Evidence that *SSR1* can act as a Hypermutable Contingency Gene in *Candida albicans*

A thesis presented in partial fulfillment of the requirement for the degree of

Master of Science

in

Microbiology

at Massey University, Palmerston North,
New Zealand

Zhuo Zhou

2010

Abstract

During adaptation to the host environment, many microorganisms undergo rapid variation in cell surface phenotype through genetic alteration in hypermutable contingency genes. One of the main mechanisms underlying these changes is alteration in the number of DNA repeat units that results in a large and flexible repertoire of similar but non-identical surface proteins. *SSR1*, a gene in the opportunistic pathogen *Candida albicans*, encodes a repeat-containing cell wall protein which may play a role in maintaining cell wall strength. This gene contains 2 regions with multiple 6 bp tandem repeat units, encoding the amino acids serine and alanine, separated by a 200 bp non-repetitive DNA region. This study investigated whether *SSR1* was a hypermutable contingency gene. Among a worldwide collection of 96 infection-causing *C. albicans* strains, 24 alleles and 40 allele combinations were identified by fluorescent-based genotyping of *SSR1* PCR products. Sequencing results confirmed that the differences in allele size were caused by variation in number of tandem repeats. Two very similar allele combinations were overrepresented (30% and 28%) among a cluster of general-purpose genotype (GPG) strains (which is the most widespread cluster) compared with non-GPG strains (Fisher's exact test, $P=0.0001$ and $P<0.0001$). Among a worldwide collection of 36 commensal GPG *C. albicans* strains, 8 allele combinations were identified by genotyping. One of the two allele combinations that were overrepresented in GPG infection-causing strains was found significantly less in GPG commensal strains (Fisher's exact test, $P=0.0004$). After culture of *C. albicans* cells in vitro for 300 generations, mutation of repeats in *SSR1* occurred, giving a high mutation rate of 1.11×10^{-4} per cell division. The results indicate that *SSR1* is a hypermutable gene and that it shows clade-specificity with the GPG cluster. Growth in a rat model did not seem to cause variation in *SSR1* and human host body sites did not seem to be associated with specific *SSR1* alleles, suggesting that *SSR1* is not used for short-term adaptation in these environments. However, the different allele distribution in commensal and infection-causing GPG strains suggest that *SSR1* may have a role in short-term adaptation in GPG strains, contributing to the change between commensalism and infection. In this case, *SSR1* may act as a hypermutable contingency gene.

Acknowledgement

I would like to convey my big thanks to my supervisors Jan Schmid and co-supervisor Zoe Jordens who have support me the whole way through, without whose guidance and encouragement this would never have happened. You are the best supervisors I could ever have, Thank you!

I would like to thank lab members. Thanks to Ningxin Zhang for her help in the lab. Huge thanks to Shuguang Zhang for his help, without whose assistance and patience to all my questions, I would have struggled to achieve results at all. Thanks to Synodalia Wattimena, Yeying Wang and Panimalar Vijayan for their help, cooperation and all the fun times.

I would also like to thank the Allan Wilson Centre for their great service and assistant.

Thanks to Jing Wang, Melissa Guo and Charlene Zhang for their mental support. Thanks to my home stay family Anne and Satyendra Ram for providing constant encouragement and caring.

Finally, I need to thank my parents who have given the best of themselves and supported me all these years. To them I dedicate this thesis.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ALS	agglutinin-like sequence
aa	amino acid
APS	ammonium persulfate
AWCGS	Alan Wilson Centre Genome services
bsn	boiled supernatant
C. albicans	*Candida albicans*
CFEM	common in several fungal extracellular membrane
CWPs	cell wall proteins
Da	Dalton
DTT	DL-Dithiothreitol
E.coli	*Escherichia coli*
EDTA	Ethylene Diamine Tetraacetic Acid
GPG	general purpose group
GPI-CWPs	glycosylphosphatidylinositol-linked cell wall proteins
IPTG	Isopropyl β-D-1-thiogalactopyranoside
LB Medium	Luria-Bertani medium
Non-GPG(ngpg)	non-general purpose group
ORF	open reading frame
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
RRs	Repeat regions
S. cerevisiae	*Saccharomyces cerevisiae*
SDS	sodium dodecyl sulfate
SERV	Sequence-based Estimation of Repeat Variability
SSR	short sequence repeats
Ssr1p	Ssr1 protein
STRs	short tandem repeats
TBE	Tris/Borate/EDTA
TEMED	Tetramethylethylenediamine
TRs	tandem repeats
X-gal (BCIG)	bromo-chloro-indolyl-galactopyranoside
YPD Medium	Yeast Extract Peptone Dextrose Medium

CHAPTER ONE – INTRODUCTION

1.1. Overview of the importance of infections caused by *Candida albicans*

The genus *Candida*, a group of diploid yeasts, contains many of the most important opportunistic fungal pathogens of human beings and other warm-blooded animals (Shepherd *et al.*, 1985; Pitarch *et al.*, 2006). It causes infection mainly in immunocompromised people, including young children, HIV-positive individuals and patients with AIDS, cancer and organ transplants. Widely indiscriminate use of antibiotics can also lead to infections with Candida spp., known as candidiasis (Rinaldi *et al.*, 1992; Perepnikhatka *et al.*, 1999; Pitarch *et al.*, 2006; Panasiti *et al.*, 2007). A host's weak immune system and the ability of these fungi to adapt leads to a wide range of infections, including skin, oral, genital, urinary and peritoneal candidiasis, hematogenous infection, endophthalmitis, meningitis, arthritis, endocarditis and neonatal candidiasis (Rinaldi *et al.*, 1992).

Within the genus, *C. albicans, C. tropicalis*, *C. parapsilosis*, and *C. glabrata* cause more and severer infections than the other species (Hobson, 2003; Pappas *et al.*, 2003; Rinaldi *et al.*, 1992) and *C. albicans* (**Figure 1.1**) is the species most commonly isolated from infections (Jackson *et al.*, 2009; Pitarch *et al.*, 2006; Rinaldi *et al.*, 1992).

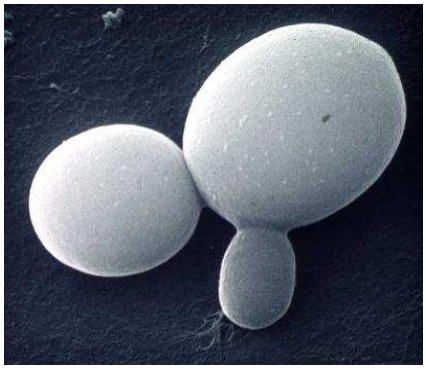


Figure 1.1. Budding *Candida albicans* cells.
(http://botit.botany.wisc.edu/toms_fungi/jan99.html)

C. albicans is often considered a normal commensal inhabitant of human hosts but the distribution of *Candida* species has found in the host varies between geographical regions (Xu & Mitchell, 2003). For example, in healthy populations, compared to North American hosts, Chinese hosts have been shown to have a larger number and more species of *Candida* in the oral cavity (Xu & Mitchell, 2003). Also, *C. albicans* strains predominant in Europe and Western Hemisphere are rarely found in China (Xu & Mitchell, 2003). Although often a commensal, over the last few decades, *C. albicans* has become a major clinical problem and, at present, this opportunistic fungal pathogen is a major health concern (McCullough *et al.*, 1996; Odds, 2010; Pitarch *et al.*, 2006). The reasons for its increase include the growing population of immunocompromised individuals and debilitated patients (Pitarch *et al.*, 2006; Shepherd *et al.*, 1985). Because of the host's weakened immune response, candidiasis often presents in patients as serious systemic disease. In some cases, the infection is difficult to eradicate, and the results are often serious organ damage or even death (Pitarch *et al.*, 2006). In addition to the personal effects of morbidity and mortality, the need for long term medication and hospitalisation caused by candidiasis have high financial costs on both individuals and the public (Hobson, 2003). Developing new diagnostic and therapeutic strategies for this disease is critically important to managing and solving such public health problems (Pitarch *et al.*, 2006). Unfortunately, the problems facing workers in the field of *C. albicans* are many. These include reduced efficacy and serious side effects of recently available medicines, drug resistance of the fungus, and the requirement for rapid, accurate diagnostic procedures (Niimi *et al.*, 1999, Pitarch *et al.*, 2006). Hence, all these reasons have stimulated current scientific research into *C. albicans*, including its biology and virulence mechanisms (Pitarch *et al.*, 2006).

Host factors such as unbalanced microflora or a compromised immune system, are considered as key contributors to *Candida* invasion (Hobson, 2003). However, these factors are believed to be less important than factors arising from the organisms themselves (Pitarch *et al.*, 2006). Opportunistic pathogens need a series of strategies to access host tissues, evade the host immune system and

survive medical treatment (Pitarch *et al.*, 2006). Despite recent research, the virulence factors of *C. albicans* remain ill-defined. Rather than a single virulence factor, a set of features are considered as putatively contributors to the organism's ability to invade the host tissue (Pitarch *et al.*, 2006). These features include its response to environmental variation, adherence to host cells and tissues, biofilm formation and production of tissue-damaging secreted enzymes (Cutler, 1991; VazquezTorres & Balish, 1997; Pitarch *et al.*, 2006). However, as pointed out by Zhang et al (Zhang *et al.*, 2009) the disruption of virtually every *C. albicans* gene has been reported to attenuate virulence in animal models, making it difficult to define the most important virulence attributes.

C. albicans isolates with different genetic backgrounds display different responses to host immune systems (Tavanti *et al.*, 2006; MacCallum *et al.*, 2009; Tavanti *et al.*, 2010). The complex virulence factors and the epidemiology of *C. albicans* indicate that its pathogenicity is complicated and inconsistent (Hobson, 2003). Therefore the study of *C. albicans* cannot focus only at the molecular level of single genes, since whole genome information hiding in phylogeny and population structure may also contribute to understanding the puzzling pathogenicity of this organism (Hunter, 1991). The analysis of particular strains of *C. albicans* can help investigations into pathogenicity and epidemiology of this yeast (Forche *et al.*, 2003; Shepherd *et al.*, 1985).

1.2. Phylogeny and population structure of *Candida albicans*

Research into the phylogeny and population structure of *Candida albicans* was initiated when molecular bioscience was still in its early stages (Hunter, 1991). Based on the frequency of nucleotide polymorphisms, the date for the last common ancestor for *C. albicans* strains is 3-16 million years ago. Different genotypes were broadly distributed geographically, suggesting that DNA sequence variation happened at the same time as early hominid evolution (Lott *et al.*, 2005). It has been suggested that the current *C. albicans* population might be the result of

variation under host immunity as a commensal species (Lott *et al.*, 2005). The phylogeny of *C. albicans* was analysed using multilocus sequence typing (MLST) a world-wide collection of 1,391 isolates with 70% of these isolates were from Europe and USA (Odds et al., 2007). One of the 23 clades identified accounted for over one-third of all the isolates, with the next 2 largest clades accounted for a further third of the isolates (Odds et al., 2007). A range of properties related to virulence of *C. albicans* showed significant positive association with the largest clade (MacCallum et al., 2009). These properties include heterozygosity at the mating type locus, acid phosphatase activity and growth rate at 40℃ (MacCallum et al., 2009).

The moderately repetitive sequence Ca3 in *C. albicans* is a probe that effectually discriminates among strains (Schmid *et al.*, 1990; Schmid *et al.*, 1995b; MacCallum *et al.*, 2009). By Southern blot hybridization based on *Eco*RI-digested Ca3 sequence allowed calculation of similarity (SAB) values between strains and assessment of relatedness by computer-assisted methods (Schmid *et al.*, 1990). Later epidemiological studies based on the same Ca3 fingerprinting method determined the relatedness of *C. albicans* isolates from infected patients (Soll *et al.*, 1991; Schmid *et al.*, 1992; Schmid *et al.*, 1993; Schmid *et al.*, 1995a). In these studies, isolates were collected from different anatomical locations of both healthy individuals and different types of patients at different geographic locations worldwide (Soll *et al.*, 1991; Schmid *et al.*, 1992; Schmid *et al.*, 1993; Schmid *et al.*, 1995a). Many studies showed there was a group of strains more frequently isolated than other groups of strains, and strains in this group were genetically similar (Schmid *et al.*, 1992; Schmid *et al.*, 1993; Schmid *et al.*, 1995a). Strains from this group were found in 30% of healthy individuals and 70% of infection-suffering patients (Schmid *et al.*, 1995a). These studies all suggested the existence of a highly successful group of *C. albicans* with a similar genetic background. Data collected from other studies also implied that there was a group of strains that was highly prevalent. In comparison with other strains, members of this genetically similar group were more resistant to a larger number of chemicals including boric acid, cetrimide, cholorhexidine, 5-fluorocytosine and high

concentrations of sodium chloride, and also adhered to surfaces significantly better (Schmid *et al.*, 1995a). This highly prevalent group was also found to include the most common infections- causing *C. albicans* strains. It was suggested that this particular group of strains may be not only the more successful *C. albicans* in human hosts, but also the most frequent strains causing disease (Schmid *et al.*, 1995a), although this investigation studies a small group of isolates.

To better understand this group of highly prevalent and genetically similar strains and its relatedness to other strains, Ca3 finger-printing was used to characterise and compare 266 infection-causing strains collected from 16 geographical regions of 6 different countries (Schmid *et al.*, 1999). The study showed that 37% of these isolates were genetically homogeneous (cluster A) while the rest of the isolates (63%, cluster B and C, see **Figure 1.2**) were genetically diverse and formed 37 groups of isolates (Schmid *et al.*, 1999). Strains in cluster A were highly prevalent in all geographical regions with a mean prevalence of 41%. This cluster was also highly prevalent in all patient types with a mean prevalence of 43% (Schmid *et al.*, 1999). These results indicated that the strains in cluster A have a general purpose genotype which has enabled a worldwide spread with predominance in all types of *C.albicans* infections affecting all patient types (Schmid *et al.*, 1999). No specificity for body site or geographical location was apparent (Schmid *et al.*, 1999). The genotype of strains in cluster A was considered more pathogenic. It causes candidiasis more often than other clusters and was found more frequently among infection-associated isolates than commensals (Schmid *et al.*, 1999).

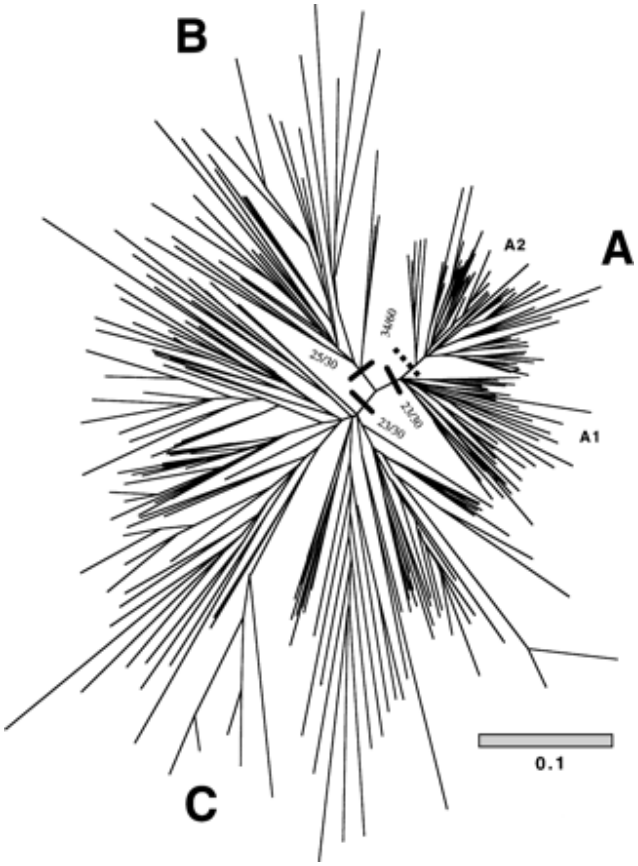


Figure 1.2. An un-rooted phylogenetic tree of 266 *C. albicans* isolates, with 3 main clusters (Schmid *et al.*, 1999). The major cluster A (also called the GPG cluster, including two sub clusters A1 and A2), B and C are labelled. The gray bar in the lower part of the figure indicates genetic distance, based on Ca3 fingerprinting.

1.3. DNA repeats as mutation units in hypermutable contingency genes

1.3.1. Micro-environment evolution and adaptation of microorganisms

In order to be successful opportunistic pathogens in the human host environment, microorganisms need to develop an adaptation system to counteract the immune system and the strong competition with other microorganisms (Moxon *et al.*, 2006). Microorganisms often have large populations and a high rate of replication (Moxon *et al.*, 2006). Thus, within the host environment or micro-environments, evolution occurs. This evolution is specific to the microorganism and its micro-environment; this is distinct from the host's environment and selection pressure that is exerted on the host (Beaumont *et al.*, 2009; Moxon *et al.*, 2006). This makes a dynamic interaction between the host and microorganism. When combined with the effects of the immune system and drug treatments, the host environment provides a stringent test of the microorganisms' adaptive potential (Bayliss *et al.*, 2001; Macia *et al.*, 2005). If the microbial population diverse properties including individuals with properties that confer fitness in the host environment and enable escape from the immune system, an adapted population may result (Bayliss *et al.*, 2001; Beaumont *et al.*, 2009; Deitsch *et al.*, 1997). The key point to understanding this "smart behaviour" lies in the gene and epigenetic mechanisms: mutations are required for evolution.

Selection of phenotypic variants leads the evolution of fitness; "Darwin's great insight" also fits the micro-environment (Moxon *et al.*, 2006). Properties such as adhesion, invasion and antigenic expression are important for the virulenc and pathogenicity of the organism (Deitsch *et al.*, 1997). To survive in the diverse and changing host environment, the organisms may vary their phenotype by regulating gene expression, or varying genes directly – the latter is often more powerful and may allow a rapid phenotypic change (Bayliss *et al.*, 2001; Beaumont *et al.*, 2009; Cutler, 1991; Deitsch *et al.*, 1997). For many microorganisms, some genes have the flexibility to provide a large repertoire of variation; these are called hypermutable contingency genes (Bayliss *et al.*, 2001; Deitsch *et al.*, 1997). An example of this is the *Escherichia coli* gene *ahqC* which encodes a peroxiredoxin

that plays a role in stress response. The gene contains 5 repeats of TCT which shows high frequencies of reversible expansion (Ritz *et al.*, 2001). Another example is the parasite *Plasmodium falciparum* which alters the structural characteristics of the repeats on a cell surface protein. These repeats may assist the parasite avoid the immune response in numerous ways, including exhibiting polymorphism, preventing maturation of the host immune response and reducing antibody binding (Ramasamy, 1998).

