Comparative Genome-Wide Characterization of Microsatellites in *Candida albicans* and *Candida dubliniensis* Leading to the Development of Species-Specific Marker

Pallavi Singh^{a,b} Ravindra Nath^b Vimala Venkatesh^c

^aDepartment of Biotechnology, Dr. A.P.J. Abdul Kalam Technical University, Lucknow, India; ^bDepartment of Computer Science & Engineering, UIET, CSJM University, Kanpur, India; ^cDepartment of Microbiology, King George's Medical University, Lucknow, India

Keywords

Candida albicans · *Candida dubliniensis* · Genome-wide mining · Microsatellites · Polymerase chain reaction

Abstract

Background: Microsatellites or simple sequence repeats (SSR) are related to genomic structure, function, and certain diseases of taxonomically different organisms. **Objective:** To characterize microsatellites in two closely related *Candida* species by searching and comparing 1–6 bp nucleotide motifs and utilizing them to develop species-specific markers. **Methods:** Whole-genome sequence was downloaded from the public domain, microsatellites were mined and analyzed, and primers were synthesized. **Results:** A total of 15,821 and 7,868 microsatellites, with mono-nucleotides (8,679) and tri-nucleotides (3,156) as most frequent microsatellites, were mined in *Candida dubliniensis* and *Candida albicans*, respectively. Chromosome size was found positively correlated with microsatellite number in both the species, whereas it was negatively correlated with the relative abundance and density of microsatellites. A number of unique motifs were also found in both the species. Overall, microsatellite frequencies of each chromosome in *C. dubliniensis* were higher

than in *C. albicans*. **Conclusion:** The features of microsatellite distribution in the two species' genomes revealed that it is probably not conserved in the genus *Candida*. Data generated in this article could be used for comparative genome mapping and understanding the distribution of microsatellites and genome structure between these closely related and phenotypically misidentified species and may provide a foundation for the development of a new set of species-specific microsatellite markers. Here, we also report a novel microsatellite-based marker for *C. dubliniensis*-specific identification.

© 2021 S. Karger AG, Basel

Introduction

The species of the genus *Candida* are well known for their ability to cause fungal colonization and infection in humans, which is commonly known as candidiasis [1]. Among them, *Candida albicans*, *Candida glabrata*, *Candida parapsilosis*, and *Candida tropicalis* are reported to be frequently occurring pathogenic species, which are responsible for superficial and systemic infections. Their distribution depends on geographic locations, patients'

cohort, and exposure to antifungal drugs [2]. In 1995, a new species named *Candida dubliniensis* was identified among HIV-infected patients of Dublin, Ireland [3]. It showed similar phenotypic characteristics, commensalism, and opportunistic infection like *C. albicans* [4]. Moran et al. [5] reported that these two species had about 96% similar gene sequences, and more than 60% are homologous. It leads to difficulties in prompt differentiation among the two species in routine laboratory tests [6] and results that may underestimate their prevalence [7]. *C. dubliniensis* is reported to develop stable resistance to fluconazole (an antifungal drug) upon exposure in vitro [8]. Thus, rapid and accurate discrimination among the two species for all the clinical treatments and epidemiological studies is crucial [9].

Phenotypic tests helped only in presumptive identification and also gave variable results, leading to difficulties in the infection management and timely treatment of these two species [10]. Earlier in different confirmation studies, *C. dubliniensis* was incorrectly identified as *C. albicans* [11]. This suggests that conventional methods were non-reliable, mainly due to phenotypic similarities between them [4, 12]. Thus, various molecular methods were developed to differentiate among these two *Candida* species [13] to overcome such problems. The PCR-based methods were mostly based on the amplification of a conserved region of genome, that is, RNA or ITS regions, and were genus-specific.

Microsatellites or simple sequence repeats (SSRs) are small stretches of 1–6 bp nucleotide motifs distributed extensively throughout the genomes of prokaryotes [14] as well as eukaryotes [15]. High frequency, codominance, and highly polymorphic nature are the significant characteristics of the microsatellites. The development of PCR technology has led to the use of microsatellites as a genetic marker system for population genetics studies [16] and paternity testing [17]. These are found distributed throughout the genome both in coding and in noncoding region [14]. The characteristics such as multiallelic nature, codominant transmission, ease of detection by PCR, relative abundance (RA), and extensive genome coverage add to its uniqueness and more value than other marker systems available [18]. The whole-genome sequence of *Candida* species is freely available, and no detailed formal analysis of microsatellites within these species is reported until now. Hence, in this paper, we report the comparative microsatellite mining among the two closely related species, namely *C. albicans* (*Ca*) and *C. dubliniensis* (*Cd*), and development of a species-specific microsatellite-based diagnostic marker for *C. dubliniensis*.

Methodology

Genome Sequences

The genome sequences of *Ca* strain SC5314 (GCA_000182965.3) [19] and *Cd* strain CD36 (GCA_000026945.1) [20] were obtained from NCBI (<http://www.ncbi.nlm.nih.gov/>) in FASTA format. The species were assembled on the level of 8 chromosomes, and every chromosome sequence was analyzed individually.

