acid, triglyceride
1400 - 1200	O=C–N–H, CH_3	γ(N–H), ν(C–N), γ(C=O), ν(C–C) and $\nu(CH_3)$	Amide III peptide, protein, collagen
~1245 - 1230	$RO–PO_2^-–OR$	$\nu_{as}(PO_2^-)$	DNA, RNA, phospholipid, phosphorylated protein
~1170	R–COO–R'	$\nu_{as}(C–O)$	Ester
~1160 and ~1120		ν(C–O)	RNA ribose
~1150	C–O, C–O–H	ν(CO), γ(COH)	carbohydrates
~1095, ~1084, ~1070	$RO–PO_2^-–OR$	$\nu_s(PO_2^-)$	DNA, RNA, phospholipid, phosphorylated protein
~1078	C–C	ν(CC)	glycogen
~1060, 1050, 1015	C–O	ν(CO)	DNA and RNA ribose
~1050	C–O–P	ν(COP)	Phosphate ester
~1028	C–O–H	def(CHO)	glycogen
~965	PO_3^{2-}	$\nu(PO_3^{2-})$	DNA and RNA ribose
~950	P–O	$\nu(PO_3^{2-})$	Phosphorylated protein
~920	C–O–P	ν(COP)	Phosphorylated protein

ν, stretching; δ, bending; γ, wagging, twisting, and rocking; def, deformation, as, antisymmetric; s, symmetric.

3.6.3. Pre-processing of FT-IR spectra

Pre-analytical and analytical variability can influence IR analyses in the spectra. This variability refers to sample collection and sample preparation procedures as well as instrumental artefacts. Protocols for spectral pre-treatment are now available [160] and in particular they focus on atmosphere compensation (CO_2, H_2O interferences) when measures are performed in air, baseline correction mostly related to scattering phenomena (resonant and non resonant scattering), normalization necessary to reduce variability among different sample thicknesses, smoothing as de-noising procedure. The

pre-treatment has a significant effect on the final results and should therefore be carefully considered [161].

FT-IR spectra of microorganisms are usually pre-processed by calculating the first or second derivative of the spectra, sometimes in combination with normalization [162]. The second derivative function usually is based on the Savitzky-Golay numerical algorithm [163]. The Savitzky-Golay procedure functions as a high-pass filter, reducing baseline vertical shift by the first derivative and slope by the second derivative. Using a selected number of smoothing points the Savitzky-Golay function calculates the polynomial fit through the data. Associated with vector normalization it is able to minimize variations transforming the spectra into a better interpretable sequence of variables [164]. Other approaches are based on Standard Normal Variate (SNV), Multiplicative Signal Correction (MSC) and Extended Multiplicative Signal Correction (EMSC) pre-processing [165]. MSC has the advantage of minimizing both additive and multiplicative interference effects. In the first pre-processing step, every spectrum is modelled with respect to a reference spectrum using a least squares fit in order to find a constant offset (baseline) and a multiplicative effect. EMSC is a reliable tool to correct additive baseline, multiplicative scaling, and interference effects. A lot of different parameters referred to physical and chemical information can be obtained and used for spectral characterization of the sample [166, 167].

3.6.4. Multivariate data analysis

The FT-IR absorbance spectrum consists of a large number of collinear input variables. A number of 900 different variables are within a 4000-400 cm^{-1} spectrum collected with 4 cm^{-1} spectral resolution. Therefore, multivariate methods are required to analyse spectral data. Those methods provide valuable insights into the chemical nature/composition of various samples as well as on their specific physical features [167]. For instance, multivariate data analysis (MVA) techniques include hierarchical cluster analysis (HCA) [159], principal component analysis (PCA) [168], partial-least-squares regression (PLSR) [169] and artificial neural networks (ANNs) [170]. Hierarchical Cluster Analysis (HCA) has become the most popular method to classify microorganisms by FT-IR [85]. The HCA follows a bottom-up strategy to discover unexpected clusters that may not be initially evident [84]. It begins with each element (whole pre-processed spectra or selected intervals) as a separate cluster and then finds clusters in a series of partitions on the basis of successively established clusters. Initially, Euclidean distances among

the comparable spectra are calculated by a standard method. Euclidean distance is a descriptor of the degree of similarity among two spectra or two clusters: the better two spectra (or clusters) match, the smaller the spectral distance. Finally, an algorithm, usually Ward's algorithm described in **Eq. (2)** performs the clustering process. Instead of determining the spectral distance, this algorithm tries to find and cluster as homogeneous groups as possible. At the end, only two groups remain:

Eq. (2) $H(r, i) = \{ [n(p) + n(i)] \cdot D(p, i) + [n(i) + n(q)] \cdot D(q, i) - n(i) \cdot D(q, i)\} / [n+n(i)]$

where *H* indicates the heterogeneity, *D* indicates distances, *n* indicates the number of spectra Subscripts "*p*" and "*q*" indicate successive clusters, whereas the "*i*" subscript designates the ith spectrum whose heterogeneity is calculated [171]. Generally, a dendrogram is the output of HCA allowing to visualize variability distances among samples. Since it does not use a *priori* knowledge for clustering (e.g.: number of classes, etc.) The "unsupervised" HCA technique has become very popular amongst microbiologists [85].

Principal Component Analysis (PCA) is a bilinear modeling method that provides an interpretable overview of the main information contained in a multidimensional table. It can be also defined as a projection method because it takes information carried by the original variables and projects onto a smaller number of latent variables called Principal Components (PC). Each PC explains a certain amount of the total information contained in the original data table with the first PC bearing the greatest. Each subsequent PC contains, in order, less information done the previous one [172]. The principal goal of applying PCA is to reduce the number of redundant variables within a confusing matrix. It is applied to extract the information in high-dimensional data set by considering only linear combinations of variables (principal components) describing most variance within the dataset [173]. The first linear combination of new variables, for instance PC_1, refers to the projection of maximal variance within the first hyperspace while the second component, PC_2, projects most of the residual variance on a second, rotated hyperspace, and so on. Being orthogonal, PCs are ranked and progressively cumulate information [101] (**Fig. 11**).

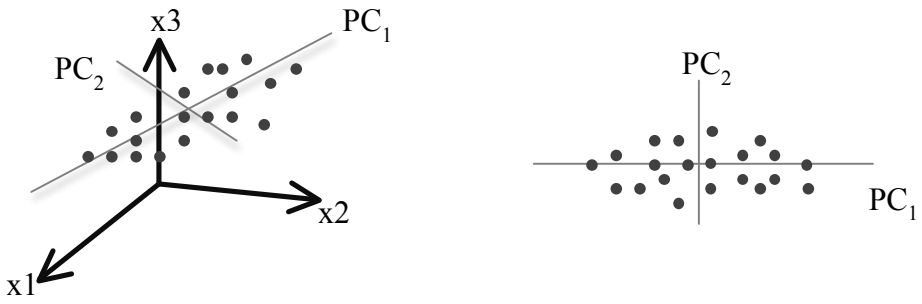


Fig. 11 Samples plotting in the new co-ordinate system of PC_1 and PC_2.

In a confusing matrix the following equation **Eq. (3)** describes the number of components:

$$\textbf{Eq. (3)}\ X = TP^T + E$$

where T represents the scores matrix, P the loadings matrix and E the residual variance. Therefore, main result outputs of PCA are scores, loadings and explained (or residual) variances. Scores describe the properties of the samples and are usually shown as a map of one PC plotted against another. Loadings describe the relationships between variables and may be plotted as a line (commonly used in spectral data interpretation) or a map (commonly used in process or sensory data analysis). Explained variance or residual variance indicates the amount of information taken into account by each PC. The combination of scores and loadings is the structured part of the data, which is the most informative. The residual, (E), is the error which represents random fluctuation not explained by the model [161]. For instance, the scores plot for PC_1 and PC_2 allowing to easily identify the number of different groups within the data set as well as to detect outliers, which could disturb the model. The closer the samples are in the scores plot, the more similar they are with respect to the two components concerned. The loadings plot describes the data structure in term of variables contributions and correlations. Each considerate variable has a loading on each PC, which reflects how much the individual variable contributes to that PC and how well the PC takes into account the variation contained in a variable [174]. In geometric terms a loading is the cosine of the angle between the variables and the current PC: the smaller the angle, the larger the loading. As a consequence, loadings can range between -1 and +1. The correlation r between two variables (vectors), x and y, is the defined by **Eq. (4)**:

Eq. (4) $r(x, y) = Cov(x, y) / s_x s_y$

where *Cov* is the covariance between *x* and *y*, *s* is the sine of *x* and *y*.

For variables, higher the loadings (e.g.: close to +1 or -1) and better and easier will be their interpretation. In addition, variables that lie close together in the loading plot are highly correlated. If both loadings have the same sign, the correlation is positive whereas in the opposite situation correlation is negative.

Specific graphical formats are used to represent correlation loadings of discrete and continuous variables. For example, the plot of discrete variables contains two ellipses indicating 100% and 50% explained variance in outer and inner ellipses, respectively (**Fig. 12**). Only variables positioning between the two ellipses contribute to the model, significantly. Those variables positioned close together in the same quadrant are positively correlated, whereas those in opposite quadrants are negatively correlated.

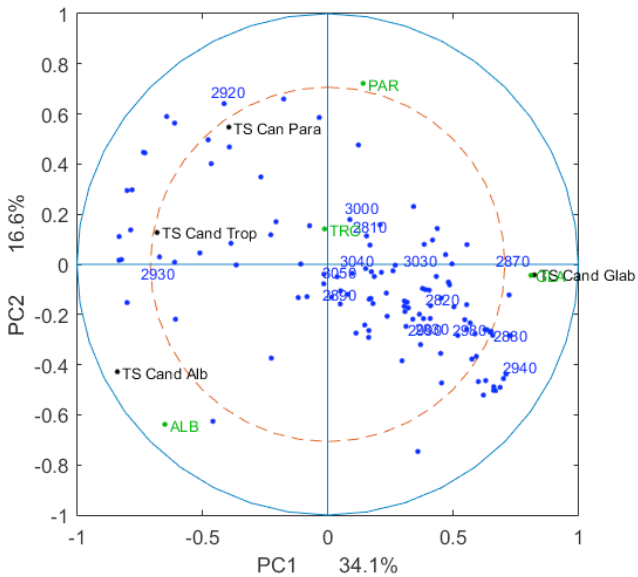


Fig. 12 Correlations Loadings plot reproduced by paper VII (Colabella et al.).

A method based on principal component analysis is the Consensus principal component analysis (CPCA). Consensus Principal Component Analysis is a multi-block method which is designed to reveal covariant patterns between and within several multivariate data sets. The computation of the parameters of this method namely, block scores, block loadings, global loadings and global scores are based on an iterative procedure [175]. It

consists of a PCA of concatenated blocks, where each block is normalized by the Frobenius norm, which is considered as a vector norm, in order to give each block the same weight. The obtained PCA scores and loadings are called global scores and global loadings. They represent the consensus of all blocks and allow studying global sample and variable variation patterns. In addition to global scores and global loadings, CPCA calculates block parameters, so-called block scores and block loadings. The block scores can be used to study the block sample variation patterns for each consensus component, i.e., the sample variation in each block that contributes to the consensus. The contribution strength of every block to the consensus can be estimated by explained block variances, which are calculated for each block. In order to study correlations between variables between and within the blocks, so-called correlation loading blocks can be used [176].

A recent technique that generalizes and combines features from principal component analysis and multiple regression is the PLSR (Partial Least Squared Regression). It is particularly useful for the prediction of a set of dependent variables from a large set of independent variables (i.e., predictors). PLSR is based on linear transition from a large number of original descriptors to a new variable space based on small number of orthogonal factors (latent variables). Factors are mutually independent (orthogonal) linear combinations of original descriptors. Unlike some similar approaches (e.g. principal component regression PCR), latent variables are chosen in such a way as to provide maximum correlation with dependent variable; thus, PLS model contains the smallest necessary number of factors [177]. With increasing number of factors, PLS model converges to ordinary multiple linear regression model. In addition, PLS approach allows one to detect relationship between activity and descriptors even if key descriptors have little contribution to the first few principal components. Therefore, PLSR can be used for the analysis of relationships between two data blocks X and Y. It owes its versatility to a combination of two aspects: bilinear approximation and linear regression. The PLSR is related to PCA, the aim is to find the variation in a data matrix X that can be used to explain the variation in a data matrix Y. The PLSR is obtained by extracting new X-variables, which are Y-relevant linear combination of the input variables from the data matrix X. Then, these new variables are used in the regression of the data matrix Y. The advantage of PLSR is that the method aims at using only the most relevant part of the variation in X for the regression of Y, while the unstable or irrelevant variation in X is left out of the calculation. The new X variables are called

PLS components; the first containing the variation in the data matrix X that is most relevant for the variation in the data matrix Y, while the second PLS component containing the variation in the data matrix X that is second most relevant for the variation in the data matrix Y and so on. As for PC's from PCA, the PLS components can be used to construct a new co-ordinate system [161, 167]. The approach to factor construction provides the description of available data using minimum number of adjustable parameters and, consequently, maximum precision and stability of regression model. However, inclusion of excessive factors in the model increases the accuracy of description but may decrease the predictivity as model starts to represent not just the true pattern of relation between descriptors and activity but also random noise and individual features of the training set. Because of this, during construction of the model its predictivity is monitored after including each successive factor by means of cross-validation procedure [178]. In cross-validation approach, computation is run several times in such a way that certain subset of the training set is not used in the model construction. Then the activity is predicted for excluded compounds using such partial model. Each compound is excluded exactly once, and normalized total error of prediction for them serves as a measure of predictivity for the full model cross-validation parameter Q^2 that is used in PLS regression to select optimal number of PLS factors. By summing up the squared errors of prediction for excluded compounds a Mean Squared Error of Cross-Validation is obtained in **Eq. (5)**:

Eq. (5) $$MSECV = \frac{1}{N}\sum_i e_i^2$$

Cross-validation parameter is defined by **Eq. (6)**:

Eq. (6) $$Q^2 = \frac{S_y^2 - MSECV}{S_y^2}$$

where S_y is the root mean square deviation of y from average value for the training set. Q^2 parameter shows to what extent the factor model constructed is better than random selection. For instance, commonly used leave-one-out cross-validation (where compounds are excluded one by one) might strongly overestimate the predictivity. Leave-one-out is the most classical exhaustive CV procedure where each data points is successively left out from the sample and used for the validation [179]. The major interest

of CV lies in the universality of the data splitting heuristics. It only assumes that data are identically distributed, and training and validation samples are independent, which can even be relaxed. Therefore, CV can be applied to almost any algorithm in almost any framework, such as regression [179, 180], density estimation [181] and classification [182, 183] among many others. This universality is not shared by most other model selection procedures, which often are specific of a framework and can be completely misleading in another one [184].

4. RESULTS AND DISCUSSION

4.1 IDENTIFICATION OF PATHOGENIC BIOFILM-FORMING STRAIN USING ITS BARCODE

The ability of some fungal species to form biofilm is considered an important factor for their persistence on medical devices and, in general, in nosocomial environment. Cells in biofilms display phenotypic traits that are dramatically different from their free-floating planktonic counterparts, such as increased resistance to antimycotic agents and protection from host defences. The biofilm forming ability has been suggested as one of the major risk factors for mortality due to species that belong to *Candida* genus. In **Paper I** a broad set of *Candida* strains, isolated from nosocomial environment, was identified and analysed for their ability to form biofilm. A first identification, according to routine clinical procedures, was carried out using CCA (Chromogenic Candida Agar) medium, which ensures a morphological and colour evaluation in a very short time. However, the use of morphological test may lead to false positives resulting in incorrect species identification. The ITS region of the rDNA operon, known as the official fungal barcode, was used to carried out the taxonomic assignments resulting in the identification of eleven species. The vast majority of the isolates belong to *C. albicans*, *C. glabrata*, *C. parapsilosis* and *C. tropicalis* species and were found in both hospitals (Pisa and Udine). The predominant species also showed different levels of biofilm forming ability, whereas the lesser present species did not produce biofilm. Species and biofilm forming ability appeared to be distributed almost randomly, although some combinations suggest a potential preference of some species or of biofilm forming strains for specific wards.

4.2 LIMIT OF ITS BARCODE IN THE DIAGNOSE OF FILAMENTOUS FUNGI

In **Paper II** a first case of endocarditis infection caused by *Trichoderma longibrachiatum* and the methodological problems inherent to its diagnosis were reported. Members of the fungal genus *Trichoderma* are saprophytic filamentous fungi, which show a large distribution in the soil, plant material, decaying vegetation and wood. Some species of *Trichoderma* can cause infections in humans. Fungal infections by *Trichoderma* spp normally cause morbidity and mortality, especially in immunocompromised patients. The definitive diagnosis of this fungus was difficult to

determine due to the lack of specific tools at the species-specific level. The current barcode ITS wasn't able to discriminate species within the *Trichoderma-Hypocrea* complex. In fact, by ITS sequencing we were not able to separate *Trichoderma longibrachiatum* from his teleomorph *Hypocrea orientalis,* even using two different databases. The identification was carried out combining molecular approach sequencing LSU rDNA genes and microbiological test using different growth media such as PSA (Potato Dextrose Agar) and SNA (Synthetic Nutrient Deficient Agar). In our case these combined approaches provided the definitive identification of this saprophytic fungal organism, requiring a time-consuming procedure.

4.3 DELIMITATION OF YEASTS FOOD/CLINIC RELATED STRAINS USING PHENOTYPIC AND MOLECULAR APPROACHES

In **Paper III** the complex *Candida/Meyerozyma guilliermondii* was described using both phenotypic and molecular approaches. *Meyerozyma guilliermondii* is known to be a colonizer of a wide natural environment and was also isolated from fruit. This species resulted of industrial interest displaying antimicrobial activity versus fruit spoiling molds. In contrast addition, *M. guilliermondii* is also known as the teleomorph of the opportunistic pathogen *Candida guilliermondii,* which causes about 2% of human blood infections. The presence of this species both in food and clinical environments poses the question on whether a selective pressure is selecting specialized strains for the different environments. In this study 96 strains isolated in both environments were analysed. The identification was firstly carried out using D1/D2 domain of the LSU. Less than 1% distances were measured within the LSU rDNA genes of all strains. The distance analysis of ITS region confirmed this low percentage allowing to conclude that all strains belong to *M. guilliermondii*. Interestingly the food isolates showed a divergence from the type strains indicating that the two groups (food, environmental/clinic) are statistically different although they belong to the same species. Differences between these two groups of isolates were also addressed by the results of both molecular and metabolomic fingerprints.

4.4 EXPLOITATION OF THE INTERNAL VARIABILITY OF THE rDNA OPERON: NGS-LIKE APPROACH

In **Paper IV** a strain isolated from an Italian vineyard was subject to sequence analysis using the two molecular marker sequences, ITS region and D1/D2 domain of the LSU

rDNA. Results indicated that this strain could not be attributed to any known species and it was described as the type strain of *Ogataea uvarum* sp.nov. The analyses conducted on the internal variants of ITS and LSU showed a significant variability. In *Ogataea uvarum* sp.nov., ITS was more variable than LSU, especially in the ITS2 region. In fact, the molecular assays showed several secondary peaks in the ITS2 sequence, but not in the LSU D1/D2 domain. In order to test whether these peaks were due to the internal heterogeneity of the rDNA operon, the region spanning from ITS1 to LSU D1/D2 was introduced in a mini library and several clones were sequenced separately. This strategy was chosen in order to determine the variations frequency among repeats, and to test whether a relation exists between the variants in the single *loci* (LSU, ITS1, 5.8S and ITS2) within the same tandem repeat copy, using an NGS-like approach. The cloning of a sample of single copy sequences showed that indeed the internal heterogeneity is present and that the process of generating a consensus using a Sanger sequencing hides a large part of it.

4.5 BRINGING THE ITS BARCODE IN THE NGS ERA

In **Paper V** we describe an innovative system of yeast strain identification using next generation sequencing of the amplicon including the region spanning from ITS1 to the D1/D2 domain of the LSU. Since the ITS region has recently been recognized as the official barcode, the D1/D2 domain of the LSU was introduced to perform NGS multi-locus sequencing. The analysis was performed on eleven pathogenic yeasts species that belong to *Candida* genus. Two different approaches, namely *de novo* assembling method and mapping against a reference method, were extensively described in terms of time and accuracy applied to assess their accuracy in identifying microorganisms. For each method two different algorithms were employed (Bowtie2 and BBMap). The statistical correlation analysis indicates that the time requested by mapping and assembly procedures are independent, whereas a weak relation exists between the algorithms employed within the same type of approach. In order to test the efficacy of mapping to a reference library, using a wide collection of sequences, we used three libraries of different size (CBS library containing only ITS sequences, CBS library with both ITS_LSU sequences and the sequences stored in the ISHAM database) using BTL and BBmap algorithms. Results showed that the time performance of the two tested algorithms varies according to the size of both the library and the FASTAq files contained the reads. For the identification of the pathogenic related yeast we developed

an easy and rapid procedure based on 3 mapping steps compared with the two algorithms employed in the previous approaches.

4.6 HT-NGS TECHNOLOGY AS A POTENTIAL TOOL FOR SNPs DETECTION

The introduction of Next Generation Sequencing leads to a deeper knowledge of the individual sequences and of the variants between the same DNA sequences located in different tandem repeats. Next Generation Sequencing offers the possibility to evaluate this heterogeneity by analyzing the Single Nucleotide Polymorphisms within the reads of an rDNA region amplified from a single strain DNA. **Paper VI** describes the internal variability of 271 strains from four prevalent yeast species of the genus *Candida*. NGS reads within the ITS-LSU amplicons were mapped against the corresponding Sanger sequence of the species type strain in order to record position and frequency relative to the references. In order to minimize background noise due to technical factors, sites with less than 1% of variants were considered non-variable. The average of the Variant Frequency (AVF) among the four species was calculated. The Variant bearing Strains Frequency (VSF), that represents the percentage of strains within each species that showed >1% of variants at each specific site of the amplicon, was also calculated. Results indicated the presence of high variability among the strains and between the species. These variants showed different distributions within the amplicons with highest concentrations in the ITS2 region. Correlation analysis between the AVF and the four rDNA region (ITS1, 5.8S, ITS2 and LSU) indicated that these four *loci* could have different rates of homogenization, probably related to different mechanisms of concerted evolution.

4.7 IDENTIFICATION OF PATHOGENIC YEASTS USING NGS BARCODING AND FT-IR JOINT-POSSIBILITIES

Correct species identification is becoming increasingly important in clinical diagnostic. To improve the quality of pathogen identification, rapid, reliable and objective identification methods are essential. Molecular information has widely contributed to the delimitation and identification of species, also for those related to infection diseases. Since yeasts infections represent a relevant problem in the various nosocomial environments, identification of pathogenic yeasts becomes crucial for mortality rates of hospitalised patients and the implementation of rapid identification may reduce both

death rates and costs related to infectious diseases. Recent advantages in the development of FT-IR spectroscopy, allow rapid identification and classification of microorganisms via chemical signatures. In **paper VII** we described a combined approach for the identification of pathogenic strains belonging to the four major species of *Candida* genus (*C. albicans*, *C. glabrata*, *C. parapsilosis* and *C. tropicalis*). All the strains were identified at species level with ITS and D1/D2 LSU marker sequences using a High-throughput Next Generation Sequencing technology. In order to evaluate relationship between the two technical approaches, consensus principal component analysis (CPCA) was employed. Using multivariate PLSR method, a classification model was also built and results showed that the most difficult group to classify was *C. glabrata*. Through the base of the PLS modelling of FT-IR spectra four distance matrices, referred to the four species, were carried out in order to calculated the distances of the strains to the taxonomic type strain (TS) and to the central strain (CS) of the distribution. Identification using both TS and CS based on PLS modelled of FT-IR spectra were also performed.

5. CONCLUSIONS AND FUTURE PROSPECTS

This thesis started with the hypothesis that ribosomal DNA variability needs to be known to improve its use in taxonomy, diagnostics and ecology. The various works presented indicate that the variability is higher than previously thought and that it can causes serious hinders to the use of these sequences in metagenomics. On the other hand, beyond the model governing the homogenization of the rDNA repeat units, the internal variability within this region can represents a source of additional information that will be useful in phylogenetic, environmental and clinical microbiology to trace the origin of the studied strains. The possibility of applying NGS offers several advantages such as the study of microbial communities independently of their viability and capacity of growing in different conditions. Furthermore, still problems exist in the exact quantification of *taxa* on the basis of the NGS outputs and a superficial approach can bring about biased conclusions. Since the ITS has been proposed as a universal barcode in fungi several limits, operational and intrinsic, were discussed and results call for the development of NGS pipelines in the application of secondary barcodes.

Another scientific hypothesis was the possibility to reconcile these DNA based markers with High-throughput FT-IR phenotypic analysis. Results showed that advanced multivariate analysis can produce significant clustering that can assume a taxonomic meaning with additional steps based on the current principle of species identification.

The possibility to use the rDNA variability for typing as well as for identification with a single analysis carried out in multiples is an interesting point that needs further understanding and work to deploy a powerful technique in all fields interested in taxonomy and biodiversity, but especially in medical diagnostics.

In conclusion, I believe that a better knowledge and an appropriate series of technological advances can be the key to open the door of the cell where this treasure is conserved.

REFERENCES

1. Mayr, E. The growth of biological thought: Diversity, evolution, and inheritance (Harvard University Press, 1982).
2. De Queiroz, K. Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences* **102**, 6600-6607 (2005).
3. De Queiroz, K. Species concepts and species delimitation. *Systematic biology* **56**, 879-886 (2007).
4. Mayr, E. The biological species concept. *Species concepts and phylogenetic theory: a debate. Columbia University Press, New York*, 17-29 (2000).
5. Naumov, G.I., Naumova, E.S. & Querol, A. Genetic study of natural introgression supports delimitation of biological species in the Saccharomyces sensu stricto complex. *Systematic and applied microbiology* **20**, 595-601 (1997).
6. Wayne, L. et al. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *International Journal of Systematic and Evolutionary Microbiology* **37**, 463-464 (1987).
7. Stackebrandt, E. & Goebel, B. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology* **44**, 846-849 (1994).
8. Rosselló-Mora, R. & Amann, R. The species concept for prokaryotes. *FEMS microbiology reviews* **25**, 39-67 (2001).
9. Hawksworth, D.L. et al. The Amsterdam declaration on fungal nomenclature. *IMA fungus* **2**, 105-112 (2011).
10. Adamowicz, S.J. & Scoles, G.J. International Barcode of Life: Evolution of a global research community. *Genome* **58**, 151-162 (2015).
11. Hebert, P.D., Cywinska, A. & Ball, S.L. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences* **270**, 313-321 (2003).
12. Casiraghi, M., Labra, M., Ferri, E., Galimberti, A. & De Mattia, F. DNA barcoding: a six-question tour to improve users' awareness about the method. *Briefings in bioinformatics*, bbq003 (2010).
13. Sanger, F.N., S.; Coulson, R. DNA sequencing with chain-terminating inhibitors. *PNAS* **74, 12**, 5463-5467 (1977).
14. Meyer, C.P. & Paulay, G. DNA barcoding: error rates based on comprehensive sampling. *PLoS biol* **3**, e422 (2005).
15. Galimberti, A. et al. DNA barcoding as a new tool for food traceability. *Food Research International* **50**, 55-63 (2013).
16. Arnot, D.E., Roper, C. & Bayoumi, R.A. Digital codes from hypervariable tandemly repeated DNA sequences in the Plasmodium falciparum circumsporozoite gene can genetically barcode isolates. *Molecular and biochemical parasitology* **61**, 15-24 (1993).
17. Floyd, R., Abebe, E., Papert, A. & Blaxter, M. Molecular barcodes for soil nematode identification. *Molecular ecology* **11**, 839-850 (2002).

18. Hebert, P.D., Ratnasingham, S. & de Waard, J.R. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences* **270**, S96-S99 (2003).
19. Seifert, K.A. et al. Prospects for fungus identification using CO1 DNA barcodes, with Penicillium as a test case. *Proceedings of the National Academy of Sciences* **104**, 3901-3906 (2007).
20. Kurtzman, C.P. & Robnett, C.J. Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. *Antonie van Leeuwenhoek* **73**, 331-371 (1998).
21. Baayen, R.P., O'Donnell, K., Breeuwsma, S., Geiser, D.M. & Waalwijk, C. Molecular relationships of fungi within the Fusarium redolens-F. hostae clade. *Phytopathology* **91**, 1037-1044 (2001).
22. Geiser, D. et al. The current status of species recognition and identification in Aspergillus. *Studies in Mycology* **59**, 1-10 (2007).
23. James, T.Y. et al. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443**, 818-822 (2006).
24. Schoch, C.L. et al. The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Systematic biology*, syp020 (2009).
25. Stielow, J. et al. 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 (2015).
26. Hofstetter, V., Miadlikowska, J., Kauff, F. & Lutzoni, F. Phylogenetic comparison of protein-coding versus ribosomal RNA-coding sequence data: a case study of the Lecanoromycetes (Ascomycota). *Molecular phylogenetics and evolution* **44**, 412-426 (2007).
27. Ertz, D. & Tehler, A. The phylogeny of Arthoniales (Pezizomycotina) inferred from nucLSU and RPB2 sequences. *Fungal Diversity* **49**, 47-71 (2011).
28. McLaughlin, D.J., Hibbett, D.S., Lutzoni, F., Spatafora, J.W. & Vilgalys, R. The search for the fungal tree of life. *Trends in microbiology* **17**, 488-497 (2009).
29. O'Donnell, K. et al. Phylogenetic analyses of RPB1 and RPB2 support a middle Cretaceous origin for a clade comprising all agriculturally and medically important fusaria. *Fungal Genetics and Biology* **52**, 20-31 (2013).
30. Druzhinina, I.S. et al. An oligonucleotide barcode for species identification in Trichoderma and Hypocrea. *Fungal Genetics and Biology* **42**, 813-828 (2005).
31. Hibbett, D.S. et al. Progress in molecular and morphological taxon discovery in Fungi and options for formal classification of environmental sequences. *Fungal Biology Reviews* **25**, 38-47 (2011).
32. Schoch, C.L. et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences* **109**, 6241-6246 (2012).
33. Fell, J.W., Boekhout, T., Fonseca, A., Scorzetti, G. & Statzell-Tallman, A. Biodiversity and systematics of basidiomycetous yeasts as determined by large-subunit rDNA D1/D2 domain sequence analysis. *International Journal of Systematic and Evolutionary Microbiology* **50**, 1351-1371 (2000).

34. Scorzetti, G., Fell, J., Fonseca, A. & Statzell-Tallman, A. Systematics of basidiomycetous yeasts: a comparison of large subunit D1/D2 and internal transcribed spacer rDNA regions. *FEMS yeast research* **2**, 495-517 (2002).
35. Seifert, K.A. Progress towards DNA barcoding of fungi. *Molecular ecology resources* **9**, 83-89 (2009).
36. Schoch, C.L. et al. Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. *Database* **2014**, bau061 (2014).
37. Richard, G.-F., Kerrest, A. & Dujon, B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews* **72**, 686-727 (2008).
38. Long, E.O. & Dawid, I.B. Repeated genes in eukaryotes. *Annual review of biochemistry* **49**, 727-764 (1980).
39. Kurtzman, C., Fell, J. & Boekhout, T. (Elsevier Sciences, Amsterdam, 1998).
40. White, T.J., Bruns, T., Lee, S. & Taylor, J. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR protocols: a guide to methods and applications* **18**, 315-322 (1990).
41. Kurtzman, C. rRNA sequence comparisons for assessing phylogenetic relationships among yeasts. *International Journal of Systematic and Evolutionary Microbiology* **42**, 1-6 (1992).
42. Kurtzman, C., Fell, J.W. & Boekhout, T. The yeasts: a taxonomic study (Elsevier, 2011).
43. Peterson, S.W. & Kurtzman, C.P. Ribosomal RNA sequence divergence among sibling species of yeasts. *Systematic and applied microbiology* **14**, 124-129 (1991).
44. Daniel, H.-M. & Meyer, W. Evaluation of ribosomal RNA and actin gene sequences for the identification of ascomycetous yeasts. *International journal of food microbiology* **86**, 61-78 (2003).
45. Shokralla, S., Spall, J.L., Gibson, J.F. & Hajibabaei, M. Next‐generation sequencing technologies for environmental DNA research. *Molecular ecology* **21**, 1794-1805 (2012).
46. Shokralla, S. et al. Next‐generation DNA barcoding: using next‐generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular ecology resources* **14**, 892-901 (2014).
47. Bellemain, E. et al. ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. *BMC microbiology* **10**, 1 (2010).
48. Simon, U.K. & Weiß, M. Intragenomic variation of fungal ribosomal genes is higher than previously thought. *Molecular biology and evolution* **25**, 2251-2254 (2008).
49. Buckler, E.S., Ippolito, A. & Holtsford, T.P. The evolution of ribosomal DNA divergent paralogues and phylogenetic implications. *Genetics* **145**, 821-832 (1997).
50. Álvarez, I. & Wendel, J.F. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular phylogenetics and evolution* **29**, 417-434 (2003).
51. Nilsson, R.H., Kristiansson, E., Ryberg, M., Hallenberg, N. & Larsson, K.H. Intraspecific ITS variability in the kingdom Fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evolutionary bioinformatics* **4** (2008).

52. Kiss, L. Limits of nuclear ribosomal DNA internal transcribed spacer (ITS) sequences as species barcodes for Fungi. *Proceedings of the National Academy of Sciences* **109**, E1811-E1811 (2012).
53. West, C., James, S.A., Davey, R.P., Dicks, J. & Roberts, I.N. Ribosomal DNA sequence heterogeneity reflects intraspecies phylogenies and predicts genome structure in two contrasting yeast species. *Systematic biology* **63**, 543-554 (2014).
54. Baldrian, P. et al. Estimation of fungal biomass in forest litter and soil. *Fungal ecology* **6**, 1-11 (2013).
55. Howlett, B.J., Rolls, B.D. & Cozijnsen, A.J. Organisation of ribosomal DNA in the ascomycete Leptosphaeria maculans. *Microbiological research* **152**, 261-267 (1997).
56. Puillandre, N., Lambert, A., Brouillet, S. & Achaz, G. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular ecology* **21**, 1864-1877 (2012).
57. Schubert, K., Braun, U., Groenewald, J. & Crous, P. Cladosporium leaf-blotch and stem rot of Paeonia spp. caused by Dichocladosporium chlorocephalum gen. nov. *Studies in Mycology* **58**, 95-104 (2007).
58. Skouboe, P. et al. Phylogenetic analysis of nucleotide sequences from the ITS region of terverticillate Penicillium species. *Mycological Research* **103**, 873-881 (1999).
59. O'Donnell, K. & Cigelnik, E. Two divergent intragenomic rDNA ITS2 types within a monophyletic lineage of the FungusFusariumAre nonorthologous. *Molecular phylogenetics and evolution* **7**, 103-116 (1997).
60. Irinyi, L. et al. 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 (2015).
61. Vu, D. et al. DNA barcoding analysis of more than 9000 yeast isolates contributes to quantitative thresholds for yeast species and genera delimitation. *Studies in Mycology* (2016).
62. Katsu, M. et al. The internal transcribed spacers and 5.8 S rRNA gene show extensive diversity among isolates of the Cryptococcus neoformans species complex. *FEMS Yeast Research* **4**, 377-388 (2004).
63. Leaw, S.N. et al. Identification of medically important yeast species by sequence analysis of the internal transcribed spacer regions. *Journal of Clinical Microbiology* **44**, 693-699 (2006).
64. Begerow, D., Nilsson, H., Unterseher, M. & Maier, W. Current state and perspectives of fungal DNA barcoding and rapid identification procedures. *Applied Microbiology and Biotechnology* **87**, 99-108 (2010).
65. Romanelli, A.M., Sutton, D.A., Thompson, E.H., Rinaldi, M.G. & Wickes, B.L. Sequence-based identification of filamentous basidiomycetous fungi from clinical specimens: a cautionary note. *Journal of clinical microbiology* **48**, 741-752 (2010).
66. Estrada-Bárcenas, D.A. et al. Genetic diversity of Histoplasma and Sporothrix complexes based on sequences of their ITS1-5.8 S-ITS2 regions from the BOLD System. *Revista Iberoamericana de Micología* **31**, 90-94 (2014).
67. Eickbush, T.H. & Eickbush, D.G. Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* **175**, 477-485 (2007).

68. Korunes, K.L. & Noor, M.A. Gene conversion and linkage: effects on genome evolution and speciation. *Molecular Ecology* (2016).
69. Nei, M. & Rooney, A.P. Concerted and birth-and-death evolution of multigene families. *Annual review of genetics* **39**, 121 (2005).
70. Nei, M. Selectionism and neutralism in molecular evolution. *Molecular biology and evolution* **22**, 2318-2342 (2005).
71. Spanu, T. et al. Direct MALDI-TOF mass spectrometry assay of blood culture broths for rapid identification of Candida species causing bloodstream infections: an observational study in two large microbiology laboratories. *Journal of clinical microbiology* **50**, 176-179 (2012).
72. Wieser, A., Schneider, L., Jung, J. & Schubert, S. MALDI-TOF MS in microbiological diagnostics—identification of microorganisms and beyond (mini review). *Applied microbiology and biotechnology* **93**, 965-974 (2012).
73. Putignani, L. et al. MALDI-TOF mass spectrometry proteomic phenotyping of clinically relevant fungi. *Molecular BioSystems* **7**, 620-629 (2011).
74. Fenselau, C. & Demirev, P.A. Characterization of intact microorganisms by MALDI mass spectrometry. *Mass Spectrometry Reviews* **20**, 157-171 (2001).
75. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207 (2003).
76. Giebel, R. et al. Microbial fingerprinting using matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS): applications and challenges. *Advances in applied microbiology* **71**, 149-184 (2010).
77. Normand, A.-C. et al. Assessment of various parameters to improve MALDI-TOF MS reference spectra libraries constructed for the routine identification of filamentous fungi. *BMC microbiology* **13**, 1 (2013).
78. Dhiman, N., Hall, L., Wohlfiel, S.L., Buckwalter, S.P. & Wengenack, N.L. Performance and cost analysis of matrix-assisted laser desorption ionization–time of flight mass spectrometry for routine identification of yeast. *Journal of clinical microbiology* **49**, 1614-1616 (2011).
79. Croxatto, A., Prod'hom, G. & Greub, G. Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *FEMS microbiology reviews* **36**, 380-407 (2012).
80. Arnold, R.J. & Reilly, J.P. Fingerprint matching of E. coli strains with matrix ‐ assisted laser desorption/ionization time ‐ of ‐ flight mass spectrometry of whole cells using a modified correlation approach. *Rapid Communications in Mass Spectrometry* **12**, 630-636 (1998).
81. Firacative, C., Trilles, L. & Meyer, W. MALDI-TOF MS enables the rapid identification of the major molecular types within the Cryptococcus neoformans/C. gattii species complex. *PLoS One* **7**, e37566 (2012).
82. Marklein, G. et al. Matrix-assisted laser desorption ionization-time of flight mass spectrometry for fast and reliable identification of clinical yeast isolates. *Journal of clinical microbiology* **47**, 2912-2917 (2009).
83. Singhal, N., Kumar, M., Kanaujia, P.K. & Virdi, J.S. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Frontiers in microbiology* **6** (2015).
84. Bellisola, G. & Sorio, C. Infrared spectroscopy and microscopy in cancer research and diagnosis. *Am J Cancer Res* **2**, 1-21 (2012).
85. Naumann, D., Helm, D. & Labischinski, H. Microbiological characterizations by FT-IR spectroscopy. *Nature* **351**, 81 (1991).

86. Erukhimovitch, V. et al. Identification of fungal phyto-pathogens by Fourier-transform infrared (FTIR) microscopy. *J Agric Technol* **1**, 145-152 (2005).
87. Erukhimovitch, V., Pavlov, V., Talyshinsky, M., Souprun, Y. & Huleihel, M. FTIR microscopy as a method for identification of bacterial and fungal infections. *Journal of pharmaceutical and biomedical analysis* **37**, 1105-1108 (2005).
88. Fischer, G., Braun, S., Thissen, R. & Dott, W. FT-IR spectroscopy as a tool for rapid identification and intra-species characterization of airborne filamentous fungi. *Journal of Microbiological Methods* **64**, 63-77 (2006).
89. Shapaval, V. et al. A high‐throughput microcultivation protocol for FTIR spectroscopic characterization and identification of fungi. *Journal of biophotonics* **3**, 512-521 (2010).
90. Büchl, N.R., Wenning, M., Seiler, H., Mietke‐Hofmann, H. & Scherer, S. Reliable identification of closely related Issatchenkia and Pichia species using artificial neural network analysis of Fourier‐transform infrared spectra. *Yeast* **25**, 787-798 (2008).
91. Kümmerle, M., Scherer, S. & Seiler, H. Rapid and reliable identification of food-borne yeasts by Fourier-transform infrared spectroscopy. *Applied and environmental microbiology* **64**, 2207-2214 (1998).
92. Adt, I., Toubas, D., Pinon, J.-M., Manfait, M. & Sockalingum, G.D. FTIR spectroscopy as a potential tool to analyse structural modifications during morphogenesis of Candida albicans. *Archives of microbiology* **185**, 277-285 (2006).
93. Toubas, D. et al. FTIR spectroscopy in medical mycology: applications to the differentiation and typing of Candida. *Analytical and bioanalytical chemistry* **387**, 1729-1737 (2007).
94. Essendoubi, M. et al. Epidemiological investigation and typing of Candida glabrata clinical isolates by FTIR spectroscopy. *Journal of microbiological methods* **71**, 325-331 (2007).
95. Essendoubi, M. et al. Rapid identification of Candida species by FT-IR microspectroscopy. *Biochimica et Biophysica Acta (BBA)-General Subjects* **1724**, 239-247 (2005).
96. Sandt, C. et al. Use of Fourier-transform infrared spectroscopy for typing of Candida albicans strains isolated in intensive care units. *Journal of clinical microbiology* **41**, 954-959 (2003).
97. Timmins, É.M., Howell, S.A., Alsberg, B.K., Noble, W.C. & Goodacre, R. Rapid differentiation of closely relatedCandida species and strains by pyrolysis-mass spectrometry and fourier transform-infrared spectroscopy. *Journal of Clinical Microbiology* **36**, 367-374 (1998).
98. Orsini, F. et al. FT-IR microspectroscopy for microbiological studies. *Journal of microbiological methods* **42**, 17-27 (2000).
99. Kohler, A. et al. High-Throughput Biochemical Fingerprinting of Saccharomyces cerevisiae by Fourier Transform Infrared Spectroscopy. *PloS one* **10**, e0118052 (2015).
100. Zhao, H., Kassama, Y., Young, M., Kell, D.B. & Goodacre, R. Differentiation of Micromonospora isolates from a coastal sediment in Wales on the basis of Fourier transform infrared spectroscopy, 16S rRNA sequence analysis, and the amplified fragment length polymorphism technique. *Applied and environmental microbiology* **70**, 6619-6627 (2004).

101. Shapaval, V. et al. Characterization of food spoilage fungi by FTIR spectroscopy. *Journal of applied microbiology* **114**, 788-796 (2013).
102. Warringer, J. & Blomberg, A. Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in Saccharomyces cerevisiae. *Yeast* **20**, 53-67 (2003).
103. Wenning, M. & Scherer, S. Identification of microorganisms by FTIR spectroscopy: perspectives and limitations of the method. *Applied microbiology and biotechnology* **97**, 7111-7120 (2013).
104. Kurtzman, C.P. & Robnett, C.J. Phylogenetic relationships among yeasts of the 'Saccharomyces complex'determined from multigene sequence analyses. *FEMS yeast research* **3**, 417-432 (2003).
105. Carle, G.F. & Olson, M.V. An electrophoretic karyotype for yeast. *Proceedings of the National Academy of Sciences* **82**, 3756-3760 (1985).
106. Török, T., Rockhold, D. & King, A. Use of electrophoretic karyotyping and DNA-DNA hybridization in yeast identification. *International journal of food microbiology* **19**, 63-80 (1993).
107. Esteve-Zarzoso, B., Belloch, C., Uruburu, F. & Querol, A. Identification of yeasts by RFLP analysis of the 5.8 S rRNA gene and the two ribosomal internal transcribed spacers. *International Journal of Systematic and Evolutionary Microbiology* **49**, 329-337 (1999).
108. Granchi, L., Bosco, M., Messini, A. & Vincenzini, M. Rapid detection and quantification of yeast species during spontaneous wine fermentation by PCR–RFLP analysis of the rDNA ITS region. *Journal of Applied Microbiology* **87**, 949-956 (1999).
109. Guillamón, J.M., Sabaté, J., Barrio, E., Cano, J. & Querol, A. Rapid identification of wine yeast species based on RFLP analysis of the ribosomal internal transcribed spacer (ITS) region. *Archives of Microbiology* **169**, 387-392 (1998).
110. Couto, M.B., Van der Vossen, J., Hofstra, H. & in't Veld, J.H. RAPD analysis: a rapid technique for differentiation of spoilage yeasts. *International journal of food microbiology* **24**, 249-260 (1994).
111. Majer, D., Mithen, R., Lewis, B.G., Vos, P. & Oliver, R.P. The use of AFLP fingerprinting for the detection of genetic variation in fungi. *Mycological Research* **100**, 1107-1111 (1996).
112. Couto, M.B., Hartog, B., in't Veld, J.H., Hofstra, H. & Van der Vossen, J. Identification of spoilage yeasts in a food-production chain by microsatellite polymerase chain reaction fingerprinting. *Food microbiology* **13**, 59-67 (1996).
113. Hierro, N., Gonzalez, A., Mas, A. & Guillamón, J. New PCR - based methods for yeast identification. *Journal of applied microbiology* **97**, 792-801 (2004).
114. Mullis, K. et al. in Cold Spring Harbor symposia on quantitative biology 263-273 (Cold Spring Harbor Laboratory Press, 1986).
115. Mullis, K. in Annales de biologie clinique 579-582 (1989).
116. Acinas, S.G., Sarma-Rupavtarm, R., Klepac-Ceraj, V. & Polz, M.F. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied and environmental microbiology* **71**, 8966-8969 (2005).
117. Su, X.-z., Wu, Y., Sifri, C.D. & Wellems, T.E. Reduced extension temperatures required for PCR amplification of extremely A+ T-rich DNA. *Nucleic acids research* **24**, 1574-1575 (1996).

118. Lahr, D.J. & Katz, L.A. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques* **47**, 857-866 (2009).
119. Markarian, S.A., Asatryan, A.M., Grigoryan, K.R. & Sargsyan, H.R. Effect of diethylsulfoxide on the thermal denaturation of DNA. *Biopolymers* **82**, 1-5 (2006).
120. Fazekas, A.J., Steeves, R. & Newmaster, S.G. Improving sequencing quality from PCR products containing long mononucleotide repeats. *Biotechniques* **48**, 277-285 (2010).
121. Irinyi, L., Lackner, M., De Hoog, G.S. & Meyer, W. DNA barcoding of fungi causing infections in humans and animals. *Fungal biology* **120**, 125-136 (2016).
122. Gardes, M. & Bruns, T.D. ITS primers with enhanced specificity for basidiomycetes - application to the identification of mycorrhizae and rusts. *Molecular ecology* **2**, 113-118 (1993).
123. Martin, K.J. & Rygiewicz, P.T. Fungal-specific PCR primers developed for analysis of the ITS region of environmental DNA extracts. *BMC microbiology* **5**, 1 (2005).
124. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463-5467 (1977).
125. Ju, J., Ruan, C., Fuller, C.W., Glazer, A.N. & Mathies, R.A. Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis. *Proceedings of the National Academy of Sciences* **92**, 4347-4351 (1995).
126. Lee, L. et al. New energy transfer dyes for DNA sequencing. *Nucleic acids research* **25**, 2816-2822 (1997).
127. Franca, L.T., Carrilho, E. & Kist, T.B. A review of DNA sequencing techniques. *Quarterly reviews of biophysics* **35**, 169-200 (2002).
128. Swerdlow, H. & Gesteland, R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research* **18**, 1415-1419 (1990).
129. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. *Genome research* **8**, 175-185 (1998).
130. Metzker, M.L. Sequencing technologies—the next generation. *Nature reviews genetics* **11**, 31-46 (2010).
131. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends in genetics* **24**, 133-141 (2008).
132. Goodwin, S., McPherson, J.D. & McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**, 333-351 (2016).
133. Hodkinson, B.P. & Grice, E.A. Next-generation sequencing: a review of technologies and tools for wound microbiome research. *Advances in wound care* **4**, 50-58 (2015).
134. Buee, M. et al. 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist* **184**, 449-456 (2009).
135. Ghannoum, M.A. et al. Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS pathog* **6**, e1000713 (2010).
136. Jumpponen, A. & Jones, K. Massively parallel 454-sequencing of Quercus macrocarpa phyllosphere fungal communities indicates reduced richness and diversity in urban environments. *New Phytol* **184**, 438-448 (2009).

137. Chakraborty, C., Doss, C.G.P., Patra, B.C. & Bandyopadhyay, S. DNA barcoding to map the microbial communities: current advances and future directions. *Applied microbiology and biotechnology* **98**, 3425-3436 (2014).
138. Santamaria, M. et al. Reference databases for taxonomic assignment in metagenomics. *Briefings in bioinformatics*, bbs036 (2012).
139. Nilsson, R.H., Ryberg, M., Abarenkov, K., Sjökvist, E. & Kristiansson, E. The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. *FEMS Microbiology Letters* **296**, 97-101 (2009).
140. Nakamura, Y., Cochrane, G. & Karsch-Mizrachi, I. The international nucleotide sequence database collaboration. *Nucleic acids research* **41**, D21-D24 (2013).
141. Nilsson, R.H. et al. Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS One* **1**, e59 (2006).
142. Robert, V. et al. MycoBank gearing up for new horizons. *IMA fungus* **4**, 371-379 (2013).
143. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
144. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).
145. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25 (2009).
146. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
147. Ferragina, P., Manzini, G., Mäkinen, V. & Navarro, G. in International Symposium on String Processing and Information Retrieval 150-160 (Springer, 2004).
148. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18**, 1851-1858 (2008).
149. Miller, J.R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315-327 (2010).
150. Cramer, R. & Dreisewerd, K. UV matrix-assisted laser desorption ionization: principles instrumentation and applications. (2007).
151. Zenobi, R. & Knochenmuss, R. Ion formation in MALDI mass spectrometry. *Mass Spectrometry Reviews* **17**, 337-366 (1998).
152. Hillenkamp, F., Karas, M., Beavis, R.C. & Chait, B.T. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical chemistry* **63**, 1193A-1203A (1991).
153. Barbuddhe, S.B. et al. Rapid identification and typing of listeria species by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Applied and Environmental Microbiology* **74**, 5402-5407 (2008).
154. Smith, W.L. et al. The retrieval of planetary boundary layer structure using ground-based infrared spectral radiance measurements. *Journal of Atmospheric and Oceanic Technology* **16**, 323-333 (1999).
155. Rossel, R.V., Walvoort, D., McBratney, A., Janik, L.J. & Skjemstad, J. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy

for simultaneous assessment of various soil properties. *Geoderma* **131**, 59-75 (2006).
156. Williams, P. & Norris, K. Near-infrared technology in the agricultural and food industries (American Association of Cereal Chemists, Inc., 1987).
157. Campbell, I.D. & Dwek, R.A. Biological spectroscopy (Benjamin/Cummings Pub. Co., 1984).
158. Helm, D., Labischinski, H. & Naumann, D. Elaboration of a procedure for identification of bacteria using Fourier-Transform IR spectral libraries: a stepwise correlation approach. *Journal of Microbiological Methods* **14**, 127-142 (1991).
159. Helm, D., Labischinski, H., Schallehn, G. & Naumann, D. Classification and identification of bacteria by Fourier-transform infrared spectroscopy. *Microbiology* **137**, 69-79 (1991).
160. Becker, S.A. et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature protocols* **2**, 727-738 (2007).
161. Kohler, A. et al. Interpreting several types of measurements in bioscience. *Biomedical Vibrational Spectroscopy. Hoboken, New Jersey, USA: John Wiley & Sons*, 333-356 (2008).
162. Kansiz, M. et al. Fourier transform infrared microspectroscopy and chemometrics as a tool for the discrimination of cyanobacterial strains. *Phytochemistry* **52**, 407-417 (1999).
163. Savitzky, A. & Golay, M.J. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry* **36**, 1627-1639 (1964).
164. Zimmermann, B. & Kohler, A. Optimizing Savitzky–Golay parameters for improving spectral resolution and quantification in infrared spectroscopy. *Applied spectroscopy* **67**, 892-902 (2013).
165. Kohler, A., Zimonja, M., Segtnan, V. & Martens, H. Standard Normal Variate, Multiplicative Signal Correction and Extended Multiplicative Signal Correction Preprocessing in Biospectroscopy-2.09. (2009).
166. Kohler, A. et al. Reducing inter-replicate variation in Fourier transform infrared spectroscopy by extended multiplicative signal correction. *Applied spectroscopy* **63**, 296-305 (2009).
167. Kohler, A., Kirschner, C., Oust, A. & Martens, H. Extended multiplicative signal correction as a tool for separation and characterization of physical and chemical information in Fourier transform infrared microscopy images of cryo-sections of beef loin. *Applied spectroscopy* **59**, 707-716 (2005).
168. Naumann, D. Infrared spectroscopy in microbiology. *Encyclopedia of analytical chemistry* (2000).
169. Abdi, H. Partial least square regression (PLS regression). *Encyclopedia for research methods for the social sciences*, 792-795 (2003).
170. Schmitt, J., Udelhoven, T., Naumann, D. & Flemming, H. in BiOS'98 International Biomedical Optics Symposium 236-244 (International Society for Optics and Photonics, 1998).
171. Murtagh, F. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal* **26**, 354-359 (1983).
172. Esbensen, K.H., Guyot, D., Westad, F. & Houmoller, L.P. Multivariate data analysis: in practice: an introduction to multivariate data analysis and experimental design (Multivariate Data Analysis, 2002).

173. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemometrics and intelligent laboratory systems* **2**, 37-52 (1987).
174. Camo Process, A. Software Unscrambler. *Oslo, Norway* (2002).
175. Hanafi, M., Kohler, A. & Qannari, E.-M. Connections between multiple co-inertia analysis and consensus principal component analysis. *Chemometrics and intelligent laboratory systems* **106**, 37-40 (2011).
176. Hassani, S. et al. Analysis of-omics data: Graphical interpretation-and validation tools in multi-block methods. *Chemometrics and Intelligent Laboratory Systems* **104**, 140-153 (2010).
177. Höskuldsson, A. PLS regression methods. *Journal of chemometrics* **2**, 211-228 (1988).
178. Wold, S., Ruhe, A., Wold, H. & Dunn, I., WJ. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing* **5**, 735-743 (1984).
179. Stone, M. Cross-validation and multinomial prediction. *Biometrika*, 509-515 (1974).
180. Geisser, S. The predictive sample reuse method with applications. *Journal of the American Statistical Association* **70**, 320-328 (1975).
181. Rudemo, M. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 65-78 (1982).
182. Devroye, L. & Wagner, T. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory* **25**, 601-604 (1979).
183. Bartlett, P.L., Boucheron, S. & Lugosi, G. Model selection and error estimation. *Machine Learning* **48**, 85-113 (2002).
184. Arlot, S. & Celisse, A. A survey of cross-validation procedures for model selection. *Statistics surveys* **4**, 40-79 (2010).

Paper I

SCIENTIFIC REPORTS

OPEN

Received: 23 February 2016
Accepted: 04 May 2016
Published: 01 June 2016

Exploring ecological modelling to investigate factors governing the colonization success in nosocomial environment of *Candida albicans* and other pathogenic yeasts

Laura Corte[1], Luca Roscini[1], Claudia Colabella[1], Carlo Tascini[2], Alessandro Leonildi[2], Emanuela Sozio[3], Francesco Menichetti[2], Maria Merelli[4], Claudio Scarparo[5], Wieland Meyer[6], Gianluigi Cardinali[1,7] & Matteo Bassetti[4]

Two hundred seventy seven strains from eleven opportunistic species of the genus *Candida*, isolated from two Italian hospitals, were identified and analyzed for their ability to form biofilm in laboratory conditions. The majority of *Candida albicans* strains formed biofilm while among the NCAC species there were different level of biofilm forming ability, in accordance with the current literature. The relation between the variables considered, i.e. the departments and the hospitals or the species and their ability to form biofilm, was tested with the assessment of the probability associated to each combination. Species and biofilm forming ability appeared to be distributed almost randomly, although some combinations suggest a potential preference of some species or of biofilm forming strains for specific wards. On the contrary, the relation between biofilm formation and species isolation frequency was highly significant (R^2 around 0.98). Interestingly, the regression analyses carried out on the data of the two hospitals separately were rather different and the analysis on the data merged together gave a much lower correlation. These findings suggest that, harsh environments shape the composition of microbial species significantly and that each environment should be considered *per se* to avoid less significant statistical treatments.

Candida bloodstream infection (BI) is an important cause of morbility and mortality in health care settings and represents the fourth cause of nosocomial sepsis in the USA and in most developed countries[1,2]. Candidemia is responsible for unacceptable percentages of attributable and overall mortality rate ranging from 30–81% and from 5–71%[3], respectively. The incidence of candidemia rose in the last decades of the 20th century due to several risk factors[4]. *Candida albicans* remains the most common species causing BI, followed by several "non *Candida albicans Candida* species" (NCAC) among which *Candida glabrata, Candida parapsilosis* and *Candida tropicalis* are usually the most common and show increasing incidence[5–7]. The various species are a further factor to interpret the origin of infection, in fact *C. albicans, C. glabrata* and *C. tropicalis* are considered predominantly commensal and therefore more present in cases of endogenous infections. On the contrary, species

[1]Department of Pharmaceutical Sciences-Microbiology, University of Perugia, Borgo 20 Giugno 74, 06121 Perugia, Italy. [2]U.O. Malattie Infettive, Azienda Ospedaliera Universitaria Pisana, Via Paradisa 2, Cisanello, 56100 Pisa, Italy. [3]U.O. Medicina d'Urgenza (Emergency Medicine Unit) Universitaria, Azienda Ospedaliera Universitaria Pisana, Via Paradisa 2, Cisanello, 56100 Pisa, Italy. [4]Infectious Diseases Clinic, Santa Maria Misericordia University Hospital, Piazzale Santa Maria della Misericordia, 15, 33100 Udine, Italy. [5]Microbiology Unit, Santa Maria Misericordia University Hospital, Piazzale Santa Maria della Misericordia, 15, 33100 Udine, Italy. [6]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, Australia. [7]CEMIN, Centre of Excellence on Nanostructured Innovative Materials, Department of Chemistry, Biology and Biotechnology, University of Perugia, Via Elce di Sotto 8, 06123 Perugia, Italy. Correspondence and requests for materials should be addressed to G.C. (email: gianluigi.cardinali@unipg.it)

found in the natural and anthropic environment, as *C. parapsilosis*, *Meyerozyma guilliermondii* (telomorph of *C. guilliermondii*) and *Wickerhamomyces anomalus*, are more present in exogenous infections[8]. The recent finding that medical and food isolates of *M. guilliermondii* cluster differently, according to the ITS and LSU markers, poses the question on whether the exogenous infection is caused by strains of the nosocomial environment or of other niches[9]. The increasing frequency of NCAC species has been extensively reported in the last years[4,6,10–13] with significant epidemiological and ecological differences among various geographic areas[7,14]. This situation represents a serious threat, complicated by a significantly lesser knowledge of biofilm and resistance mechanisms in NCAC then in *C. albicans*.

The ability of some fungal species to form biofilm is considered an important factor for their persistence on medical devices[15,16], and in general in the nosocomial environment[17], however, the actual extent of the fungal persistence in the hospital is still unclear. Furthermore, the fungal biofilm displays higher resistance to drugs[18–20], with a complex and yet not totally understood mechanism[7], involving the Hsp90, a protein also responsible for cell dispersion[21], making the biofilm a system to persist in harsh environments and resist to the associated stresses.

From what briefly reported above, the biofilm forming ability represents a severe risk factor[1,22] and adds more problems to the need of a timely and appropriate therapy to reduce the mortality[23].

Altogether, *Candida* infections represent a serious problem and their ability to form biofilm seems to represent not only a medical, but also an ecological problem. In fact, the infecting cells can be present in different niches spanning from the devices to the surfaces, the air, some foods and the patients themselves. The cell circulation in the environment is an essential point to understand the complex ecology represented by the interaction of fungal cells with patients, different substrates and drugs. Furthermore, only a good ecological insight can lead to the actual possibility of "catching" these pathogens in their actual niches before the infection. In fact, once the infection occurred there are relatively few therapeutics to treat these diseases successfully, whereas environmental treatments with harsh biocidal compounds can be effective and decrease significantly the incidence and the mortality caused by these fungi.

With the above rationale, the present study has been designed around two hypotheses: *i.* the hospital and the various departments, i.e. specific environments, are key factors for the frequency of candidemias; *ii.* the ability to form biofilm has a measurable effect on the incidence of these diseases.

For this purpose, 277 strains of eleven *Candida* species have been isolated from the various departments of two Italian hospitals (Pisa and Udine) 450 km apart, identified at the species level and tested for biofilm formation.

Results

Distribution of the studied characters. *Species in the hospitals and wards.* Four species were isolated in both hospitals: *C. albicans, C. glabrata, C. parapsilosis* and *C. tropicalis* (Fig. 1, panel a,b). In the Pisa hospital were isolated specifically *Pichia kudriavzevii* (telomorph of *Candida krusii*), *C. rugosa* and some strains yet to be attributed to a possibly new yeast species (hereinafter referred to as *Candida. spp*) (panel a). The species isolated uniquely in Udine were *Clavispora lusitaniae* (telomorph of *Candida lusitaniae*), *C. sake, Cyberlindnera jadinii* (telomorph of *Candida utilis*) and *Meyerozyma guilliermondii* (telomorph of *Candida guilliermondii*) (panel b). The four most represented species present in both hospitals accounted for 86.36% and 96.81% in Pisa and Udine hospitals, respectively. As expected, *C. albicans* was the most frequently isolated species (panel c), absent only from the Rehabilitation department of Pisa (panel d). The second most frequent species was *C. parapsilosis,* followed by *C. glabrata* and *C. tropicalis*. These data are in good agreement with the current literature[24,25]. The departments with the highest incidences were the specialized medicine, Surgery and ICU in Pisa, while general medicine in Udine was by far the ward with more isolations, followed by specialized medicine, Surgery and ICU (panel d,e). Merging those data produced a picture very similar to that described for Udine from which came 188 out of the 276 strains analyzed in this study (panel f).

Biofilm forming strains in hospital departments. The various departments in the two hospital under study showed different frequencies of biofilm forming strains (Fig. 2). Namely, in Pisa the Rehabilitation ward had only non-biofilm forming isolates, whereas the other four departments showed frequencies ranging from 60 to more than 80% (panel a). The situation in Udine was more variable with frequencies of biofilm forming strains ranging from 20% (Rehabilitation) to 80% (Surgery) (panel b). Merging the data from the two hospitals gave a synthetic view with biofilm forming strain frequency ranging from less than 20% (Rehabilitation) to the 75% of the Surgery departments (panel c).

