Current Trends in Biotechnology and Pharmacy
Vol. 16 (2)195 - 202, April 2022, ISSN 0973-8916 (Print), 2230-7303 (Online)
10.5530/ctbp.2022.2.18

A Comparative *In Silico* Analysis on Frequency and Distribution of Microsatellites in Whole Genome Sequences of Three Pathogenic *Candida* species and Development of SSR markers for Diversity Analysis

Pallavi Singh[1,2,3], Ravindra Nath[2], Vimala Venkatesh[3], Prashant Gupta[3]

[1] Department of Biotechnology, Dr APJ Abdul Kalam Technical University, Lucknow, UP, India
[2] Department of Computer Science & Engineering, UIET, CSJM University, Kanpur, UP, India;
[3] Department of Microbiology, King George's Medical University, Lucknow, UP, India

Abstract

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. 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 infection. In the present study, we surveyed the microsatellites in the available whole-genome sequences of three pathogenic species viz. *C.albicans, C.glabrata* and *C.tropicalis*. The relative abundance and density was higher in *C.albicans* when compared to *C.glabrata* and *C.tropicalis*. Thirty microsatellite primers were designed, ten from each species, for genetic characterization of *Candida* isolates belonging to six *Candida* species. Of the 30 markers, only nineteen showed amplifications. A total of 32 alleles were amplified by 19 primers with an average of two alleles per marker. Ten markers showed 100% polymorphism. The markers were found to be more polymorphic (75%) in *C.albicans* as compared to *C.glabrata and C.tropicalis,* however, polymorphic information content was the maximum (0.75) in CgSSR 4. Twelve polymorphic markers obtained in this study clearly demonstrate the utility of newly developed markers in establishing genetic relationships among different isolates of *Candida.*

Keywords: *Candida*, microsatellites, polymorphic markers, genetic diversity

Introduction

Candidiasis has emerged as an alarming opportunistic disease with the increase in number of patients who are immunocompromised, aged, receiving prolonged antibacterial and aggressive cancer chemotherapy or undergoing invasive surgical procedures and organ transplantation (1,2) In India, *Candida albicans* along with *C. parapsilosis, C. dubliniensis, C. glabrata* and *C. tropicalis* are considered to be the commonest and most virulent pathogenic species of the genus *Candida.* Delay in speciation of *Candida* isolates by conventional methods and resistance to antifungal drugs in various *Candida* species are some of the factors responsible for the increase in morbidity and mortality. So, the rapid detection and identification of *Candida* isolates is very important for the proper management of patients having candidiasis.

DNA markers are reported to be used for genetic diversity, evolutionary studies (3,4).With the increased availability of sequenced genome information in public domain, microsatellites have become a marker of choice because of their reproducibility, multi-allelic nature, codominant inheritance, relative abundance and high genome coverage (5,6) SSR patterns with in a genome could reflect

the variability up to species level, which make possible to increase the resolution of existing markers to discriminate individual strain or species. In the present study an *in-silico* approach was used to compare the frequency and distribution of SSRs in the sequenced genomes of three *Candida* species viz. *Candida albicans*, *C. glabrata* and *C. tropicalis*. After retrieving all the SSRs, primers were designed for diversity analysis and polymorphism studies among the *Candida* species.

Materials and Methods

Source of EST and annotated transcript sequences

The available genome sequences of *Cg, Cp* and *Ct* were downloaded from National Center for Biotechnology Information (www.ncbi.nln1.nih.gov). The identification of microsatellites was carried out using Rabin Karp Algorithm (7). All SSRs were analyzed for their frequency of occurrence, density, and relative abundance.

Thirty SSR primers representing ten from each species were randomly selected for PCR amplification to study their utility in revealing polymorphism. Primers complimentary to the flanking regions of selected microsatellites were designed using the program primer 3 online software (frodo.wi.mit.edu/).

Fungal isolates

A total of 24 different *Candida* isolates including ATCC strains and clinical isolates, which include five each of *C.albicans, C.glabrata, C.tropicalis , C,parapsilosis* and two each of *C.gulliermondii* and *C.dubliniensis* (Table 1.) were obtained from microbiology Department of King George's Medical University.