Identification of Microsatellites

The microsatellites were identified using the online search software WebSat [21], which searches both perfect and compound microsatellites. The nucleotide repeats longer than 12 bp were considered microsatellites, which means that there should be at least 6 occurrences of a di-nucleotide repeat, 4 occurrences of a tri-nucleotide repeat, and 3 occurrences of tetra-, penta-, and hexa-nucleotide repeats. Other parameters used for the analysis of microsatellites were relative density (RD) and RA. RA is the number of microsatellites per Mb of the sequence analyzed, and relative density (RD) is the length (in bp) of microsatellites per Mb of the sequence analyzed. Shared and unique motifs were analyzed using the method of Mahfooz et al. [22]. Briefly, if one motif was present in both species, it was deemed common, and on the other hand, the motif that did not have any match was considered unique. These unique microsatellite motifs for each species were then chosen for designing primers using the program PRIMER 3 online software (frodo.wi.mit.edu/), considering all the crucial factors (product size, T_m, GC content, dimerization, etc.) needed for a good primer designing.

Statistical Analysis

The Pearson correlation coefficient was calculated using the SPSS package (SPSSV16.0; SPSS Inc., Chicago, IL, USA) for correlation analysis between two variables, including RA, RD, GC content, and chromosome sequence size.

Species-Specific PCR

For evaluating primer specificity of *Cd*-specific primers, DNA of various pathogenic fungi and major pathogenic *Candida* species were used as the template. For this purpose, template DNA of *C. albicans* ATCC 90028, *C. dubliniensis* (type strain CD36) and *C. dubliniensis* (CBS7987), *C. glabrata* ATCC-MYA 2950, *C. parapsilosis* ATCC 22019, *C. tropicalis* ATCC 750, *Candida guilliermondii* ATCC 6260, and *Candida krusei* ATCC 14243 were extracted using the Burst n' Grab protocol [23]. DNA from *Trichosporon* spp., *Cryptococcus*, *Fusarium* spp., and *Aspergillus* spp. was isolated using the QIAamp® DNA Mini Kit (Qiagen, India).

Polymerase Chain Reaction

PCR was performed in 10 µL reaction volumes containing 2 µL of 10× PCR buffer (Merck), 200 µM each of dNTP (Merck), 0.2 U of Taq DNA polymerase (Merck), 10 pM each of forward and reverse primers (*Cd*-specific primers, Table 1), and 10 ng of genomic DNA template. Thermocycling conditions were 94°C for 5 min, 35 cycles of 94°C for 1 min, 55°C for 1 min, and 72°C for 1 min, followed by a final extension step of 72°C for 7 min.

Electrophoresis and Imaging

All PCR products were electrophoresed in agarose gel (3%) for 1 h at 100 V at room temperature in TBE buffer (89 mM Trizma base, 89 mM boric acid, and 2.5 mM EDTA) stained with ethidium bromide, and visualized with UV light (360 nm).

Table 1. Details of *Candida dubliniensis*-specific primer

Primer name	Species-specific	Forward primer sequence (5'–3')	Reverse primer sequence (5'–3')	Motif	Motif length	Microsatellite
PSCD	<i>C. dubliniensis</i>	GTGGATGGATGGAAGAGT	GGATGGATGGATGGATGG	AGGTCT	6	AGGTCTAGGTCTAGGTCT

Table 2. Overview of microsatellites in two *Candida* species

Species	<i>Candida albicans</i>		<i>Candida dubliniensis</i>	
	whole genome	transcripts	whole genome	transcripts
Size, Mb	14.28	9.09	14.04	9.03
GC content, %	33.48	35.0	33.1	34.7
Microsatellites, <i>n</i>	7,868	4,229	15,821	4,587
Perfect microsatellites	7,323	3,661	15,033	3,890
Compound + imperfect microsatellites	545	568	788	697
Length of microsatellites, bp	136,583	78,412	297,009	92,603
% length	0.96	0.86	2.03	1.02
Relative density	9,355	8,626	21,154.4	10,255.0
Relative abundance	519.7	465.1	1,126.6	507.9

Results

The Number, RA, RD, and GC Content of Microsatellites in Two Genomes

A higher frequency of the microsatellites was observed in *Cd* (15,821) than in *Ca* (7,868). The variation among the frequency of microsatellites could be due to the difference in the genome size; hence, we have calculated the RA and RD of microsatellites. However, this, too, did not make any difference (Table 2). In transcript sequences, the maximum frequency of microsatellites was observed in *Cd* (4,587) compared with *Ca* (4,229). Similarly, the RD pattern of microsatellites in transcript sequences was higher in *Cd*, as in whole-genome sequence (Table 2).

The number, length, RA, and RD of individual microsatellite class are shown in Table 3. The most frequent microsatellites were tri-nucleotides (40.1%) in *Ca*, while in the *Cd* genome, mono-nucleotides accounted for the highest (52.1%) frequency. The second most frequent microsatellites included mono- and tri-nucleotides, respectively, for *Ca* (34.08%) and *Cd* (27.2%). Tetra- and hexa-nucleotides followed the order, and least were found for penta-nucleotides (only around 3% of the whole genome) (Fig. 1). We further calculated the GC content of all classes of microsatellites. Overall, the high-

er GC content was observed in *Ca* (23.1%) than in *Cd* (16.8%). Among the different microsatellites, the GC content of hexa-nucleotide motifs was higher (36.4% for *Ca* and 32.57% for *Cd*). The second highest was observed in tri-nucleotide motifs, followed by tetra- and penta-motifs, which were highly similar (ranging from 24.23 to 26.72%). The GC content of mono- and di-nucleotide repeats was least.