1.3.2. Genes containing tandem repeats (TRs)

One type of hypermutable contingency gene is found at loci containing tandem DNA repeats with the repeat size of 2-200 nucleotides (Bayliss *et al.*, 2001; Levdansky *et al.*, 2008). The repeat units can be identical or partially degenerate (Strand *et al.*, 1993). Tandem repeats are known as microsatellites or simple sequence repeats (SSRs) when the repeat units are shorter than 10 nucleotides, or as minisatellites when the repeat units are 10-200 nucleotides long (Levdansky *et al.*, 2008). Variation in the numbers of tandem repeats leads to length polymorphisms (Al-Aidan *et al.*, 2007; Levdansky *et al.*, 2008). Variation in the number of repeats is caused by three different genetic mechanisms: 1) nucleotide strand slippage during replication; 2) genetic recombination; and 3) double strand repair in which the repair is based on the homologous chromosome (Levdansky *et al.*, 2008). The repeat regions can occur in non-coding or coding region of a gene. Repeat regions have been found in the genes of all living microorganisms including viruses, Archaea, Bacteria and eukaryotes (Levdansky *et al.*, 2008; van Belkum, 1999). As tandem repeats can be in a coding sequence or a promoter, an altered numbers of repeats can result either in the variation of an ORF or a change in the activity of a promoter (Bayliss *et al.*, 2001).

Several algorithms have been developed for detecting tandem repeats (Levdansky *et al.*, 2008), including ETANDEM (Rice *et al.*, 2000), mREPS (Kolpakov *et al.*, 2003), and Tandem Repeat Finder (Benson, 1999). Legendre and coworkers

developed the program SERV that also predicts the degree of mutability of a repeat as a so-called VARscore. They demonstrated correlation between VAR scores and the frequency with which mutations in a genetically engineered repeat region arose in yeast (Legendre *et al.*, 2007).

In fungal cells, DNA tandem repeats encoding amino acid repeats are more frequently found in cell wall proteins, including GPI proteins, than in the other classes of proteins (Levdansky *et al.*, 2008). Cell wall GPI proteins (GPI-CWPs) contain a signal peptide sequence and a glycosylphosphatidylinositol (GPI) anchor motif, and are located at the plasma membrane and/or cell wall (Levdansky *et al.*, 2008). For a given repeat-containing DNA sequence, the numbers of repeat units are often found to vary among strains (Levdansky *et al.*, 2008), resulting in amino acid sequence changes. By varying the amino acid sequence of the cell wall protein, the tandem repeat-containing genes potentially have the ability to modulate interactions between the fungal cells and the surrounding environment (Levdansky *et al.*, 2008; Nather & Munro, 2008; Rudd *et al.*, 2009). For example, many cell wall proteins involved in adhesion contain tandem repeats (Levdansky *et al.*, 2008; Rudd *et al.*, 2009). Thus, CPI-CWPS are believed to provide microorganisms with a large and flexible collection of similar but nonidentical surface proteins (Al-Aidan *et al.*, 2007; Braun *et al.*, 2005; Zhang *et al.*, 2003). Genes containing tandem repeats can therefore result in a large collection of genotypes and confer phenotypes that help the microorganism survive a dynamic host environment (Al-Aidan *et al.*, 2007; Moxon *et al.*, 2006).

Although the tandem repeats may allow faster adaptation to a variable environment than non-repeat regions, genome analysis has shown that the production and maintenance of short repeats tends to be regulated but the mechanisms for this remain to be elucidated (Astolfi *et al*., 2003).

1.3.3. Genes containing tandem repeats (TRs) in *Candida albicans*

Compared with another ascomycete yeast *Saccharomyces cerevisiae*, *C. albicans* has a much higher proportion of TR sequences in its protein-coding sequences; this has been considered a probable cause for its pathogenic potential in humans (Braun *et al.*, 2005; Levdansky *et al.*, 2008). Only a few TR-containing genes of *C. albicans* have been studied, including *EAP1*, *PIR1*, *CEK1*, *HYR1*, *HYR2*, *HWP1*, *RLM1* and the *ALS* (agglutinin like sequence) family (Lott et al., 1999; Hoyer, 2001; Staab et al., 2004; Oh et al., 2005; Sumita et al., 2005; Zhao et al., 2007; Li & Palecek, 2008). Further research on TR sequences will contribute to a better understanding of the *C. albicans* genome (Braun *et al.*, 2005). The *C. albicans* genome harbours many short sequence repeats (Al-Aidan *et al.*, 2007), and repeat-dependent modulation of gene expression could play a crucial role in its virulence. Scanning of the C*. albicans* genome for repeat units of sizes of two to five nucleotides detected 1,940 repeats in all the ORFs (Braun *et al.*, 2005; Jones *et al.*, 2004).

TRs have been identified in genes with a variety of functions in *C. albicans*. For example, the ALS proteins (*ALS*, agglutinin-like sequence) belong to the gene family that encodes adhesion proteins of *C. albicans*. They show considerable repeat region-based allelic diversity that give functionally relevant variability (Zhang *et al* 2003, Rauceo *et al* 2006). The gene *ALS7* has been shown to contain TRs (Zhang *et al.*, 2003; Chaffin, 2008). Gene *HWP1*, involved in hyphal growth and biofilm formation, contains repeats encoding Gln/Pro and Ser/Thr (Padovan *et al.*, 2009). Al-Aidan and coworkers studied the TRs W2, W4, W5, W7 and W10 (Al-Aidan *et al.*, 2007). They showed that W2 is in the gene encoding Efg1, a protein that regulates hypha formation (Stoldt *et al.*, 1997) and controls white-opaque phenotypic switching (Sonneborn *et al.*, 1999; Srikantha *et al.*, 2000). The W4 TRs are part of the gene encoding hyphal wall protein 1 (Hwp1) which works as a germ tube-specific surface protein (Staab *et al.*, 1999), and the W5 TRs are in the gene for an integrin-like protein which is crucial for filamentation. The W7 TRs lie in the *ZNF1* gene which regulates filamentous growth under special conditions (Field *et*

al., 1996). The W10 TRs lie in the *MNT1* gene which encodes the mannosyltransferase, and which plays a crucial role in adherence and virulence (Al-Aidan *et al.*, 2007; Munro *et al.*, 2005; Rouabhia *et al.*, 2005). However, studies which compared the five 5 repeat loci W2-W10 in colonized and invasively infected patients, only repeat region W2 showed variation (Al-Aidan *et al.*, 2007).

ALS proteins play an important role in the adherence of *C. albicans*. It has been claimed that the *ALS* gene family member *ALS7* is a hypermutable contingency gene because sixty alleles, differing in repeat regions, and 49 genotypes were detected in 66 strains (Zhang *et al.*, 2003). However some of the same authors recently presented evidence that suggests that the rapid generation of new alleles in many repeat-containing ORFs may be of limited benefit or even detrimental (Zhang *et al.*, 2010). Zhang *et al.*, (2010) observed that GPG-specific alleles exist not only in *ALS7* but also in other *C. albicans* repeat-containing ORFs. They also experimentally verified a high *in vitro* mutation rate in the repeat-containing ORF, *PNG2*, which has a GPG-specific allele combination. On the basis of these observations they argued that GPG-specific alleles of *PNG2* and other repeat-containing ORFs can only be retained in the presence of the observed mutation rate if the genetic background of GPG strains selects against survival of clones with new alleles (Zhang *et al.*, 2010). Thus, the constant generation of new alleles, commonly assumed to be the main benefit and reason for the existence of hypermutable contingency genes (Moxon *et al.*, 2006), may be only of limited benefit or even detrimental in the case of many *C. albicans* repeat-containing ORFs. Hence, *ALS7* and other *C. albicans* repeat-containing genes, while hypermutable, may indeed not function as contingency genes.

1. 4. Cell wall proteins of *Candida albicans*

1. 4. 1. Molecular organization of the cell wall of *Candida albicans*

The cell wall of *Candida albicans* contains 3 principal kinds of components: glucan, chitin and mannoproteins (Klis *et al.*, 2001; Netea *et al.*, 2008; Ruiz-Herrera *et al.*, 2006). The cell wall has 2 layers; the inner layer consists of a combination of glucan and chitin (microfibrillar polymers) and is the structural element of the wall, providing the cell strong physical properties including mechanical strength (Bates *et al.*, 2007; Chaffin *et al.*, 1998; Chaffin, 2008). Glucan is the main constituent which accounts for 60% of the wall dry weight (Klis *et al.*, 2001). Chitin only accounts for about 1-2% of wall dry weight but it is an important component of the cell wall. For example, it plays a major role in bud formation (Chaffin *et al.*, 1998; Klis *et al.*, 2001; Netea *et al.*, 2008). The outer layer is dominated by mannoproteins. These are the major part of the cell wall matrix on the surface, accounting for 35-40% of wall dry weight. Mannoproteins play a crucial role in the host-*C. albicans* interaction (Klis *et al.*, 2001; Klis *et al.*, 2002).

Cell wall proteins can be divided into 3 different groups according to their method of extraction (Bates *et al.*, 2007). Group 1 can be extracted with detergents such as SDS, suggesting a loose association with the cell wall. Group 2 can be extracted using reducing agents such as DL-Dithiothreitol. The third group can only be separated from the cell wall when glucan and chitin are degraded by enzymes, such as glucanases and chitinase. Only proteins in the third group are covalently bound to glucan and chitin. Glycosylphosphatidylinositol-linked cell wall proteins (GPI-CWPs) belong to this group. Some GPI-CWPs are expressed in amounts that depend on growth conditions (Bates *et al.*, 2007).

GPI proteins are the major class of cell surface mannoproteins (Klis *et al.*, 2001) They have some structures in common which include: 1) an N-terminal signal sequence; 2) a conserved domain; 3) a variable-length Ser/Thr-rich region, and 4) a C-terminal GPI-anchor (Bates *et al.*, 2007; De Groot *et al.*, 2003; De Groot *et al.*, 2004; Garcera *et al.*, 2003; Klis *et al.*, 2001; Klis *et al.*, 2002). *In silico* genome

analysis suggests that there are potentially 115 GPI proteins in *C. albicans*, which is twice the number in the ascomycete yeast *Saccharomyces cerevisiae* (Plaine *et al.*, 2008). By 2007, 12 GPI-CWPs had been identified by proteomic analysis from yeast-form *C. albicans* cells (Bates *et al.*, 2007; De Groot *et al.*, 2004). The 9 GPI-CWPs for which function was identified were: Cht2, Crh11, Pga4, Phr1, Scw1, the ALS family (Als1 and Als4), Sod4 and Rbt5 (De Groot *et al.*, 2004; Weissman & Kornitzer, 2004). The functions included carbohydrate-active enzymes, adhesion proteins, superoxide dismutase and haemoglobin-iron utilization (Bates *et al.*, 2007; De Groot *et al.*, 2004; Weissman & Kornitzer, 2004). The functions of the other 3 proteins - Ywp1, Ecm33, and Ssr1 - remain unknown (Bates *et al.*, 2007; De Groot *et al.*, 2004; Ruiz-Herrera *et al.*, 2006).

1. 4.2. Roles of cell wall proteins of *Candida albicans*

Cell wall proteins (CWPs) are crucial for the survival of *C. albicans* (Netea *et al.*, 2006; Sosinska *et al.*, 2008) and there are five mains roles of CWPs (Ruiz-Herrera *et al.*, 2006). Some CWPs have an enzymatic role and are involved in the degradation of large nutritional molecules or degradation of other cell wall polymers. A second role of CWPs is in cell interaction with the host, including biochemical systems for attachment and adherence. The ability to adhere to the host and other substrates is crucial for *C. albicans* colonisation and is mediated by a group of cell surface proteins called adhesins (Linder & Gustafsson, 2008). A third role of CWPs is in antigenicity and antigenic variability of the cell wall. This has been considered to be one of the many putative virulence factors (Cutler, 1991; Netea *et al.*, 2008). CWPs have also been shown to have a role in pathogenicity as they are involved in the establishment of the pathogen for invasion of the host (Fradin *et al.*, 2005; Nather & Munro, 2008; Netea *et al.*, 2008). Another major role of CWPs is in wall structure and morphogenesis, such as responses to cell wall stress and the yeast and hyphal growth phase (Klis *et al.*, 2001; Ruiz-Herrera *et al.*, 2006; Tsarfaty *et al.*, 2000; Whiteway & Bachewich, 2007).

Biofilm formation also involves cell surface interaction mediated by CWPs (Chaffin, 2008; Douglas, 2003). Biofilm formation of *C. albicans* is involved in serious systemic diseases such as heart disease (Douglas, 2003). *C. albicans* cells grow in both yeast and hyphal forms which combine to make a complicated bilayer structure, Furthermore *C. albicans* biofilms are resistant to the currently used antifungal drugs (Chandra *et al.*, 2001; Douglas, 2003).

1. 5. *SSR1* gene and Ssr1 protein

1.5.1. Genomic study of *SSR1* gene

C. albicans was one of the first eukaryotic human pathogens selected for sequencing and the diploid genome sequence of *C. albicans* strain SC5314 was published by Stanford University as Assembly 19 in 2004 (Jones *et al.*, 2004). SC5314 was selected for sequencing because it was virulent, widespread, commonly used in laboratory analysis and had a standard diploid karyotype (Jones *et al.*, 2004). it has eight chromosomes (chromosomes 1-7, R) with a total haploid genome size of 14,851 kb, that encodes 11,615 putative coding sequences (Braun *et al.*, 2005). This sequencing and its publication was a landmark in *C. albicans* research and provided a platform for new genetic studies (Odds *et al.*, 2004). Many research studies have been based on the genome sequence of SC5314. These include a study of *SSR1* gene (orf19.7030) on chromosome 7, which encodes the protein Ssr1p, (Braun *et al.*, 2005). The *SSR1* gene in SC5314 encodes a polypeptide of 234 amino acids with a molecular mass of 22,553 Da (for primary sequence see appendix III). The most recent assembly (Assembly 21) confirmed the genome locus for *SSR1* (van het Hoog *et al.*, 2007).

The structure of Ssr1p has been determined, based on an earlier Candida genome assembly (Garcera *et al.*, 2003). It contains four main domains; a hydrophobic N-terminal signal sequences, a neutral region in the central part of the protein which includes a cysteine rich domain and a Ser/Thr rich domain (38%), and a hydrophobic GPI-attachment site in the hydrophobic C-terminal domain (**Figure**

1.3). These domains indicated that Ssr1p is potentially a GPI cell wall protein. This protein also contains the amino acids Ala, Glu, Gly, Leu, Lys and Val (Garcera *et al.*, 2003)

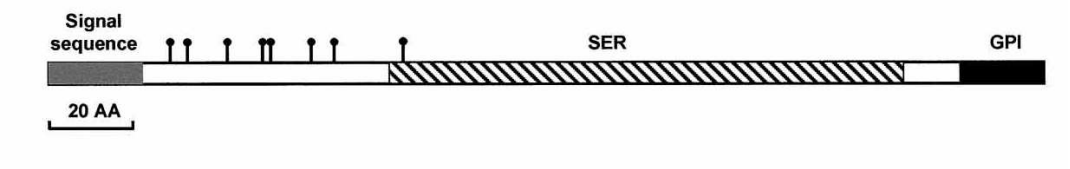


Figure 1.3. Four main domains in *SSR1* (Garcera *et al.*, 2003). From left to right: the hydrophobic N-terminal signal peptides, followed by a neutral region including a cysteine rich domain (8 staked out = 8 cysteines) and a serine rich domain (hatched), and a hydrophobic GPI-attachment site (Garcera *et al.*, 2003).

1.5.2. Ssr1p is a repeat-containing cell wall protein of *Candida albicans* with a role in wall stability

The function of Ssr1p remains unclear and the information about Ssr1p is limited. Northern blot analysis found the *SSR1* transcript was present in both yeast and mycelial forms of *C. albicans* cells. This result showed that the expression of Ssr1p did not vary in the two morphological states (Garcera *et al*. 2003).

Previous research shows Ssr1p might contribute to cell wall structure (Garcera *et al.*, 2003; Richard *et al.*, 2002). The function of Ssr1p in *C.albicans* was tested using a *SSR1* null mutant constructed by disrupting the *SSR1* gene. Analysis of the mutant phenotype showed decreased strength of the cell wall (Garcera *et al.*, 2003). Apart from this, the *C. albicans SSR1* null mutant cells behaved normally during growth with normal growth rate, morphology and germ tube formation. Thus the Ssr1 protein is not essential for *C. albicans* survival (Garcera *et al*. 2003). The *SSR1* null mutant cell wall was more sensitive than the wild type, and that a lack of

Ssr1 protein leads to a defective cell wall (Garcera *et al.*, 2003). Disruption and over-expression of the *SSR1* gene increased the sensitivity of *C. albicans* cells to a range of reagents, including calcofluor white, Congo red, zymolyase and caspofungin (Garcera *et al.*, 2003; Garcera *et al.*, 2005; Plaine *et al.*, 2008). These molecules affect fungal growth by causing cell lysis or inhibiting chitin and glucan syntheses (Roncero & Duran, 1985). Ssr1p must be kept at an appropriated concentration for a normal cell wall and deviation in the concentration of Ssr1p change cell wall stability (Garcera *et al.*, 2003). In a rich growth environment without disturbance by harsh chemicals, Ssr1 protein is not an absolutely necessary component for *C. albicans* survival (Garcera *et al.*, 2003).

The *C. albicans SSR1* null mutant upregulated by at least 1.4-fold seven other genes including *SPR54*, *IPF29* and *PTR3* (Garcera *et al.*, 2005). *SRP54* encodes a signal recognition particle subunit; *IPF29* encodes a zinc finger protein and *PTR3* encodes a transcriptional regulator. The *SSR1* null mutant also leads to down regulation of 27 other genes by at least 0.7-fold, including gene encoding proteins IPF6318 (a β-glucosidase) and SOU1 (a sorbitol utilization protein) (Garcera *et al.*, 2005). These studies were done with a transcription profile of 6039 *C. albicans* genes (Garcera *et al.*, 2005). They suggest that *SSR1* and Ssr1p may have multiple roles in cell wall function.

1. 6. Hypothesis and aim

There is currently an absence of evidence that tests for the existence of typical hypermutable contingency genes in *C. albicans* i.e. genes which have the flexibility to provide a large repertoire of variation for adaptation. However, few of the large number of TR-containing ORFs in this organism have been investigated (see **1.3.3**). Since hypermutable contingency genes mainly encode cell surface proteins, a TR- containing gene known to encode a cell surface protein was selected to assess if it may be a contingency gene and to learn more about the biological significance of TRs in *C. albicans* surface proteins.

This project investigates the hypothesis that the *SSR1* gene of *C. albicans* is a hypermutable contingency gene.

The objectives of the projects are to answer the following questions.

- Does *SSR1* show the high level allelic diversity expected of contingency genes?
- Does *SSR1* mutate at the high rates expected of contingency genes?
- Is allelic variation in the repeat regions of *SSr1* biologically significant and is this used for the transient, short-term adaptation expected of a contingency gene?

CHAPTER TWO – MATERIALS AND METHODS

For suppliers of chemicals, solutions and equipments, see appendix I.

All methods were performed as described in *Current Protocols in Molecular Biology* (Ausubel *et al.,* 2008), unless stated otherwise.

2.1. Location of tandem repeats

The software program matrix plot (MacVector software program) was used to plot base pair against base pair for the *SSR1* gene sequence from the genome sequence database strain SC5314 (assembly 19), obtained from http://www. candidagenome.org/cgi-bin/ locus.pl?locus=ssr1.