Biofilm forming strains and species. The four prominent species (*C. albicans, C. glabrata, C. parapsilosis* and *C. tropicalis*) showed different levels of biofilm forming ability, whereas the lesser present species (*C. rugosa*, *P. kudriavzevii*, *C. jadinii, C. sake, C. lusitaniae* and *M. guilliermondii*) did not produce biofilm (Fig. 2, panel f). *C. albicans* biofilm forming strains (hereinafter referred to as BF, in contrast with non BF referred to as NBF) were 97.78% and 87.71% in Pisa and Udine, respectively (panels d,e). Since these figures are particularly high, the biofilm formation test was repeated with two different methods using XTT and Tetrazonium Blu as indicators, obtaining no significant differences (data not shown). *C. glabrata* showed quite different frequencies of BF strains in Pisa (25%) and in Udine (4.3%), similarly *C. tropicalis* had all BF strains in Pisa and some 55% in Udine. Finally, *C. parapsilosis* showed 35% and 57% in Udine and Pisa, respectively. In general, Pisa had 75% BF strains and Udine 63.63%.

Contingency analysis studied characters. *Species vs. hospital departments.* Different opportunistic yeast species can be distributed randomly in the various hospital departments or show specific preferences for the environment represented by the various wards. In order to test the null hypothesis, i.e. that species are distributed

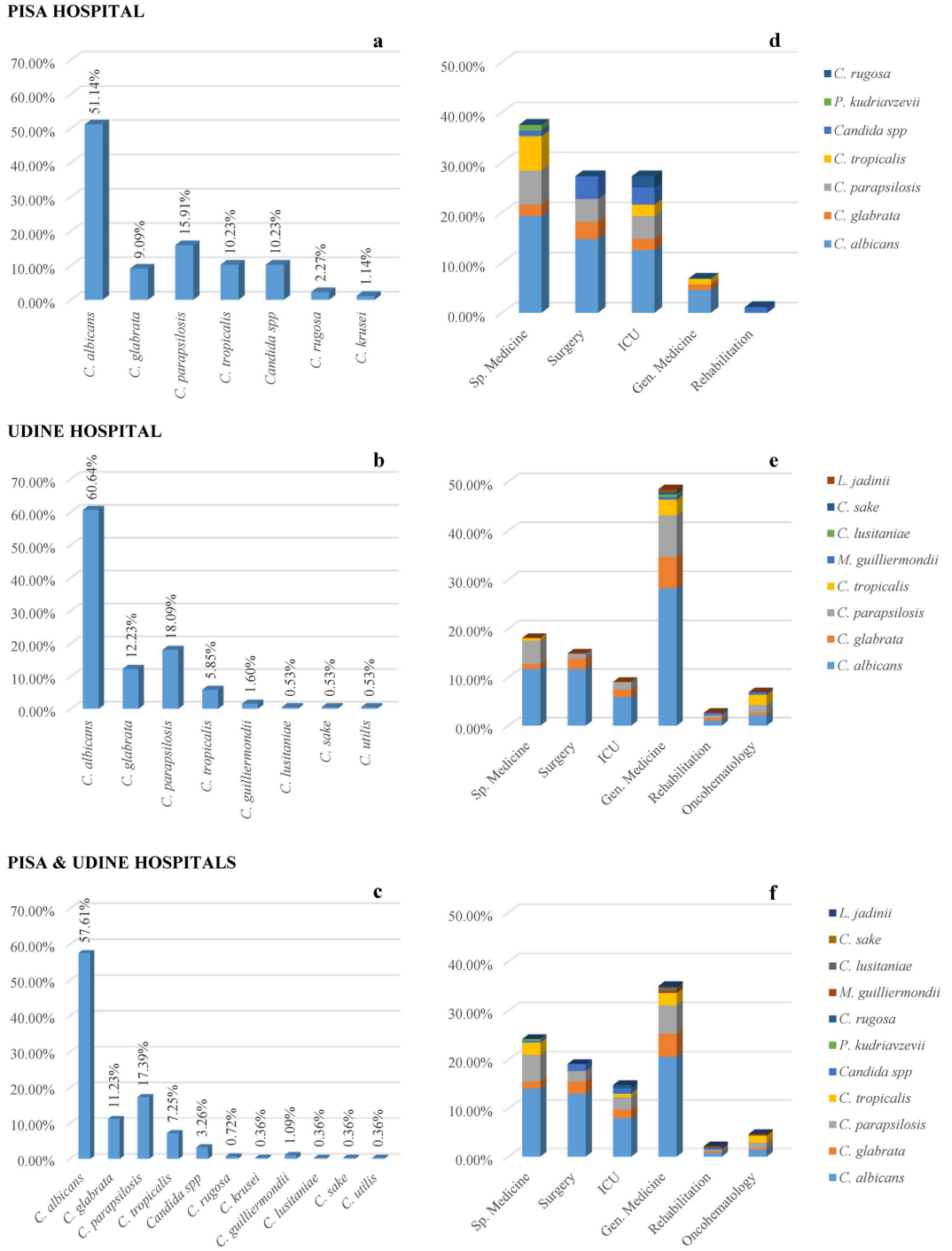


Figure 1. Species presence in the two hospitals studied and relatively frequency of isolation of the different species in the hospitals wards.

randomly within the hospital, a two-way contingency analysis was carried out, following the rationale described in Legendre & Legendre[26], as outlined in the Experimental Procedures section. This analysis is the gold standard in ecological studies when non continuous qualitative data are employed, as in our case, and allows to calculate the probability associated to the χ^2 (p-value) to accept or reject that the association between species and hospital department are independent. When two descriptors are non-independent, then their combination is indicative of some sort of specific occurrence. This test was carried out considering the two hospitals separately and then by merging all data for a joint analysis (Table 1).

The general χ^2 test for the two-way contingency table, obtained with the data of the Pisa hospital, showed no statistical significance ($p = 0.32$), indicating that the distribution of the yeast species was largely due to random effect. However, the frequency of *C. rugosa* in the ICU, and of *Candida. spp* in the Rehabilitation ward, were significantly non independent with 0.0489 and 0.0049 p-values, respectively. These data suggested that these species

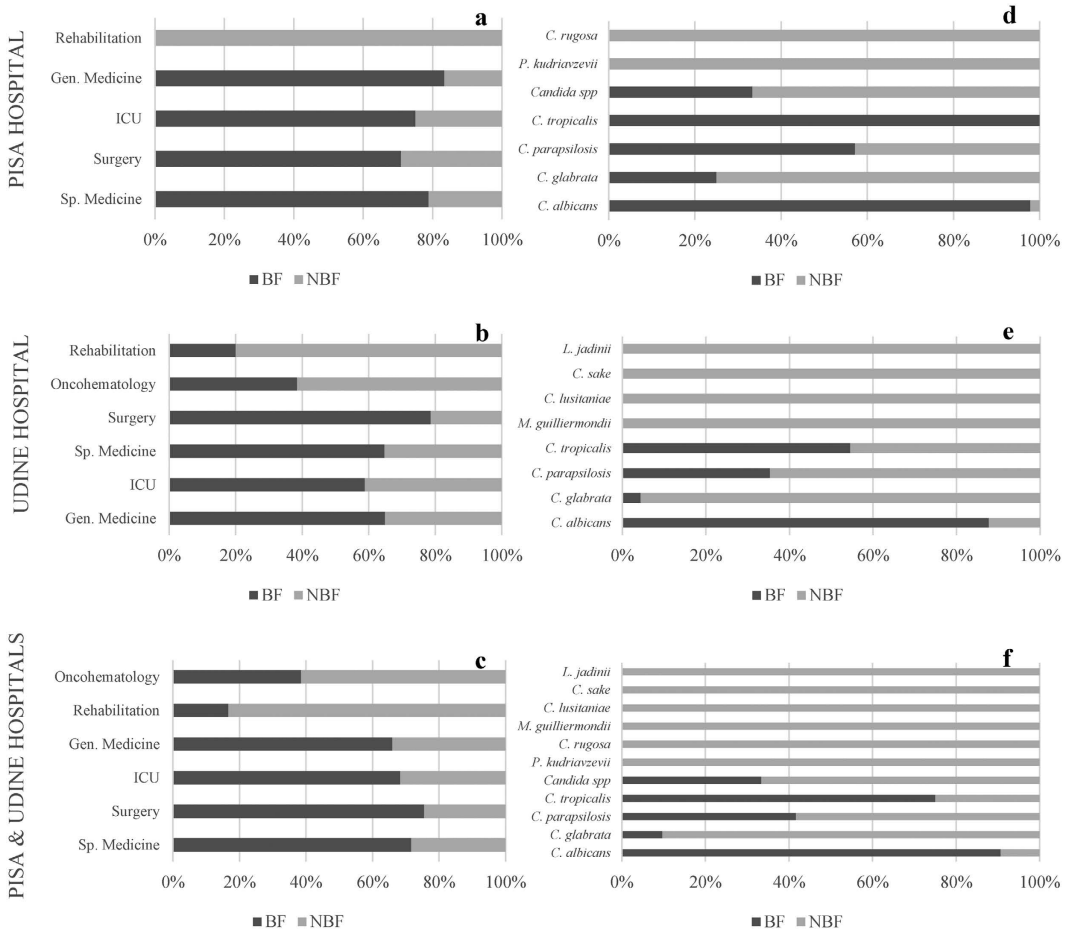


Figure 2. Relative occurrence of biofilm forming capability in different hospital departments and distribution of biofilm forming strains among species.

have a certain level of specificity for these environments. Almost significant ($p = 0.11741$) was the frequency of isolation of *C. tropicalis* in the Surgery department (panel a).

The general χ^2 test carried out with all the data from the Udine hospital gave 0.1189 p-value indicating independence between species and wards, assuming $p = 0.10$ as the minimum for statistical significance. However, the frequency of isolation of *M. guilliermondii* in Oncohematology and Rehabilitation showed high χ^2 values corresponding to $p = 0.081$ and $p = 0.0011$, respectively.

C. tropicalis in Onchohaematology showed a significantly non-independence with $p = 0.0002$. Interestingly, three species (*C. albicans, C. parapsilosis, C. tropicalis*) in the Surgery ward displayed elevated χ^2 values with p values ranging from 0.17 to 0.22, indicating that their frequency of isolation cannot be entirely ascribed to a randomness (panel b).

Merging the data of the two hospitals, the general χ^2 test was significant with $p = 0.00819$, indicating that the hypotheses of random distribution of the species in the departments of the two hospitals can be rejected. In fact, several species-department combinations showed high and significant χ^2 values, such as *C. tropicalis* in Surgery and Oncohaematology and *M. guilliermondii* in Rehabilitation and Oncohaematology (panel c). The difference between the χ^2 tests of the two hospitals considered separately and then jointly can be ascribed to the fact that merging all data together produces larger numbers and a consequently more solid statistics. Interestingly, expectable random effects explain the frequency of isolation of the various species (null hypothesis) only in the specialized medicine departments. *C. tropicalis* was overrepresented in Oncohaematology and underrepresented in the Surgery department ($p = 0.0504$ and 0.0016 respectively). *M. guilliermondii* was isolated with a frequency higher than expected in Rehabilitation and Oncohaematology ($p = 0.0002$ and 0.022 respectively), although the absolute frequencies are low, that is typical for this rather ubiquitous species found on fruit[9].

Most of the cases in which the frequency of isolation were different than expected (null hypothesis rejected) interested mostly the less representative species which account for ca. 13% in Pisa and 3% in Udine Hospitals. Based on this observation, the cases of statistically supported over- and under-isolation represent only few cases, maybe of little epidemiological importance. Yet these preferences are worth better insight, because these few cases could underline yet to be described phenomena of preferential environmental (nosocomial) colonization.

Biofilm forming strains vs. hospital departments. The question on whether the biofilm forming strains can dwell preferentially in some hospital departments was addressed with the same contingency approach as above

a) Pisa	Sp. Med.	Surgery	ICU	Gen. Med.	Rehabilit.		d) Pisa	BF	NBF	g) Pisa	BF	NBF
C. albicans	0.98	0.84	0.72	0.59	0.47		Sp. Medicine	0.80	0.66	*C. albicans*	0.39	0.13
C. glabrata	0.56	0.58	0.90	0.54	0.76		Surgery	0.81	0.68	*C. glabrata*	0.15	0.01
C. parapsilosis	0.74	0.93	0.93	0.33	0.69		ICU	1.00	1.00	*C. parapsilosis*	0.39	0.13
C. tropicalis	0.15	0.12	0.77	0.62	0.75		Gen. Medicine	0.81	0.68	*C. tropicalis*	0.22	0.03
Candida spp	0.20	0.32	0.73	0.43	<0.01		Rehabilitation	0.39	0.13	*Candida spp*	0.08	0.01
P. kudriavzevii	0.31	0.60	0.60	0.79	0.92					*P. kudriavzevii*	1.00	0.32
C. rugosa	0.39	0.46	0.05	0.71	0.88					*C. rugosa*	1.00	0.16
b) Udine	**Gen. Med.**	**ICU**	**Sp. Med.**	**Surgery**	**Oncohemat.**	**Rehabilit.**	**e) Udine**	**BF**	**NBF**	**h) Udine**	**BF**	**NBF**
C. albicans	0.77	0.83	0.76	0.22	0.17	0.55	Gen. Medicine	0.85	0.81	*C. albicans*	<0.01	<0.01
C. glabrata	0.79	0.52	0.29	0.76	0.64	0.62	ICU	0.82	0.76	*C. glabrata*	<0.01	<0.01
C. parapsilosis	0.91	0.97	0.25	0.17	0.67	0.92	Sp. Medicine	0.92	0.89	*C. parapsilosis*	0.04	0.01
C. tropicalis	0.77	0.32	0.48	0.20	<0.01	0.59	Surgery	0.31	0.18	*C. tropicalis*	0.72	0.63
M. guilliermondii	0.71	0.60	0.46	0.50	0.08	<0.01	Oncohematology	0.26	0.14	*M. guilliermondii*	0.17	0.07
C. lusitaniae	0.46	0.76	0.67	0.70	0.79	0.87	Rehabilitation	0.22	0.11	*C. lusitaniae*	0.43	0.30
C. sake	0.46	0.76	0.67	0.70	0.79	0.87				*C. sake*	0.43	0.30
L. jadinii	0.46	0.76	0.67	0.70	0.79	0.87				*L. jadinii*	0.43	0.30
c) Udine & Pisa	**Sp. Med.**	**Surgery**	**ICU**	**Gen. Med.**	**Rehabilit.**	**Oncohemat.**	**f) Udine & Pisa**	**BF**	**NBF**	**j) Udine & Pisa**	**BF**	**NBF**
C. albicans	0.96	0.33	0.73	0.90	0.43	0.20	Sp. Medicine	0.65	0.52	*C. albicans*	<0.01	<0.01
C. glabrata	0.20	0.66	0.85	0.52	0.69	0.71	Surgery	0.46	0.29	*C. glabrata*	<0.01	<0.01
C. parapsilosis	0.32	0.29	0.97	0.84	0.97	0.62	ICU	0.93	0.90	*C. parapsilosis*	0.03	<0.01
C. tropicalis	0.33	0.05	0.58	1.00	0.51	<0.01	Gen. Medicine	0.89	0.84	*C. tropicalis*	0.67	0.54
Candida spp	0.43	0.08	0.15	0.08	0.07	0.52	Rehabilitation	0.13	0.03	*Candida spp*	0.22	0.08
P. kudriavzevii	0.12	0.66	0.70	0.55	0.88	0.83	Oncohematology	0.21	0.07	*P. kudriavzevii*	0.41	0.24
C. rugosa	0.49	0.54	<0.01	0.40	0.84	0.76				*C. rugosa*	0.25	0.10
M. guilliermondii	0.39	0.45	0.51	0.96	<0.01	0.02				*M. guilliermondii*	0.16	0.04
C. lusitaniae	0.62	0.66	0.70	0.27	0.88	0.83				*C. lusitaniae*	0.41	0.24
C. sake	0.62	0.66	0.70	0.27	0.88	0.83				*C. sake*	0.41	0.24
L. jadinii	0.62	0.66	0.70	0.27	0.88	0.83				*L. jadinii*	0.41	0.24

Table 1. Probabilities associated with the χ^2 statistics calculated on contingency tables. Legend. Data are probability values associated with the χ^2 statistics and assess the hypothesis that the relationship between the two descriptors is random. When the probability value is low the null hypothesis of independence is rejected and therefore the combination between the two descriptors is not considered random but caused by some phenomenon. *p* values below 0.10 are reported in boldface; all data have been rounded to the second decimal digit. Sp. Med: Specialized Medicine; ICU: Intensive Care Unit; Gen. Med.: General Medicine; Rehabilit: Rehabilitation; Oncohaemat.: Oncohaematology; BF: Biofilm Forming; NBF: Non Biofilm Forming.

(Table 1). The general χ^2 test for the two-way contingency table, obtained with the data of the Pisa hospital, showed no statistical significance ($p = 0.44$), once again indicating that the combination of biofilm forming strains and hospital departments were largely due the randomness. In fact, no combination showed *p* values below 0.10 (panel d). On the contrary, in the Udine hospital, the χ^2 test was associated with $p = 0.0607$, indicating that the frequencies of biofilm forming strains and the various departments of isolation cannot be considered totally independent. Although no combination showed significant *p* values below 0.10, some weak signal of non-independence existed for those departments with less biofilm to isolates as in Rehabilitation and Oncohaematology of the Udine hospital ($p = 0.11$ and 0.14, respectively) (panel e).

Merging the two datasets led to a two-way contingency table for which the hypothesis of independence cannot be rejected ($p = 0.14$). In general, these tests indicated the frequency of isolation of biofilm forming strains is based on the randomness, with some significant non independence for the combinations of the Udine hospital presented above (panel f).

Biofilm forming strains vs. species. The χ^2 analysis of two-way contingency tables between the frequency of biofilm forming strains and species allowed to reject the null hypothesis with $p = 3 \times 10^{-8}$, 8.7×10^{-6} and 7.2×10^{-22} for Pisa, Udine and merged data, respectively. This extremely high statistical significance indicates that the frequency of isolation of the single species is influenced by their ability to form biofilm. More specifically, in the Pisa hospital, the random effect could be excluded for the BF strains of *Candida spp.* and for the NBF isolates of *C. glabrata, C. tropicalis* and *Candida. spp*. Particularly interesting the fact that *C. glabrata* had three times more NBF than expected on the basis of a random effect. In Udine, BF and NBF strain frequencies were significantly associated with *C. albicans, C. parapsilosis* and *C. glabrata*. Namely, *C. albicans* had 28% more BF than expected, while *C. glabrata and C. parapsilosis* had respectively 14 times more and 50% less BF strains than expected. The situation

of both hospitals considered together was substantially similar to that of Udine, given the preponderance of the strains deriving from that hospital.

Modelling the biofilm vs. isolation frequencies. The biofilm formation frequency was compared with the frequency of isolation of the single species because the combination of these two characters were shown to be non-random with extremely high statistical significance. These two descriptors had 98.51% and 97.37% Pearson correlation (R) in Pisa and Udine hospitals, respectively. This was confirmed by the linear correlation analysis with R^2 of 97.04%, and 94.8% for Pisa and Udine, but with two totally different correlation equations. In fact, in Udine, the correlation curve could be described by the equation

$$\textit{Formula 3} \quad IF = 0.942\ BF + 0.0419$$

in which IF and BF indicate, respectively, the isolation frequency of single species and the biofilm formation frequency. Formula 3 specifies that the frequency of isolation of each species is 94.2% of the frequency of BF strains of the same species.

In Udine, the correlation analysis produced the same R^2 as in Pisa, but a different equation:

$$\textit{Formula 4} \quad IF = 1.081\ \ BF + 0.0392$$

This equation indicates that in the Udine hospital the frequency of isolation of each yeast species is 108% of its frequency of BF strains. This observation indicates that indeed the ability to form biofilm plays a key role in the presence of the species in the hospital environment, but with different dynamics in Pisa and Udine. It was therefore not surprising that the correlation analysis of the Pisa and Udine data merged together produced an intermediate situation, described by Formula 5 with a lower R^2 (60.94%) than those obtained with the two hospitals analyzed separately.

$$\textit{Formula 5} \quad IF = 0.4005\ \ BF = 0.0002$$

The differences between the two hospitals can be pin-pointed by comparing the IF/BF ratio of the single species. *C. albicans* showed 103% and 114% IF/BF ratios in Pisa and Udine respectively. This means that, given a frequency of BF *C. albicans* strains, the probability of isolating this species in Pisa is slightly lower than in Udine. *C. parapsilosis* had 175% and 283% IF/BF ratios in Pisa and Udine, showing that the biofilm forming ability influences the frequency of isolation more in Udine than in Pisa. A similar situation was observed for *C. tropicalis* with 100% and 183% IF/BF ratios in Pisa and Udine. Finally, *C. glabrata* showed 400% and 2300% IF/BF ratios in the two hospitals, making it the species for which the biofilm formation ability has the maximum effect on the frequency of isolation.

Discussion

The distribution of opportunistic species of the genus *Candida* presented in this work is largely in accordance with the current literature. In fact, the number of isolates of *C. albicans* ranged from 51–61% in Pisa and Udine, showing figures similar to those previously reported[5,25]. *C. tropicalis* and *C. parapsilosis* showed 12.5% average frequency and *C glabrata* 10.6% with a significant difference from the 20% of the formers and the 5% of the latter recently reported in Brazil[7]. The seven species isolated in only one of the two hospitals showed low frequencies around 1%, again in accordance with most of the current literature[27]. These differences can be justified by the different geographic places where the studies were carried out, and by a general increase of NCAC species vs *C. albicans*[14], indicating that epidemiological data differ significantly over the time and the geography. The question on how the geography and the hospital ward influence these species frequencies is difficult to address, due to a literature concentrating either on aggregated data or on specific wards, with very few papers reporting the distribution within the hospital. This suggests that more detailed reports will be necessary in future to track these aspects.

Candida infections have an endogenous origin based on the growth of the cells already dwelling on or in the body of the patient as a commensal, while the exogenous origin derives from the surrounding environment[28]. Early studies on these aspects suggested that different areas of the hospital can be more interested to one of the two types of *Candida* infection[29]. Exogenous nosocomial infections can be triggered by the ability of these fungi to persist in the environment, by forming biofilm. Kramer and colleagues[30] reported ca 4 months persistance for *C. albicans* and *C. glabrata*, whereas these figures dropped to a few days according to an older study[31]. The persistence of *C. parapsilosis* was estimated on a couple of weeks by the above two studies. There is currently little if any information on whether the cells move as pieces of biofilm (sessile cells) or as planktonic cells liberated during the maturation of the biofilm and in their way to colonize another surface, initiating a new biofilm plaque[32]. In both cases, dispersed cells have showed enhanced adherence and produce a more robust biofilm than planktonic cells not deriving from a biofilm[33]. The biofilm architecture and resistance varies among species[34–37] and sometimes with the genetic setting[38], making any generalization quite difficult. The whole situations seems to be further complicated by the diversity of resistance mechanisms in young and mature biofilms, whereas the former is relatively well elucidated and based on multidrug pumps, the latter seems based on a dormancy mechanism leading to the production of "persister" cells[5,24,25,39]. The fungal biofilm displays higher resistance to drugs[18–20], with a complex and yet not totally understood mechanism[7], involving the Hsp90, a protein also responsible for cell dispersion[21]. The double function of this protein suggests that the complex regulation of the biofilm formation[40] is responsible for dissemination and resistance: two key factors for the success of these cells in the hospital environment.

The Biofilm forming ability has been suggested as one of the major risk factors for mortality due to *C. albicans*[1] and in general to the other yeast species analyzed in this article[7]. However, it seems that, in some settings, the biofilm forming ability does not increase the mortality nor the probability of getting catheter related candidemia[16].

The study of the factors facilitating the formation of biofilm is an important and little investigated aspect[22] to understand whether the biofilm formation is triggered by the environmental conditions. In this paper, we have considered different hospital departments to test if they have an influence in the biofilm formation. The contingency analysis has shown that there is no differential effect of the various departments in the presence of biofilm forming strains. Some instances, such as the Oncohematology of Udine, showed less biofilm than expectable. The statistical significance of these variations is relatively good when considering the aggregated data of the two hospitals ($p = 0.031$), whereas there is not significance when considering the disaggregated data of Udine alone. This suggests that we have only hints of an effect exerted by some departments, but higher number of strains for each hospital will be necessary for a careful determination of the statistical significance.

In our analytical conditions, *C. albicans, C. tropicalis, C. parapsilosis* and *C. glabrata* formed biofilm in 90%, 75%, 42% and 10% of the isolates with 67% among all the studied strains. These figures are in agreement with a recent study carried out in Scotland, in which the biofilm formation was reported as quartiles of the spectrophotometric quantification after Crystal Violet staining[1]. The high and intermediate producers accounted for 67%, 59 and 100% of the *C. albicans, C. tropicalis* and *C. parapsilosis* strains, whereas in both cases *C. glabrata* produced biofilm poorly. Interestingly, the high and intermediate producers reported for these three species in this study match exactly with the 67 of biofilm formers of our investigation. The differences in the portion of biofilm formers within the various species can be due to different environmental conditions.

These high levels of BF vs NBF strains can be due to at least three causes: the environment or the niche, methods and clonality. The study of the environmental effect is of great importance in these infections, very often caused by exogenous environmental contaminations. The other two factors, briefly outlined below, can seriously interfere with the effective determination of environmental effects on biofilm formation. The Crystal Violet, the XTT and the SYTO9 methods showed correlations ranging from 0.8 to 0.4, meaning poor R^2 regression values ranging from 0.64 to 0.16[1]. This fact indicates that the three methods yield rather independent measures of biofilm formation and the results obtained with different methods can diverge significantly. The clonality can account for the same strain being isolated repeatedly as different isolates. This problem has not an easy solution, because very accurate analyses are necessary to rule out the hypothesis of two strains being identical. Even the current barcode marker ITS[41] is scarcely effective differentiating isolates and combinations with more marker genes can be necessary to address the clonality problem effectively[9,42]. Until more light will not be shaded on this issue, we can only state that the isolates of these studies are very frequently able to form biofilm, regardless of their genetic relationships.

In the two hospital studied, the ability to form biofilm was directly correlated with the different frequency of the various isolated species according to solid linear correlations. The regression between species frequencies and biofilm formation studied separately in the two hospitals had better R^2 (0.97 and 0.95 for Pisa and Udine, respectively) than when all data were merged together ($R^2 = 0.6$).

This evidence indicates that the dynamics governing yeasts biofilm presence in the two hospitals considered are quite different. Although this situation cannot be generalized on the basis of a single study, care should be taken in epidemiological studies to analyze the data pertinent to the specific place from which they derive by condensing data because merging data from different environments (hospitals, cities etc.) might be induce serious bias.

Biofilm formation is considered of paramount importance in medicine and in a number of environmental applications. Altogether, this study has demonstrated that, in the two hospitals analyzed, the biofilm is the major factor triggering the persistence of the yeast species in these environments. A number of questions are still open, as the problem of clonality and the definition of a comprehensive working model to explain the role played by biofilm in persistence, resistance and spreading of the cells in the four most common opportunistic pathogens of the genus *Candida.*

The finding that biofilm formation can be an important factor to favor the presence of microbial cells in harsh environments further improves its general and applied biological importance.

Methods

Strains and growth condition. 277 strains belonging to opportunistic species of the *Candida* genus, isolated in a clinical (medical) environment, were employed in this study (Tables S1 and S2, Supplementary materials). All strains were isolated from patient bloodcultures, with the exception of *C tropicalis* 6184a and 6184b isolated from peritoneal fluid, *C. parapsilosis* 6551 from pharyngeal swab and *C. albicans* 8158 and 8158/C from vascular prosthesis. Isolates are kept frozen at −80 °C in 17% glycerol. Short term storage was carried out on YEPDA (YEPD added with 1.7% agarose) at 4 °C. Strains were grown in YEPD (Yeast Extract 1%, Peptone 1%, dextrose 1% - all products from Biolife- http://www.biolifeitaliana.it/) at 37 °C with 150 rpm shaking.

Molecular analysis ITS and bioinformatics tools. Genomic DNA was extracted as indicated by Cardinali *et al.*[43]. ITS1, 5.8S, ITS2 rDNA genes were amplified with FIREPol® Taq DNA Polymerase (Solis BioDyne, Estonia), using ITS1 (5′-TCCGTAGGTGAACCTGCGG) - ITS4 (TCCTCCGCTTATTGATATGC) primers.

The amplification protocol was carried out as follows. Initial denaturation at 95 °C for 4 min, 35 amplification cycles (94 °C for 1 min, 53 °C for 1 min and 72 °C for 1 min) and final extension at 72 °C for 10 min. Amplicons were purified with the GFX PCR DNA purification kit (GE Healthcare) and subject to electrophoresis on 1.5% agarose gel (Gellyphor, EuroClone, Italy). Amplicons were sequenced in both directions with ABI PRISM technology by MACROGEN (www.macrogen.com) with the same primers used for the generation of the amplicons. Consensus sequences for each strain and trimming of the ends with low sequencing quality were carried out with Geneious R6 (v. 6.17, Biomatters, Auckland, New Zealand, www.geneious.com). ITS-based species identification was carried out with BLAST search[44] in GenBank (www.ncbi.nlm.nih.gov/genbank/) and refined with specialized databases, RefSeq[45] and ISHAM-ITS reference database (ref. 46).

Biofilm protocol. Biofilm activity was assessed with an XTT method[47] and using Resazurin with slight modification on the XTT protocol. Briefly, each strain was grown over night in bottles containing YEPD (Yeast Extract 1%, Peptone 1%, Dextrose 2% - Difco Laboratories, USA) medium, at 30 °C in an orbital shaker at 150–180 rpm and then harvested and centrifuged at 3,000 × *g* for 5 minutes at 4 °C. The supernatant was removed and the pellet was washed twice with PBS. Washed cells were then resuspended in RPMI-1640 medium (Sigma Aldrich), previously warmed at 37 °C, in order to obtain a final density of 1.0×10^6 cells/mL. 100 μL of this standardized cell suspension were seeded in each selected well of 96-well microtiter plate; the wells on column 12 remained unseeded, in order to act as negative background control for the subsequent steps.

The microtiter plate was closed, sealed and incubated for 24 h at 37 °C. After biofilm formation, the medium in each well was removed carefully with a multi-channel pipette, taking care of not disrupting the biofilm; each well was subsequently washed three times with PBS. After each washing step, the plate was drained in an inverted position by blotting with paper towels.

Using a multichannel pipette, 100 μl of fresh 0.001 mg/mL resazurin solution was added to each well of the drained plated, included the negative control wells, to assess the biofilm formation. After 1 h incubation at 37 °C, the plate was visually inspected to highlight the presence of pink color gradient, resulting from the reduction of the blue dye resazurin to the pink dye resorufin by living biofilm-forming cells. Both systems produced quite similar results. Strains were considered able to form biofilm when a visible color change was detectable.

Statistical data analysis. *Contingency analysis and chi squared test*. Contingency analysis and chi squared test is the gold standard approach in ecology for the treatment of qualitative non-continuous data, as those treated in this paper. Contingency analysis was carried out according to Legendre & Legendre[26], as detailed in Supplementary materials and depicted in the four parts of Table S3.

References

1. Rajendran, R. *et al.* Biofilm formation is a risk factor for mortality in patients with *Candida albicans* bloodstream infection—Scotland, 2012–2013. *Clinical Microbiology and Infection* **22.1,** 87–93 (2015).
2. Wisplinghoff, H. *et al.* Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study. *Clinical infectious diseases* **39,** 309–317 (2004).
3. Horn, D. L. *et al.* Epidemiology and outcomes of candidemia in 2019 patients: data from the prospective antifungal therapy alliance registry. *Clinical infectious diseases* **48,** 1695–1703 (2009).
4. Sobel, J. D. The emergence of non-albicans *Candida* species as causes of invasive candidiasis and candidemia. *Current infectious disease reports* **8,** 427–433 (2006).
5. Dimopoulos, G., Ntziora, F., Rachiotis, G., Armaganidis, A. & Falagas, M. E. *Candida albicans* versus non-albicans intensive care unit-acquired bloodstream infections: differences in risk factors and outcome. *Anesthesia & Analgesia* **106,** 523–529 (2008).
6. Miceli, M. H., Díaz, J. A. & Lee, S. A. Emerging opportunistic yeast infections. *The Lancet infectious diseases* **11,** 142–151 (2011).
7. Sardi, J., Scorzoni, L., Bernardi, T., Fusco-Almeida, A. & Giannini, M. M. *Candida* species: current epidemiology, pathogenicity, biofilm formation, natural antifungal products and new therapeutic options. *Journal of medical microbiology* **62,** 10–24 (2013).
8. Merseguel, K. B. *et al.* Genetic diversity of medically important and emerging *Candida* species causing invasive infection. *BMC infectious diseases* **15,** 57, doi: 10.1186/s12879-015-0793-3 (2015).
9. Chen, S. C. A. *et al.* Candidaemia with uncommon *Candida* species: predisposing factors, outcome, antifungal susceptibility, and implications for management. *Clinical Microbiology and Infection* **15,** 662–669, doi: 10.1111/j.1469-0691.2009.02821.x (2009).
10. Trofa, D., Gácser, A. & Nosanchuk, J. D. *Candida parapsilosis*, an emerging fungal pathogen. *Clinical microbiology reviews* **21,** 606–625 (2008).
11. Gácser, A., Schäfer, W., Nosanchuk, J. S., Salomon, S. & Nosanchuk, J. D. Virulence of *Candida parapsilosis, Candida orthopsilosis*, and *Candida metapsilosis* in reconstituted human tissue models. *Fungal Genetics and Biology* **44,** 1336–1341, doi: 10.1016/j.fgb.2007.02.002 (2007).
12. Pfaller, M. *et al. Candida guilliermondii*, an opportunistic fungal pathogen with decreased susceptibility to fluconazole: geographic and temporal trends from the ARTEMIS DISK antifungal surveillance program. *Journal of clinical microbiology* **44,** 3551–3556 (2006).
13. Papon, N., Courdavault, V., Clastre, M. & Bennett, R. J. Emerging and emerged pathogenic *Candida* species: beyond the *Candida albicans* paradigm. *PLoS Pathog* **9,** e1003550 (2013).
14. Kojic, E. M. & Darouiche, R. O. *Candida* infections of medical devices. *Clinical microbiology reviews* **17,** 255–267 (2004).
15. Guembe, M. *et al.* Is biofilm production a predictor of catheter-related candidemia? *Medical Mycology* **52,** 407–410 (2014).
16. Weinstein, R. A. & Hota, B. Contamination, disinfection, and cross-colonization: are hospital surfaces reservoirs for nosocomial infection? *Clinical infectious diseases* **39,** 1182–1189 (2004).
17. Ramage, G., Rajendran, R., Sherry, L. & Williams, C. Fungal biofilm resistance. *International journal of microbiology* 2012, doi: 10.1155/2012/528521 (2012).
18. d'Enfert, C. Biofilms and their role in the resistance of pathogenic *Candida* to antifungal agents. *Current drug targets* **7,** 465–670 (2006).
19. Mukherjee, P. K. & Chandra, J. *Candida* biofilm resistance. *Drug Resistance Updates* **7,** 301–309, doi: 10.1016/j.drup.2004.09.002 (2004).
20. Robbins, N. *et al.* Hsp90 governs dispersion and drug resistance of fungal biofilms. *PLoS pathogens* **7,** e1002257 (2011).
21. Tumbarello, M. *et al.* Risk factors and outcomes of candidemia caused by biofilm-forming isolates in a tertiary care hospital. *PloS one* **7,** e33705 (2012).
22. Bassetti, M., Molinari, M., Mussap, M., Viscoli, C. & Righi, E. Candidaemia in internal medicine departments: the burden of a rising problem. *Clinical Microbiology and Infection* **19,** E281–E284 (2013).
23. Leroy, O. *et al.* Epidemiology, management, and risk factors for death of invasive *Candida* infections in critical care: a multicenter, prospective, observational study in France (2005–2006). *Critical care medicine* **37,** 1612–1618 (2009).
24. Weinberger, M. *et al.* Characteristics of candidaemia with *Candida albicans* compared with non-albicans *Candida* species and predictors of mortality. *Journal of Hospital Infection* **61,** 146–154 (2005).
25. Legendre, P. & Legendre, L. F. Numerical ecology. *Second English Edition* Ch. 6, 207–245 (Elsevier, 1998).
26. Corte, L. *et al.* Phenotypic and molecular diversity of *Meyerozyma guilliermondii* strains isolated from food and other environmental niches, hints for an incipient speciation. *Food Microbiology* **48,** 206–215 (2015).
27. Lortholary, O. *et al.* Worrisome trends in incidence and mortality of candidemia in intensive care units (Paris area, 2002–2010). *Intensive care medicine* **40,** 1303–1312 (2014).

28. Gallè, F., Catania, M. & Liguori, G. Nosocomial *Candida* infections: epidemiology of candidaemia. *Journal of preventive medicine and hygiene* **47,** 119–126 (2015).
29. Fridkin, S. K. & Jarvis, W. R. Epidemiology of nosocomial fungal infections. *Clinical microbiology reviews* **9,** 499–511 (1996).
30. Kramer, A., Schwebke, I. & Kampf, G. How long do nosocomial pathogens persist on inanimate surfaces? A systematic review. *BMC infectious diseases* **6,** 130, doi: 10.1186/1471-2334-6-130 (2006).
31. VandenBergh, M. F. Q., Verweij, P. E. & Voss, A. Epidemiology of nosocomial fungal infections: invasive aspergillosis and the environment. *Diagnostic Microbiology and Infectious Disease* **34,** 221–227, doi: 10.1016/S0732-8893(99)00026-7 (1999).
32. Nikolaev, Y. A. & Plakunov, V. Biofilm—"City of microbes" or an analogue of multicellular organisms? *Microbiology* **76,** 125–138 (2007).
33. Uppuluri, P. *et al.* Dispersion as an important step in the *Candida albicans* biofilm developmental cycle. *PLoS Pathog* **6,** e1000828 (2010).
34. Silva, S. *et al.* Adherence and biofilm formation of *non-Candida albicans Candida* species. *Trends in microbiology* **19,** 241–247 (2011).
35. Silva, S. *et al.* Biofilms of *non-Candida albicans Candida* species: quantification, structure and matrix composition. *Medical Mycology* **47,** 681–689 (2009).
36. Parahitiyawa, N. *et al.* Interspecies variation in *Candida* biofilm formation studied using the Calgary biofilm device. *Apmis* **114,** 298–306 (2006).
37. Hawser, S. P. & Douglas, L. J. Biofilm formation by *Candida* species on the surface of catheter materials *in vitro*. *Infection and immunity* **62,** 915–921 (1994).
38. Park, Y.-N., Daniels, K. J., Pujol, C., Srikantha, T. & Soll, D. R. *Candida albicans* forms a specialized "sexual" as well as "pathogenic" biofilm. *Eukaryotic cell* **12,** 1120–1131 (2013).
39. LaFleur, M. D., Kumamoto, C. A. & Lewis, K. *Candida albicans* biofilms produce antifungal-tolerant persister cells. *Antimicrobial agents and chemotherapy* **50,** 3839–3846 (2006).
40. Nobile, Clarissa J. *et al.* A Recently Evolved Transcriptional Network Controls Biofilm Development in *Candida albicans*. *Cell* **148,** 126–138, doi: 10.1016/j.cell.2011.10.048 (2012).
41. Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 6241 –6246, doi: 10.1073/pnas.1117018109 (2012).
42. Stielow, J. *et al.* One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes. *Persoonia-Molecular Phylogeny and Evolution of Fungi* **35,** 242–263, doi: 10.3767/003158515X689135 (2015).
43. Cardinali, G., Bolano, A. & Martini, A. A DNA extraction and purification method for several yeast genera. *Annals of Microbiology* **51,** 121–130 (2001).
44. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215,** 403–410 (1990).
45. Schoch, C. L. *et al.* Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. *Database- the journal of biological datrabase and curation* **2014,** doi: 10.1093/database/bau061 (2014).
46. Irinyi, L. *et al.* 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* **53,** 313–337 (2015).
47. Pierce, C. G. *et al.* A simple and reproducible 96-well plate-based method for the formation of fungal biofilms and its application to antifungal susceptibility testing. *Nature protocols* **3,** 1494–1500 (2008).

Acknowledgements

The work was partly supported by the project of National interest (PRIN) number 2010WZ2NJN_002, funded by the Italian Ministry of Research and University. LR was partly supported by a Grant of the Fondazione Cassa di Risparmio di Perugia (Italy); CC was partly supported by a PhD Grant of the Italian Ministry of Research and University. The authors thank Mrs. Cecilia Lucarelli and Mr. Emanuele Pitari for their technical work.

Author Contributions

L.C. participated in planning of the study and experimental design, carried out the experimental activities, participated in data analysis and interpretation, and drafted the manuscript. L.R. participated in the experimental activities and data analysis and interpretation. C.C. participated in the experimental activities. C.T. participated in planning of the study and experimental design, carried out the experimental activities, participated in data analysis and interpretation, and drafted the manuscript. A.L. participated in the experimental activities and data analysis and interpretation. E.S. participated in the experimental activities and data analysis and interpretation. F.M. participated in planning of the study and experimental design, carried out the experimental activities, participated in data analysis and interpretation, and drafted the manuscript. M.M. commented on the manuscript. C.S. commented on the manuscript. W.M. commented on the manuscript. C.G. participated in planning of the study, experimental design, data analysis and interpretation, and commented on the manuscript. M.B. participated in planning of the study, experimental design, data analysis and interpretation, and commented on the manuscript. All authors read and approved the final version.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: In the past 2 years, Dr. Carlo Tascini has been paid for lectures on behalf of Pfizer, Novartis, Merck Astra, Angelini, Gilead and Astellas.

How to cite this article: Corte, L. *et al.* Exploring ecological modelling to investigate factors governing the colonization success in nosocomial environment of *Candida albicans* and other pathogenic yeasts. *Sci. Rep.* **6**, 26860; doi: 10.1038/srep26860 (2016).

Paper II

Mycopathologia (2016) 181:297–303
DOI 10.1007/s11046-015-9961-7

First Case of *Trichoderma longibrachiatum* CIED (Cardiac Implantable Electronic Device)-Associated Endocarditis in a Non-immunocompromised Host: Biofilm Removal and Diagnostic Problems in the Light of the Current Literature

Carlo Tascini · Gianluigi Cardinali · Valentina Barletta · Antonello Di Paolo · Alessandro Leonildi · Giulio Zucchelli · Laura Corte · Claudia Colabella · Luca Roscini · Augusta Consorte · Maria Bruna Pasticci · Francesco Menichetti · Maria Grazia Bongiorni

Received: 29 July 2015 / Accepted: 28 October 2015 / Published online: 20 November 2015

Abstract

Background *Trichoderma* species are saprophytic filamentous fungi producing localized and invasive infections that are cause of morbidity and mortality, especially in immunocompromised patients, causing up to 53 % mortality. Non-immunocompromised patients, undergoing continuous ambulatory peritoneal dialysis, are other targets of this fungus. Current molecular diagnostic tools, based on the barcode marker ITS, fail to discriminate these fungi at the species level, further increasing the difficulty associated with these infections and their generally poor prognosis.

Case Report We report on the first case of endocarditis infection caused by *Trichoderma longibrachiatum* in a 30-year-old man. This patient underwent the implantation of an implantable cardioverter defibrillator in 2006, replaced in 2012. Two years later, the patient developed fever, treated successfully with amoxicillin followed by ciprofloxacin, but an echocardiogram showed large vegetation onto the ventricular lead. After CIED extraction, the

C. Tascini · A. Leonildi · F. Menichetti
Infectious Diseases Unit, Azienda Ospedaliera Universitaria Pisana, Via Paradisa 2, Cisanello, 56100 Pisa, Italy

G. Cardinali (✉) · L. Corte · C. Colabella · L. Roscini
Department of Pharmaceutical Sciences – Microbiology, University of Perugia, Borgo 20 Giugno 74, 06121 Perugia, Italy
e-mail: gianluigi.cardinali@unipg.it

G. Cardinali
Department of Chemistry, Biology and Biotechnology, CEMIN, Centre of Excellence on Nanostructured Innovative Materials, University of Perugia, Via Elce di Sotto 8, 06123 Perugia, Italy

V. Barletta · G. Zucchelli · M. G. Bongiorni
Cardiovascular Medicine Unit 2, Azienda Ospedaliera Universitaria Pisana, Via Paradisa 2, Cisanello, 56100 Pisa, Italy

A. Di Paolo
Division of Pharmacology, Department of Clinical and Experimental Medicine, University of Pisa, Via Roma 55, 56126 Pisa, Italy

A. Consorte
Infectious Diseases Unit, "Spirito Santo" Hospital, Pescara, Italy

M. B. Pasticci
Infectious Diseases Unit, Department of Medicine, University of Perugia, Piazzale Gambuli, 1, 06132 Perugia, Italy

patient had high-grade fever. The culturing of the catheter tip was positive only in samples deriving from sonication according to the 2014 ESCMID guidelines, whereas the simple washing failed to remove the biofilm cells from the plastic surface. Subsequent molecular (ITS sequencing) and microbiological (macromorphology) analyses showed that the vegetation was due to *T. longibrachiatum.*
Conclusions This report showed that *T. longibrachiatum* is an effective threat and that sonication is necessary for the culturing of vegetations from plastic surfaces. Limitations of the current barcode marker ITS, and the long procedures required by a multistep approach, call for the development of rapid monophasic tests.

Keywords *Trichoderma longibrachiatum* · CIED · Endocarditis · Molecular diagnosis

Introduction

The members of the fungal genus *Trichoderma* are saprophytic filamentous fungi with worldwide distribution in the soil, plant material, decaying vegetation and wood. Some species of *Trichoderma* can cause infections in humans, so far 36 cases of human infections are described in the literature [1]. Localized infections such as pulmonary mycetoma, peritonitis, sinusitis, otitis, cellulitis or brain abscess, and fatal disseminated disease, especially in immunocompromised host, were described [2]. Fungal infections by *Trichoderma* spp. normally cause morbidity and mortality, especially in immunocompromised patients: Among solid organ transplantation, nine cases of infections with seven deaths are described (seven in liver transplant recipients and one in renal and lung transplant recipients, respectively); among patients with hematological malignancies, seven cases with two deaths are described [3]. As for non-immunocompromised patients, infections were described mainly among those undergoing continuous ambulatory peritoneal dialysis: 11 patients with 7 deaths [4, 5]. Furthermore, *T. longibrachiatum* was reportedly the cause of allergic sinusitis and external malignant otitis [6, 7]. Here we report the first case of cardiac implantable electronic device (CIED)-associated endocarditis in a non-immunocompromised patient, caused by this fungus, and discuss the methodological problems inherent to its diagnosis in the light of the current literature.

Case Report

A 30-year-old man (70 kg, 1.75 m) underwent implantable cardioverter defibrillator (ICD) implantation in 2006 for ventricular tachycardia, and in 2008, a new lead was added due to a malfunction of the device. In 2012, an elective ICD replacement was complicated by a pneumothorax and a *Staphylococcus epidermidis* bacteremia, treated for 14 days with linezolid. After 2 months, the ICD was replaced again for several inappropriate shocks. In May 2014, the patient developed fever, treated successfully with amoxicillin followed by ciprofloxacin. A month later, an echocardiogram showed a large vegetation along the leads. A new blood culture revealed a methicillin-susceptible *S. epidermidis*. The patient was treated with teicoplanin and amikacin for 2 weeks and then with ciprofloxacin and rifampin for another 2 weeks. Four months before this episode of fever, the patient underwent a tattoo on the right leg, in which black ink has been prevalently used, but no sign of inflammation was subsequently noted by the patient. The patient was referred to Pisa Hospital, the Italian reference center for CIED extraction. Here, a transesophageal echocardiogram showed again a large vegetation of 2 × 2 cm on the ventricular lead. According to the Pisa Hospitals internal protocol about antibiotic therapy for transvenous lead extraction, daptomycin (6 mg/kg) therapy was started and the patient underwent an uncomplicated lead extraction procedure. Early after lead removal, the patient had high-grade fever and a PET scan showed an embolic lesion at left lung (Fig. 1) without intra-cardiac tracer accumulation. The tips of the leads were cultured with standard procedure. Furthermore, the same tips underwent sonication in sterile conditions. The container with catheter tip was vortexed for 30 s and then sonicated for 1 min at a frequency of 40 kHz and power density of 0.22 W/cm^2 in an ultrasonic bath (BactoSonic, Bandelin GmbH, Berlin, Germany). A total of 0.1 mL of the resulting sonication fluid was inoculated onto aerobic and anaerobic sheep blood agar plates, mannitol salt agar and Sabouraud agar (Biomérieux, Milan, Italy) and incubated at 37 °C for 7 days.

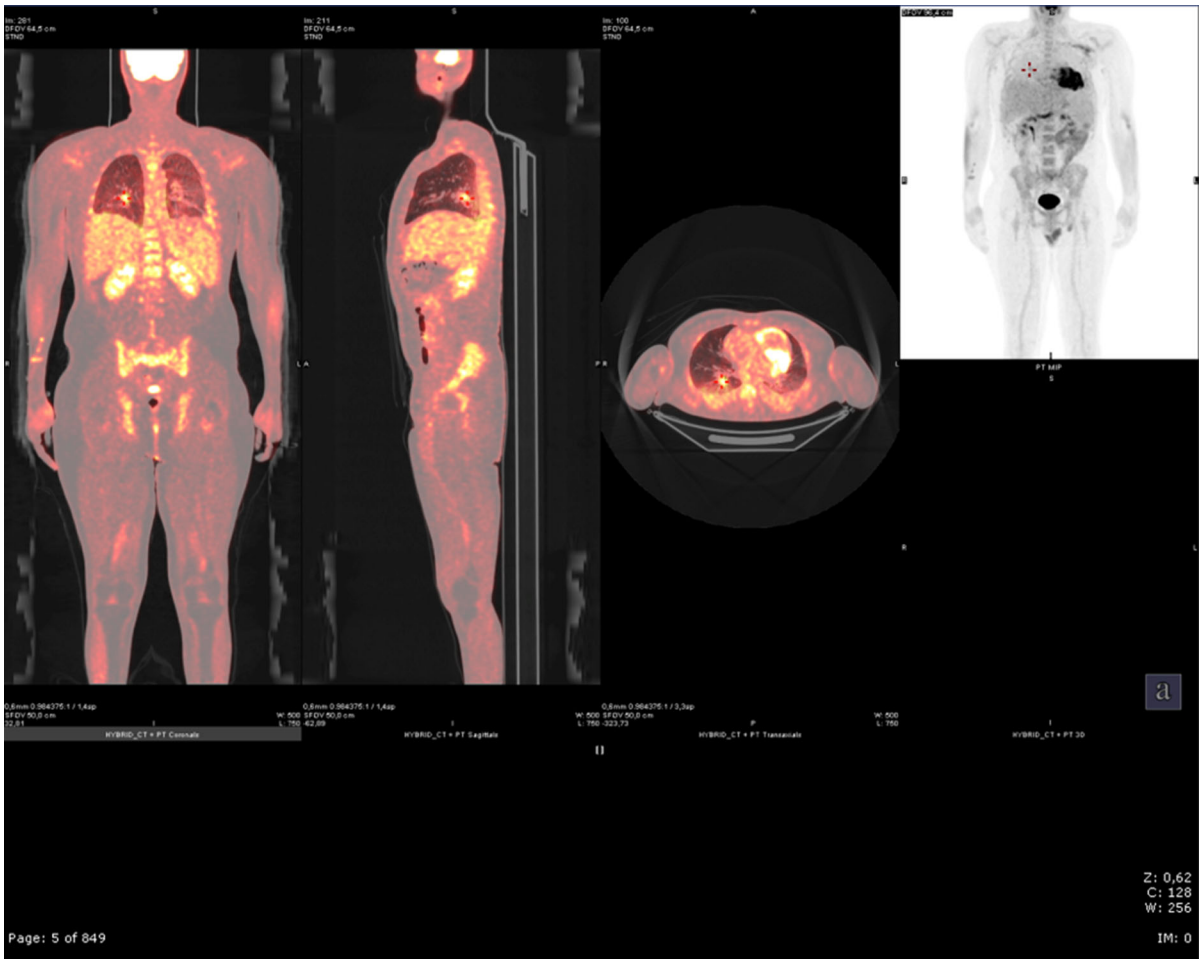


Fig. 1 PET demonstrating embolic pneumonia after ICD transvenous removal

After 48 h of culture, methicillin-susceptible *S. epidermidis* was identified. After further 24 h, a filamentous fungus colony, later identified as *Trichoderma longibrachiatum*, was visible. The latter was present only in the culture derived from the sonicated material. Using the E-test method, the following MIC values were obtained: voriconazole 0.5 µg/L, amfotericin B 2 µg/L, caspofungin 1 µg/L.

Morphological analyses and ITS1–ITS2 sequencing confirmed the identity of the fungal species. The morphological identification was carried out using the method described by Samuels et al. [8]. In all the tested conditions, the isolate displayed a surface mycelium disposed in rays. At 35 °C, the colony was covered by white conidia, slowly turning to dark green (Fig. 2).

Genomic DNA was extracted, and the internal transcribed spacer (ITS) region of the rDNA gene cluster was amplified using ITS1 and ITS4 primers. Amplicons were sequenced in both directions by Macrogen (www.macrogen.com). ITS sequences identification queries were fulfilled by BLAST search in GenBank (www.ncbi.nlm.nih.gov/genbank/) and using the ITS-based barcoding tool TrichOKEY (www.isth.info). The clean and correct ITS sequence was deposited in GenBank under the accession number KP636421.

A therapy with 200 mg bid oral voriconazole was initiated, after a loading dose of 400 mg bid for only the first day. After 3 days, the peak and through concentration of voriconazole was measured by using a commercially available kit (Chromsystems, Grafelfing, Germany) on a Waters TQD liquid chromatography–mass spectrometry instrument (Waters, Milford, MA). Surprisingly, plasma concentrations of voriconazole

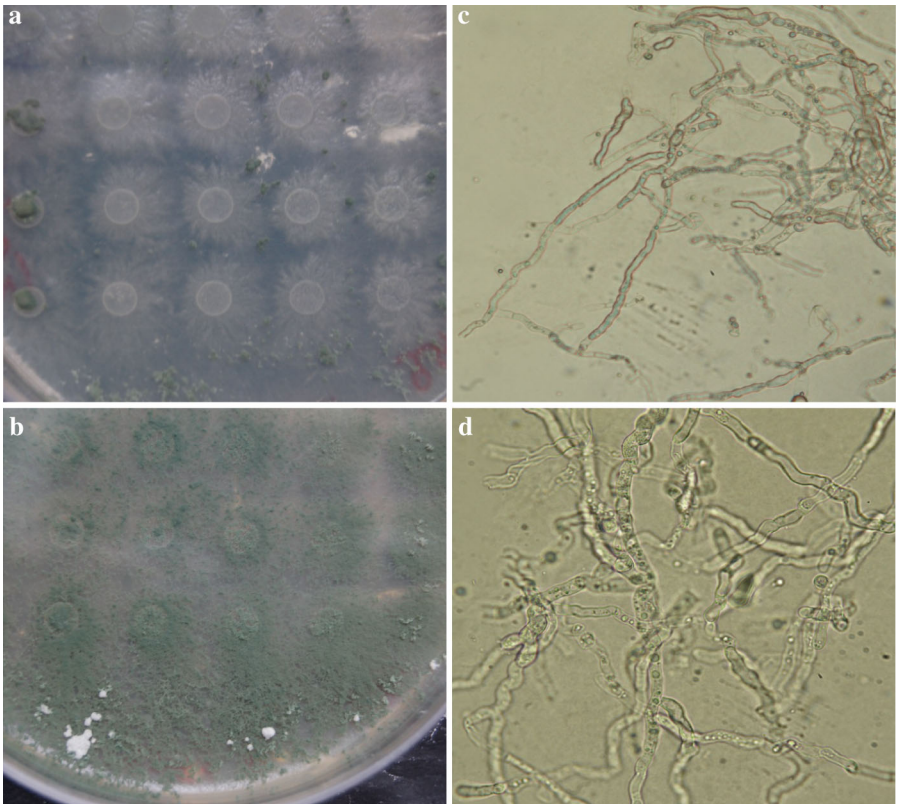


Fig. 2 Macroscopic (**a**, **b**) and light microscopic (**c**, **d**) morphology of *Trichoderma longibrachiatum*. Isolates grown on CMD and PDA are *white* in the initial phase of growth (**a**) then slowly turn to *dark green* (**b**), with conidia disposed in rays. (Color figure online)

were below the sensitivity limit (0.01 mg/L). Plasma measurement of drug concentrations was performed again, but results were unchanged, even if voriconazole was switched to i.v. administration. Therefore, we switched to liposomal amphotericin B. A subcutaneous ICD was implanted 10 days after lead extraction, and the patient was transferred to the referring hospital. Here, the amphotericin B therapy was complicated by a renal insufficiency and the drug was stopped after a total of 25 days of therapy with complete recovery of the renal function after 2 weeks from the antifungal therapy interruption.

Review of the Literature and Discussion

Eight species of the genus *Trichoderma* [*T. longibrachiatum, T. harzianum, T. koningii, T. pseudokoningii, T. orientale* (*syn. Hypocrea orientalis*) *T. viride*, *T. atroviride* and *T. citrinoviride*] have been identified as etiologic agents of infections in immunocompromised hosts [3, 9]. *T. longibrachiatum* is the main human pathogenic species within the genus and was isolated with increasing frequency in recent years [2].

Among cardiac infections, *T. longibrachiatum* was described so far only in two patients. The former underwent an ascending aorta replacement for a De Bakey's type II aortic dissection and a subsequent surgery for endocarditis of aortic conduit complicated by three consecutive embolic events (in the lower limb, brain and spleen). During surgery, several mobile vegetations attached to both proximal and distal sutures lines of the aortic conduit were found. Cultures of the removed material were positive for *Trichoderma* species. The patient received antifungal drugs in the postoperative period and was discharged without complications [10]. The latter was a man suffering from short bowel and receiving home parenteral nutrition who developed an endocarditis over catheter. A fragment of the catheter remained on

the right atrial, and after the surgical removal, the cultures were positive for *S. epidermidis, Ochrobactrum anthropi* and *T. longibrachiatum*. The extraction of the infected catheter along with antibiotic and antifungal therapy led to the complete recovery of the subject [11]. Overall, survival of *T. longibrachiatum* cases was around 47 %, including the immunocompromised patients with a generally poor prognosis for *Trichoderma* infection.

The definitive diagnosis of this fungus is difficult to achieve due to the lack of specific tools at the species-specific level. In fact, the current barcode ITS cannot separate species within the *Trichoderma–Hypocrea* complex. Namely, for our case, *T. longibrachiatum* cannot be separated from *Hypocrea orientalis* by ITS sequencing [12]. The species identification can currently benefit from a combined microbiological and molecular approach [8]. For our patient, such procedure provided the definitive identification of this saprophytic fungal organism, but required a time-consuming procedure. A polyphasic (i.e., multistep) procedure requires time and can be hampered by the presence of other species in the culture, as described in the case of the short bowel patient. This situation calls for rapid monophasic tests to determine the presence of *Trichoderma* at the species level. Since the current barcoding marker ITS [13] is not able to discriminate among species efficiently, even using two different databases, new *loci* recently identified, such as translation elongation factor 1 alpha (hereinafter reported as TEF1α) and calmodulin genes [1, 14]. The presence of *Trichoderma* spp. in mixed cultures could be easily detected using innovative NGS (next-generation sequencing) strategy by multiplex sequencing of several genes, in order to define all the species putatively causing the infection.

Potential virulence factors of *T. longibrachiatum* as an opportunistic pathogen include its ability of mycelial growth up to 40° C, hemolytic ability, toxicity to mammalian cells and the resistance to pH values ranging from 2 to 9 [15]. Moreover, it has been reported to produce extracellular proteases [16] and to display high levels of resistance to antifungal compounds including fluconazole, itraconazole and in some cases amphotericin B [17].

Patients with this infection were usually treated with amphotericin B. Other therapeutic options are available, such as voriconazole and caspofungin. In the case described in this article, voriconazole was not detected in patients serum; therefore, the patient was treated with liposomial amphotericin B. The therapy was stopped for renal toxicity and the renal function normalized in few weeks after the withdrawal. It is possible to hypothesize that voriconazole was not detected in the serum due to a very active detoxification metabolism of the patient. Plasma levels below the limit of quantitation have been already described in patients receiving voriconazole at doses ranging from 2 up to 12 mg/kg in 25 patients [18]; hence, a possible explanation for our therapeutic drug monitoring (TDM) results could be related to the wide interpatient variability. For our patient, susceptibility tests of *T. longibrachiatum* isolate were performed with the use of the E-test. The MICs for the patient's isolates were as follows: amphotericin B, 1 μg/mL, voriconazole, 0.5 μg/mL, and caspofungin, 1 μg/mL. Fluconazole was not tested. These results are in accordance with the reported cases [19]. In the literature, voriconazole was the drug of choice for this kind of infection, but also posaconazole among azoles, echinocandins and amphotericin B were considered alternative options when the MIC is similar to serum concentrations achievable with standard dose of these antifungal drugs (Kredics).

Trichoderma longibrachiatum is known to produce biofilm on different surfaces, including nylon [20]. Biofilm removal is a harsh operation carried out by prolonged vortexing, bead beating and sonication, as recommended by ESCMID [21]. Sonication has been advocated as a reference method for microbiology of prosthetic material, especially for orthopedic devices [22]. In the field of CIED microbiology, there are conflicting results; in fact, Viola and coworkers have found that traditional culture techniques are as effective as sonication in culturing microorganisms from removed CIEDs [23]. On the other hand, Oliva et al. reported that sonication is able to improve microbial detection in cardiac device infection [24]. In the case of our patient, only the hemoculture from sonication produced cells growth, suggesting that indeed, a biofilm was formed on the catheter tip. This observation confirms the findings by Oliva et al. and highlights the importance of the ESCMID guidelines on the sonication as a safer way than simple washing to start the diagnostic procedures. In fact, the latter procedure can cause underestimating the presence of biofilm-forming fungi with very few, if any, planktonic cells detached from the biofilm.

We could not find any correlation between tattooing and *T. longibrachiatum* infection, normally ascribed to *Acremonium* spp. contamination, and therefore, it should be ruled out that in this case, tattoo in patients with implanted CIED would be a risk factor for life-threatening endocarditis.

To our knowledge, this is the first case of CIED fungal endocarditis due to *T. longibrachiatum*. In conclusion, *Trichoderma* infection could be misdiagnosed as other types of hyalohyphomycosis. As demonstrated in other device-related infections, the removal of the CIED is pivotal and mandatory in order to cure this kind of difficult-to-treat infections such as fungal endocarditis. Clinical judgment has to be used in order to correctly interpret microbiological results, particularly until a gold standard will be established for *T. longibrachiatum* identification. The impossibility of a species-specific diagnosis with the current barcode marker ITS [13] calls for introducing new marker loci such as TEF1α [14, 25].