Conservation, Diversity, and Repeat Times of Different Microsatellite Categories

The prominent repeat motifs shared within two species were (A)_{*n*} (T)_{*n*} (AT)_{*n*} (AAC)_{*n*} (GTT)_{*n*} (AAG)_{*n*} (AAT)_{*n*} (CTT)_{*n*} (ATG)_{*n*} (ATC)_{*n*} (GGT)_{*n*} (AAT)_{*n*} (ACC)_{*n*} (AAAG)_{*n*} (CTTT)_{*n*} and (AAACA)_{*n*} (Table 4). A motif comparison study was performed to identify the repeat motifs unique for each species (Fig. 2). The highest number of shared motifs belongs to penta-nucleotides (34), whereas the highest number of unique motifs belongs to hexa-nucleotides (158–192) (Fig. 2; see online suppl. Table 1; for all online suppl. material, see www.karger.com/doi/10.1159/000512087).

The maximum repeat times of different microsatellite classes were very close among the two genomes. The predominant repeat times were 12–50 for mono-nucleotides,

Table 3. Number, length, abundance, and density of 6 types of microsatellites

Repeat type	Species	Total number	Total length, bp	Abundance	Density, bp/Mb	GC	
						length, bp	content, %
Mono-nucleotide	<i>Ca</i>	2,682	36,576	187.78	2,560.8	3,191	2.36
	<i>Cd</i>	8,679	134,766	593.7	9,218.9	2,123	5.80
Di-nucleotide	<i>Ca</i>	673	11,448	47.12	801.5	3,543	15.85
	<i>Cd</i>	1,271	22,732	86.95	1,555.0	1,976	17.26
Tri-nucleotide	<i>Ca</i>	3,156	58,395	220.97	4,088.5	27,685	30.90
	<i>Cd</i>	4,528	89,586	309.75	6,128.2	19,467	33.33
Tetra-nucleotide	<i>Ca</i>	764	14,688	53.49	1,028.3	5,927	26.72
	<i>Cd</i>	1,110	22,176	75.93	1,516.9	3,831	26.08
Penta-nucleotide	<i>Ca</i>	239	5,720	16.73	400.4	2,822	24.33
	<i>Cd</i>	483	11,645	33.04	796.5	1,439	25.15
Hexa-nucleotide	<i>Ca</i>	354	9,756	24.79	683.0	5,246	32.57
	<i>Cd</i>	578	16,104	39.54	1,101.6	3,552	36.40
Total	<i>Ca</i>	7,868	136,583	550.88	9,562.8	48,414	16.30
	<i>Cd</i>	15,821	297,009	1,138.91	20,317.4	32,388	23.71

Ca, *Candida albicans*; *Cd*, *Candida dubliniensis*.

7–21 for di-nucleotides, 5–53 for tri-nucleotides, 4–27 for tetra-nucleotides, 4–17 for penta-nucleotides, and 4–21 for hexa-nucleotides for both species (Fig. 3; online suppl. Table 2). *Cd*'s maximum repeat times were noted to be higher (4–5 repeats) than those of *Ca* in almost every type of microsatellite except the penta-nucleotides (equal).

The Distribution of Microsatellites within Chromosomes of Two Species

We analyzed the distribution of microsatellites among the chromosomes of both the *Candida* species. The number of microsatellites in *Cd* was higher than in *Ca* in all chromosomes (Fig. 4). The maximum microsatellite count was recorded in chromosome 1 (1,604 and 3,742). The highest microsatellite frequency order in *Ca* was chromosome 1 > 2 > R > 3 > 4 > 7 > 5 > 6, whereas in *Cd*, it was chromosome 1 > 2 > R > 3 > 4 > 5 > 6 > 7. The least frequent microsatellite was found in chromosome 6 (581 loci, 7.38%) and 7 (1,089 loci, 6.5%), respectively, in *Ca* and *Cd*. The RA of various microsatellite types had a similar distributional pattern among the two genomes (Fig. 5). *Ca* and *Cd* had tri-nucleotide (226.42 loci/Mb) and mono-nucleotide (532.72 loci/Mb), respectively, as their highest frequent microsatellite types. The frequency order was followed as mono- > tetra- > di- > hexa-nucleotide in all the chromosomes of *Ca* (except chromosome

3, which has di-nucleotide repeats greater than tetra-nucleotide). Similarly, the order was tri- > di- > tetra- > hexa-nucleotide (except chromosome 1, which has a higher abundance of di-nucleotide repeats) in *Cd*. Penta-nucleotide microsatellite frequency was lowest in each chromosome of both the species (except for chromosome 1 of *Cd*).

Chromosome size was significantly positively correlated with the number of microsatellites (Pearson, $r = 0.990$, $p < 0.01$), both *Ca* and *Cd*. However, surprisingly, the size was negatively correlated with RA and RD of both the species (Fig. 6). For instance, chromosome 1 was highest in *Ca* and *Cd*, but the maximum RA and RD were observed in chromosome 7, which is having the smallest chromosome size.

The GC Content of All Microsatellites in Each Chromosome

We calculated the proportion of each microsatellite category's GC content to the total size of all microsatellite for each chromosome (Fig. 7). Results showed that the overall GC content of microsatellites for both the *Candida* species is almost the same (23.9 and 24.4%). Hexa-nucleotide microsatellites showed the highest GC content in most chromosomes of two species (ranging from 31 to 43%). However, tri-nucleotide microsatellites had the

Fig. 1. Percentage of six categories of microsatellites in whole genome of two *Candida* species.

Fig. 2. Venn diagram representing the number of shared and unique motifs in each class of repeats.

highest GC content in chromosome 3 of *Ca* and chromosomes 1, 3, and 5 of *Cd*. The same microsatellites also had equivalent GC content for each species except in chromosome 1 of the two species. Di-nucleotide (10–20%) and mono-nucleotide (2–10%) microsatellites had the lowest GC content.