DNA isolation and SSR amplification

Total genomic DNA extraction was carried out from 24 isolates using Bust 'n' Grab method (8). The PCR mixture was prepared as described in (9). PCR program was as initial denaturation at 95 °C for 3 min, subsequently, five touch-clown PCR cycles comprising of 94 °C for 20 s, 60/ 35 °C (depending on the marker as given in Table 2) for 20 s, and 72 °C for 30 s were performed. These cycles were followed by 40 cycles of denaturation at 94 °C for 20 s with constant annealing temperature of 56/31 °C (depending on marker) for 20 s, and extension at 72 °C for 2s, and a final extension it 72 °C for 20 min.

All PCR amplicons were resolved by electrophoresis on 3% agarose gel to identify the informative SSR loci across all the isolates. Generuler 100-bp DNA ladder (MBI Fermentas) was used to estimate the allele size.

Statistical analysis

The amplification data generated by SSR markers were analyzed using SIMQUAL route to generate Jaccard's similarity coefficient (10) Using NTSYS-PC software version 2.1 (11). These similarity coefficients were used to construct a dendrogram depicting genetic relationships among the isolates by employing the Unweighted Paired Group Method of Arithmetic Averages (UPGMA) algorithm and SAHN clustering. The robustness of the dendrogram was evaluated with a bootstrap analysis performed on the binary dataset using WINBOOT software (version. 2.0).

Evaluation of polymorphism

The allelic diversity or Polymorphism information Coefficient (PIC) was measured as described by Botstein et. al. (12). PIC is defined as the probability that two randomly chosen copies of gene will be different alleles within a population. The PIC value was calculated with the formula as follows:

$$PIC_i = 1 - \sum_{j=1}^{n} P_{ij}^2$$

where, P represents the frequency of the j^{th} pattern for marker i, and summation extends over r^i patterns.

Results and Discussion

The frequency of repeat motifs in the genome sequence of three species was assessed, and both perfect and compound SSRs were selected with a minimum acceptable length of 12 bp for di, tri, and tetra-nucleotide motifs (13). Only SSRs with a minimum of three repeats were included in the analyses of penta- and hexa-nucleotide repeats. Maximum number of SSR (10,268) was identified in *Ca* followed by *Ct* (9307) and *Cg* (1937). The higher number of SSRs in *Ca* and *Ct* was expected because the genome size was higher (14.46 Mb and 14.86 Mb) compared to *Ct*, as reported by Tian et al (14) genome nucleotide content might infuence the frequency of microsatellites. To compare the SSR count between all three *Candida* species the complete length of each set of sequences was analyzed, and thus, total relative abundance anal total relative density was calculated and depicted in Table 2. It was found that relative abundance of SSRs in *Ca* (710.09) was higher than *Ct* (626.312) and *Cg* (153.73). Similarly, the relative density of SSR was also higher in *Ca* (13673.58) in comparison with *Ct* (11363) and *Cg* (2588.25) (Table 2, Figure 1).

The maximum frequency of SSRs among all three sequence sets was of tri-nucleotide repeats (50.84%). Tetra-nucleotide repeats constituted the second most frequent motif (20.37%) followed by di-nucleotide

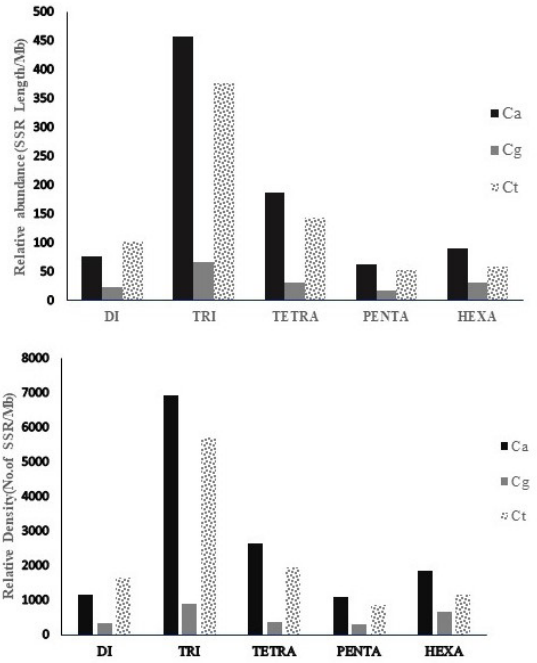

Fig.1 Graphical representation of (a) relative abundance and (b) relative density of different SSR type found in whole-genome sequences of three Candida species

(11.42%) hexa-nucleotide (9.94%). The least penta-nucleotide repeats represented 7.4% repeat motifs in sequences of all three *Candida* species. This agrees with the results from other eukaryotes, where tri-nucleotide repeats are overrepresented in coding region (9,13). Kim et.