The software program SERV (Legendre *et al.*, 2007), available online at http://hulsweb1.cgr.harvard.edu/S-ERV/, was also used to detect the repeat regions of the *SSR1* gene, based on the sequence above.

2.2. Biological materials

2.2.1 *Candida albicans* strains

The origin of strains used in this study is shown in table 2.1.

Table 2.1. *Candida albicans* strains used in this study.

Isolate	Infection-causing/ Commensal	Group*	Country of origin	Site of isolation	Reference
Au1	infection-causing	GPGA1	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
Au19	infection-causing	GPGA2	New Zealand	Urine	Schmid *et al*, 1995a.
Au27	infection-causing	GPGA1	New Zealand	Sputum	Schmid *et al*, 1995a.
Au39	infection-causing	GPGA1	New Zealand	skin/wounds	Schmid *et al*, 1995a.
Au90	infection-causing	GPGA2	New Zealand	skin/wounds	Schmid *et al*, 1995a.
CH14	infection-causing	GPGA1	New Zealand	skin/wounds	Schmid *et al*, 1995a.
CH35	infection-causing	GPGA2	New Zealand	urine	Schmid *et al*, 1995a.
CH42	infection-causing	GPGA2	New Zealand	skin/wounds	Schmid *et al*, 1995a.
CHOB5	infection-causing	GPGA2	New Zealand	skin/wounds	Schmid *et al*, 1995a.
CLB42	infection-causing	GPGA2	Colombia	vagina/vulva	Schmid *et al*, 1999.
CLB53	infection-causing	GPGA2	Colombia	skin/wounds	Schmid *et al*, 1999.
cour-c	infection-causing	GPGA2	USA	respiratory/oral	Schmid *et al*, 1999.
FJ23	infection-causing	GPGA1	Fiji	urine	Schmid *et al*, 1999.
FJ26	infection-causing	GPGA1	Fiji	skin/wounds	Schmid *et al*, 1999.
FJ9	infection-causing	GPGA1	Fiji	respiratory/oral	Schmid *et al*, 1999.
HUN122	infection-causing	GPGA1	Great Britain	sterile	Schmid *et al*, 1999.
HUN127	infection-causing	GPGA2	Great Britain	sterile	Schmid *et al*, 1999.
HUN93	infection-causing	GPGA2	Great Britain	sterile	Schmid *et al*, 1999.
HUN95	infection-causing	GPGA2	Great Britain	sterile	Schmid *et al*, 1999.
HUN96	infection-causing	GPGA2	Great Britain	sterile	Schmid *et al*, 1999.
jam-2c	infection-causing	GPGA2	USA	anal	Schmid *et al*, 1995a.

Isolate	Infection-causing/ Commensal	Group*	Country of origin	Site of isolation	Reference
ko-2c	infection-causing	GPGA1	USA	stool	Schmid *et al*, 1995a.
OD8826	infection-causing	GPGA2	Great Britain	respiratory/oral	Schmid *et al*, 1999.
OD8911	infection-causing	GPGA2	Great Britain	respiratory/oral	Schmid *et al*, 1999.
OD8916	infection-causing	GPGA2	Great Britain	respiratory/oral	Schmid *et al*, 1999.
OD9014	infection-causing	GPGA1	Great Britain	respiratory/oral	Schmid *et al*, 1999.
RIHO10	infection-causing	GPGA2	USA	Sterile	Schmid *et al*, 1999.
RIHO13	infection-causing	GPGA2	USA	sterile	Schmid *et al*, 1999.
RIHO16	infection-causing	GPGA2	USA	sterile	Schmid *et al*, 1999.
RIHO9	infection-causing	GPGA2	USA	sterile	Schmid *et al*, 1999.
Rolo	infection-causing	GPGA1	USA	respiratory/oral	Schmid *et al*, 1999.
sim-c	infection-causing	GPGA1	USA	respiratory/oral	Schmid *et al*, 1999.
var1.10vag	infection-causing	GPGA2	USA	vagina/vulva	Schmid *et al*, 1999.
var1.1vag	infection-causing	GPGA1	USA	vagina/vulva	Schmid *et al*, 1999.
var1.3vag	infection-causing	GPGA1	USA	vagina/vulva	Schmid *et al*, 1999.
var1.4vag	infection-causing	GPGA2	USA	vagina/vulva	Schmid *et al*, 1999.
var1.8vag	infection-causing	GPGA2	USA	vagina/vulva	Schmid *et al*, 1999.
W132	infection-causing	GPGA2	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
W134	infection-causing	GPGA1	New Zealand	sputum	Schmid *et al*, 1995a.
W26	infection-causing	GPGA2	New Zealand	anal	Schmid *et al*, 1995a.
W3	infection-causing	GPGA2	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
W43	infection-causing	GPGA2	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
W59	infection-causing	GPGA1	New Zealand	sputum	Schmid *et al*, 1995a.
W68	infection-causing	GPGA2	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
YASM73	infection-causing	GPGA2	Malaysia	respiratory/oral	Schmid *et al*, 1999.
YASU568	infection-causing	GPGA2	Malaysia	urine	Schmid *et al*, 1999.
YASU649	infection-causing	GPGA2	Malaysia	urine	Schmid *et al*, 1999.
YASU751	infection-causing	GPGA1	Malaysia	urine	Schmid *et al*, 1999.
Au11	infection-causing	ngpg_C	New Zealand	skin	Schmid *et al*, 1995a.
Au134	infection-causing	ngpg_C	New Zealand	respiratory/oral	Schmid *et al*, 1995a.

Isolate	Infection-causing/ Commensal	Group*	Country of origin	Site of isolation	Reference
Au33	infection-causing	ngpg_C	New Zealand	skin/wounds	Schmid *et al*, 1995a.
Au36	infection-causing	ngpg_C	New Zealand	catheter	Schmid *et al*, 1995a.
CH20	infection-causing	ngpg_C	New Zealand	vagina/vulva	Schmid *et al*, 1995a.
CH3	infection-causing	ngpg_C	New Zealand	urine	Schmid *et al*, 1995a.
CH41.1	infection-causing	ngpg_C	New Zealand	urine	Schmid *et al*, 1995a.
CH9	infection-causing	ngpg_B	New Zealand	vagina/vulva	Schmid *et al*, 1995a.
CLB44	infection-causing	ngpg_C	Colombia	skin/wounds	Schmid *et al*, 1999.
CLB45	infection-causing	ngpg_C	Colombia	skin/wounds	Schmid *et al*, 1999.
CLB49	infection-causing	ngpg_C	Colombia	respiratory/oral	Schmid *et al*, 1999.
FJ12	infection-causing	ngpg_C	Fiji	respiratory/oral	Schmid *et al*, 1999.
FJ27	infection-causing	ngpg_C	Fiji	catheter	Schmid *et al*, 1999.
FJ3	infection-causing	ngpg_C	Fiji	respiratory/oral	Schmid *et al*, 1999.
gaymc-c	infection-causing	ngpg_C	USA	respiratory/oral	Schmid *et al*, 1999.
HUN123	infection-causing	ngpg_B	Great Britain	sterile	Schmid *et al*, 1999.
HUN61	infection-causing	ngpg_B	Great Britain	respiratory/oral	Schmid *et al*, 1999.
HUN64	infection-causing	ngpg_C	Great Britain	skin/wounds	Schmid *et al*, 1999.
HUN66	infection-causing	ngpg_C	Great Britain	skin/wounds	Schmid *et al*, 1999.
HUN68	infection-causing	ngpg_C	Great Britain	respiratory/oral	Schmid *et al*, 1999.
HUN91	infection-causing	ngpg_C	Great Britain	sterile	Schmid *et al*, 1999.
HUN92	infection-causing	ngpg_B	Great Britain	sterile	Schmid *et al*, 1999.
OD8824	infection-causing	ngpg_C	Great Britain	respiratory/oral	Schmid *et al*, 1999.
OTG1	infection-causing	ngpg_C	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
OTG10	infection-causing	ngpg_B	New Zealand	anal	Schmid *et al*, 1995a.
OTG18	infection-causing	ngpg_B	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
OTG2	infection-causing	ngpg_C	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
OTG4	infection-causing	ngpg_C	New Zealand	vagina/vulva	Schmid *et al*, 1995a.
OTG6	infection-causing	ngpg_C	New Zealand	skin/wounds	Schmid *et al*, 1995a.
RIHO2	infection-causing	ngpg_C	USA	no information available	Schmid *et al*, 1999.

Isolate	Infection-causing/ Commensal	Group*	Country of origin	Site of isolation	Reference
RIHO5	infection-causing	ngpg_C	USA	no information available	Schmid *et al*, 1999.
sw-17c	infection-causing	ngpg_C	USA	sterile	Schmid *et al*, 1999.
var1.5vag	infection-causing	ngpg_B	USA	vagina/vulva	Schmid *et al*, 1999.
var1.7vul	infection-causing	ngpg_B	USA	vagina/vulva	Schmid *et al*, 1999.
W137	infection-causing	ngpg_B	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
W142	infection-causing	ngpg_C	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
W17	infection-causing	ngpg_C	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
W53	infection-causing	ngpg_C	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
W55	infection-causing	ngpg_C	New Zealand	anal	Schmid *et al*, 1995a.
YASM1	infection-causing	ngpg_C	Malaysia	sterile	Schmid *et al*, 1999.
YASM42	infection-causing	ngpg_C	Malaysia	respiratory/oral	Schmid *et al*, 1999.
YASU123	infection-causing	ngpg_C	Malaysia	urine	Schmid *et al*, 1999.
YASU363	infection-causing	ngpg_C	Malaysia	urine	Schmid *et al*, 1999.
YASU63	infection-causing	ngpg_C	Malaysia	urine	Schmid *et al*, 1999.
YASU709	infection-causing	ngpg_C	Malaysia	urine	Schmid *et al*, 1999.
cfr2.10vul	Commensal	GPG	USA	vulva	Schmid *et al*, 1993.
cfr2.1vul	Commensal	GPG	USA	vulva	Schmid *et al*, 1993.
cfr2.3vul	Commensal	GPG	USA	vulva	Schmid *et al*, 1993.
cfr2.4vul	Commensal	GPG	USA	vulva	Schmid *et al*, 1993.
cfr2.8vag	Commensal	GPG	USA	vagina	Schmid *et al*, 1993.
cfr2.9vag	Commensal	GPG	USA	vagina	Schmid *et al*, 1993.
COD21	Commensal	GPG	UK	respiratory/oral	Schmid *et al*, 1992.
cpr2.2fec	Commensal	GPG	USA	stool	Schmid *et al*, 1993.
HMHc1	Commensal	GPG	USA	respiratory/oral	Schmid *et al*, 1990.
HMHc2	Commensal	GPG	USA	respiratory/oral	Schmid *et al*, 1990.
HMHc4	Commensal	GPG	USA	respiratory/oral	Schmid *et al*, 1990.
HMHc5	Commensal	GPG	USA	respiratory/oral	Schmid *et al*, 1990.
HMHc6	Commensal	GPG	USA	respiratory/oral	Schmid *et al*, 1990.

Isolate	Infection-causing/ Commensal	Group*	Country of origin	Site of isolation	Reference
HMHc9	Commensal	GPG	USA	respiratory/oral	Schmid *et al*, 1990.
hp10bt	Commensal	GPG	USA	respiratory/oral	Schmid *et al*, 1990.
hp11vw	Commensal	GPG	USA	vagina	Schmid *et al*, 1990.
hp12bt	Commensal	GPG	USA	respiratory/oral	Schmid *et al*, 1990.
hp13vu	Commensal	GPG	USA	vulva	Schmid *et al*, 1990.
hp13vw	Commensal	GPG	USA	vagina	Schmid *et al*, 1990.
hp2bt	Commensal	GPG	USA	respiratory/oral	Schmid *et al*, 1990.
hp31an	Commensal	GPG	USA	anal	Schmid *et al*, 1990.
hp31vw	Commensal	GPG	USA	vagina	Schmid *et al*, 1990.
hp33vw	Commensal	GPG	USA	vagina	Schmid *et al*, 1990.
hp42bt	Commensal	GPG	USA	respiratory/oral	Schmid *et al*, 1990.
hp42vp	Commensal	GPG	USA	vagina	Schmid *et al*, 1990.
hp55bt	Commensal	GPG	USA	respiratory/oral	Schmid *et al*, 1990.
W104	Commensal	GPG	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
W105	Commensal	GPG	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
W106	Commensal	GPG	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
W107	Commensal	GPG	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
W108	Commensal	GPG	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
W109	Commensal	GPG	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
W111	Commensal	GPG	New Zealand	respiratory/oral	Schmid *et al*, 1995a.
SC5314	infection-causing	GPG	laboratory strain		

* For infection-causing strains the Ca3-based classification into two GPG sub-branches A1, and A2, and two major non-GPG (ngpg) branches, B and C is included (Schmid *et al*., 1999).

2.2.2. Other biological materials

Other biological materials used were *E.coli* DH5α as competent cells, and pLUG Multi TA-cloning vector (iNtRON, Daejoen city, Korea).

DNA Ladders:

1kb+ DNA ladder, 40ng/µl

(500µl) 400µl distilled water, 80µl loading dye, 20µl 1kb+DNA ladder.

100bp DNA ladder, 83ng/µl

2.3. Media, buffers and solutions

2.3.1. Media

LB agar with ampicillin (per litre) 15g agar; ampicillin final concentration 100µg/ml

LB medium (per litre) 10g Bacto®-tryptone; 5g Bacto®-yeast extract; 5g NaCl

YPD agar (per litre) YPD medium; 15g agar

YPD medium (per litre) 20g Glucose; 20g Bacto®-tryptone; 10g Bacto®-yeast extract

2.3.2. Buffers and solutions

Ammonium persulfate (APS) (10%) 0.1g/1ml store at 4℃ < 1 month

Ampicillin solution 4mg/ml, filter sterilized

$CaCl_2$ solution (500ml) 3.3g $CaCl_2$; 93.5 ml glycerol; 1.5g PIPES. adjust pH to 7 with NaOH 8M.

(600µl) 400µl distilled water, 100µl loading dye, 100µl 100bpDNA ladder.

EDTA 0.5M 186g NaEDTA·2 H_2O in 700ml H_2O; adjust pH to 8.0 with 10M NaOH (~50ml); add distilled water to 1 litre.

Gel loading buffer (10X) 20% Ficoll 400; 0.1M disodium EDTA, pH 8; 1% SDS; 0.25% Bromophenol blue; 0.25% Xylene cyanol

Glycerol solution 65% glycerol, 0.1M Mg SO4, 0.025 M Tris, pH 8

IPTG stock solution (0.1M) 1.2g IPTG add water to 50ml final volume. Filter-sterilize and store at 4℃.

Polyacrylamide gel (8%) (5 ml) mQ H_2O 3.7ml; long ranger acrylamide-bis 0.8ml; 10×TBE 0.5ml; TEMED 3µl; 10%APS, 30µl

X-Gal (2ml) 100mg 5-bromo-4-chloro-3-indolyl-ß-D-galactoside dissolved in 2ml N,N'-dimethyl-formamide. Cover with aluminum foil and store at -20℃.

TBE (10×) (per litre) 108g Tris; 55g Boric acid; 0.5M EDTA, pH 8.0, 40ml

2.4. Growth and maintenance of cultures

All media were sterilized for 20 min at 121℃ prior to use. MilliQ water was used in all preparations. Solidified media contained agar (1.5%) and as allowed to cool to approximately 50℃ before antibiotics were added an d plates poured. In a laminar flow hood, approximately 25 ml YPD agar was poured into petri dishes and left to set. Agar plates were kept in a plastic bag to avoid agar drying out and stored at room temperature for no longer than one week before use.

2.4.1. Growth conditions of *Candida albicans* cells on agar plates

After inoculation of *C.albicans* cells, YPD agar plates were incubated at 37℃ overnight, and then stored at 4℃ for up to two mon ths.

2.4.2. Preparation of *C. albicans* stocks on YPD slants

Approximately 15 ml melted YPD agar was poured into universal bottles and sterilized. The caps were tightened and the agar left to harden in a slanted position at room temperature. After inoculation of *C.albicans* cells, YPD agar slants were incubated at 37℃ overnight then stored at 4℃ for up to six months.

2.4.3. Glycerol stock for *Candida albicans*

In a 2 ml plastic cryovial with a screw cap, a 0.5 ml overnight culture of *C.albicans* in YPD was added to 0.5 ml glycerol solution (65% glycerol, 0.1M $MgSO_4$, 0.025 M Tris, pH8). The cap was tightened, the contents mixed and the tube stored at 80℃.

2.4.4. Growth conditions of *Escherichia coli* on LB agar plates containing ampicillin (for blue-white selection)

Sterilised LB agar medium was cooled to 50℃ and filter sterilised ampicillin added to a final concentration of 100µg/ml. Approximately 25 ml of LB-ampicillin agar was poured into petri dishes in a laminar flow hood and the agar left to harden. 100µl of 100mM IPTG (Isopropyl β-D-1-thiogalactopyranoside) and 20 µl of 50mg/ml X-Gal (bromo-chloro-indolyl-galactopyranoside) were spread over the surface of the plates, and allowed to absorb for 30 minutes at 37℃ before use. After inoculation with *E. coli* cells, the LB plates were incubated at 30℃ overnight and stored at 4℃ for up to two weeks.

2.5. Preparation of template for PCR

2.5.1. Colony PCR

In most cases cells from a colony were used directly to provide template DNA. A colony was picked with a sterile 10-µl white pipette tip, and mixed well into a 20µl PCR reaction mix (see **Table 2.2** for details of reaction mixes). For these templates the first step of the PCR protocol included extra time to release DNA (see **2.6.3.1** for details of PCR reaction conditions).

2.5.2. Extraction of DNA (boiled supernatants)

Approximately 5 small colonies (or 1 big colony) were picked with a sterile white pipette tip and mixed with 15µl sterilized water in an eppendorf tube. The tubes were boiled for 4 min, cooled on ice immediately, and then centrifuged at 13,000×g for 1 min. 1µl of the resulting boiled supernatant (bsn) was used as DNA template in each PCR reaction.

2.6. PCR

2.6.1. PCR reagents

The suppliers and quantities of reagents used in DNA amplification by PCR are given in **Table 2.2**.

Table 2.2. Details of reagents used in PCR reactions

Reagent	Manufacturer	Working Concentration	Volume/reaction (colony PCR)	Volume/reaction (bsn PCR)
Taq polymerase	Roche	5units/µl	0.2 µl	0.2µl
Taq Buffer	Roche	-	2µl	2.5µl
dNTPs	Roche	100pmoles/ µl	2µl	2.5µl
Q buffer	Qiagen	-	4µl	-
$MgCl_2$	Invitrogen	25nmoles/µl	-	0.5µl
Primer (forward)	Invitrogen/ Gibco BRL	10pmoles/µl	1µl	1µl
Primer (reverse)	Invitrogen/ Gibco BRL	10pmoles/µl	1µl	1µl
Water	-	-	9µl	16.5µl
DNA template	-	-	Single colony	1 µl
Total	-	-	~20 µl	~25 µl

2.6.2. Primers used in experiment

Primers were designed based on the *SSR1* gene sequence of strain SC5314 using software Vector NTI (**Figure 2.1**). Primers SSR1-6 and SSR1-7 were designed to amplify the whole *SSR1* ORF, giving a length of 783 bp. After cloning in the Multi TA-cloning vector, the inserts were amplified using vector-borne sequences M13F and M13R, giving a product of 965 bp. Primers F1 and B2 were designed to amplify the two repeat regions in a single product, with a length of 465 bp. Primers F1 and SSR1-1 were designed to amplify repeat region 1, giving a 205 bp product. Primers SSR1-2 and SSR1-3 were designed to amplify repeat region 2, giving a product length 277 bp. The sequences of each primer is given in Appendix II.