Species-Specific PCR

The identified unique motifs in both *Candida* species were further explored for the possibility of designing species-specific markers. Taking advantage of other sequenced

genomes of *Ca* (GCA_003454735.1, ASHC00000000.1, GCA_005890765.1, GCA_000149445.2, etc.) and *Cd* (GCA_000647555.1 and GCA_000647575.1), we confirmed the presence of these unique motifs in those whole-genome sequences as well. Primers were designed (taking into consideration their T_m , GC%, self-annealing, etc.) to amplify these unique motifs. Additionally, we performed a BLAST search of the primer sequences to confirm their uniqueness and selected 5 primers from each species for primer synthesis to check their potential as species-specific markers. One of the *Cd* markers yielded a single amplified

Table 4. Number, length, frequency, and density of the predominant motifs for each microsatellite type in two species

Motif	<i>Candida</i> species	Total numbers	Total length, bp	Average length, bp	Frequency, loci/Mb	Density, bp/Mb
A	<i>Ca</i>	1,286	17,427	13.55	90.04	1,220.15
	<i>Cd</i>	4,254	66,019	15.52	291	4,516.15
T	<i>Ca</i>	1,248	17,026	13.64	87.38	1,192.07
	<i>Cd</i>	4,197	65,556	15.62	287.1	4,484.48
AT	<i>Ca</i>	427	7,442	17.43	29.9	521.05
	<i>Cd</i>	860	15,646	18.19	58.83	1,070.29
GT	<i>Ca</i>	83	1,352	16.29	5.81	94.66
	<i>Cd</i>	145	2,578	17.78	9.92	176.35
AC	<i>Ca</i>	66	1,106	16.76	4.62	77.44
	<i>Cd</i>	137	2,354	17.18	9.37	161.03
AAC	<i>Ca</i>	556	10,455	18.8	38.93	732.01
	<i>Cd</i>	689	14,271	20.71	47.13	976.23
GTT	<i>Ca</i>	518	9,840	19	36.27	688.95
	<i>Cd</i>	683	14,346	21	46.72	981.36
AAG	<i>Ca</i>	263	4,731	17.99	18.41	331.24
	<i>Cd</i>	353	6,522	18.48	24.15	446.15
AAT	<i>Ca</i>	237	4,779	20.16	16.59	334.6
	<i>Cd</i>	439	9,495	21.63	30.03	649.52
AAAG	<i>Ca</i>	81	1,528	18.86	5.67	106.98
	<i>Cd</i>	108	2,244	20.78	7.39	153.5
CTTT	<i>Ca</i>	66	1,280	19.39	4.62	89.62
	<i>Cd</i>	117	2,376	20.31	8	162.53
AAAT	<i>Ca</i>	60	1,148	19.13	4.2	80.38
	<i>Cd</i>	62	1,232	19.87	4.24	84.28
GTTTT	<i>Ca</i>	11	255	23.18	0.77	17.85
	<i>Cd</i>	22	555	25.23	1.5	37.97
ATTTT	<i>Ca</i>	11	240	21.82	0.77	16.8
	<i>Cd</i>	23	560	24.35	1.57	38.31
ACAGCA	<i>Ca</i>	10	306	30.6	0.7	21.42
	<i>Cd</i>	12	318	26.5	0.82	21.75
GGTTGT	<i>Ca</i>	8	222	27.75	0.56	15.54
ATTTTG	<i>Cd</i>	11	282	25.64	0.75	19.29

Ca, *Candida albicans*; *Cd*, *Candida dubliniensis*.

243-bp product with two *Cd* reference strains only. The specificity of the primer was tested on some other pathogenic *Candida* and fungal strains along with *Cd*. The primer pair was specific for *Cd* as none of the other strains tested could yield any amplicon under identical conditions of amplification (Fig. 8).

Discussion

Similarity and Variation of Microsatellites Distribution

Our results suggested that microsatellites' distributional pattern is not very similar as *Cd* harbors more

Fig. 3. Distribution of microsatellites for different repeat times in each type of microsatellite class (a–f). Red bar denotes *Candida albicans* and green bar *Candida dubliniensis*.

than twice microsatellites compared to *Ca*. Mahfooz et al. [22] also reported that two species of *Fusarium*, namely *F. verticilloides* and *F. solani*, having the same transcript size showed the difference of approximately

twice the number of microsatellites. The possible explanation of more number of repeated sequences in *Cd* is might be due to the presence of the higher number of clusters of repeated sequences [24]. In turn, these se-

Fig. 4. Comparison of microsatellite numbers in different chromosomes.

Fig. 5. Frequency of different repeat types in each chromosome. The above bar represents microsatellite frequency in *Candida dubliniensis* and the following bar represents microsatellite frequency in *Candida albicans*.

quences undergo a reorganization process (duplication and deletion) twice as frequently in *Cd*, leading to karyotypic instability in its genome. Magee et al. [25] also reported that the higher frequency of translocation in *Cd* is clearly due to the higher number of RPS sub-

units on the major repeat sequence of the chromosomes.

The possible cause of a higher percentage of microsatellites in the whole genome of these *Candida* species (1–2%) might be because the Ascomycetes fungus adopts much more microsatellites than other fungi [26]. Lim et al. [27] reported that *C. albicans* possessed the highest number of each type of repeats among 14 fungi they have studied. The higher number of microsatellites in *Candida* is due to its ability to tolerate more repeat expansions in coding regions [28]. The difference in repeat numbers within *Ca* and *Cd* might be due to the mechanisms of mutation or repair of specific motifs [29].