Table 1: Number and distribution of SSRs in whole genome sequences of three *Candida* species

	Candida albicans	Candida glabrata	Candida tropicalis
Genome size (Mb)	14.6	12.3	14.4
Number of SSR identified	10,268	1937	9307
Perfect SSRs	8477 (82.55%)	1793(92.56%)	7985(85.79%)
Compound SSRs	1756(17.10%)	133(6.86%)	1302(13.98%)
Total length of SSRs (bp)	197720	32612	168855
Percent length	1.36%	0.26	1.13
Relative density (bp per Mb)	710.096	153.730	626.312
Relative abundance (SSR per Mb)	13673.58	2588.25	11363.0

Table 2. Percentage, relative abundance, and relative density of SSR sequence sets of three Candida species.

	Class	Count	Percentage	RA	RD
Ca	Di	1110	WG	76.76	1156.43
	Tri	6602	52.31	456.57	6920.75
	Tetra	2703	21.41	186.93	2643.71
	Penta	910	7.21	62.93	1097.51
	Hexa	1297	10.28	89.70	1855.19
Cg	Di	297	13.85	23.57	345.71
	Tri	846	39.44	67.14	910.47
	Tetra	391	18.23	31.03	382.22
	Penta	218	10.16	17.30	293.65
	Hexa	393	18.321	31.19	656.19
Ct	Di	1527	13.99	102.76	1643.88
	Tri	5609	51.41	377.46	5701.21
	Tetra	2137	19.59	143.81	1965.28
	Penta	774	7.09	52.09	878.53
	Hexa	864	7.92	58.14	1174.16

Table 3. Details of locus, primer sequence, *Tm,* Motif, no. of alleles, alleles size, percentage polymorphism, and PIC value of different primers used to evaluate genetic diversity within *Candida species*

Primer Name/ locus	Forward primer	Reverse primer	Tm (C)	Motif	No. of alleles	Allele size	% Polymorphism	PIC
CaSSR2	TTCGTGTACGTCAAGGATAA	TTCGTGTACGTCAAGGATAA	54.3	(GA)6	2	183	100	0.33
CaSSR3	TTGTTCAACTCCATGACTGA	TAGATCCAACCAAGGACAAC	54.9	(TA)6	2	196	-	
CaSSR4	TCAACGAAGAATCCACTGA	AATGATTAGCGAATTGGTTG	55	(GAC)5	1	261	100	0.498
CaSSR6	TTACGTACGGTATCGTGTTG	ACGAGCTTCTGTAGTTCGAC	54.7	(TGC)4	2	223	100	0.594
CaSSR7	ATTGCTTGAGGAATACTTGC	TCGTCTATTAATGAGCTGGC	55,6	(CGA)4	2	183	-	
CaSSR8	CAGAGGACGAAGTGGAGAG	CAACACCTCTGTCATCTGC	54.9	(GGA)4	3	296	100	0.443
CaSSR9	CAATCCCAATCTCAAGCTAC	TCATCGTTGTCATCTTCATC	54.3	(ATG)4	4	292	100	0.358
CaSSR10	AGTCTAGTTGGTGTTGGTGC	GCGTAATCGAAGTTCTCATC	55	(ATGC)3	2	233	50	0.082
CgSSR 1	TCCAGAAAAATTTCCAAAAA	CGTAGCTGGTGATATGGTTT	55	(GT)8	1	275	-	
CgSSR 4	TAGTAATCCTTGGCTCCTGA	GTTCATTCACTTCGGATGAT	55	(ATC)4	2	181	50	0.75
CgSSR5	CTGTGAAGAACAATGCAAAA	CAAGTTCTGAACCAGCCTAC	55	(TAA)4	1	161	100	0.234
CgSSR8	CTTTTCAAATGGAGAGCAAC	AATGCAATCATAGCCTTTGT	55	(AAG)5	1	190	-	
CgSSR9	GAAAGCTAGACCCAGTGAAA	CCTTTTATCCCATTTTCCTT	54.9	(AGC)8	2	232	100	0.594
CgSSR10	GAAAGGACACGACATCAACT	TGCAGTCTTGAAAGGAATCT	55	(GTAC)3	2	166	-	
CtSSR1	ACCAGTACCTACCATAGATGC	GAGGGGGTTATACCCATACT	53.3	(TA)7	1	298	100	0.707
CtSSR2	TACTCGAAGAGCAGGAAAAG	TAATTACGCATTGACTGTCG	54.9	(AG)6	1	262	100	0.498
CtSSR6	CTTGGAATCAACTTGGTCAT	GTTTCTGACTTCTTCAACGG	55.03	(CAA)4	1	200	-	
CtSSR8	GTTTTCTTTGGTGTTGTCGT	CAAGAAGGTCAAGGTTGAAG	55.19	(CGT)4	1	292	100	0.234
CtSSR9	GACCAATTGGAGTACTTGGA	CCTTCTTCTTGTCTTCATCG	55.03	(GAA)4	1	202	-	-