Compliance with Ethical Standards

Conflict of interest Dr. TASCINI reports personal fees from PFIZER, personal fees from MERCK, personal fees from ASTELLAS, personal fees from ASTRAZENECA, personal fees from NOVARTIS, personal fees from GILEAD, personal fees from ANGELINI, outside the submitted work.

References

1. Kredics L, Hatvani L, Manczinger L, Vágvölgyi C, Antal Z. *Trichoderma*: molecular detection of human fungal pathogens. London: Taylor & Francis Group; 2011. p. 509–62.
2. Trabelsi S, Hariga D, Khaled S. First case of *Trichoderma longibrachiatum* infection in a renal transplant recipient in Tunisia and review of the literature. Tunis Med. 2010;88(1):52–7.
3. Hatvani L, Manczinger L, Vágvölgyi C, Kredics L. *Trichoderma* as a human pathogen. In: *Trichoderma*: biology and applications; 2013. p 292–313.
4. Aroca T, Piontelli L, Cruz C. Case report: *Trichoderma longibrachiatum* infections in a pediatric patient with peritoneal dialysis. Bol Micol. 2004;19:13–7.
5. Lee HJ, Kim DW, Cho HS, Lim MH, Jung EY, Lee DW, et al. Peritonitis and intra-abdominal abscess by *Trichoderma longibrachiatum* in a patient undergoing continuous ambulatory peritoneal dialysis (CAPD). Korean J Nephrol. 2007;26(2):254–7.
6. Hennequin C, Chouaki T, Pichon J, Strunski V, Raccurt C. Otitis externa due to *Trichoderma longibrachiatum*. Eur J Clin Microbiol Infect Dis. 2000;19(8):641–2.
7. Tang P, Mohan S, Sigler L, Witterick I, Summerbell R, Campbell I, et al. Allergic fungal sinusitis associated with *Trichoderma longibrachiatum*. J Clin Microbiol. 2003;41(11):5333–6.
8. Samuels GJ, Ismaiel A, Mulaw TB, Szakacs G, Druzhinina IS, Kubicek CP, et al. The *Longibrachiatum* Clade of *Trichoderma*: a revision with new species. Fungal Divers. 2012;55(1):77–108.
9. Gautheret A, Dromer F, Bourhis JH, Andremont A. *Trichoderma pseudokoningii* as a cause of fatal infection in a bone marrow transplant recipient. Clin Infect Dis. 1995;20(4):1063–4.
10. Bustamante-Labarta MH, Caramutti V, Allende GN, Weinschelbaum E, Torino AF. Unsuspected embolic fungal endocarditis of an aortic conduit diagnosed by transesophageal echocardiography. J Am Soc Echocardiogr. 2000;13(10):953–4.
11. Rodríguez Peralta LI, Mañas Vera M, García Delgado MJ, Pérez De la Cruz AJ. Endocarditis por *Trichoderma longibrachiatum* en paciente con nutrición parenteral domiciliaria. Nutrición Hospitalaria. 2013;28(3):961-4.
12. Sandoval-Denis M, Sutton DA, Cano-Lira JF, Gene J, Fothergill AW, Wiederhold NP, et al. Phylogeny of the clinically relevant species of the emerging fungus Trichoderma and their antifungal susceptibilities. J Clin Microbiol. 2014;52(6):2112–25.
13. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci. 2012;109(16):6241–6.
14. Druzhinina IS, Komon-Zelazowska M, Kredics L, Hatvani L, Antal Z, Belayneh T, et al. Alternative reproductive strategies of Hypocrea orientalis and genetically close but clonal *Trichoderma longibrachiatum*, both capable of causing invasive mycoses of humans. Microbiology. 2008;154(Pt 11):3447–59.
15. Antal Z, Kredics L, Pakarinen J, Dóczi I, Andersson M, Salkinoja-Salonen M, et al. Comparative study of potential virulence factors in human pathogenic and saprophytic *Trichoderma longibrachiatum* strains. Acta Microbiol Immunol Hung. 2005;52(3):341–50.
16. Kredics L, Antal Z, Szekeres A, Manczinger L, Dóczi I, Kevei F, et al. Production of extracellular proteases by human pathogenic *Trichoderma longibrachiatum* strains. Acta Microbiol Immunol Hung. 2004;51(3):283–95.
17. Dóczi I, Dósa E, Varga J, Antal Z, Kredics L, Nagy E. Etest for assessing the susceptibility of filamentous fungi. Acta Microbiol Immunol Hung. 2004;51(3):271–81.
18. Miyakis S, van Hal SJ, Ray J, Marriott D. Voriconazole concentrations and outcome of invasive fungal infections. Clin Microbiol Infect. 2010;16(7):927–33.
19. Kratzer C, Tobudic S, Schmoll M, Graninger W, Georgopoulos A. In vitro activity and synergism of amphotericin B, azoles and cationic antimicrobials against the emerging pathogen *Trichoderma* spp. J Antimicrob Chemother. 2006;58(5):1058–61.
20. Cobas M, Ferreira L, Tavares T, Sanromán MA, Pazos M. Development of permeable reactive biobarrier for the removal of PAHs by *Trichoderma longibrachiatum*. Chemosphere. 2013;91(5):711–6.
21. Høiby N, Bjarnsholt T, Moser C, Bassi GL, Coenye T, Donelli G, et al. ESCMID guideline for the diagnosis and treatment of biofilm infections 2014. Clin Microbiol Infect. 2015;21(Supplement 1(0)):S1–25.

22. Trampuz A, Piper KE, Jacobson MJ, Hanssen AD, Unni KK, Osmon DR, et al. Sonication of removed hip and knee prostheses for diagnosis of infection. N Engl J Med. 2007;357(7):654–63.
23. Viola GM, Mansouri MD, Nasir N, Darouiche RO. Incubation alone is adequate as a culturing technique for cardiac rhythm management devices. J Clin Microbiol. 2009;47(12):4168–70.
24. Oliva A, Nguyen BL, Mascellino MT, D'Abramo A, Iannetta M, Ciccaglioni A, et al. Sonication of explanted cardiac implants improves microbial detection in cardiac device infections. J Clin Microbiol. 2013;51(2):496–502.
25. Stielow J, Lévesque C, Seifert K, Meyer W, Irinyi L, Smits D, et al. One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes. Pers Mol Phylogeny Evol Fungi. 2015;35:242–263.

Paper III

Food Microbiology 48 (2015) 206–215

Contents lists available at ScienceDirect

Food Microbiology

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

Phenotypic and molecular diversity of *Meyerozyma guilliermondii* strains isolated from food and other environmental niches, hints for an incipient speciation

Laura Corte [a], Raffaella di Cagno [b], Marizeth Groenewald [c], Luca Roscini [a], Claudia Colabella [a], Marco Gobbetti [b], Gianluigi Cardinali [a, d, *]

[a] Department of Pharmaceutical Sciences – Microbiology, University of Perugia, Borgo 20 Giugno 74, 06121 Perugia, Italy
[b] Department of Plant Protection and Applied Microbiology, University of Bari, via Amendola 165/a, 70126 Bari, Italy
[c] CBS-KNAW, Fungal Biodiversity Centre, Utrecht, the Netherlands
[d] CEMIN, Centre of Excellence on Nanostructured Innovative Materials, Department of Chemistry, Biology and Biotechnology, University of Perugia, via Elce di Sotto 8, I-06123 Perugia, Italy

ARTICLE INFO

Article history:
Received 3 June 2014
Received in revised form
17 November 2014
Accepted 16 December 2014
Available online 20 January 2015

Keywords:
Food environment
Fruits
Meyerozyma guilliermondii
Candida guilliermondii
Species diversity
Yeast

ABSTRACT

Meyerozyma guilliermondii is a yeast species widely isolated from several natural environments and from fruit; in medical microbiology it is known as the teleomorph of the opportunistic pathogen *Candida guilliermondii*, which causes about 2% of the human blood infections. This yeast is also promising in a variety of biotechnological applications as vitamins production and post-harvest control. The question if isolates from different sources are physiologically and genetically similar, or if the various environments induced significant differences, is crucial for the understanding of this species structure and to select strains appropriate for each application. This question was addressed using LSU and ITS sequencing for taxonomic assignment, i-SSR ($GACA_4$) for the molecular characterization and FTIR for the metabolomic fingerprint. All data showed that fruit and environmental isolates cluster separately with a general good agreement between metabolomics and molecular analysis. An additional RAPD analysis was able to discriminate strains according to the isolation position within the pineapple fruit. Although all strains are members of the *M. guilliermondii* species according to the current standards, the distribution of large variability detected suggests that some specialization occurred in the niches inhabited by this yeast and that food related strains can be differentiated from the medical isolates.

1. Introduction

Meyerozyma guilliermondii was firstly described as *Endomycopsis guilliermondii* (Wickerham and Burton, 1954), later coopted to the genus *Pichia* in 1966 as *Pichia guilliermondii* by Wickerham in 1966 (Wickerham, 1966) and has been recently renamed after a phylogenetic study by Kurtzman (Kurtzman and Suzuki, 2010). The history of this species started at the beginning of the last century, when Castellani described *Endomyces guilliermondii*, a yeast unable to sporulate, isolated from the sputum of a patient affected by chronic bronchitis (Castellani, 1912). This species was later reclassified in six different genera until 1938 when Langeron and Guerra brought it into the yeast genus *Candida* (Langerhorn and Guerra, 1938). The 1952 edition of "The yeast – a taxonomic study" reported a research carried out on six strains, three of which of human origin and one isolated from a horse (Lodder and Kreger-van Rij, 1952). The 1984 edition of "The yeast-a taxonomic study" described *Pichia* and *Candida guilliermondii* on the basis of several strains of which only two *P. guilliermondii* isolates were of clinical origin, whereas eight *C. guilliermondii* derived from clinical and animal samples. In both species the environmental strains predominated and a small number of food-related strains were reported (Kurtzman, 1984). Finally, it was moved to the *Meyerozyma* genus on the basis of the LSU and SSU sequence analysis (Kurtzman and Suzuki, 2010).

The complex *Candida*/*Meyerozyma guilliermondii* resulted rather ubiquitous in a number of ecological surveys. It has been isolated from

* Corresponding author. Dpt. of Pharmaceutical Sciences – Microbiology, Borgo 20 Giugno, 74, I – 06121 Perugia, Italy. Tel.: +39 075 585 6478; fax: +39 075 585 6470.
E-mail address: gianluigi.cardinali@unipg.it (G. Cardinali).

http://dx.doi.org/10.1016/j.fm.2014.12.014
0740-0020/

deep-sea hydrothermal systems of the Mid-Atlantic Rift (Gadanho and Sampaio, 2005) to wastewater treatment plants (Lahav et al., 2002), from insect surfaces (Suh and Blackwell, 2004) to maize wounds (Nout et al., 1997). *M. guilliermondii* was often isolated at relatively high densities in sugary fruit such as pineapple (Chanprasartsuk et al., 2010; Di Cagno et al., 2010), grapes and wine (Chavan et al., 2009; Li et al., 2010; Lopes et al., 2009) and resulted to be the predominant species on different varieties of apple and pear (Pelliccia et al., 2011). Its presence in food of animal origin, such as milk and salmon, was also detected (Seker, 2010; Yoshikawa et al., 2010).

This species resulted of industrial interest already in the first half of the last century, when it was observed that some strains can synthesize large quantities of riboflavin (Burkholder, 1943; Protchenko et al., 2000; Tanner jr., 1945). Later it was studied for its ability to absorb heavy metals from various sources (Balsalobre et al., 2003; de Siloniz et al., 2002; Junghans and Straube, 1991; Kaszycki et al., 2004; Ksheminska et al., 2003). More recently these two species were investigated for xylitol production and polycyclic aromatic hydrocarbons degradation (El-Latif Hesham et al., 2006; Mussatto et al., 2006).

The ability of *M. guilliermondii* to contrast *Penicillium expansum* and other fruit spoiling molds has been exploited by including this yeast in several study of non-chemical post-harvest control (Droby et al., 1997; Richards et al., 2004; Zhao et al., 2009, 2010) and by proposing its toxin to protect fruit (Coelho et al., 2009). The antimicrobial activity displayed by *M. guilliermondii* cells has been also proposed to protect wheat (Petersson and Schnurer, 1995), bread (Coda et al., 2013) and olives (Hernandez et al., 2008).

The interest in *C. guilliermondii* as potential opportunistic pathogen is present from the early literature (Castellani, 1912), although it was reported as the etiological agent of some 2.4% of the nosocomial pathogenic yeasts (Prasad et al., 1999). *C. guilliermondii* participates to the increasing interest in Non-*albicans Candida* (NAC) yeasts (Hospenthal et al., 2006; Krcmery and Barnes, 2002), as it is the sixth most frequent yeast isolated in clinical environment (Pfaller et al., 2006), causing up to 11.7% of the candidemia episodes (Girmenia et al., 2006).

The presence of this species in so many and different substrates poses the question on whether any selective pressure is selecting specialized strains for the various environments. This is particularly important if fruit and non-fruit isolates, hereinafter referred to as F and NF respectively, differ significantly, since the NF include isolates from pathogenic situations. This question is critical especially due to the presence of this species in both food and clinical environments and because several studies proposed the use of this yeast in a number of food applications. The molecular and metabolomics fingerprint carried out in this study intends to elucidate if any significant variation exists between strains of different origin and, in case, which markers can be readily employed to select strains not related to the clinical environment and safely of food origin.

2. Materials and methods

2.1. Strains and growth conditions

In this study, ninety six authentic strains, i.e. unambiguously classified with state of the art technology, of *M. guilliermondii* from different environments (Table 1) were analyzed. Twenty nines strains were provided by the CBS culture collection while the remaining by the internal microbial collection of the Microbial Genetics and Phylogenesis Laboratory of DSF (Department of Pharmaceutical Science, University of Perugia). Pineapple strains were included in the DSF collection after a study (Di Cagno et al., 2010) on the pineapple microbiota for which strains were isolated from the most inner part of the fruit (core), from the mid part of the pulp (pulp) and from the outer part of the pulp (external part). All strains were frozen stored at −80 °C in 17% glycerol.

2.2. Sequence analyses

2.2.1. LSU sequence analysis

All the strains were re-identified by LSU (D1/D2 26S) rDNA and ITS sequence analysis.

LSU analysis. Genomic DNA was extracted from yeast cells grown on YEPDA (Yeast Extract 1%, Peptone 1%, Dextrose 2%, Agarose 1.7%) Petri dishes following a protocol for colony extraction adjusted from the original one previously appeared in Cardinali et al. (Cardinali et al., 2001). The genomic DNA was amplified with FIREPol® Taq DNA Polymerase (Solis BioDyne, Estonia), using NL-1 (5′-GCAT ATCAATAAGCGGAGGAAAAG) and NL-4 (5′-GGTCCGTGTTTCAAGA CGG) (O'Donnell, 1993) primers in order to amplify the D1/D2 domain of 26S rDNA. The amplification protocol first appeared in Kurtzman and Robnett (Kurtzman and Robnett, 1998), as follows: initial denaturation at 95 °C for 4 min, 35 amplification cycles (94 °C for 1 min, 53 °C for 1 min and 72 °C for 1 min) and final extension at 72 °C for 10 min. Amplicons were purified using the GFX PCR DNA purification kit (GE Healthcare) while the electrophoresis was performed on 1.5% agarose gels (Gellyphor, EuroClone, Italy). Amplicons were sequenced in both directions with ABI PRISM technology by MACROGEN (www.macrogen.com) with the same primers used for the generation of the amplicons. Sequencing electropherograms data were processed with Geneious. D1/D2 LSU rDNA sequences identification queries was fulfilled by BLAST search (Altschul et al., 1990) in GenBank (www.ncbi.nlm.nih.gov/genbank/).

2.2.2. ITS sequence analysis

Genomic DNA was extracted as indicated by Cardinali et al. (Cardinali et al., 2001). ITS1, 5.8S, ITS2 rDNA genes were amplified with FIREPol® Taq DNA Polymerase (Solis BioDyne, Estonia), using ITS1 (5′-TCCGTAGGTGAACCTGCGG) - ITS4 (TCCTCCGCTTATTGAT ATGC) primers according to the same protocol explained for LSU amplification. Amplicons were purified with the GFX PCR DNA purification kit (GE Healthcare) and subject to electrophoresis on 1.5% agarose gel (Gellyphor, EuroClone, Italy). Amplicons were sequenced in both directions with ABI PRISM technology by MACROGEN (www.macrogen.com) with the same primers used for the generation of the amplicons. Consensus sequences for each strain and trimming of the ends with low sequencing quality were carried out with Geneious R6 (v. 6.17, Biomatters, Auckland, New Zealand, www.geneious.com).

2.2.3. LSU and ITS phylogenetic analysis

Alignment of the ITS and D1/D2 domain of the 26S rDNA (LSU) sequences was carried out with MUSCLE (Edgar, 2004) in MEGA6 (Tamura et al., 2013). Distances were inferred with the Maximum Composite Likelihood method and expressed as number of base substitutions per site. This procedure has been chosen because it assumes equal substitution patterns and rates among lineages and sites, conditions considered appropriate for a recent and ongoing separation phenomenon. Both transitions and transversions were considered. The Neighbor-Joining method (Saitou and Nei, 1987), was used to reconstruct the tree with 1000 bootstrap reiterations.

The distance analysis was performed in R environment (http://www.R-project.org) on the basis of the genetic distances calculated with MEGA6 as described above.

2.3. i-SSR $(GACA)_4$ and RAPD analysis

$(GACA)_4$ i-SSR PCR amplification was performed on genomic DNA, extracted as indicated by Cardinali et al. (Cardinali et al.,

Table 1
Strains of *Meyerozyma guilliermondii* (n = 96) employed in this study.

Strain number	Source of isolation	
CBS 463	Unknown	NF
CBS 566	Sputum (man)	NF
CBS 1909	Flowers of *Gentiana imbricata*	NF
CBS 2021	Unknown	NF
CBS 2024	Unknown	NF
CBS 2025	Butter milk	NF
CBS 2030[T]	Insect frass on *Ulmus americana*	NF
CBS 2031	Unknown	NF
CBS 2033	Mulberry bush	NF
CBS 2077	Lung (man)	NF
CBS 2082	Man	NF
CBS 2083	Blood of woman with ulcerated cheek	NF
CBS 2084	Sputum of bronchial patient	NF
CBS 2086	Atmosphere	NF
CBS 2672	Case of cystitis	NF
CBS 2830	Unknown	NF
CBS 2891	Spoiled leather	NF
CBS 4236	Kidney of child, together with *Candida albicans*	NF
CBS 5059	*Cossidae larvae*	NF
CBS 5241	Sake starter culture	NF
CBS 5483	Culture contaminant	NF
CBS 6021	Soil	NF
CBS 6316	Sewage	NF
CBS 6557	Pozol, Mexican fermented maize dough	NF
CBS 7099	Milk on isoprene nipple of baby's feeding bottle	NF
CBS 7369	Frass of *Synoxylon rufficorne*, in *Dichrostachys cinerea*	NF
CBS 8105	Production of citric acid	NF
CBS 8417	Brine bath in cheese factory	NF
CBS 9751	Hindgut of *Reticulitermes santonensis*	NF
LCF 1076	*Pyrus communis* cv. Abate Fetel	F
LCF 1077	*Pyrus communis* cv. Kaiser	F
LCF 1079	*Malus domestica* cv. Golden Delicious	F
LCF 1081	*M. domestica* cv. Golden Delicious	F
LCF 1087	*P. communis* cv. Kaiser	F
LCF 1088	*P. communis* cv. Kaiser	F
LCF 1089	*P. communis* cv. Kaiser	F
LCF 1090	*P. communis* cv. Kaiser	F
LCF 1091	*P. communis* cv. Kaiser	F
LCF 1102	*P. communis* cv. Abate Fetel	F
LCF 1103	*P. communis* cv. Abate Fetel	F
LCF 1104	*P. communis* cv. Abate Fetel	F
LCF 1105	*P. communis* cv. Abate Fetel	F
LCF 1106	*Malus domestica* cv. Fuji	F
LCF 1108	*M. domestica* cv. Fuji	F
LCF 1109	*M. domestica* cv. Fuji	F
LCF 1113	*M. domestica* cv. Fuji	F
LCF 1131	*P. communis* cv. Abate Fetel	F
LCF 1132	*M. domestica* cv. Golden Delicious	F
LCF 1133	*M. domestica* cv. Golden Delicious	F
LCF 1140	*M. domestica* cv. Golden Delicious	F
LCF 1352	*Ananas comosus* (external part)[a]	F
LCF 1353	*A. comosus* (external part)	F
LCF 1354	*A. comosus* (external part)	F
LCF 1355	*A. comosus* (external part)	F
LCF 1356	*A. comosus* (external part)	F
LCF 1357	*A. comosus* (external part)	F
LCF 1358	*A. comosus* (external part)	F
LCF 1359	*A. comosus* (external part)	F
LCF 1360	*A. comosus* (external part)	F
LCF 1361	*A. comosus* (external part)	F
LCF 1362	*A. comosus* (external part)	F
LCF 1363	*A. comosus* (external part)	F
LCF 1364	*A. comosus* (external part)	F
LCF 1365	*A. comosus* (external part)	F
LCF 1366	*A. comosus* (external part)	F
LCF 1367	*A. comosus* (external part)	F
LCF 1368	*A. comosus* (external part)	F
LCF 1369	*A. comosus* (external part)	F
LCF 1371	*A. comosus* (pulp)	F
LCF 1372	*A. comosus* (pulp)	F
LCF 1373	*A. comosus* (pulp)	F
LCF 1374	*A. comosus* (pulp)	F
LCF 1375	*A. comosus* (pulp)	F
LCF 1376	*A. comosus* (pulp)	F
LCF 1377	*A. comosus* (pulp)	F
LCF 1378	*A. comosus* (pulp)	F
LCF 1379	*A. comosus* (pulp)	F
LCF 1381	*A. comosus* (core)	F
LCF 1382	*A. comosus* (core)	F
LCF 1383	*A. comosus* (core)	F
LCF 1384	*A. comosus* (core)	F
LCF 1385	*A. comosus* (core)	F
LCF 1386	*A. comosus* (core)	F
LCF 1387	*A. comosus* (core)	F
LCF 1388	*A. comosus* (core)	F
LCF 1389	*A. comosus* (core)	F
LCF 1390	*A. comosus* (core)	F
LCF 1391	*A. comosus* (core)	F
LCF 1392	*A. comosus* (core)	F
LCF 1393	*A. comosus* (core)	F
LCF 1394	*A. comosus* (core)	F
LCF 1395	*A. comosus* (core)	F
LCF 1396	*A. comosus* (core)	F
LCF 1397	*A. comosus* (core)	F
LCF 1398	*A. comosus* (core)	F
LCF 1399	*A. comosus* (core)	F

The abbreviations F and NF indicate the fruit or not-fruit source of isolation, respectively.
[a] "External part" was used as a generic term in order to identify the jagged-edged bract surface subtending each single berry composing the pineapple fruit. This sample also involved the underneath fleshy layer in which berries are imbedded.

2001), following the protocol previously described by Andrade et al. as "microsatellite primers $(GACA)_4$" (Andrade et al., 2006), using EuroTaq enzyme (EuroClone, Italy) in an OnGradient Thermal Cycler apparatus (EuroClone, Italy). Genomic DNA from a restricted set of strains isolated from pineapple (*Ananas comosus* L. Merr.) was also amplified by means of a duplex RAPD amplification carried out with primers M13m (5′ – GAG GGT GGC GGT TC – 3′) and Rp 11 (5′ – GAA ACT CGC CAA G – 3′). DNA was amplified for 34 cycles (denaturation, 94 °C 1 min; annealing, 40 °C 1:10 min; extension, 72 °C 1:10 min) followed by a single 15 min extension at 72 °C.

$(GACA)_4$ i-SSR and RAPD profiles were converted in binary (0/1) matrixes with the ClassMaker 1.27 software (Cardinali et al., 2003), following the procedures described in the paper and with an additional check that similar bands were included in the same band class. Euclidean distances among the strains were calculated from the binary matrices and used to build the DIANA tree with the Cluster Package (Kaufman and Rousseeuw, 1990) in the R statistical environment.

2.4. FTIR analysis and spectra pre-processing

Pre-cultures of the 96 strains of *M. guilliermondii* were inoculated at $OD_{600} = 0.3$ in 100 ml bottles containing 20 ml YEPD medium (Yeast Extract 1%, Peptone 1%, Dextrose 2%- Difco Laboratories, USA) and were grown for 18 h at 25 °C, with 130 rpm shaking. For each strain 105 μl volume was sampled for three independent FTIR readings [35 μl each, according to the technique suggested by Manfait and colleagues (Essendoubi et al., 2005)]. FTIR measurements were performed in transmission mode. All spectra were recorded in the range between 4000 and 400 cm^{-1} with a TENSOR 27 FTIR spectrometer, equipped with HTS-XT accessory for rapid automation of the analysis (BRUKER Optics GmbH, Ettlingen, Germany). Spectral resolution was set at 4 cm^{-1}, sampling 256 scans per sample. OPUS version 6.5 software (BRUKER Optics GmbH, Ettlingen, Germany) was used to carry out the quality test, baseline correction, vector normalization and the calculation of the first and second derivatives of spectral values.

2.5. Data processing

2.5.1. FTIR cluster analysis

The first part of this analysis was performed using the OPUS software (Bruker GmbH, Ettlingen – Germany). To compare the spectra of the different samples, cluster analysis using the second derivatives of the original spectra as input was carried out for different spectral regions. Dendrograms were obtained using the Euclidean algorithm to calculate distances and the Pearson coefficient for the correlations among spectra. Heterogeneity within the dendrogram (reported as y-scale of the dendrogram) has been defined according to the Ward's algorithm according to Formula 1:

$$\textit{Formula 1} \quad H(r,i) = \{[n(p)+n(i)]\cdot D(p,i) + [n(i) + n(q)]\cdot D(q,i) - n(i)\cdot D(q,i)\}/[n+n(i)]$$

where H indicates the heterogeneity, D indicates distances, n indicates the number of spectra Subscripts "p" and "q" indicate successive clusters, whereas the "i" subscript designates the ith spectrum whose heterogeneity is calculated. Spectra were classified by using the OPUS cluster analysis based on a hierarchical classification algorithm. The procedure has gone as follows: vectorial normalization, and the calculation of the second derivative using a Savitsky–Golay algorithm, with nine smoothing points. This pre-processing was carried out for all spectra on the spectral region with biologically relevant information (cm^{-1} $[3200-2800]+[1800-700]$). The derivation of the spectra to the second order was used to increase the number of discriminant features present in the spectra. The spectra were classified by using the OPUS hierarchical cluster analysis based on Ward's classification algorithm. The function used, minimized the variance intraclass of the spectra and represented this in a cluster, according to their similarities. The spectral windows were chosen to obtain a consistent classification of the strains.

2.5.2. Correlation analysis

Spectra second derivatives data were exported as an ASCII file from OPUS and used in the "R" environment (http://cran.r-project.org/, 2011) to carry out normalization with range spanning from 0 to 1 (Huang et al., 2006), and spectra averaging. Spectra correlation analyses were performed by subdividing the whole spectrum in five different regions: fatty acids (W1) from 3000 to 2800 cm^{-1}, amides (W2) from 1800 to 1500 cm^{-1}, mixed region (W3) from 1500 to 1200 cm^{-1}, carbohydrates (W4) from 1200 to 900 cm^{-1} and typing region (W5) from 900 to 700 cm^{-1} (Kummerle et al., 1998).

The matrices with the LSU, ITS and i-SSR data were imported in the same R-environment to carry out distance, ANOVA and Mantel (with 999 permutations) analyses. A correlation analysis between i-SSR matrix and the six spectral matrices (whole spectra plus the five regions taken independently) was performed.

All other statistical analyses were carried out in R using the "base" package.

3. Results

3.1. LSU and ITS analysis

The distance analysis carried out on the D1/D2 (LSU) domain sequences indicated that all strains of this study (Table 1) belong to *M. guilliermondii*, according to the widely accepted concept that the strains of one species should show less than 1% distance in the 26S rDNA gene (Kurtzman and Robnett, 1998). In fact, the distances from the type strain (CBS 2030^T) ranged from 0 to 0.7% with a 0.4% mean and 0.24 standard deviation. The ITS marker, recently suggested as a general barcode marker for Fungi (Schoch et al., 2014, 2012), corroborated the LSU results, yielding distances from CBS 2030^T spanning from 0 to 0.7% with 0.34% mean and 0.16% standard deviation. The environmental and medical isolates showed 0.103% and 0.213% mean distance from the type strain, using LSU and ITS respectively. Interestingly, the food isolates diverged from the type strain by 0.529% and 0.400%, indicating that the two groups are statistically different ($p < 0.0001$) although both belong to the same species.

A Mantel test between the distance matrixes obtained with the two markers gave 0.75 r, with 0.001 significance of the null hypothesis. These data together indicate that there is an overall agreement between the two markers, allowing to use their assemblage for the construction of a phylogenetic tree.

The phylogenetic tree resulting from the LSU and ITS sequences concatenated confirmed that the environmental and medical isolates differ from the fruit strains (Fig. 1). Isolates from fruit clustered in a clade with 67% bootstrap support, excluding LCF 1090, LCF 1091 and LCF 1103, all isolated from pears. The strains isolated from pineapple clustered in a series of subclades without any strong phylogenetic support, indicating a relatively independent evolution of these strain groups within the species and within the F strains. The NF strains clustered with low (36%) bootstrap support. The medically related strains (CBS 2077, 2082, 2083, 2672) were scattered among the other NF isolates. These observations indicated that the F and NF isolates could be separated, although no obvious discrimination could be found between NF isolates.

3.2. Molecular characterization

The $(GACA)_4$ primers were chosen for the i-SSR analysis after a series of preliminary tests in which the $(GAC)_5$ and $(GTG)_5$ (Andrade et al., 2006) primers showed to produce little if any resolution among the tested strains (data not shown). According to the distribution obtained from the analysis of the i-SSR banding the 96 strains formed four major clades (Fig. 2). The large clade (Clade1) accommodates many of the isolates from pineapple and only one (LCF 1088) from Kaiser pears isolated before the industrial washings. Strain LCF 1385 from pineapple core was positioned as outlier of Clade 1. Clade 2 was articulated in four subclades of which one was composed by three strains derived from pears (LCF 1076, 1087 and 1131), one by three other strains derived from pineapple exterior and core. A third subgroup included five strains isolated from pears and apples and finally the last subclade comprised strains from apples and pears, one strain isolated from sake (CBS 5241) and one from insect larvae (CBS 5059). Clade 3 was entirely composed of NF strains isolated from various substrates such as spoiled leather (CBS 2891), soil (CBS 6021), Mexican maize dough (CBS 6557), frass of insect (CBS 7369), cheese brine (CBS 8417), insect gut (CBS 9751) and a strain from man sputum (CBS 566). The fourth clade (Clade 4) included strains of clear medical origin and other from various sources, among which buttermilk (CBS 2025), air (CBS 2086) and pears (LCF 1090, 1091 and 1103). Interestingly, the three strains from pears in clade CL4 (LCF 1090, 1091 and 1103) are the same that clustered apart from the others in the fruit strain clade according to the LSU and ITS sequences, supporting that these three strains differ significantly from the other fruit isolates. In general, the distance matrix obtained with i-SSR gave Mantel r 0.613 r ($p = 0.001$) with LSU and 0.354 ($p = 0.001$) with ITS, indicating a good level of correlation between the i-SSR and the LSU, but not between the i-SSR and the ITS.

The variability among the strains from pineapple according to both 26S and i-SSR analyses was relatively high, especially considering that all isolates derived from few fruits. This suggested carrying out further molecular analyses to better characterize the

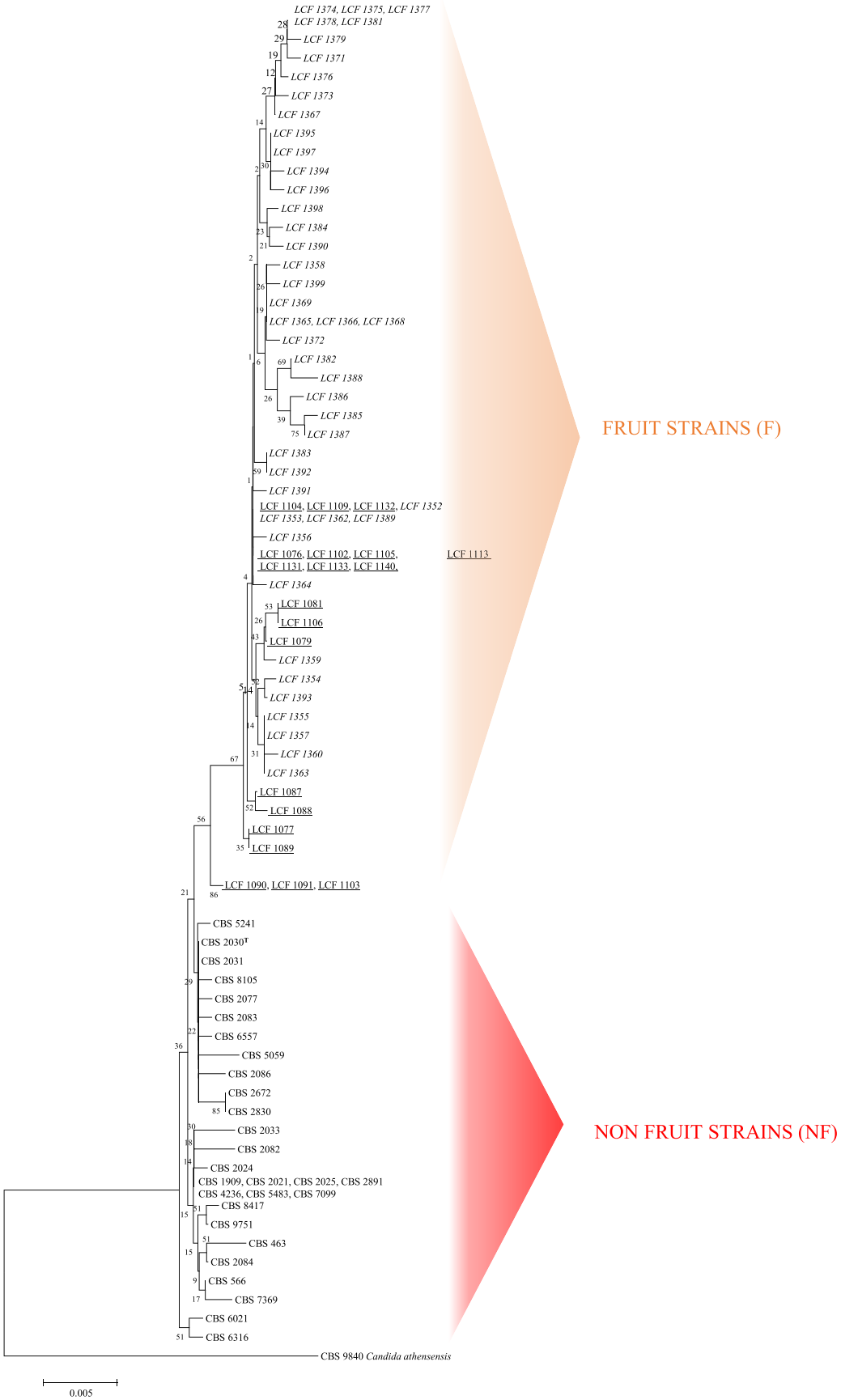


Fig. 1. Neighbor-Joining phylogenetic tree based on the ITS and LSU aligned sequences. **Legend**. The evolutionary distances were computed using the Maximum Composite Likelihood method expressed as number of base substitutions per site. Neighbor-Joining trees were obtained using the Maximum Composite Likelihood with MEGA6; bootstrap support values were calculated with 1000 replicates and are shown next to the branches. GenBank accession numbers are reported in Table S1. Plain character: medical/environmental strains; Italic character: pineapple strains; Underlined character: pear/apple strains.

variability of these strains isolated from sound fruit without contact with the external environment and therefore constituting an interesting model of selection in restricted conditions. The RAPD and i-SSR pattern combination produced a distribution of the pineapple isolates in two major clades (Fig. 3), of which Clade A included strains isolated from the core of the pineapple slices, with three exceptions isolated from the pulp (LCF 1377, 1378 AND 1379). The strains in Clade B were mostly isolated from the external part of the slice with only one exception derived from the pulp (LCF 1372). Five strains were not included in these two groups. Three of them (LCF 1352, 1355 and 1389) isolated from the external and pulp part of the slice formed a small distinct clade, confirming their difference with the other pineapple strains already detected with the sole i-SSR profiles (Fig. 2). Apart from a few exceptions, the isolates from the external part of the fruit were well separated from those present in the core, the strains of the slice pulp being distributed in both clades. These data suggested that the pineapple fruit is either a favorable environment to induce and select variability or that the various conditions throughout the fruit select different strains, maybe deposited by the insects during the blooming.

3.3. FTIR typing

The yeast whole cells subject to spectroscopic analysis revealed differences when comparing the spectra second derivatives (Fig. 4) in the W3 region, especially in the range 1200 to 1150 cm^{-1} normally attributed to DNA, RNA and phospholipids (P=O asymmetric stretches around 1240 cm^{-1}) and carbohydrates (1200–1000 cm^{-1}) (Yu and Irudayaraj, 2005). The so-called typing region (W5) showed some differences, particularly between medical/environmental isolates and fruit isolates in the spectral range around 800 cm^{-1}. These data confirmed the importance of the carbohydrates and typing region in strains discrimination (Kummerle et al., 1998; Naumann et al., 1991). The dendrogram obtained with the spectral second derivative data displayed a strains distribution into three major groups, two (FCL1 and FCL2) that include fruit isolates, while FCL3 is composed by strains of different origin including those of clinical interest (Fig. 4). Most of the medical isolates of this clade gave the same clustering together as was found for the LSU/ITS dendrogram. FCL2 included mostly pineapple isolates with few exceptions, whereas FCL1 had mostly apple and pears isolates with some pineapple strains partly grouped in a sub-clade. No discrimination could be obtained between the strains of the FCL1 and FCL2 clades, according to the isolation source (apple and pear vs. pineapple), nor according to the isolation position in the pineapple slice. The strains of the FCL3 clade were mostly environmental and medical with the exception of LCF 1090 and 1103 deriving from fruit. This fact confirms that these isolates were somehow intermediate between the medical/environmental and the food strains, as shown by their presence in the i-SSR clade CL4, formed by medical and environmental yeasts (Fig. 2), in a separate sub-clade according to the LSU/ITS dendrogram (Fig. 1). Clinical and environmental isolates could not be differentiated as also by the FTIR analysis (Fig. 4) with any combination of spectral regions, indicating that indeed these strains are extremely similar from the overall metabolomics viewpoint, as it was previously shown with the molecular analyses (Figs. 2 and 3).

4. Discussion

M. guilliermondii is one of the most widespread yeasts in nature and particularly on fruit surfaces (Pelliccia et al., 2011). It has been isolated from several sources in different geographical areas and often on or in some fruit like pears, apples and pineapples (Chanprasartsuk et al., 2010). It is considered one of the main agents for organic post harvest control of fruit, due to its ability to inhibit the growth of moulds on the fruit surfaces and a promising agent in some biotechnological processes (Coda et al., 2013; Matos et al., 2013). In spite of the "one fungus one name" (Taylor, 2011) it retains the epithet *C. guilliermondii* for its imperfect state (anamorph), with which it is normally designated in medical literature being a known, opportunistic pathogen.

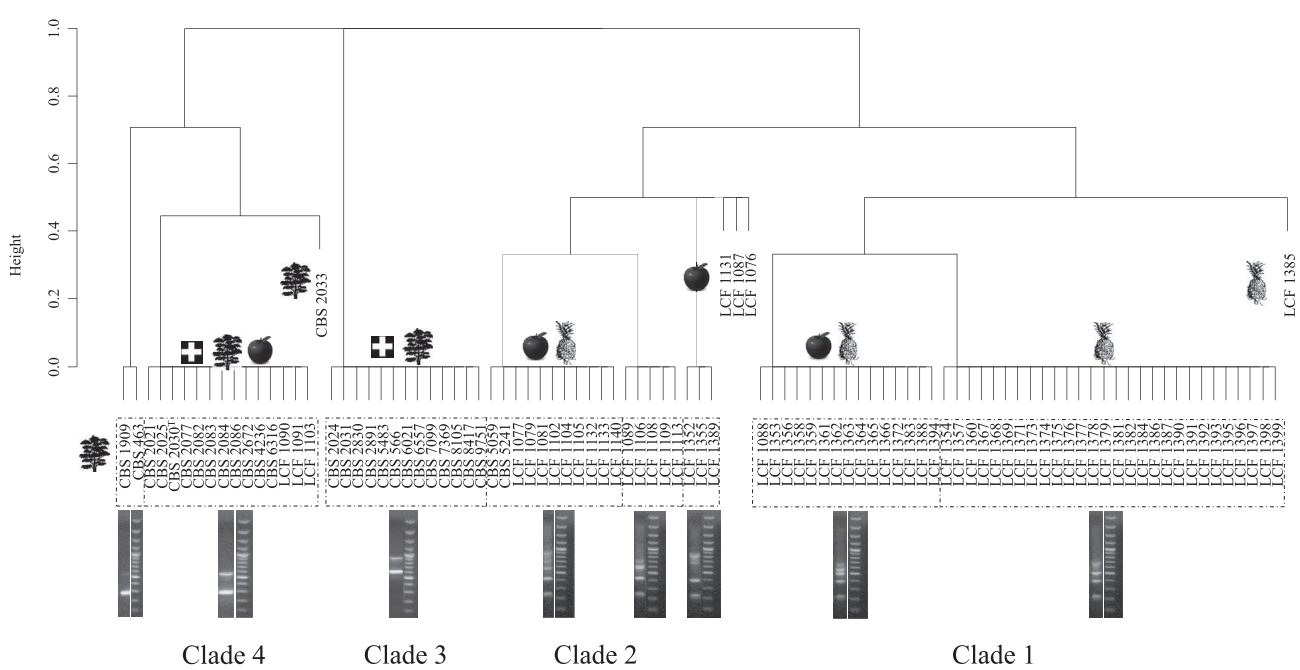


Fig. 2. Hierarchical dendrogram obtained from the $GACA_4$ distance matrix of the 96 strains employed in the study. **Legend**. The four major clades represent the distribution obtained from the i-SSR banding analysis. Each cluster is described with a symbol, indicating different source of strain isolation: pears, apples, pineapple, various natural environments (tree) and clinical situations (cross).

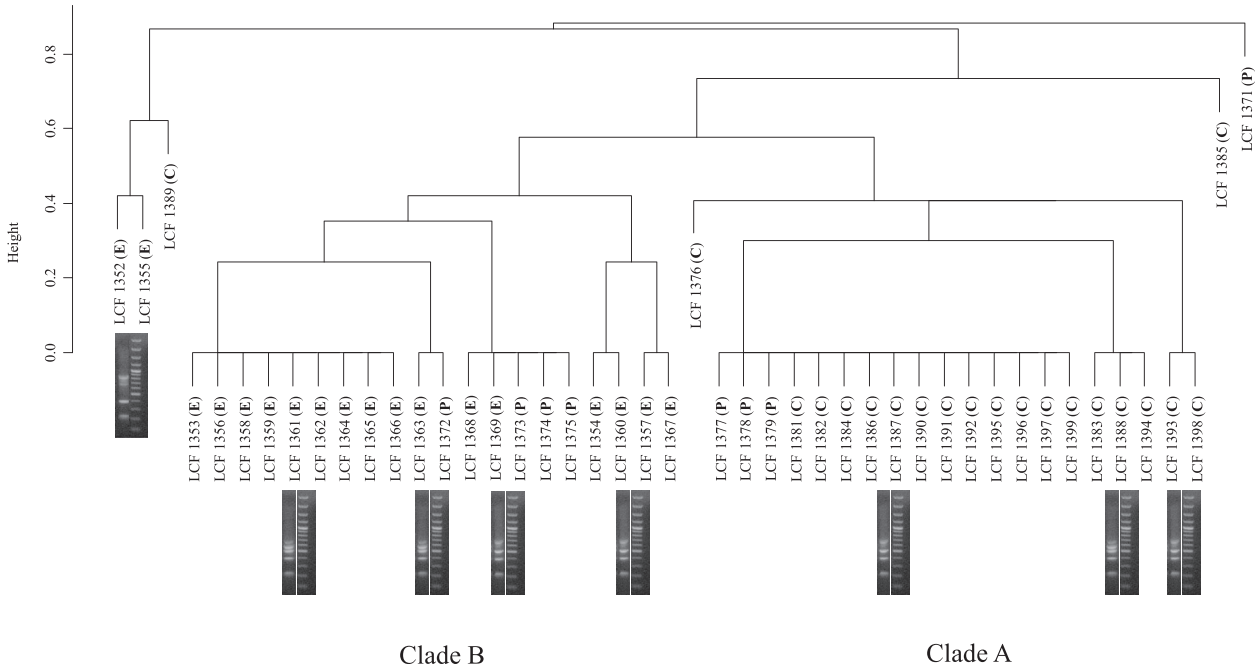


Fig. 3. RAPD and iSSR combined clustering of pineapple isolates, differentiated according to the colonized fruit portion. **Legend**. The pattern profiles clustered together according to the portion of fruit from which strains were isolated. **E)** strains isolated from the fruit external part, **P)** strains isolated from fleshy pulp portion and **C)** strains isolated from the inner core of the fruit.

Due to the above reasons, *M. guilliermondii* should have a questionable GRAS (Generally Regarded As Safe) or QPS (Qualified Presumption of Safety) status, in fact the European legislation at the moment is not considering this species neither in the QPS nor in the non-QPS list (http://www.efsa.europa.eu/en/efsajournal/doc/3449.pdf). The type strain and the patented strains of this species were reported to be non pathogenic in animal model tests, although the same author reported that this species is the first cause of fungemia in cancer patients (Sibirny and Boretsky Yu, 2009).

Given this complex and potentially dangerous situation, new tools to discriminate among strains of different origin are necessary to avoid harmful applications of medical isolates in food and biotechnological industries.

The two accepted DNA markers (LSU and ITS) in yeast taxonomy (Kurtzman and Robnett, 1998; Schoch et al., 2012) produced a separation of fruit (F) and non fruit (NF) strains.

The phylogenetic tree based on these two genes showed relatively low bootstrap values. Particularly, the F clade had 56% or 67% support depending on whether or not were included the three pear isolates (LCF 1090, LCF 1091 and LCF 1103), which were shown to be somehow anomalous F strains by all other analyses. These bootstrap figures are somehow lower than the 70% recommended by Hillis and Bull (Hillis and Bull, 1993) as a reference value to state that the clade is real. However, it must be stressed that this study was carried out based on genetic distances much larger than those found between the *M. guilliermondii* strains. It is also possible to hypothesize that an ongoing speciation event can be characterized by low bootstrapping support, as found in other papers in which low levels of variability observed between separated species were correlated with low bootstrap support (Diekmann et al., 2001; Tryfonopoulos et al., 2008).

The distances between strains and type strain never exceeded the 1% threshold, commonly used to define different species, and therefore *M. guilliermondii* cannot be split into two species. This fact generated the need to find characterization markers to discriminate the strains within the species, in order to avoid the dangerous contaminations of NF strains in food-related industry, as stated above.

Two completely different fingerprinting techniques such as i-SSR and FTIR provided an extensive description of these strains, confirming the discrimination between F and NF isolates observed with LSU and ITS. The separation between F and NF strains could be explained both as genetic drift or as the result of selective pressure exerted by the fruit environment. Dissecting between these two hypotheses is beyond the scope of the present work, but the fact that the two groups F and NF isolates are more differentiated in terms of FTIR fingerprint than according to the molecular characterizations, is an indication that the selective hypothesis is more likely.

Only three F strains, derived from pears, clustered with NF isolates, according to both i-SSR and FTIR. This finding is in line with the peripheral positioning of these three strains obtained with the two taxonomic markers LSU and ITS. The presence of medically related isolates in fruits is not uncommon as shown in the paper describing the isolation of these three strains (Pelliccia et al., 2011) in which the medical relevant yeasts *Wickerhamomyces anomalus* and *Candida famata* (*Debaryomyces hansenii*) have been found (Chan et al., 2013; Feng et al., 2014). However an environmental contamination of fruits during the washings or other microbial circulation mechanisms deserve more insight to elucidate the relation between strains of the same species derived from separate habitats. Furthermore a possible circulation of strains would rule out the genetic drift hypothesis to explain F and NF isolates. None of the tools employed in this work has been able to discriminate between medical and environmental isolates within the NF group, although medical strains tended to cluster together in the FTIR dendrogram. This evidence confirms the high sensitivity of the metabolomic fingerprint in discriminating strains subject to different stressing conditions (Corte et al., 2012; Perromat et al., 2003).

This observation indicates that an active circulation of strains between natural and medical environment exists and is more active

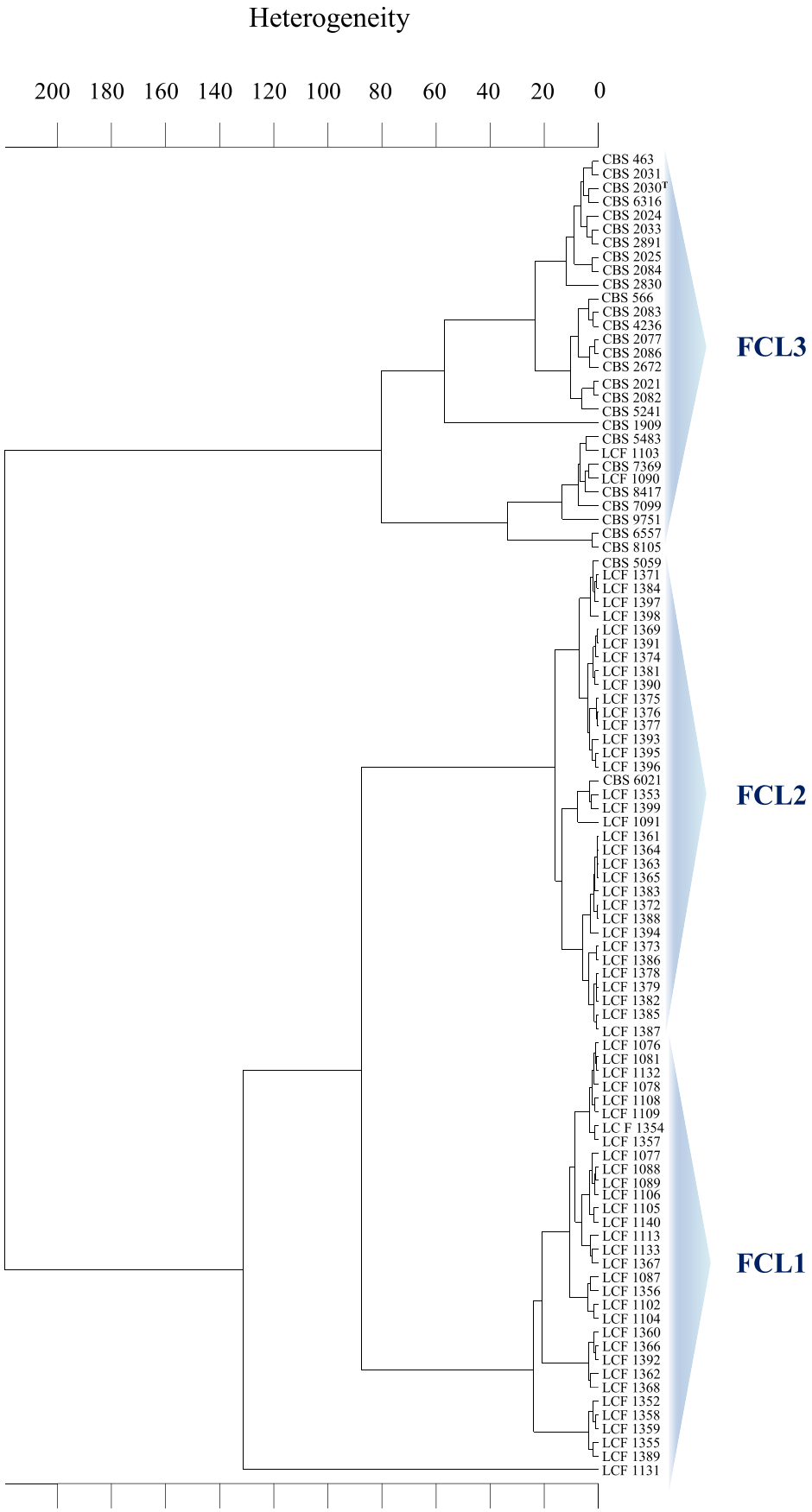


Fig. 4. FTIR analysis of the 96 *M. guillermondii* strains. **Legend**. Hierarchical cluster analysis of 96 strains, obtained calculating the distance between the second derivatives spectra, considering a combination of the W3–W4 region, ranging from 1489 to 1150 cm^{-1}, and W5 region, and assigning relative weights of 0.05 and 1 respectively.

than that observed between F and NF habitats, further reinforcing the concept that the F vs NF strains separation is likely due to a selective pressure. The ability of this yeast to ferment and assimilate a very large range of carbon sources (Kurtzman, 2011; Kurtzman and Suzuki, 2010) is an additional element to justify its presence in various substrates and environments.

A peculiar type of variability was found within the pineapple fruits among the strains isolated from the core, the pulp and the periphery of the fruit slice. Very few strains from the pulp were similar to those found in the core (LCF 1371, 1377, 1378 and 1379) and in the periphery (LCF 1373, 1374 and 1375). Only strain LCF 1389, isolated from the core, was similar to the strains deriving from the external part. These few exceptions to an otherwise discrimination, can be justified by the inevitable carry over produced by the slicing operations.

Pineapple fruits consist of coalesced berries derived each from a different flower. Hummingbirds, butterflies and bees are known to visit pineapple flowers ad may possibly deposit onto the flowers different yeast strains, which then remain trapped in the growing fruit (de Queiroz Piacentini and Varassin, 2007). According to the fruit structure and evolution, the strain variability, observed within the pineapple fruit, could be explained by multiple inoculations of the many pineapple flowers. This mechanism does not fully explain the variability found among pineapple isolates, which could be due also by the presence of environmental conditions, favoring the strain differentiation. Should this explanation hold, the species would be particularly plastic from the genetic point of view and able to adapt even to micro-environmental conditions such as those found within few centimeters within pineapple fruit.

All together it seems that *M. guilliermondii* is very adaptable and that some sort of variation is occurring between F and NF strains. This evidence should invite to a careful monitoring of the isolates, especially when mass productions are carried out for biotechnological purposes. The possibility to discriminate among strains with molecular and metabolomic analyses is an additional tool to empower this monitoring and to gain further knowledge on the genetic variations of this species, which is also a potential and interesting starter for fruit and vegetable storage.

Acknowledgments

The work was supported partly by a grant of the Italian Ministry of Agriculture (MiPAF). CC was supported by PhD fellowships of the Italian Ministry of Education and Research (MURST), 29th cycle; LR was partially supported by a PRIN grant of the MURST (prot. 2010WZ2NJN): “Microorganisms in foods and in humans: study of the microbiota and the related metabolome as affected by omnivore, vegetarian or vegan diets”.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.fm.2014.12.014.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Andrade, M.J., Rodriguez, M., Sanchez, B., Aranda, E., Cordoba, J.J., 2006. DNA typing methods for differentiation of yeasts related to dry-cured meat products. Int. J. Food Microbiol. 107, 48–58.
Balsalobre, L., De Siloniz, M.I., Valderrama, M.J., Benito, T., Larrea, M.T., Peinado, J.M., 2003. Occurrence of yeasts in municipal wastes and their behaviour in presence of cadmium, copper and zinc. J. Basic Microbiol. 43, 185–193.
Burkholder, P.R., 1943. Synthesis of riboflavin by a yeast. Proc. Nat. Acad. Sci. U. S. A. 29, 166–172.
Cardinali, G., Bolano, A., Martini, A., 2001. A DNA extraction and purification method for several yeast genera. Ann. Microbiol. 51, 121–130.
Cardinali, G., Maraziti, F., Selvi, S., 2003. Electrophoretic data classification for phylogenetics and biostatistics. Bioinformatics 19, 2163–2165.
Castellani, A., 1912. Note on the importance of hyphomycetes and other fungi in tropical pathology. Br. Med. J. 2, 1208–1212.
Chan, A.W., Cartwright, E.J., Reddy, S.C., Kraft, C.S., Wang, Y.F., 2013. *Pichia anomala* (*Candida pelliculosa*) fungemia in a patient with sickle cell disease. Mycopathologia 176, 273–277.
Chanprasartsuk, O.O., Prakitchaiwattana, C., Sanguandeekul, R., Fleet, G.H., 2010. Autochthonous yeasts associated with mature pineapple fruits, freshly crushed juice and their ferments; and the chemical changes during natural fermentation. Bioresour. Technol. 101, 7500–7509.
Chavan, P., Mane, S., Kulkarni, G., Shaikh, S., Ghormade, V., Nerkar, D.P., Shouche, Y., Deshpande, M.V., 2009. Natural yeast flora of different varieties of grapes used for wine making in India. Food Microbiol. 26, 801–808.
Coda, R., Rizzello, C.G., Di Cagno, R., Trani, A., Cardinali, G., Gobbetti, M., 2013. Antifungal activity of *Meyerozyma guilliermondii*: identification of active compounds synthesized during dough fermentation and their effect on long-term storage of wheat bread. Food Microbiol. 33 (2), 243–251.
Coelho, A.R., Tachi, M., Pagnocca, F.C., Nobrega, G.M., Hoffmann, F.L., Harada, K., Hirooka, E.Y., 2009. Purification of *Candida guilliermondii* and *Pichia ohmeri* killer toxin as an active agent against *Penicillium expansum*. Food Addit. Contam. Part A Chem. Anal. Control Expo. Risk Assess. 26, 73–81.
Corte, L., Roscini, L., Zadra, C., Antonielli, L., Tancini, B., Magini, A., Emiliani, C., Cardinali, G., 2012. Effect of pH on potassium metabisulphite biocidic activity against yeast and human cell cultures. Food Chem. 134 (3), 1327–1336.
de Queiroz Piacentini, V., Varassin, I.G., 2007. Interaction network and the relationships between bromeliads and hummingbirds in an area of secondary Atlantic rain forest in southern Brazil. J. Trop. Ecol. 23, 663.
de Siloniz, M.I., Balsalobre, L., Alba, C., Valderrama, M.J., Peinado, J.M., 2002. Feasibility of copper uptake by the yeast *Pichia guilliermondii* isolated from sewage sludge. Res. Microbiol. 153, 173–180.
Di Cagno, R., Cardinali, G., Minervini, G., Antonielli, L., Rizzello, C.G., Ricciuti, P., Gobbetti, M., 2010. Taxonomic structure of the yeasts and lactic acid bacteria microbiota of pineapple (*Ananas comosus* L. Merr.) and use of autochthonous starters for minimally processing. Food Microbiol. 27, 381–389.
Diekmann, O., Bak, R., Stam, W., Olsen, J., 2001. Molecular genetic evidence for probable reticulate speciation in the coral genus *Madracis* from a Caribbean fringing reef slope. Mar. Biol. 139, 221–233.
Droby, S., Wisniewski, M.E., Cohen, L., Weiss, B., Touitou, D., Eilam, Y., Chalutz, E., 1997. Influence of $CaCl_{(2)}$ on *Penicillium digitatum*, Grapefruit Peel Tissue, and Biocontrol activity of *Pichia guilliermondii*. Phytopathology 87, 310–315.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
El-Latif Hesham, A., Khan, S., Liu, X., Zhang, Y., Wang, Z., Yang, M., 2006. Application of PCR-DGGE to analyse the yeast population dynamics in slurry reactors during degradation of polycyclic aromatic hydrocarbons in weathered oil. Yeast 23, 879–887.
Essendoubi, M., Toubas, D., Bouzaggou, M., Pinon, J.M., Manfait, M., Sockalingum, G.D., 2005. Rapid identification of *Candida* species by FT-IR microspectroscopy. Biochim. Biophys. Acta 1724, 239–247.
Feng, X., Wu, J., Ling, B., Yang, X., Liao, W., Pan, W., Yao, Z., 2014. Development of two molecular approaches for differentiation of clinically relevant yeast species closely related to *Candida guilliermondii* and *Candida famata*. J. Clin. Microbiol. 52, 3190–3195.
Gadanho, M., Sampaio, J.P., 2005. Occurrence and diversity of yeasts in the mid-atlantic ridge hydrothermal fields near the Azores Archipelago. Microb. Ecol. 50, 408–417.
Girmenia, C., Pizzarelli, G., Cristini, F., Barchiesi, F., Spreghini, E., Scalise, G., Martino, P., 2006. *Candida guilliermondii* fungemia in patients with hematologic malignancies. J. Clin. Microbiol. 44, 2458–2464.
Hernandez, A., Martin, A., Cordoba, M.G., Benito, M.J., Aranda, E., Perez-Nevado, F., 2008. Determination of killer activity in yeasts isolated from the elaboration of seasoned green table olives. Int. J. Food Microbiol. 121, 178–188.
Hillis, D.M., Bull, J.J., 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42, 182–192.
Hospenthal, D.R., Beckius, M.L., Floyd, K.L., Horvath, L.L., Murray, C.K., 2006. Presumptive identification of *Candida* species other than *C. albicans*, *C. krusei*, and *C. tropicalis* with the chromogenic medium CHROMagar *Candida*. Ann. Clin. Microbiol. Antimicrob. 5, 1.
Huang, W.E., Hopper, D., Goodacre, R., Beckmann, M., Singer, A., Draper, J., 2006. Rapid characterization of microbial biodegradation pathways by FT-IR spectroscopy. J. Microbiol. Methods 67, 273–280.
Junghans, K., Straube, G., 1991. Biosorption of copper by yeasts. Biol. Met. 4, 233–237.
Kaszycki, P., Fedorovych, D., Ksheminska, H., Babyak, L., Wojcik, D., Koloczek, H., 2004. Chromium accumulation by living yeast at various environmental conditions. Microbiol. Res. 159, 11–17.
Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data. Wiley.
Krcmery, V., Barnes, A.J., 2002. Non-*albicans Candida* spp. causing fungaemia: pathogenicity and antifungal resistance. J. Hosp. Infect. 50, 243–260.
Ksheminska, H., Jaglarz, A., Fedorovych, D., Babyak, L., Yanovych, D., Kaszycki, P., Koloczek, H., 2003. Bioremediation of chromium by the yeast *Pichia guilliermondii*: toxicity and accumulation of Cr (III) and Cr (VI) and the influence of riboflavin on Cr tolerance. Microbiol. Res. 158, 59–67.