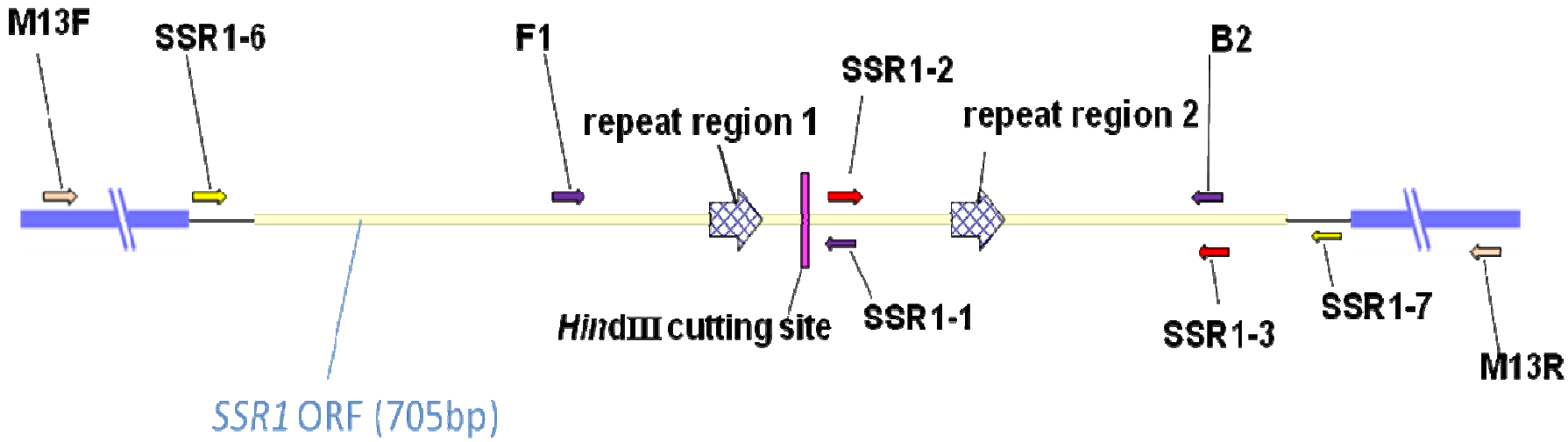


Figure 2.1. The relative annealing positions of the primers in the *SSR1* gene of strain SC5314. The blue segments are plasmid sequences.

2.6.3. PCR reaction conditions

PCR reaction conditions were designed according the type of DNA templates used, the primer characteristics and the product lengths. The extension time was based on the expected product lengths (based on *SSR1* of strain SC5314). In PCR reactions for genotyping, step 5 (72℃) was set to 20 min to give a sharp peak in genotyping.

2.6.3.1. Colony PCR of *C. albicans*: amplification of *SSR1* with primers SSR1-6 and SSR1-7

Product length: 783 bp

Step 1. 96℃ 5min
Step 2. 94℃ 45sec
Step 3. 55℃ 40sec } 30 cycles from step 2 to 4
Step 4. 72℃ 50sec
Step 5. 72℃ 10min
Step 6. 4℃ hold

2.6.3.2. Colony PCR of *E. coli*: amplification of *SSR1* inserts with primers M13F and M13R

Product length: 965 bp

Step 1. 94℃ 3min
Step 2. 94℃ 30sec
Step 3. 63℃ 30sec } 30 cycles from step 2 to 4
Step 4. 72℃ 1min
Step 5. 72℃ 5min
Step 6. 4℃ hold

2.6.3.3. Colony PCR of *C. albicans*: amplification of repeat regions 1 and 2 with primers F1 and B2

Product length: 465 bp

Step 1. 94℃ 3min
Step 2. 94℃ 40sec
Step 3. 53℃ 40sec } 30 cycles from step 2 to 4
Step 4. 72℃ 50sec
Step 5. 72℃ 20min
Step 6. 4℃ hold

2.6.3.4. PCR of extracted DNA (bsn): amplification of repeat region 1 with primers F1 and SSR1-1; amplification of repeat region 2 with primers SSR1-2 & SSR1-3

Product length: 205 bp; 277 bp

Step 1. 94℃ 2min
Step 2. 94℃ 30sec
Step 3. 55℃ 30sec } 30 cycles from step 2 to 4
Step 4. 72℃ 30sec
Step 5. 72℃ 20min
Step 6. 4℃ hold

2.6.3.5. PCR of extracted DNA (bsn): amplification of repeat region 1 and 2 with primers F1 and SSR1-3

Product length: 459 bp

Step 1. 94℃ 2min
Step 2. 94℃ 45sec
Step 3. 55℃ 40sec } 30 cycles from step 2 to 4
Step 4. 72℃ 30sec
Step 5. 72℃ 20min
Step 6. 4℃ hold

2.7. Endonuclease digestion of PCR product

The two repeat regions in *SSR1* amplification products were separated by digestion with *Hin*dIII. The *Hin*dIII restriction site was found between the two repeat regions with expected fragment lengths of 162 bp and 303 bp from a PCR

obtained product using primers F1 and B2. The PCR product was digested at 37℃ overnight for digestion (**Table 2.3**).

Table 2.3. Reagents for restriction of F1/B2 PCR products with *Hin*dIII

Component	Volume (µl)
*Hin*dIII (10U/µl)	1
PCR product	5-10 (depending on amount of PCR product)
Distilled water	Adjusted
10× Buffer	2
Total volume	20

2.8. Generation of competent *Escherichia coli* cells

A single colony of *E. coli* (strain DH5α) was inoculated into 50 ml LB medium in a 500 ml flask and incubated overnight at 37℃ with m oderate shaking. 4 ml of this culture was inoculated into 400ml LB medium in a 2L flask and incubated, at 37℃ with moderate shaking, to an OD_{590} = 0.375. The culture was dispensed into eight 50-ml sterile polypropylene tubes and the tubes left on ice for 10 min. After centrifugation at 3000 × g for 7min at 4℃, the sup ernatant was discarded and the pellet suspended in 10 ml ice-cold $CaCl_2$ solution. The tubes were centrifuged at 2500×g for 5min at 4℃, the supernatant was discard ed, the pellet resuspended in 10 ml ice-cold $CaCl_2$ solution and the tubes set on ice for 10 min. After centrifugation at 2500 × g for 5 min at 4℃, the su pernatant was discarded, the pellet resuspended in 2 ml ice cold $CaCl_2$ solution, and the tubes then left on ice at 4℃ for 24 hours. Finally, 250µl sample of cells was dispensed into pre-chilled sterile eppendorf tubes and stored immediately at -80℃.

2.9. Gel electrophoresis

2.9.1. Agarose gel electrophoresis

2.9.1.1. Gel preparation

The relatively small sizes of DNA fragments required the use of high percentage (1.5-2%) agarose gels. Between 1.5 and 2g of agarose were added to 100ml 1×TBE buffer, the suspension heated in a microwave oven with regular mixing until dissolved, cooled to 50℃, and poured into a gel casting unit.

2.9.1.2. Gel loading and running

After the gel had set, the gel comb was removed and the gel was placed in the electrophoresis tank. 1×TBE buffer was added until the gel surface was covered by the buffer. 8-10 µl of each sample was mixed with 1µl of 10× gel loading buffer, and loaded into a well. For each agarose gel, ~8µl 1Kb+ DNA ladder was added as a size marker. The voltage was set to 85V ~95V for an appropriate time period, depending on the gel size and equipment used.

2.9.1.3. Gel staining and illumination

The gel electrophoresis was stopped when the loading buffer dye had migrated a distance which indicated the appropriate distribution of DNA fragments. The gel was stained by covering with ethidium bromide (0.5µg/ml in water) for 20min, and destained with water for 10min. The gel was photographed using a Gel Doc UV transilluminator and Quantity One software programme.

2.9.2. Polyacrylamide gel electrophoresis (PAGE)

A Mini Protean® 3 system (Bio-Rad) was used in all PAGE experiments.

2.9.2.1. Gel preparation

Gel plates were cleaned thoroughly with 5% SDS and their surfaces were brushed gently to remove dust or any residual gel, then rinsed with 70% ethanol and left to dry. The gel casting apparatus was assembled, making sure the glass plates would not leak.

An 8% polyacrylamide gel was used to resolve the DNA fragments. Each polyacrylamide gel was made with 5ml of polyacrylamide mixture. For each gel, the components used were as shown below in **Table 2.4.**

Table 2.4. Components used for each 8% polyacrylamide gel

Component	Volume
Long ranger™ (polyacrylamide)	0.8ml
10×TBE	0.5ml
mQ H_2O	3.7ml
10% APS	30µl
TEMED	3µl
Total volume	~ 5ml

The H_2O, 10×TBE and Long ranger™ were mixed in a 50ml-beaker. After 5 mins, TEMED and 10% APS were added (10% APS was stored up to 1 month at 4 ºC before use) and the beaker was swirled to mix the reagents well. Blue tips (1-ml) were used to place the gel mixture between the gel plates and the 10-teeth comb was then inserted. Care was taken to avoid any air bubbles. The gel was

polymerised for 2 h before use.

2.9.2.2. Gel loading and running

After polymerization was complete, the combs were removed and the gels placed in the electrophoresis tank. 1× TBE buffer was poured in the space between the two sets of glass plates to cover the wells, and then samples were mixed with 1 µl of 10× gel loading buffer (the same gel loading buffer as used for agarose gels) and added to the wells. For each polyacrylamide gel, ~6µl 1×100bp DNA ladder was loaded as a size marker. 1×TBE buffer was then poured into the space around the gels to half way up the container. Electrophoresis was carried out at 65V for 2.2 hours.

2.9.2.3. Gel staining and illumination

The gel electrophoresis was stopped when the loading buffer dye had migrated a distance which indicated the appropriate distribution of DNA fragments (e.g. down to the bottom of the gel). The polyacrylamide gel was carefully removed from the glass plate and stained by covering with ethidium bromide (0.5µg/ml in water) for 20min, then destained with water for 10min. The gel was placed on a UV transilluminator and photographed with a Gel Doc system and Quality One software programme.

2.10. Genotyping

For genotyping, fluorescent dye 5'-end labeled primers F1-FAM™ (blue) and SSR1-2- HEX™ (green) were used with primers SSR1-1 and SSR1-3, respectively, to give different coloured products for each repeat region of *SSR1* (**Fig 2.2**).The PCR conditions were as described for non-fluorescent labeled primers (**2.6.3.4** and **2.6.3.5**). Separate reactions were carried out for each primer pair. Samples (8-10µl)

of reaction mixtures were run on agarose gels to determine if the PCR had been successful and the dilution required for genotyping (usually 3-5 fold, depending on the amount of PCR products shown on the gel). 1 µl of a mixture containing both diluted PCR products was amplified through the genotyping for microsatellite analysis service at the Alan Wilson Centre Genome Services (AWCGS), Massey University using a 3730 Genetic Analyzer. The size standard 500 LIZc (Applied Biosystems) provided the internal standard for fragment sizing. This size standard accurately determine the size of fragments between 50 and 500 bp. Genotyping results were analyzed with the software *Peak Scanner* (Applied Biosystems).

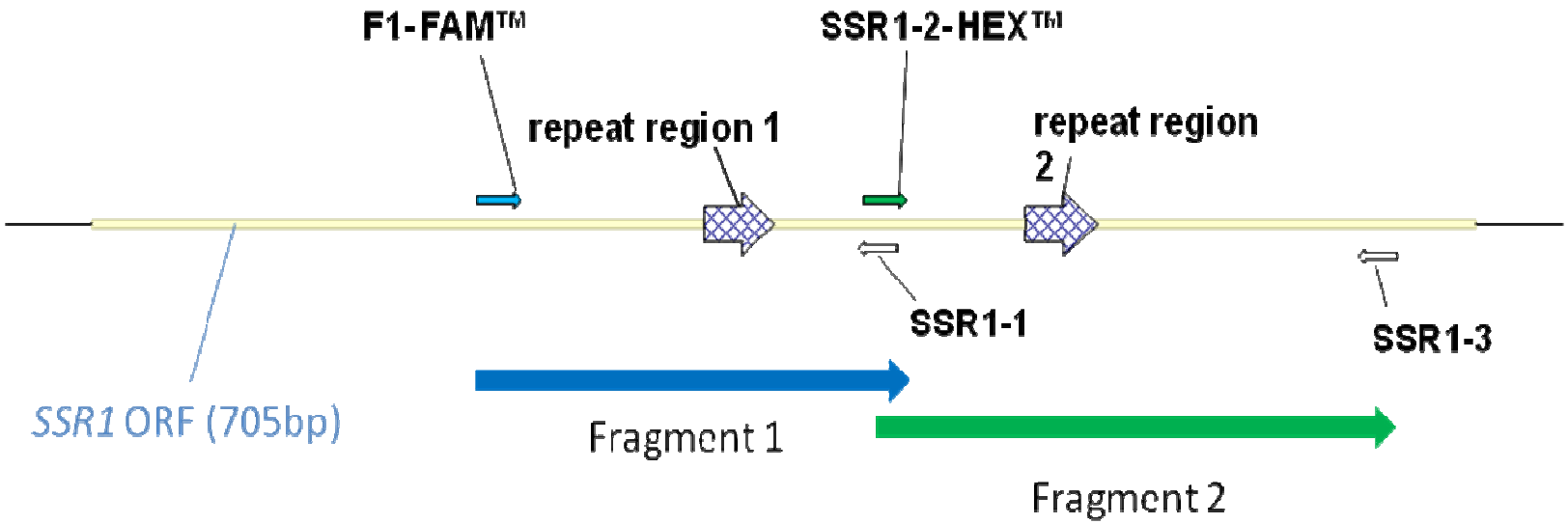


Figure 2.2. The relative annealing positions of primers used for genotyping and their PCR products for *SSR1* in strain SC5314.

2.11. Cloning of *SSR1* for DNA sequencing

As *C. albicans* is diploid, cloning was necessary for DNA sequencing. Primers for sequencing were designed to amplify the whole gene (SSR1-6 and SSR1-7) (**Figure 2.1**). The plasmid used for cloning was pLUG. After ligation with purified PCR products of *SSR1* (**Figure 2.3**), recombinant pLUG was cloned by transformation into competent cells of *E. coli* DH5α. The *E. coli* cells containing *SSR1* inserts were detected by blue-white selection as white colonies, the *SSR1* inserts amplified by colony PCR and analysed by DNA sequencing at the AWCGS.

2.11.1. PCR product purification

PCR products were purified with High Pure PCR Product Purification Kit (Roche), used according to the manufacturer's instructions with the exception that the final was in 30 µl.

2.11.2. Ligation

2.11.2.1. Determination of DNA concentration

The DNA concentration of PCR products was determined using Quant-iT™ dsDNA BR Assay Kit and QuBit® fluorometer, according to the manufacturer's instructions. If the concentration was not high enough, the sample was air dried at 37℃ and sterile H_2O was added to give the required concentration of DNA.

2.11.2.2. TA-Cloning

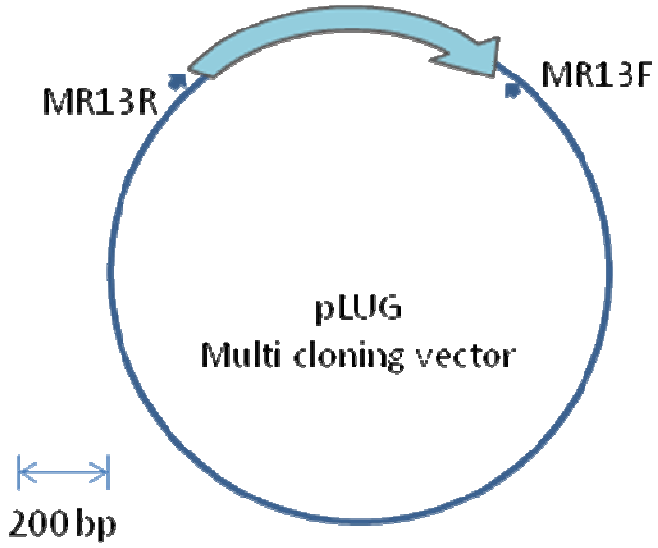


Figure 2.3. pLUG®-Multi TA-Cloning Vector and vector-borne primers M13F and M13R, amplifying the insert (the light blue arrow).

Ligation was achieved with pLUG®-Multi TA-Cloning Vector Kit (iNtRON) (**Figure 2.3**). For the ligation reaction, the molar ratio of PCR product to vector had to be optimized. A 5-10 fold molar excess of PCR product/vector was recommended in the protocol, the pLUG®-Multi TA-cloning vector is 2276 bp and the protocol suggested adding 1 µl (50 ng) of the vector. A 5-fold excess of insert required 87 ng of insert (800 bp).

The following equation calculates the amount of PCR product required for 50ng vector:

ng PCR product required
= [50 ng× PCR product size (bp) ×molar ratio]/[vector size (bp)]
In this ligation reaction, ng PCR product required
= [50 ng× 800 bp ×5]/ [2276 (bp)]
=87ng

Ligation reaction mixtures were prepared according to the scheme in **Table 2.5**. Ligation reaction mixtures were mixed by pipetting and the tubes were incubated overnight at 4℃.

Table 2.5. Reagent volumes for ligation reactions

Component	Volume (µl)
TA-cloning vector	1
PCR product	1-4 (depends on concentration of PCR products)
Distilled H_2O	Adjusted
10×Ligation Buffer	1
T4 DNA ligase	1
Total volume	10

2.11.3. Transformation

For each transformation, 100 µl of competent cells was added to an eppendorf tube containing 10 µl of ligation mixture, mixed by pipetting and incubated on ice for 10 min. The solution was then heat-shocked by incubating the tubes in a 42℃ water bath for 2 min, and then put on ice immediately. After 2 min, 1 ml LB medium was added and the tubes were incubated at 37℃ with shaking for 1 hour.

2.11.4. Blue-white selection

LB-Amp plates, with 100 µl of 100 mM IPTG and 20 µl of 50 mg /µl X-gal on the agar surface (see **2.4.4**), were dried for 30 min at 37℃. 100 µl of transfo rmed cells (**2.11.3**) was spread on the surface of each plate, and the plates were then incubated at 37℃ for 18 hours. Ligation and transf ormation are shown by white colonies (containing inserts) and blue colonies (no inserts) growing on the agar

plates. The contents of white colonies were checked by amplifying the insert region with primers M13F and M13R, using a colony PCR reaction (see **2.5.1** and **2.6.3.2** for details). 8 µl of products were separated on 2% agarose gels and the colonies which contained the right size of inserts identified. The rest of the PCR products (12 µl) were purified and sequenced.