able 4. A comparison between CaSSR, CgSSR and CtSSR markers to estimate the level ofpoly-morphism revealed by them.

	CaSSR markers	CgSSR markers	CtSSR markers	Total markers
Markers used	10	10	10	30
Marker amplified	8 (80%)	6(60%)	5(50%)	19 (63.33%)
No. of monomorphic markers	2 (25%)	3(50%)	2(40%)	7(36.84%)
No. of polymorphic markers	6 (75%)	3(50 %)	3(60%)	12 (63.15%)
Average PIC value	0.384	0.526	0.479	0.443
No. of alleles amplified	18	9	5	32
Similarity coefficient value (Avg.)	0.68	0.64	0.57	0.63

al. (15) reported that the tri-nucleotides in the coding regions are translated into amino-acid repeats which contribute to biological function to proteins. Lower number of penta-nucleotides repeats in fungal genome was also reported by Mahfooz et al (16).

DNA polymorphism

A total of nineteen SSR markers (eight from *Ca,* six from *Cg,* and five from *Ct)* amplified easily scorable bands ranged from 166 to 296 bp in all the isolates. Of the nineteen markers, five amplified di-nucleotide repeats, twelve amplified tri-nucleotide repeats, and only two markers were able to amplify tetra-nucleotide repeat. We used three indexes (percentage of polymorphic SSRs, number of alleles per locus and PIC value) to indicate SSR polymorphism level. Among all the markers, twelve markers (63.15%) were polymorphic, whereas rest seven markers (36.84%) were monomorphic. A total of 32 alleles were amplified by nineteen markers. We detected 1-4 alleles per microsatellite locus with an average of 1.68 per marker. *CaSSR* markers simplified eighteen alleles with 2.28 allele per locus, whereas *CgSSR* markers detected nine alleles with 1.5 alleles per locus and *CtSSR* markers detected five alleles with one allele per locus. Maximum number of alleles (4) were simplified by *CaSSR*9, while minimum one allele was amplified with nine markers *viz*. *CaSSR*4, *CgSSR*1, *CgSSR*5, *CgSSR*8, *CtSSR*1, *CtSSR*2, *CtSSR*6, *CtSSR*8 and *CtSSR*9 (Table 3). Of nineteen amplified markers polymorphic markers, ten showed 100% polymorphism and two showed 50% polymorphism (CaSSR10 and CgSSR4). On comparison of polymorphism potential of markers derived from each *Candida* species, of six SSR markers from *CgSSR* and five SSR markers from *CtSSR,* only three (50%) and two (40%) markers were found polymorphic, respectively (Table 4). *CaSSR* markers exhibited highest percentage of polymorphisms (75%), as six out of ten markers were found polymorphic. Among the polymorphic markers, the maximum PIC value was obtained with *CgSSR* (0.526) and minimum with *CaSSR* (0.384), the average being 0.443. The low value of PIC in our study maybe contributed to the fact that SSRs represent the coding region of genome which is generally conserved. The number of alleles per locus varied according to the origin of the marker. Markers with PIC values of > 0.50, such as *CaSSR6*, *CgSSR9* (0.594), *CgSSR10* (0.70) and *CgSSR4* (.75), will be highly informative for genetic studies and arc extremely useful in distinguishing the polymorphism rate of the marker at specific locus. High levels of polymorphism associated with microsatellites are expected because of the unique mechanism responsible for generating microsatellite allelic diversity by replication slippage rather than by simple mutations or insertions/deletions (17). Also, the transferability of markers within *Candida* species demonstrated that the regions flanking these microsatellites are conserved

enough to allow locus amplification.