Kummerle, M., Scherer, S., Seiler, H., 1998. Rapid and reliable identification of food-borne yeasts by Fourier-transform infrared spectroscopy. Appl. Environ. Microbiol. 64, 2207–2214.
Kurtzman, C., 2011. *Meyerozyma* Kurtzman & M. Suzuki (2010). The Yeasts: a Taxonomic Study, fifth ed. Elsevier, pp. 621–624.
Kurtzman, C.P., 1984. *Pichia* Hansen. In: Kreger-van Rij, N.J.W. (Ed.), The Yeasts-a Taxonomic Study, third ed. Elsevier Science Publishers, Amsterdam, pp. 329–330.
Kurtzman, C.P., Robnett, C.J., 1998. Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. Ant. Van Leeuwenhoek 73, 331–371.
Kurtzman, C.P., Suzuki, M., 2010. Phylogenetic analysis of ascomycete yeasts that form coenzyme Q-9 and the proposal of the new genera *Babjeviella, Meyerozyma, Millerozyma, Priceomyces*, and *Scheffersomyces*. Mycoscience 51, 2–14.
Lahav, R., Fareleira, P., Nejidat, A., Abeliovich, A., 2002. The identification and characterization of osmotolerant yeast isolates from chemical wastewater evaporation ponds. Microb. Ecol. 43, 388–396.
Langerhorn, M., Guerra, P., 1938. Nouvelles recherches de zymologie médical. Ann. Parasitol. Hum. Comparée 16, 429–476.
Li, S.S., Cheng, C., Li, Z., Chen, J.Y., Yan, B., Han, B.Z., Reeves, M., 2010. Yeast species associated with wine grapes in China. Int. J. Food Microbiol. 138, 85–90.
Lodder, J., Kreger-van Rij, N.J.W., 1952. The Yeasts, a Taxonomic Study, first ed. North-Holland, Amsterdam.
Lopes, C.A., Jofre, V., Sangorrin, M.P., 2009. Spoilage yeasts in Patagonian wine-making: molecular and physiological features of *Pichia guilliermondii* indigenous isolates. Rev. Argent. Microbiol. 41, 177–184.
Matos, I.T.S.R., Cassa-Barbosa, L.A., Galvão, R.d.S.M., Nunes-Silva, C.G., Astolfi Filho, S., 2013. Isolation, taxonomic identification and investigation of the biotechnological potential of wild-type *Meyerozyma guilliermondii* associated with amazonian termites able to ferment D-xylose. Biosci. J. 30, 260–266.
Mussatto, S.I., Silva, C.J., Roberto, I.C., 2006. Fermentation performance of *Candida guilliermondii* for xylitol production on single and mixed substrate media. Appl. Microbiol. Biotechnol. 72, 681–686.
Naumann, D., Helm, D., Labischinski, H., 1991. Microbiological characterizations by FT-IR spectroscopy. Nature 351, 81–82.
Nout, M.J., Platis, C.E., Wicklow, D.T., 1997. Biodiversity of yeasts from Illinois maize. Can. J. Microbiol. 43, 362–367.
O'Donnell, K., 1993. Fusarium and its near relatives. In: Reynolds, D.R., Taylor, J.W. (Eds.), The Fungal Holomorph: Mitotic, Meiotic and Pleomorphic Speciation in Fungal Systematics. CAB International, Wallingford, United Kingdom, pp. 225–233.
Pelliccia, C., Antonielli, L., Corte, L., Bagnetti, A., Fatichenti, F., Cardinali, G., 2011. Preliminary prospection of the yeast biodiversity on apple and pear surfaces from Northern Italy orchards. Ann. Microbiol. 61 (4), 965–972.
Perromat, A., Melin, A.M., Lorin, C., Deleris, G., 2003. Fourier transform IR spectroscopic appraisal of radiation damage in *Micrococcus luteus*. Biopolymers 72, 207–216.
Petersson, S., Schnurer, J., 1995. Biocontrol of mold growth in high-Moisture wheat stored under Airtight conditions by *Pichia anomala, Pichia guilliermondii* and *Saccharomyces cerevisiae*. Appl. Environ. Microbiol. 61 (3), 1027–1032.
Pfaller, M.A., Diekema, D.J., Mendez, M., Kibbler, C., Erzsebet, P., Chang, S.C., Gibbs, D.L., Newell, V.A., 2006. *Candida guilliermondii*, an opportunistic fungal pathogen with decreased susceptibility to fluconazole: geographic and temporal trends from the ARTEMIS DISK antifungal surveillance program. J. Clin. Microbiol. 44, 3551–3556.
Prasad, K.N., Agarwal, J., Dixit, A.K., Tiwari, D.P., Dhole, T.N., Ayyagari, A., 1999. Role of yeasts as nosocomial pathogens & their susceptibility to fluconazole & amphotericin B. Indian J. Med. Res. 110, 11–17.
Protchenko, O.V., Boretsky Yu, R., Romanyuk, T.M., Fedorovych, D.V., 2000. Over-synthesis of riboflavin by yeast *Pichia guilliermondii* in response to oxidative stress. Ukr. Biokhim. Zh 72, 19–23.
Richards, G.M., Buck, J.W., Beuchat, L.R., 2004. Survey of yeasts for antagonistic activity against *Salmonella Poona* in cantaloupe juice and wounds in rinds coinfected with phytopathogenic molds. J. Food Prot. 67, 2132–2142.
Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–406.
Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., Bolchacova, E., Voigt, K., Crous, P.W., 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc. Natl. Acad. Sci. 109, 6241–6246.
Schoch, C.L., Robbertse, B., Robert, V., Vu, D., Cardinali, G., Irinyi, L., Meyer, W., Nilsson, R.H., Hughes, K., Miller, A.N., 2014. Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. Database, 2014, bau061.
Seker, E., 2010. Identification of *Candida* species isolated from bovine mastitic milk and their in vitro hemolytic activity in Western Turkey. Mycopathologia 169, 303–308.
Sibirny, A.A., Boretsky Yu, R., 2009. Pichia guillermondii. In: Satyanarayana, T.-K.G. (Ed.), Yeast Biotechnology: Diversity and Applications.
Suh, S.O., Blackwell, M., 2004. Three new beetle-associated yeast species in the *Pichia guilliermondii* clade. FEMS Yeast Res. 5, 87–95.
Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729.
Tanner jr., F.W., 1945. Riboflavin production by *Candida* species. Science 101, 180–181.
Taylor, J.W., 2011. One fungus = one name: DNA and fungal nomenclature twenty years after PCR. IMA Fungus: Glob. Mycol. J. 2, 113.
Tryfonopoulos, G., Thanou, E., Chondropoulos, B., Fraguedakis-Tsolis, S., 2008. mtDNA analysis reveals the ongoing speciation on Greek populations of *Microtus (Terricola) thomasi* (Arvicolidae, Rodentia). Biol. J. Linn. Soc. 95, 117–130.
Wickerham, L.J., 1966. Validation of the species *Pichia guilliermondii*. J. Bacteriol. 92, 1269.
Wickerham, L.J., Burton, K.A., 1954. A clarification of the relationship of *Candida guilliermonii* to other yeasts by a study of their mating types. J. Bacteriol. 68, 594–597.
Yoshikawa, S., Yasokawa, D., Nagashima, K., Yamazaki, K., Kurihara, H., Ohta, T., Kawai, Y., 2010. Microbiota during fermentation of chum salmon (*Oncorhynchus keta*) sauce mash inoculated with halotolerant microbial starters: analyses using the plate count method and PCR-denaturing gradient gel electrophoresis (DGGE). Food Microbiol. 27, 509–514.
Yu, C., Irudayaraj, J., 2005. Spectroscopic characterization of microorganisms by Fourier transform infrared microspectroscopy. Biopolymers 77, 368–377.
Zhao, Y., Tu, K., Su, J., Tu, S., Hou, Y., Liu, F., Zou, X., 2009. Heat treatment in combination with antagonistic yeast reduces diseases and elicits the active defense responses in harvested cherry tomato fruit. J. Agric. Food Chem. 57, 7565–7570.
Zhao, Y., Tu, K., Tu, S., Liu, M., Su, J., Hou, Y.P., 2010. A combination of heat treatment and *Pichia guilliermondii* prevents cherry tomato spoilage by fungi. Int. J. Food Microbiol. 137, 106–110.

Paper IV

Running head: Internal Variability of rDNA Operon

Travel Into the Internal Variability of Cloned rDNA Operon

CLAUDIA COLABELLA[1], LUCA ROSCINI[1], MARIANA TRISTEZZA[2], LAURA CORTE[1], CARLA PERROTTA[3], PATRIZIA RAMPINO[3], GIANLUIGI CARDINALI[1,4]*, AND FRANCESCO GRIECO[2]

[1] *Department of Pharmaceutical Sciences - Microbiology, University of Perugia, Perugia (Italy).*

[2] *Institute of Sciences of Food Production (ISPA), National Research Council (CNR), Lecce (Italy).*

[3] *Department of Biological and Environmental Sciences and Technologies, University of Salento, Lecce (Italy).*

[4] *CEMIN, Centre of Excellence on Nanostructured Innovative Materials, Department of Chemistry, Biology and Biotechnology, University of Perugia, Via Elce di Sotto 8, 06123 Perugia, Italy.*

*__Corresponding Author__: Dr. Gianluigi Cardinali

Dept. of Pharmaceutical Sciences – Microbiology

Borgo 20 Giugno, 74

I – 06121 PERUGIA

ITALY

e.mail: gianluigi.cardinali@unipg.it

phone +39 075 585 6478

fax +39 075 585 6470

ABSTRACT

A yeast strain isolated during a large-scale study on vineyard-associated yeast strains from Apulia (Southern Italy) was subjected to sequence analysis of the large subunit (LSU) and internal transcribed spacer (ITS) domains of its DNA operon encoding for the ribosomal RNA (rDNA). The two molecular marker sequences indicated that this strain could not be attributed to any known species and it was described as the type strain of *Ogataea uvarum* sp.nov. Moreover, the molecular assays showed several secondary peaks in the ITS2 sequence, but not in the LSU D1/D2. In the aim to test whether these peaks were due to the internal heterogeneity of the rDNA operon, the region spanning from ITS1 to LSU D1/D2 was introduced in a mini library and several clones were sequenced separately. The analyses on the internal variants of ITS and LSU showed a significant variability, although within that predictable among different strains of the same yeast species. In this *Ogataea uvarum* sp.nov., ITS was more variable than LSU, especially in the ITS2 region. The heterogeneity revealed by this strain was then judged in the frame of its potential consequence in Next Generation Sequencing-based environmental metagenomic studies, in which the variability among operons can lead to biodiversity overestimation and to incorrect identification at the species level. The above findings are discussed in the light of the diverse analytical approaches for fungi identification based on sequence similarity. The results of this study show that the internal variability of the rDNA operon requires careful consideration before being used in future metagenomic investigations and emphasizes the need of specific models to interpret the concept of fungal species, when the reproductive barriers represented by exclusively sexual reproduction are not present.

KEYWORDS: ITS, rDNA, Variability, Yeast, *Ogataea uvarum*

INTRODUCTION

The ribosomal RNA encoding region (rDNA) is widely recognized as useful for both phylogenetic and species identification, particularly ITS (Internal Transcribed Spacer) has been proposed as a universal "barcode" for Fungi, after a multi-laboratory work (Schoch et al., 2012) and a massive bioinformatics analysis of candidate *loci* (Robert, V., Szöke, S., et al. 2011). This marker increases the possibility already offered by the D1/D2 domain of the Large Subunit (LSU) of the rRNA encoding, previously proposed as species marker sequence (Kurtzman, C.P. and Robnett, C.J. 1998). The rDNA genes are arranged in a large operon with over 100 tandem repeated copies per genome, which have been demonstrated to be somehow heterogeneous (Korabecna, M. 2007, James, S.A., O'Kelly, M.J., et al. 2009). This is due to the presence of different nucleotides in the same position of different copies (Henry, T. et al., 2000) and this is sometimes used as the basis for species discrimination.

However, a considerable problem in using this barcode can often be found in the internal variations among copies, which have been detected for a number of fungal species (Kårén et al. 1997, Aanen et al. 2001, Smith et al. 2007). This evidence makes it fundamental to study a range of ITS region variants among different species (Horton 2002, Nilsson et al. 2008). The question goes beyond the interest in fungi because it involves various phylogenetic groupings, e.g. prokaryotes (Stewart and Cavanaugh 2007), dinoflagellates (Thornhill et al. 2007), mycetes (Connell et al. 2010, Huang et al. 2010, Santos et al. 2010) and different animals (Wörheide et al. 2004, Sánchez and Dorado 2008, Elderkin 2009) .

Internal heterogeneity is considered a transient situation that will be fixed by concerted evolution which will homogenize all copies to the most predominant one (West, C., James, S.A., et al. 2014). This model would predict that the intra-genomic heterogeneity is higher in newly formed species and decreases rapidly (Kobayashi, T. 2011), providing additional phylogenetic information on the history of the strains and of the taxa (West, C., James, S.A., et al. 2014). Concerted evolution is supposed to homogenize the tandem repeats via gene conversion during meiosis (Naidoo, K.,

Steenkamp, E.T., et al. 2013). The fact that this phenomenon is present also in non-sexual fungi, implies that either the homogenization occurred in the early-stage of the species life in which they were still able to sporulate and to have a high efficient gene conversion, or that via the far less frequent mitotic recombination.

During a large-scale study on vineyard-associated yeast strains from Apulia (Southern Italy) (Tristezza et al., 2013), we isolated a strain from "Negroamaro" grape berries, here described as a new yeast species of the genus *Ogataea*. Initially the genus *Ogataea* was proposed on the basis of the type species *Ogataea minuta* (Yamada et al., 1994). To date, more than 17 species have already been described (Ji and Bai, 2008; Limtong et al., 2008; Péter et al., 2007a, b, 2008).

This yeast strain was subjected to sequence analysis of the large subunit (LSU) and internal transcribed spacer (ITS) domains of its DNA operon encoding for the ribosomal RNA (rDNA). On the basis of morphological characteristics and the two molecular marker sequences, this strain could not be attributed to any known species and it was described as the type strain of *Ogataea uvarum* sp.nov. From the time when the genus *Ogataea* was molecular phylogenetically defined by Suh et al. (2006), numerous new species of *Ogataea* have been proposed (Ji and Bai, 2008; Limtong et al., 2008; Péter et al., 2007a, b, 2008). The first description of the genus *Ogataea* produced by Yamada and coworkers (1994) showed the assimilation of the potassium nitrate and the presence of asci containing one to four ascospores of pileiform shape.

Interestingly, the molecular assays of *Ogataea uvarum* sp.nov showed several secondary peaks in the ITS2 sequence, but not in the LSU D1/D2. In the aim to test whether these peaks were due to the internal heterogeneity of the DNA operon encoding for the rDNA, the two domains themselves and about fifty clones from them, derived after PCR amplification, were sequenced.

107 ## MATERIALS AND METHODS

108

109 ### *Grape Sampling and Yeast Isolation*

110 Healthy undamaged Primitivo (*Vitis vinifera*) grape bunches were sampled in a vineyard located
111 at Cutrofiano (Lecce, Southern Italy). Individual grape berries were randomly and aseptically
112 selected from the bunches, to get a 25 g working sample. Epiphytic yeasts were isolated from the
113 sample by washing berries in 250 mL of sterile water on a rotary shaker at 200 rpm for 30 min
114 (Bleve, G., Grieco, F., et al. 2006). The sample was centrifuged at 5000 × g for 10 min and the
115 sediment was recovered and suspended in 1 mL of Yeast Peptone Dextrose medium (YPD; Yeast
116 extract 1%, Peptone 1% and Dextrose 2%, Sigma-Aldrich, USA). Sample dilutions from 10^{-1} to
117 10^{-4} were spread onto YPD agar plates. After incubation at 28°C for 48h yeast colonies were
118 submitted to molecular procedures for identification.

119

120 ### *Enzymatic Activity*

121 Appropriate dilutions of yeast cultures were plated on solid media containing different substrates
122 for the detection of the enzymatic activities. β-glucosidase, aminoacid decarboxylase, protease,
123 pectinase, glucanase and xylanase activity associated with the non-*Saccharomyces* isolates were
124 determined by specific plate assays as previously described (De Benedictis, M., Bleve, G., et al.
125 2011, Tristezza, M., Vetrano, C., et al. 2013). Acetic acid, H_2S and SO_2 productions were
126 determined as described by Belarbi and Lemaresquier (Belarbi, M. and Lemaresquier, M. 1994).

127

128 ### *DNA Extraction and Sequencing*

129 The genomic DNA representative of each single morphology was extracted (Bolano, A., Stinchi,
130 S., et al. 2001, Cardinali, G., Bolano, A., et al. 2001). The isolates were firstly identified according
131 to the length of the rDNA region spanning the 5.8S rRNA gene and flanking the internal transcribed
132 spacers 1 and 2 (De Benedictis, M., Bleve, G., et al. 2011). The ITS region was amplified by
133 polymerase chain reaction (PCR) using ITS1 (5′ TCCGTAGGTGAACCTGCGG 3′) and ITS4 (5′

134 TCCTCCGCTTATTGATATGC 3′) primers, following the procedure described by
135 (Chanchaichaovivat, A., Ruenwongsa, P., et al. 2007, Tristezza, M., Vetrano, C., et al. 2013). The
136 PCR products and their restriction fragments were separated on 1% agarose gels, with 1X TAE
137 buffer (45 mM Tris–borate, 1 mM EDTA, pH 8). After electrophoresis, gels were stained with
138 ethidium bromide (5 µg/mL) and visualized under UV light (300 nm). The D1/D2 regions of the
139 LSU rDNA from the investigated strain were then amplified using NL1 and NL4 primers
140 (O'Donnell, K. 1993, Kurtzman, C.P. and Robnett, C.J. 1998). The PCR conditions were the
141 following: denaturation at 94°C for 4 min; 35 cycles at 94°C for 60 sec, 48°C for 60 sec and 72°C
142 for 1.5 min, with a final incubation at 72°C for 10 min. The final products were analyzed as
143 described above. The rDNA fragment (ca. 1400 bp), that included the ITS1-5.8S rDNA-ITS2
144 regions and the 5-terminal region (ca. 600 bp) of the ribosomal large subunit gene (26S rDNA), was
145 amplified using the ITS1 and NL4 primer pair as described by Alves and colleagues (Alves, A.,
146 Phillips, A.J., et al. 2005). In order to obtain a DNA template suitable for direct sequencing, the
147 PCR products were purified by the PCR Purification Spin Kit (Invitrogen, USA) and quantified by
148 agarose gel analysis. The PCR sequencing mix (final volume, 20 µl) contained 2µl 10X Ready Mix
149 (Applied Biosystems, USA), 4µl 10X reaction buffer, 1µl of 3.2 µM sequencing primer and
150 3ng/100bp amplicon DNA. Reactions were run using a PCR Express System (Hybaid, U.S.A.), for
151 an initial denaturation at 96 °C for 2.5 min and for 25 cycles of 10 s at 96 °C, 10 s at 56 °C, and 4
152 min at 60 °C. After PCR reactions, the sample was purified and then sequenced by the ABI PRISM
153 3130 sequencer (Applied Biosystems, USA). Data output were analyzed by the Chromas program
154 version 1.45 and sequences were identified by a database similarity search in the GENBANK
155 Collection using the BLAST software (http://www.ncbi.nlm.nih.gov/BLAST/).
156
157 ***PCR Product Cloning and Screening***
158 PCR products were cloned in pGEM-T Easy vector (Promega) included in pGEM-T Easy Vector
159 System (Promega), following the supplier's instructions, and ligation reactions (10 µl final volume)

160 were incubated overnight at 4°C. Transformation of *E. coli* DH5α (F− Φ80lacZΔM15 Δ (lacZYA-
161 argF) U169 recA1 endA1 hsdR17 (rK−, mK+) phoA supE44 λ− thi-1 gyrA96 relA1) competent
162 cells was performed using standard procedures and cells were plated onto LB/Ampicillin/IPTG/X-
163 Gal plates. Detection of positive clones was performed by colony PCR. Each reaction (25 µl)
164 contained: Emerald Amp MAX HS PCR Master Mix 2×Premix (Takara) 12.5 µl, M13 forward
165 primer 0.2 µM, M13 reverse primer 0.2 µM, and sterilized distilled water up to 25 µl. Amplification
166 reactions were performed using the following conditions: 2 min at 98°C (1 hold), 10 s at 98°C, 30 s
167 at 55°C and 45 s at 72°C (25 cycles), followed by a final step of 10 min at 72°C. Plasmid DNA was
168 purified using the Eurogold Plasmid Miniprep Kit I (Euroclone, Italy) and the inserted fragment
169 was sequenced by ABI PRISM 3730xl with primer SP6 (5' ATTTAGGTGACACTATAG 3') and
170 T7 (5' TAATACGACTCACTATAGGG 3'), for LSU clones, and M13 forward (5'
171 GTTTTCCCAGTCACGAC 3') and M13 reverse (5' CAGGAAACAGCTATGACC 3'), for ITS
172 clones. Consensus sequences for each strain and trimming of the ends with low sequencing quality
173 were carried out with Geneious R6 (v. 6.17, Biomatters, Auckland, New Zealand,
174 www.geneious.com).
175

176 ***LSU and ITS Phylogenetic Analysis***

177 Alignment of the ITS and D1/D2 domain of the 26S rDNA (LSU) sequences was carried out in
178 Geneious R6 with Geneious Alignment tool (Bast, F. 2013). Distances were inferred in MEGA6
179 (Tamura, K., Stecher, G., et al. 2013) using the Maximum Composite Likelihood method and
180 expressed as number of base substitutions per site. This procedure has been chosen because it
181 assumes equal substitution patterns and rates among lineages and sites, conditions considered
182 appropriate for a recent and ongoing separation phenomenon. Both transitions and transversions
183 were considered. The Neighbour-Joining method (Saitou, N. and Nei, M. 1987), was used to
184 reconstruct the tree with 1000 bootstrap reiterations. Statistical analyses were performed in R

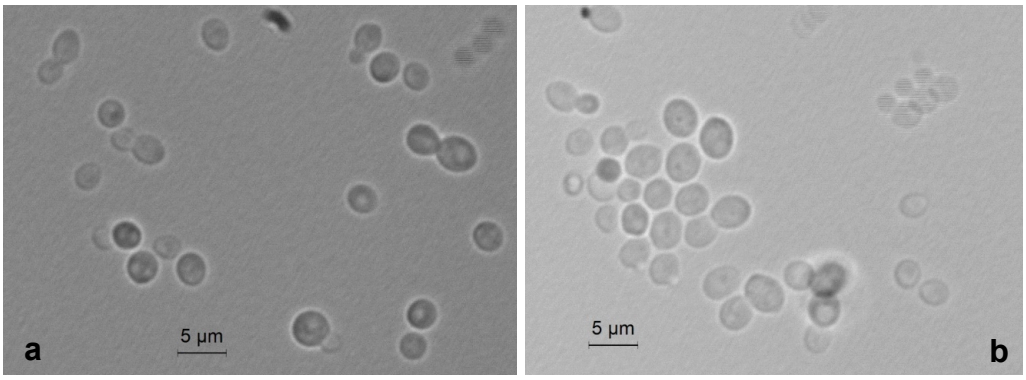
environment (http://www.R-project.org), on the basis of the genetic distances calculated with MEGA6 as described above.

RESULTS AND DISCUSSION

Description of **Ogataea uvarum** *Colabella, Grieco, Corte, Roscini, Cardinali Sp. Nov.*

Ogataea uvarum (*u'va'rum*, *L. n.f.,* pertaining to grapevine, referring to the Latin name of the plant, where the yeast has been isolated the first time). After growth in YM broth at 25° C for 3 days, the cells were elliptic shaped (2-4 x 3-5 µm) and occurred singly or in pairs (Fig. 1).

Figure 1.



Vegetative reproduction occurred by multilateral budding. Sediment was not present. After 5 days at 25 °C on YM agar, streak cultures showed round colonies with regular edges and a matt white color. On Dalmau slide cultures with corn meal agar or rice extract agar after 5 days at 25 °C, pseudomycelium was not formed, neither under the cover glass nor without cover glass. Sporulation doesn't occurred on McClary's acetate agar, Yeast Extract-Malt Extract (YM) agar at 17° C and 25° C after 10 days. Glucose was not fermented.

D-glucose, [α]-trehalose, glycerol, erythritol, D-mannitol, D-sorbitol, glucosamine, D-xylose, D-ribose, adonitol, L-sorbose, ethanol, 2 keto-D-gluconate, nitrate, glucono-δ-lactone, citric acid,

211 methanol and lysine were assimilated. Other carbon compounds tested in this study, including,
212 soluble starch, succinic acid, N-acetyl glucosamine, maltose, nitrite and malic acid were weakly
213 assimilated. D-galactose, sucrose, cellobiose, lactose, melibiose, raffinose, melizitose, inulin, L-
214 arabinose, D-arabinose, L-rhamnose, dulcitol, salicin, Dl-lactic acid, inositol, glucuronic acid, α -
215 methyl-D-glucoside, ethylamine and hexadecane were not assimilated. Growth on 50% glucose and
216 12.5% NaCl were negative.

217 Growth occurred on 5% NaCl, in presence of 0.1, 1 and 10 ppm cycloheximide and weakly on
218 10% NaCl. Growth occurred at 25°C, 37 °C and 42 °C but not at 4°C. No starch-like substance was
219 produced. Urea hydrolysis and Diazonium blue B reaction were negative.

220 Lipase activity was negative. Proteinase activity was weak. Enzyme production assays revealed
221 that this strain was able to decarboxylate histidine and to produce SO_2 and H_2S. It showed β-
222 glucosidase activity on arbutin agar. No xylanase was detected. Moreover, this strain was able to
223 degrade 1,3-β-D-glucan (pachyman) and 1,3:1,4-β-D-glucan (lichenan). Growth carried out on
224 grape-skin and grape-seed agar medium produced dark hazel colonies. Type strain was isolated
225 from grape bunches in a southern Italian region. The culture was deposited in the collection of the
226 Centraalbureau voor Schimmelcultures (CBS), Utrecht (The Netherlands) as CBS 12829, in the
227 Phaff Yeast Culture as UCDFST 14-401, in the Mycoteque de l`Universite Catholique de Louvain
228 (MUCL) collection as MUCL 54959 and in the MycoBank database (MB) as MB 810217.

229 According to LSU and ITS rDNA sequences (deposited in GenBank under the accession
230 numbers reported in Table S1), the new species *O. uvarum* was placed in a well bootstrap supported
231 clade (100%), including members of the genera *Ogataea* and *Candida* globally named *Ogataea*
232 clade (Fig. S1).

233 The closest relatives were *Ogataea philodendra* (9 substitutions equivalent to 1.55%
234 difference), *Ogataea polymorpha* (27 substitutions equivalent to 4.63% difference) *Ogataea*
235 *angusta* (29 substitutions equivalent to 4.98% difference) and *Ogataea dorogensis* (26 substitutions
236 equivalent to 4.47% difference). Members of the clade rather distant to the new species were

Ogataea kodamae, known as a species associated with insects (41 substitutions equivalent to 7.04% difference) (Mikata, K. and Yamada, Y. 1995), *Ogataea naganishii*, a species isolated from plant exudates and rotted logs (50 substitutions equivalent to 8.59% difference) (Kurtzman, C.P. and Robnett, C.J. 2010, Kurtzman, C.P., Fell, J.W., et al. 2011) and *Candida pignaliae*, another yeast species associated with plants (44 substitutions equivalent to 7.56% difference) (Péter, G., Tornai-Lehoczki, J., et al. 2010). The assimilation and fermentation profile of the proposed species differ for several traits from the closest species of the clade (Table 1); in fact, it assimilates 2-Keto-D-Gluconate and does not assimilates D-ribose, D-xylose and ribitol, unlike most members of the clade. It also does not sporulate like *Candida nemodendra* and *Candida pignaliae*.

Table 1. Comparison of the assimilation profile of selected substrates of species phylogenetically close to *Ogataea uvarum*.

	CBS Number	L-Sorbose	D-Glucosamine	D-Ribose	D-Xylose	Ribitol	2-Keto-D-Gluconate	Succinate	Nitrite	Ethylamine	Glucosamine (N)	Growth at 42°C	Spores
O. uvarum	**CBS 12829^T**	**+**	**+**	**-**	**-**	**-**	**+**	**w**	**w**	**-**	**+**	**+**	**-**
O. philodendra	CBS 6075^T	d	-	+	+	+	-	+	+	+	-	-	+
O. minuta	CBS 1708^T	-	-	+	+	+	-	-, +	nd	nd	nd	-	+
O. polymorpha	CBS 4732^I	v	-	+	v	+	-	v	+	+	-	+	+
O. nonfermentans	CBS 5764^I	-		v	v	+	-	+	+	+	-	nd	+
O. naganishi	CBS 6429^I	-	d	+	+	+	-	-	-	+	-	-	+
O. angusta	CBS 7073^T	d	-	+	d	+	-	+	+	-	-	+	+
O. kodamae	CBS 7081^I	-, +	-	+	+	+	-	+	nd	nd	nd	nd	+
O. dorogensis	CBS 9260^I	+	-	+	+	+	-	+	-	nd	nd	nd	+
C. pignaliae	CBS 6071^T	d	-	v	+	+	-	+	+	+	-	-	-
C. nemodendra	CBS 6280^T	+	-	+	+	+	-	+	-	+	-	-	-

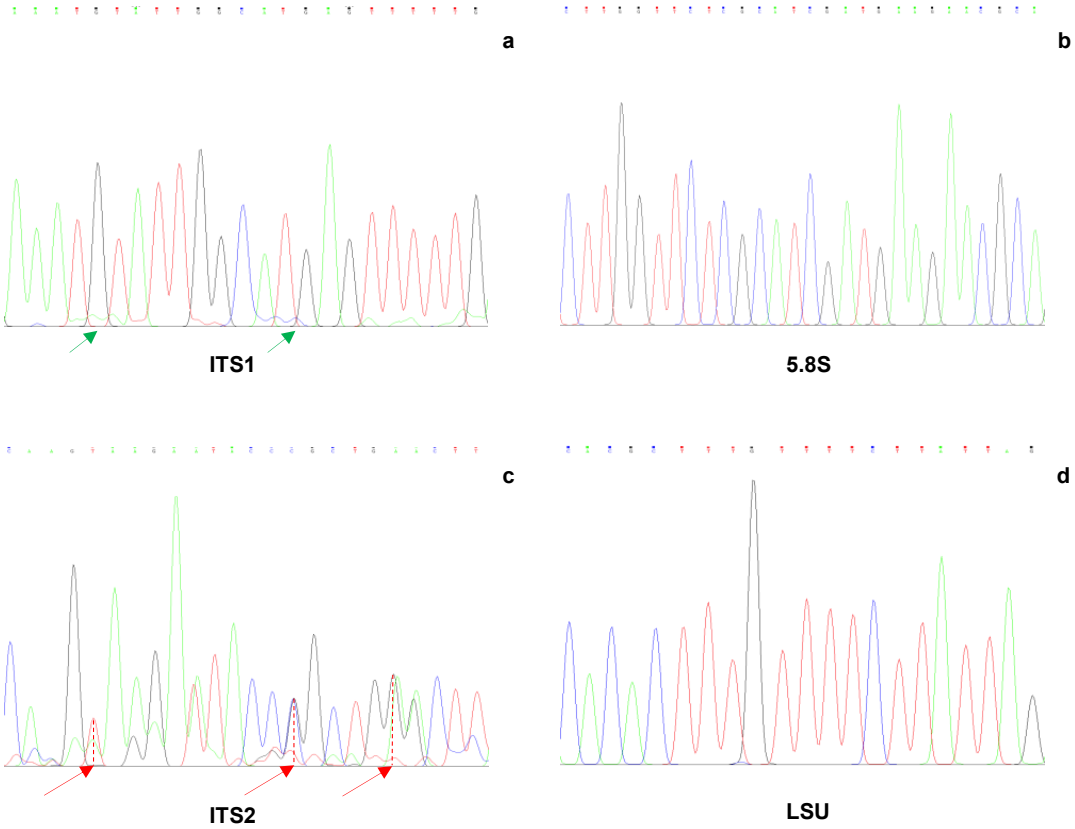
Notes: + = growth, - = no growth, w = weak growth, d = delayed growth, v = variable growth, nd = not determined.

255 ***The Consensus and the Use of Reference Sequences Hide the Variations***

256 The ITS and LSU D1/D2 sequences of the strain CBS 12829^{T} were obtained with standard
257 procedure by Sanger sequencing of the respective amplicons. These sequences will be hereinafter
258 referred to as "reference sequences", in order to distinguish them from the cloned sequences. Their
259 electropherograms displayed some peaks of lower intensity (hereinafter referred to as secondary
260 peaks) exactly below those of normal height, especially in the ITS2 region (Fig. 2).

261

262 **Figure 2.**



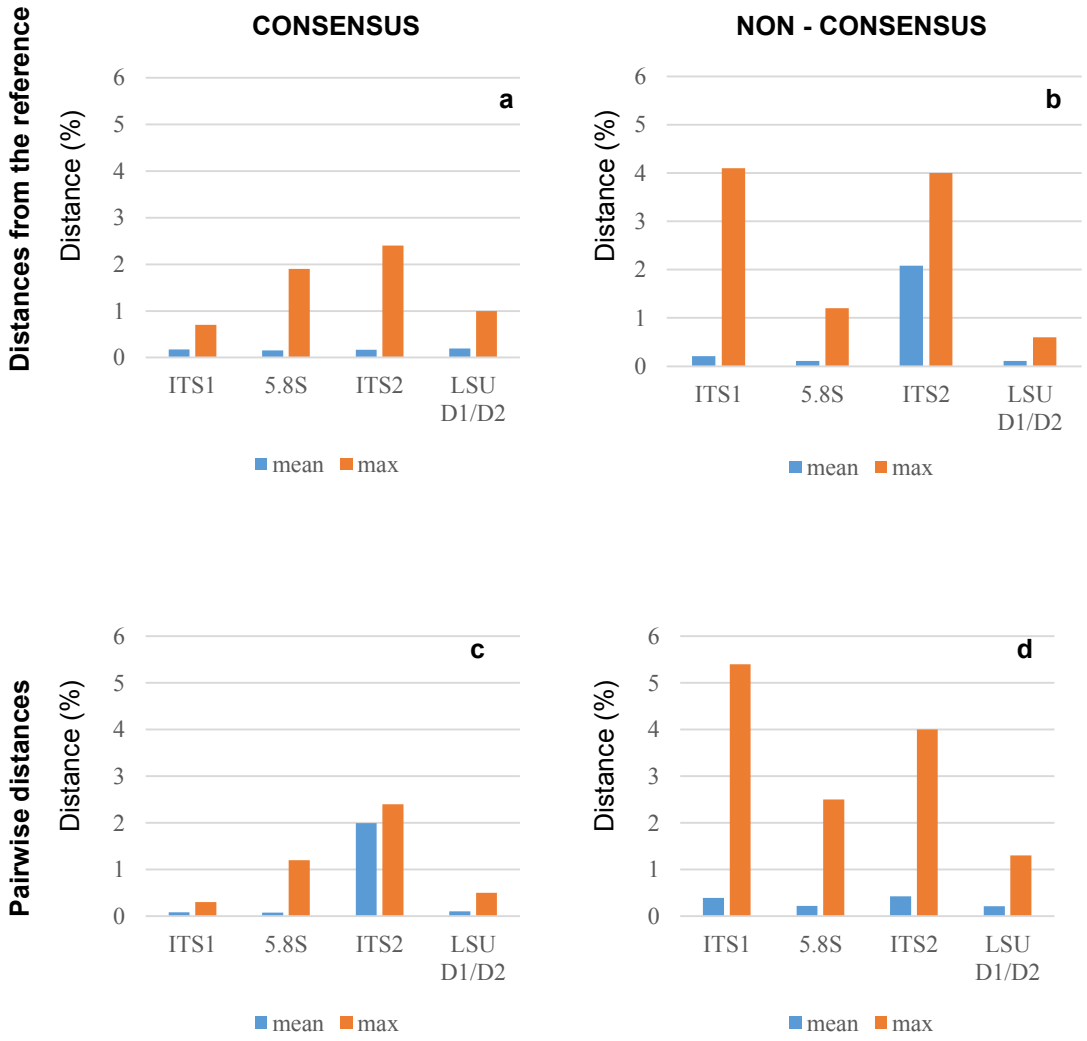
263

264

265 The normal routine with Sanger sequencing is to compare the two strands in order to produce a
266 consensus. During this process, possible variations are resolved in one of the two possible
267 alternatives, perhaps hiding the presence of these internal variants. In some cases, the process is
268 favored by the comparison with the type strain sequence, yielding conservative consensus very
269 similar to that of the type strain, further hiding the level of variation. In order to test the effect
270 caused by consensus sequence, distances of ITS and LSU were calculated from both original and

consensus sequences. The hypothesis on the effect of the use of a reference sequences was simulated by comparing all the pairwise distances among the clones with those of each strain with its reference. When consensus cloned sequences where compared to the reference, all four tested *loci* showed a relatively low mean distances below 1%, whereas the maximum differences ranged from less than 1% (ITS1) to more than 2% (ITS2) confirming the visual inspection of the electropherograms (Fig. 3).

Figure 3.



The use of non-consensus sequences, hereinafter referred to as "original sequences", produced an increase of mean distances from the reference in all *loci* excluded LSU. The ITS1 and ITS2 maxima

283 reached values close to 4% (Fig. 3**b**). When all cloned sequences were compared in a pairwise
284 manner, means and maxima of all *loci* increased (Fig. 3**c** and 3**d**). Once again, the distances from
285 non-consensus sequences increased more than those obtained with consensus sequences, with
286 maxima spanning from 1 to 5% (Fig. 3**d**).
287

288 ***Independence of the Variations among the Four Loci***

289 Since the rDNA operon is constituted by over 100 tandem repeats in the yeast genome
290 (Dammann, R., Lucchini, R., et al. 1995), with some degree of variability already studied with
291 different approaches (James, S.A., O'Kelly, M.J., et al. 2009, West, C., James, S.A., et al. 2014),
292 these secondary peaks were tentatively attributed to the heterogeneity among repeats. In order to
293 test the relative frequency of variant repeats, the ITS-LSU region was cloned and plasmid borne
294 repeats were sequenced separately in both directions and consensus sequences were obtained. This
295 strategy was chosen to determine the actual frequency of variation among repeats and to test
296 whether a relation exists between the variants in the single *loci* (LSU, ITS1, 5.8S and ITS2) within
297 the same tandem repeat copy. In order to determine the correlation among *loci*, the distance
298 between each clone and the reference sequences was calculated for both consensus and original
299 sequences. The variations among the four *loci* showed independence as indicated by Pearson
300 correlation moments close to 0 and very high *p* values (Table 2).
301

302 **Table 2.** Correlation tables among the four *loci* sequences.

a)	ITS1	5.8S	ITS2	LSU
ITS1		0.8066	0.6063	0.0088
5.8S	0.0371		0.9562	0.3376
ITS2	0.0780	0.0090		0.6336
LSU	0.3817	-0.1446	0.0722	

b)	ITS1	5.8S	ITS2	LSU
ITS1		0.0007	0.9419	0.1309
5.8S	0.3590		0.6699	0.1395
ITS2	0.0080	0.0466		0.1375
LSU	0.1642	-0.1607	0.1614	

303

Notes: The lower triangles report the correlations among the distances between the reference sequence and the consensus cloned (**a**) or the original (**b**) sequences. Upper triangles report the *p*-value of the corresponding correlations.

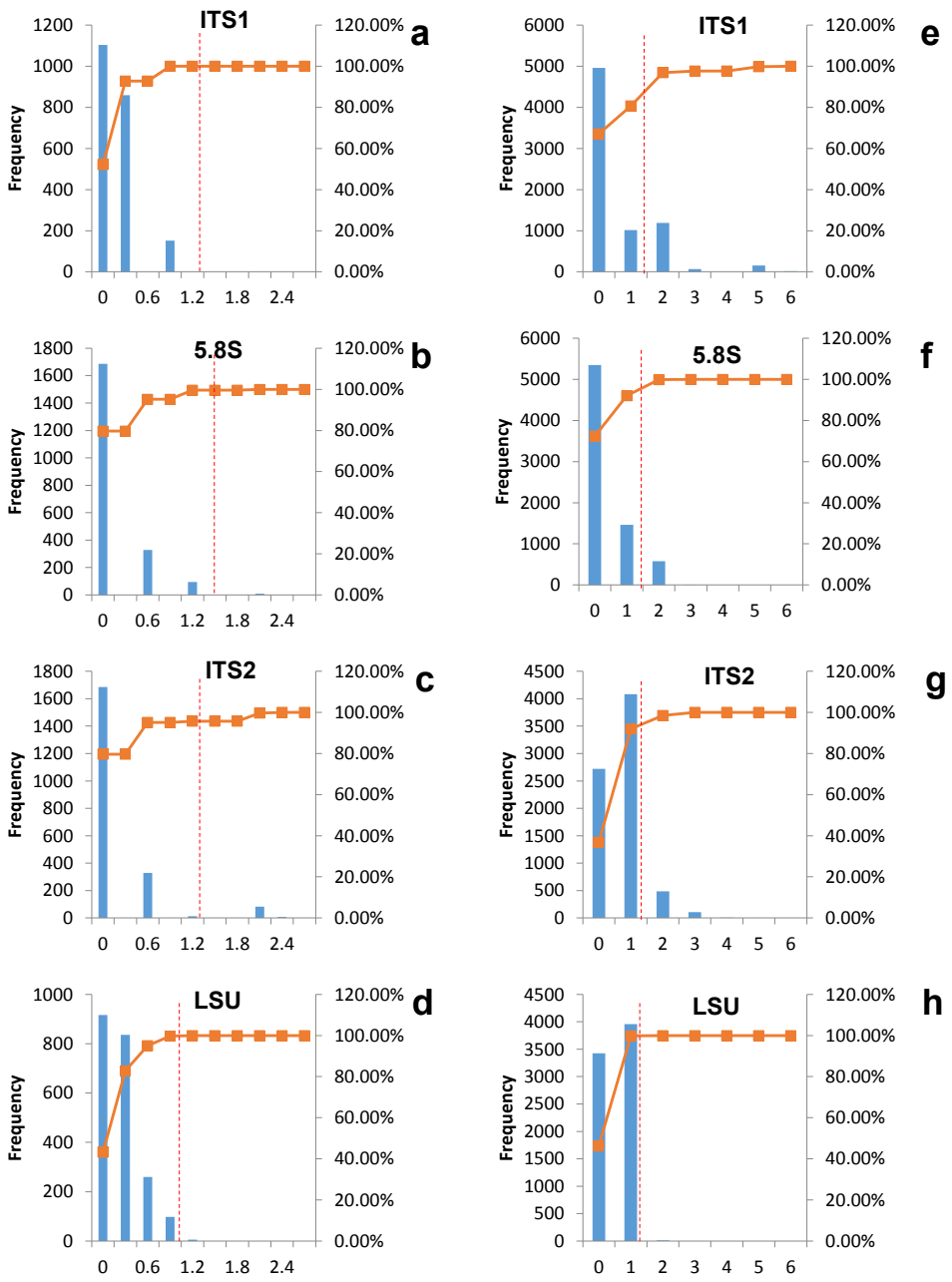
Interestingly, the LSU and the 5.8S *loci* were poorly, but negatively correlated in the two conditions studied (Tab. 2**a** and 2**b**). ITS1 correlated relatively well with the LSU of the consensus (0.381) and with the 5.8S (0.359) of the original sequences, in both cases with an excellent support of the *p* values. Altogether, these data support the idea that the variations occurring within the various regions are independent, although some weak pattern has been detected as the negative correlation between the 5.8S and the LSU *loci*.

Possible Effects of the Heterogeneity on the Identification and on the Biodiversity Estimate in a Metagenomics Scenario

The application of next generation sequencing NGS technologies to metagenomics has opened effective ways for the determination of the microbial diversity, overcoming several drawbacks bound to the cultivation of microbes. ITS is one of the most used sequences for species identification and has been proposed as a universal barcoding marker. LSU was introduced almost two decades ago for species identification and phylogenetic analysis. The application of NGS in metagenomics using these two markers with high copy numbers and high heterogeneity can cause severe over-estimations of the actual biodiversity.

For this reason, an analysis on the distribution of the cloned sequences of the four *loci* was carried out, in order to estimate the effect of these marker sequences in a NGS environment. Values of 1% for LSU (Kurtzman, C.P. and Robnett, C.J. 1998) and 1.4% for ITS (Vu, D., Groenewald, M., et al. 2016) have been suggested as distance thresholds for the species identification.

329 In metagenomic studies carried out with NGS, the single reads would be used for the
330 identification of the species within the sample, creating a question on the possible effects caused by
331 the internal heterogeneity of the studied *loci*. This point was addressed calculating distance matrices
332 of both the original and consensus sequences of the four *loci*, in order to evaluate their distribution.
333 Cloned consensus sequences showed that all the distances calculated follow within the threshold
334 range, with the only exception of 4.25% distances of the ITS2 region (Fig. 4, panels **a** to **d**).
335
336 **Figure 4.**



337

When the original sequences were not subject to the treatment to obtain a consensus, all four *loci* displayed some extent of distances beyond the threshold limits. Namely, 3.14% of ITS1 and 5% of ITS2 distances were larger than 1.4%, whereas these figures were around 1% for LSU (Fig. 4, panels **e** to **h**). In general, the internal heterogeneity of this strain would produce little if any mis-positioning within the *Ogataea* species tree (Fig. S1).

DISCUSSION

This study refers to a single strain of a newly described species, in which the Sanger sequences of ITS and LSU showed a relatively large number of double peaks, suggesting internal heterogeneity among the copies of the DNA encoding for the ribosomal DNA. The cloning of a sample of single copy sequences showed that indeed the internal heterogeneity is present and that the process of generating a consensus hides a large part of it. This finding is in good agreement with that found by James and colleagues about the rDNA sequence variation that exists within individual genomes of 34 *Saccharomyces cerevisiae* strains (James, S.A., O'Kelly, M.J., et al. 2009).

The most concerned *locus* is the ITS2, whereas the LSU and the 5.8S displayed a moderate amount of variability. Cloning the whole region spanning from ITS1 to LSU D1/D2 allowed to compare the level of variability of the single *loci* within each clone, showing a low degree of correlation that suggests, that the variations occur independently among the single *loci* within the same copy.

The impact of NGS in metagenomics studies allows to believe that these *loci* will be increasingly used to describe the species present in the samples and the extent of alpha-diversity. According to our data and analysis, the internal heterogeneity can produce very moderate over-estimations of biodiversity when the LSU D1/D2 *locus* is employed, whereas the use of ITS1, and especially ITS2, would produce more serious overestimates.

As long as fungal taxonomic descriptions will be restricted to isolated strains, this internal heterogeneity is not expected to produce problems of misidentification, neither with Sanger, nor with NGS sequencing. In fact, the former requires a thorough process that purges the consensus sequences from most if not all the effects of the variants. If NGS is applied as an alternative to Sanger to sequence single strain *loci*, the heterogeneity is expected to be displayed, but once again purged by the process of generating a consensus. The real problem is expected to arise when the NGS would be used within metagenomics strategies to explore the vast amount of yet non-described fungal diversity, maybe accounting for some 98% of the total (Taylor, D.L., Hollingsworth, T.N., et al. 2014). Whether other species would show the same extent of the problem or will exhibit different figures is an issue requiring further investigation with more strains and species.

For the current understanding, the internal heterogeneity is a sort of internal noise within otherwise quite similar copies of the rDNA genes. The fact that different species show significantly different sequences of both LSU (Kurtzman, C.P. and Robnett, C.J. 1998) and ITS (Schoch, C.L., Seifert, K.A., et al. 2012) led to their use as taxonomy and barcoding tools. The question on how the various copies change more or less simultaneously in a newly formed species has been long debated and mechanisms of concerted evolution spanning from gene conversion to unequal crossing over have been considered (Nei, M. and Rooney, A.P. 2005). Whereas only few people sustain the hypothesis of birth-and-death model (Rooney, A.P. and Ward, T.J. 2005).

The present study was not intended to elucidate this aspect, but to confirm the presence of an internal variability, partly due to sequences resembling of the phylogenetically close species *Ogataea philodendra* and gives hints to consider this variability when using rDNA in metagenomics studies and in species delimitation analyses.

390 **REFERENCES**

391

392 Alves A, Phillips AJ, Henriques I, Correia A. 2005. Evaluation of amplified ribosomal DNA
393 restriction analysis as a method for the identification of *Botryosphaeria* species. FEMS Microbiol
394 Lett, 245:221-229.

395 Bast F. 2013. Sequence similarity search, multiple sequence alignment, model selection, distance
396 matrix and phylogeny reconstruction. Nature Protocol Exchange.

397 Belarbi M, Lemaresquier M. 1994. La caratterizzazione dei lieviti. Vignevini, 21:57-59.

398 Bleve G, Grieco F, Cozzi G, Logrieco A, Visconti A. 2006. Isolation of epiphytic yeasts with
399 potential for biocontrol of *Aspergillus carbonarius* and *A. niger* on grape. Int J Food Microbiol,
400 108:204-209.

401 Bolano A, Stinchi S, Preziosi R, Bistoni F, Allegrucci M, Baldelli F, Martini A, Cardinali G.
402 2001. Rapid methods to extract DNA and RNA from *Cryptococcus neoformans*. FEMS Yeast
403 Research, 1:221-224.

404 Cardinali G, Bolano A, Martini A. 2001. A DNA extraction and purification method for several
405 yeast genera. Annals of Microbiology, 51:121-130.

406 Chanchaichaovivat A, Ruenwongsa P, Panijpan B. 2007. Screening and identification of yeast
407 strains from fruits and vegetables: Potential for biological control of postharvest chilli anthracnose
408 (*Colletotrichum capsici*). Biological Control, 42:326-335.

409 Dammann R, Lucchini R, Koller T, Sogo JM. 1995. Transcription in the yeast rRNA gene locus:
410 distribution of the active gene copies and chromatin structure of their flanking regulatory sequences.
411 Molecular and Cellular Biology, 15:5294-5303.

412 De Benedictis M, Bleve G, Grieco F, Tristezza M, Tufariello M. 2011. An optimized procedure
413 for the enological selection of non-*Saccharomyces* starter cultures. Antonie Van Leeuwenhoek,
414 99:189-200.

415 James SA, O'Kelly MJ, Carter DM, Davey RP, van Oudenaarden A, Roberts IN. 2009.
416 Repetitive sequence variation and dynamics in the ribosomal DNA array of *Saccharomyces*
417 *cerevisiae* as revealed by whole-genome resequencing. Genome research, 19:626-635.
418 Kobayashi T. 2011. Regulation of ribosomal RNA gene copy number and its role in modulating
419 genome integrity and evolutionary adaptability in yeast. Cellular and Molecular Life Sciences,
420 68:1395-1403.
421 Korabecna M. 2007. The variability in the fungal ribosomal DNA (ITS1, ITS2, and 5.8 S rRNA
422 gene): its biological meaning and application in medical mycology. Communicating current
423 research and educational topics and trends in applied microbiology, 105:783-787.
424 Kurtzman CP, Fell JW, Boekhout T. 2011. The Yeasts: A Taxonomic study. V ed. Amsterdam,
425 Elsevier Science Publishers.
426 Kurtzman CP, Robnett CJ. 1998. Identification and phylogeny of ascomycetous yeasts from
427 analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. Antonie Van
428 Leeuwenhoek, 73:331-371.
429 Kurtzman CP, Robnett CJ. 2010. Systematics of methanol assimilating yeasts and neighboring
430 taxa from multigene sequence analysis and the proposal of *Peterozyma* gen. nov., a new member of
431 the *Saccharomycetales*. FEMS Yeast Research, 10:353-361.
432 Mikata K, Yamada Y. 1995. Ogataea kodamae, a new combination for a methanol-assimilating
433 yeast species, *Pichia kodamae* van der Walt et Yarrow. Res Commun Inst Ferment, 17:99-101.
434 Naidoo K, Steenkamp ET, Coetzee MP, Wingfield MJ, Wingfield BD. 2013. Concerted
435 evolution in the ribosomal RNA cistron. PLoS one, 8:e59355.
436 Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families.
437 Annual review of genetics, 39:121.
438 O'Donnell K. 1993. *Fusarium* and its near relatives. In: Taylor RRaJW editor. The fungal
439 holomorph: mitotic, meiotic and pleomorphic speciation in fungal systematics. Wallingford, United
440 Kingdom, CBA International, p. 225-233.

Péter G, Tornai-Lehoczki J, Dlauchy D. 2010. *Ogataea pignaliae* sp. nov., the teleomorph of *Candida pignaliae*. Int J Syst Evol Microbiol, 60:2496-2500.

Robert V, Szöke S, Eberhardt U, Cardinali G, Seifert KA, Lévesque CA, Lewis CT, Meyer W. 2011. The Quest for a General and Reliable Fungal DNA Barcode The Open Applied Informatics Journal, 5:45-61.

Rooney AP, Ward TJ. 2005. Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. Proceedings of the National Academy of Sciences of the United States of America, 102:5084-5089.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol, 4:406-425.

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W. 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.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol, 30:2725-2729.

Taylor DL, Hollingsworth TN, McFarland JW, Lennon NJ, Nusbaum C, Ruess RW. 2014. A first comprehensive census of fungi in soil reveals both hyperdiversity and fine□scale niche partitioning. Ecological Monographs, 84:3-20.

Tristezza M, Vetrano C, Bleve G, Spano G, Capozzi V, Logrieco A, Mita G, Grieco F. 2013. Biodiversity and safety aspects of yeast strains characterized from vineyards and spontaneous fermentations in the Apulia Region, Italy. Food Microbiol, 36:335-342.

Vu D, Groenewald M, Szöke S, Cardinali G, Eberhardt U, Stielow B, de Vries M, Verkley GJM, Crous PW, Boekhout T*, 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.

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.

492 **FIGURE LEGENDS**

493

494 **Figure 1.** Light microscopic morphology of cells of *Ogataea uvarum* CBS 12829.

495 **Notes: a)** *O. uvarum* cells in YM broth; **b)** *O. uvarum* cells in YEPD medium.

496 **Figure 2.** Internal variability on *Ogataea uvarum*^{T} ITS and LSU sequences.

497 **Notes:** Examples of the variability found on the four barcoding genes. Green arrows identify
498 position with low variability degree; red arrows identify position with high variability degree.

499 **Figure 3.** Mean and maximum distances from reference Sanger sequence of four different
500 analytical settings.

501 **Notes:** The four panels report the distances between the reference Sanger Sequence and **a**) the
502 cloned consensus sequences and **b**) the original (non-consensus) cloned sequences. Panel **c** and **d**
503 report all the pairwise distances between the consensus and the original sequences, respectively.

504 **Figure 4.** Distance distribution of consensus and original cloned sequences.

505 **Notes:** Panels **a-d** report the histogram of the distances between consensus cloned sequences.
506 Panels **e-h** report histogram and accumulation curves of the distances between original cloned
507 sequences. Red dotted lines represent the thresholds suggested for species identification

508

509

510

511

512

513

514

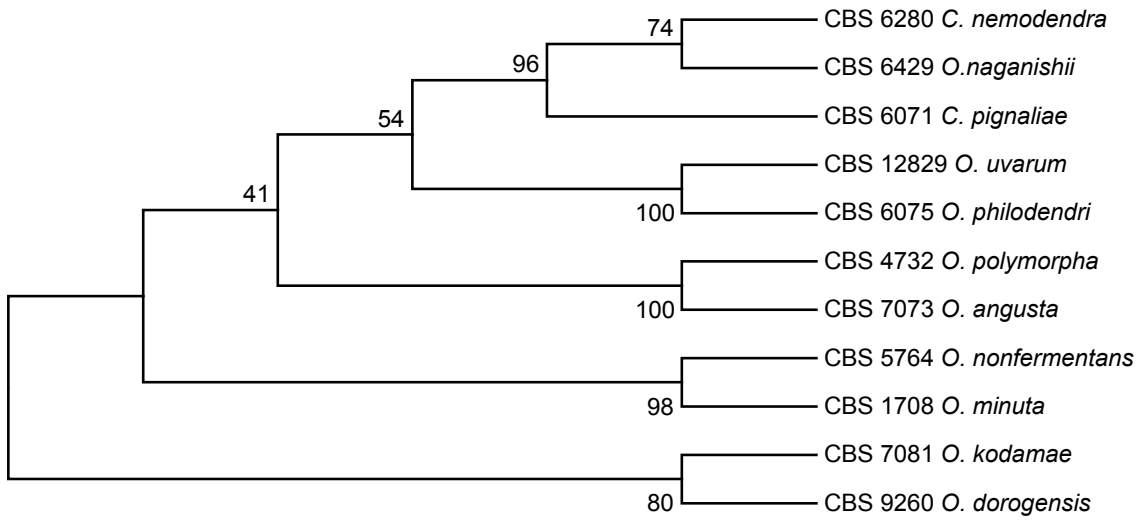
515

516

517

APPENDICES

Figure S1. Evolutionary relationships of 11 taxa related to *O. uvarum*.



Notes: Phylogenetic tree reconstructed using the Maximum Parsimony (MP) method on previously aligned and concatenated ITS and LSU sequences. The optimal tree is shown, with bootstrap support (1000 replicates) shown next to the branches. The MP tree was obtained using the Subtree-Pruning-Regrafting (SPR) algorithm in MEGA6. Type strain sequences were retrieved from GenBank and CBS databases.

Table S1. GenBank accession numbers of the strains belonging to the *Ogataea* clade.

Strain number	LSU sequence ID	ITS sequence ID
CBS 6280	CBS 6280 ex no 10202 lsu	CBS 6280 ex 10202 its
CBS 6429	CBS 6429_ex32166_42146_LSU	cr - CBS 6429
CBS 6071	U70183	cr - CBS 6071
CBS 6075	CBS 6075_ex46790_116657_LSU	cr - CBS 6075
CBS 12829	LN849460	HE965024
CBS 4732	FJ914932	FJ914915
CBS 7073	FJ914931	JF756588
CBS 5764	U75518	cr - CBS 5764
CBS 7081	U75525	cr - CBS 7081
CBS 9260	AF403146	CBS 9260_ex21244_6403_ITS

Paper V

Moving to NGS barcode sequencing for identification and diagnostics, an application in "*Candida*" pathogenic yeasts

Claudia Colabella[1], Laura Corte[1], Luca Roscini[1], Matteo Bassetti[2], Carlo Tascini[3], Joe Mellor[4], Wieland Meyer [5] and Gianluigi Cardinali[1,6*].

[1]Department of Pharmaceutical Sciences, Microbiology section - University of Perugia-Italy; [2]Udine Hospital - Udine - Italy; [3]Cotugno Hospital Napoli - Italy; [4]seqWell, Inc. 376 Hale Street - Beverly; [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, Australia; [6]CEMIN Research Centre of Excellence - University of Perugia - Italy

Running Title: NGS barcoding of *Candida*

Key Words: NGS, Sanger, ITS, LSU, MALDI-TOF, *Candida*.

***Corresponding author:** Dr. Gianluigi Cardinali

Dept. of Pharmaceutical Sciences - Microbiology

Borgo 20 Giugno, 74

I - 06121 PERUGIA (ITALY)

e.mail: *gianluigi.cardinali@unipg.it*

phone +39 075 585 6478; fax +39 075 585 6470

ABSTRACT

Species identification of yeasts and other fungi is currently carried out with Sanger sequences of selected molecular markers, mainly from the operon encoding the ribosomal DNA, characterized by hundreds of tandem repeats of the 18S, ITS1, 5.8S, ITS2 and LSU *loci*. The ITS locus (including ITS1, 5.8S, ITS2) has been proposed as a primary barcode marker making this region the most used one in taxonomy, phylogeny and diagnostics. The introduction of NGS sequencing 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 paper, 286 strains belonging to 11 species of pathogenic yeasts were analysed with MALDI-TOF and NGS sequencing 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.