2.11.5. DNA Sequencing

PCR products (12 µl) shown to contain the right size of insert (**2.11.4**) were purified in an eppendorf tube by adding 80 µl of butanol and centrifuging 13,000 × g for 2 min. The supernatant was discarded, the tubes were centrifuged again at 13,000 × g for 2 min, and the remaining supernatant was discarded by pipetting. The eppendorf tube was air dried at 37℃, 15 µl sterile H_2O was added dissolve the pellet and then the concentration of the DNA were estimated using QuBit® fluorometer (**2.11.2.1**).

The concentration of DNA required for the sequencing by AWCGS is 2 ng/100 bp/ 15µl. The length of the DNA fragments sequenced was about 1,000 bp, therefore the final amount of DNA needed to be about 20 ng. The DNA sequencing reactions were made as shown in **Table 2.6** and sent to AWCGS for analysis on the ABI3730 DNA Analyzer.

Table 2.6. Reagent volumes for sequencing

Component	Volume (µl)
DNA (20 ng-23 ng)	2-10 µl
1 pmol/µl primer (SSR1-6)	3.2 µl
distilled H_2O	Adjusted
Total volume	15 µl

2.12. Subculturing of *Candida albicans* for 300 generations

The strains HUN91 and RIHO30, from glycerol stocks, were grown on YPD agar plates at 37℃ overnight. A single colony from the agar plate was used to inoculate 2 ml of YPD medium in a test tube and this culture was incubated for 37℃ for 12 hours. Ten µl culture sample were transferred into a new set of sterile 2 ml YPD media and incubated for 12 or 24 hours. This process was repeated until cells had grown for 300 generations. The number of transfers required was calculated based on the increase in optical density at 600 nm measured with a spectrophotometer (NOVA TECH) in the first 8 cultures. It was concluded that in 12 hours, HUN91 have grown 7.38 generations and RIHO30 have grown 7.46 generations; in 24 hours, HUN91 have grown 7.95 generations and RIHO30 have grown 7.92 generations.

After the strains had been subcultured for 300 generations, 200 µl of culture from the last tube were taken, spread over the surfaces of a YPD agar plate and the plates were incubated at 37℃ overnight. Most of the cells on the agar surface were then transferred into a 10-ml tube containing 5 ml of distilled H_2O, mixed, and the culture was diluted to 1 cell/µl according to a microscope cell counting result. 100 µl (~100 cells) of the diluted culture were spread on an agar plate, with 3 plates for each strain. These plates were incubated at 37℃ overnight. 60 single colonies were chosen from these agar plates and allele sizes determined by genotyping (**2.10**). Mutation rates were calculated as follows:

Mutation rates per cell division =
[Number of mutated cells] / [number of tested colonies × generations]

Mutation rates per allele=
[Number of mutated cells] / [number of tested colonies × generations×2]

2.13. Subculturing of *Candida albicans* rat samples

Samples from rats inoculated 4 weeks earlier with pairs of isolates at the University of Otago (**Table 2.7**) were plated. The number of C. albicans colonies growing on the plates from each sample is shown in Table 2.7. A suspension was made from all C. albicans colonies on the plates from each sample, and a loopful of this suspension was used to inoculate a slant. The growth pattern on the slant suggests that the inoculums contained > 300 viable cells. Slants were sent to Massey, and all of the cells from the slant were mixed with 1 ml of sterilized H_2O and followed by vortex-mixing. Subsequently 1 µl of the suspension was used as the DNA template in a PCR reaction amplified with primers F1/SSR1-1 and SSR1-2/SSR1-3 (**2.6.3.4**).

Also, suspensions of 60 colonies of each sample were made from a culture of the original 1 ml mixture and the *SSR1* allele's colony PCR (**2.5.1**) amplified with primers F1/SSR1-1 and SSR1-2/SSR1-3 (**2.6.3.4**), analysed by PAGE and compared with the alleles of the original strains.

Table 2.7. Number of *C. albicans* cells from rat host.

Rat host	Strains	Number of cells from rat host
A1	YsU63, W43	320
A2	YsU63, W43	180
A3	YsU63, W43	320
B2	Au90, Hun97	48
B3	Au90, Hun97	88
C1	Au90, FJ11	130
C2	Au90, FJ11	3300
C3	Au90, FJ11	200

CHAPTER THREE – RESULTS

3.1. Choice of the *SSR1* gene for this investigation

This project is part of the Schmid laboratory's attempts to understand the biological significance of repeat-containing ORFs in *C. albicans* and to determine if some of them function as contingency genes in short-term adaptation. Since contingency genes often encode cell surface proteins (**1.3**), *SSR1* seemed a good candidate for such investigations. As described in the introduction section, Ssr1p is a GPI protein, the protein class that form the major part of the cell wall mannoproteins (Klis *et al.*, 2001; Chaffin, 2008; Plaine *et al.*, 2008), and the *SSR1* gene (orf19.7030) of the data base strain SC5314 contains short tandem repeats (Braun *et al.*, 2005), as shown below in **Table 3.1**.

Table 3.1. Short tandem repeats in *SSR1* determined by Braun et al.

Name of repeat*	Positions (bp)	DNA sequence
BR 1	272-295	**CATCTG AATCTG AATCTG AATCGG**
BR 2	311-345	**CTTCTG CTTCCG CTTCTG CTTCTG CTTCTT CATCT**
BR 3	476-495	**CTTCTG CTGCTG CTTCTG CT**

*(Braun *et al.*, 2005)

3.2. *SSR1* has variable repeats that may generate proteins which differ in the number and arrangement of amino acid repeats

Braun and coworkers' detected repeat regions in *SSR1* in a genome-wide screen aimed at determining the frequency of repeat regions (Braun *et al.*, 2005) rather than during a focused investigation of *SSR1*. The first goal of this project was

therefore to analyze *SSR1* repeats more closely and to determine if they varied in length between strains.

Two significant repeat regions in the ORF, containing BR2 and BR3, were identified using a matrix plot (**Figure 3.1 A**, solid line ovals). The positions of the two areas highlighted with dashed line ovals indicated that the two repeat sequences were similar to each other. BR1 was not identified as a repeat-containing region in this matrix plot (**Figure 3.1 A**), even though it contains a region of 4 repeat units of 6 base pairs each (mainly AATCTG) (**Figure 3.1 B**). A short repeat region not detected by Braun et al. (2005) (square in **Figure 3.1 A**) was identified in this analysis as a repeat region between (bp 260 - bp 272 with the sequence CTCAAT CTTCATC).

An additional way used to identify repeat regions is the SERV system (Legendre *et al.*, 2007). In the SERV system, the VARscore gives an estimation of repeat mutability, which means a larger VARscore indicated a higher repeat variability of the repeat region (Legendre *et al.*, 2007). Repeat regions SERV2 and SERV3 have higher VARscores than the SERV1 and SERV4 regions (**Table 3.2** and **Figure 3.1 C**), therefore the variability between strains of region 2 and 3 are expected to be higher.

Table 3.2. Short tandem repeats in *SSR1* determined with SERV software program

Name of repeats	Position from the origin of ORF(bp)	DNA sequences	VARscore
SERV1	273-293	**ATCTGA ATCTGA ATCTGA ATC**	-0.62
SERV2	311-348	**CTTCTG CTTCCG CTTCTG CTTCTG CTTCTT CATCTG CT**	**-0.07**
SERV3	476-512	**CTTCTG CTGCTG CTTCTG CTTCTG CTTCTG CTTCTG C**	**0.10**
SERV4	658-690	**GTTGGTCTT GGTGCTTTA GTTGGTTTA GTTGGT**	-0.78

Alignment of the repeat regions detected by the matrix plot, Braun *et al*. (2005) and SERV, showed the presence of overlapping and non-overlapping repeat regions (**Figure 3.1**). Combining all this information, for on strain SC5314, suggest regions 313-348 and 478-513 were most likely to be repeat regions. This was confirmed by visual inspection of the sequence. In both repeat regions the repeat units are mainly TCTGCT, encoding amino acid repeats SerAla (SA, **Figure 3.2**).

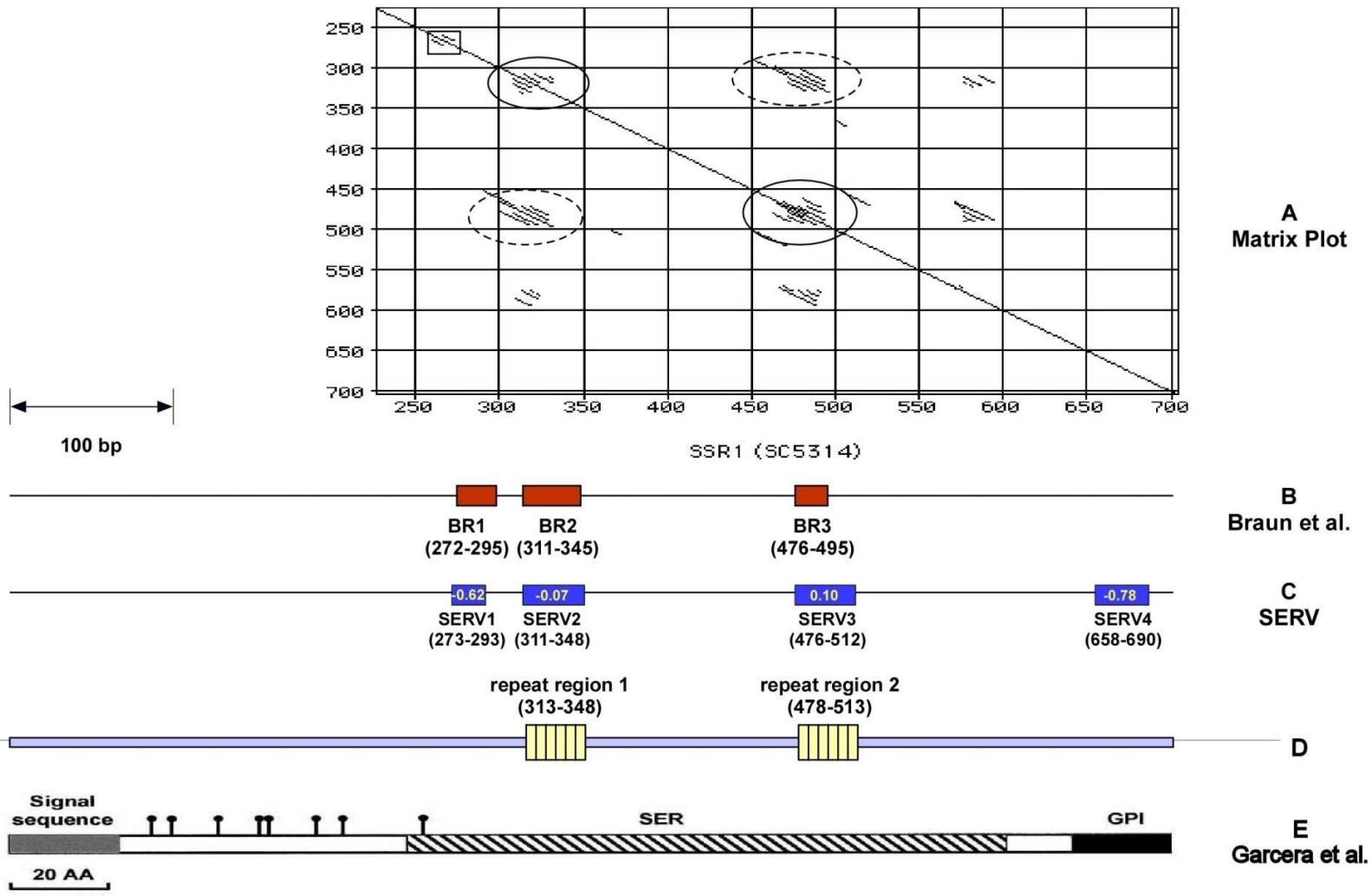

250
300
350
400
450
500
550
600
650
700
250 300 350 400 450 500 550 600 650 700
SSR1 (SC5314)
A
Matrix Plot
100 bp
BR1
(272-295)
BR2
(311-345)
BR3
(476-495)
B
Braun et al.
-0.62
-0.07
0.10
-0.78
SERV1
(273-293)
SERV2
(311-348)
SERV3
(476-512)
SERV4
(658-690)
C
SERV
repeat region 1
(313-348)
repeat region 2
(478-513)
D
Signal
sequence
SER
GPI
20 AA
E
Garcera et al.

Figure 3.1. Gene structure and repeat regions identified in *SSR1* by Matrix Plot (A), Braun *et al*. (2005) (B) and SERV(C). VARscores of regions determined by SERV system are labeled in (C). The repeat regions 1 and 2 confirmed by visual inspection of sequence are labeled in (D). Ssr1 ploy peptide structural characteristics shown in (E), including the two hydrophobic domains (grey and black bars), the Ser/Thr-rich domain (hatched bar) and Cys rich residues (8 stake outs) (Garcera *et al.*, 2003).

```
  1  ATGGCTTCAT TTTTAAAGAT TTCTACTTTG ATTGCAATTG TTTCTACTTT
 51  ACAAACCACT TTAGCTGCTC CACCAGCTTG TTTATTAGCT TGTGTTGCTA
101  AAGTTGAAAA AGGTTCTAAA TGTTCAGGTT TAAATGATTT AAGTTGTATT
151  TGTACTACTA AGAATTCTGA CGTTGAAAAA TGTTTGAAAG AGATTTGTCC
201  AAATGGTGAT GCTGATACTG CCATTTCTGC TTTTAAGAGT TCTTGTTCTG
251  GTTATAGTTC TCAATCTTCA TCATCTGAAT CTGAATCTGA ATCGGCTTCA
301  AGTGAAGAAT CTTCTGCTTC CGCTTCTGCT TCTGCTTCTT CATCTGCTGG
351  TAAATCTTCA AATGTTGAAG CTTCTACTAC TAAAGAATCT AGTTCAGCCA
401  AGGCTTCTTC TTCCGCTGCC GGCTCATCTG AAGCTGTTTC TTCTGCTACT
451  GAAACTGCTT CTACTGAAGA ATCATCTTCT GCTGCTGCTT CTGCTTCTGC
501  TTCTGCTTCT GCCACTAAAG AATCTTCTTC TGAAGCTGCT AGTTCCACTT
551  CATCTACCTT AAAAGAATCT AAAACTTCTA CTACTGCTGC TGCTTCATCT
601  TCTGAATCTA CTACTGCAAC TGGTGTTTTG ACTCAATCTG AAGGTTCTGC
651  TGCTAAGGTT GGTCTTGGTG CTTTAGTTGG TTTAGTTGGT GCTGTTTTAT
701  TGTAA

  1  MASFLKISTL IAIVSTLQTT LAAPPACLLA CVAKVEKGSK CSGLNDLSCI
 51  CTTKNSDVEK CLKEICPNGD ADTAISAFKS SCSGYSSQSS SSESESESAS
101  SEESSASASA SASSSAGKSS NVEASTTKES SSAKASSSAA GSSEAVSSAT
151  ETASTEESSS AAASASASAS ATKESSSEAA SSTSSTLKES KTSTTAAASS
201  SESTTATGVL TQSEGSAAKV GLGALVGLVG AVLL
```

Figure 3.2. Sequence of *SSR1* gene and translated sequence showing repeat units in strain SC5314. http://www.candidagenome.org/.

3.3. Detection of variation in *SSR1* repeat regions strains by DNA sequencing

To determine if either of the repeat regions varied in the number or arrangement of repeats between strains, sequencing of the entire ORF in a small selection of 14 strains (both GPG and non-GPG strains) was performed.

For the DNA sequencing, primers SSR1-6 and SSR1-7 were designed to amplify the whole ORF of *SSR1* (**Figure 2.1**). *C. albicans* is a diploid fungus and the PCR analysis using these primers showed that the *SSR1* genes were heterozygous in some strains (**Figure 3.3**). Therefore, these strains could not be sequenced directly and the two alleles needed to be separated by cloning. After cloning of purified PCR products of *SSR1* in pLUG, plasmids were used to transform competent cells of *Escherichia coli DH*5α (see section **2.11**). The *E. coli* colonies containing inserts were selected and the inserts identified by colony PCR using primers M13F and M13B (**2.11**; **Figure 2.3**). The expected size of the PCR products was about 1000bp. The PCR result showed that the recombinant plasmids contained inserts of the expected size (**Figure 3.4**). After purification of the PCR products, DNA samples were sequenced using the primer SSR1-6 (see **2.11.5**).

Alignment of sequencing results confirmed that repeat regions 1 and 2 (containing BR2 and BR3) were variable regions, that differed in the number of repeat units, while the sequences in other regions of *SSR1* (including BR1, SERV1 and SERV4) were constant among strains, with the exception of a few point mutations (**Figure 3.5**). The sequencing data was unambiguous, with the exception of strain HUN91, where the sequencing results for the last three repeat units were not clear. Because DNA sequencing results confirmed that repeat regions 1 and 2 were highly variable among strains (**Figure 3.1** and **3.5**), the investigation of variable repeats in this project focused on these regions only.

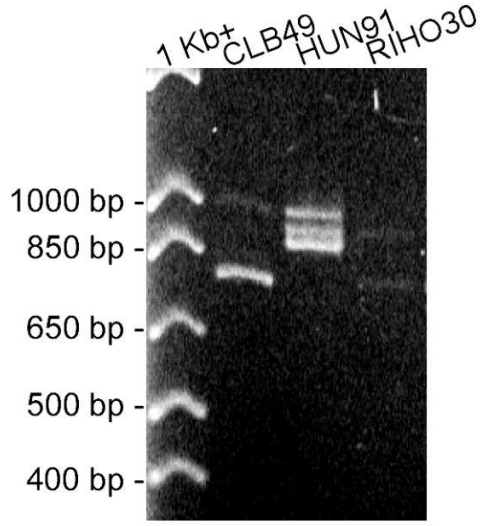


Figure 3.3. Example of *SSR1* heterozygosity in strains. The whole *SSR1* ORF was amplified using primers SSR1-6 and SSR1-7 and PCR products were resolved on a 1.5% agarose gel. Note the 2 bands present in CLB 49 and RIHO30. HUN91 has three bands in this photo; one of them is probably an artificial band from the PCR reaction (Bradley & Hillis, 1997). A 1kb+ ladder was included as a size standard (leftmost lane). Numbers on the left are the size of some of the 1 kb + ladder bands. Gel was run at 80 V for 2 hours (Hoefer HE33 mini horizontal agarose unit).

The alignments showed that variation in number of tandem repeats in repeat regions 1 and 2 (mainly TCTGCT, encoding amino acids Ser and Ala) caused different allele sizes (**Figure 3.5**). Because these 6 nucleotides repeats encode 2 amino acid repeats, the different alleles will have different lengths of amino acid repeats, rather than introducing frame shifts that change the amino acid sequence or introduce premature stop codons. Sequencing results also showed that differences in alleles can be caused by variation in sequence of similar repeat units, with not all repeat units encoding the same amino acids (**Figure 3.5**). However, there are some conserved arrangements of the units in region 2 (the second and last units), with the exception of HUN 91.