Diversity and cluster analysis

The similarity coefficient value between isolates lies between 0.16 to 0.96 with the mean of 0.63 for all 276 isolate combinations used in the present study. For microsatellite markers derived from *CaSSR* markers, the similarity coefficient values between isolates ranged from 0.25 to 1.00 with average genetic diversity of 32%. Similarly, with *CgSSR*-derived SSR markers, the similarity coefficients between isolates ranged from 0.5 to 1.00 with 36% genetic diversity. For *CtSSR* markers, similarity coefficient value ranged from 0.2 to 1.0 with an average diversity being 42.7% (Table 4).

The highest similarity coefficient (0.96) was observed between three groups of isolates Cg1-Cg3, Cgul1-Cgul2 and Ct5-Ct4 which was closely followed by two isolates Cp5-Cp3 and Ca1-Cd2 (0.94). The similarity was expected because all the above isolates belong to same species group (with the exception of Ca1-Cd2). The most diverse (similarity coefficient value 0.16) isolates were Ca3 and Cg1. The dendrogram constructed based on similarity index resulted in two major clusters (Fig. 2). High boot- strap values were recorded with internodes, which indicate the robustness of the clustering. The first major cluster has been exclusively composed of *C.glabrata* isolates, which is further divided into two subclusters having three and two *C.glabrata* isolates with the exception of one *C.tropicalis* isolates. The second cluster having different subclusters comprises a mix of all other *Candida* species taken into this study except *C.glabrata.*

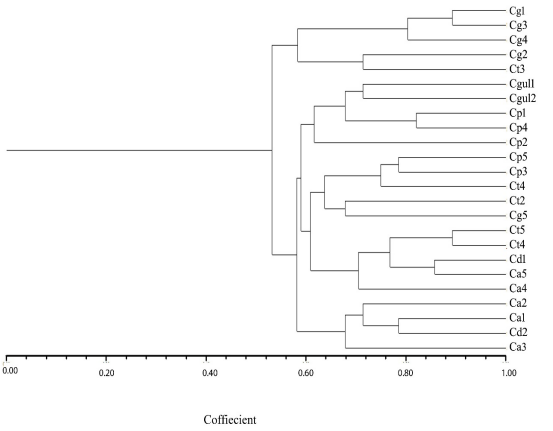

Fig. 2. Dendrogram showing genetic relationship among the Candida isolates based on 30 microsatellites markers. Scale indicates Jaccard's coefficient of similarity.

The dendrogram obtained based on the similarity index values clearly justifies and clusters the various isolates according to the earlier evolutionary studies. The second subcluster contains two sub-clusters containing *C.albicans* and *C.dubliniensis* species together depicting their close evolutionary relatedness as also proven by a study from Imran (18). Similarly, various earlier reports (19,20.21) justify the occurrence of *C.tropicalis*, *C.gulliermondii* and *C. parapsilosis* isolates on same cluster along with *C.albicans*.

It was surprising that isolate Ct3 clustered in a separate cluster along with *C. glabrata* isolates, that may be due to some error in sampling or DNA mixing while performing the PCR experiments.

Conclusion

To our knowledge, this is the first attempt to extensively develop SSR markers from the whole genome sequences of different *Candida* species. This study clearly demonstrates the utility of newly developed SSR markers in establishing genetic relationships among different species of *Candida*. Although the number of useful markers was low, all the isolates could be differentiated from each other. These markers can be further utilized for addressing genetic relatedness in other species of *Candida* because SSR markers have a reputation of being highly transferable.

References

1 Sakamoto S and Miura Y. (1993). Improved

survival from fungaemia in patients with hematological malignancies: Analysis of risk factor for death and usefulness of early antifungal therapy. European Journal of Haematology, 51: 156-160.