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 18S, ITS1, 5.8S, ITS2, LSU and 5S (Dujon, Sherman et al. 2004). These arrays vary in number from a few dozen to hundreds (Maleszka and Clark−Walker 1993, Amend, Seifert et al. 2010). In *Candida albicans*, the operon is 12,756 bp long and the diploid genomes contains some 110 repeats in a single locus (Jones, Federspiel et al. 2004), whereas in *C. glabrata* these sequences are dispersed in two subtelomeric regions (Maleszka and Clark−Walker 1993, Dujon, Sherman et al. 2004). The sequences of these genes have been largely used in the last decades in taxonomy and phylogenetic studies thanks to their high conservation (Kurtzman and Robnett 1998, Groenewald, Robert et al. 2011), by means of Sanger sequencing that

53 produces a single sequence of each gene. However, secondary peaks have been observed suggesting
54 some level of heterogeneity among the various copies of the tandem repeats in fungi (Korabecna
55 2007, Woo, Leung et al. 2010, Vydryakova, Van et al. 2012, Li, Sun et al. 2014), ciliates (Gong,
56 Dong et al. 2013) and in some plants (Wang, Ma et al. 2015). The extent of this variability is critical
57 for the exact understanding governing the homogenization of multigene family, typically attributed
58 to concerted evolution by means of gene conversion or asymmetric crossing-over (Liao 1999, Nei
59 and Rooney 2005, Ganley and Kobayashi 2007, Naidoo, Steenkamp et al. 2013). However, the
60 possibility to explain this homogenization of the rDNA loci with birth-and-death model is still
61 matter of debate and some authors claimed that the variation observed fit more to this model than to
62 concerted evolution (Nei and Rooney 2005). In general it is possible that different mechanisms act
63 in different *taxa* and maybe even in different regions (Vydryakova, Van et al. 2012). A mixed
64 model of evolution involving both models simultaneously was considered, although not for this
65 gene family (Nei and Rooney 2005). The major differences between these models are on the fact
66 that concerted evolution is expected to produce scarce heterogeneity, whereas birth-and-death
67 mechanism should yield more Intra-Genomic Polymorphisms (IGP) (Ganley and Kobayashi 2007).
68 From the above observations it looks as if the concerted evolution is not a satisfactory model when
69 IGPs frequency is particularly high (Simon and Weiß 2008).
70 Beyond the model governing the homogenization of the repeat units, the internal variability within
71 the rDNA is a source of additional information useful in phylogenetic, environmental and clinical
72 microbiology to trace the origin of the studied strains (West, James et al. 2014). As long as Sanger
73 sequencing was the sole or predominant technology, the sequence reported the most frequent
74 nucleotides hiding the least frequent, occasionally visible as secondary peaks (Woo, Leung et al.
75 2010).
76 Since ITS has been proposed as universal barcode for fungi (Schoch, Seifert et al. 2012), the
77 possibility of applying Next Generation Sequencing (NGS) for these *loci* offers several advantages
78 such as the possibility of studying microbial communities, independently of their viability and

79 capacity of growing on existing media (Bokulich and Mills 2012, Hajibabaei 2012). Still problems
80 exist in the exact quantification of *taxa* on the basis of NGS reads abundance (Amend, Seifert et al.
81 2010), and care should be taken in data analysis because the internal heterogeneity could cause an
82 inflation of the species richness (Lindner and Banik 2011, Lindner, Carlsen et al. 2013). Since
83 database quality and completeness are mandatory for NGS analyses (Bokulich and Mills 2012), the
84 presence of few alternative barcode markers proposed and under evaluation (Stielow, Lévesque et
85 al. 2015) conveys to use rDNA genes for identification, although a much deeper understanding of
86 the problems and effective analytical pipelines are necessary (Medinger, Nolte et al. 2010).
87 NGS approach is now mature to move from specialized research centres to environmental and
88 clinical laboratories. However the massive amount of data and the internal heterogeneity can be a
89 serious problem to get sound and rapid high-throughput identifications (Ahmed 2016), especially
90 when a more complex metagenomics approach is taken to describe microbial communities in
91 patients and healthy controls. (Imabayashi, Moriyama et al. 2016). Among the various approaches
92 described in literature, BLAST search and assembly followed by BLAST have been recently
93 described and proposed (Ahmed 2016, Imabayashi, Moriyama et al. 2016).
94 In this paper we describe an innovative system of yeast strain identification using next generation
95 sequencing of the amplicons including the region spanning from ITS1 to the D1/D2 domain of the
96 LSU. The rationale of this strategy is that the amplicons includes the three loci (ITS1, 5.8S, ITS2)
97 proposed as universal barcode in fungi (Schoch, Seifert et al. 2012) and now included in highly
98 curated databases (Schoch, Robbertse et al. 2014, Irinyi, Serena et al. 2015). Furthermore, the
99 D1/D2 domain of the LSU was introduced to explore the possibilities offered by NGS in terms of
100 multi-locus sequencing, that can be carried out and analysed easily with the procedure proposed
101 here (Kurtzman and Robnett 2013, Susca, Perrone et al. 2013, Yurkov, Guerreiro et al. 2015). The
102 analysis was performed with 286 strains of pathogenic yeasts isolated from two Italian hospitals,
103 previously studied to show that the success of these strains in the hospital environment is strictly
104 related to their ability to form biofilm (Corte, Roscini et al. 2016). This set of strains is large enough

105 to represent the identification routine occurring in clinical setting as well as in other environmental
106 studies and to pave the way to further studies on the composition of the repeats present in the rDNA
107 region.
108

109 **MATERIALS AND METHODS**

110 **Strains and growth conditions**

111 In this study 286 strains were analysed (Tab. S1), they belong to opportunistic species of *Candida*
112 genus, isolated from two Italian Hospitals (Pisa and Udine). All strains were isolated from patient
113 blood cultures and were included in the Cemin Microbial Collection of the Microbial Genetics and
114 Phylogenesis Laboratory of Cemin (Centre of Excellence on Nanostructured Innovative Materials
115 for Chemicals, Physical and Biomedical Applications - University of Perugia). Over twelve species
116 were isolated in both hospitals, among which four, *C. albicans*, *C. glabrata*, *C. parapsilosis* and *C.*
117 *tropicalis*, represented the vast majority of the isolates. All the strains were stored at -80°C in 17%
118 glycerol immediately after isolation. First step of cultivation was carried out on YEPDA (YEPD
119 added with 1.7% agar) at 37°C, following the current procedures. When a biomass was necessary,
120 the strains were grown in YEPD (Yeast extract 1%, Peptone 1%, Dextrose 1% all products from
121 Biolife - http://www.biolifeitaliana.it/) at 37°C with 150 rpm shaking.
122

123 **MALDI-TOF analysis**

124 *Candida* strains were grown over night in YEPDA (YEPD added with 1.7% agar). For each strain
125 one colony was suspended in 300 µl of distilled and sterilised water. Ethanol absolute, 900 µl (Carlo
126 Erba reagents - http://www.carloerbareagents.com/) was added and mixed carefully, and the sample
127 was centrifuged (13,000 x g for 2 min). Supernatant was discarded and the pellet was dried at 37°C
128 for 5 min. Formic acid (70%; 20 µl) and acetonitrile (20 µl) (Carlo Erba reagents) were added to the
129 pellet, mixed and centrifuged again (13,000 x g for 2 min). The supernatant (1 µl), containing
130 ribosomal proteins, was deposited on the 96 wells MALDI plate and dried at room temperature.

131 Samples were overlaid with 1 µl of matrix solution (Bruker Matrix HCCA α-Cyano-4-
132 hydroxycinnamic acid, Bruker Daltonik, GmbH - http://www.bruker.com/) and then dried at room
133 temperature. For each sample two spots were generated. Measurements were performed with a
134 Microflex mass spectrometer (Bruker Daltonik, Bremen, Germany) using FlexControl software (v.
135 3.4.85 Bruker, Germany). Spectra were recorded in the positive linear mode (laser frequency, 60
136 Hz; ion source 1 voltage, 20 kV; ion source 2 voltage, 18.24 kV; lens voltage, 0,01 kV; mass range,
137 2,000 to 20,000 Da). For each spectrum 200 shots in 40-shot steps from different positions of the
138 target spot (automatic mode) were collected and analysed. Spectra were internally calibrated by
139 using *Escherichia coli* ribosomal proteins (Bruker IVD Bacterial Test Standard). Spectra were
140 imported into BioTyper software (v. 3.1 Bruker, Germany) and analysed by standard pattern
141 matching with default settings, using the database included in the software and regularly updated by
142 the Bruker Company. The accuracy of the results was expressed with scores ranging from 0 to 3.
143 Scores below 1.7 could not be considered a reliable identification, scores ≥ 1.7 were recognized as
144 identification to genus level while scores of ≥ 2.0 were considered useful for species identification.
145 When the scores of two duplicates of a same sample matched exactly the identification was
146 considered correct. Another criterion was based on a difference of at least 0.4 between the first and
147 the second species listed in the score hit list.
148
149 **DNA extraction and molecular techniques**
150 Genomic DNA was extracted as indicated by Cardinali et al (Cardinali, Bolano et al. 2001). ITS1,
151 5.8S, ITS2 rDNA genes and D1/D2 domain of the LSU were amplified with FIREPole® Taq DNA
152 Polymerase (Solis BioDyne, Estonia), using ITS1 (5'-TCCGTAGGTGAACCTGCGG) - NL4
153 (GGTCCGTGTTTCAAGACGG) primers. The amplification protocol was carried out as follows:
154 initial denaturation at 94°C for 3 min, 30 amplification cycles (94°C for 1 min, 54°C for 1 min and
155 72°C for 1 min) and final extension at 72°C for 5 min. Amplicons were subjected to electrophoresis
156 on 1.5% agarose gel (Gellyphor, EuroClone, Italy). Amplicons were sequenced with NGS

157 PlexWell™ technologies (http://www.seqwell.com/) with the same primers used for the generation
158 of the amplicons. The reads of each strain, contained in FASTAq file, were analysed with Geneious
159 R9 (v. 9.1.5, Biomatters, Auckland, New Zealand - http://www.geneious.com/).
160
161 **Bioinformatics analysis**
162 *Mapping against a reference vs de novo assembling*
163 Mapping and assembling analyses were carried out on 12 strains, representative of 6 species
164 showing different number of reads, ranging from 19,388 to 53,497. For the mapping analysis 6
165 ITS_LSU concatenate rDNA sequences of type strains (CBS562 *C. albicans*, CBS138 *C. glabrata*,
166 CBS604 *C. parapsilosis*, CBS10906 *C. orthopsilosis*, CBS94 *C. tropicalis* and CBS2030 *M.*
167 *guilliermondii*) were used as references. The 12 FASTAq files were filtered to remove reads shorter
168 than 140 bp. Contigs were obtained using two algorithms: Bowtie2 (Langmead and Salzberg 2012)
169 (hereinafter referred to as BTL) setting "local", searching only the best match, and BBMap (BBm)
170 (Bushnell 2014) mapping multiple best matching in random mode. Mappings were performed using
171 High Sensitivity mode and no trimming of the sequences. *De novo* assembling was performed using
172 Geneious assembler after testing other algorithms, with High and Low sensitivity without trimming.
173 Contigs identification was carried out with BLAST search using a local library containing 15
174 ITS_LSU concatenate rDNA sequences of *Candida* type strains (CBS562 *C. albicans*, CBS7987 *C.*
175 *dubliniensis*, CBS1795 *C. famata*, CBS138 *C. glabrata*, CBS573 *I. orientalis*, CBS10907 *C.*
176 *metapsilosis*, CBS10906 *C. orthopsilosis* CBS604 *C. parapsilosis*, CBS1010 *C. pararugosa*,
177 CBS613 *C. rugosa*, CBS159 *C. sake*, CBS94 *C. tropicalis*, CBS621 *C. utilis*, CBS6936 *C.*
178 *lusitaniae* and CBS2030 *M. guilliermondii*) and CBS1171 *S. cerevisiae* type strain as outgroup. The
179 highest similarity matches was carried out using Megablast. All output parameters will be discussed
180 in the following paragraphs.
181
182 *Mapping against a reference using large libraries*

183 Mapping procedures were performed using two strains CMC 1793 and CMC 1818 with 21,238 and
184 58,263 reads respectively and three selected and accurate libraries: CBS fungal collection database
185 containing ITS sequences of type strains, CBS ITS_LSU database and the ISHAM database
186 containing ITS sequences of medical related strains. The two FASTAq files were mapped against
187 the different libraries using the two algorithms BTL setting "local" and BBm with High Sensitivity
188 mode and no trimming of the sequences. All output parameters were discussed in the results
189 section.

190

191 *Mapping against a selected library (M1)*

192 All the 286 FASTAq files were filtered to remove reads shorter than 140 bp and were mapped
193 against the local library of 16 ITS_LSU concatenate rDNA sequences of type strains using two
194 algorithms (BTL and BBm) with High Sensitivity mode and no trimming. Results were exported
195 from Geneious R9 software in Microsoft Excel®. With a build in macro six indexes where
196 calculated for both algorithms and data where assembled in order to give an easy taxonomic reading
197 of the results.

198 • *Iread* index is the ratio of the number of the reads attributed to each member of the reference
199 library and therefore indicates the share of reads (R_i) attributable to the species represented
200 by the type strain (SR).

$$Iread = \frac{R_i}{\sum SR}$$

201

202 • *Inuc*: this index is the ratio of the nucleotides mapped with type strain (N_i) on the total
203 number of nucleotides present on the reads of the FASTAq file (SN).

$$Inuc = \frac{N_i}{\sum SN}$$

204

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

$$Icov = \frac{C_i}{\sum C}$$

210

211 • *Iref*: refseq percentage.

$$Iref = \%\, refseq * \frac{N_i}{\sum SN}$$

212

213 • *Isim*: similarity percentage.

$$Isim = \%\, pairwise\ identity * \frac{N_i}{\sum SN}$$

214

215 • *Isyn*: This index represents the ratio of the sum of all the five indexes of one species (ΣIsp)
216 on the sum of the indexes of all the species (ΣItot).

$$Isyn = \frac{\sum Isp}{\sum Itot}$$

217

218 *Mapping against the type strain of the presumptive species and the role of the unused reads (M2-*
219 *M3)*

220 All the 286 FASTAq files were mapped against the ITS_LSU concatenated sequences of the type
221 strains resulting in the first mapping (M1). Mapping procedure was performed using the two
222 algorithms (BTL and BBm) with High Sensitivity mode and no trimming. Results were exported
223 from Geneious R9 software in Microsoft Excel®. The six indexes applied for the first mapping were

224 used in order to give an easy and certain taxonomic meaning of the results. The reads that did not
225 match with sequences of the local library were filtered in order to remove reads shorter than 140 bp
226 and re-mapped against the local library using the same settings and indexes of the first and second
227 mapping (M1-M2).
228
229 **RESULTS**
230 **Mapping against a reference vs *de novo* assembly: different efficiency in terms of time**
231 The reads contained in a FASTAq file can be analysed with the *de novo* assembly or by mapping to
232 a reference approaches, each with a variety of algorithms and settings. The former approach does
233 not theoretically require any *a priori* knowledge. Moreover, it could produce contigs of the reads
234 without the bias due to a possibly wrong reference sequence. On the other hand, assembling
235 thousands of reads deriving from hundreds of repeats without a reference could produce inaccurate
236 assemblies. We carried out a comparison between these two different approaches by comparing the
237 accuracy obtained and the computation time requested. The mapping was carried out using an ad
238 hoc library containing the ITS-LSU region of the type strains of 16 yeast species, most of which are
239 known pathogens. These two analyses were carried out on 12 strains, representative of 6 species
240 and exhibiting different number of reads, ranging from 19,388 to 53,497. The output of these
241 analyses showed that the *de novo* assembly takes much more than the mapping in terms of total
242 time necessary to carry out the operation using an i7 Intel processor with 8Gb Ram and the
243 Geneious 9 interface (Tab. 1). The CPU time necessary for the two types of treatment diverges by
244 two or three orders of magnitude (Fig. 1a). Within the two different approaches, BTL (setting
245 "Local") showed lower processing times than BBm, in fact the former carried out the operation with
246 an average of 0.329 milliseconds per read, vs the 0.544 milliseconds for the latter. More
247 interestingly, the standard deviations of the BTL and BBm treatments were 0.120 and 0.226
248 milliseconds respectively (corresponding to 0.31 and 0.42 variation coefficient).

249 Using the *de novo* approach with the Geneious algorithm, the CPU time required was 282.38 and
250 91.27 milliseconds per read, with the high and low sensitivity settings (HS and LS), respectively.
251 Even in this case a large difference was observed between settings; in fact, the HS had a standard
252 deviation of 91.9 milliseconds (variation coefficient 0.33) and the LS 34.9 standard deviation,
253 corresponding to 0.38 variation coefficient.
254 Altogether, these data indicate that the *de novo* assembly takes much more computational time than
255 the mapping against a reference. The variability observed between the four algorithms
256 performances posed the question on the influence of the number of reads on the operational time
257 required. Surprisingly there was a low correlation between the CPU time and the number of reads:
258 0.67 BTL, 0.56 BBm, 0.52 HS and 0.19 LS. Finally the correlation analysis of the computational
259 time required by the four algorithms showed relatively high correlation values between BTL and
260 BBm (0.741) and between the two *de novo* procedures (0.827) (Fig. 1b). These data indicate that the
261 time required by mapping and assembly are independent, whereas a weak relation exists between
262 the algorithms employed within the same type of approach.
263
264 **Contigs quality obtained with "mapping against a reference" and "*de novo* assembly"**
265 The contigs obtained with the two methods were analysed with a local blasting using an *ad hoc*
266 library containing the type strains of 16 yeast species. Typical results (Fig. 2a and 2b) showed that
267 high levels of nucleotide similarity were obtained between the single contig of mapping with most
268 of the library members. These nucleotide similarities spanned from approximately 80% to the
269 99.6% of the correctly identified species (Fig. 2a). In this analysis no major differences could be
270 observed between the BTL and the BBm algorithms. On the contrary, blasting the several contigs
271 derived from the *de novo* assembly produced several identification with homologies spanning from
272 almost 0% to 13% approximately, whereas the correct species displayed 69.4% and 74.6%
273 homology with the high and low sensitivity algorithms. These types of results deriving from the two
274 approaches were confirmed in all the strains analysed. For the mapping approach the similarities

275 with the correct species were in the range between 97.88% and 99.8%, whereas the second more
276 similar species homology varied between 89.9% and 98.8% using BTL and BBm algorithms (Fig.
277 2c). The blasting of the contigs of the 12 strains obtained with *de novo* assembly gave homologies
278 with the correct species ranging from 52.4% and 87.5% and from 48.6 to 90.5% with the high and
279 the low sensitivity algorithms, respectively (Fig. 2d). The second most similar species showed
280 homologies in the ranges 3.8% - 38% and 3.8% - 34%, respectively with the HS and LS algorithms.
281

282 **Feasibility of "mapping against a reference" with large libraries**

283 From the results shown above, it was clear that the use of *de novo* assembly is time consuming and
284 produced relatively low homologies to the correct species. These two aspects suggested analysis of
285 the data in more detail to determine the possibilities offered by the mapping when the reference is
286 represented by a large library of sequences. Ideally, these reference sequences should contain the
287 sequences from the type strains of all known species, in order to avoid the presence of misidentified
288 strains that would lead not only to a incorrect identification, but to an inflation of misidentifications.
289 For this reason, ad hoc, highly curated libraries are produced and maintained, such as the CBS
290 fungal collection (http://www.westerdijkinstitute.nl/Collections/Biolomics) fungal reference library
291 within the NIH-GenBank (Schoch, Robbertse et al. 2014), or the dedicated ITS library of ISHAM
292 for medical identifications (Irinyi, Serena et al. 2015).
293 In order to test the efficacy of mapping against a reference library using large collections of
294 sequences, we used three libraries of different size: i. a curated library of type strains from CBS
295 containing only ITS sequences, ii. a curated CBS library with both ITS and LSU D1/D2 sequences
296 and the ISHAM database. The strains CMC 1793 and CMC 1818 with 21,238 and 58,263 reads
297 respectively were mapped against the three libraries using BTL and BBm algorithms. This scheme
298 produced twelve mapping combinations that were tested in order to define the feasibility of using
299 large libraries and the performance parameters.

300 Using an i7 Intel processor with 8Gb Ram and the Geneious 9 interface the minimum CPU time
301 was 13.59 seconds and the maximum 393 seconds. The average time to mapping CMC 1793 was
302 52.65 sec., whereas CMC 1818 (with almost three times more reads) required an average time of
303 131.02 sec. The time performances of the three libraries varied as expected with their size expressed
304 as number of sequences. The relatively small ISHAM-ITS library required an average of 22.31 sec.,
305 whereas the average CPU times of the two CBS libraries were 61.43 and 191.75 seconds for the
306 CBS-ITS and CBS-ITS-LSU.
307 The time performance of the two tested algorithms varies according to the size of both the library
308 and the FASTAq file (Fig. 3a) BBm was faster than BTL, especially when processing the large
309 CMC 1818 file with over 58,000 reads, whereas it was slightly slower with the smaller FASTAq
310 file from CMC 1793. The processing rate was obviously conditioned by the library size as well.
311 Both algorithms showed the largest reads/second values with the ISHAM ITS and the smallest with
312 the very large CBS-ITS-LSU library. Taking together these observations, we tested the hypothesis
313 that the reads/sec. processing rate may be function of the number of reads of the FASTAq file and
314 the number of sequences of the library used. The regression analysis, carried out for BTL and BBm
315 separately, yielded 0.9702 and 0.8354 R2 values for BTL and BBm, respectively (Fig. 3b). Taken
316 together, these data indicate that mapping against a reference is feasible in terms of time and that it
317 is more convenient than *de novo* assembly, even if large libraries and FASTAq files are employed.
318 This test showed that when using a large library as reference, homology values ranging from, 60%
319 to 100% (Tab. 2), indicating that a careful analytical protocol is necessary to discriminate the
320 taxonomically positive identifications.
321
322 **A pipeline to optimize the mapping against a reference**
323 The tests described before showed that, even using large libraries, mapping is faster than *de novo*
324 assembling. Furthermore, the levels of homology typically found with Sanger sequencing are closer
325 to the data yielded by mapping than those produced by *de novo* assembling. These characteristics

326 suggested to develop a pipeline to take into consideration all aspects emerging from the tests
327 showed above and accounting for the typical conditions in which identifications are carried out. The
328 first step is mapping against a reference (mapping M1) carried out to indicate the most likely
329 species to which the strain belongs to. Since mappings can be used with a wide range of parameters
330 and their examination is quite difficult especially when the libraries and the FASTAq files are large,
331 we developed a series of indexes where a final synthetic index *Isyn* is calculated (Tab. 3a and 3b).
332 These indexes are reported as percentages, for easier reading. All the calculated indexes are
333 consistently higher with the correct species (*C. glabrata* in the case shown in Tab. 3a) as indicated
334 by standard deviations not higher than 1.5%. The reads of the conserved sequences are often very
335 similar to more than one member of the library. This produces a biased decrease of all indexes of
336 the species to whom the unknown strain should be attributed. This means that when the library
337 increases in size, and includes a large number of entries similar to the strain under identification, all
338 indexes and the *Isyn* will show relatively low values. The M1 mapping does not yield a definitive
339 identification at the species level, but rather gives an indication of the most likely species and the
340 type strain that should be used in the next M2 mapping as reference. M2 produces the same
341 parameters as M1 and normally sets aside a relatively number of "unused reads" that usually ranges
342 from 5% to over 30% approximately. Some of these reads were highly homologous (i.e. > 98%
343 homology) to the rDNA of other species. These considerations led us to propose a third mapping
344 (M3) similar to M1 in which all the unused reads are mapped against the same selected library. To
345 ease the reading of the output parameters the indexes were calculated on the outputs of M2 and M3
346 jointly (Tab. 3b). The major difference between the M1 and the M2-M3 mappings is that the
347 conserved sequences are attributed to the most likely species in the latter, whereas are distributed
348 randomly and evenly in the former. The whole M2-M3 procedure led to homology values
349 comparable to those usually observed with the Sanger sequencing.
350 These findings suggested the pipeline depicted in Fig. 4, in which the preliminary attribution to one
351 species is carried out with M1 or with any other presumptive identification. If the M2-M3 mapping

352 does not produce a high level of homology with the type strain of a known species, the
353 identification should be questioned and the possibility of describing a new species should be
354 considered. The residual unused reads after the M2-M3 mapping, can be considered as a mere
355 background noise or subjects of further investigation.
356
357 **Validation of the whole procedure with a large group of strains**
358 The described procedure was tested with a set of 286 strains, isolated as opportunistic pathogenic
359 yeasts from two Italian hospitals. Both BTL and BBm algorithms were used. Since the yeasts were
360 supposedly part of the known pathogenic yeasts attributed to the genus *Candida*, a restricted
361 reference library of the 16 type strains was employed. The whole analyses produced the output
362 values with the members of the library that led to the calculation of the indexes including *Isyn* (Tab.
363 3). For simplicity of language, the identification characterized by the highest *Isyn* will be hereinafter
364 referred to as "correct identification", whereas the others as "incorrect identifications".
365 It must be highlighted here that the *Isyn* value gives an overview of the homology of the reads in a
366 FASTAq file with the members of the library. Even a high *Isyn* value does not preclude that some
367 of the reads showed a high homology with another type strains that does not represents the species
368 which the unknown strain belongs to. This is, for instance, the case of the *C. glabrata* CMC 1912
369 strain, which included some 2.86% of the reads with over 98% of homology with *C. albicans*.
370 Results of the BTL mapping showed that the majority of the correct identifications ranged from
371 80% to 95% and no strain showed *Isyn* higher that 95% with the M1 mapping. The majority of the
372 M2-M3 mapping was between 95% and 100% with only 10% of the strains showing less than 95%
373 *Isyn* (Fig. 4a). Similar results were obtained with the BBm algorithm, although some 4% of the
374 strains had an *Isyn* higher than 95% and only some 5% of the strain had less than 95% *Isyn* with the
375 M2-M3 mapping (Fig. 4b).

376 The distribution of the "incorrect identification" was studied with both algorithms showing a
377 decrease of their maximum frequency from *Isyn* = 20% to *Isyn*=5% respectively with the M1 and
378 M2-M3 mappings (Fig. 4c and 4d).
379 One of the major differences between the M1 and the M2-M3 mapping is that in the former most of
380 the reads display very high homology (> 95%), whereas in the latter this frequency decreases to
381 around 5% (Tab. 3a and 3b). This phenomenon was observed in all the analysed strains and
382 indicates that M1 alone can produce highly biased results, especially if the homology (e.g. pairwise
383 identity) is used directly without any other correction. Altogether, it seems that the *Isyn* from the
384 M2-M3 mapping produces data reliable and quite comparable to the Sanger homology levels used
385 typically by taxonomists when identifying strains.
386 The reads with high homology (> 97%) to one of the library's type strains were 23.96% with BTL
387 Local and 18.31% with BBm. These relatively high frequencies could be due to the reciprocal
388 similarity among the members of the library, because similar reference would share most of the
389 common conserved regions. In order to test this hypothesis, we analysed the behaviour of the strains
390 belonging to the sister species *C. parapsilosis* and *C. orthopsilosis.* These strains had 23.97% and
391 20.64% reads with < 97% of homology, when using BTL and BBm, respectively. This information
392 indicates that the "incorrect identifications" would be increased by libraries containing highly
393 related type strains, when using the M1 mapping.
394 These analyses showed that the sole application of the mapping M1 produced a high rate of
395 incorrect identifications, generating an over 20% overestimate of false positive identifications, that
396 can be effectively corrected by the M2-M3 mapping.
397
398 **DISCUSSION**
399 The increasing expansion of Next Generation Sequencing and the unparalleled wealth of output
400 reads matching with rapidly decreasing prices, call for a consideration of this technology in the
401 identification of fungal strains. The cost and the throughput possibilities of NGS are already more

402 competitive than the traditional Sanger sequencing in many analytical settings. Our initial
403 hypothesis was that, in the case of multigene families, the sequencing depth obtainable with
404 amplicons based-NGS could produce information on the real extent of the internal heterogeneity,
405 possibly masked by the peculiarities of Sanger sequencing. Considering the number of reads
406 deriving from a single NGS analysis, and the relatively short sequences of the amplicons, this
407 technique has the potential for multiplexing. The technique used in this paper allowed sequencing
408 of two different marker genes LSU and ITS that represent a low level multiplexing, but the depth
409 obtained indicated that there is room for more markers, especially if not represented by multi-copy
410 genes, as those used here. Another application of NGS is the direct exploration of the environmental
411 biodiversity, taking a metagenomic, culture-independent approach. One of the aims of this paper
412 was to explore the heterogeneity of these popular markers at the single strain level to assess the
413 potential effect of the rDNA heterogeneity when the template DNA derives from a variety of strains
414 and species.
415 The results showed in this paper indicate that the relatively high heterogeneity present can hamper
416 the use of these sequences for the identification of single strains and even more of complex
417 microbial mixtures.
418 These observations point out that the metagenomics studies could be affected by species inflection
419 at levels higher than currently expected (Lindner, Carlsen et al. 2013). However, the possibilities
420 offered by current NGS techniques, and their future developments, promise to shed more light on
421 the rDNA composition and to transform its internal variability in a powerful tool.
422

423 **REFERENCES**

424 Ahmed, A. (2016). "Analysis of Metagenomics Next Generation Sequence Data for Fungal ITS Barcoding:
425 Do You Need Advance Bioinformatics Experience?" Frontiers in Microbiology **7**: 1061.
426 Amend, A. S., et al. (2010). "Quantifying microbial communities with 454 pyrosequencing: does read
427 abundance count?" Molecular Ecology **19**(24): 5555-5565.

428 Bokulich, N. A. and D. A. Mills (2012). "Next-generation approaches to the microbial ecology of food
429 fermentations." BMB reports **45**(7): 377-389.
430 Bushnell, B. (2014). BBMap: A Fast, Accurate, Splice-Aware Aligner, Ernest Orlando Lawrence Berkeley
431 National Laboratory, Berkeley, CA (US).
432 Cardinali, G., et al. (2001). "A DNA extraction and purification method for several yeast genera." Annals of
433 Microbiology **51**(1): 121-130.
434 Corte, L., et al. (2016). "Exploring ecological modelling to investigate factors governing the colonization
435 success in nosocomial environment of Candida albicans and other pathogenic yeasts." Scientific Reports **6**:
436 26860.
437 Dujon, B., et al. (2004). "Genome evolution in yeasts." Nature **430**(6995): 35-44.
438 Ganley, A. R. and T. Kobayashi (2007). "Highly efficient concerted evolution in the ribosomal DNA repeats:
439 total rDNA repeat variation revealed by whole-genome shotgun sequence data." Genome research **17**(2):
440 184-191.
441 Gong, J., et al. (2013). "Extremely high copy numbers and polymorphisms of the rDNA operon estimated
442 from single cell analysis of oligotrich and peritrich ciliates." Protist **164**(3): 369-379.
443 Groenewald, M., et al. (2011). "The value of the D1/D2 and internal transcribed spacers (ITS) domains for
444 the identification of yeast species belonging to the genus Yamadazyma." Persoonia **26**: 40-46.
445 Hajibabaei, M. (2012). "The golden age of DNA metasystematics." Trends in genetics **28**(11): 535-537.
446 Imabayashi, Y., et al. (2016). "Molecular analysis of fungal populations in patients with oral candidiasis
447 using next-generation sequencing." Scientific Reports **6**: 28110.
448 Irinyi, L., et al. (2015). "International Society of Human and Animal Mycology (ISHAM)-ITS reference
449 DNA barcoding database—the quality controlled standard tool for routine identification of human and
450 animal pathogenic fungi." Medical Mycology: myv008.
451 Jones, T., et al. (2004). "The diploid genome sequence of Candida albicans." Proceedings of the National
452 Academy of Sciences of the United States of America **101**(19): 7329-7334.
453 Korabecna, M. (2007). "The variability in the fungal ribosomal DNA (ITS1, ITS2, and 5.8 S rRNA gene): its
454 biological meaning and application in medical mycology." Communicating current research and educational
455 topics and trends in applied microbiology **105**: 783-787.

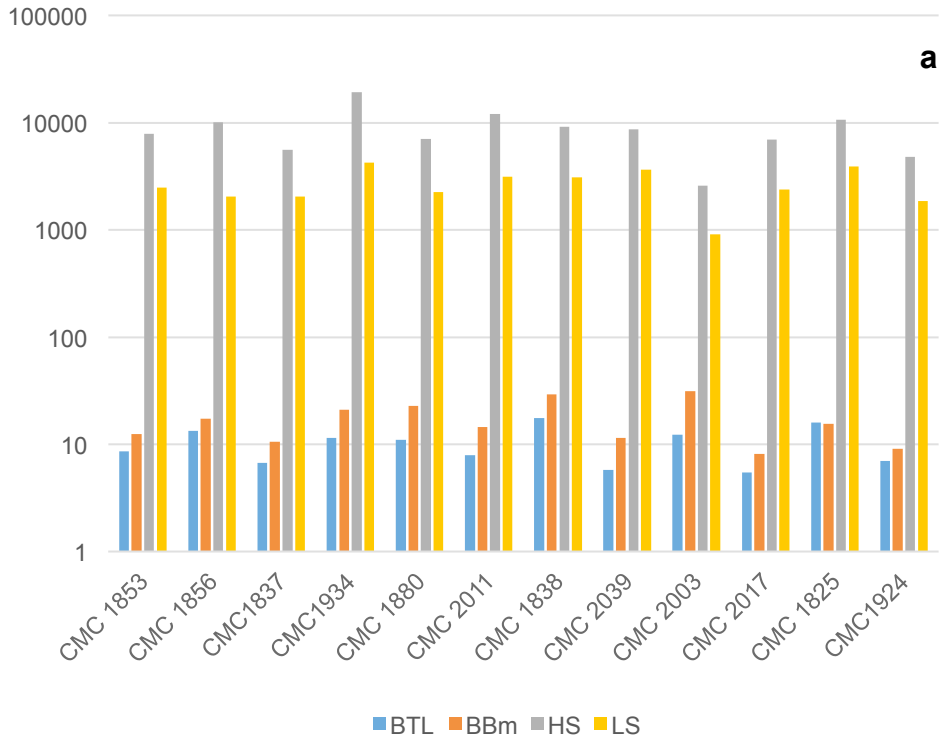
456 Kurtzman, C. P. and C. J. Robnett (1998). "Identification and phylogeny of ascomycetous yeasts from
457 analysis of nuclear large subunit (26S) ribosomal DNA partial sequences." Antonie Van Leeuwenhoek **73**(4):
458 331-371.
459 Kurtzman, C. P. and C. J. Robnett (2013). "Relationships among genera of the Saccharomycotina
460 (Ascomycota) from multigene phylogenetic analysis of type species." FEMS yeast research **13**(1): 23-33.
461 Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nature methods **9**(4):
462 357-359.
463 Li, W., et al. (2014). "The heterogeneity of the rDNA-ITS sequence and its phylogeny in Rhizoctonia
464 cerealis, the cause of sharp eyespot in wheat." Current genetics **60**(1): 1-9.
465 Liao, D. (1999). "Concerted evolution: molecular mechanism and biological implications." The American
466 Journal of Human Genetics **64**(1): 24-30.
467 Lindner, D. L. and M. T. Banik (2011). "Intragenomic variation in the ITS rDNA region obscures
468 phylogenetic relationships and inflates estimates of operational taxonomic units in genus Laetiporus."
469 Mycologia **103**(4): 731-740.
470 Lindner, D. L., et al. (2013). "Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in
471 the internal transcribed spacer rDNA region in fungi." Ecology and evolution **3**(6): 1751-1764.
472 Maleszka, R. and G. Clark−Walker (1993). "Yeasts have a four−fold variation in ribosomal DNA copy
473 number." Yeast **9**(1): 53-58.
474 Medinger, R., et al. (2010). "Diversity in a hidden world: potential and limitation of next−generation
475 sequencing for surveys of molecular diversity of eukaryotic microorganisms." Molecular Ecology **19**(s1):
476 32-40.
477 Naidoo, K., et al. (2013). "Concerted evolution in the ribosomal RNA cistron." PLoS one **8**(3): e59355.
478 Nei, M. and A. P. Rooney (2005). "Concerted and birth-and-death evolution of multigene families." Annual
479 review of genetics **39**: 121.
480 Schoch, C. L., et al. (2014). "Finding needles in haystacks: linking scientific names, reference specimens and
481 molecular data for Fungi." Database- the journal of biological database and curation **2014**.
482 Schoch, C. L., et al. (2012). "Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA
483 barcode marker for Fungi." Proc Natl Acad Sci U S A **109**(16): 6241-6246.

484 Simon, U. K. and M. Weiß (2008). "Intragenomic variation of fungal ribosomal genes is higher than
485 previously thought." Molecular biology and evolution **25**(11): 2251-2254.
486 Stielow, J., et al. (2015). "One fungus, which genes? Development and assessment of universal primers for
487 potential secondary fungal DNA barcodes." Persoonia-Molecular Phylogeny and Evolution of Fungi.
488 Susca, A., et al. (2013). "Multilocus sequence analysis of Aspergillus Sect. Nigri in dried vine fruits of
489 worldwide origin." International journal of food microbiology **165**(2): 163-168.
490 Vydryakova, G. A., et al. (2012). "Intergenomic and intragenomic ITS sequence heterogeneity in
491 Neonothopanus nambi (Agaricales) from Vietnam." Mycology **3**(2): 89-99.
492 Wang, W., et al. (2015). "Astonishing 35S rDNA diversity in the gymnosperm species Cycas revoluta
493 Thunb." Chromosoma: 1-17.
494 West, C., et al. (2014). "Ribosomal DNA sequence heterogeneity reflects intraspecies phylogenies and
495 predicts genome structure in two contrasting yeast species." Systematic biology **63**(4): 543-554.
496 Woo, P. C., et al. (2010). "Internal transcribed spacer region sequence heterogeneity in Rhizopus
497 microsporus: implications for molecular diagnosis in clinical microbiology laboratories." Journal of clinical
498 microbiology **48**(1): 208-214.
499 Yurkov, A., et al. (2015). "Multigene assessment of the species boundaries and sexual status of the
500 basidiomycetous yeasts Cryptococcus flavescens and C. terrestris (Tremellales)." PloS one **10**(3): e0120400.
501
502
503
504
505
506
507
508
509
510

511 **FIGURE CAPTIONS.**

512 **Figure 1.** Evaluation of the computational time requested for the two different approaches

513 (mapping against a reference and *de novo* assembly).



b

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

514

515

516 **Legend.** CPU time needed by the two different types of procedures with the four settings (**a**);

517 Correlation between the four different algorithms (**b**).

518

519

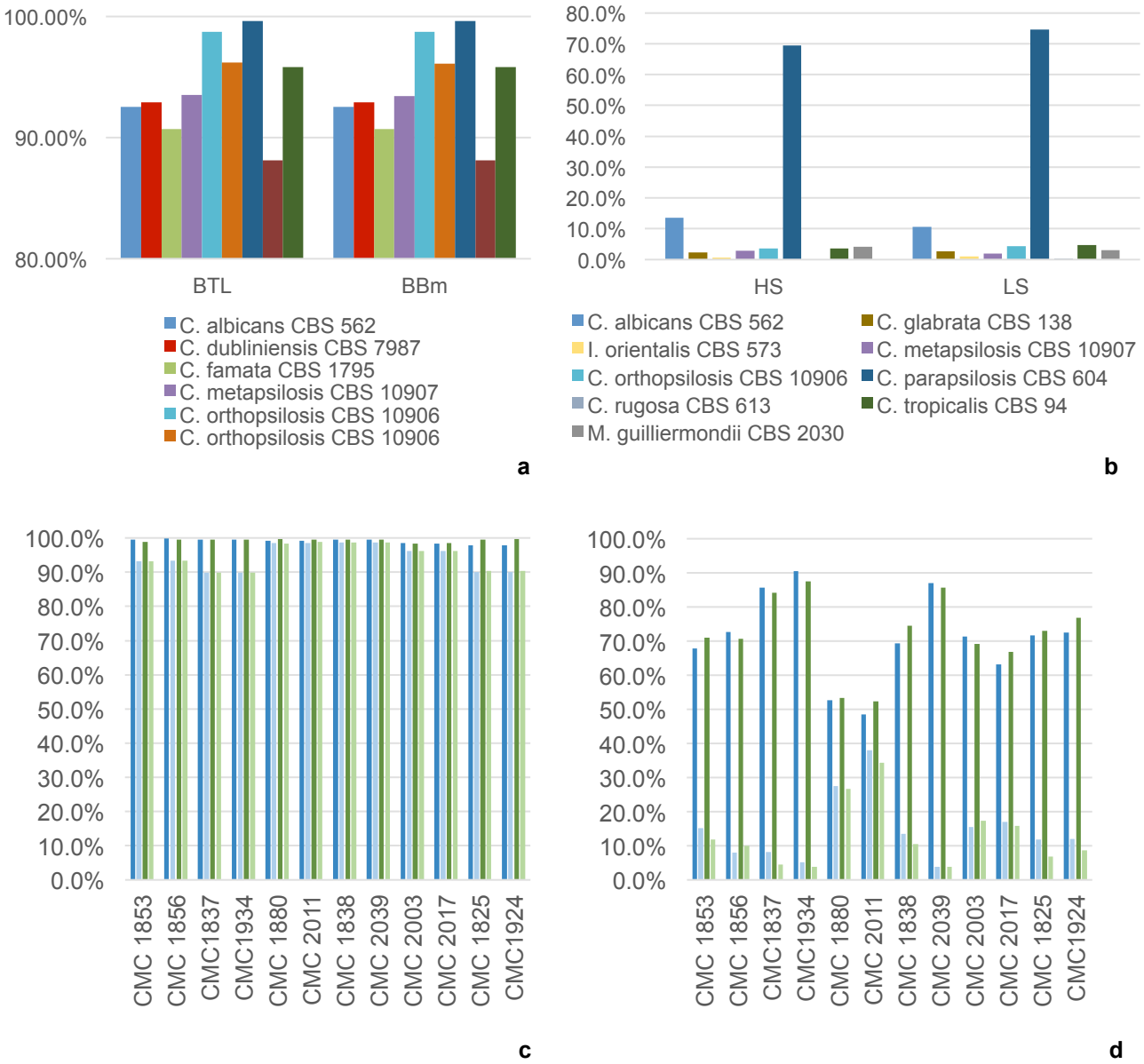
520

521

522

523

524 **Figure 2.** Analysis of contigs quality.



525

526

527 **Legend.** Similarity between a single contig of each of the two mapping algorithms and the

528 members of the reference library (**a**); Variation of the similarity with the members of the library

529 using High Sensitivity (HS) or Low Sensitivity (LS) algorithms (**b**); Homology of the contigs with

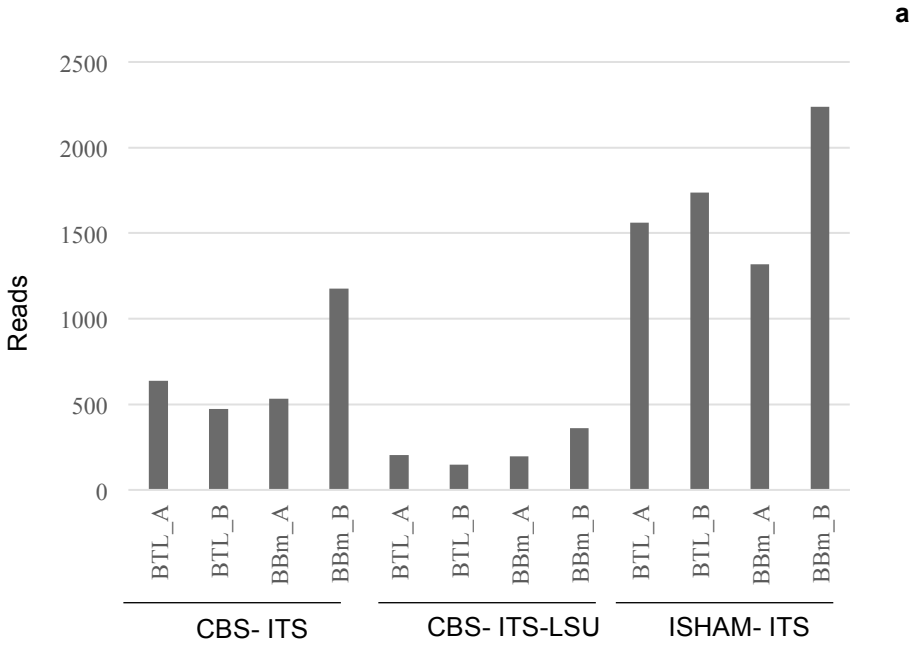
530 the first and the second most similar species using BTL (dark-light blue) and BBm (dark-light

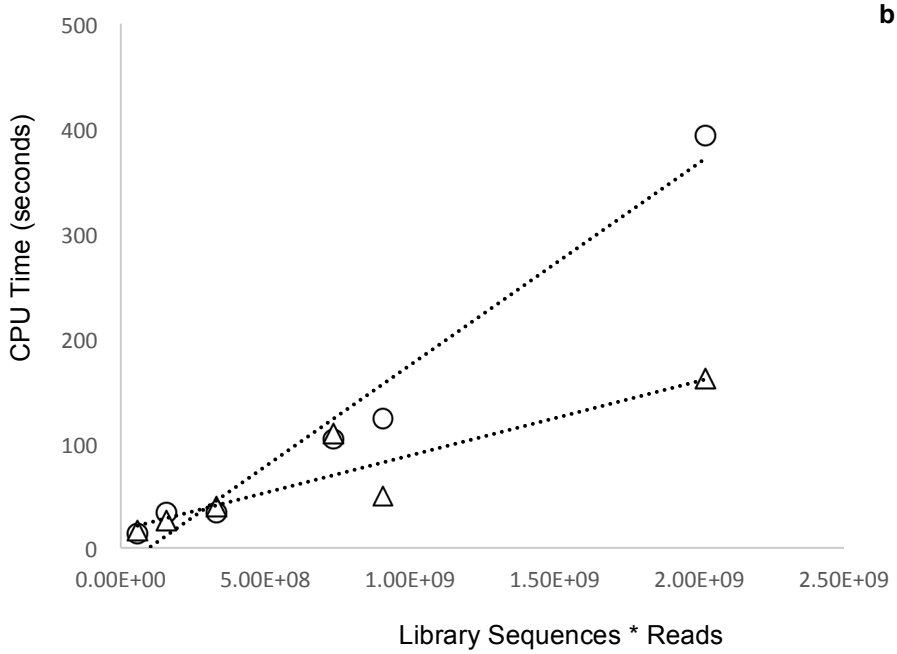
531 green); (**c**) Homology of the contigs analysed with High Sensitivity (dark-light blue) or Low

532 Sensitivity (dark-light green) algorithms (**d**).

533

534 **Figure 3.** Time performance of the mapping algorithms against three large libraries.





535

536

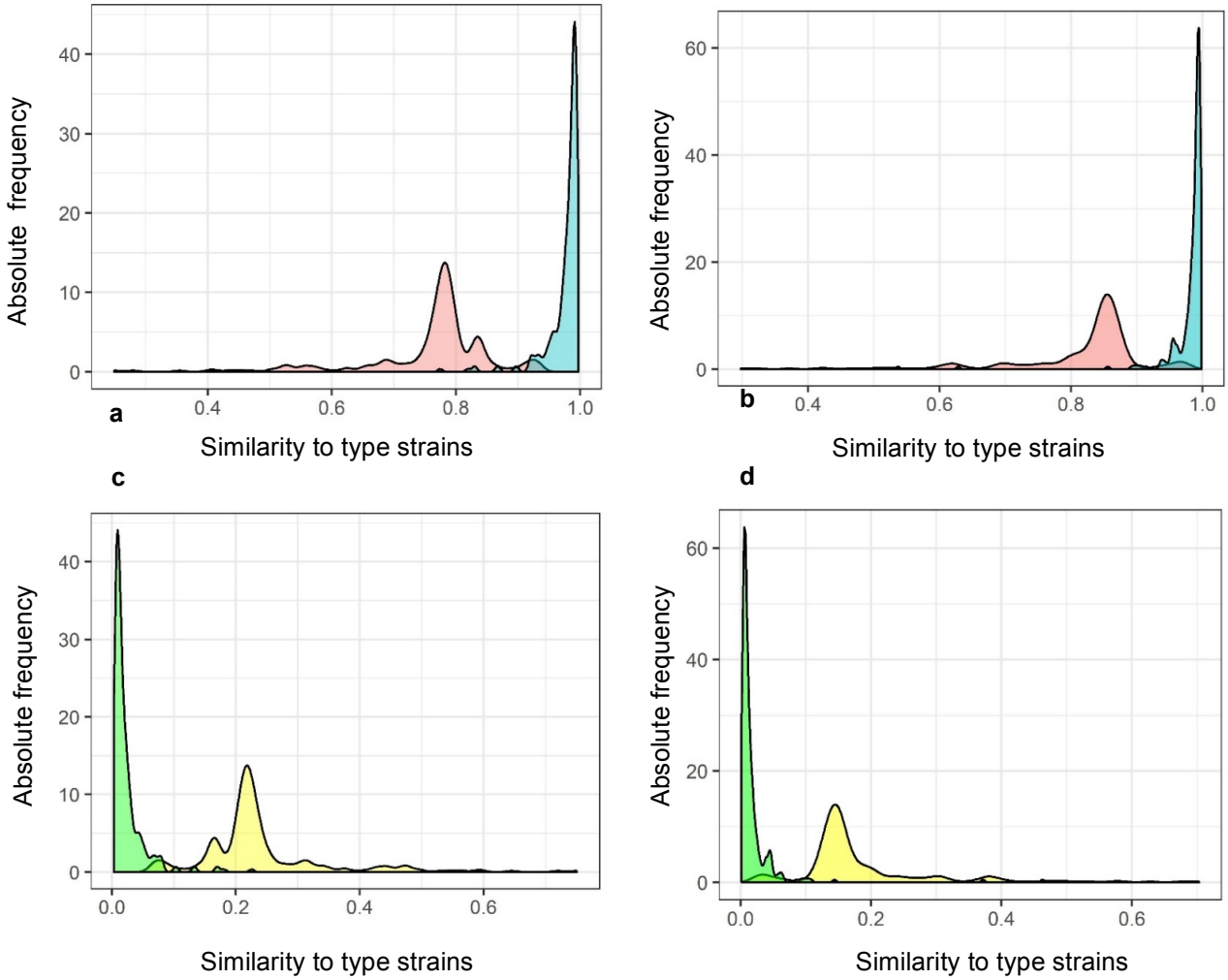
537 **Legend.** Variation of the time performance using references and files of different size (**a**);

538 regression analysis between the time performance, the dimension of library and FASTAq files using

539 the BTL (circle) and BBm (triangle) algorithms (**b**).

540

541 **Figure 4.** Distribution of the similarity to type strains with different analytical combinations.



542

543

544 **Legend.** (**a**): BTL Local - similarity to the correct species; (**b**): BBm - similarity to the correct

545 species. (**c**): BTL Local - similarity to the incorrect species; (**d**): BBm - similarity to the incorrect

546 species.

547 **Light Red** = mapping M1; **Light Blue** = mappings M2 and M3;

548 **Yellow** = mapping M1; **Green** = Mappings M2 and M3.

549

550

551

552 **TABLES**

553 **Table S1.** Strains employed in the study.

Strain Number	Species	Ward	City	Strain Number	Species	Ward	City
CMC 1730	*C. albicans*	Sp. Medicine	PI	**CMC 1926**	*C. albicans*	Gen. Medicine	Ud
CMC 1966	*C. albicans*	Sp. Medicine	Pi	**CMC 1927**	*C. albicans*	Surgery	Ud
CMC 1965	*C. albicans*	Sp. Medicine	Pi	**CMC 1928**	*C. albicans*	Sp. Medicine	Ud
CMC 1968	*C. albicans*	Sp. Medicine	Pi	**CMC 1931**	*C. albicans*	Surgery	Ud
CMC 1969	*C. albicans*	Sp. Medicine	Pi	**CMC 1932**	*C. albicans*	Gen. Medicine	Ud
CMC 1970	*C. albicans*	Sp. Medicine	Pi	**CMC 1936**	*C. albicans*	Sp. Medicine	Ud
CMC 1971	*C. albicans*	Sp. Medicine	Pi	**CMC 1937**	*C. albicans*	Gen. Medicine	UD
CMC 1974	*C. albicans*	Sp. Medicine	Pi	**CMC 1940**	*C. albicans*	Gen. Medicine	Ud
CMC 1977	*C. albicans*	Sp. Medicine	Pi	**CMC 1941**	*C. albicans*	Gen. Medicine	Ud
CMC 1980	*C. albicans*	Surgery	Pi	**CMC 1942**	*C. albicans*	Surgery	Ud
CMC 1982	*C. albicans*	Sp. Medicine	Pi	**CMC 1946**	*C. albicans*	Sp. Medicine	Ud
CMC 1983	*C. albicans*	ICU	Pi	**CMC 1952**	*C. albicans*	Gen. Medicine	Ud
CMC 1985	*C. albicans*	Sp. Medicine	Pi	**CMC 1954**	*C. albicans*	Surgery	Ud
CMC 1986	*C. albicans*	ICU	Pi	**CMC 1957**	*C. albicans*	Gen. Medicine	Ud
CMC 1987	*C. albicans*	Surgery	Pi	**CMC 1958**	*C. albicans*	Surgery	Ud
CMC 1990	*C. albicans*	ICU	Pi	**CMC 1959**	*C. albicans*	Surgery	Ud
CMC 1991	*C. albicans*	Surgery	Pi	**CMC 1960**	*C. albicans*	Gen. Medicine	Ud
CMC 1992	*C. albicans*	ICU	Pi	**CMC 1962**	*C. albicans*	Gen. Medicine	Ud
CMC 1994	*C. albicans*	Sp. Medicine	Pi	**CMC 1963**	*C. albicans*	Rehabilitation	Ud
CMC 1995	*C. albicans*	Surgery	Pi	**CBS 562**	*C. albicans*	Sp. Medicine	Type strain
CMC 1998	*C. albicans*	ICU	Pi	**CMC 1727**	*C. glabrata*	Sp. Medicine	Pi
CMC 2000	*C. albicans*	Sp. Medicine	Pi	**CMC 1726**	*C. glabrata*	Sp. Medicine	Pi
CMC 2001	*C. albicans*	Sp. Medicine	Pi	**CMC 1731**	*C. glabrata*	Sp. Medicine	Pi
CMC 2008	*C. albicans*	Sp. Medicine	Pi	**CMC 1976**	*C. glabrata*	Sp. Medicine	Pi
CMC 2010	*C. albicans*	ICU	Pi	**CMC 1989**	*C. glabrata*	ICU	Pi
CMC 2019	*C. albicans*	ICU	Pi	**CMC 2007**	*C. glabrata*	Sp. Medicine	Pi
CMC 2020	*C. albicans*	Surgery	Pi	**CMC 2015**	*C. glabrata*	Gen. Medicine	Pi
CMC 2021	*C. albicans*	ICU	Pi	**CMC 2018**	*C. glabrata*	ICU	Pi
CMC 2023	*C. albicans*	Gen. Medicine	Pi	**CMC 2027**	*C. glabrata*	Surgery	Pi
CMC 2025	*C. albicans*	ICU	Pi	**CMC 1782**	*C. glabrata*	ICU	Ud
CMC 2026	*C. albicans*	Surgery	Pi	**CMC 1781**	*C. glabrata*	Oncohematology	Ud
CMC 2029	*C. albicans*	Sp. Medicine	Pi	**CMC 1796**	*C. glabrata*	Sp. Medicine	Ud
CMC 2030	*C. albicans*	ICU	Pi	**CMC 1807**	*C. glabrata*	Gen. Medicine	Ud
CMC 2031	*C. albicans*	Surgery	Pi	**CMC 1813**	*C. glabrata*	Gen. Medicine	Ud
CMC 2032	*C. albicans*	Surgery	Pi	**CMC 1817**	*C. glabrata*	Gen. Medicine	Ud
CMC 2033	*C. albicans*	Surgery	Pi	**CMC 1830**	*C. glabrata*	Surgery	Ud
CMC 2034	*C. albicans*	Gen. Medicine	Pi	**CMC 1837**	*C. glabrata*	Gen. Medicine	Ud
CMC 2035	*C. albicans*	Sp. Medicine	Pi	**CMC 1846**	*C. glabrata*	Gen. Medicine	Ud
CMC 2036	*C. albicans*	Surgery	Pi	**CMC 1857**	*C. glabrata*	Gen. Medicine	Ud
CMC 2037	*C. albicans*	Gen. Medicine	Pi	**CMC 1861**	*C. glabrata*	Gen. Medicine	Ud
CMC 2042	*C. albicans*	ICU	Pi	**CMC 1864**	*C. glabrata*	ICU	Ud

CMC 2043	*C. albicans*	Gen. Medicine	Pi	**CMC 1865**	*C. glabrata*	Surgery	Ud
CMC 2045	*C. albicans*	Sp. Medicine	Pi	**CMC 1884**	*C. glabrata*	Gen. Medicine	Ud
CMC 2046	*C. albicans*	ICU	Pi	**CMC 1895**	*C. glabrata*	Surgery	Ud
CMC 2047	*C. albicans*	Sp. Medicine	Pi	**CMC 1912**	*C. glabrata*	Gen. Medicine	Ud
CMC 2048	*C. albicans*	Sp. Medicine	Pi	**CMC 1916**	*C. glabrata*	ICU	Ud
CMC 2049	*C. albicans*	Surgery	Pi	**CMC 1933**	*C. glabrata*	Gen. Medicine	Ud
CMC 2053	*C. albicans*	Sp. Medicine	Pi	**CMC 1934**	*C. glabrata*	Surgery	Ud
CMC 1768	*C. albicans*	Surgery	Ud	**CMC 1938**	*C. glabrata*	Gen. Medicine	Ud
CMC 1769	*C. albicans*	Gen. Medicine	Ud	**CMC 1950**	*C. glabrata*	Sp. Medicine	Ud
CMC 1770	*C. albicans*	Gen. Medicine	Ud	**CMC 1964**	*C. glabrata*	Sp. Medicine	Ud
CMC 1771	*C. albicans*	Gen. Medicine	Ud	**CMC 1967**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1773	*C. albicans*	Gen. Medicine	Ud	**CMC 1972**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1774	*C. albicans*	ICU	Ud	**CMC 1973**	*C. parapsilosis*	Surgery	Pi
CMC 1776	*C. albicans*	Gen. Medicine	Ud	**CMC 1975**	*C. parapsilosis*	Rehabilitation	Pi
CMC 1778	*C. albicans*	Sp. Medicine	Ud	**CMC 1979**	*C. parapsilosis*	ICU	Pi
CMC 1780	*C. albicans*	Gen. Medicine	Ud	**CMC 1981**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1785	*C. albicans*	Sp. Medicine	Ud	**CMC 1984**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1786	*C. albicans*	Sp. Medicine	Ud	**CMC 1993**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1788	*C. albicans*	Surgery	Ud	**CMC 1997**	*C. parapsilosis*	Surgery	Pi
CMC 1790	*C. albicans*	Gen. Medicine	Ud	**CMC 1999**	*C. parapsilosis*	Surgery	Pi
CMC 1794	*C. albicans*	Surgery	Ud	**CMC 2005**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1795	*C. albicans*	Gen. Medicine	Ud	**CMC 2006**	*C. parapsilosis*	ICU	Pi
CMC 1797	*C. albicans*	Oncohematology	Ud	**CMC 2012**	*C. parapsilosis*	ICU	Pi
CMC 1799	*C. albicans*	ICU	Ud	**CMC 2013**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1802	*C. albicans*	ICU	Ud	**CMC 2014**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1803	*C. albicans*	ICU	Ud	**CMC 2016**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1804	*C. albicans*	Surgery	Ud	**CMC 2022**	*C. parapsilosis*	Surgery	Pi
CMC 1806	*C. albicans*	Surgery	Ud	**CMC 2038**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1811	*C. albicans*	Gen. Medicine	Ud	**CMC 2039**	*C. parapsilosis*	Surgery	Pi
CMC 1815	*C. albicans*	Gen. Medicine	Ud	**CMC 2040**	*C. parapsilosis*	Surgery	Pi
CMC 1816	*C. albicans*	Sp. Medicine	Ud	**CMC 2044**	*C. parapsilosis*	ICU	Pi
CMC 1818	*C. albicans*	Gen. Medicine	Ud	**CMC 2050**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1819	*C. albicans*	Surgery	Ud	**CMC 2051**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1820	*C. albicans*	Sp. Medicine	Ud	**CMC 1772**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1821	*C. albicans*	Surgery	Ud	**CMC 1783**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1822	*C. albicans*	Surgery	Ud	**CMC 1787**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1823	*C. albicans*	Gen. Medicine	Ud	**CMC 1791**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1824	*C. albicans*	Surgery	Ud	**CMC 1793**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1828	*C. albicans*	Surgery	Ud	**CMC 1800**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1829	*C. albicans*	Sp. Medicine	Ud	**CMC 1801**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1831	*C. albicans*	Surgery	Ud	**CMC 1805**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1833	*C. albicans*	Surgery	Ud	**CMC 1809**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1834	*C. albicans*	Gen. Medicine	Ud	**CMC 1814**	*C. parapsilosis*	Oncohematology	Ud
CMC 1835	*C. albicans*	Sp. Medicine	Ud	**CMC 1838**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1840	*C. albicans*	Surgery	Ud	**CMC 1841**	*C. parapsilosis*	Surgery	Ud
CMC 1842	*C. albicans*	Surgery	Ud	**CMC 1849**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1843	*C. albicans*	Oncohematology	Ud	**CMC 1851**	*C. parapsilosis*	Sp. Medicine	Ud

CMC 1844	*C. albicans*	Gen. Medicine	Ud	**CMC 1859**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1845	*C. albicans*	Gen. Medicine	Ud	**CMC 1867**	*C. parapsilosis*	ICU	Ud
CMC 1847	*C. albicans*	Gen. Medicine	Ud	**CMC 1882**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1848	*C. albicans*	Sp. Medicine	Ud	**CMC 1892**	*C. parapsilosis*	Rehabilitation	Ud
CMC 1850	*C. albicans*	ICU	Ud	**CMC 1897**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1852	*C. albicans*	Gen. Medicine	Ud	**CMC 1899**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1853	*C. albicans*	Gen. Medicine	Ud	**CMC 1902**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1854	*C. albicans*	Sp. Medicine	Ud	**CMC 1917**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1856	*C. albicans*	ICU	Ud	**CMC 1929**	*C. parapsilosis*	ICU	Ud
CMC 1858	*C. albicans*	Gen. Medicine	Ud	**CMC 1930**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1860	*C. albicans*	Gen. Medicine	Ud	**CMC 1935**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1862	*C. albicans*	Gen. Medicine	Ud	**CMC 1939**	*C. parapsilosis*	Surgery	Ud
CMC 1863	*C. albicans*	Gen. Medicine	Ud	**CMC 1945**	*C. parapsilosis*	Oncohematology	Ud
CMC 1866	*C. albicans*	Surgery	Ud	**CMC 1948**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1868	*C. albicans*	Gen. Medicine	Ud	**CMC 1949**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1869	*C. albicans*	Sp. Medicine	Ud	**CMC 1951**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1870	*C. albicans*	Sp. Medicine	Ud	**CMC 2004**	*C. orthopsilosis*	ICU	Pi
CMC 1871	*C. albicans*	Sp. Medicine	Ud	**CMC 2011**	*C. orthopsilosis*	Sp. Medicine	Pi
CMC 1872	*C. albicans*	Sp. Medicine	Ud	**CMC 1808**	*C. orthopsilosis*	Gen. Medicine	Ud
CMC 1873	*C. albicans*	Gen. Medicine	Ud	**CMC 1812**	*C. orthopsilosis*	Gen. Medicine	Ud
CMC 1875	*C. albicans*	Gen. Medicine	Ud	**CMC 1826**	*C. orthopsilosis*	Gen. Medicine	Ud
CMC 1876	*C. albicans*	Gen. Medicine	Ud	**CMC 1880**	*C. orthopsilosis*	Sp. Medicine	Ud
CMC 1877	*C. albicans*	ICU	Ud	**CMC 1922**	*C. orthopsilosis*	Gen. Medicine	Ud
CMC 1878	*C. albicans*	Gen. Medicine	UD	**CMC 1978**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1879	*C. albicans*	Gen. Medicine	Ud	**CMC 2003**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1881	*C. albicans*	Gen. Medicine	Ud	**CMC 2009**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1885	*C. albicans*	Surgery	Ud	**CMC 2017**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1886	*C. albicans*	Gen. Medicine	Ud	**CMC 2024**	*C. tropicalis*	ICU	Pi
CMC 1887	*C. albicans*	ICU	Ud	**CMC 2041**	*C. tropicalis*	Gen. Medicine	Pi
CMC 1888	*C. albicans*	Sp. Medicine	Ud	**CMC 2052**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1889	*C. albicans*	Gen. Medicine	Ud	**CMC 1784**	*C. tropicalis*	Oncohematology	Ud
CMC 1890	*C. albicans*	Gen. Medicine	Ud	**CMC 1792**	*C. tropicalis*	ICU	Ud
CMC 1891	*C. albicans*	Sp. Medicine	Ud	**CMC 1798**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1893	*C. albicans*	Oncohematology	Ud	**CMC 1810**	*C. tropicalis*	Oncohematology	Ud
CMC 1896	*C. albicans*	Gen. Medicine	Ud	**CMC 1827**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1898	*C. albicans*	Gen. Medicine	Ud	**CMC 1836**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1900	*C. albicans*	Gen. Medicine	Ud	**CMC 1839**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1901	*C. albicans*	Gen. Medicine	Ud	**CMC 1855**	*C. tropicalis*	Oncohematology	Ud
CMC 1903	*C. albicans*	Gen. Medicine	Ud	**CMC 1874**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1904	*C. albicans*	Sp. Medicine	Ud	**CMC 1943**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1905	*C. albicans*	Gen. Medicine	Ud	**CMC 1953**	*C. tropicalis*	Oncohematology	Ud
CMC 1906	*C. albicans*	Gen. Medicine	Ud	**CMC 1956**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1907	*C. albicans*	Gen. Medicine	Ud	**CMC 1961**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1909	*C. albicans*	Sp. Medicine	Ud	**CMC 1777**	*C. dubliniensis*	ICU	Ud
CMC 1910	*C. albicans*	ICU	Ud	**CMC 1908**	*C. dubliniensis*	ICU	Ud
CMC 1911	*C. albicans*	Gen. Medicine	Ud	**CMC 1988**	*I. orientalis*	Surgery	Pi
CMC 1913	*C. albicans*	Gen. Medicine	Ud	**CMC 2002**	*I. orientalis*	Sp. Medicine	Pi

CMC 1914	*C. albicans*	Surgery	Ud	**CMC 1894**	*I. orientalis*	Rehabilitation	Ud
CMC 1915	*C. albicans*	Sp. Medicine	Ud	**CMC 1944**	*C. lusitaniae*	Sp. Medicine	Ud
CMC 1918	*C. albicans*	Rehabilitation	Ud	**CMC 1996**	*C. rugosa*	ICU	Pi
CMC 1919	*C. albicans*	Gen. Medicine	Ud	**CMC 1832**	*M. guilliermondii*	Oncohematology	Ud
CMC 1920	*C. albicans*	Gen. Medicine	Ud	**CMC 1825**	*M. guilliermondii*	Oncohematology	Ud
CMC 1921	*C. albicans*	Gen. Medicine	Ud	**CMC 1883**	*M. guilliermondii*	Rehabilitation	Ud
CMC 1923	*C. albicans*	Gen. Medicine	Ud	**CMC 1924**	*M. guilliermondii*	Gen. Medicine	Ud
CMC 1925	*C. albicans*	Sp. Medicine	Ud	**CMC 1775**	*P. jadinii*	Gen. Medicine	Ud

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
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. Performances of algorithms in mapping FASTAq files of different size with large reference libraries.