Our results indicated that mono-nucleotide microsatellites were the most frequent repeat category, consistent with previous studies on eukaryotes [30] and fungal genomes [26, 27, 31]. Similar to the latter reports, we, too, have found the longest stretches of mono-nucleotide repeats (up to 50) mainly with a higher content of polyA and polyT and di-nucleotides with the highest frequency for AT repeats (online suppl. Table 1). This higher A/T frequency may be due to higher AT/TA content in most yeasts and fungi genomes [32]. In addition to that, Testa et al. [33] reported that repeat-induced point mutations in the repetitive DNA of fungal taxa reduce GC content by converting cytosine to thymine residues.

The tri-nucleotides were the second highest occurring microsatellites in these *Candida* species, similar to other fungi [22, 34]. The higher number of tri-nucleotide repeats may be attributed to the fact that these contribute

Fig. 6. Comparative analysis of number (a), relative frequency (b), and relative density (c) of microsatellites in two species with respect to different chromosome sizes.

to the amino acids that may translate into proteins to perform various cellular processes and thus are overrepresented in coding regions [35]. Another important finding was that despite being an AT-rich genome, GC-rich trinucleotides were equally abundant in these species, found in other fungi [31]. The presence of a higher number of (AAC)_n, (AAG)_n, and (AAT)_n motifs (online suppl. Table 1) in our analysis was also reported in *Ca*, in accordance with previously published report [27].

Further mining of *Ca* and *Cd* transcript sequences revealed the presence of a high number of glutamine- and threonine-encoding tri-nucleotide repeats. According to earlier reports, Tec1p is the TEA transcription factor homolog of *Ca* [36], and recently, Rayyan et al. [37] found

that Tec1p harbors 744 amino acids and includes 4 structural regions. The N-terminal glutamine-rich regions are responsible for the DNA binding activity of Tec1p. The C-terminal glutamine-rich regions, on the other hand, are responsible for hyphal activation, that is, mediating biological functions of morphological development, pathogenesis, and invasive growth of *Ca*. Higher threonine-encoding repeats may also be justified as threonine-rich regions are considered the site of O-glycosylation, an essential factor in fungal biology and virulence, and also help in adhesion properties [38]. The frequency of tetra-, penta-, and hexa-nucleotide microsatellites was found to be very low, covering only 3–4% of microsatellites in the whole genome, as in other yeasts and fungi [26].

Fig. 7. Comparison of GC content of microsatellites in the chromosomes of two *Candida* species. The left column represents GC content of microsatellite in *Candida albicans* and the right column represents GC content of microsatellite in *Candida dubliniensis*.

Fig. 8. Gel picture showing the amplification pattern of *Candida dubliniensis*-specific marker. M – 100 bp ladder. 1 – *C. dubliniensis* (type strain CD36), 2 – *C. dubliniensis* (CBS7987), 3 – *Candida albicans* ATCC 90028, 4 – *Candida glabrata* ATCC-MYA 2950, 5 – *C. parapsilosis* ATCC 22019, 6 – *Candida tropicalis* ATCC 750, 7 – *Candida guilliermondii* ATCC 6260, 8 – *Candida krusei* ATCC 14243, 9 – *Trichosporon* spp., 10 – *Cryptococcus* spp., 11–12 – *Fusarium* spp., 13 – *Aspergillus* spp., 14 – negative control.

The GC Content and Repeat Times of Microsatellites

It has already been reported that GC content has a very significant role on the genomic level. The gene concentration varies from lowest in GC-poor to highest in GC-rich regions [39]. The high GC content repeats are related to some diseases in humans and the pathogenesis of some

microorganisms. Similarly, virulence and pathogenesis of *Chlamydia pneumoniae* were influenced by the (G)_n repeats in membrane protein-gene pmp10 [40], and the (C)_n in outer membrane proteins was involved in the pathogenesis of *C. pneumoniae* [41]. However, GC content also plays a significant role in the entire viral genome, such as herpes simplex virus type 1 [42]. The reactivation of herpes simplex virus was reported to be associated with G-string mutants in the thymidine kinase gene. According to our data, a significant portion of microsatellites belong to transcript sequences (Table 1). Also, hexa- and tri-nucleotides were found to share the highest GC content among all the microsatellite types both in the whole-genome and chromosome-wise analysis. These results correlate with many earlier reports saying that CDS are preferentially selected with tri- and hexa-nucleotide microsatellite motifs [43]. Thus, it suggests that tri- and hexa-nucleotides play a significant role in gene encoding sequences in *Ca* and *Cd*. Therefore, these species can serve as an excellent model organism for studying microsatellite evolution and investigating the expansion of specific tri-nucleotide repeats implicated in humans' pathogenic effect.

Several studies reported that microsatellite repeat count influences gene expression. For instance, Miret et al. [44] found that an increase in (CAG)_n repeats from 25 to 30 may lead to drug resistance in *Saccharomyces cere-*

visiae; similarly, *lacZ* expression is hindered in *E. coli* when the number of repeats of GAA is very high [45]. Our study found that the repeat times of microsatellites had some differences within and between 6 microsatellite types, and the maximum repeat times of different microsatellite types were distinct among the two species. We also report that as the length of the repeat increases, the frequency of repeats decreases in each microsatellite class, as seen in other yeasts. It may be because longer repeats have higher mutation rates and, hence, could be less stable. In human colorectal cells, replication errors at (G)₁₆ repeat were 30-fold higher than (G)₁₀, and at (CA)₂₆ were 10-fold higher than (CA)₁₃ [46]. Overall, we can say that the GC content and repeat counts of microsatellites may play significant roles in most species.