2. Merlino J., Tambosis E. and Veal (1998). Chromogenic tube test for presumptive identification or confirmation of isolates as *Candida albicans.* Journal of Clinical Microbiology, 36: 1157-1159.

3. Parvin S., N, V., T, V., and Md, S. A. (2022). Study of the genetic variations in different variants of *Tribulus terrestris* L. in Rayalaseema region in Andhra Pradesh. Current Trends in Biotechnology and Pharmacy, 15(6), 104–107.

4. Mahfooz S., Singh P. and Akhter Y. (2022). A comparative study of microsatellites among crocodiles and development of genomic resources for the critically endangered Indian gharial. Genetica. 150(1):67-75.

5. Varshney R.K., Graner A. and Sorrells M.E. (2005). Genic microsatellite markers in plants: features and applications. Trends Biotechnology, 23**:** 48-55.

6. Chaudhary P. and Sharma P.E. (2022). Microsatellite polymorphism in relation to geographical distribution and adaptation of Seabuckthorn (*Hippophaerhamnoides* L.) in the Indian Himalayas. Current Trends in Biotechnology and Pharmacy.16(1): 1-13.

7. Singh P. and Nath R. (2021) Rabin-Karp algorithm-based microsatellite searching in whole-genome. Bioinfolet, 18 (1 A): 25 – 31.

8. Harju S., Fedosyuk H., Peterson K.R. (2004) Rapid isolation of yeast genomic DNA: Bust n' grab. BMC Biotechnology,4:8.

9. Singh P., Nath R. and Venkatesh V. (2021) Comparative Genome-Wide characterization of microsatellites in *Candida albicans* and *Candida dubliniensis* leading to the development of species-specific marker. Public Health Genomics, 24:1-13.

10. Jaccard P. (1908) Nouvelle recherches sur La distribution fiorale. *Buh Soc* Scud Ser Nat 44*:* 223-270.

11. Rohlf F.J. (1998) NTSYS-PC Numerical Taxonomy and Multivariate Analysis System Version 2.02h. Exeter Software, Applied Biostatistics, New York.

12. Botstein D., White K.L., Skolnick M., Davis R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Animal and Human Genetics, 32: 314 -331.

13. Garnica D.P., Pinzón A.M., Ocampo L.M.Q., Bernal A.J., Barreto E., Grunwald N.J. and Restrepo S. (2006) Survey and analysis of microsatellites from transcript sequences in *Phytophthora* species: frequency, distribution, and potential as markers for the genus. BMC Genomics, 7: 245.

14. Tian X., Strassmann J.E. et al (2011) Genome nucleotide composition shapes variation in simple sequence repeats. Molecular Biology and evolution, 28(2):899–909.

15. Kim T.S., Booth J.G., Gauch H.G. Jr, Sun Q., Park J., Lee Y.H. and Lee K. (2008) Simple sequence repeats in *Neurospora crassa:* distribution, polymorphism and evolutionary inference. BMC Genomics, 9: 31.

16. Mahfooz S., Maurya D.K., Srivastava A.K., Kumar S., Arora D.K. (2012.) A comparative *in silico* analysis on frequency and distribution of microsatellites in coding regions of three formae speciales of *Fusarium oxysporum* and development of EST-SSR markers for polymorphism studies. FEMS Microbiology Letters ,328(1):54–60.

17. Tautz D. (1989) Hyper variability of simple sequences as a general source of

polymorphism DNA markers. Nucleic Acids Research, 17: 6463-6471.

18 Imran Z.K. (2015) *Candida albicans* ssp. *dubliniensis* stat. et comb. nov., a new combination for *Candida dubliniensis* based on genetic criteria. African Journal Microbiology Research ,1205–14.

19. Butler G., Rasmussen M., Lin, M. *et al.* (2009) Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. Nature, 45: 657–662.

20. Brenda A. M. and David C.C. (2014) Molecular epidemiology, phylogeny and evolution of *Candida albicans* Infection, Genetics and Evolution, 21:166-178.

21. Mancera E., Frazer C., Porman A.M., Ruiz-Castro S., Johnson A.D. and Bennett R.J. (2019) Genetic modification of closely related *Candida* species. Frontiers in Microbiology, 10:357.