Library		**Algorithm**	**FASTAq**		**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**
CBS ITS	15,565	BTL	A	21,238	63.0	33.34	18,274	2,964	684	74.0%	100%
			B	58,263	99.0	123.00	54,822	3,441	661	91.0%	100%
		BBm	A	21,238	80.0	39.82	18,323	2,915	626	68.2%	100%
			B	58,263	114.0	49.57	55,675	2,588	708	62.3%	100%
CBS ITS-LSU	34,683	BTL	A	21,238	186.0	104.00	19,163	2,075	2051	73.4%	100%
			B	58,263	273.0	393.00	57,140	1,123	2,457	79.5%	100%
		BBm	A	21,238	204.0	109.00	19,000	2,238	1,939	58.8%	100%
			B	58,263	311.0	161.00	56,930	1,333	2,445	63.6%	100%
ISHAM ITS	2,727	BTL	A	21,238	16.7	13.59	13,904	7,334	319	74.1%	100%
			B	58,263	30.3	33.50	36,416	21,847	297	82.0%	100%
		BBm	A	21,238	37.7	16.13	13,665	7,573	280	74.5%	100%
			B	58,263	45.2	26.03	35,653	22,610	305	85.7%	100%

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

571 **Table 3.** Example of a M1-M2-M3 mapping against an *ad hoc* library of pathogenic yeasts (CMC

572 1912 strain).

573 **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	212814	1426	99.90%	99.00%	203.48	6.14%	6.14%	9.26%	6.14%	6.08%	6.80%
C. dubliniensis	12362	80	53.20%	99.30%	10.01	0.34%	0.36%	0.46%	0.19%	0.35%	0.34%
C. famata	1863	6	24.40%	98.80%	0.56	0.03%	0.05%	0.03%	0.01%	0.05%	0.03%
C. glabrata	3004133	20223	100.00%	99.00%	1813.15	87.06%	86.74%	82.51%	86.74%	85.87%	**86.41%**
I. orientalis	2689	12	42.20%	96.60%	1.30	0.05%	0.08%	0.06%	0.03%	0.07%	0.06%
C. metapsilosis	3886	20	43.10%	99.00%	1.93	0.09%	0.11%	0.09%	0.05%	0.11%	0.09%
C. orthopsilosis	8590	52	92.30%	99.10%	6.34	0.22%	0.25%	0.29%	0.23%	0.25%	0.25%
C. parapsilosis	70672	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	0.00	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
C. rugosa	1034	2	4.50%	92.90%	0.05	0.01%	0.03%	0.00%	0.00%	0.03%	0.01%
C. sake	105986	703	35.00%	99.00%	69.76	3.03%	3.06%	3.17%	1.07%	3.03%	2.69%
C. tropicalis	23711	154	96.40%	99.20%	20.51	0.66%	0.68%	0.93%	0.66%	0.68%	0.73%
C. utilis	1705	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	8425	52	97.70%	98.90%	6.21	0.22%	0.24%	0.28%	0.24%	0.24%	0.25%
S. cerevisiae	4642	25	19.00%	99.20%	2.32	0.11%	0.13%	0.11%	0.03%	0.13%	0.10%

574

575 **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	91888	606	96.40%	98.20%	89.0	2.86%	2.90%	4.62%	2.33%	2.88%	3.13%
C. dubliniensis	0	0	0.00%	0.00%	0.0	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
C. famata	0	0	0.00%	0.00%	0.0	0.01%	0.04%	0.01%	0.00%	0.03%	0.02%
C. glabrata	3231287	21757	100.00%	87.70%	2413.9	96.13%	95.85%	94.02%	95.85%	94.32%	**95.74%**
I. orientalis	1221	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	0.02%	0.05%	0.01%	0.00%	0.05%	0.02%
C. orthopsilosis	2355	9	44.50%	97.80%	1.1	0.06%	0.09%	0.08%	0.06%	0.09%	0.08%
C. parapsilosis	18410	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	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
C. rugosa	0	0	0.00%	0.00%	0.0	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
C. sake	0	0	0.00%	0.00%	0.0	0.09%	0.12%	0.08%	0.01%	0.12%	0.09%
C. tropicalis	6670	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	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
C. lusitaniae	0	0	0.00%	0.00%	0.0	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
M. guilliermondii	4147	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	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

576

577

578 **Legend.** (**a**) Mapping of a FASTAq file against a selected library of 16 type strains of pathogenic

579 yeasts; (**b**) Mapping of the FASTAq file against the type strains of the presumptive species and the

580 resulting mapping of the residual unused reads.

Paper VI

High Depth Next Generation Sequencing of single colony DNA reveals large variation levels of the Ribosomal DNA region ITS-LSU D1/D2 in the four prevalent pathogenic species of the genus *Candida*

Claudia Colabella[1], Laura Corte[1], Luca Roscini[1], Debora Casagrande Pierantoni[1], Matteo Bassetti[2], Carlo Tascini[3], and Gianluigi Cardinali [1,4*].

[1]Department of Pharmaceutical Sciences, Microbiology section - University of Perugia-Italy; [2]Udine Hospital - Udine - Italy; [3]Cotugno Hospital Napoli - Italy; [4]CEMIN Excellence Research Centre - University of Perugia - Italy

Running Title: Internal rDNA heterogeneity revealed by NGS

Key Words: NGS, Sanger, ITS, LSU, Variants, *Candida*, concerted evolution, gene conversion, unequal crossing over.

***Corresponding author:** Dr. Gianluigi Cardinali
Dept. of Pharmaceutical Sciences - Microbiology
Borgo 20 Giugno, 74
I - 06121 PERUGIA (ITALY)
e.mail: *gianluigi.cardinali@unipg.it*
phone +39 075 585 6478; fax +39 075 585 6470

ABSTRACT

Ribosomal RNA in fungi is encoded by a series of genes and spacers included in a large operon present in 100-200 tandem repeats, normally in a single locus. The multigene nature of the rDNA was somehow masked by Sanger sequencing, which produces a single sequence reporting the prevalent nucleotide of each site. The introduction of Next Generation Sequencing leads to a deeper knowledge of the individual sequences and therefore of the variants between the same DNA sequence located in different tandem repeats. The use of an innovative NGS technique allowed the high-throughput high-depth of the ITS1-LSU D1/D2 amplicons sequencing of 271 strains belonging to the four prevalent yeast species of the genus *Candida*. Results showed the presence of a large extent of variability among the strains and between the species. These variants were differently distributed throughout the analysed regions with an higher concentration within the ITS2 region. The variant profiles of strain isolated from two different hospitals showed more than 0.9 correlation in *Candida glabrata, C. parapsilosis* and *C. tropicalis*, whereas the correlation of the *C. albicans* isolates was 0.8.

These data indicate that the concerted evolution was not able to homogenize totally these sequences. Furthermore, the variation level and localization suggest that gene conversion is the most likely mechanism to remove the variants, but its action differ among the species and the four DNA marker sequences employed. Finally, the detected variability can be considered as a typing tool to characterize yeast strains.

INTRODUCTION

The question on the genetic mechanisms leading to the homogenization of the multigene families has been long debated since 1972 when Brown proposed the unequal crossing over as a potential explanation of the fact that copies of the same gene in the same genome seem to evolve in a concerted manner (Brown, Wensink et al. 1972). Seven years later, Jeffries proposed gene conversion as the mechanism able to clear variant copies within human globin multigene families

53 (Jeffreys 1979) and this mechanism became the most accepted to explain the whole phenomenon.
54 Last in the debate, the birth-and-death evolution was proposed as an alternative to the other two
55 models by Nei (Nei and Rooney 2005). For ribosomal DNA there is a large consensus in favour of
56 gene conversion, although birth and death evolution was sustained to be present in some
57 filamentous fungi (Rooney and Ward 2005).
58 In recent years, cloning has been used to unveil the variability levels (Simon and Weiß 2008),
59 finding high variation rates among fungi. Similarly, NGS has been applied, but with a relatively low
60 level of coverage in most instances (Ganley and Kobayashi 2007, Torres-Machorro, Hernández et
61 al. 2010, West, James et al. 2014). Results are often contrasting due to differences in the rDNA
62 region studied, technique employed, and level of coverage. The last parameter is probably critical
63 because with over 100 repeats (Torres-Machorro, Hernández et al. 2010), a coverage of at least
64 1000X is necessary to randomly sample each copy ten times. Lower values are likely to
65 underestimate this internal variability.
66 The aim of this paper is to elucidate the extent and the location of the rDNA variants, by using an
67 high-depth NGS sequencing of a rDNA region interesting not only for its functions, but also
68 because it contains two important taxonomic markers proposed also as barcodes (Schoch, Seifert et
69 al. 2012, Kurtzman and Robnett 2013, Schoch, Robbertse et al. 2014, Irinyi, Serena et al. 2015).
70 The work was carried out on 271 strains of the four prevalent pathogenic species of the genus
71 *Candida*, including also the four type strains for taxonomic control. This strain set was chosen to
72 give a reliable and representative view of the sequence variability with a relevant of freshly isolated
73 strains that did not have time to undergo laboratory induced variations. Finally, the importance of
74 the selected species in the medical environment requires a good knowledge of these markers
75 variability, since it can easily impact on diagnostics.
76
77
78

79 **MATERIALS AND METHODS**

80 **Strains and growth conditions**

81 In this study 286 strains were isolated from two Italian Hospitals (Pisa and Udine). All the strains
82 belong to opportunistic species of *Candida* genus and were isolated from patient blood cultures.
83 The strains were included in the Cemin Microbial Collection of the Microbial Genetics and
84 Phylogenesis Laboratory of Cemin (Centre of Excellence on Nanostructured Innovative Materials
85 for Chemicals, Physical and Biomedical Applications - University of Perugia) and extensively
86 described in a medical ecology paper (Corte, Roscini et al. 2016). Over twelve species were isolated
87 in both hospitals, among which four, *C. albicans*, *C. glabrata*, *C. parapsilosis* and *C. tropicalis*,
88 represented the vast majority of the isolates. 271 strains of these four major species were employed
89 in this study (Tab. 1). All the strains were stored at -80°C in 17% glycerol immediately upon
90 isolation. First step of cultivation was carried out on YEPDA (YEPD added with 1.7% agar) at
91 37°C, following the current procedures. When a biomass was necessary, the strains were grown in
92 YEPD (Yeast extract 1%, Peptone 1%, Dextrose 1% all products from Biolife -
93 http://www.biolifeitaliana.it/) at 37°C with 150 rpm shaking.

94

95 **DNA extraction and molecular techniques**

96 Genomic DNA was extracted as indicated by Cardinali et al (Cardinali, Bolano et al. 2001). ITS1,
97 5.8S, ITS2 rDNA genes and D1/D2 domain of the LSU were amplified with FIREPole® Taq DNA
98 Polymerase (Solis BioDyne, Estonia), using ITS1 (5'-TCCGTAGGTGAACCTGCGG) - NL4
99 (GGTCCGTGTTTCAAGACGG) primers. The amplification protocol was carried out as follows:
100 initial denaturation at 94°C for 3 min, 30 amplification cycles (94°C for 1 min, 54°C for 1 min and
101 72°C for 1 min) and final extension at 72°C for 5 min. Amplicons were subjected to electrophoresis
102 on 1.5% agarose gel (Gellyphor, EuroClone, Italy). Amplicons were sequenced with NGS
103 PlexWell™ technologies (http://www.seqwell.com/) with the same primers used for the generation
104 of the amplicons. The reads of each strain, contained in FASTAq file, were analysed with Geneious

105 R9 software (v. 9.1.5, Biomatters, Auckland, New Zealand - http://www.geneious.com/).
106 Identification was carried out as indicated by Colabella et al (Colabella et al. in submission).
107
108 **Variants analysis**
109 All the 271 FASTAq files were mapped against the relative type strains using Bowtie2 algorithm
110 setting "local" with High Sensitivity mode and no trimming of the sequences. The detection of the
111 SNPs was performed on each contigs in Geneious software as follows: Annotate&Predict function,
112 FindVariations/SNPs with a Minimum Variant Frequency of 1% and separating annotations for
113 each variant at a position. A csv table recording positions, variant frequency, average quality and
114 variant P-value was exported and analysed with a build in macro in Microsoft Excel®. The
115 frequency of the variants was calculated for all the strains at each amplicons position. The analysis
116 were first carried out considering the four species separately and then all the strains were grouped
117 trough the base of their site of isolation (Pisa and Udine hospitals). The same analysis were also
118 performed considering the minimum variant frequency of 5%. Among the species the average of the
119 Variant Frequency (hereinafter referred to as AVF) and the Variant bearing Strains Frequency
120 (hereinafter referred to as VSF) were calculated.
121
122 **Statistics**
123 For the four major species a correlation analysis of the AVF value among the four rDNA region
124 (ITS1, 5.8S, ITS2, LSU) was carried out in R environment software (http://www.R-project.org/)
125 using (vegan) library and cor.test function for the estimation of the *p*-value. A correlation between
126 the AVF and the two different sampling sites (Pisa and Udine) was also calculated showing an high
127 Pearson correlation values that ranged from 0.90 to 0.93. In order to evaluate the distribution of the
128 strains isolated from the two different places, a PLS analysis was carried out using R software. For
129 the sparse multivariate model the mixOmics package (http://www.mixOmics.org/) was used.
130

131 **RESULTS**

132 **Heterogeneity of the rDNA within the genomes**

133 In fungi ribosomal RNA is encoded by 100 to 200 tandem repeats known to have some extent of
134 internal variation between the copies of the same gene. Next Generation Sequencing offers the
135 possibility to evaluate this heterogeneity by analysing the Single Nucleotide Polymorphisms within
136 the reads of a rDNA region amplified from a single strain. Theoretically, variants between the
137 various copies could be obtained by SNPs calling, of a *de novo* assembly procedure. The limits of
138 this approach are the difficulty to compare SNPs of different strains and the computer intensity
139 required by the *de novo* assembly. In our hands, mapping the reads against a reference was some
140 100 times faster than *de novo* assembly and produced more reliable results (Colabella et al. in
141 submission). On the basis of these evidences, the reads deriving from the NGS of the ITS-LSU
142 amplicons were mapped against the corresponding Sanger sequences of the species type strain, in
143 order to record position and frequency of differences relative to the reference sequence.

144 This approach allowed to calculate the Variant Frequency (hereinafter referred to as VF) i.e. the
145 proportion of the nucleotides that showed variants in comparison to the type strain used as
146 reference, calculated for each position of the sequence. Sites with less than 1% of variants were
147 considered non variable, in order to avoid background noise due to technical factors. The average
148 of the VF (AVF) among a set of strain (e.g. the species) was calculated to account for the
149 phenomenon at the species level and indicates the frequency of each single site among all the strains
150 of a species. Another measure of heterogeneity, the Variant bearing Strains Frequency (VSF), is the
151 percentage of strains, within each species, that showed > 1% variants at each specific site of the
152 amplicons.

153 The AVF and the VSF of the four species strains analysed showed a large variability among the
154 species with AVF varying from 1.20% (*C. albicans*) to 3.75% (*C. parapsilosis*) and VSF from
155 2.50% (*C. albicans*) to 4.44% in *C. tropicalis* (Tab. 2). The variability within the species was
156 particularly high ranging from 4.93% standard deviation for *C. albicans* AVF to 17.05% of *C.*

157 *tropicalis* VSF. Finally, the differences between the two cities of isolation (Pisa and Udine) were
158 not statistically significant (p> 0.1), due to the very high standard deviations.
159 Histograms, describing this internal variability, showed different distributions between the species
160 of both AVF and VSF (Fig. 1). Excluding the sites with no variants to the type strain (columns "0
161 %"), *C. albicans* showed 163 sites with 1% variants (AVF) and few with variations up to 40%.
162 Similarly, 187 sites bore at least 2% of strains with > 1% of variants and few up to 20%. *C.*
163 *glabrata* and *C. tropicalis* showed in general less sites with variants, although the former had 161
164 nucleotides with at least 4% strains carrying > 1% of variants. Finally, *C. parapsilosis* showed 69%
165 of invariant sites and several sites with up to 20% AVF and VSF. In this species, there were 8
166 strains with more than 90% AVF (Fig. 1e). In the same species, 12 sites had between 90% and
167 100% of the strains with more than 1% variants.
168
169 **Positioning the internal heterogeneity of rDNA within the four regions**
170 The analysis of the distribution of the variants throughout the four regions under study, showed that
171 the occurrence of the internal heterogeneity is scattered throughout the whole region with different
172 distribution in the four species considered. *C. albicans* showed several variant sites with AVF peaks
173 in the ITS2 and LSU regions. The percentage of strains of this species carried variants in the area
174 between the 3'end of the ITS and the 5' of the LSU (Fig. 2a and 2b). *C. glabrata* and *C.*
175 *parapsilosis* with several variants site distributed throughout the whole region (Fig. 2c to 2f). In *C.*
176 *parapsilosis* very high AVF was found in the 5.8S gene that is characterized by low AVF in *C.*
177 *albicans* and *C. tropicalis*, which showed that the vast majority of the strains carried variants in the
178 region between the 5.8S and the 5' of the LSU. In all the species, the central part of the LSU
179 showed few strains carrying variants and relatively low AVF.
180 A more stringent analysis was carried out considering only the variants with frequency > 5% among
181 the reads of the same strain. With this condition, the number of strains carrying variants decreased
182 drastically (Fig. 3), although the general disposition remained quite similar to that showed with the

183 1% threshold. The AVF distribution at 1% and 5% threshold were similar and no drastic decrease of
184 frequency was observed because the variants present in the sites with more than 5% were in any
185 case very high (Fig. 2 and Fig. 3).
186 A summary of these results was obtained by condensing the AVF and VSF in a single average value
187 for each of the four regions (Fig. 4). *C. parapsilosis* showed the highest average variability among
188 all species with all four regions over 3% of AVF, whereas *C. albicans* was the least variant (Fig.
189 4a). ITS2 displayed higher AVF than the other three regions, ranging from 1.72% (*C. albicans*) to
190 5.37% in *C. parapsilosis*. The least variant regions were the ITS1 and the LSU with very close
191 levels of variability within the species, although in general ranged from 1% to 3.68%. The 5.8S
192 gene had intermediate AVF, ranging from 1.19% to 4.22%.
193 The percentage of strains carrying variants (VSF) displayed a totally different pattern than AVF,
194 with *C. tropicalis* displaying the highest values, followed by *C. parapsilosis, C. glabrata* and *C.*
195 *albicans* (Fig. 4b). According to this metric, most of the strains carried variants in the ITS2 and
196 LSU genes, whereas few variations were observed in the ITS1. Once again, the 5.8S had an
197 intermediate trend.
198 The ratio between AVF and VSF is the mean number of variants (MNV) in the variable sites. *C.*
199 *tropicalis* and *C. albicans* showed very low MNV in all the four regions with very little variations
200 among the regions i.e. from 0.18 to 0.31 for the former and from 0.37 to 0.88 for the latter. The
201 other two species displayed much higher MNV values and larger differences among the genes, for
202 example in *C. glabrata* differences ranged from 0.31 of the LSU to 2.37 of the ITS 1 and ITS2.
203 Interestingly, the MVN of ITS1 and ITS2 within these two species were quite similar: 2.37 and 2.36
204 for *C. glabrata;* 1.70 and 1.71 in *C. parapsilosis*.
205 These data indicated that the four DNA regions have different rates of homogenization, possibly
206 due to different mechanisms of concerted evolution. In order to verify this hypothesis, the data
207 reported in Fig. 4a were subject to correlation analysis, showing that the AVF of ITS2 is poorly
208 correlated with the other three DNA regions (Tab. 3). Furthermore the high *p* values of these three

209 correlations corroborate the concept that ITS2 variation rate differs significantly from that of the
210 other genes and ITS1. According to this analysis, ITS1 is well correlated with the neighbour
211 sequence of the 5.8S (correlation 0.949, *p* 0.05) and slightly more with that of the D1/D2 domain of
212 the LSU (correlation 0.991, *p* 0.008).
213
214 **Relationship of the heterogeneity of rDNA between strains isolated in different places**
215 The variability presented above was calculated on all strains together, independently of the fact that
216 were isolated from two different hospitals, Pisa and Udine, located some 450 Km apart. Splitting
217 the AVF and VSF data between the two strain sets produced quite different distribution patterns
218 using both 1% (Supplementary Figure S1) and 5% (Supplementary Figures S2). For instance, *C.*
219 *albicans* isolates from Pisa had a series of large AVF peaks in a small region of the ITS2, whereas
220 large peaks were found also in the LSU gene when the strains from Udine were analysed (Fi S1b
221 and S2b). Conversely, the 5' half of the *C. parapsilosis* strains from Pisa had several peaks with up
222 to 80% AVF. *C. glabrata* and *C. parapsilosis* strains from Pisa showed larger AVF values than
223 those from Udine and the variant sites were scattered in a rather uniform way, whereas in Udine
224 strains, most of the variants were clustered together leaving large areas with little variations, as
225 described above.
226 Using a 5% threshold, the higher AVF of Pisa in comparison to Udine remained substantially
227 unchanged (Supplementary Fig. S2). Interestingly, *C. tropicalis* showed little (Pisa) or no (Udine)
228 variants in the central part of the LSU gene.
229 The AVF between the strains of the two hospitals showed high Pearson correlation values ranging
230 from 0.90 to 0.93 (see legend of Fig. 5). The linear regression analysis of *C. albicans* was described
231 by the equation F1
232
233 *F1* $$AVF_U = 0.5617 * AVF_P + 0.0041$$
234

235 where AVF_U and AVF_P indicate the AVF values from Udine and Pisa, respectively. This regression
236 indicates that in *C. albicans* the variants found in Udine were generally little more than 50% of
237 those retrieved in Pisa. On the other hand, the relatively good R^2 (0.8) indicates that most of the
238 variations found in the two hospitals were correlated and that the two strain sets share the
239 evolutionary history that caused either the presence of variants, or their partial homogenization.
240 The same analysis, carried out for the other three species indicated not only a better correlation but
241 also that AVF_U and AVF_P were almost similar, as indicated by the slope coefficient close to 1.
242 Interestingly, some sites showed much larger variant frequency among the Udine than the Pisa
243 strains, such as those enclosed in an ellipse in Fig. 5b and 5d, regarding *C. glabrata* and *C.*
244 *tropicalis* respectively. The opposite situation occurred in *C. albicans* (Fig. 5a). These data indicate
245 that some sites had larger AVF of the same sequences dwelling in strains of different origin. These
246 evidences suggest that the evolutionary process, leading to the copy homogenization, might differ
247 according to the species and to the place of isolation, opening the possibility that these markers can
248 be used as typing tools when next generation sequencing is used.
249
250 **Is it possible to use the internal heterogeneity of rDNA for strain typing?**
251 The variability of the distribution of the sequence variants among the genes, the species and the city
252 of isolation suggested to test the applicability of the AVF profiles to typing. The result of a PLS
253 analysis, carried out for the four species separately, showed that the strains from Udine cover a
254 larger PLS space than those from Pisa and that these profiles cannot discriminate the origin of
255 isolation (Fig. 6). Only *C. glabrata* displayed a partial separation of the strains from the two
256 hospitals. However, some strain clustering was observed in all four species as that within a circle in
257 Fig. 6a, related to *C. albicans*. A careful examination of these tightly positioned strains revealed a
258 very close date of isolation in the same hospital, suggesting that indeed some of them can be a copy
259 of the same isolate or close relatives. A hierarchical dendrogram from the same AVF (data not
260 shown) confirmed these findings and corroborated the idea that putatively similar isolates are

261 placed closer in a PLS space. These data must be confirmed by other experiments with different
262 typing systems compared to the NGS variant proposed here. Furthermore, a fine-tuning of variant
263 thresholds (e.g. 1% vs 5%) and of other analytical parameters must be carried out to carefully check
264 whether this approach can produce an effective stain typing system.
265

266 **DISCUSSION**

267 The discussion relative to the genetic mechanisms of concerted evolution has recently focused on
268 the level of variability (Ganley and Kobayashi 2007). Relatively low levels of variants are expected
269 with gene conversion, whereas larger variability should be expected from unequal crossing over and
270 birth and death evolution. Despite the interest on this topic, a consensus on the quantitative levels of
271 "low" and "high" variability has never been reached, for what we can tell at this time. Another
272 problem consists in the technique employed to determine the variation levels. In fact, some works
273 rely on cloning some tens of copies and other use NGS with a lower depth than that reached in this
274 work, in which levels of depth between 3,000 and 15,000 were common. The amount of the
275 variability when studying yeast populations can be carried out with different metrics and starting
276 from various approaches. The *de novo* assembly in our hands produced several consensus sequences
277 with high divergence levels to the Sanger sequence of the species type strain and even divergent to
278 the Sanger sequence of rDNA region of the very same strain. It was therefore clear that, at least
279 with our setting, the *de novo* assembly approach is biased by high level variability, large part of the
280 algorithm chosen. The mapping against a reference was much more consistent in our hands and
281 showed great fidelity in reconstructing the sequence, in terms of variants and global length. In fact,
282 *de novo* assemblies were in average much longer than the expected sequences, whereas the
283 mappings matched the Sanger sequence length. The major problem with mapping is the definition
284 of a reliable reference sequence. Theoretically, the best reference would be the Sanger sequence of
285 the same strain used for the high-depth next generation sequencing. This approach is tedious and
286 has the problem of lacking a common reference to align all the NGS mappings. In other terms, this

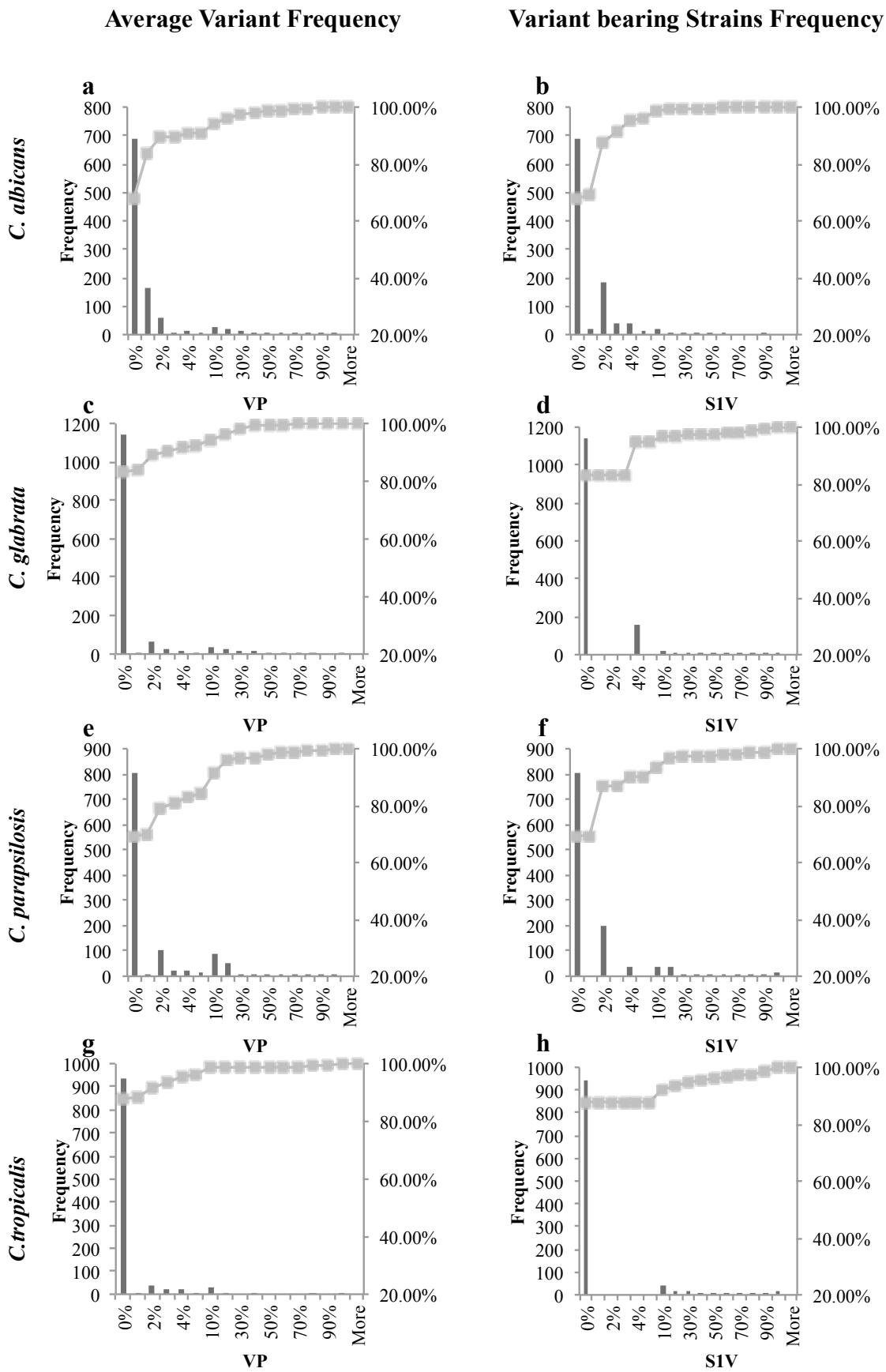
287 solution would produce difficult to compare data at the species or population level. On the basis of
288 these considerations, we chose to map all strains NGS sequences to the respective species type
289 strain.
290 The variability found ranged from 12% in *C. tropicalis* to 32% approximately in *C. albicans* (Fig.
291 1) with a level of variability among species and genes largely described in the results section. These
292 figures are higher than those presented for other yeast species, but the coverage of the sequences
293 used in that paper was never higher than 8x (Ganley and Kobayashi 2007), suggesting that an
294 increase on the sequencing depth is crucial to uncover an otherwise hidden variability. This
295 relatively high variation rate points to unequal crossing over or birth-and-death mechanisms, but the
296 fact that some areas were less densely populated by polymorphisms suggests that not even gene
297 conversion can be ruled out.
298 The genetic mechanisms for concerted evolution do not differ only in terms of expected variability,
299 but also as range of action. In fact, gene conversion is expected to be effective within less than 1.5
300 Kb, whereas the unequal crossing over could cover several kb length (Hillis and Dixon 1991). The
301 rDNA region chosen in this paper, containing the two most important barcode markers proposed for
302 taxonomy in the last two decades (Schoch, Seifert et al. 2012, Kurtzman and Robnett 2013), is
303 characterized by variable length among the species and ranges between 1020 and 1350 bp. This
304 length is close to the theoretical limits of gene conversion and well inside those of unequal crossing-
305 over, giving the possibility to evaluate the effects of both mechanisms. Considering the number of
306 repeats and their length (Torres-Machorro, Hernández et al. 2010), both mechanisms are eligible for
307 operating concerted evolution and therefore to maintain one or very few copies, purging the genome
308 from the variants (Nei and Rooney 2005). This consideration poses the question on whether the
309 variations are more likely to occur in spacer regions than in ribosomal RNA encoding genes, being
310 the variation of the latter constrained by their function. Our data suggest that the ITS2 region had
311 more variants in general and with larger levels of variant reads per polymorphic site, a metric
312 particularly high in the ITS1 too (Fig. 4). These data suggest that the occurrence of variants is

313 somehow limited in the regions encoding for rRNA, maybe due to the effect of a purging selection
314 that is expected in birth-and-death evolution (Nei and Rooney 2005) , but that could occur in all the
315 three models. However, these differences of variants found among the DNA regions can be justified
316 also by the polarity effect (Nicolas and Petes 1994), although sharp gradients where not obvious
317 from our data.
318 *C. tropicalis* showed an AVF similar to *C. albicans*, a larger VSF and the smallest level of variants
319 per site. Whether this phenomenon derives from the relatively lower number of *C. tropicalis*
320 isolates considered in the work is unclear, but it seems unlikely, since this metric relies more on the
321 internal variability within the reads of each single strain than on the number of strains employed.
322 On the other hand, *C. glabrata* showed the highest level of variants per site ca. fourfold that of *C.*
323 *tropicalis.* Interestingly, *C. glabrata* is the only species, out of the four analysed, that underwent the
324 whole genome duplication (WGD) as *S. cerevisiae* (Dujon, Sherman et al. 2004). Moreover, *C.*
325 *glabrata* has two rDNA *loci* (Maleszka and Clark−Walker 1993), suggesting that the observed
326 variation can be the sum of the variants occurring in the two *loci* repeats, that could not be separated
327 in our analyses.
328 The extent of observed variation in these four species of medical importance poses a series of
329 practical questions in addition to the general and theoretical aspects outlined above. First of all this
330 variability can hamper in some cases a clear species identification with consequent problems for
331 diagnosis. Secondly, the strains of the four species seem to have a quite different variation profile,
332 suggesting that a rather independent evolution is occurring within each single strain. As a matter of
333 fact, in asexual organisms, such as those studied in this paper, the evolution is expected to occur
334 within each single genome with few random exchanges due to horizontal gene transfer. In this
335 scenario, the variation of the rDNA could be used to shed light in the different evolutionary tracks
336 followed by these organisms. This is already visible in the PLS analyses showing some strains
337 highly different from the majority of the species members in almost all species, but in particular in
338 *C. albicans* , i.e. the species with more isolates. Finally, the possibility to employ the internal

339 heterogeneity as a typing tool is particularly tempting because the same markers would be used
340 simultaneously in identification, considering the consensus sequence, and in typing using the
341 internal variants.
342
343 **REFERENCES**

344 Brown, D. D., et al. (1972). "A comparison of the ribosomal DNA's of Xenopus laevis and Xenopus
345 mulleri: the evolution of tandem genes." Journal of molecular biology **63**(1): 57-73.
346 Cardinali, G., et al. (2001). "A DNA extraction and purification method for several yeast genera."
347 Annals of Microbiology **51**(1): 121-130.
348 Corte, L., et al. (2016). "Exploring ecological modelling to investigate factors governing the
349 colonization success in nosocomial environment of Candida albicans and other pathogenic yeasts."
350 Scientific Reports **6**: 26860.
351 Dujon, B., et al. (2004). "Genome evolution in yeasts." Nature **430**(6995): 35-44.
352 Ganley, A. R. and T. Kobayashi (2007). "Highly efficient concerted evolution in the ribosomal
353 DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data."
354 Genome research **17**(2): 184-191.
355 Hillis, D. M. and M. T. Dixon (1991). "Ribosomal DNA: molecular evolution and phylogenetic
356 inference." Quarterly Review of Biology: 411-453.
357 Irinyi, L., et al. (2015). "International Society of Human and Animal Mycology (ISHAM)-ITS
358 reference DNA barcoding database—the quality controlled standard tool for routine identification
359 of human and animal pathogenic fungi." Medical Mycology: myv008.
360 Jeffreys, A. J. (1979). "DNA sequence variants in the Gγ-, Aγ-, δ-and β-globin genes of man." Cell
361 **18**(1): 1-10.
362 Kurtzman, C. P. and C. J. Robnett (2013). "Relationships among genera of the Saccharomycotina
363 (Ascomycota) from multigene phylogenetic analysis of type species." FEMS Yeast Research **13**(1):
364 23-33.

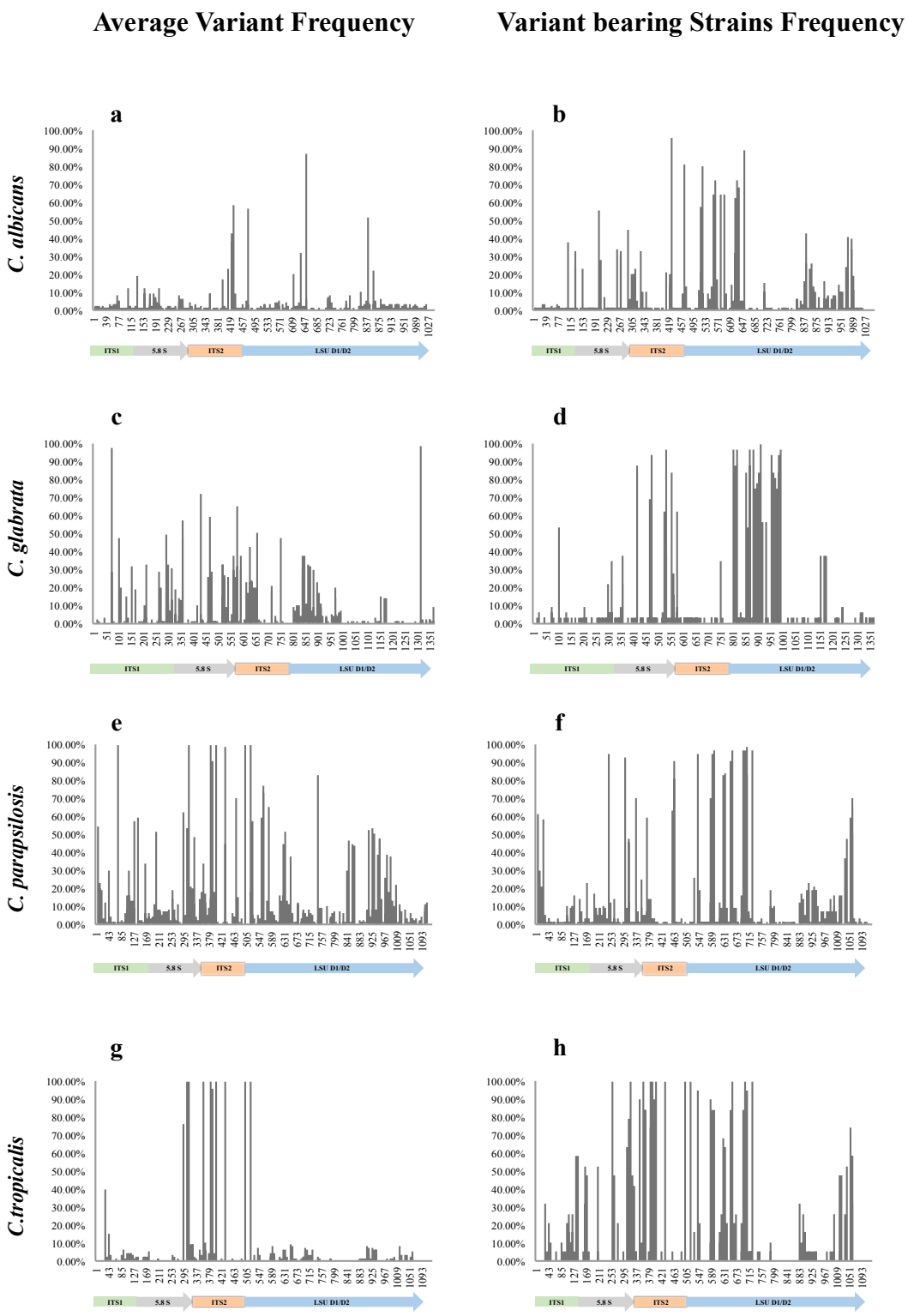
365 Maleszka, R. and G. Clark-Walker (1993). "Yeasts have a four-fold variation in ribosomal DNA
366 copy number." Yeast **9**(1): 53-58.
367 Nei, M. and A. P. Rooney (2005). "Concerted and birth-and-death evolution of multigene families."
368 Annual review of genetics **39**: 121.
369 Nicolas, A. and T. D. Petes (1994). "Polarity of meiotic gene conversion in fungi: Contrasting
370 views." Experientia **50**(3): 242-252.
371 Rooney, A. P. and T. J. Ward (2005). "Evolution of a large ribosomal RNA multigene family in
372 filamentous fungi: birth and death of a concerted evolution paradigm." Proceedings of the National
373 Academy of Sciences of the United States of America **102**(14): 5084-5089.
374 Schoch, C. L., et al. (2014). "Finding needles in haystacks: linking scientific names, reference
375 specimens and molecular data for Fungi." Database - the journal of biological database and curation
376 **2014**.
377 Schoch, C. L., et al. (2012). "Nuclear ribosomal internal transcribed spacer (ITS) region as a
378 universal DNA barcode marker for Fungi." Proc Natl Acad Sci U S A **109**: 6241 -6246.
379 Simon, U. K. and M. Weiß (2008). "Intragenomic variation of fungal ribosomal genes is higher than
380 previously thought." Molecular Biology and Evolution **25**(11): 2251-2254.
381 Torres-Machorro, A. L., et al. (2010). "Ribosomal RNA genes in eukaryotic microorganisms:
382 witnesses of phylogeny?" FEMS microbiology reviews **34**(1): 59-86.
383 West, C., et al. (2014). "Ribosomal DNA sequence heterogeneity reflects intraspecies phylogenies
384 and predicts genome structure in two contrasting yeast species." Systematic biology **63**(4): 543-554.
385
386
387
388 **FIGURE CAPTIONS.**
389 **Figure 1.** Distribution of the heterogeneity within the four studied species.



Legend. Average Variant Frequency (AVF) Variant bearing Strain frequency (VSF). Frequencies are absolute values and are reported as number of nucleotides along the sequences. VP indicates the variant percentage per site. S1V indicates the percentage of strains carrying > 1% variants per site.

394 **Figure 2.** Location of the variant and SNP site frequencies - 1% threshold.

395



396

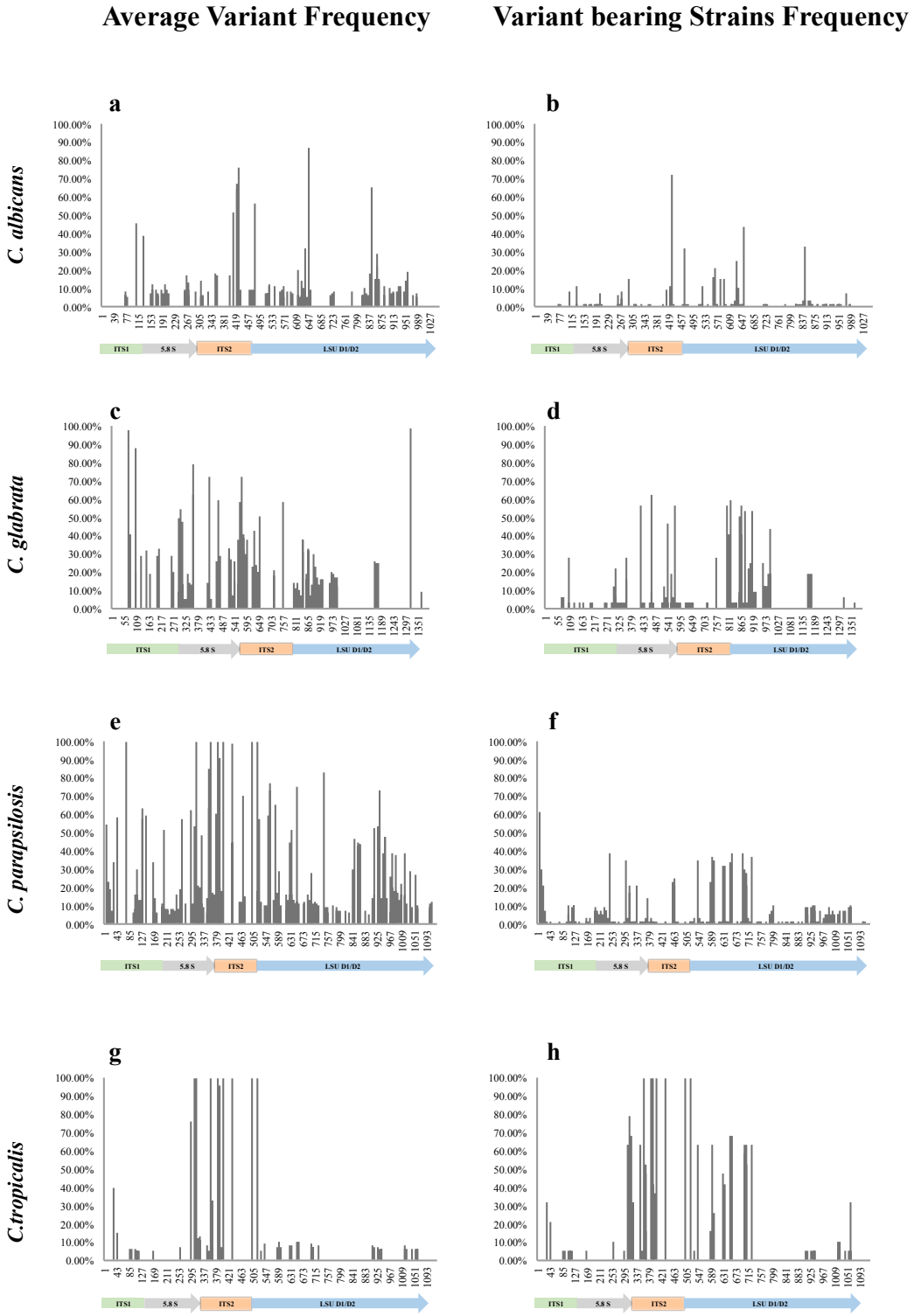
397 **Legend.** Panel (**a**, **c**, **e**, **g**) refers to AVF; Panel (**b**, **d**, **f**, **h**) refers to VSF. The length of the rDNA

398 *loci* of each species (*C. albicans* **a**, **b**; *C. glabrata* **c**, **d**; *C. parapsilosis* **e**, **f**; *C. tropicalis* **g**, **h**) were

399 collected from GenBank database (https://www.ncbi.nlm.nih.gov/).

400 **Figure 3.** Location of the variant and SNP site frequencies - 5% threshold.

401



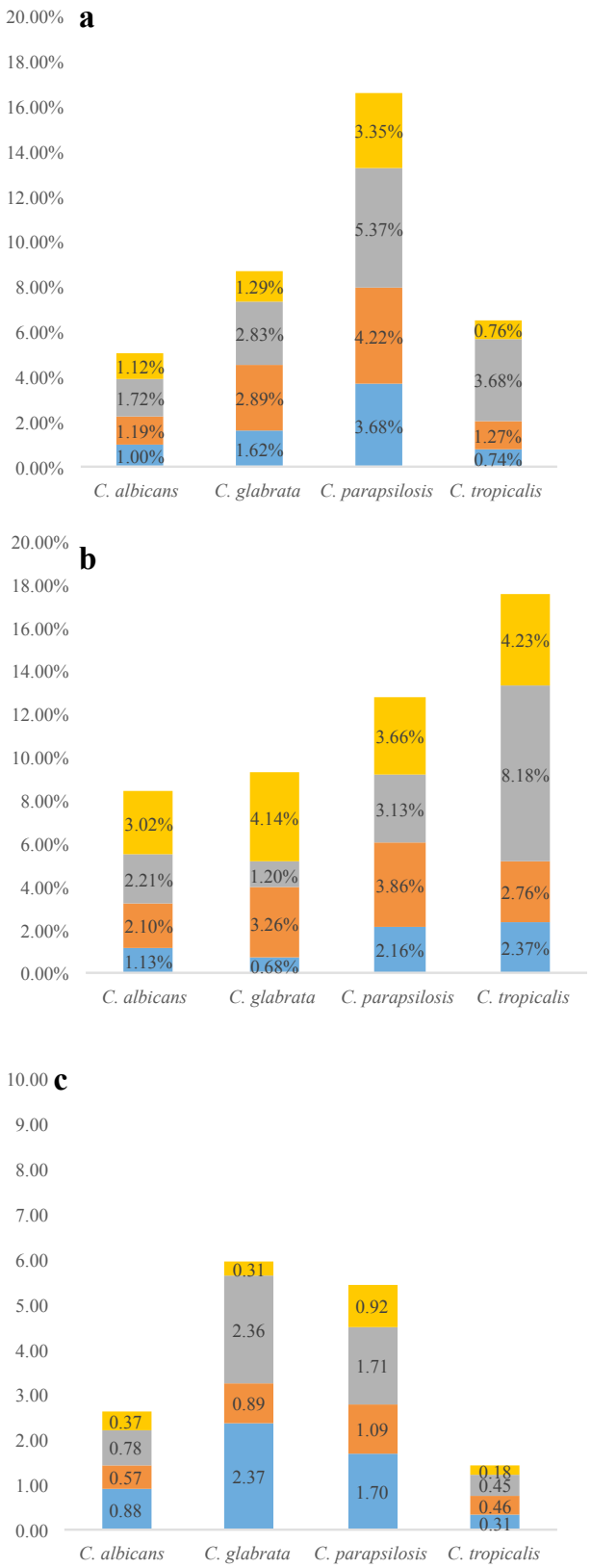
402

403 **Legend.** Panel (**a**, **c**, **e**, **g**) refers to AVF; Panel (**b**, **d**, **f**, **h**) refers to VSF. The length of the rDNA

404 *loci* of each species (*C. albicans* **a**, **b**; *C. glabrata* **c**, **d**; *C. parapsilosis* **e**, **f**; *C. tropicalis* **g**, **h**) were

405 collected from GenBank database.

406 **Figure 4.** Variability of AVF, VSF and mean variant per site throughout the ITS-LSU region.

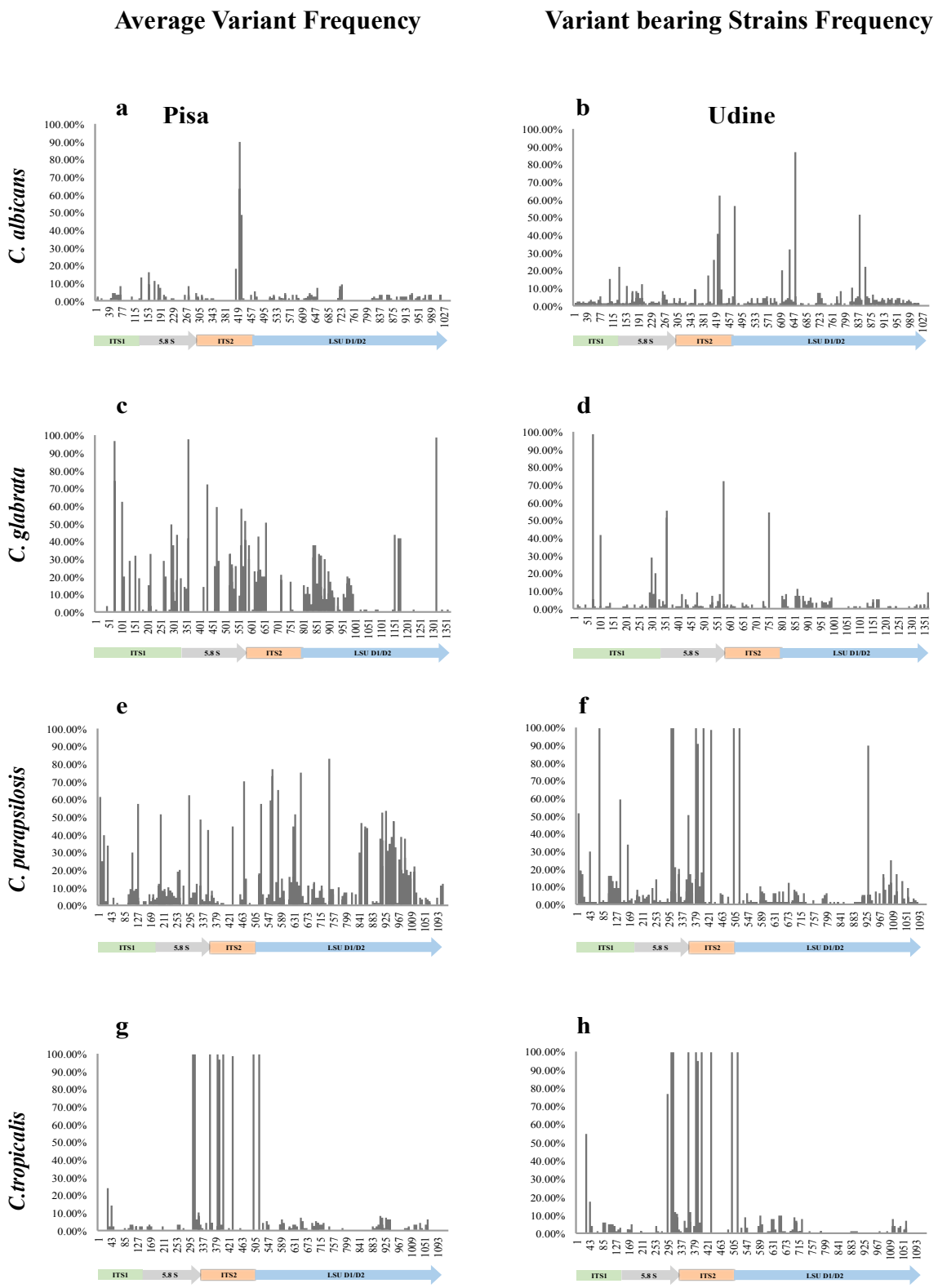


407

408 **Legend.** ITS1: light blue; 5.8S: orange; ITS2: grey; LSU: yellow.

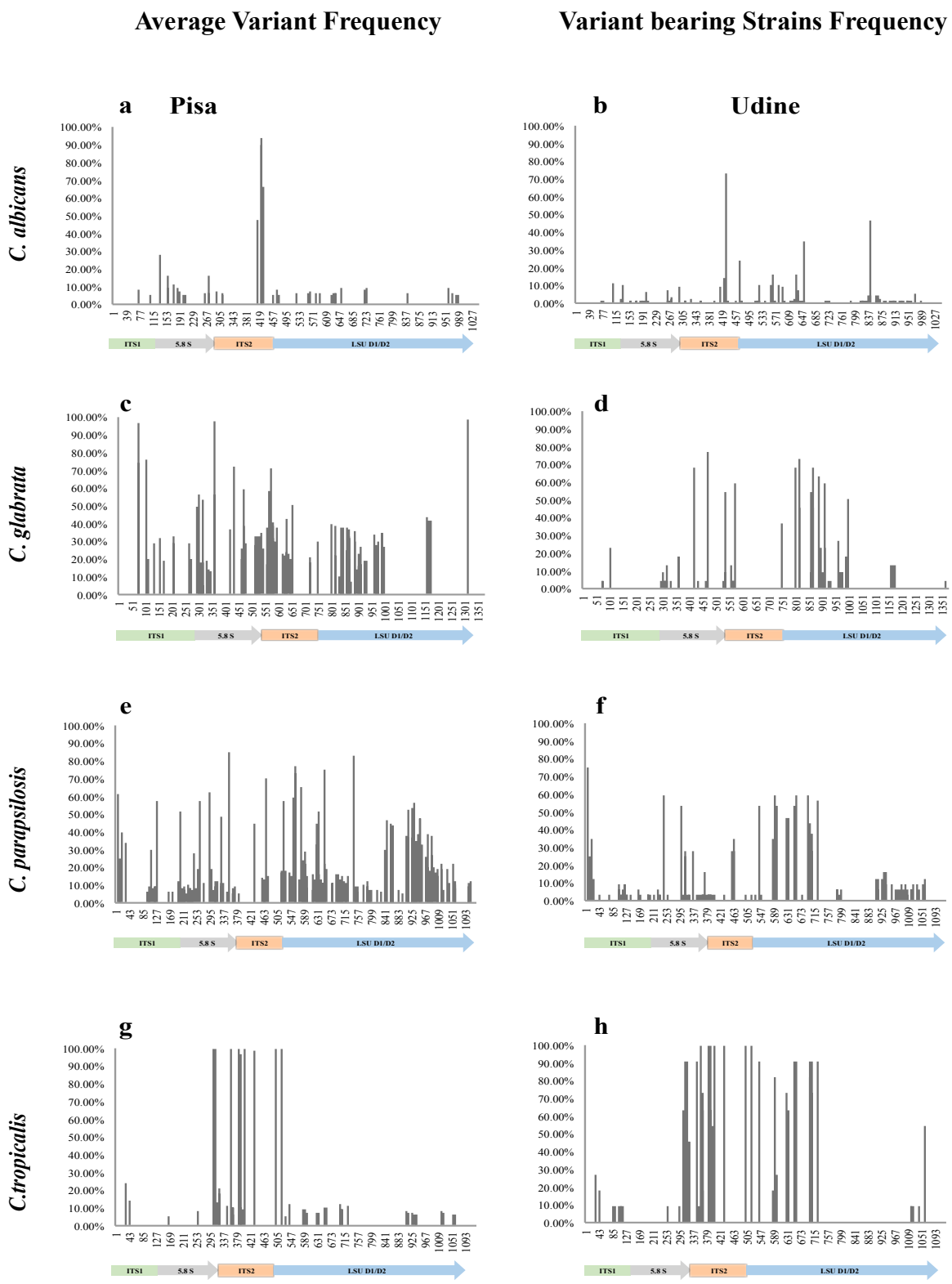
409 Panel **a**, AVG; panel **b** VSF; panel **c** AVG/VSF i.e. average number of variants per site (MNV).

Figure S1. Location of the AVF among isolates form Pisa and Udine Hospital - 1% threshold.



Legend. Panel (**a**, **c**, **e**, **g**) refers to AVF; Panel (**b**, **d**, **f**, **h**) refers to VSF. The length of the rDNA *loci* of each species (*C. albicans* **a**, **b**; *C. glabrata* **c**, **d**; *C. parapsilosis* **e**, **f**; *C. tropicalis* **g**, **h**) were collected from GenBank database.

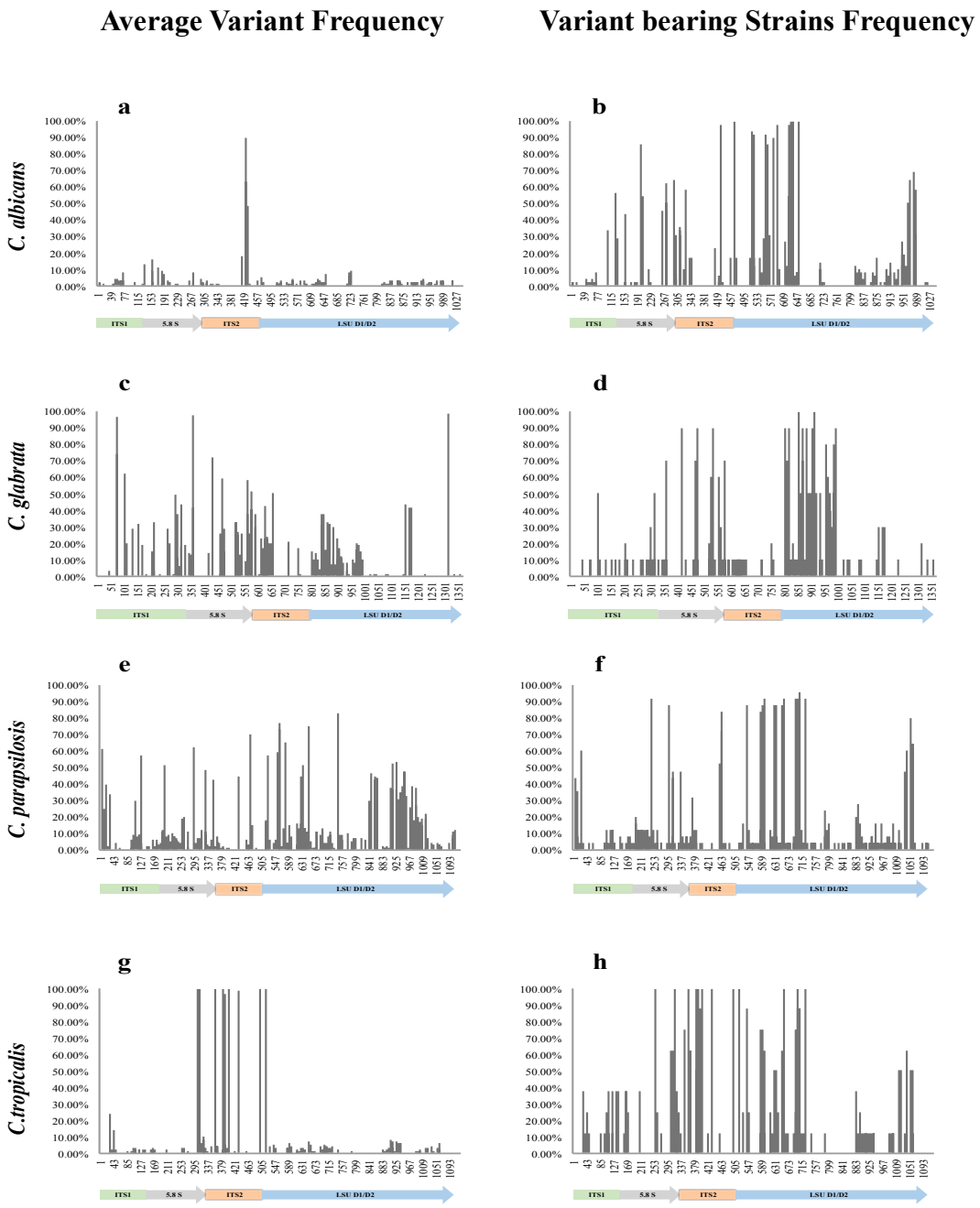
Figure S2. Location of the AVF among isolates form Pisa and Udine Hospital - 5% threshold.



Legend. Panel (**a**, **c**, **e**, **g**) refers to AVF; Panel (**b**, **d**, **f**, **h**) refers to VSF. The length of the rDNA *loci* of each species (*C. albicans* **a**, **b**; *C. glabrata* **c**, **d**; *C. parapsilosis* **e**, **f**; *C. tropicalis* **g**, **h**) were collected from GenBank database.