Conservation of Motifs

The total number of shared and unique motifs for each species was identified to understand the evolutionary link between these two species. The maximum number of motifs shared between two species was penta-nucleotide repeats (34, 27.9%), followed by tetra (31, 25.4%) and hexa (29, 23.9%) repeats. A total of 16.4% of motifs were found conserved among the two species. The possible reason for the lower percentage of conserved motifs is that *Cd* contains a more diverse type of repeats (due to a higher number of RPS clusters) than *Ca* [24]. Our study supports the results of earlier phylogenetic relationship studies [47]. The disparity in genome level is also due to the evolutionary processes that have occurred during or soon after the two species have diverged from their common ancestor [48].

The two *Candida* species analyzed showed a good number of unique microsatellite motif patterns. Lim et al. [27] found that two closely related species of *C. neoformans* have an entirely different profile for all the higher nucleotide repeats. Thus, the diversity of microsatellite motifs gives each species a unique “signature” of repeat distribution. It may be due to the rapid evolution and de novo production and loss of microsatellites in living organisms. Such unique microsatellite motifs were also reported in a few more fungal species [49, 50]. We utilized these unique microsatellite motifs for the development of species-specific markers.

RA and RD of Microsatellites

Our results revealed that chromosome size was significantly positively correlated with the number of microsatellites but negatively correlated with the RA and RD of microsatellites. The lower abundance microsatellite in *Ca*

is maybe because it has almost less than half the number of mono-nucleotide microsatellites than *Cd*, present in each chromosome. According to Toth et al. [32], RA of microsatellites differs and varies depending upon the species examined. Sometimes, microsatellites get overrepresented in genomes due to differences or biases during the genomes’ mismatch repairing system. Thus, their number gets increased, as may be in the case of *Cd*. The negative correlation between RD and chromosome size was also found in the earlier report [31, 51]. Therefore, it may be possible that microsatellites might not have contributed significantly to the expansion of the genome in evolution.

Although, except chromosome R, there was no significant difference between the chromosome sizes of the two species, the number of microsatellites in *Cd* was about 2-fold more than that in *Ca* for each chromosome. This possible reason could be the karyotype variability among these two species [20, 48]. Different types of microsatellites in both the genome showed significant variation in frequency. Subramanian et al. [52] reported that the frequency of microsatellites is influenced by their secondary structures and the effect on DNA replication.

C. dubliniensis-Specific PCR

We also studied our primers’ specificity by performing the PCR on the template DNA of different *Candida* species along with *Cd*. It is important as differences in *Ca* and *Cd* is either not noticed or often they are misidentified while performing the germ tube test in serum and other routine conventional methods used in the laboratories [53]. In our PCR experiment, we have used the DNA of two reference strains of *Cd*, which gave the desired product with our *Cd*-specific primer. The absence of 243-bp product in other *Candida* species and few other pathogenic fungal species (which are likely to be present in routine patient samples) confirmed that these primers could be utilized selectively and specifically to identify *Cd*. This marker will not only help in early detection of infection in the course of the disease but also has advantages like less number of steps involved to give result in a short time, ease of use by anyone with limited training, more comfortable in analysis or interpretation, and last and foremost important its low cost [54]. All this is important because many of the pathology laboratories still cannot afford to use expensive instruments or chemicals needed in methods like real-time PCR and MALDI-TOF. Here, we tried for a cost-effective approach for *Candida* detection, but the marker may also be used for the qPCR experiments depending upon the availability of a real-time PCR machine in the diagnostic laboratory.

Conclusion

In this study, we observed that microsatellites' distributional pattern is not similar in these phenotypically closely related species. The differentiation of microsatellites also emphasizes the genetic diversity and phylogenetic relationship between these two species. These data will help understand the microsatellite distribution and provide strong support for developing and identifying more species-specific microsatellite markers. Our study concluded that this *Cd*-specific marker could be successfully used in the identification of these species.

Acknowledgements

We thank Dr. Prashant Gupta, Department of Microbiology, KGMU, Lucknow, and Dr. Gulnaz Bashir, Additional Professor, Microbiology, Sher-i-Kashmir Institute of Medical Sciences, Srinagar, Kashmir, J&K, for sharing reference strains used in our PCR analysis and primer validation work.

Statement of Ethics

Our study is exempted from ethics approval as it does not require any human or animal samples. Only the ATCC strains were included in the tests, also the research laboratory has been approved from institute's Biosafety and Ethical Committee.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

All the expenses during the study were borne by P.S., as the work does not have any external funding.

Author Contributions

This work was carried out in collaboration with all authors. P.S. designed the study, performed all the analyses and wrote the manuscript's first draft. R.N. helped in designing and corrections. V.V. managed the wet laboratory experiment and content improvement. All authors read and approved the final manuscript.