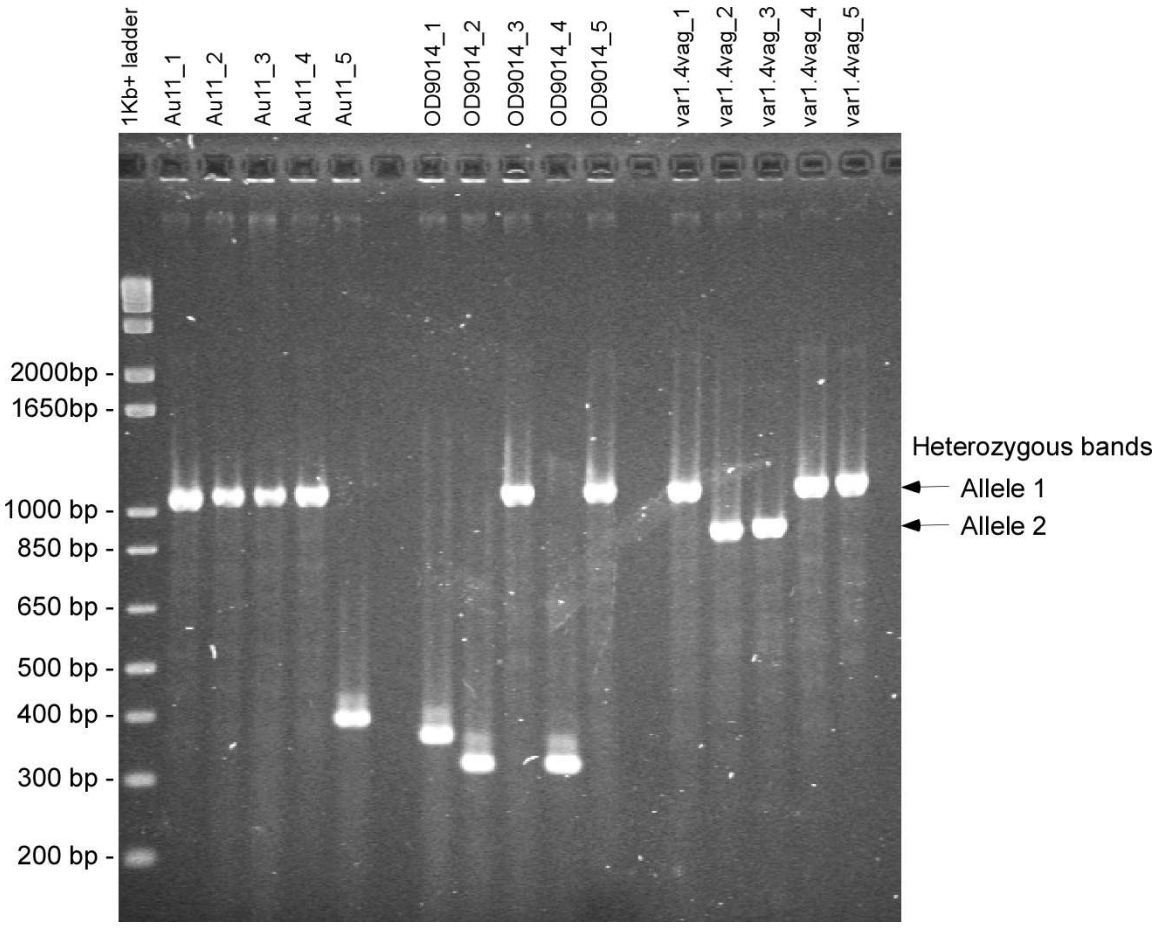


Figure 3.4. Examples of heterozygous alleles separated by TA cloning. After blue-white selection, 5 white *E. coli* colonies were chosen from each strain and the *SSR1* inserts were amplified by colony PCR. This photo showed strain var1.4vag was clearly heterozygous, while strain Au11 might also be heterozygous. The bands of sizes about 1000bp and 850bp are likely to be the cloned *SSR1* products. Other smaller fragments (400bp and 300bp) imply that the white colonies also contained no inserts or other inserts with different lengths (2% agarose gel). The gel was run at 90 V for 2.5 hours (Owl D2 wide-gel unit).

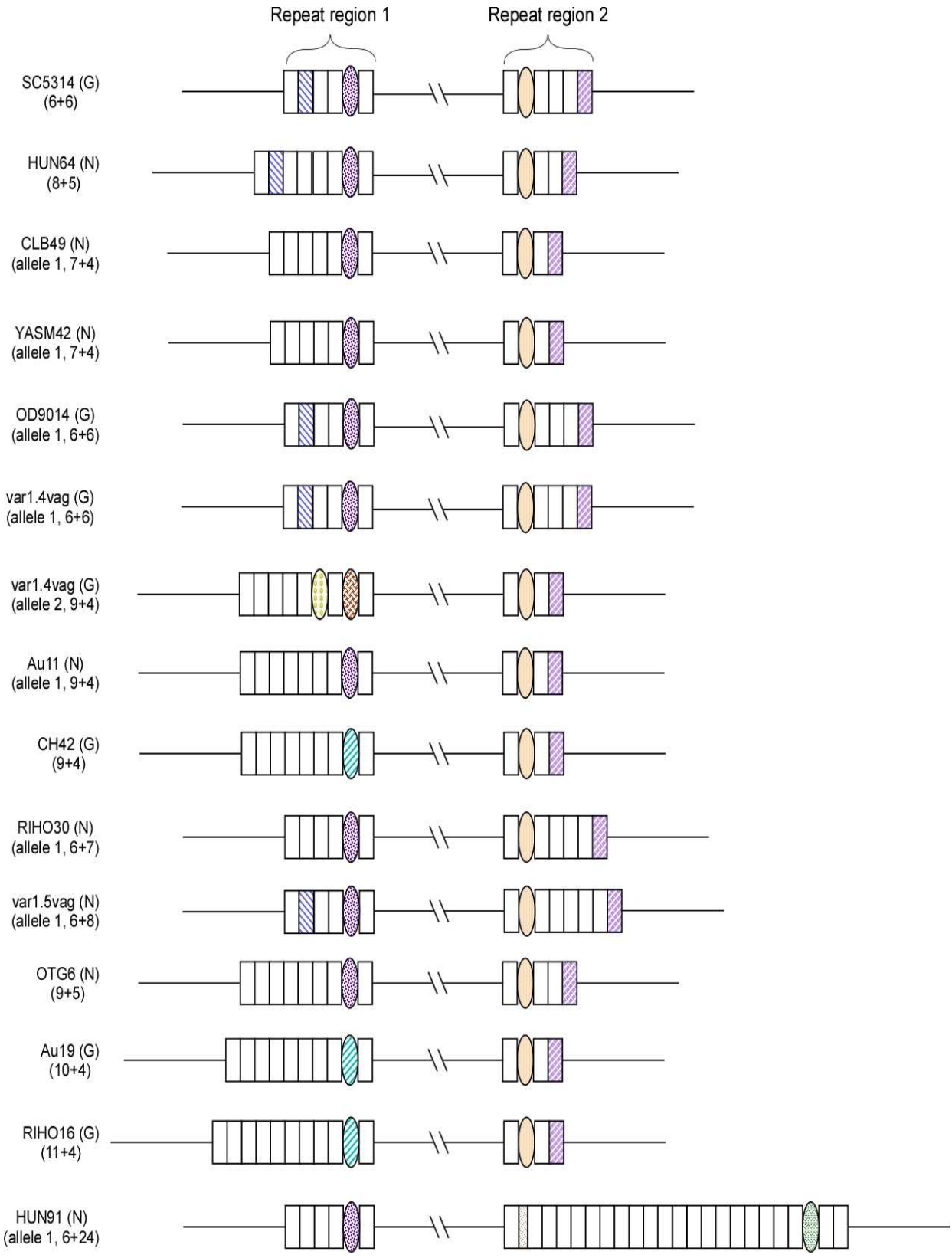

Repeat region 1
Repeat region 2
SC5314 (G)
(6+6)
HUN64 (N)
(8+5)
CLB49 (N)
(allele 1, 7+4)
YASM42 (N)
(allele 1, 7+4)
OD9014 (G)
(allele 1, 6+6)
var1.4vag (G)
(allele 1, 6+6)
var1.4vag (G)
(allele 2, 9+4)
Au11 (N)
(allele 1, 9+4)
CH42 (G)
(9+4)
RIHO30 (N)
(allele 1, 6+7)
var1.5vag (N)
(allele 1, 6+8)
OTG6 (N)
(9+5)
Au19 (G)
(10+4)
RIHO16 (G)
(11+4)
HUN91 (N)
(allele 1, 6+24)

Figure 3.5. Fourteen sequenced *SSR1* alleles from this investigation compared with strain SC5314. (G) is GPG strain and (N) is non-GPG strain. The squares represent the repeat units encoding amino acids Ser and Ala, while the ovals show that point mutation-containing repeat units encoding other amino acids. The sequences of repeat units: TCTGCT; encode amino acids SerAla, TCCGCT, encode amino acids SerAla; TCTGCC, encode amino acids SerAla; GCT, encode amino acid Ala; TCTGTT, encode amino acids SerVal; TCTTCA, encode amino acids SerSer; GCTGCT, encode amino acids AlaAla; CCTGCT, encode amino acids ProAla; TTTTTA, encode amino acids PheLeu; TCTTTA , encode amino acids SerLeu.

3.4. Detection of *SSR1* allele variability among *C. albicans* strains by gel electrophoresis and genotyping

A survey of a larger number of *SSR1* alleles was necessary to define the degree of variability and to gain insights into its role in short and/or long term adaptation. Assuming mutation rates typical of repeat-containing genes, association of specific alleles with specific genetic backgrounds would argue against a role in short term adaptation (**1.3.3**). Absence of such association and/or association of specific alleles with environmental parameters (such as body location, or differences in allele prevalence between commensal and disease-causing isolates) would indicate a role in short term adaptation. Variation in *SSR1* could also have no role in adaptation but just be random variation. In this case alleles would not be preferably associated with either genetic background or specific conditions. To determine if *SSR1* variation is used for long term adaptation, short term adaptation or has no role in adaptation, PCR amplification was used to determine the lengths of repeat regions 1 and 2 in a collection of 132 *C.albicans* strains (**2.2.1** and **2.6**). The sizes of the amplified target regions were determined by gel electrophoresis and genotyping (**2.9** and **2.10**).

The DNA sequence of a selection of alleles (**3.3**) had revealed that, with the exception of one extremely long allele, all allelic variation in the repeat regions was caused by addition of 6 bp units encoding pairs of amino acids and therefore should be more efficiently detected by determining the sizes of repeat units than by cloning and sequencing.

3.4.1. Identification of PCR product length by agarose gel electrophoresis

Two primers (F1 and B2) were designed to amplify a region of *SSR1* containing both of the repeat regions. To determine the length of each repeat region, the two repeat regions needed to be separated. There were a few endonuclease cutting sites found between these regions in strain SC5314. *Hin*dIII was chosen to cut the PCR products and separate the two repeat regions (**2.7**).

For strain SC5314, the PCR product from primers F1/B2 was expected to be 466bp in length. On agarose gel electrophoresis, strain SC5314 gave two products of the approximate length expected (**Figure 3.6**) and was used as a positive control. The detection of other PCR products clearly showed there were differences among strains, although the basis for the variation in the two repeat regions could not be detected. After *Hin*dIII digestion, the resulting fragments obtained for strain SC5314 were expected to be 162bp (fragment A) and 303bp (fragment B). However, because of the limited resolution of agarose electrophoresis, the smaller fragments presented in the gel with misty edges (**Figure 3.7**). On these photos, the variation of the length of the smaller *Hin*dIII fragment containing repeat region 1 was difficult to see clearly, and it was not possible to detect one unit (6bp) variation.

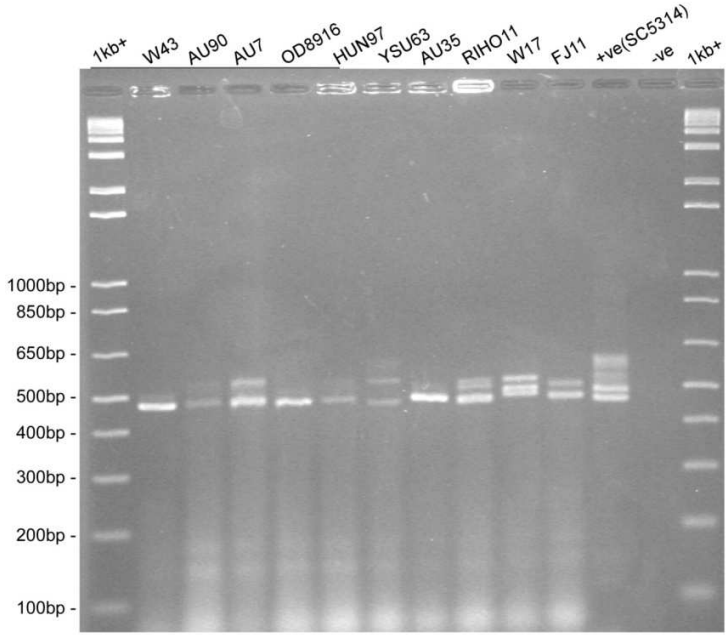


Figure 3.6. PCR products of *SSR1* with primers F1/B2 (before *Hin*dIII digestion) on 1.5% agarose gel. Sizes of the marker fragments are shown to the left of the figure. The gel was run at 65 V for 3 hours (Hoefer HE 99 MAX agarose gel unit).

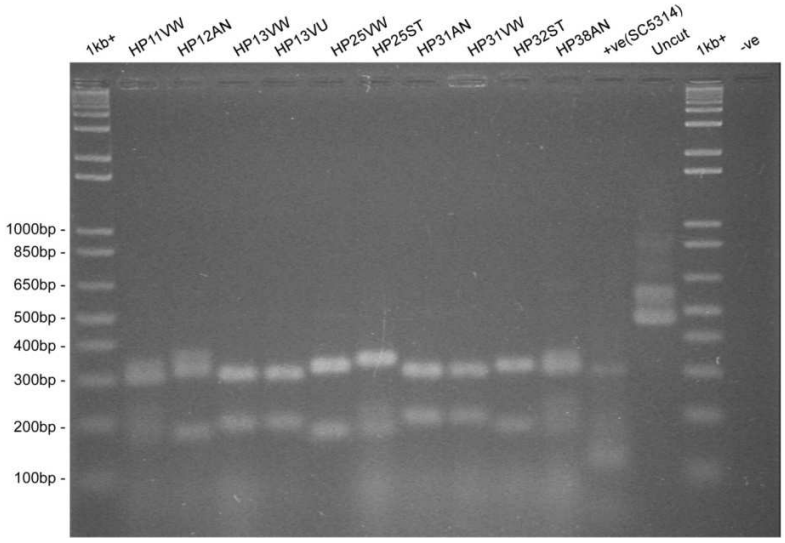


Figure 3.7. PCR products of *SSR1* digested with *Hin*dIII and separated in a 2% agarose gel, 30 V, overnight (Hoefer HE 99 MAX agarose gel unit). Sizes of the marker fragments are shown to the left of the figure.

3.4.2. Identification of PCR product length by polyacrylamide gel electrophoresis

PAGE (polyacrylamide gel electrophoresis) was used to determine the size of smaller DNA fragments. Compared with agarose gel electrophoresis, it has higher resolution and small DNA products often appear as sharp bands (Chang & Yeung, 1995). The extra endonuclease digestion step was avoided by designing primers to amplify the two repeat regions separately. PCR products of the expected sizes were obtained for strain SC5314 and other strains (fragment 1, **Figure 3.8.a** and fragment 2, **Figure 3.8.b**). Two bands were clearly visible in the expected size range for some strains, for example, HUN97 (**Figure 3.8.a** and **Figure 3.8.b**) and RIHO30 (**Figure 3.8.b**). This indicated that *SSR1* is heterozygous in these strains.

As expected, the resolution on PAGE was much higher than on agarose and the PCR products gave sharp bands. To check if PAGE could successfully detect the

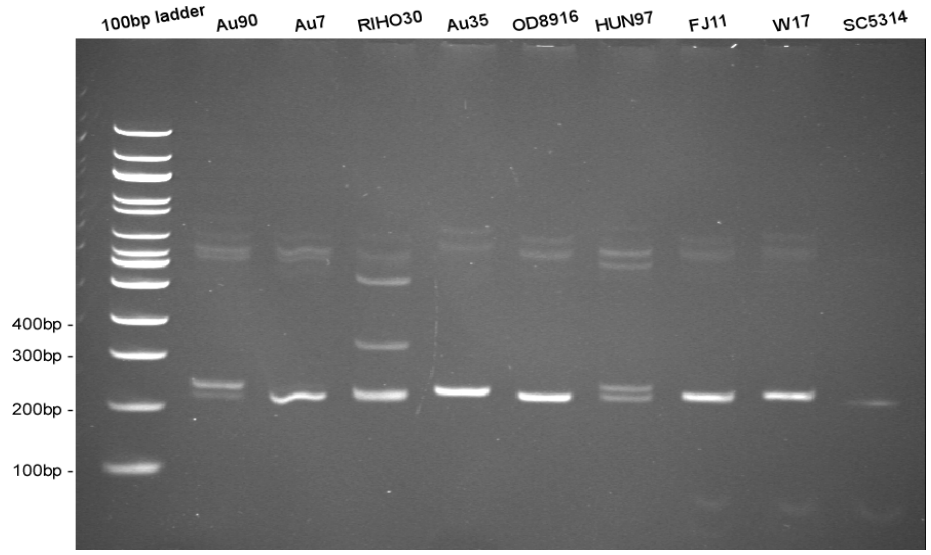
single repeat unit (6 bp) difference, the two sequenced strains SC5314 and HUN64 were compared using 8% PAGE (**Table 3.3**, **Figure 3.9**).

Table 3.3. The characteristics of strains used to check size discrimination on PAGE.

Strains	Group	Genotype	Repeat region 1	Repeat region 2
SC5314(lab strain)	GPG	homozygous	6 units	6 units
HUN64	GPG	homozygous	8 units	5 units

Under the experimental conditions used, the minimum difference between PCR fragments detected on 8% polyacrylamide gels was 2 units (12 bp). A one unit (6 bp) difference in the larger product could not be distinguished. In addition, the distribution of bands in PAGE is affected significantly by other factors such as gel setting conditions and of loading buffer concentration. It was conclude that was that PAGE was not sensitive enough to determine the variation of *SSR1* among strains. Because of the requirement for high accuracy, and the large number of strains to be detected in a limited time, genotyping was used determine the length of the repeat regions.