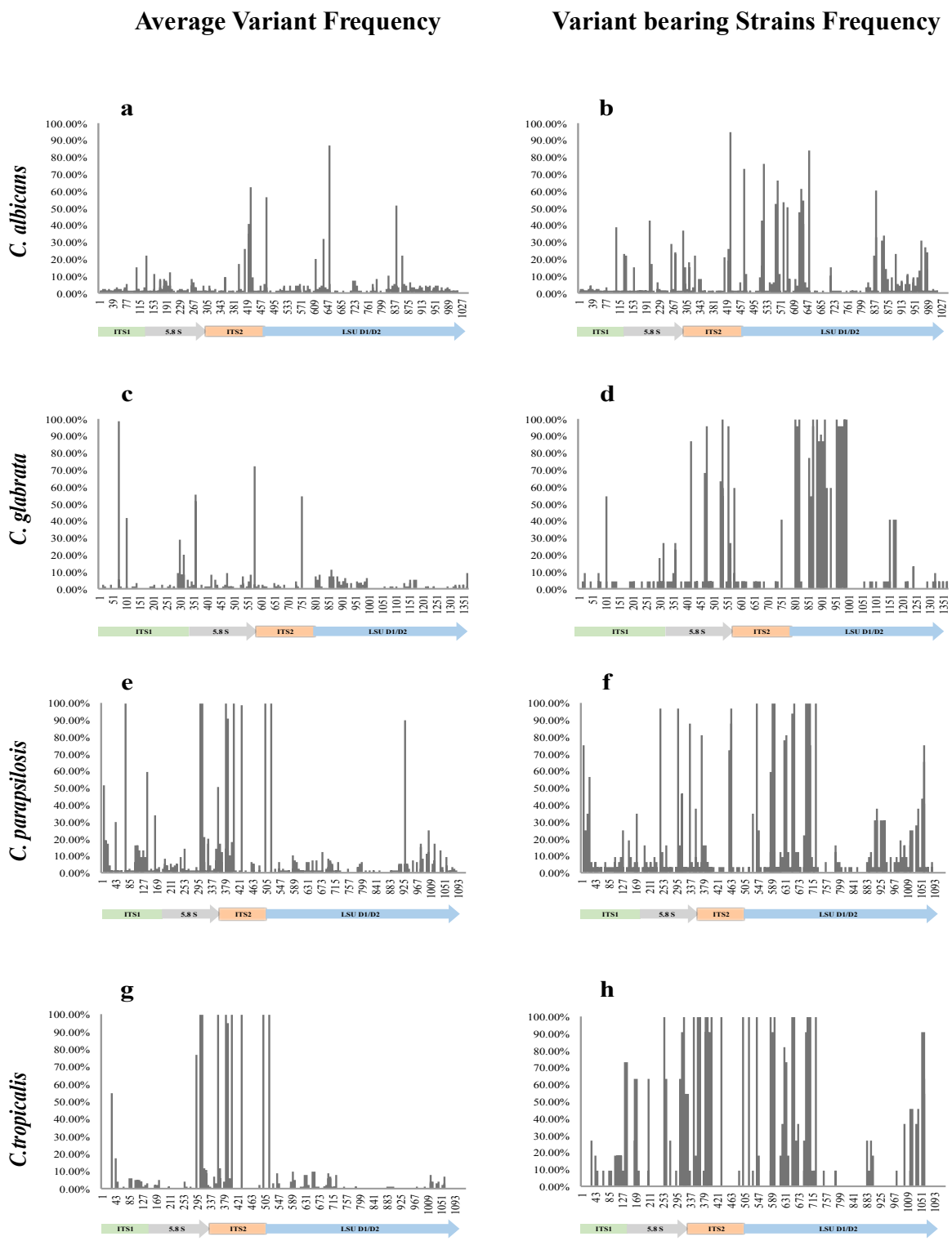
The following four figures were condensed in **Fig. S1** and **Fig. S2**, but data and images can be used in future communications and publications.

Figure S1a. Location of the variant and SNP site frequencies in Pisa hospital - 1% threshold.



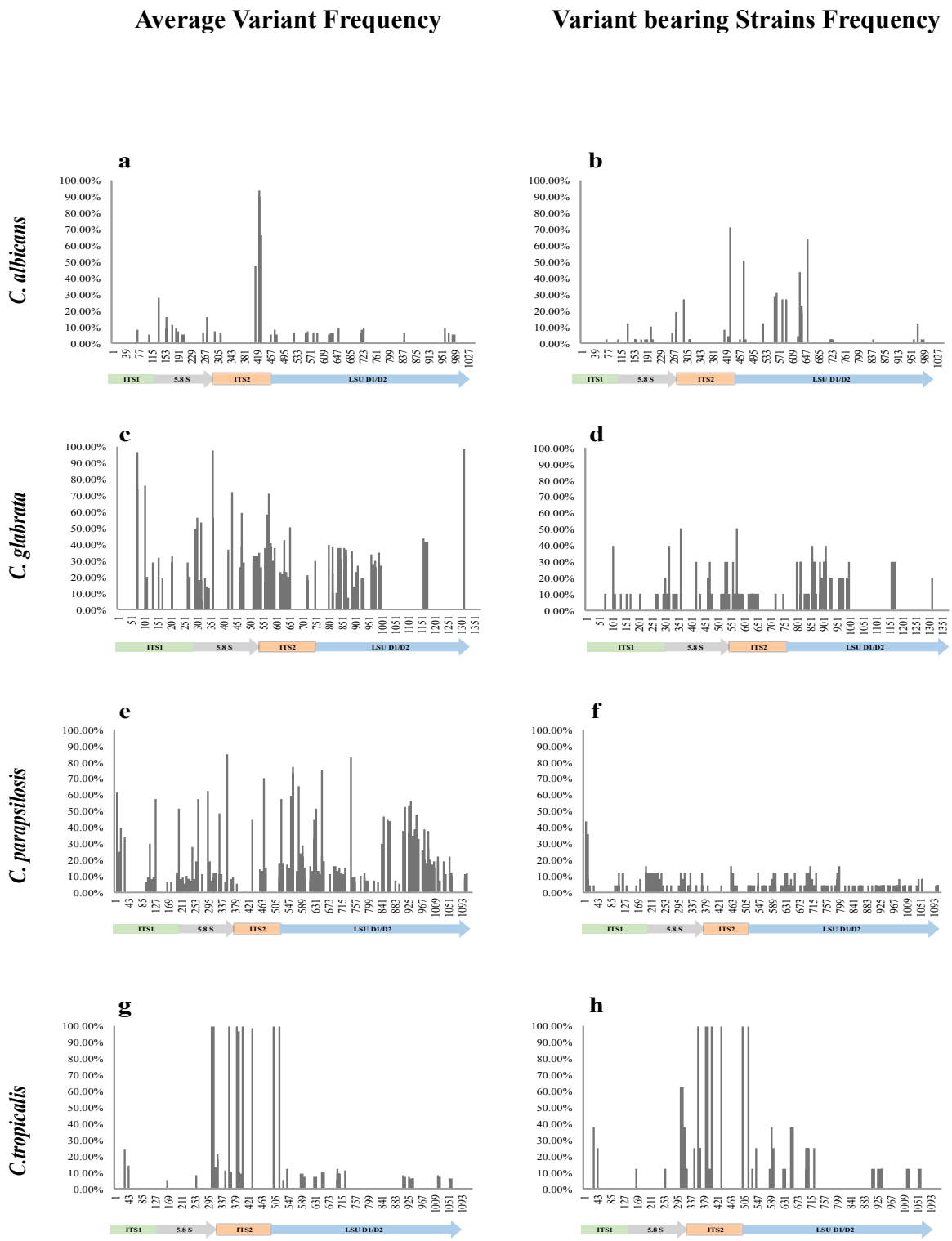
Legend. Panel (**a**, **c**, **e**, **g**) refers to AVF; Panel (**b**, **d**, **f**, **h**) refers to VSF. The length of the rDNA *loci* of each species (*C. albicans* **a**, **b**; *C. glabrata* **c**, **d**; *C. parapsilosis* **e**, **f**; *C. tropicalis* **g**, **h**) were collected from GenBank database.

Figure S1b. Location of the variant and SNP site frequencies in Udine hospital - 1% threshold.



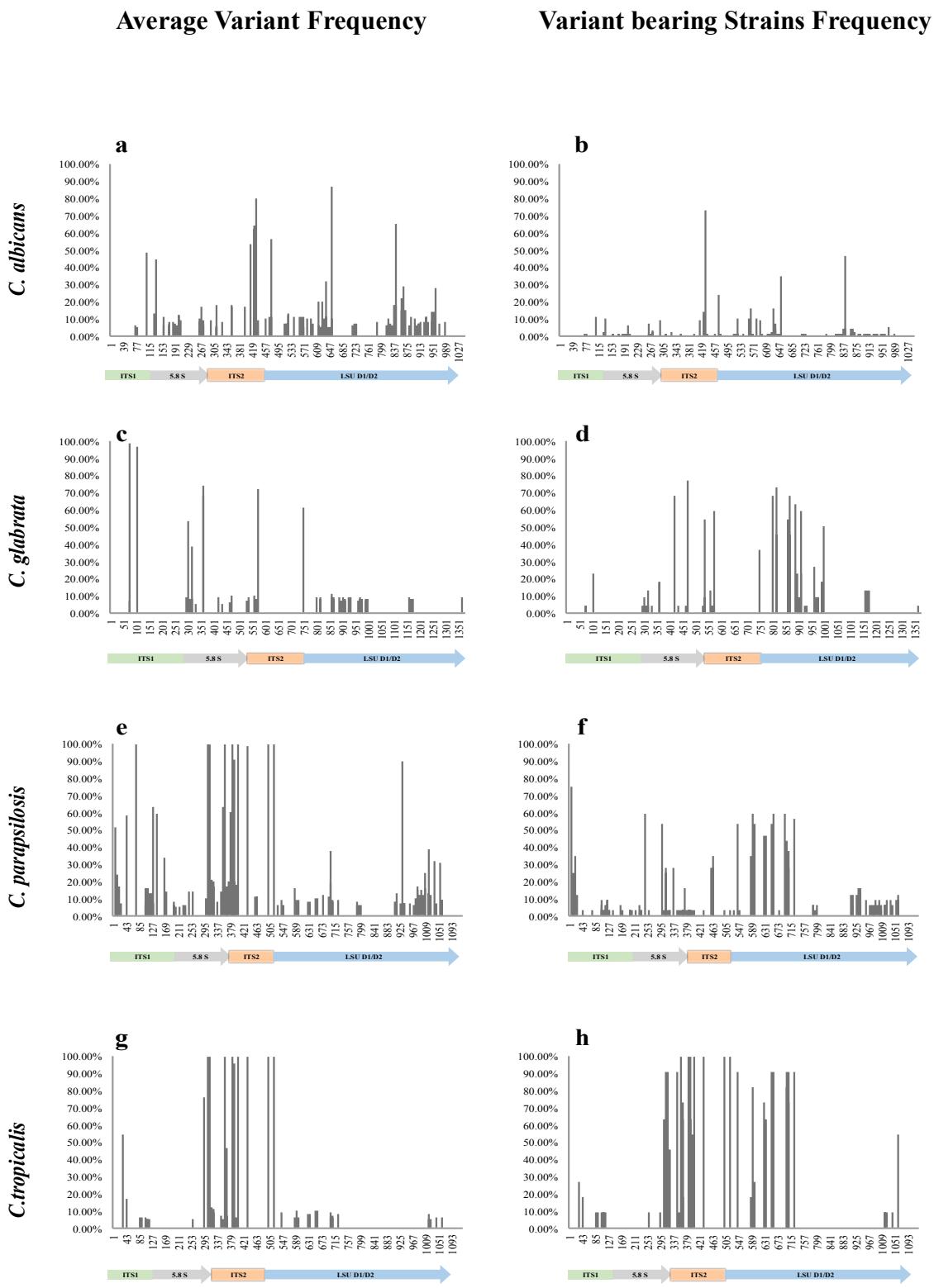
Legend. Panel (**a**, **c**, **e**, **g**) refers to AVF; Panel (**b**, **d**, **f**, **h**) refers to VSF. The length of the rDNA *loci* of each species (*C. albicans* **a**, **b**; *C. glabrata* **c**, **d**; *C. parapsilosis* **e**, **f**; *C. tropicalis* **g**, **h**) were collected from GenBank database.

Figure S2a. Location of the variant and SNP site frequencies in Pisa hospital - 5% threshold.



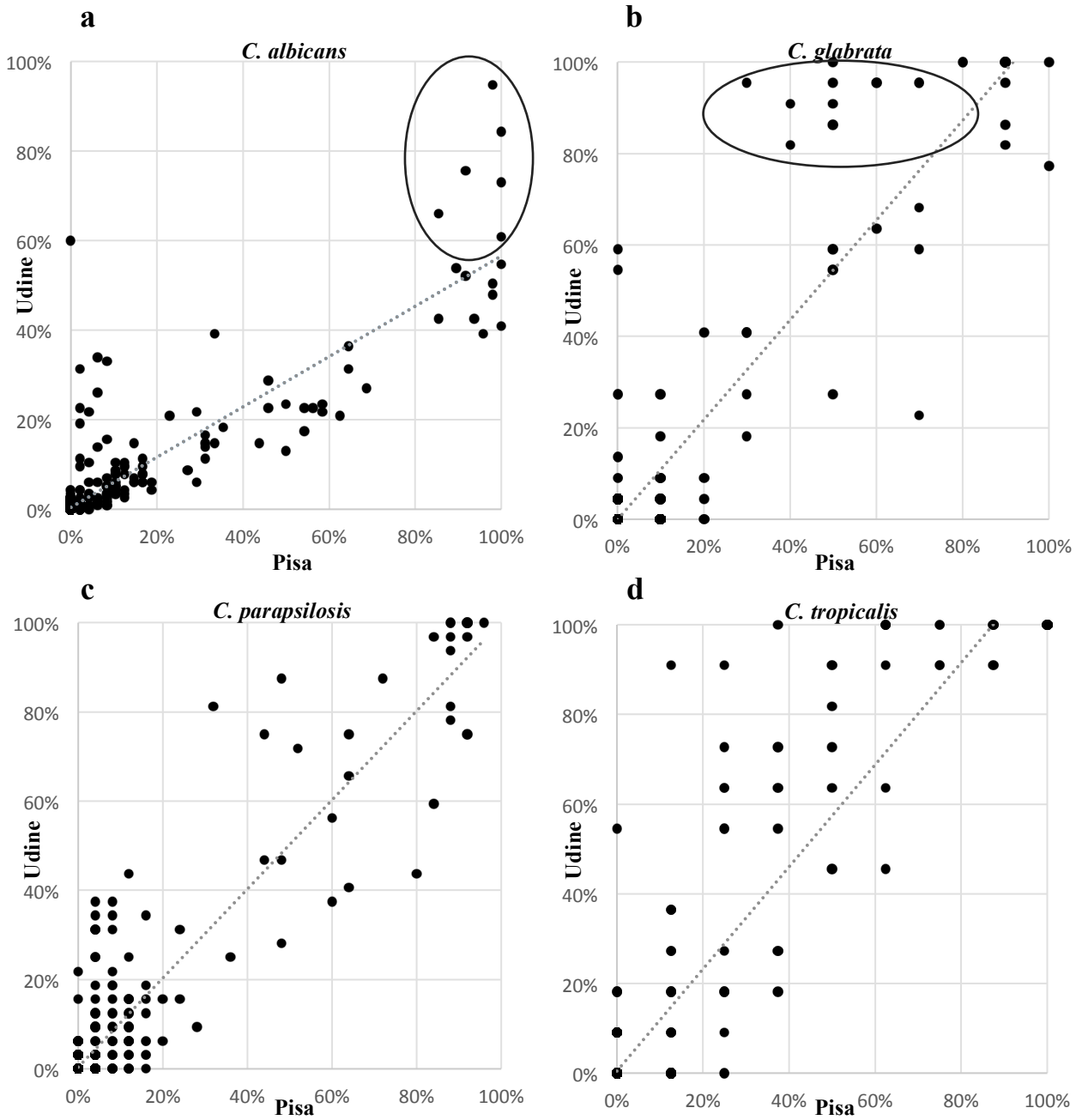
Legend. Panel (**a**, **c**, **e**, **g**) refers to AVF; Panel (**b**, **d**, **f**, **h**) refers to VSF. The length of the rDNA *loci* of each species (*C. albicans* **a**, **b**; *C. glabrata* **c**, **d**; *C. parapsilosis* **e**, **f**; *C. tropicalis* **g**, **h**) were collected from GenBank database.

Figure S2b. Location of the variant and SNP site frequencies in Udine hospital - 5% threshold.



Legend. Panel (**a**, **c**, **e**, **g**) refers to AVF; Panel (**b**, **d**, **f**, **h**) refers to VSF. The length of the rDNA *loci* of each species (*C. albicans* **a**, **b**; *C. glabrata* **c**, **d**; *C. parapsilosis* **e**, **f**; *C. tropicalis* **g**, **h**) were collected from GenBank database.

453 **Figure 5.** Regression of the SNP site frequency between two different sampling sites.



454

455

456 **Legend.** *C. albicans* correlation = 0.89, R^2= 0.80, *C. glabrata* correlation = 0.90, R^2= 0.81,

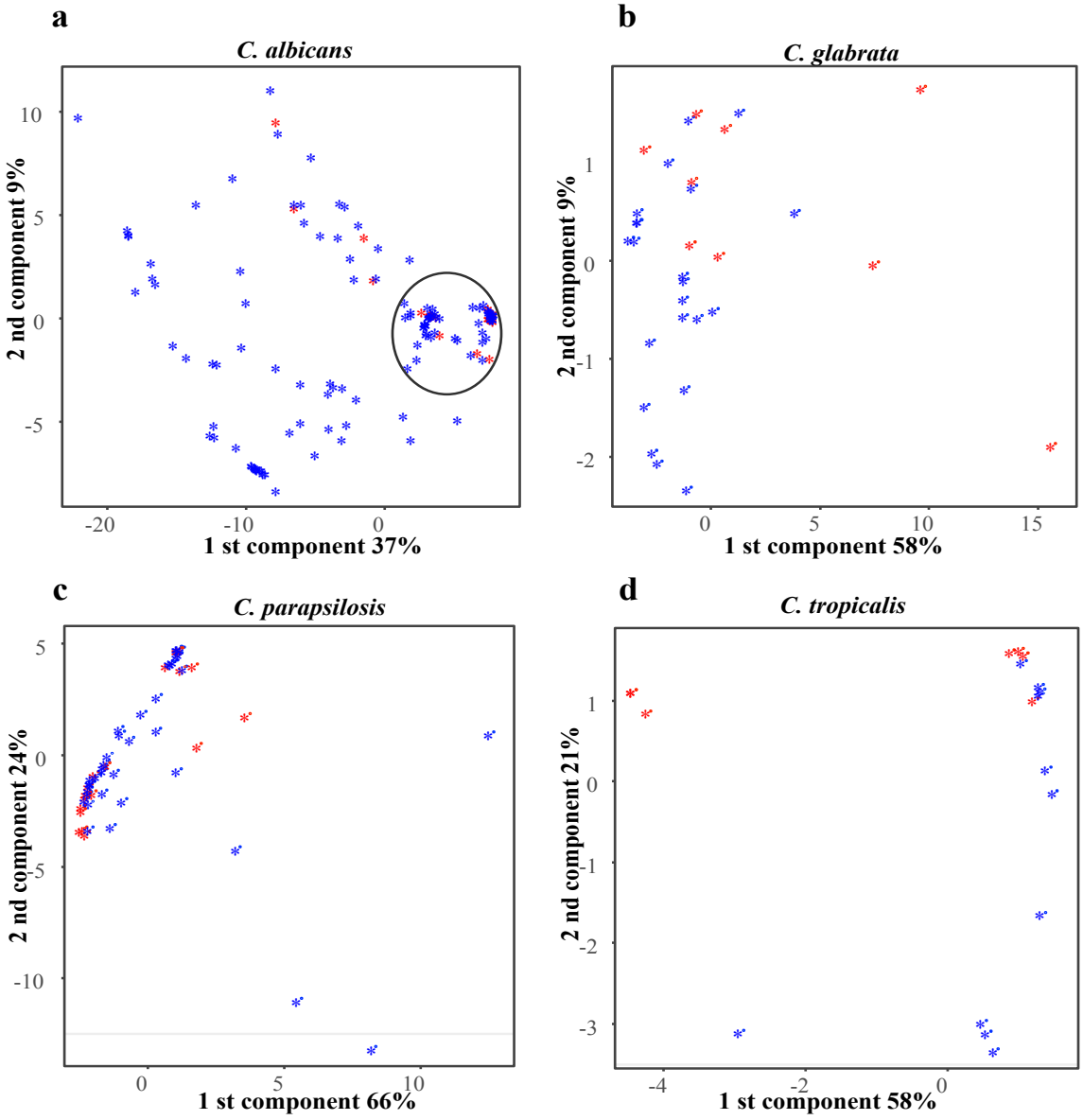
457 *C. parapsilosis* correlation = 0.93, R^2= 0.86, *C. tropicalis* correlation = 0.93, R^2= 0.86.

458

459

460

461 **Figure 6.** PLS distribution of strains isolates from two different places on the basis of their SNPs.



462

463

464 **Legend.** Red spots refer to samples from Pisa hospital, blue refer to Udine hospital.

465

466

467

468 **TABLES.**

469 **Table 1.** 271 strains employed in the study.

Strain Number	Species	Ward	City	Strain Number	Species	Ward	City
CMC 1730	*C. albicans*	Sp. Medicine	Pi	**CMC 1914**	*C. albicans*	Surgery	Ud
CMC 1966	*C. albicans*	Sp. Medicine	Pi	**CMC 1915**	*C. albicans*	Sp. Medicine	Ud
CMC 1965	*C. albicans*	Sp. Medicine	Pi	**CMC 1918**	*C. albicans*	Rehabilitation	Ud
CMC 1968	*C. albicans*	Sp. Medicine	Pi	**CMC 1919**	*C. albicans*	Gen. Medicine	Ud
CMC 1969	*C. albicans*	Sp. Medicine	Pi	**CMC 1920**	*C. albicans*	Gen. Medicine	Ud
CMC 1970	*C. albicans*	Sp. Medicine	Pi	**CMC 1921**	*C. albicans*	Gen. Medicine	Ud
CMC 1971	*C. albicans*	Sp. Medicine	Pi	**CMC 1923**	*C. albicans*	Gen. Medicine	Ud
CMC 1974	*C. albicans*	Sp. Medicine	Pi	**CMC 1925**	*C. albicans*	Sp. Medicine	Ud
CMC 1977	*C. albicans*	Sp. Medicine	Pi	**CMC 1926**	*C. albicans*	Gen. Medicine	Ud
CMC 1980	*C. albicans*	Surgery	Pi	**CMC 1927**	*C. albicans*	Surgery	Ud
CMC 1982	*C. albicans*	Sp. Medicine	Pi	**CMC 1928**	*C. albicans*	Sp. Medicine	Ud
CMC 1983	*C. albicans*	ICU	Pi	**CMC 1931**	*C. albicans*	Surgery	Ud
CMC 1985	*C. albicans*	Sp. Medicine	Pi	**CMC 1932**	*C. albicans*	Gen. Medicine	Ud
CMC 1986	*C. albicans*	ICU	Pi	**CMC 1936**	*C. albicans*	Sp. Medicine	Ud
CMC 1987	*C. albicans*	Surgery	Pi	**CMC 1937**	*C. albicans*	Gen. Medicine	Ud
CMC 1990	*C. albicans*	ICU	Pi	**CMC 1940**	*C. albicans*	Gen. Medicine	Ud
CMC 1991	*C. albicans*	Surgery	Pi	**CMC 1941**	*C. albicans*	Gen. Medicine	Ud
CMC 1992	*C. albicans*	ICU	Pi	**CMC 1942**	*C. albicans*	Surgery	Ud
CMC 1994	*C. albicans*	Sp. Medicine	Pi	**CMC 1946**	*C. albicans*	Sp. Medicine	Ud
CMC 1995	*C. albicans*	Surgery	Pi	**CMC 1952**	*C. albicans*	Gen. Medicine	Ud
CMC 1998	*C. albicans*	ICU	Pi	**CMC 1954**	*C. albicans*	Surgery	Ud
CMC 2000	*C. albicans*	Sp. Medicine	Pi	**CMC 1957**	*C. albicans*	Gen. Medicine	Ud
CMC 2001	*C. albicans*	Sp. Medicine	Pi	**CMC 1958**	*C. albicans*	Surgery	Ud
CMC 2008	*C. albicans*	Sp. Medicine	Pi	**CMC 1959**	*C. albicans*	Surgery	Ud
CMC 2010	*C. albicans*	ICU	Pi	**CMC 1960**	*C. albicans*	Gen. Medicine	Ud
CMC 2019	*C. albicans*	ICU	Pi	**CMC 1962**	*C. albicans*	Gen. Medicine	Ud
CMC 2020	*C. albicans*	Surgery	Pi	**CMC 1963**	*C. albicans*	Rehabilitation	Ud
CMC 2021	*C. albicans*	ICU	Pi	**CMC 1726**	*C. glabrata*	Sp. Medicine	Pi
CMC 2023	*C. albicans*	Gen. Medicine	Pi	**CMC 1727**	*C. glabrata*	Sp. Medicine	Pi
CMC 2025	*C. albicans*	ICU	Pi	**CMC 1731**	*C. glabrata*	Sp. Medicine	Pi
CMC 2026	*C. albicans*	Surgery	Pi	**CMC 1976**	*C. glabrata*	Sp. Medicine	Pi
CMC 2029	*C. albicans*	Sp. Medicine	Pi	**CMC 1989**	*C. glabrata*	ICU	Pi
CMC 2030	*C. albicans*	ICU	Pi	**CMC 2007**	*C. glabrata*	Sp. Medicine	Pi
CMC 2031	*C. albicans*	Surgery	Pi	**CMC 2015**	*C. glabrata*	Gen. Medicine	Pi
CMC 2032	*C. albicans*	Surgery	Pi	**CMC 2018**	*C. glabrata*	ICU	Pi
CMC 2033	*C. albicans*	Surgery	Pi	**CMC 2027**	*C. glabrata*	Surgery	Pi
CMC 2034	*C. albicans*	Gen. Medicine	Pi	**CMC 1781**	*C. glabrata*	Oncohematology	Ud
CMC 2035	*C. albicans*	Sp. Medicine	Pi	**CMC 1782**	*C. glabrata*	ICU	Ud
CMC 2036	*C. albicans*	Surgery	Pi	**CMC 1796**	*C. glabrata*	Sp. Medicine	Ud
CMC 2037	*C. albicans*	Gen. Medicine	Pi	**CMC 1807**	*C. glabrata*	Gen. Medicine	Ud
CMC 2042	*C. albicans*	ICU	Pi	**CMC 1813**	*C. glabrata*	Gen. Medicine	Ud

CMC 2043	*C. albicans*	Gen. Medicine	Pi	**CMC 1817**	*C. glabrata*	Gen. Medicine	Ud
CMC 2045	*C. albicans*	Sp. Medicine	Pi	**CMC 1830**	*C. glabrata*	Surgery	Ud
CMC 2046	*C. albicans*	ICU	Pi	**CMC 1837**	*C. glabrata*	Gen. Medicine	Ud
CMC 2047	*C. albicans*	Sp. Medicine	Pi	**CMC 1846**	*C. glabrata*	Gen. Medicine	Ud
CMC 2048	*C. albicans*	Sp. Medicine	Pi	**CMC 1857**	*C. glabrata*	Gen. Medicine	Ud
CMC 2049	*C. albicans*	Surgery	Pi	**CMC 1861**	*C. glabrata*	Gen. Medicine	Ud
CMC 2053	*C. albicans*	Sp. Medicine	Pi	**CMC 1864**	*C. glabrata*	ICU	Ud
CMC 1768	*C. albicans*	Surgery	Ud	**CMC 1865**	*C. glabrata*	Surgery	Ud
CMC 1769	*C. albicans*	Gen. Medicine	Ud	**CMC 1884**	*C. glabrata*	Gen. Medicine	Ud
CMC 1770	*C. albicans*	Gen. Medicine	Ud	**CMC 1895**	*C. glabrata*	Surgery	Ud
CMC 1771	*C. albicans*	Gen. Medicine	Ud	**CMC 1912**	*C. glabrata*	Gen. Medicine	Ud
CMC 1773	*C. albicans*	Gen. Medicine	Ud	**CMC 1916**	*C. glabrata*	ICU	Ud
CMC 1774	*C. albicans*	ICU	Ud	**CMC 1933**	*C. glabrata*	Gen. Medicine	Ud
CMC 1776	*C. albicans*	Gen. Medicine	Ud	**CMC 1934**	*C. glabrata*	Surgery	Ud
CMC 1778	*C. albicans*	Sp. Medicine	Ud	**CMC 1938**	*C. glabrata*	Gen. Medicine	Ud
CMC 1780	*C. albicans*	Gen. Medicine	Ud	**CMC 1950**	*C. glabrata*	Sp. Medicine	Ud
CMC 1785	*C. albicans*	Sp. Medicine	Ud	**CMC 1964**	*C. glabrata*	Sp. Medicine	Ud
CMC 1786	*C. albicans*	Sp. Medicine	Ud	**CMC 1967**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1788	*C. albicans*	Surgery	Ud	**CMC 1972**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1790	*C. albicans*	Gen. Medicine	Ud	**CMC 1973**	*C. parapsilosis*	Surgery	Pi
CMC 1794	*C. albicans*	Surgery	Ud	**CMC 1975**	*C. parapsilosis*	Rehabilitation	Pi
CMC 1795	*C. albicans*	Gen. Medicine	Ud	**CMC 1979**	*C. parapsilosis*	ICU	Pi
CMC 1797	*C. albicans*	Oncohematology	Ud	**CMC 1981**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1799	*C. albicans*	ICU	Ud	**CMC 1984**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1802	*C. albicans*	ICU	Ud	**CMC 1993**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1803	*C. albicans*	ICU	Ud	**CMC 1997**	*C. parapsilosis*	Surgery	Pi
CMC 1804	*C. albicans*	Surgery	Ud	**CMC 1999**	*C. parapsilosis*	Surgery	Pi
CMC 1806	*C. albicans*	Surgery	Ud	**CMC 2005**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1811	*C. albicans*	Gen. Medicine	Ud	**CMC 2006**	*C. parapsilosis*	ICU	Pi
CMC 1815	*C. albicans*	Gen. Medicine	Ud	**CMC 2012**	*C. parapsilosis*	ICU	Pi
CMC 1816	*C. albicans*	Sp. Medicine	Ud	**CMC 2013**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1818	*C. albicans*	Gen. Medicine	Ud	**CMC 2014**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1819	*C. albicans*	Surgery	Ud	**CMC 2016**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1820	*C. albicans*	Sp. Medicine	Ud	**CMC 2022**	*C. parapsilosis*	Surgery	Pi
CMC 1821	*C. albicans*	Surgery	Ud	**CMC 2038**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1822	*C. albicans*	Surgery	Ud	**CMC 2039**	*C. parapsilosis*	Surgery	Pi
CMC 1823	*C. albicans*	Gen. Medicine	Ud	**CMC 2040**	*C. parapsilosis*	Surgery	Pi
CMC 1824	*C. albicans*	Surgery	Ud	**CMC 2044**	*C. parapsilosis*	ICU	Pi
CMC 1828	*C. albicans*	Surgery	Ud	**CMC 2050**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1829	*C. albicans*	Sp. Medicine	Ud	**CMC 2051**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1831	*C. albicans*	Surgery	Ud	**CMC 1772**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1833	*C. albicans*	Surgery	Ud	**CMC 1783**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1834	*C. albicans*	Gen. Medicine	Ud	**CMC 1787**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1835	*C. albicans*	Sp. Medicine	Ud	**CMC 1791**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1840	*C. albicans*	Surgery	Ud	**CMC 1793**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1842	*C. albicans*	Surgery	Ud	**CMC 1800**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1843	*C. albicans*	Oncohematology	Ud	**CMC 1801**	*C. parapsilosis*	Sp. Medicine	Ud

CMC 1844	*C. albicans*	Gen. Medicine	Ud	**CMC 1805**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1845	*C. albicans*	Gen. Medicine	Ud	**CMC 1809**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1847	*C. albicans*	Gen. Medicine	Ud	**CMC 1814**	*C. parapsilosis*	Oncohematology	Ud
CMC 1848	*C. albicans*	Sp. Medicine	Ud	**CMC 1838**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1850	*C. albicans*	ICU	Ud	**CMC 1841**	*C. parapsilosis*	Surgery	Ud
CMC 1852	*C. albicans*	Gen. Medicine	Ud	**CMC 1849**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1853	*C. albicans*	Gen. Medicine	Ud	**CMC 1851**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1854	*C. albicans*	Sp. Medicine	Ud	**CMC 1859**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1856	*C. albicans*	ICU	Ud	**CMC 1867**	*C. parapsilosis*	ICU	Ud
CMC 1858	*C. albicans*	Gen. Medicine	Ud	**CMC 1882**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1860	*C. albicans*	Gen. Medicine	Ud	**CMC 1892**	*C. parapsilosis*	Rehabilitation	Ud
CMC 1862	*C. albicans*	Gen. Medicine	Ud	**CMC 1897**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1863	*C. albicans*	Gen. Medicine	Ud	**CMC 1899**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1866	*C. albicans*	Surgery	Ud	**CMC 1902**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1868	*C. albicans*	Gen. Medicine	Ud	**CMC 1917**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1869	*C. albicans*	Sp. Medicine	Ud	**CMC 1929**	*C. parapsilosis*	ICU	Ud
CMC 1870	*C. albicans*	Sp. Medicine	Ud	**CMC 1930**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1871	*C. albicans*	Sp. Medicine	Ud	**CMC 1935**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1872	*C. albicans*	Sp. Medicine	Ud	**CMC 1939**	*C. parapsilosis*	Surgery	Ud
CMC 1873	*C. albicans*	Gen. Medicine	Ud	**CMC 1945**	*C. parapsilosis*	Oncohematology	Ud
CMC 1875	*C. albicans*	Gen. Medicine	Ud	**CMC 1948**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1876	*C. albicans*	Gen. Medicine	Ud	**CMC 1949**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1877	*C. albicans*	ICU	Ud	**CMC 1951**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1878	*C. albicans*	Gen. Medicine	Ud	**CMC 1808**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1879	*C. albicans*	Gen. Medicine	Ud	**CMC 1812**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1881	*C. albicans*	Gen. Medicine	Ud	**CMC 1826**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1885	*C. albicans*	Surgery	Ud	**CMC 1880**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1886	*C. albicans*	Gen. Medicine	Ud	**CMC 1978**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1887	*C. albicans*	ICU	Ud	**CMC 2003**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1888	*C. albicans*	Sp. Medicine	Ud	**CMC 2009**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1889	*C. albicans*	Gen. Medicine	Ud	**CMC 2017**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1890	*C. albicans*	Gen. Medicine	Ud	**CMC 2024**	*C. tropicalis*	ICU	Pi
CMC 1891	*C. albicans*	Sp. Medicine	Ud	**CMC 2041**	*C. tropicalis*	Gen. Medicine	Pi
CMC 1893	*C. albicans*	Oncohematology	Ud	**CMC 2052**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1896	*C. albicans*	Gen. Medicine	Ud	**CMC 1784**	*C. tropicalis*	Oncohematology	Ud
CMC 1898	*C. albicans*	Gen. Medicine	Ud	**CMC 1792**	*C. tropicalis*	ICU	Ud
CMC 1900	*C. albicans*	Gen. Medicine	Ud	**CMC 1798**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1901	*C. albicans*	Gen. Medicine	Ud	**CMC 1810**	*C. tropicalis*	Oncohematology	Ud
CMC 1903	*C. albicans*	Gen. Medicine	Ud	**CMC 1827**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1904	*C. albicans*	Sp. Medicine	Ud	**CMC 1836**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1905	*C. albicans*	Gen. Medicine	Ud	**CMC 1839**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1906	*C. albicans*	Gen. Medicine	Ud	**CMC 1855**	*C. tropicalis*	Oncohematology	Ud
CMC 1907	*C. albicans*	Gen. Medicine	Ud	**CMC 1874**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1909	*C. albicans*	Sp. Medicine	Ud	**CMC 1943**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1910	*C. albicans*	ICU	Ud	**CMC 1953**	*C. tropicalis*	Oncohematology	Ud
CMC 1911	*C. albicans*	Gen. Medicine	Ud	**CMC 1956**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1913	*C. albicans*	Gen. Medicine	Ud	**CMC 1961**	*C. tropicalis*	Gen. Medicine	Ud

470 **Table 2.** Variant Frequency and Variant carrying Strains frequencies within the species.
471

		VF		VSF	
Species	**City**	**mean**	**sd**	**mean**	**sd**
C. albicans	Pisa	0.58%	4.00%	3.19%	13.63%
	Udine	1.18%	4.96%	2.21%	8.57%
	Tot	1.20%	4.93%	2.50%	9.80%
C. glabrata	Pisa	2.08%	8.73%	2.60%	11.39%
	Udine	0.62%	4.52%	2.76%	13.73%
	Tot	1.92%	7.77%	2.71%	12.75%
C. parapsilosis	Pisa	2.86%	9.66%	3.15%	12.68%
	Udine	1.98%	9.88%	3.53%	13.63%
	Tot	3.75%	12.31%	3.36%	12.98%
C. tropicalis	Pisa	1.16%	9.13%	3.87%	15.34%
	Udine	1.30%	9.57%	4.86%	18.78%
	Tot	1.34%	9.49%	4.44%	17.05%

472

473

474 **Table 3.** Correlation analysis of the AVF among the four rDNA region.

	ITS1	5.8S	ITS2	LSU
ITS1		0.050	**0.210**	0.008
5.8S	0.949		**0.251**	0.100
ITS2	**0.789**	**0.749**		**0.222**
LSU	0.991	0.899	**0.778**	

475

476

477 **Legend.** Lower triangular matrix reports the correlations; the upper the *p* value.

478

Paper VII

Approaches and tools for species delimitation with FT-IR and NGS in the four prevalent species of *Candida* pathogenic yeasts

CLAUDIA COLABELLA [1*], LAURA CORTE [1&], LUCA ROSCINI [1&], VOLHA SHAPAVAL[2¶], ACHIM KOHLER [2¶], VALERIA TAFINTSEVA [2¶] and GIANLUIGI CARDINALI [1,3]

[1] *Department of Pharmaceutical Sciences - Microbiology, University of Perugia, Perugia (Italy).*

[2]*Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, 1432 Ås, Norway.*

[3] *CEMIN, Centre of Excellence on Nanostructured Innovative Materials, Department of Chemistry, Biology and Biotechnology, University of Perugia, Via Elce di Sotto 8, 06123 Perugia, Italy.*

[*]**Corresponding Author**: Claudia Colabella

Dept. of Pharmaceutical Sciences – Microbiology

Borgo 20 Giugno, 74

I – 06121 PERUGIA

ITALY

e.mail: claudia.colabella@gmail.com

phone +39 075 585 6484

fax +39 075 585 6470

[¶] These authors contributed equally to this work.

[&] These authors also contributed equally to this work.

Abstract

The rapid and accurate identification of pathogen species is of crucial importance in clinical diagnosis of yeast infections. These are becoming a problem of increasing relevance in hospitals and nosocomial environments. Therefore, new rapid, high-throughput, reliable and objective identification methods are required. Among several, FT-IR spectroscopy associated with multivariate data analysis may be considered for the rapid identification and objective classification of microorganisms allowing to administer most appropriate therapy to patients in order to reduce their death rates as well as global treatment costs.

In this work we described a combined approach based on High-throughput (HT) Next Generation Sequencing (NGS) and HT Fourier Transform (FT) Infrared (IR) absorbance spectroscopy (FT-IR) applied to improve the identification of pathogenic yeast strains. A collection of 256 strains of *Candida* genus isolated in two Italian Hospitals was used. Strains were initially identified by biochemical criteria validated by the results of MALDI-TOF analysis. Then, ITS and D1/D2 LSU marker regions amplified by PCR were sequenced by NGS, aligned and classified on the bases of the differences to reference type strains (TS).

The average FT-IR absorbance spectra of whole microorganisms were acquired within the 4000-400 cm^{-1} intervals in the corresponding representative samples. After pre-processing, multivariate data analysis (MVA) by Consensus Principal Component Analysis (CPCA) was carried out. Partial Least Squares Regression (PLSR) was applied to build a classification model based on most relevant IR variables. The model was then

23 cross-validated. Initially, MVA was carried out in the NGS and FT-IR data-sets,
24 separately.

25 Four principal species were classified calculating the distances of the strains to the
26 taxonomic type strain (TS) namely *Candida albicans*, *Candida parapsilosis*, *Candida*
27 *glabrata* and *Candida tropicalis*, in decreasing order of frequency, respectively.

28 In order to improve the ability of single methods, inter- and intra-species variability was
29 then investigated by consensus principal component analysis (CPCA) which combines
30 high-dimensional data of the two complementary analytical approaches in concatenated
31 PCA blocks normalized to the same weight. Block 4 in the FT-IR model corresponding to
32 variables within 1200-700 cm^{-1} and block 5 composed of NGS distances gave similar
33 results although they showed different abilities as suggested by higher PC scores for NGS
34 classification than FT-IR classification, respectively.

35 For NGS identification type strains (TS) were used. A similar approach was
36 evaluated for FT-IR data where identification was performed considering TS and the
37 central strain (CS) of PLS model. Considering the matching to both the TS and the CS,
38 the total percentage of correct identification reached around 97.4% for *C. albicans* and
39 74% for *C. parapsilosis* while the other two species showed lower identification rates
40 when using the TS compared to using the CS. Results suggested that the identification
41 success could be due to the number of strains actually used in the PLS analysis.

42 We concluded that the absence of reliable FT-IR libraries might represent a limitation
43 to FTIR-based identification of strains. The reliable FT-IR libraries should include
44 several tens of strains for each relevant species, possibly over 50, according to our data.
45 At the same time, the panel of strains needs to be composed of well-identified strains,

possibly deriving from diverse sources and collected over an extensive time period. This implies a multidisciplinary effort of specialists working in strain isolation and maintenance, molecular taxonomy, FT-IR technique and chemo-metrics, data management and data basing.

Introduction

The correct identification and classification of fungi is essential for basic biological research such as the assessment of biodiversity, conservation, taxonomy and evolutionary biology and for those applications in which humanity and biodiversity intersect (agriculture, ecology, bioremediation and pathology) [1, 2].

To understand the biodiversity, the ecological roles and the geographical distribution of pathogenic fungi, DNA barcoding was proven to be a powerful tool with enormous potential [3]. DNA barcoding is a global initiative designed to provide rapid, accurate, and automatable species identifications by using short, standardized gene regions as internal species markers [4]. The critical issue underlying barcoding is accuracy, defined in taxonomic terms as the capability of unbiased and unequivocal identification at the species level. Accuracy depends especially on the extent of, and the separation between, intra-specific variation and inter-specific divergence in the selected marker creating a significant barcoding ‘‘gap’’ between intra- and inter-specific variation [5]. The sequences that are unique for a single species make identification easier, but their lack of universality hamper their amplification and therefore the whole procedure [3, 6].

67 Many barcode markers have been described for fungi [7-17]. For yeasts, the D1/D2
68 domain of the nuclear large ribosomal subunit (LSU) was adopted for characterizing
69 species long before the concept of DNA barcoding was promoted [8, 18, 19]. Among the
70 region of the ribosomal operon, the internal transcribed spacer (ITS) showed a relatively
71 good level of identification, displaying the most clearly defined barcoding gap between
72 intra and interspecific variation for most species of fungi and has been adopted as their
73 universal standard barcoding region [17]. In addition, ITS displays high robust PCR
74 amplification fidelity (>90% success rate), a Probability of Correct Identification (PCI) of
75 about 70% and pertinence to a broad range of sample conditions [13]. The rDNA operon
76 consist of multiple copies ranging from around 50 to 100 per haploid genome in fungi
77 [20, 21].
78 Different processes can occur within individual sequence heterogeneity in the
79 ribosomal repeat. In some cases, these can complicate the analysis using ITS sequencing,
80 such as intra- and inter-taxon hybridization with the loss of the homogenization of the
81 ribosomal repeat in a broad range of species. In order to increase the accuracy of species
82 identification robust primers for secondary barcodes were explored [13, 22].
83 Alongside with the development of genetic techniques, phenotyping techniques also
84 undergo enormous development. Currently, there is a few modern phenotyping
85 techniques that based on their robustness and sensitivity could be considered as Next
86 Generation Phenotyping (NGP) techniques. One of them is FT-IR spectroscopy, which is
87 an emerging technique to characterize and identify fungi in many different fields like
88 food microbiology, medical diagnostics and microbial ecology [23-26]. The method has
89 been successfully applied for the identification of fungal genera such as *Penicillum* and

90 *Fusarium* spp [24], fungal phyto-pathogenes [23], for the differentiation of *Aspergillus*
91 and *Penicillium* at species and strain level [25], yeast food-related strains [27, 28] and for
92 pathogenic strains belonging to the *Candida* genus [29-33]. FT-IR spectroscopy
93 represents thus a multi-molecular method to apply in a clinical setting alone or in
94 combination with other analytical techniques.

95 FT-IR spectroscopy was established for microbial identification by Naumann and co-
96 workers in the 90ies [34]. The basic principle of FT-IR absorbance spectroscopy is the
97 absorption of vibrating chemical bonds in sample molecules at specific frequencies is
98 represented by the infrared spectrum. FT-IR spectroscopy is a high-throughput technique
99 which does not require extended sample manipulation and allowing to achieve massive
100 and rapid molecular information of samples at very low running costs [30, 34-36].

101 Recent advances in the development of high-throughput sample preparation
102 techniques, allow cultivation of fungi in 96-microwell plates and measurement by high-
103 throughput FT-IR spectroscopy employing 384-well plates for FT-IR measurements after
104 one day growth for yeasts and five days growth for filamentous fungi [37-39]. Since the
105 FT-IR phenotype represents a biochemical fingerprint of the cells, growth media and
106 growth conditions need to be controlled strictly [38, 40-42]. Contrarily, the high
107 sensitivity of the FT-IR biochemical fingerprint towards phenotypic changes offers a
108 great opportunity to elucidate taxonomy by acquiring FT-IR spectra of microorganisms
109 using different, but defined and tailored growth media [38], an approach which is used
110 for genome-wide phenotyping via growth parameters [43].

111 Identification of microorganisms via FT-IR fingerprints can be accomplished by the
112 use of spectral databases. Comprehensive databases have been established covering a

113 large range of species and genera by the use of reference strains [44]. When suitable
114 databases are established, spectra of unknown strains can be compared with the reference
115 database spectra and strains can be rapidly identified at genus, species and sometimes
116 even at the strain level. Since identification of pathogenic yeast is crucial for mortality
117 rates of hospitalised patients [32], the implementation of rapid identification via FT-IR
118 spectroscopy may reduce death rates in patients and social costs related infectious
119 diseases [45, 46].

120 The aim of this paper is to compare identification and classification of pathogenic
121 *Candida* yeasts species via FT-IR spectroscopy and DNA barcoding as well as to develop
122 more accurate and rapid approaches to identify pathogens. To this purpose, a large set of
123 pathogenic yeasts was employed including the three species, *C. albicans*, *C. glabrata* and
124 *C. tropicalis*, which are commonly found in medical environments. In addition, *C.*
125 *parapsilosis* strains that are commonly found in natural and food environments [47] were
126 analysed by FT-IR spectroscopy and DNA barcoding.

127 In order to evaluate the capability of FT-IR spectroscopy and DNA barcoding in
128 describing inter- species and intra-species variability, results of both methods were
129 integrated into one data model by so-called consensus principal component analysis
130 (CPCA). In order to account for the problems deriving from the inherent species
131 variability, a panel of several strains per species was used. We further evaluated to what
132 extend FT-IR spectroscopy and DNA barcoding can be used for identification. In order to
133 accomplish this, we performed identification based on FT-IR spectroscopy via
134 multivariate analysis and we considered taxonomic issues, such as the relation of strains
135 to the type strain.

136

137 # Materials and Methods

138

139 ## Strains and growth conditions

140 We analysed a collection of 286 strains, belonging to opportunistic species of *Candida*
141 genus isolated from two Italian Hospitals (Pisa and Udine) and included in the Cemin
142 Microbial Collection of the Microbial Genetics and Phylogenesis Laboratory of Cemin
143 (Centre of Excellence on Nanostructured Innovative Materials for Chemicals, Physical
144 and Biomedical Applications - University of Perugia). All strains were isolated from
145 patient blood cultures and extensively described in a medical ecology paper [48].

146 Twelve species were isolated from both hospitals, with *C. albicans*, *C. glabrata*, *C.*
147 *parapsilosis* and *C. tropicalis*, representing the vast majority of the isolates. 256 strains of
148 these four species and the respective type strains were employed in this study (Table 1).

149

150 **Table 1.** 256 strains employed in the study.

151

Strain Number	Species	Ward	City	Strain Number	Species	Ward	City
CMC 1965	*C. albicans*	Sp. Medicine	Pi	**CMC 1913**	*C. albicans*	Gen. Medicine	Ud
CMC 1966	*C. albicans*	Sp. Medicine	Pi	**CMC 1914**	*C. albicans*	Surgery	Ud
CMC 1969	*C. albicans*	Sp. Medicine	Pi	**CMC 1915**	*C. albicans*	Sp. Medicine	Ud
CMC 1970	*C. albicans*	Sp. Medicine	Pi	**CMC 1918**	*C. albicans*	Rehabilitation	Ud
CMC 1974	*C. albicans*	Sp. Medicine	Pi	**CMC 1919**	*C. albicans*	Gen. Medicine	Ud
CMC 1977	*C. albicans*	Sp. Medicine	Pi	**CMC 1920**	*C. albicans*	Gen. Medicine	Ud
CMC 1980	*C. albicans*	Surgery	Pi	**CMC 1921**	*C. albicans*	Gen. Medicine	Ud
CMC 1982	*C. albicans*	Sp. Medicine	Pi	**CMC 1923**	*C. albicans*	Gen. Medicine	Ud
CMC 1983	*C. albicans*	ICU	Pi	**CMC 1925**	*C. albicans*	Sp. Medicine	Ud

CMC 1985	*C. albicans*	Sp. Medicine	Pi	**CMC 1926**	*C. albicans*	Gen. Medicine	Ud
CMC 1986	*C. albicans*	ICU	Pi	**CMC 1927**	*C. albicans*	Surgery	Ud
CMC 1987	*C. albicans*	Surgery	Pi	**CMC 1928**	*C. albicans*	Sp. Medicine	Ud
CMC 1988	*C. albicans*	Surgery	Pi	**CMC 1931**	*C. albicans*	Surgery	Ud
CMC 1990	*C. albicans*	ICU	Pi	**CMC 1932**	*C. albicans*	Gen. Medicine	Ud
CMC 1991	*C. albicans*	Surgery	Pi	**CMC 1936**	*C. albicans*	Sp. Medicine	Ud
CMC 1992	*C. albicans*	ICU	Pi	**CMC 1937**	*C. albicans*	Gen. Medicine	Ud
CMC 1994	*C. albicans*	Sp. Medicine	Pi	**CMC 1940**	*C. albicans*	Gen. Medicine	Ud
CMC 1998	*C. albicans*	ICU	Pi	**CMC 1941**	*C. albicans*	Gen. Medicine	Ud
CMC 2000	*C. albicans*	Sp. Medicine	Pi	**CMC 1942**	*C. albicans*	Surgery	Ud
CMC 2001	*C. albicans*	Sp. Medicine	Pi	**CMC 1946**	*C. albicans*	Sp. Medicine	Ud
CMC 2008	*C. albicans*	Sp. Medicine	Pi	**CMC 1952**	*C. albicans*	Gen. Medicine	Ud
CMC 2019	*C. albicans*	ICU	Pi	**CMC 1954**	*C. albicans*	Surgery	Ud
CMC 2020	*C. albicans*	Surgery	Pi	**CMC 1957**	*C. albicans*	Gen. Medicine	Ud
CMC 2023	*C. albicans*	Gen. Medicine	Pi	**CMC 1958**	*C. albicans*	Surgery	Ud
CMC 2025	*C. albicans*	ICU	Pi	**CMC 1959**	*C. albicans*	Surgery	Ud
CMC 2026	*C. albicans*	Surgery	Pi	**CMC 1960**	*C. albicans*	Gen. Medicine	Ud
CMC 2029	*C. albicans*	Sp. Medicine	Pi	**CMC 1962**	*C. albicans*	Gen. Medicine	Ud
CMC 2030	*C. albicans*	ICU	Pi	**CMC 1963**	*C. albicans*	Rehabilitation	Ud
CMC 2031	*C. albicans*	Surgery	Pi	**CMC 1976**	*C. glabrata*	Sp. Medicine	Pi
CMC 2033	*C. albicans*	Surgery	Pi	**CMC 1989**	*C. glabrata*	ICU	Pi
CMC 2034	*C. albicans*	Gen. Medicine	Pi	**CMC 2007**	*C. glabrata*	Sp. Medicine	Pi
CMC 2035	*C. albicans*	Sp. Medicine	Pi	**CMC 2015**	*C. glabrata*	Gen. Medicine	Pi
CMC 2036	*C. albicans*	Surgery	Pi	**CMC 2018**	*C. glabrata*	ICU	Pi
CMC 2037	*C. albicans*	Gen. Medicine	Pi	**CMC 2027**	*C. glabrata*	Surgery	Pi
CMC 2042	*C. albicans*	ICU	Pi	**CMC 2032**	*C. glabrata*	Surgery	Pi
CMC 2043	*C. albicans*	Gen. Medicine	Pi	**CMC 1782**	*C. glabrata*	ICU	Ud
CMC 2045	*C. albicans*	Sp. Medicine	Pi	**CMC 1807**	*C. glabrata*	Gen. Medicine	Ud
CMC 2046	*C. albicans*	ICU	Pi	**CMC 1813**	*C. glabrata*	Gen. Medicine	Ud
CMC 2048	*C. albicans*	Sp. Medicine	Pi	**CMC 1817**	*C. glabrata*	Gen. Medicine	Ud
CMC 2049	*C. albicans*	Surgery	Pi	**CMC 1830**	*C. glabrata*	Surgery	Ud
CMC 2053	*C. albicans*	Sp. Medicine	Pi	**CMC 1832**	*C. glabrata*	Oncohematology	Ud
CMC 1768	*C. albicans*	Surgery	Ud	**CMC 1837**	*C. glabrata*	Gen. Medicine	Ud
CMC 1769	*C. albicans*	Gen. Medicine	Ud	**CMC 1846**	*C. glabrata*	Gen. Medicine	Ud
CMC 1770	*C. albicans*	Gen. Medicine	Ud	**CMC 1857**	*C. glabrata*	Gen. Medicine	Ud
CMC 1771	*C. albicans*	Gen. Medicine	Ud	**CMC 1860**	*C. glabrata*	Gen. Medicine	Ud
CMC 1773	*C. albicans*	Gen. Medicine	Ud	**CMC 1861**	*C. glabrata*	Gen. Medicine	Ud
CMC 1774	*C. albicans*	ICU	Ud	**CMC 1864**	*C. glabrata*	ICU	Ud
CMC 1776	*C. albicans*	Gen. Medicine	Ud	**CMC 1865**	*C. glabrata*	Surgery	Ud
CMC 1777	*C. albicans*	ICU	Ud	**CMC 1884**	*C. glabrata*	Gen. Medicine	Ud
CMC 1778	*C. albicans*	Sp. Medicine	Ud	**CMC 1894**	*C. glabrata*	Rehabilitation	Ud

CMC 1780	*C. albicans*	Gen. Medicine	Ud	**CMC 1895**	*C. glabrata*	Surgery	Ud
CMC 1785	*C. albicans*	Sp. Medicine	Ud	**CMC 1912**	*C. glabrata*	Gen. Medicine	Ud
CMC 1786	*C. albicans*	Sp. Medicine	Ud	**CMC 1916**	*C. glabrata*	ICU	Ud
CMC 1788	*C. albicans*	Surgery	Ud	**CMC 1933**	*C. glabrata*	Gen. Medicine	Ud
CMC 1790	*C. albicans*	Gen. Medicine	Ud	**CMC 1934**	*C. glabrata*	Surgery	Ud
CMC 1794	*C. albicans*	Surgery	Ud	**CMC 1938**	*C. glabrata*	Gen. Medicine	Ud
CMC 1795	*C. albicans*	Gen. Medicine	Ud	**CMC 1950**	*C. glabrata*	Sp. Medicine	Ud
CMC 1797	*C. albicans*	Oncohematology	Ud	**CMC 1964**	*C. glabrata*	Sp. Medicine	Ud
CMC 1799	*C. albicans*	ICU	Ud	**CMC 1972**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1802	*C. albicans*	ICU	Ud	**CMC 1979**	*C. parapsilosis*	ICU	Pi
CMC 1803	*C. albicans*	ICU	Ud	**CMC 1973**	*C. parapsilosis*	Surgery	Pi
CMC 1804	*C. albicans*	Surgery	Ud	**CMC 1981**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1806	*C. albicans*	Surgery	Ud	**CMC 2006**	*C. parapsilosis*	ICU	Pi
CMC 1811	*C. albicans*	Gen. Medicine	Ud	**CMC 2012**	*C. parapsilosis*	ICU	Pi
CMC 1815	*C. albicans*	Gen. Medicine	Ud	**CMC 2013**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1816	*C. albicans*	Sp. Medicine	Ud	**CMC 2014**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1818	*C. albicans*	Gen. Medicine	Ud	**CMC 2016**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1819	*C. albicans*	Surgery	Ud	**CMC 2022**	*C. parapsilosis*	Surgery	Pi
CMC 1820	*C. albicans*	Sp. Medicine	Ud	**CMC 2038**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1821	*C. albicans*	Surgery	Ud	**CMC 2039**	*C. parapsilosis*	Surgery	Pi
CMC 1822	*C. albicans*	Surgery	Ud	**CMC 2040**	*C. parapsilosis*	Surgery	Pi
CMC 1823	*C. albicans*	Gen. Medicine	Ud	**CMC 2044**	*C. parapsilosis*	ICU	Pi
CMC 1824	*C. albicans*	Surgery	Ud	**CMC 2050**	*C. parapsilosis*	Sp. Medicine	Pi
CMC 1828	*C. albicans*	Surgery	Ud	**CMC 1772**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1829	*C. albicans*	Sp. Medicine	Ud	**CMC 1781**	*C. parapsilosis*	Oncohematology	Ud
CMC 1831	*C. albicans*	Surgery	Ud	**CMC 1783**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1833	*C. albicans*	Surgery	Ud	**CMC 1787**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1834	*C. albicans*	Gen. Medicine	Ud	**CMC 1791**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1835	*C. albicans*	Sp. Medicine	Ud	**CMC 1792**	*C. parapsilosis*	ICU	Ud
CMC 1840	*C. albicans*	Surgery	Ud	**CMC 1793**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1842	*C. albicans*	Surgery	Ud	**CMC 1796**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1843	*C. albicans*	Oncohematology	Ud	**CMC 1800**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1844	*C. albicans*	Gen. Medicine	Ud	**CMC 1801**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1845	*C. albicans*	Gen. Medicine	Ud	**CMC 1805**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1847	*C. albicans*	Gen. Medicine	Ud	**CMC 1808**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1848	*C. albicans*	Sp. Medicine	Ud	**CMC 1809**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1849	*C. albicans*	Sp. Medicine	Ud	**CMC 1812**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1850	*C. albicans*	ICU	Ud	**CMC 1814**	*C. parapsilosis*	Oncohematology	Ud
CMC 1852	*C. albicans*	Gen. Medicine	Ud	**CMC 1826**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1853	*C. albicans*	Gen. Medicine	Ud	**CMC 1838**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1854	*C. albicans*	Sp. Medicine	Ud	**CMC 1841**	*C. parapsilosis*	Surgery	Ud

CMC 1856	*C. albicans*	ICU	Ud	**CMC 1851**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1858	*C. albicans*	Gen. Medicine	Ud	**CMC 1859**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1862	*C. albicans*	Gen. Medicine	Ud	**CMC 1867**	*C. parapsilosis*	ICU	Ud
CMC 1863	*C. albicans*	Gen. Medicine	Ud	**CMC 1880**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1866	*C. albicans*	Surgery	Ud	**CMC 1892**	*C. parapsilosis*	Rehabilitation	Ud
CMC 1868	*C. albicans*	Gen. Medicine	Ud	**CMC 1899**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1869	*C. albicans*	Sp. Medicine	Ud	**CMC 1909**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1870	*C. albicans*	Sp. Medicine	Ud	**CMC 1922**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1871	*C. albicans*	Sp. Medicine	Ud	**CMC 1902**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1872	*C. albicans*	Sp. Medicine	Ud	**CMC 1917**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1873	*C. albicans*	Gen. Medicine	Ud	**CMC 1929**	*C. parapsilosis*	ICU	Ud
CMC 1875	*C. albicans*	Gen. Medicine	Ud	**CMC 1930**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1876	*C. albicans*	Gen. Medicine	Ud	**CMC 1935**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1877	*C. albicans*	ICU	Ud	**CMC 1939**	*C. parapsilosis*	Surgery	Ud
CMC 1878	*C. albicans*	Gen. Medicine	Ud	**CMC 1945**	*C. parapsilosis*	Oncohematology	Ud
CMC 1879	*C. albicans*	Gen. Medicine	Ud	**CMC 1948**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1881	*C. albicans*	Gen. Medicine	Ud	**CMC 1949**	*C. parapsilosis*	Gen. Medicine	Ud
CMC 1885	*C. albicans*	Surgery	Ud	**CMC 1951**	*C. parapsilosis*	Sp. Medicine	Ud
CMC 1886	*C. albicans*	Gen. Medicine	Ud	**CMC 1978**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1887	*C. albicans*	ICU	Ud	**CMC 2003**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1888	*C. albicans*	Sp. Medicine	Ud	**CMC 2009**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1889	*C. albicans*	Gen. Medicine	Ud	**CMC 2017**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1890	*C. albicans*	Gen. Medicine	Ud	**CMC 2024**	*C. tropicalis*	ICU	Pi
CMC 1891	*C. albicans*	Sp. Medicine	Ud	**CMC 2041**	*C. tropicalis*	Gen. Medicine	Pi
CMC 1893	*C. albicans*	Oncohematology	Ud	**CMC 2052**	*C. tropicalis*	Sp. Medicine	Pi
CMC 1896	*C. albicans*	Gen. Medicine	Ud	**CMC 1784**	*C. tropicalis*	Oncohematology	Ud
CMC 1897	*C. albicans*	Rehabilitation	Ud	**CMC 1798**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1898	*C. albicans*	Gen. Medicine	Ud	**CMC 1810**	*C. tropicalis*	Oncohematology	Ud
CMC 1900	*C. albicans*	Gen. Medicine	Ud	**CMC 1827**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1901	*C. albicans*	Gen. Medicine	Ud	**CMC 1836**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1903	*C. albicans*	Gen. Medicine	Ud	**CMC 1839**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1905	*C. albicans*	Gen. Medicine	Ud	**CMC 1855**	*C. tropicalis*	Oncohematology	Ud
CMC 1906	*C. albicans*	Gen. Medicine	Ud	**CMC 1874**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1907	*C. albicans*	Gen. Medicine	Ud	**CMC 1904**	*C. tropicalis*	Sp. Medicine	Ud
CMC 1908	*C. albicans*	ICU	Ud	**CMC 1953**	*C. tropicalis*	Oncohematology	Ud
CMC 1910	*C. albicans*	ICU	Ud	**CMC 1956**	*C. tropicalis*	Gen. Medicine	Ud
CMC 1911	*C. albicans*	Gen. Medicine	Ud	**CMC 1961**	*C. tropicalis*	Gen. Medicine	Ud

153 All strains were stored at -80°C in 17% glycerol right after isolation. Cultivation was
154 carried out on YEPDA (YEPD with 1.7% agar) at 37°C, following the current
155 procedures. To generate cell biomass needed for the analysis, the strains were grown in
156 YEPD broth (Yeast extract 1%, Peptone 1%, Dextrose 1%; all chemicals from Biolife,
157 Italy - http://www.biolifeitaliana.it/) at 37°C with 150 rpm shaking.
158

159 Molecular analysis and bioinformatics tools

160 Genomic DNA was extracted as indicated by Cardinali et al [49] ITS1, 5.8S, ITS2
161 rDNA genes and D1/D2 domain of the LSU were amplified with FIREPole® Taq DNA
162 Polymerase (Solis BioDyne, Estonia), using ITS1 (5'-TCCGTAGGTGAACCTGCGG) -
163 NL4 (GGTCCGTGTTTCAAGACGG) primers. The amplification protocol was carried
164 out as follows: initial denaturation at 94°C for 3 min, 30 amplification cycles (94°C for 1
165 min, 54°C for 1 min and 72°C for 1 min) and final extension at 72°C for 5 min.
166 Amplicons were subjected to electrophoresis on 1.5% agarose gel (Gellyphor, EuroClone,
167 Italy). Amplicons were sequenced with NGS PlexWell™ technologies
168 (http://www.seqwell.com/) with the same primers used for the generation of the
169 amplicons. The reads of each strain, contained in FASTAq file, were analysed with
170 Geneious R9 (v. 9.1.5, Biomatters, Auckland, New Zealand - http://www.geneious.com/).
171 Identification was carried out according to the criteria indicated in the taxonomic papers
172 dealing with LSU [8] and ITS [17, 22, 50, 51].
173 In order to obtain distance matrices for the four major species, all the consensus
174 sequences were aligned with the corresponding type strain using pairwise alignment in
175 Geneious software (Biomatters, New Zeland). The distance matrices were calculated

through the base of percentage of identical bases/residues and exported as tsv files in Microsoft Excel ®.