References

- 1 Mary EB. *Candida* and candidiasis. *Emerg Infect Dis J*. 2002;8:876.
- 2 Ruhnke M. Epidemiology of *Candida albicans* infections and role of non-*Candida albicans* yeasts. *Curr Drug Targets*. 2006;7(4):495–504.
- 3 Sullivan DJ, Westerneng TJ, Haynes KA, Bennett DE, Coleman DC. *Candida dubliniensis* sp. nov.: phenotypic and molecular characterization of a novel species associated with oral candidosis in HIV-infected individuals. *Microbiology*. 1995;141(Pt 7):1507–21.
- 4 Ellepola AN, Khan ZU. Rapid differentiation of *Candida dubliniensis* from *Candida albicans* by early d-xylose assimilation. *Med Princ Pract*. 2012;21(4):375–8.
- 5 Moran G, Stokes C, Thewes S, Hube B, Coleman DC, Sullivan D. Comparative genomics using *Candida albicans* DNA microarrays reveals absence and divergence of virulence-associated genes in *Candida dubliniensis*. *Microbiology*. 2004;150(Pt 10):3363–82.
- 6 Sullivan D, Coleman D. *Candida dubliniensis*: characteristics and identification. *J Clin Microbiol*. 1998;36(2):329–34.
- 7 Livério HO, Ruiz LDS, Freitas RS, Nishikaku A, Souza AC, Paula CR, et al. Phenotypic and genotypic detection of *Candida albicans* and *Candida dubliniensis* strains isolated from oral mucosa of aids pediatric patients. *Rev Inst Med Trop Sao Paulo*. 2017;59:e14.
- 8 Moran GP, Sanglard D, Donnelly SM, Shanley DB, Sullivan DJ, Coleman DC. Identification and expression of multidrug transporters responsible for fluconazole resistance in *Candida dubliniensis*. *Antimicrob Agents Chemother*. 1998;42(7):1819–30.
- 9 Neppelenbroek KH, Seó RS, Urban VM, Silva S, Dovigo LN, Jorge JH, et al. Identification of *Candida* species in the clinical laboratory: a review of conventional, commercial, and molecular techniques. *Oral Dis*. 2014;20(4):329–44.
- 10 Pincus DH, Coleman DC, Pruitt WR, Padhye AA, Salkin IF, Geimer M, et al. Rapid identification of *Candida dubliniensis* with commercial yeast identification systems. *J Clin Microbiol*. 1999;37(11):3533–9.
- 11 Jabra-Rizk MA, Falkler WA, Merz WG, Baqui AA, Kelley JI, Meiller TF. Retrospective identification and characterization of *Candida dubliniensis* isolates among *Candida albicans* clinical laboratory isolates from human immunodeficiency virus (HIV)-infected and non-HIV-infected individuals. *J Clin Microbiol*. 2000;38(6):2423–6.
- 12 Ells R, Kock JL, Pohl CH. *Candida albicans* or *Candida dubliniensis*? *Mycoses*. 2011;54(1):1–16.
- 13 Liguori G, Di Onofrio V, Lucariello A, Gallé F, Signoriello G, Colella G, et al. Oral candidiasis: a comparison between conventional methods and multiplex polymerase chain reaction for species identification. *Oral Microbiol Immunol*. 2009;24(1):76–8.
- 14 Mahfooz S, Srivastava A, Yadav MC, Tahoor A. Comparative genomics in phytopathogenic prokaryotes reveals the higher relative abundance and density of long-SSRs in the smallest prokaryotic genome. *3 Biotech*. 2019;9(9):340.
- 15 Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Vieira MLC. Origin, evolution and genome distribution of microsatellites. *Genet Mol Biol*. 2006;29(2):294–307.
- 16 Sunnucks P. Efficient genetic markers for population biology. *Trends Ecol Evol*. 2000;15(5):199–203.
- 17 Schlotterer C. Evolutionary dynamics of microsatellite DNA. *Chromosoma*. 2000;109(6):365–71.
- 18 Datta S, Mahfooz S, Singh P, Choudhary AK, Singh F, Kumar S. Cross-genera amplification of informative microsatellite markers from common bean and lentil for the assessment of genetic diversity in pigeonpea. *Physiol Mol Biol Plants*. 2010;16(2):123–34.
- 19 Muzzey D, Schwartz K, Weissman JS, Sherlock G. Assembly of a phased diploid *Candida albicans* genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure. *Genome Biol*. 2013;14(9):R97.