(**a**)



(**b**)

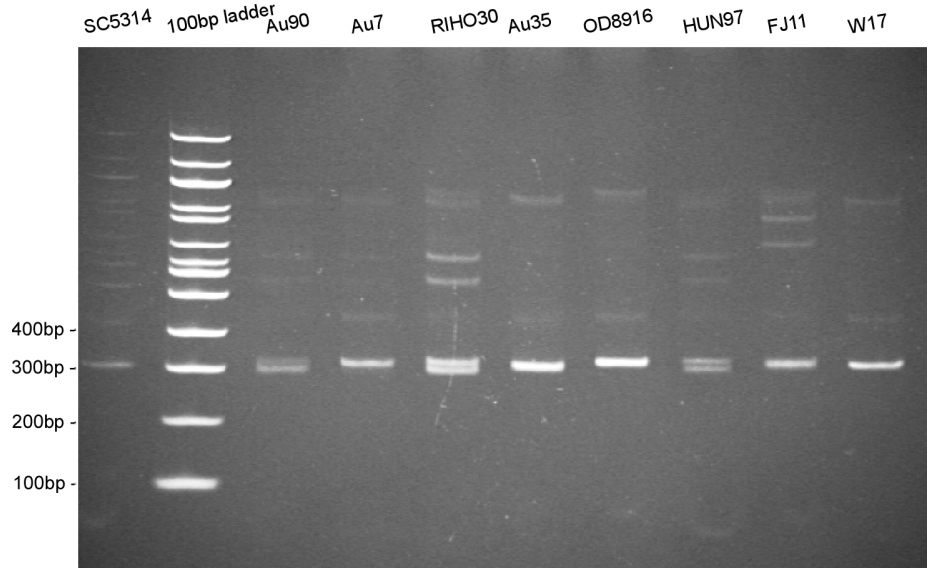


Figure 3.8. PCR products of *SSR1* with primers F1/ SSR1-1 and SSR1-2/SSR1-3 separated in an 8% polyacrylamide gel. (**a**) PCR products of *SSR1* fragment 1 (Primers: F1 and SSR1-1). (**b**) PCR products of *SSR1* fragment 2 (Primers: SSR1-2 and SSR1-3). Very faint large bands were also apparent in most lanes. These are probably artificial bands from PCR reactions (Bradley & Hillis, 1997). In PCR reactions of repeat-containing genes, these artificial bands occur commonly, and they are obvious only in high resolution gels such as polyacrylamide.

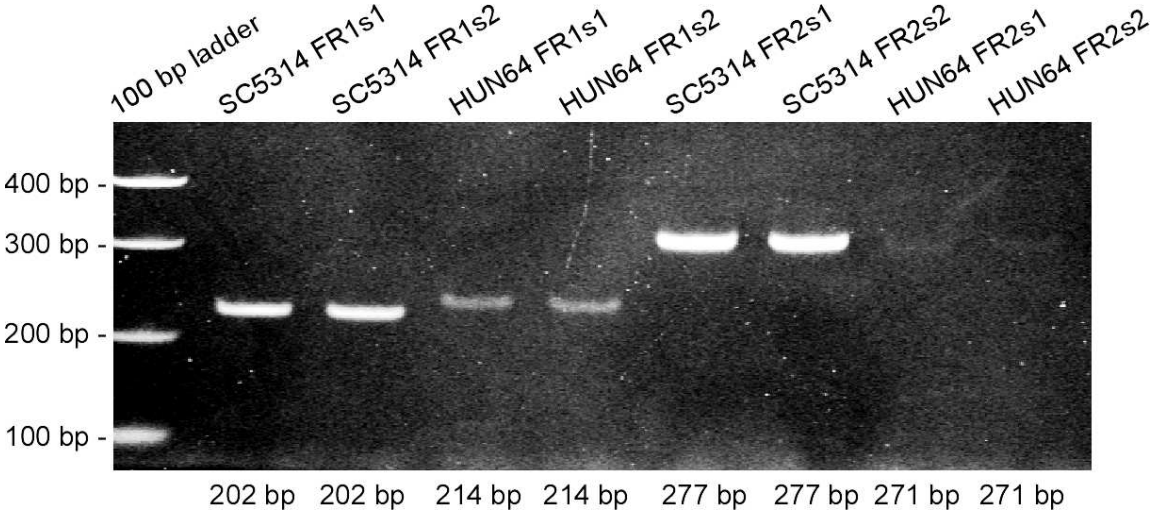


Figure 3.9. The repeat-containing PCR products from SC5314 and HUN64 separated in 8% PAGE. In the lane heading, FR1 means fragment 1 and FR2 means fragment 2; s1 means sample 1 while s2 means a duplicate sample. Between these two strains, there is two units difference in fragment 1, and one unit difference in fragment 2.

3.4.3. Identification of different alleles and allele combinations in *C.albicans* strains by genotyping

3.4.3.1. Identification of length of *SSR1* repeat regions in *C.albicans* strains by genotyping

For genotyping, the two primers pairs F1-FAM/SSR1-1 and SSR1-2-Hex/SSR1-3 were used to amplify the two repeat regions (**Figure 2.1**). The lengths of both fragments could be determined in a single genotyping reaction because the products would be labeled with different fluorescent dyes. In the DNA sequence of strain SC5314, the length of fragment one is 205 bp, fragment two is 277 bp, and is 459 bp when both regions are amplified using the F1-FAM/SSR1-3 primer pair. The lengths of different alleles were then calculated by adding or decreasing

fragment length 6 bp per repeat unit (**Table 3.4).** When determining the number of repeat units, fragment lengths from genotyping were considered as the nearest matching number in **Table 3.4**. Genotyping of repeat-containing fragments from the sequenced strains gave repeat region sizes closely matching, 36 fragments have been matched and the mean difference is 1.59 bp. The matching of genotyping reading and sequencing results are shown in **Figure 3.10**. This showed that the genotyping gives a reliable result in this experiment. The sequence based-sizes and, unlike gel-based techniques, could clearly distinguish alleles differing by one 6bp unit (**Figure 3.11**). Genotyping was therefore chosen for the survey of a large collection of strains.

Table 3.4. Numbers of repeat units and corresponding fragment sizes.

No. of repeat units	Fragment 1 (bp) F1/SSR1-1	Fragment 2 (bp) SSR1-2/SSR1-3	F1/SSR1-3 Fragment (bp) *
4	193	265	(411)
5	199	271	(417)
6	205	277	(423)
7	211	283	(429)
8	217	289	435
9	223	295	441
10	229	301	447
11	235	307	453
12	241	313	459
13	247	319	465

* Fragment sizes in brackets were not actually observed among the strains tested.

Table 3.5 Repeat-containing fragments' sizes in genotyping and in the sequencing results.

Size range in genotyping (bp)	Expected size of sequence (bp)
205.92-206.53	205
212.68-212.79	211
218.84	217
224.36-225.38	223
231.59	229
238.4	235
266.89-267.49	265
272.76-272.96	271
278.4-278.8	277
284.71	283
290.06	289
454.08-454.16	453
459.63-459.87	459
465.58-465.68	465
471.39	471

In this thesis, allele combinations in *SSR1* are written in the form "a+b/c+d" where "a" is the number of repeat units in fragment 1 of allele 1, "b" is the number of repeat units in fragment 2 of allele 1, "c" is the number of repeat units in fragment 1 of allele 2 and "d" is the number of repeat units in fragment 2 of allele 2. To order two different alleles, for fragment 1, the larger number is "a" while the smaller number is "c"; if "a" and "c" are the same, for fragments 2, the larger number is "b" while the smaller number is "d".

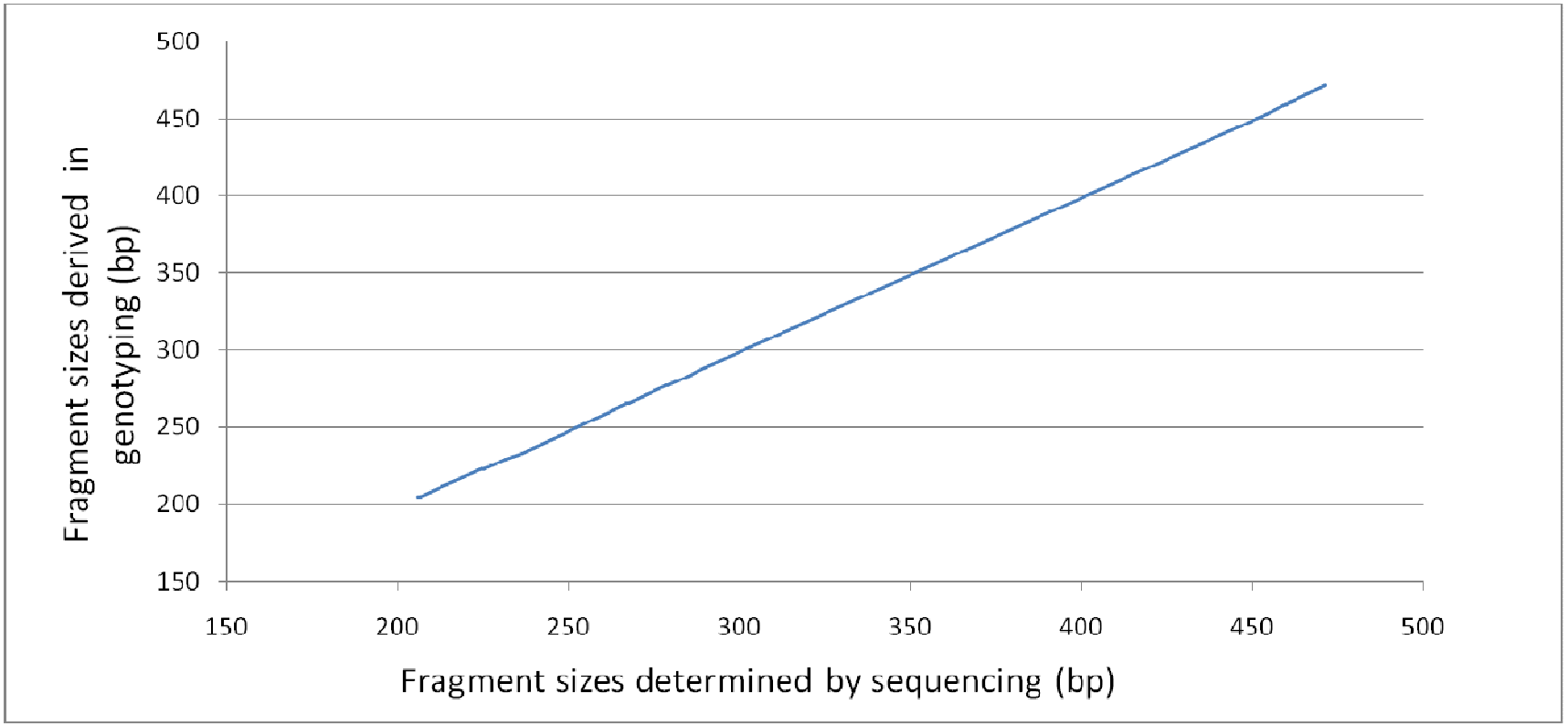


Figure 3.10. Matching of 36 fragments' sequencing results and their genotyping readings. The both results are closely matched.

3.4.3.2. Identification of allele combinations

Because *C.albicans* is a diploid fungus, genotyping of strains heterozygous for *SSR1* was expected to show more than one blue and green peak. For example, for a strain with two different alleles in fragment 1 and two different alleles in fragment 2 (**Figure 3.11.a**), the genotyping should show 2 different coloured peaks for each fragment but give no allele-related information which repeat region1 length is paired with which repeat region 2 length.

To determine the allele combinations in heterozygous *C.albicans* strains, the total fragment length containing both repeat regions was determined using a single primer labeled with a fluorescent dye: F1-FAM. The total fragment was amplified with the primer pair F1-FAM/SSR1-3 (see **2.6.3.5**), and genotyped as described in **2.10**. The PCR product of each allele then presented as a single blue peak in the genotyping results (**Figure 3.11.b**).

Once the total length of the combined regions was known, the allele combinations could be determined by adding the numbers of repeat units, as shown in **Figure 3. 11**. In the figure, genotyping of a strain shows repeat region 1 sizes of 6 and 8 repeat units (blue peaks in **Figure.3.11.a**) and repeat region 2 sizes of 5 and 6 repeat units (green peaks in **Figure.3.11.a**). The combined number of repeats in both repeat regions is 12 in one allele and 13 in the other (**Figure 3.11.b**). Therefore one allele must have a combination of 6 repeats in each region, and the second allele must have 5 units in region 1 and 8 units in region 2, since only these combinations add up to the totals of 12 and 13 units, respectively.

This method of determining allele combinations is possible because the error in genotyping is less than 1.5 bp (**3.4.3.1**), and the sequencing of 14 *SSR1* genes showed that variation only occurs in 6 bp repeat regions (**3.3**). Based on this method, the *SSR1* allele combinations in all 132 strains were determined using this genotyping method. The genotyping results for these are given in Appendix IV.

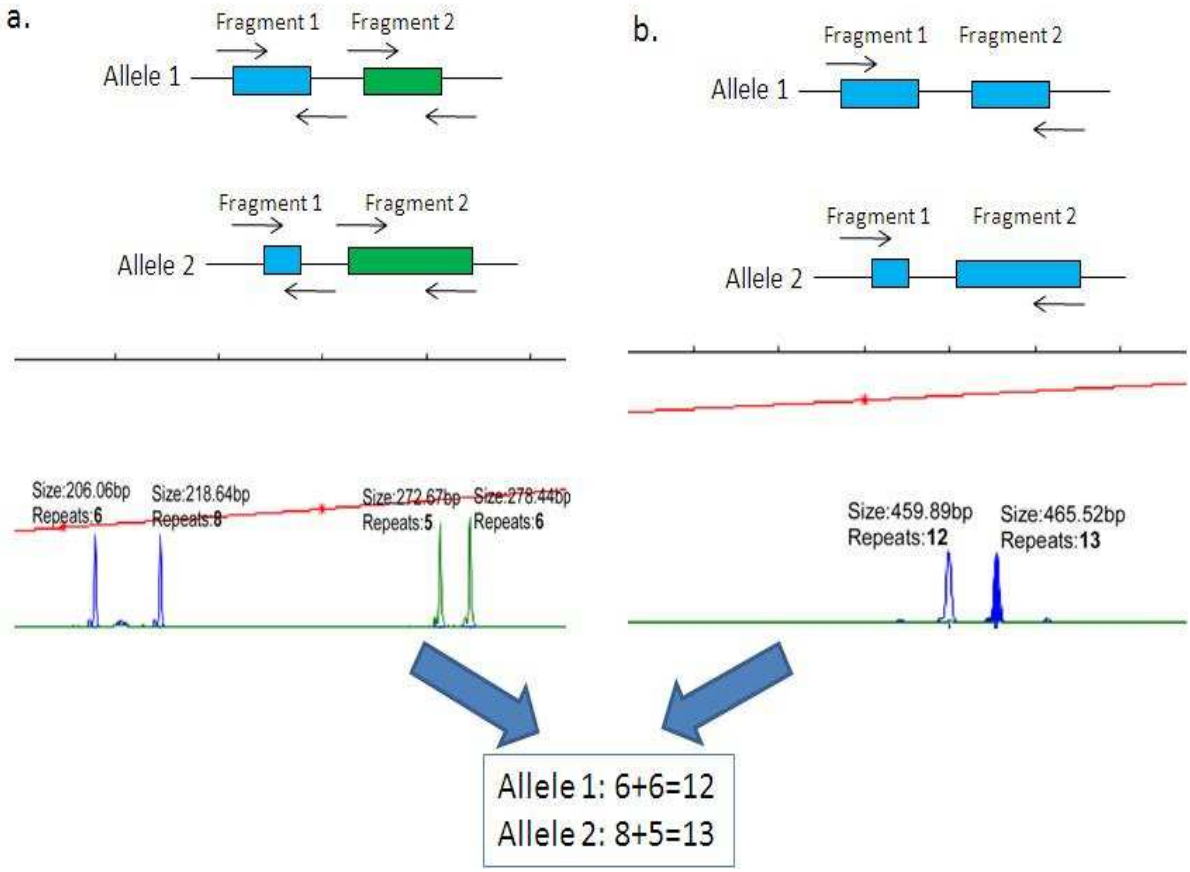


Figure 3.11. The allele/fragment combinations and the relationship with the colour peaks from genotyping a heterozygous strain. (**a**) Primer pairs F1/SSR1-1 and SSR1-2/SSR1-3 give PCR products each containing one repeat region. Two blue peaks are the fragments labeled with FAM; two green peaks are the fragments labeled with Hex. (**b**) Genotyping results of the whole length containing both repeat regions. Primer pairs F1/SSR1-3 give PCR products containing both repeat regions and are labeled with FAM (blue). The red lines in at the background are the slope of the size standard used in this genotyping, with the red stars showing the positions of the actual size standard DNA molecules.

3.4.3.3. Variation of *SSR1* repeat regions among infection-causing strains

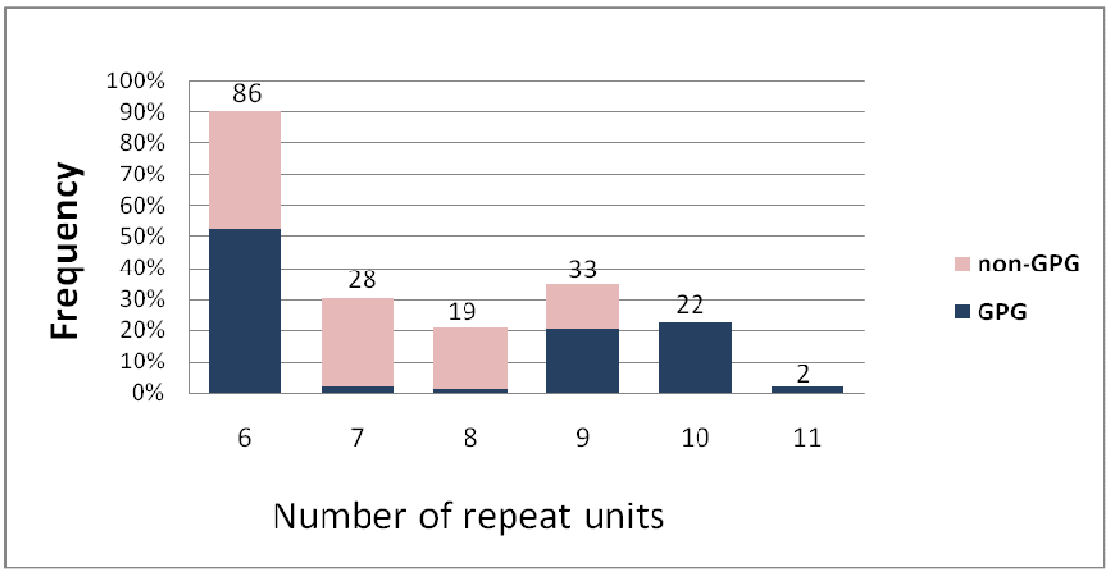
Among a worldwide collection of 96 infection-causing *C.albicans* isolates, including 50 GPG strains and 46 non-GPG (ngpg) strains, 6 lengths of repeat region 1 (**Figure 3.12.a**) and 12 lengths of repeat region 2 (**Figure 3.12.b**), 24 alleles (**Figure 3.13**) and 41 allele combinations (**Figure 3.14**) were identified. GPG strains had 11 different alleles (45.8% of allele types) and 14 allele combinations (35% of combinations) while non-GPG strains had 21 alleles (87.5% of allele types) and 30 allele combinations (73.1% of combinations). The results of genotyping showed:

1) In repeat region 1, 6 repeat units were most prevalent (**Figure 3.12.a**); in repeat region 2, 4 repeat units and 6 repeat units were the most prevalent (**Figure 3.12.b**).