FT-IR measurements

For FT-IR analysis, the selected strains were grown over night in YNB (added with 2% dextrose, 1.7% agar - all products from Biolife). For each sample one colony was transferred with a calibrated platinum loop from the plate to Eppendorf tubes containing 200µl of pure water (HPLC Gradient Grade - J.T. Baker - http://www.jtbaker.com/). Of each suspension, 35µl were transferred to an IR-light-transparent silicon 96-well microtiter plate (Bruker, Germany). The samples were dried at 42°C to form films of uniform thickness in order to minimize the interference of scattering effects during the acquisition of FT-IR absorbance spectra.

FT-IR measurements were performed in transmission mode. For each sample three technical replicates were performed. All spectra were recorded in the range between 4000 and 400 cm^{-1} with a TENSOR 27 FT-IR spectrometer, equipped with HTS-XT accessory for rapid automation of the analysis (Bruker, Germany). Spectral resolution was set at 4 cm^{-1}, sampling 128 scans per sample. The OPUS version 6.5 software (Bruker, Germany) was used to carry out the quality test and to obtain a matrix of raw spectra, which was subsequently exported as ASCII file.

Pre-processing of FT-IR data

The measured FT-IR raw data consisted of 780 spectra including technical replicates. The data set was reduced to 260 spectra by averaging over technical replicates. Whole

raw spectra were pre-processed with the second derivative function by the Savitzky-Golay algorithm and 15 smoothing points [52]. The 3050-2800 cm^{-1} and 1800-700 cm^{-1} intervals were considered.

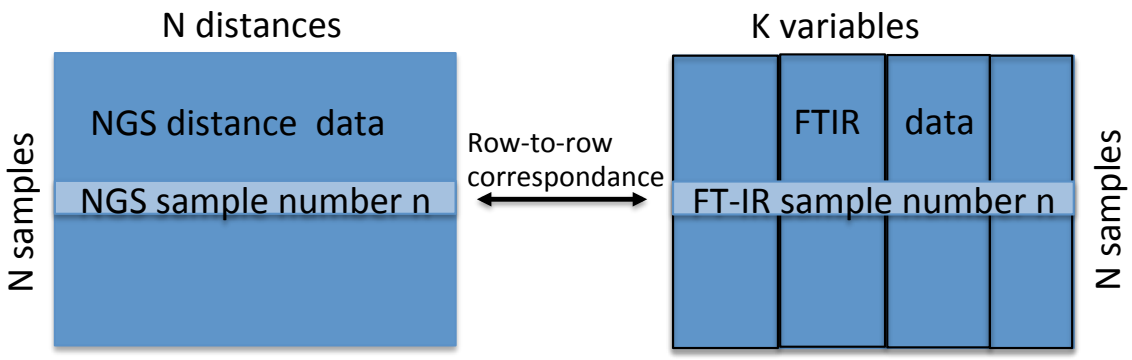
Then, Extended Multiplicative Signal Correction (EMSC) taking into account linear and quadratic components was applied [53]. Pre-processing by second derivative and EMSC is done, in order to remove physical variations such as baseline variations and variations due to the thickness of the film of microbial cells used for FT-IR transmission spectroscopy [54].

Multivariate analysis of FT-IR and NGS data

Consensus Principal Component Analysis

Consensus principal component analysis (CPCA) was applied in order to integrate NGS distance data and FT-IR absorbance data in one data model. CPCA is a so-called multiblock method that can be used to connect different types of data [55]. For CPCA, the data was organized such that a row-to-row correspondence was obtained for NGS distance data and FT-IR absorbance data (Fig 1).

Fig 1. For CPCA of FT-IR and NGS data a row-to-row correspondence needs to be obtained.



Legend. In order to integrate NGS data and FT-IR data in one data model, consensus principal component analysis is applied (CPCA). For CPCA, the data is organized such that a row-to-row correspondence between the different data blocks is obtained. The NGS distance data matrix contains N samples as rows and as columns the variables, which are the distances to all samples. The FT-IR data contains N samples as rows and as columns the absorbance values at different wavenumbers. The FT-IR data is further split into four data blocks according to groups of chemicals.

The FT-IR data was further split into four blocks, namely the FT-IR absorbance values from 3050-2800 cm^{-1} were defined as block one, the region from 1800-1500 cm^{-1} as block two, the region from 1500-1200 cm^{-1} as block three and the region from 1200-700 cm^{-1} as block four. The NGS distance matrix was defined as block five. Each block was normalized to unit variance in order to set all blocks on the same footing. This is done by normalizing each block by the Frobenius norm, thus achieving the same weight for each block [56]. CPCA is based on principal component analysis. CPCA is equivalent to performing a PCA on all five concatenated blocks, where all blocks are normalized by the Frobenius norm. By CPCA, global scores and block scores are calculated. The global

238 scores are equivalent to PCA scores obtained on concatenated and normalized blocks.
239 They represent the consensus of all blocks and allow studying global sample and variable
240 variation patterns. In addition to global scores, CPCA calculates block parameters, so-
241 called block scores and block loadings.

242 The block scores can be used to study the block sample variation patterns for each
243 consensus component, i.e. the sample variation in each block that contributes to the
244 consensus. How strongly every block contributes to the consensus can be estimated by
245 explained block variances, which are calculated for each block. In order to study
246 correlations between variables between and within the blocks, correlation loading blocks
247 can be used [57]. In correlation loading plots, the correlations between the global scores
248 and the FT-IR and genetic distances matrices are plotted. In addition, correlations
249 between global scores and group variables for each species are visualized in the same
250 plots. Species groups are represented by so-called indicator variables. Each species is
251 represented by one column of indicator variables, where a strain obtains the value one if
252 it belongs to a species and zero otherwise.

253 For the correlation loading plots, we multiply the genetic distance variables and the
254 FT-IR second derivative data by minus one in order to facilitate the interpretation of the
255 correlations with the group indicator variables. In second derivative spectra, bands appear
256 as negative peaks and are thus inversely correlated to concentrations of chemical
257 compounds. The multiplication by minus one turns the negative correlation into a positive
258 correlation, which facilitates interpretation. A similar argument applies for the genetic
259 distance matrices.

260

261 **Identification by Partial Least Squares Discriminant Analysis**

262 Partial Least Squares Regression (PLSR) [58] was used to establish classification
263 models to differentiate four groups of species. In order to establish models, the data
264 matrix X of FT-IR spectra is regressed on a matrix Y of indicator variables containing
265 group labels. When PLSR is used together with a matrix of group indicator variable a
266 matrix Y, it is called Partial Least Squares Discriminant Analysis (PLSDA) and widely
267 used for the identification of microorganisms.

268 The ***optimization of the model*** was done via cross-validation. A leave-one-out cross-
269 validation (CV) procedure was used, where one strain was taken out at a time and used
270 for validation. The four type strains were always included in the calibration model and
271 never used for validation. Therefore, the CV contained 260 segments. The optimal
272 number of principal components (AOpt) was determined as the one, which did not yield
273 significantly higher MCR than the model with the minimum MCR. The MCR was
274 calculated as a fraction of the misclassified samples by the total number of samples. The
275 statistical significance was evaluated by a binomial test.

276 To ***validate the established model***, a cross-model-validation (CMV) was done [59]. A
277 leave-one-out CMV was performed in the following manner. In each step of CMV one
278 strain was left aside and a leave-one-out CV model was established on the rest of the
279 samples as described above. Thus, the left-out strain was not included in the calibration
280 model and identification was performed on the basis of similarity to other strains
281 belonging to the same species. The sample, which is left aside, was used for validation
282 and the misclassification rate was stored. As for the CV, type strains were not taken out
283 and used for validation. Thus, the CMV consisted of 260 segments, where at each

284 segment a 259-fold CV was done. The final CMV error is the mean of all the errors. The
285 CMV error allows the control of stability and reliability of established classification
286 models. A stable model is expected to show a CMV error, which is comparable to the CV
287 error of the model.
288

289 Correlation analysis between NGS and FT-IR distance matrices

290 In order to correlate the two different data-sets, Mantel test analysis were carried out.
291 Mantel's test is an approach that overcomes some of the problems inherent in explaining
292 species-environment relationships. Mantel's test is a regression in which the variables are
293 themselves distance or dissimilarity matrices summarizing pairwise similarities among
294 sample locations. One advantage of Mantel's test is that, because it proceeds from a
295 distance (dissimilarity) matrix, it can be applied to variables of different logical type
296 (categorical, rank, or interval-scale data) [60]. In general, a Mantel test measures the
297 correlation between two matrices typically containing measure of distance.
298 Correlation were performed using R environment software (http://www.R-project.org/)
299 ade4 library, mantel.rtest and cor.test function for the estimation of the *p*-value with 9999
300 permutations using distance matrix calculated on the basis of the ITS and LSU markers
301 and for the FT-IR matrices before and after the PLS model.
302

303 Identification by distances to type strains and central strains

304 Each strain was originally attributed to one of the four species by means of rapid
305 clinical identification (CHROMagar) followed by MALDI-TOF, sequencing of the ITS
306 and LSU D1/D2 regions and FT-IR spectroscopy. For molecular data, distance matrices

307 were calculated through the base of percentage of identical bases/residues. For FT-IR
308 data, distance matrixes were obtained on the basis of PLS loadings. After the PLS
309 modeling as described above, a squared distance matrix was obtained with all the
310 distances among objects in the PLS hyperspace.

311 For both sequencing and FT-IR data, four distance matrices, referred to as species
312 matrices, were obtained. The species matrices contain the distances among members of
313 the strains of each of the four species. As described elsewhere [61], the central strain (CS)
314 of each species distribution was identified as that with the minimum sum of distances
315 from all other strains. Type strains and central strains were named jointly as "reference
316 strains". The distances between each studied strain and: ***i.*** the four type strains (DTS); ***ii.***
317 the four central strains (DCS) and ***iii.*** the eight reference strains (DRS) were calculated.

318 Two identification approaches were tested:

319 **a. Single match approach**. The correct species attribution requires that the DTS or DCS
320 be the lowest among the eight DRS values of each stain.

321 **b. Double match approach**. For each strain two identifications are carried, one with the
322 TS and one with the CS. In both cases, the correct species attribution requires that the
323 DTS or DCS is the lowest among respectively the four DTS and DCS values of each
324 stain.

325 This can be summarized by the following logical expression.

326 *If* $D(Si\text{-}TSj) < D(Si\text{-}TS) \Rightarrow \text{Match} = 1$

327 $D(Si\text{-}TSj) \geq D(Si\text{-}TS) \Rightarrow \text{Match} = 0$

328

329

Results

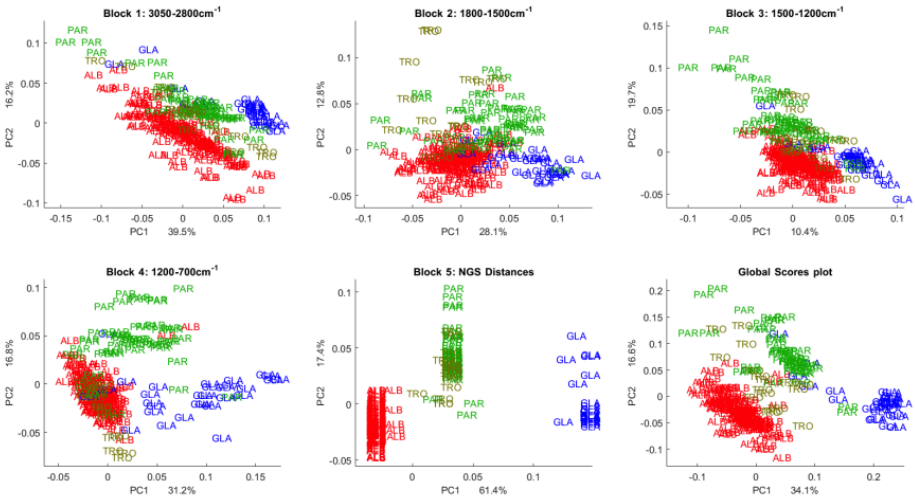
Connecting FT-IR and NGS data by consensus principal component analysis and grouping patterns in FT-IR data

In order to connect FT-IR data and DNA sequencing data in one data model, we applied consensus principal component analysis (CPCA) [55]. To this purpose, the FT-IR data was split into four blocks: The region from 3050-2800 cm^{-1} was defined as block one, the region from 1800-1500 cm^{-1} as block two, the region from 1500-1200 cm^{-1} as block three and the region from 1200-700 cm^{-1} as block four. The sequencing relationship matrix was defined as block five. In order to investigate global and block grouping patterns, block and global score plots are used [57].

The results showed that higher similarity between *C. albicans* and *C. tropicalis* was indicated in FT-IR data than NGS distance, whereas *C. tropicalis* seems to be more closely related to *C. parapsilosis*. These results are presented by the correlation loading plot (Fig 4, panel d), where it could be seen that wavenumbers from different spectral regions are responsible for the separation of the *C. glabrata* from other three species, whereas *C. albicans*, *C. parapsilosis* and *C. tropicalis* are better separated by the linear sequence of variables within the 1200-700 cm^{-1} polysaccharide region.

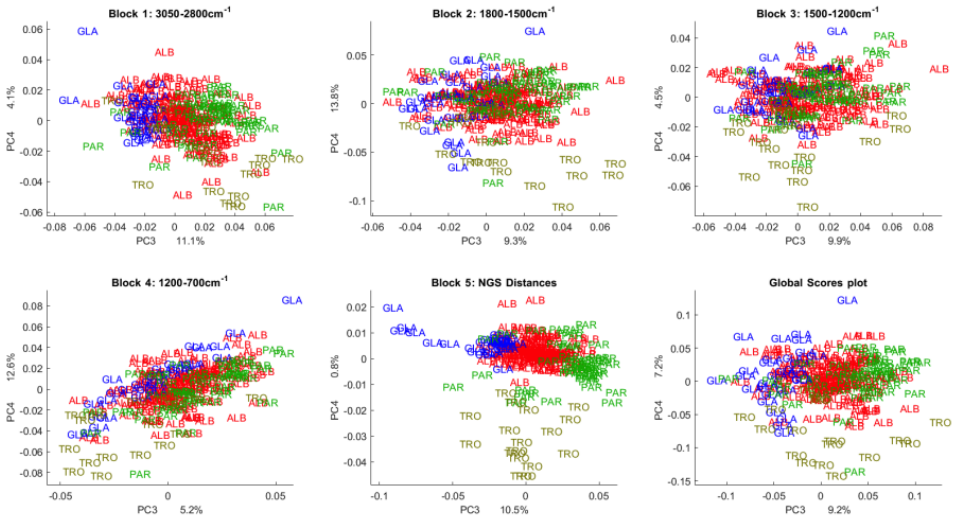
CPCA of FT-IR data indicated that PC1 and PC2 in the first three data blocks separated *C. albicans* – ALB, *C. glabrata* – GLA, *C. parapsilosis* – PAR whereas *C. tropicalis* – TRO required additional PC3 and PC4 within the fourth data block to be clearly distinguished (Fig 3). Moreover, the first four PCs describe only 77.4% total variance, which required six additional components (up to PC10) to reach 92.9% value.

Fig 2. Score plots of CPCA (PC1 and PC2) analysis of genetic-NGS and phenotypic-FT-IR spectroscopic data of strains from four *Candida* species - *C. albicans*, *C. parapsilosis*, *C. glabrata* and *C. tropicalis.*



Legend. The score plots of blocks 1-4 of CPCA analysis of FT-IR spectroscopy data, where block 1 is for lipid region (3050-2800 cm^{-1}), block 2 is for mixed lipid and protein region (1800-1500 cm^{-1}), block 3 is for mixed lipid, protein and polysaccharide region (1500-1200 $cm^{-}1$) and block 4 is for polysaccharide region (1200-700 cm^{-1}). The score plot of block 5 is for NGS data. The score plot of block 6 represents the global score plot of CPCA components one and two indicating the consensus of all blocks.

Fig 3. Score plots of CPCA (PC3 and PC4) analysis of genetic-NGS and phenotypic-FT-IR spectroscopic data of strains from four *Candida* species - *C. albicans*, *C. parapsilosis*, *C. glabrata* and *C. tropicalis.*



Legend. The score plots of block 1-4 are for FT-IR spectroscopy data, where block 1 refers to the lipid region (3050-2800 cm^{-1}), block 2 to the mixed lipid and protein region (1800-1500 cm^{-1}), block 3 to the mixed lipid, protein and polysaccharide region (1500-1200 cm^{-1}) and block 4 to the polysaccharide region (1200-700 cm^{-1}). The score plot of block 5 refers NGS data. The score plot of block 6 represents the global score plot of CPCA components three and four indicating the consensus of all blocks.

From the block score plots (Fig 2 and Fig 3) it is obvious that intra-species variation captured by the FT-IR data (block one to four) is much larger than the NGS intra-species variation (block five). As shown previously, the phenotyping variability identified by FT-IR can be explained by a real chemical variability between the strains, and not by an instrumental variability, which is negligible [37]. Further, it was shown that the chemical variability between strains is mainly due to inherent chemical differences between strains if cultivation conditions are controlled strictly [37].

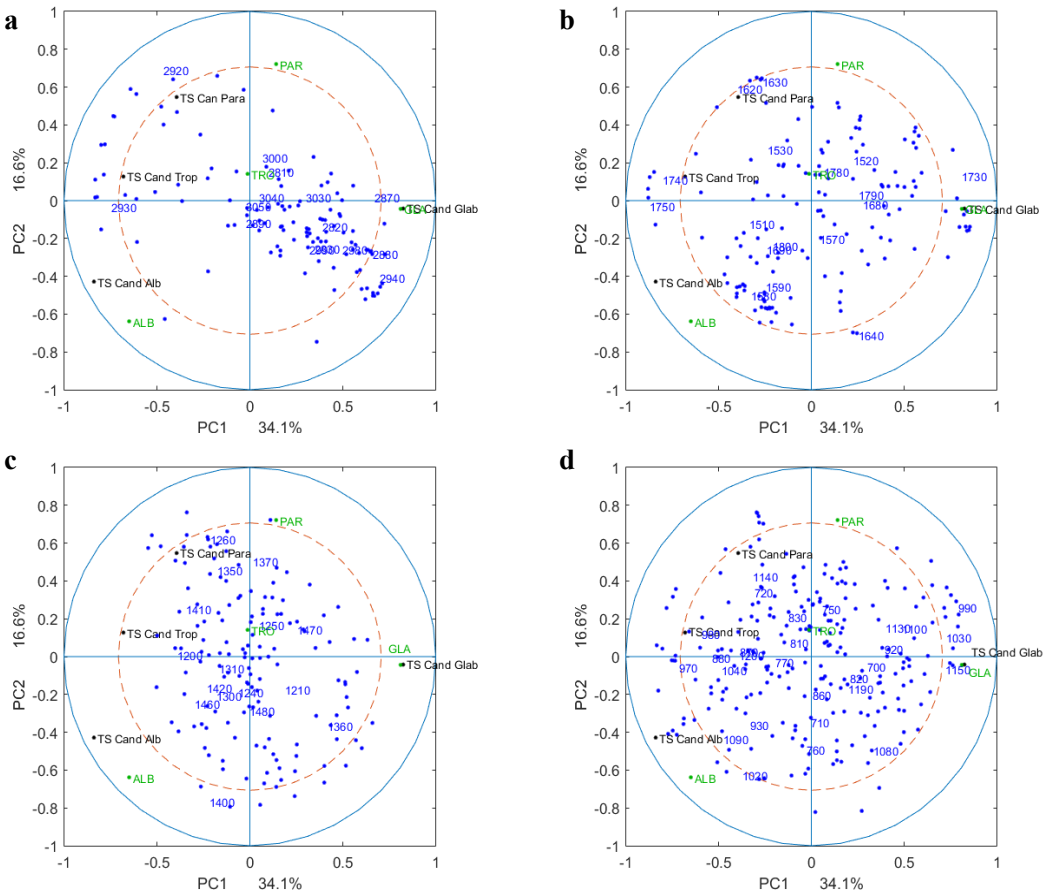
All block score plots for the first and second component (Fig 2) show a distinct separation of the three species - *C. albicans*, *C. parapsilosis*, *C. glabrata* - for both

387 genetic and FT-IR data. It is interesting to note, that in the global score plot the species *C.*
388 *tropicalis* is separated and located between *C. albicans* and *C. parapsilosis* (Fig 2, block
389 6) while, in all block score plots for FT-IR and NGS data, *C. tropicalis* is overlapping
390 with other species. A possible explanation is that whereas in block one, two, three and
391 five, *C. tropicalis* is mixed with *C. parapsilosis*, block four shows an overlap of *C.*
392 *parapsilosis* and *C. albicans.* Therefore, when combining the discriminant information
393 contained in all four FT-IR blocks, a separation of all four species is to a large extent
394 possible by only using the first two components of the FT-IR data. While *C. glabrata*, *C.*
395 *albicans* and *C. parapsilosis* can be discerned by the first two components of all blocks,
396 including lipid, protein and polysaccharide region, *C. tropicalis* is overlapping with other
397 species for all blocks including the NGS data.

398 For all blocks, except the polysaccharide region, *C. tropicalis* is overlapping with *C.*
399 *parapsilosis* for the first two components, while in the polysaccharide region *C. tropicalis*
400 is overlapping with *C. albicans*. This can suggest that *C. glabrata*, *C. albicans* and *C.*
401 *parapsilosis* are phenotypically and biochemically very different, while *C. tropicalis* and
402 *C. parapsilosis* appear phenotypically and biochemically very similar. The separation
403 between *C. tropicalis* and *C. parapsilosis* supported by the results of CPCA within the
404 polysaccharide region can suggest that major differences are in the cell wall, which
405 associates the majority of the cellular polysaccharides. It is interesting to note that *C.*
406 *albicans* and *C. tropicalis*, although separated in all other FT-IR blocks, are similar in
407 their polysaccharide profile, which is revealed in the block score plot of the first two
408 components of the region 1200-700 cm^{-1}. Further, the score plots of principal component

three and four show clear separation of the *C. tropicalis* species for both NGS and FT-IR data (Fig 3).

Fig 4. Correlation loading plot (PC1 and PC2) of NGS and FT-IR (block 1, 2, 3 and 4) with global scores.



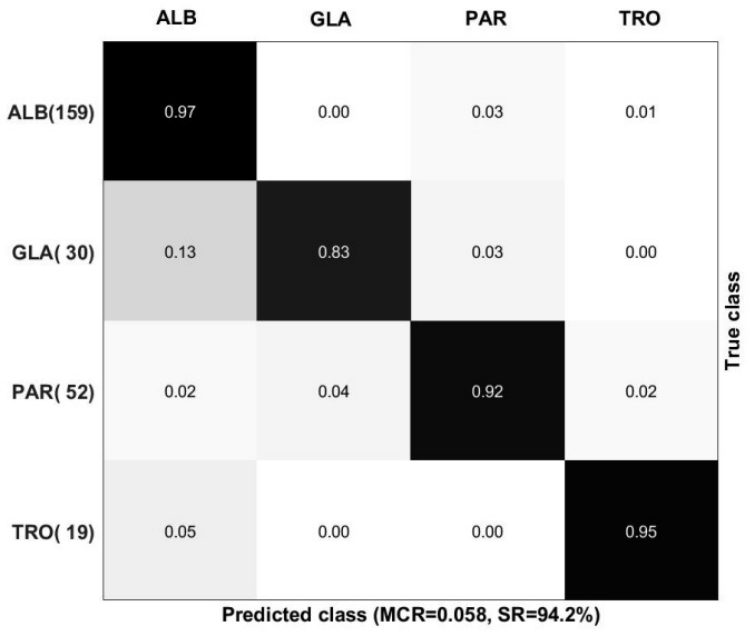
Legend. The correlation loading plots showing the correlation between the global scores of the CPCA analysis with the four different FT-IR blocks (**a-d**) and the distance matrix of the genetic data. In addition, the correlations between the global scores and the genetic distance matrix and the indicator variables for each species are visualized in each

421 plot. Panel **a**. Correlations between global scores and the lipid region (block 1, 3050-2800
422 cm^{-1}), the genetic distance matrix and the species indicator variables; panel **b**.
423 correlations between the global scores and the mixed lipid and protein region (block 2,
424 1800-1500 cm^{-1}), the genetic distance matrix and the species indicator variables; panel **c**.
425 correlations between the global scores and the mixed lipid, protein and polysaccharide
426 region (1500-1200 cm^{-1}), the genetic distance matrix and the species indicator variables
427 and panel **d**. correlations between the global scores and the polysaccharide region (1200-
428 700 cm^{-1}), the genetic distance matrix and the species indicator variables. Blue dots
429 represent FT-IR wavelengths; black dots distances to TS Cand alb, TS Cand para, TS
430 Cand glab and TS Cand trop represent type strains (TS) of the four species and green dots
431 namely ALB, PAR, GLA and TRO represent the group variables (indicator variables).
432
433 The global score plot of CPCA components one and two represents the consensus of
434 all blocks involved. We can see that the first two components of the global scores
435 representing the consensus of all blocks, separate all species of *Candida* namely *C.*
436 *albicans*, *C. parapsilosis*, *C. glabrata* and *C. tropicalis* very well.
437 Further, comparing the corresponding block score plots for all FT-IR regions and the
438 genetic-NGS, we observe that all block score plots show a similar tendency as the global
439 scores, but there are also clear differences in grouping patterns and explained variances,
440 i.e. the contributions of each block to the global pattern.
441 In the correlation loadings plot between the global scores of the FT-IR and the NGS
442 distance matrix, the genetic distance matrix are nicely correlated with the group indicator
443 variables (Fig 4, panel a-d). For instance, the first component showed significant

444 difference between *C. glabrata* and the other three species. Furthermore, by the second
445 component the difference between the other three species (*C. albicans*, *C. parapsilosis*
446 and *C. tropicalis*) is explained. In Fig 4 panel a the fatty acid region of FT-IR explains
447 mainly the difference between *C. glabrata* and the other species. Considering the first
448 component, the ester band (around 1750) explains very well the difference between *C.*
449 *glabrata* and the other species (Fig 4, panel b) while by the second component the protein
450 bands explain the difference among the other three species. The mixed region (Fig 4,
451 panel c) showed differences between *C. parapsilosis*, *C. tropicalis* and *C. albicans* while
452 the carbohydrates region (Fig 4, panel d) explain difference among all the four species
453 considered.
454

455 ## Classification based on discriminant PLSR

456 A classification model was built by the PLSR method and optimized by cross-
457 validation (CV). The established model contained six PLS components and a total
458 success rate (SR) value of 94.2% was achieved.
459 The corresponding confusion matrix is *C. glabrata* species with the SR equal to 83%.
460
461 **Fig 5.** Confusion matrix for the cross-validated classification model.
462



Legend. Errors are given as misclassification rate (MCR), which is the fraction of misclassified samples over the total number of samples. The success rate (SR) is given in percentage and equals SR = (1-MCR)*100. The number of samples in each group is specified in the left column with the true group affiliations. The predicted group is specified on the top of the matrix.

The CMV was done in order to test the model performance and the error stability. The CMV error repeats exactly the CV error and the CMV success rate equals to 94.2%. This is an important property of the model suggesting that the cross-validated model is reliable and could perform well when used for prediction of a new strain.

It is important to put in mind that strains used for validation where not present in the dataset of strains used to establish the model. Both for CV and CMV validation was done by taking a strain completely out. Taking a strain completely out is the most stringent test that can be performed for validating the model and corresponds to the actual situation,

478 where unknown strains need to be identified in hospitals or in source tracking in food
479 industry.
480

481 Correlation analysis between NGS and FT-IR distance 482 matrices

483 In order to correlate the two different data-sets, Mantel test analysis were performed
484 using distance matrix based on ITS and LSU markers and distances obtained with FT-IR
485 in different conditions (Table 2).
486

487 **Table 2.** Mantel test analysis between NGS and FT-IR.
488

Conditions	Mantel *r*	*p* value
FT-IR whole spectrum	0.5725	0.0001
FT-IR block 1	0.2143	0.0001
FT-IR block 2	0.4729	0.0001
FT-IR block 3	0.3289	0.0001
FT-IR block 4	0.5465	0.0001
FT-IR block 1+2	0.4445	0.0001
FT-IR block 1+3	0.3002	0.0001
FT-IR block 1+4	0.5388	0.0001
FT-IR block 2+3	0.4761	0.0001
FT-IR block 2+4	0.5911	0.0001
FT-IR block 3+4	0.5485	0.0001
FT-IR - PLS PC 1	0.6456	0.0001
FT-IR - PLS PC 1-2	0.7062	0.0001
FT-IR - PLS PC 1-3	0.7121	0.0001
FT-IR - PLS PC 1-4	0.7089	0.0001
FT-IR - PLS PC 1-5	0.7071	0.0001

FT-IR - PLS PC 1-6	0.7090	0.0001
FT-IR - PLS PC 1-7	0.7075	0.0001
FT-IR - PLS PC 1-8	0.7072	0.0001
FT-IR - PLS PC 1-9	0.7018	0.0001
FT-IR - PLS PC 1-10	0.6955	0.0001

Legend. Mantel test data report the correlation between the distance matrix among strains calculated on the basis of LSU-ITS and the distance matrices among strains calculated with the FT-IR data in the conditions indicated in column 1. The *p* value reports the error probability of the corresponding mantel test. All Mantel analyses were carried out with 9999 permutations. FT-IR - PLS PC1 indicates that only the first principal component was used. Similarly, FT-IR - PLS PC 1:n indicates that all principal components from 1 to n were used to calculate the distance matrix.

Taxonomic analyses are carried out on the basis of the distance matrix among strains calculated on the basis of the sequences from well established molecular markers such as LSU and ITS. The idea of using the FT-IR technique as a sort of phenotypically proxy of the molecular markers relays on the possibility of applying cluster and factor analysis to select the right wavelengths for this use.

In order to test the quality of the FT-IR techniques compared to ITS and LSU sequencing, a series of Mantel tests were carried out between the distance matrix for all the strains obtained with the two molecular markers and the distance matrices obtained with different treatments of the FT-IR spectra. This test calculates the correlation (*r*) between distance matrices of the same size and gives also a *p* value on the quality of the correlation.

The whole FT-IR spectrum, with only basic pre-treatments, yielded 0.57 Mantel *r* that was higher than the values obtained with single blocks of IR region, and slightly lower than the combination of blocks 2 and 4 (*r*= 0.59) (Table 2). The distance matrix obtained with the first 10 principal components from the PLS analysis gave a 0.64 Mantel *r*. This analysis was repeated using the first principal component and then the combinations of consequent components from 1 to 10 (e.g PC1 and 2, PC 1 thru 3 etc).

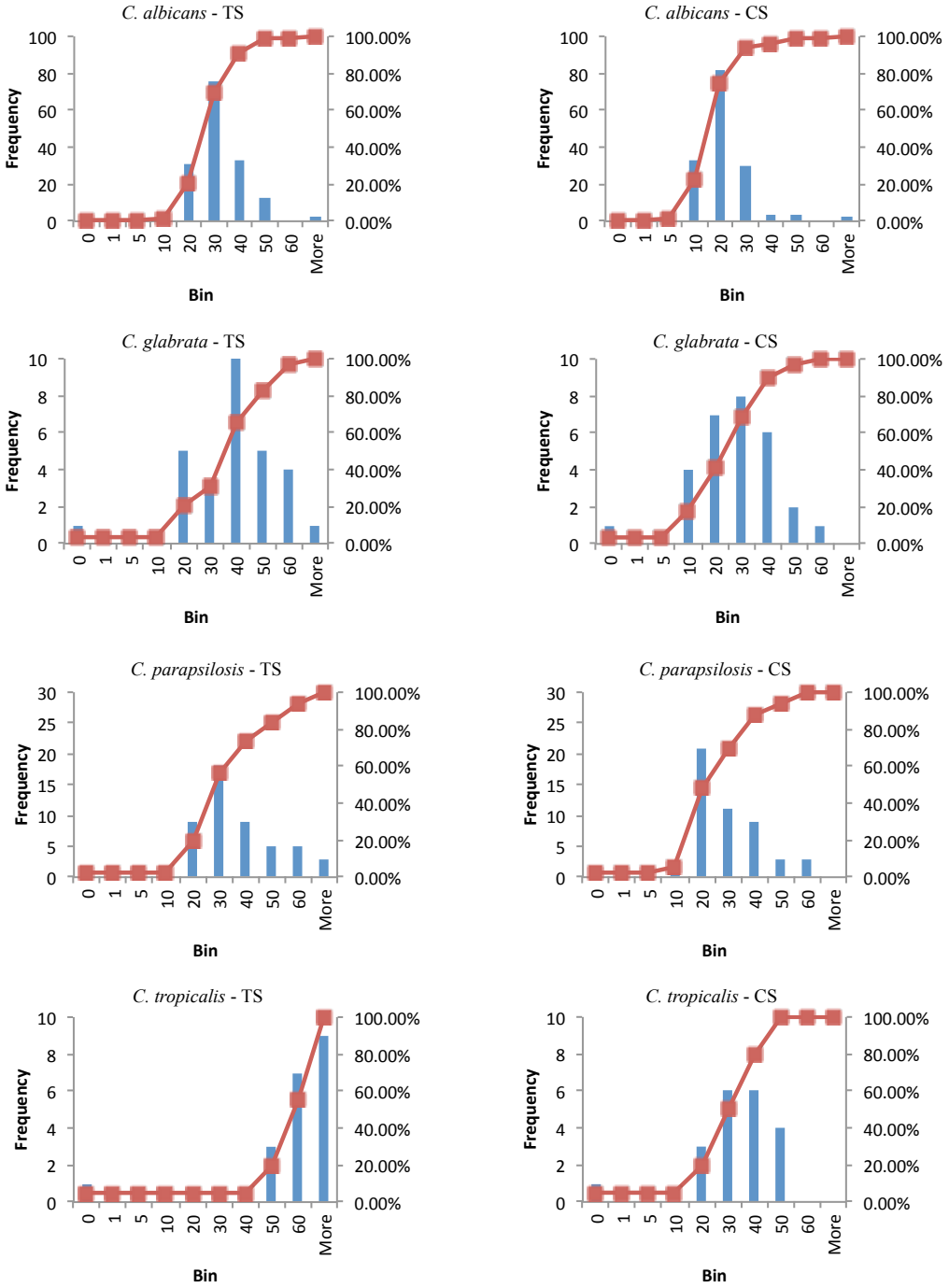
The results of these tests showed that the best correlation with the LSU and ITS sequencing was obtained by using the first three principal components (*r*= 0.71). The last seven components contained a part of the overall spectral variability, but this was not correlated to the variations among species as detected by molecular markers.

Distribution of the strains around the Central and Type strain: TS are not central

In order to determine distance matrixes for FT-IR spectra, distances between all strains were estimated on the basis of PLS scores. To this purpose a PLSDA model was established and the optimal number of components was estimated according to the procedure described above. The following distances were estimated: (1) the distances of all strains to the Type Strain (TS), which is used as a reference strain in NGS approach and (2) the distances of all strains to the Central Strain (CS), which was previously demonstrated to be the optimal reference strain, when a distance approach is used to species delimitation and identification [61] (Fig 6).

532 **Fig 6.** Distribution of the strains distances to TS (type strain) and CS (central strain).

533



534

535

536 **Legend.** Distribution of strains distances reference spectra of the four *Candida* species
537 respect to TS (**a**, **c**, **e** and **g**) and CS (**b**, **d**, **f** and **h**), respectively.
538
539 The distances of the strains of each species from the CS and the taxonomic TS showed
540 different distributions for each of the four species. In general, the vast majority of the
541 strains showed a short distances to the CS of the PLS distribution rather than the TS. For
542 *C. tropicalis* for example, 40% of the strains were distributed around the CS, while the
543 majority of them showed huge distances to the TS. The fact that TS is not central in this
544 case could be due to the small number of strains within the *C. tropicalis* subset (19
545 strains) as well as an overlapping of the species *C. tropicalis* with *C. parapsilosis* in the
546 NGS-based distance matrix (Fig 2, block 5). Also the *C. albicans* subset with more than
547 150 strains did not reveal a central positioning of the type strain in the strain distribution;
548 similar observations were made for *C. parapsilosis* and *C. glabrata*.
549

550 Taxonomic usage of PLS modeled FT-IR data

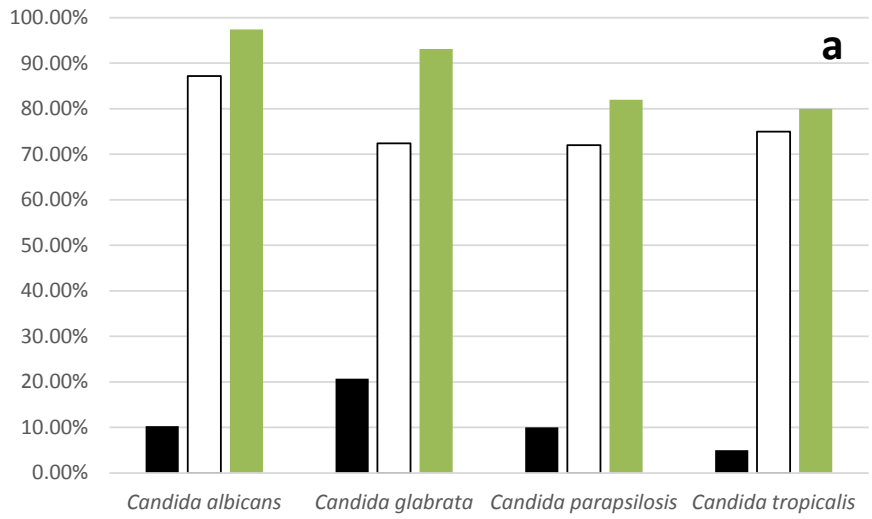
551 The rationale in assigning an unknown strain to a species by phenotypic approaches is
552 that the shortest the distance and the closest will be the microorganisms to that species.
553 Since every species is represented by several strains, one option is to define a single
554 reference strain and determine the distance of an unknown strain to this reference strain.
555 From a taxonomic point of view, the reference strain should be represented by the type
556 strain (TS). Notwithstanding, if the type strain is far from the centre of the strain
557 distribution in a given species, the use of the type strain as reference strain may result in
558 miss-identifications.

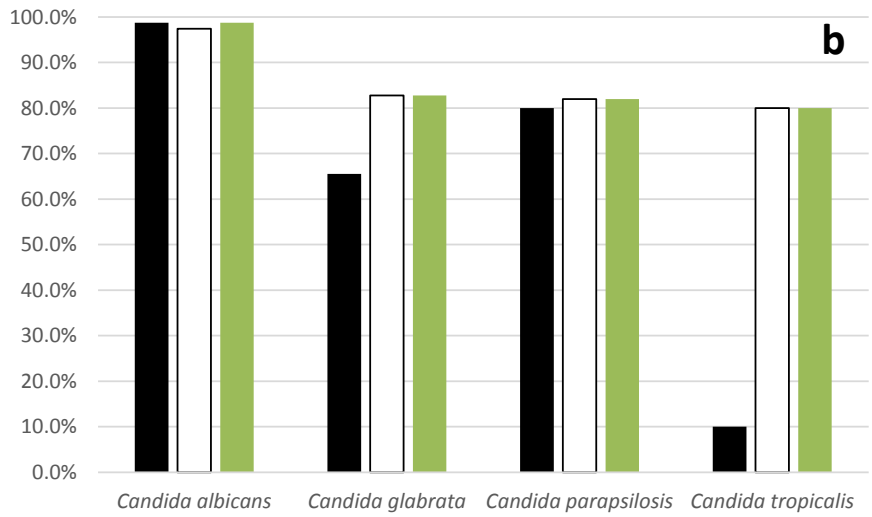
559 For these reasons, we compared the outputs of the identifications carried out with both
560 the type and the central strain. Two procedures based on a single possible match and on
561 two matches were testes, respectively. The former requires that the distance of a strain to
562 the TS or the CS of its species is the shortest among the distances to the TSs and CSs of
563 all species considered. The latter requires that the distance to the TS of the species is the
564 shortest among all distances to the various TSs. The CSs distances are calculated in a
565 similar way.

566 With the single match procedure, the number of correct identifications was higher
567 using CS than TS (Fig 7, panel a).

568

569 **Fig 7.** Comparison of single and double match approaches in classify *Candida* strains.





Legend. Panel **a.** single match analysis with CHROMagar; panel **b.** double match analysis with CHROMagar and MALDI-TOF. Black columns represent the percentage of matchings' to the Type Strain (TS); white columns report the matchings' to the central strain (CS). Green columns report the percentage of the sum of matchings' in the single match analysis (**a**) and the maximum obtainable percentage of correct matchings' in the double match analysis (**b**).

578

579 Summing the results of both the TS and the CS distances obtained with the single
580 match procedure, the total percentage of correct identification reached 97.4% for *C.*
581 *albicans*, 93.1% for *C. glabrata,* while *C. parapsilosis* and *C. tropicalis* achieved
582 respectively 82% and 80% correct identifications. With the double match procedure, the
583 successful identification to the TS and CS were very similar for *C. albicans* and *C.*
584 *parapsilosis* (Fig 7, panel b). The other two species showed lower identification rates
585 with the double matching procedure. Interestingly, in *C. albicans* 97.4% of the strains
586 were correctly identified with two matches. The proportion of strains with two matches
587 decreased to 74% in *C. parapsilosis*, 62.1% in *C. glabrata* and was only 15% in *C.*
588 *tropicalis.*

589 A possible interpretation of those results was that the identification success could be
590 due to the number of strains actually used in the PLS analysis. Therefore, we investigated
591 the correlation between the number of correct identifications and the number of strains in
592 a given species. In the single match case, the correlation between the number of strains
593 and the percentage of positive identifications was 0.7060. This poor correlation value,
594 sometimes resulting even in lower value when all strains were considered, can be
595 probably ascribed to the fact that *C. glabrata* showed more matching that expected.
596 Considering only three species *C. albicans*, *C. parapsilosis* and *C. tropicalis*, the
597 correlation between the number of correctly identified strains and the number of strains in
598 each species resulted 0.9943. For the double match algorithm, the correlation between the
599 number of strains and the percentage of correct identification was 0.9849 when all
600 species were included. These results demonstrate that the quality of the identification

601 depends strongly on the number of correctly identified reference strains used to create the
602 PLS model [34, 44].
603

604 # DISCUSSION AND CONCLUSION

605 Taxonomy of fungi is subject to the code of nomenclature (http://www.iapt-
606 taxon.org/nomen/main.php), requiring that "The application of names of taxonomic
607 groups is determined by means of nomenclatural types" (Principle II). The type is defined
608 in the article 7.2 as follows: "A nomenclatural type (*typus*) is that element to which the
609 name of a taxon is permanently attached, whether as the correct name or as a synonym.
610 The nomenclatural type is not necessarily the most typical or representative element of a
611 taxon". A living "type strain" represents the type in microbiology. The fact that the type
612 strain is not necessarily the most representative, poses serious problems, when a distance-
613 based approach is applied for the classification of unknown strain. In fact, it was
614 demonstrated that serious problems in identification could be due to reference strains far
615 from the centre of the strain distribution, extreme closeness of the species and width of
616 their distribution. The worst situation is present when two or more species are closer than
617 their mean variation.

618 Recent papers demonstrated that the type strain is central in many species when using
619 the ITS as taxonomic marker. The same situation was mostly present when the analysis
620 was focused on fungal species of medical interest. For the four species considered in this
621 paper, the type strain was not central. We have shown that the quality of the identification
622 depends very strongly on the number of strains employed with an acceptable minimum of
623 at least 50, as in the case of *C. parapsilosis*. Even *C. albicans,* with more than 150 strains,

624 did not show the centrality of type strain in the strain distribution. Taking into account
625 these considerations, one may hypothesize that increasing the number of strains leads to
626 an improvement of identification quality, but it is unlikely that the type strain converges
627 towards the centre of the PLS distribution. These evidences raise the question for the
628 rational of this phenomenon and pose a practical problem related to the correct
629 identification procedure using FT-IR technology.

630 The rational behind those differences can be due to the strong independence between
631 the FT-IR metabolomics and the ITS-LSU D1/D2 description of the strains. In fact, there
632 is no evidence that the metabolome and the sequence of these DNA markers should be
633 biologically linked. On the other hand, the evolutionary divergence between species can
634 have varied at similar pace for both DNA markers and metabolome, making the two
635 systems comparable, although not biologically dependent as, for instance, a protein
636 sequence with the DNA sequence of its encoding gene.

637 From a practical point of view, it seems that the right procedure to employ FT-IR as an
638 effective system in strain identification relays on three major points: i. a large database of
639 reference spectra of strains identified correctly with state of the art methods, ii. an
640 efficient statistical modeling, iii. the simultaneous use of the central and type strain with
641 *ad hoc* tailored algorithms as those described in this paper or more advanced algorithms
642 based on pattern recognition [61].

643 The data shown indicate that when high numbers of strains are considered, the lack of
644 centrality of the type strain plays a secondary role. Moreover, the application of the
645 double match algorithm allows for a more careful identification. In fact, strains scoring

646 "1" should be double checked with other analyses, whereas the identification of strains
647 scoring "2" matching can be considered highly satisfactory.
648 As compared to other high-throughput techniques such as MALDI-TOF and NGS, FT-
649 IR has several advantages; included easy and fast sample preparation as well as low costs
650 for consumables, which makes this spectroscopic technique very attractive and suitable
651 for medical diagnostics. However, the current limitation to its use seems to be the
652 absence of reliable and validated libraries linked to taxonomically sound identification
653 procedure. In principle, libraries should include several tens of strains for each relevant
654 species, possibly over 50, according to our data. At the same time, the panel of strains
655 needs to be composed of well-identified strains, possibly deriving from diverse sources
656 and collected over an extensive time period. This implies a multidisciplinary effort of
657 specialists working in strain isolation and maintenance, molecular taxonomy, FT-IR
658 technique and chemo-metrics, data management and data basing.
659
660 **REFERENCES**
661

662 1. De Queiroz K. Species concepts and species delimitation. Systematic
663 biology. 2007;56(6):879-86.
664 2. Adamowicz SJ, Scoles GJ. International Barcode of Life: Evolution of a
665 global research community. Genome. 2015;58(5):151-62. doi: 10.1139/gen-2015-
666 0094.
667 3. Hebert PD, Stoeckle MY, Zemlak TS, Francis CM. Identification of birds
668 through DNA barcodes. PLoS Biol. 2004;2(10):e312.
669 4. Hebert PD, Cywinska A, Ball SL. Biological identifications through DNA
670 barcodes. Proceedings of the Royal Society of London B: Biological Sciences.
671 2003;270(1512):313-21.
672 5. Meyer CP, Paulay G. DNA barcoding: error rates based on comprehensive
673 sampling. PLoS biol. 2005;3(12):e422.
674 6. Hebert PD, Ratnasingham S, de Waard JR. Barcoding animal life:
675 cytochrome c oxidase subunit 1 divergences among closely related species.

676 Proceedings of the Royal Society of London B: Biological Sciences.
677 2003;270(Suppl 1):S96-S9.
678 7. Seifert KA, Samson RA, Houbraken J, Lévesque CA, Moncalvo J-M,
679 Louis-Seize G, et al. Prospects for fungus identification using CO1 DNA
680 barcodes, with Penicillium as a test case. Proceedings of the National Academy
681 of Sciences. 2007;104(10):3901-6.
682 8. Kurtzman CP, Robnett CJ. Identification and phylogeny of ascomycetous
683 yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial
684 sequences. Antonie van Leeuwenhoek. 1998;73(4):331-71.
685 9. Baayen RP, O'Donnell K, Breeuwsma S, Geiser DM, Waalwijk C.
686 Molecular relationships of fungi within the Fusarium redolens-F. hostae clade.
687 Phytopathology. 2001;91(11):1037-44.
688 10. Geiser D, Klich M, Frisvad JC, Peterson S, Varga J, Samson RA. The
689 current status of species recognition and identification in Aspergillus. Studies in
690 Mycology. 2007;59:1-10.
691 11. James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, et al.
692 Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature.
693 2006;443(7113):818-22.
694 12. Schoch CL, Sung G-H, López-Giráldez F, Townsend JP, Miadlikowska J,
695 Hofstetter V, et al. The Ascomycota tree of life: a phylum-wide phylogeny clarifies
696 the origin and evolution of fundamental reproductive and ecological traits.
697 Systematic biology. 2009:syp020.
698 13. Stielow J, Lévesque C, Seifert K, Meyer W, Iriny L, Smits D, et al. One
699 fungus, which genes? Development and assessment of universal primers for
700 potential secondary fungal DNA barcodes. Persoonia: Molecular Phylogeny and
701 Evolution of Fungi. 2015;35:242.
702 14. Hofstetter V, Miadlikowska J, Kauff F, Lutzoni F. Phylogenetic comparison
703 of protein-coding versus ribosomal RNA-coding sequence data: a case study of
704 the Lecanoromycetes (Ascomycota). Molecular phylogenetics and evolution.
705 2007;44(1):412-26.
706 15. O'Donnell K, Rooney AP, Proctor RH, Brown DW, McCormick SP, Ward
707 TJ, et al. Phylogenetic analyses of RPB1 and RPB2 support a middle Cretaceous
708 origin for a clade comprising all agriculturally and medically important fusaria.
709 Fungal Genetics and Biology. 2013;52:20-31.
710 16. Druzhinina IS, Kopchinskiy AG, Komoń M, Bissett J, Szakacs G, Kubicek
711 CP. An oligonucleotide barcode for species identification in Trichoderma and
712 Hypocrea. Fungal Genetics and Biology. 2005;42(10):813-28.
713 17. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA,
714 et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal
715 DNA barcode marker for Fungi. Proceedings of the National Academy of
716 Sciences. 2012;109(16):6241-6.
717 18. Fell JW, Boekhout T, Fonseca A, Scorzetti G, Statzell-Tallman A.
718 Biodiversity and systematics of basidiomycetous yeasts as determined by large-
719 subunit rDNA D1/D2 domain sequence analysis. International Journal of
720 Systematic and Evolutionary Microbiology. 2000;50(3):1351-71.

19. Scorzetti G, Fell J, Fonseca A, Statzell-Tallman A. Systematics of basidiomycetous yeasts: a comparison of large subunit D1/D2 and internal transcribed spacer rDNA regions. FEMS yeast research. 2002;2(4):495-517.
20. Baldrian P, Větrovský T, Cajthaml T, Dobiášová P, Petránková M, Šnajdr J, et al. Estimation of fungal biomass in forest litter and soil. Fungal ecology. 2013;6(1):1-11.
21. Howlett BJ, Rolls BD, Cozijnsen AJ. Organisation of ribosomal DNA in the ascomycete Leptosphaeria maculans. Microbiological research. 1997;152(3):261-7.
22. Irinyi L, Serena C, Garcia-Hermoso D, Arabatzis M, Desnos-Ollivier M, Vu D, et al. 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. 2015:myv008.
23. Erukhimovitch V, Tsror L, Hazanovsky M, Talyshinsky M, Mukmanov I, Souprun Y, et al. Identification of fungal phyto-pathogens by Fourier-transform infrared (FTIR) microscopy. J Agric Technol. 2005;1(1):145-52.
24. Erukhimovitch V, Pavlov V, Talyshinsky M, Souprun Y, Huleihel M. FTIR microscopy as a method for identification of bacterial and fungal infections. Journal of pharmaceutical and biomedical analysis. 2005;37(5):1105-8.
25. Fischer G, Braun S, Thissen R, Dott W. FT-IR spectroscopy as a tool for rapid identification and intra-species characterization of airborne filamentous fungi. Journal of Microbiological Methods. 2006;64(1):63-77.
26. Shapaval V, Møretrø T, Suso HP, Åsli AW, Schmitt J, Lillehaug D, et al. A high−throughput microcultivation protocol for FTIR spectroscopic characterization and identification of fungi. Journal of biophotonics. 2010;3(8−9):512-21.
27. Büchl NR, Wenning M, Seiler H, Mietke−Hofmann H, Scherer S. Reliable identification of closely related Issatchenkia and Pichia species using artificial neural network analysis of Fourier−transform infrared spectra. Yeast. 2008;25(11):787-98.
28. Kümmerle M, Scherer S, Seiler H. Rapid and reliable identification of food-borne yeasts by Fourier-transform infrared spectroscopy. Applied and environmental microbiology. 1998;64(6):2207-14.
29. Adt I, Toubas D, Pinon J-M, Manfait M, Sockalingum GD. FTIR spectroscopy as a potential tool to analyse structural modifications during morphogenesis of Candida albicans. Archives of microbiology. 2006;185(4):277-85.
30. Toubas D, Essendoubi M, Adt I, Pinon J-M, Manfait M, Sockalingum GD. FTIR spectroscopy in medical mycology: applications to the differentiation and typing of Candida. Analytical and bioanalytical chemistry. 2007;387(5):1729-37.
31. Essendoubi M, Toubas D, Lepouse C, Leon A, Bourgeade F, Pinon J-M, et al. Epidemiological investigation and typing of Candida glabrata clinical isolates by FTIR spectroscopy. Journal of microbiological methods. 2007;71(3):325-31.
32. Essendoubi M, Toubas D, Bouzaggou M, Pinon J-M, Manfait M, Sockalingum GD. Rapid identification of Candida species by FT-IR

767 microspectroscopy. Biochimica et Biophysica Acta (BBA)-General Subjects.
768 2005;1724(3):239-47.
769 33. Sandt C, Sockalingum G, Aubert D, Lepan H, Lepouse C, Jaussaud M, et
770 al. Use of Fourier-transform infrared spectroscopy for typing of Candida albicans
771 strains isolated in intensive care units. Journal of clinical microbiology.
772 2003;41(3):954-9.
773 34. Naumann D, Helm D, Labischinski H. Microbiological characterizations by
774 FT-IR spectroscopy. Nature. 1991;351(6321):81.
775 35. Timmins ÉM, Howell SA, Alsberg BK, Noble WC, Goodacre R. Rapid
776 differentiation of closely relatedCandida species and strains by pyrolysis-mass
777 spectrometry and fourier transform-infrared spectroscopy. Journal of Clinical
778 Microbiology. 1998;36(2):367-74.
779 36. Orsini F, Ami D, Villa A, Sala G, Bellotti M, Doglia S. FT-IR
780 microspectroscopy for microbiological studies. Journal of microbiological
781 methods. 2000;42(1):17-27.
782 37. Kohler A, Böcker U, Shapaval V, Forsmark A, Andersson M, Warringer J,
783 et al. High-throughput biochemical fingerprinting of Saccharomyces cerevisiae by
784 Fourier transform infrared spectroscopy. Plos One.
785 2015;10(2):e0118052.doi:10.1371/journal.pone.
786 38. Shapaval V, Walczak B, Gognies S, Møretrø T, Suso H-P, Åsli AW, et al.
787 FTIR spectroscopic characterization of differently cultivated food related yeasts.
788 The Analyst. 2013;138(14):4129-38. doi: http://dx.doi.org10.1039/c3an00304c.
789 39. Shapaval V, Schmitt J, Møretrø T, Suso H, Skaar I, Åsli AW, et al.
790 Characterization of food spoilage fungi by FTIR spectroscopy. Journal of Applied
791 Microbiology. 2013;114(3):788-96. doi: http://dx.doi.org10.1111/jam.12092.
792 40. Shapaval V, Moretro T, Suso HP, Asli AW, Schmitt J, Lillehaug D, et al. A
793 high-throughput microcultivation protocol for FTIR spectroscopic characterization
794 and identification of fungi. Journal of Biophotonics. 2010;3(8-9):512-21. PubMed
795 PMID: ISI:000281055600004.
796 41. Oust A, Moretro T, Kirschner C, Narvhus JA, Kohler A. Evaluation of the
797 robustness of FT-IR spectra of lactobacilli towards changes in the bacterial
798 growth conditions. Fems Microbiol Lett. 2004;239(1):111-6. doi: DOI
799 10.1016/j.femsle.2004.08.024. PubMed PMID: WOS:000224317000015.
800 42. Zhao H, Kassama Y, Young M, Kell DB, Goodacre R. Differentiation of
801 Micromonospora isolates from a coastal sediment in Wales on the basis of
802 Fourier transform infrared spectroscopy, 16S rRNA sequence analysis, and the
803 amplified fragment length polymorphism technique. Applied and environmental
804 microbiology. 2004;70(11):6619-27.
805 43. Warringer J, Blomberg A. Automated screening in environmental arrays
806 allows analysis of quantitative phenotypic profiles in Saccharomyces cerevisiae.
807 Yeast. 2003;20(1):53-67. PubMed PMID: WOS:000180219900006.
808 44. Wenning M, Scherer S. Identification of microorganisms by FTIR
809 spectroscopy: perspectives and limitations of the method. Applied microbiology
810 and biotechnology. 2013;97(16):7111-20.

811 45. Doern GV, Vautour R, Gaudet M, Levy B. Clinical impact of rapid in vitro
812 susceptibility testing and bacterial identification. Journal of clinical microbiology.
813 1994;32(7):1757-62.
814 46. Maquelin K, Kirschner C, Choo-Smith L-P, Ngo-Thi N, Van Vreeswijk T,
815 Stämmler M, et al. Prospective study of the performance of vibrational
816 spectroscopies for rapid identification of bacterial and fungal pathogens
817 recovered from blood cultures. Journal of Clinical Microbiology. 2003;41(1):324-
818 9.
819 47. Deak T, Beuchat L. Yeasts associated with fruit juice concentrates.
820 Journal of Food Protection®. 1993;56(9):777-82.
821 48. Corte L, Roscini L, Colabella C, Tascini C, Leonildi A, Sozio E, et al.
822 Exploring ecological modelling to investigate factors governing the colonization
823 success in nosocomial environment of Candida albicans and other pathogenic
824 yeasts. Nature Publishing Group Scientific Reports. 2016;6:26860.
825 49. Cardinali G, Bolano A, Martini A. A DNA extraction and purification method
826 for several yeast genera. Annals of microbiology. 2001;51(1):121-30.
827 50. Schoch CL, Robbertse B, Robert V, Vu D, Cardinali G, Irinyi L, et al.
828 Finding needles in haystacks: linking scientific names, reference specimens and
829 molecular data for Fungi. Database. 2014;2014:bau061.
830 51. Vu D, Groenewald M, Szöke S, Cardinali G, Eberhardt U, Stielow B, et al.
831 DNA barcoding analysis of more than 9000 yeast isolates contributes to
832 quantitative thresholds for yeast species and genera delimitation. Studies in
833 Mycology. doi: http://dx.doi.org/10.1016/j.simyco.2016.11.007.
834 52. Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified
835 least squares procedures. Anal Chem. 1964;36:1627ff.
836 53. Martens H, Stark E. Extended multiplicative signal correction and spectral
837 interference subtraction: new preprocessing methods for near infrared
838 spectroscopy. J Pharm Biomed Anal. 1991;9(8):625-35. PubMed PMID:
839 1790182.
840 54. Zimmermann B, Kohler A. Optimizing Savitzky-Golay Parameters for
841 Improving Spectral Resolution and Quantification in Infrared Spectroscopy. Appl
842 Spectrosc. 2013;67(8):892-902. doi: Doi 10.1366/12-06723. PubMed PMID:
843 WOS:000322559700010.
844 55. Kohler A, Hanafi M, Bertrand D, Oust Janbu A, Møretrø T, Naderstad K, et
845 al. Interpreting several types of measurements in bioscience. Modern concepts in
846 biomedical vibrational spectroscopy. 2007:333-56.
847 56. Kohler A, Hanafi M, Bertrand D, Quannari M, Oust Janbu A. Interpreting
848 several types of measurements in bioscience. Biomedical Vibrational
849 Spectroscopy Hoboken, New Jersey, USA: John Wiley & Sons. 2008:333-56.
850 57. Hassani S, Martens H, Qannari M, Hanafi M, Borge GI, Kohler A. Analysis
851 of –omics data: Graphical interpretation- and validation tools in multi–block
852 methods. Chemometrics and Intelligent Laboratory Systems. 2010;104:140-53.
853 58. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of
854 chemometrics. Chemometr Intell Lab. 2001;58(2):109-30.

59. Anderssen E, Dyrstad K, Westad F, Martens H. Reducing over-optimism in variable selection by cross-model validation. Chemometr Intell Lab. 2006;84(1):69-74.
60. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer research. 1967;27(2 Part 1):209-20.
61. Antonielli L, Robert L, Corte L, Roscini L, Ceppitelli R, Cardinali G. Centrality of objects in a multidimensional space and its effects on distancebased biological classifications. Open Appl Inform J. 2011;5:11-9.