- 20 Jackson AP, Gamble JA, Yeomans T, Moran GP, Saunders D, Harris D, et al. Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. *Genome Res.* 2009;19(12):2231–44.
- 21 Martins WS, Lucas DC, Neves KF, Bertoli DJ. Websat: a web software for microsatellite marker development. *Bioinformatics.* 2009; 3(6):282–3.
- 22 Mahfooz S, Srivastava A, Srivastava AK, Aroora DK. A comparative analysis of distribution and conservation of microsatellites in the transcripts of sequenced *Fusarium* species and development of genic-SSR markers for polymorphism analysis. *FEMS Microbiol Lett.* 2015;362(17):fnn131.
- 23 Harju S, Fedosyuk H, Peterson KR. Rapid isolation of yeast genomic DNA: Bustin' grab. *BMC Biotechnol.* 2004;4:8.
- 24 Sullivan DJ, Moran GP, Coleman DC. *Candida dubliniensis*: ten years on. *FEMS Microbiol Lett.* 2005;253(1):9–17.
- 25 Magee BB, Sanchez MD, Saunders D, Harris D, Berriman M, Magee PT. Extensive chromosome rearrangements distinguish the karyotype of the hypovirulent species *Candida dubliniensis* from the virulent *Candida albicans*. *Fungal Genet Biol.* 2008;45(3):338–50.
- 26 Li CY, Liu L, Yang J, Li JB, Su Y, Zhang Y, et al. Genome-wide analysis of microsatellite sequence in seven filamentous fungi. *Interdiscip Sci.* 2009;1(2):141–50.
- 27 Lim S, Notley-McRobb L, Lim M, Carter DA. A comparison of the nature and abundance of microsatellites in 14 fungal genomes. *Fungal Genet Biol.* 2004;41(11):1025–36.
- 28 Young ET, Sloan JS, Van Riper K. Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. *Genetics.* 2000; 154(3):1053–68.
- 29 Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet.* 2004; 5(6):435–45.
- 30 Sharma PC, Grover A, Kahl G. Mining microsatellites in eukaryotic genomes. *Trends Biotechnol.* 2007;25(11):490–8.
- 31 Karaoglu H, Lee CM, Meyer W. Survey of simple sequence repeats in completed fungal genomes. *Mol Biol Evol.* 2005;22(3):639–49.
- 32 Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome research.* 2000;10:967–81.
- 33 Testa AC, Oliver RP, Hane JK. Occultercut: a comprehensive survey of AT-rich regions in fungal genomes. *Genome Biol Evol.* 2016; 8(6):2044–64.
- 34 Singh R, Pandey B, Danishuddin M, Sheoran S, Sharma P, Chatrath R. Mining and survey of simple sequence repeats in wheat rust *Puccinia* sp. *Bioinformatics.* 2011;7(6):291–5.
- 35 Cavagnaro PF, Senalik DA, Yang L, Simon PW, Harkins TT, Kodira CD, et al. Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.). *BMC Genomics.* 2010;11:569.
- 36 Shareck J, Nantel A, Belhumeur P. Conjugated linoleic acid inhibits hyphal growth in *Candida albicans* by modulating Ras1p cellular levels and downregulating *TEC1* expression. *Eukaryot Cell.* 2011;10(4):565–77.
- 37 Rayyan WA, Singh A, Al-Jaafreh A, Dayyih WA, Bustami M, Salem S, et al. The role of glutamine-rich region of *Candida albicans* *Tec1p* in mediating morphological transition and invasive growth. *Int J Med Res Health Sci.* 2019;13:150–8.
- 38 Willaert RG. Adhesins of yeasts: protein structure and interactions. *J Fungi.* 2018 Oct 27;4(4):119.
- 39 Ren L, Gao G, Zhao D, Ding M, Luo J, Deng H. Developmental stage related patterns of codon usage and genomic GC content: searching for evolutionary fingerprints with models of stem cell differentiation. *Genome Biol.* 2007;8(3):R35.
- 40 Grimwood J, Olinger L, Stephens RS. Expression of *Chlamydia pneumoniae* polymorphic membrane protein family genes. *Infect Immun.* 2001;69(4):2383–9.
- 41 Rocha EP, Blanchard A. Genomic repeats, genome plasticity and the dynamics of mycoplasma evolution. *Nucleic Acids Res.* 2002; 30(9):2031–42.
- 42 Ouyang Q, Zhao X, Feng H, Tian Y, Li D, Li M, et al. High GC content of simple sequence repeats in herpes simplex virus type 1 genome. *Gene.* 2012;499(1):37–40.
- 43 Mayer C, Leese F, Tollrian R. Genome-wide analysis of tandem repeats in *Daphnia pulex*: a comparative approach. *BMC Genomics.* 2010;11:277.
- 44 Miret JJ, Pessoa-Brandão L, Lahue RS. Orientation-dependent and sequence-specific expansions of CTG/CAG trinucleotide repeats in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A.* 1998;95(21):12438–43.
- 45 Liu L, Dybvig K, Panangala VS, van Santen VL, French CT. GAA trinucleotide repeat region regulates M9/PMGA gene expression in *Mycoplasma gallisepticum*. *Infect Immun.* 2000;68(2):871–6.
- 46 Campregher C, Scharl T, Nemeth M, Honeder C, Jascur T, Boland CR, et al. The nucleotide composition of microsatellites impacts both replication fidelity and mismatch repair in human colorectal cells. *Hum Mol Genet.* 2010;19(13):2648–57.
- 47 Imran ZK. *Candida albicans* ssp. *dubliniensis* stat. et comb. nov., a new combination for *Candida dubliniensis* based on genetic criteria. *Afr J Microbiol Res.* 2015;9:1205–14.
- 48 Brenda B, James A, Benjamin DK. Neonatal candidiasis. *J Pediatr.* 2014;79:123–5.
- 49 Mahfooz S, Singh SP, Mishra N, Mishra A. A comparison of microsatellites in phytopathogenic *Aspergillus* species in order to develop markers for the assessment of genetic diversity among its isolates. *Front Microbiol.* 2017; 8:1774.
- 50 Mahfooz S, Singh SP, Rakh R, Bhattacharya A, Mishra N, Singh PC, et al. A comprehensive characterization of simple sequence repeats in the sequenced trichoderma genomes provides valuable resources for marker development. *Front Microbiol.* 2016;7:575.
- 51 Srivastava S, Avvaru AK, Sowpati DT, Mishra RK. Patterns of microsatellite distribution across eukaryotic genomes. *BMC Genomics.* 2019;20(1):153.
- 52 Subramanian S, Mishra RK, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol.* 2003; 4(2):R13.
- 53 Abaci Ö, Halki-Uztan A, Ates M. Specific identification of *Candida albicans* and *Candida dubliniensis* by PCR using species-specific primers. *Ann Microbiol.* 2008;58:325.
- 54 Singh P, Nath R, Venkatesh V. Molecular methods for diagnosing candida infection: a comparative study and review of literature. *Asian J Microbiol Biotechnol.* 2019;4:84–107.