2) Non-GPG isolates had an allele distribution that was significantly more variable than that of GPG isolates (Chi square test, P<0.0001, **Figure 3.13**). This is not surprising because GPG strains are genetically similar and non-GPG strains are very diverse. The distribution of allele 6+6 is significantly different between GPG (45%) and non-GPG (10.8%) strains (Chi-square test, P<0.0001).

3) Non-GPG isolates had a distribution of allele combinations that was significantly more variable than that of GPG isolates (Chi square test, P<0.0001). Among the GPG strains, two very similar allele combinations were over-represented; in both combinations one allele had 6+6 repeats, and the other allele had either 9+4 repeats or 10+4 repeats (**Figure 3.14**). The distributions of these two allele combinations in GPG strains and non-GPG strains were significantly different: the frequency of the allele combination 6+6/9+4 was 30% in GPG and 2.2% in non-GPG strains (Fisher's exact test, P=0.0001); and the frequency of the allele combination 6+6/10+4 was 28% in GPG and 0% in non-GPG strains (Fisher's exact test, P<0.0001).

(**a**)



(**b**)

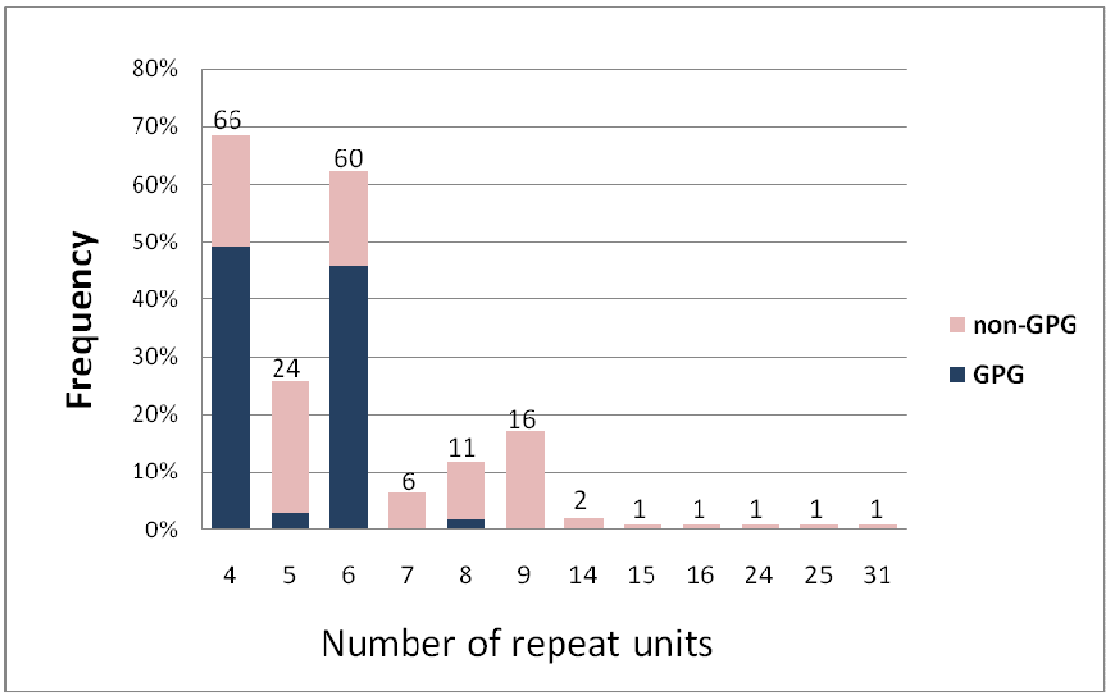


Figure 3.12. Proportions of repeat numbers in repeat region 1 (**a**) and repeat region 2 (**b**). The addition of the pink bar (non-GPG) and the blue bar (GPG) gives the total proportion of that repeat number. The numbers above the bars indicate number of the alleles analysed.

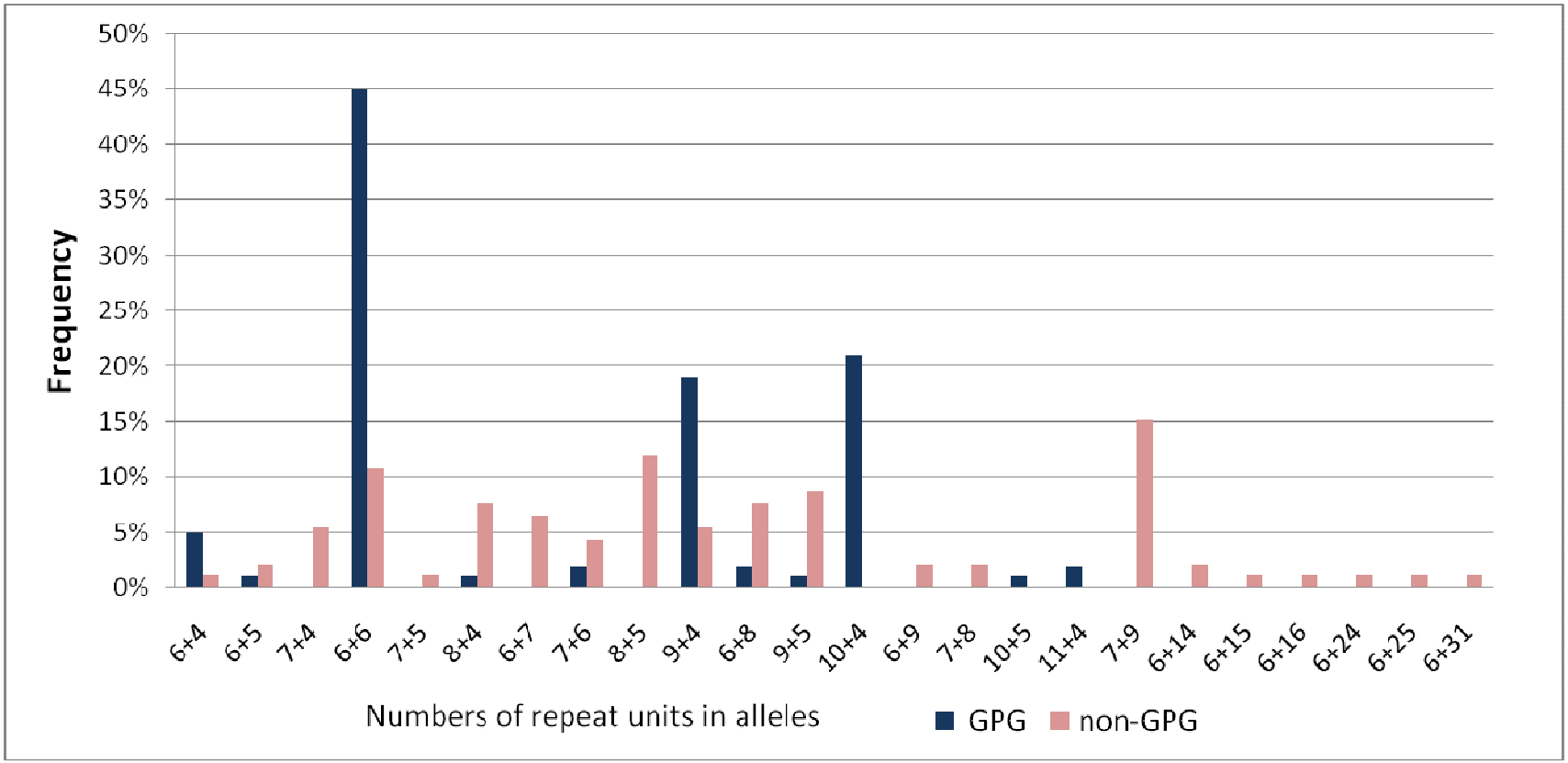


Figure 3.13. Frequencies of *SSR1* alleles in infection-causing strains. Alleles are labeled as repeat numbers in fragment 1 + repeat numbers in fragment 2. A total of 92 non-GPG and 100 GPG alleles were analyzed and 192 alleles altogether.

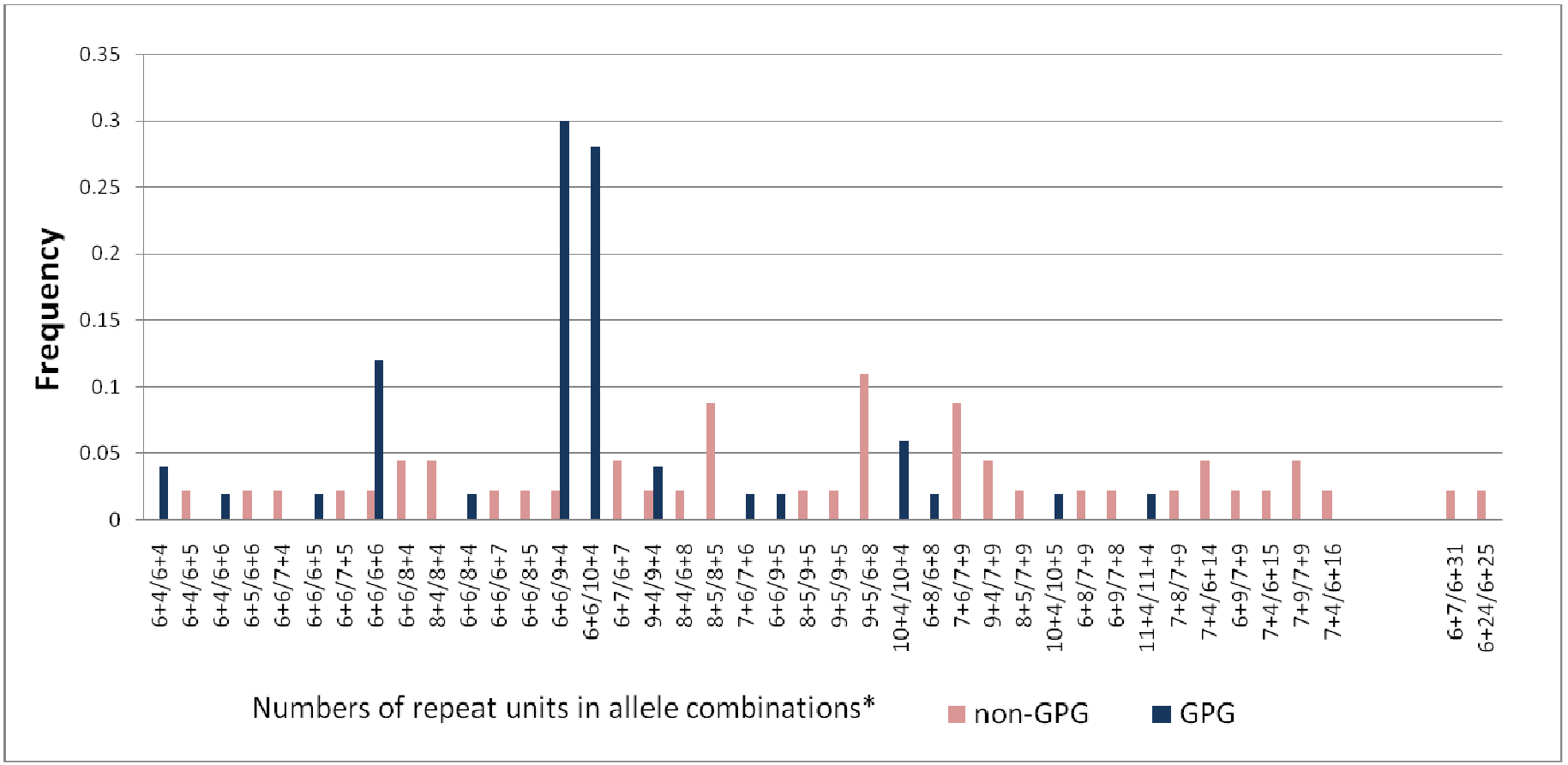


Figure 3.14. Frequencies of *SSR1* allele combinations in infection-causing strains. *Allele combinations are labeled with repeat numbers in allele 1 (fragment 1+fragment 2) followed by repeat numbers in allele 2 (fragment 1+fragment 2). A total of 46 non-GPG and 50 GPG strains were analyzed and 96 allele combinations altogether.

3.4.3.4. Different alleles and allele combinations in commensal *C. albicans* strains

To find out if there were differences in *SSR1* genotypes between infection-causing strains and commensal strains, and determine if hypermutability might be used for short term adaptation, *SSR1* allele combinations were compared between commensal and infection-causing GPG strains. Difference between *SSR1* genotypes in commensal and infection-causing isolates would suggest that *SSR1* might be used in short-term adaptation to different interactions with the host.

Only GPG strains were compared in this experiment because they have similar genetic backgrounds whereas non-GPG strains are genetically diverse. This minimizes the number of variables (as much as is possible with wild-type strains), so any associations detected are more likely to be due to changes in *SSR1*. Another reason for choosing GPG commensal strains was the small numbers for all other genotypes which would cast doubt on any results.

Among the 36 commensal GPG strains investigated, 8 allele combinations were identified. In contrast to GPG infection-causing strains, GPG commensal strains had only one over-represented allele combination. The allele combination 6+6/10+4 accounted for 50% of GPG commensal strains (**Figure 3.15**). The 6+6/9+4 allele combination was less frequent in GPG commensal strains compared to infection-causing strains with a frequency 5.6% in commensals and 30% in infection-causing strains. The distributions of allele combination 6+6/9+4 in GPG infection-causing strains and commensal strains are significantly different (Fisher's exact test, P=0.004). The 6+6/10+4 allele was more frequent in GPG commensal isolates compared to infection-causing isolates, with a frequency of 50% in commensals and 28% in infection-causing strains. The distributions of allele combination 6+6/10+4 in GPG infection-causing strains and commensal strains are significantly different (Fisher's exact test, P=0.031). These results suggested the possibility that a variation of one repeat unit in *SSR1* might be under

environmental selection, and may be used as short-term adaptation for *C. albicans* between infection-causing and commensal states.

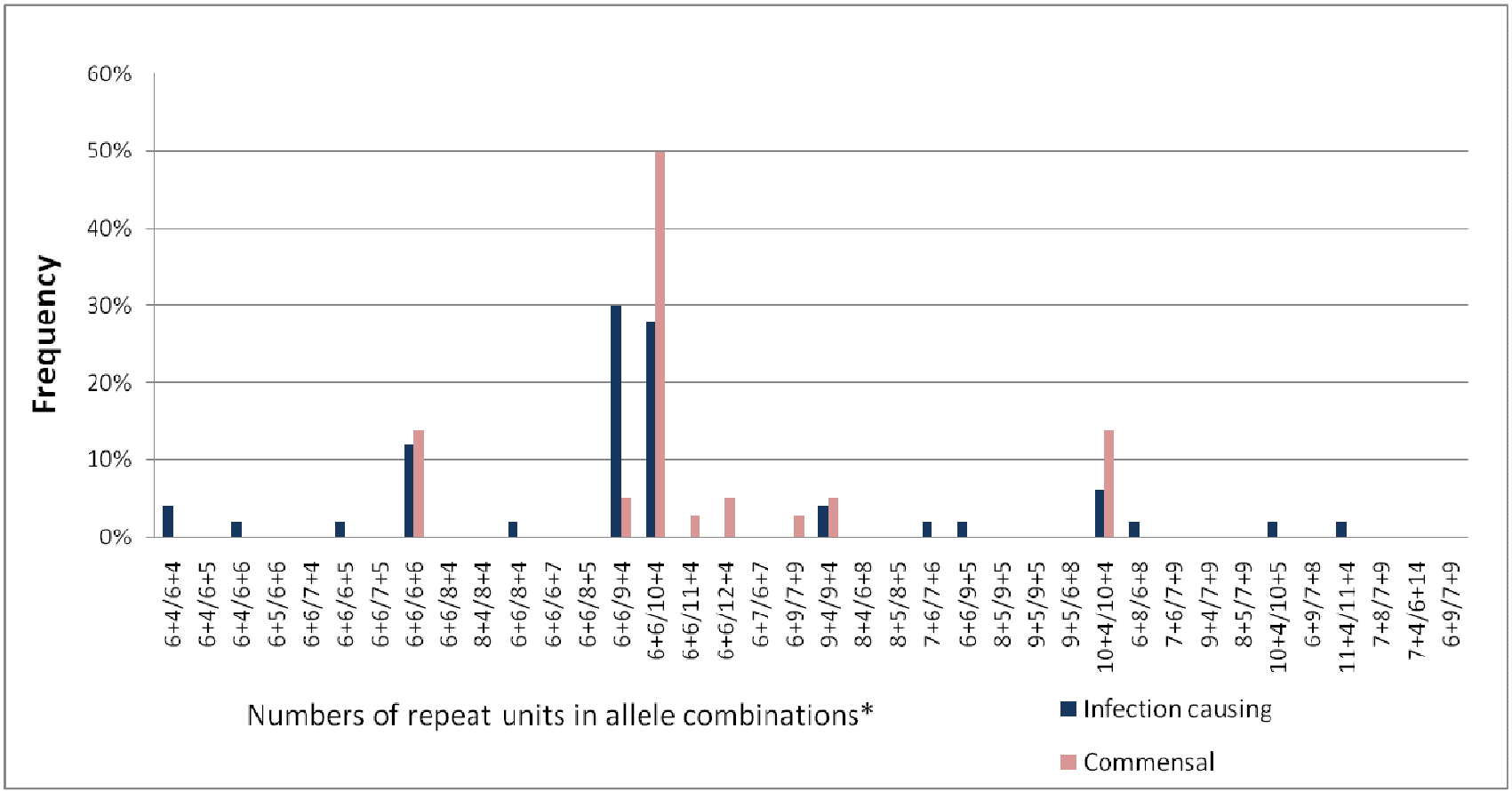


Figure 3.15. Allele combinations in commensal and infection-causing GPG strains. *Allele combinations are labeled with repeat numbers in allele 1 (fragment 1+fragment 2) followed by repeat numbers in allele 2 (fragment 1+fragment 2). 50 infection – causing strains and 36 commensal strains were analyzed and 86 allele combinations altogether.

3.4.3.5. Site of isolation and alleles and allele combinations

To find out if there was any association between *SSR1* genotypes and body sites of isolation, and determine if the hypermutability may be used for short term adaptation to different host body sites, *SSR1* alleles in all strains were investigated. The genotyping of infection-causing strains showed no clear association between body site of isolation and repeat regions, alleles or allele combinations (**Figure 3.16. a-d**). However, it is interesting to note that the two non-GPG strains with extremely long alleles are both from sterile sites: RIHO30 (6+7/6+31) and HUN91 (6+24/6+25) (**Figure 3.16.d**). And in GPG strains, one third of the allele combinations from sterile sites isolates are not found at any other body site (**Figure 3.17.a**). They are (11+4/11+4) (10+4/10+5) and (6+6/9+5). Comparison of allele combinations of GPG infection-causing strains and GPG commensal strains showed no obvious association between site of isolation and allele combination (**Figure 3.17.a-b**). These results suggested that *SSR1* alleles are not used for short-term adaptation at different body sites to a degree sufficient to verify this within the sample size used.

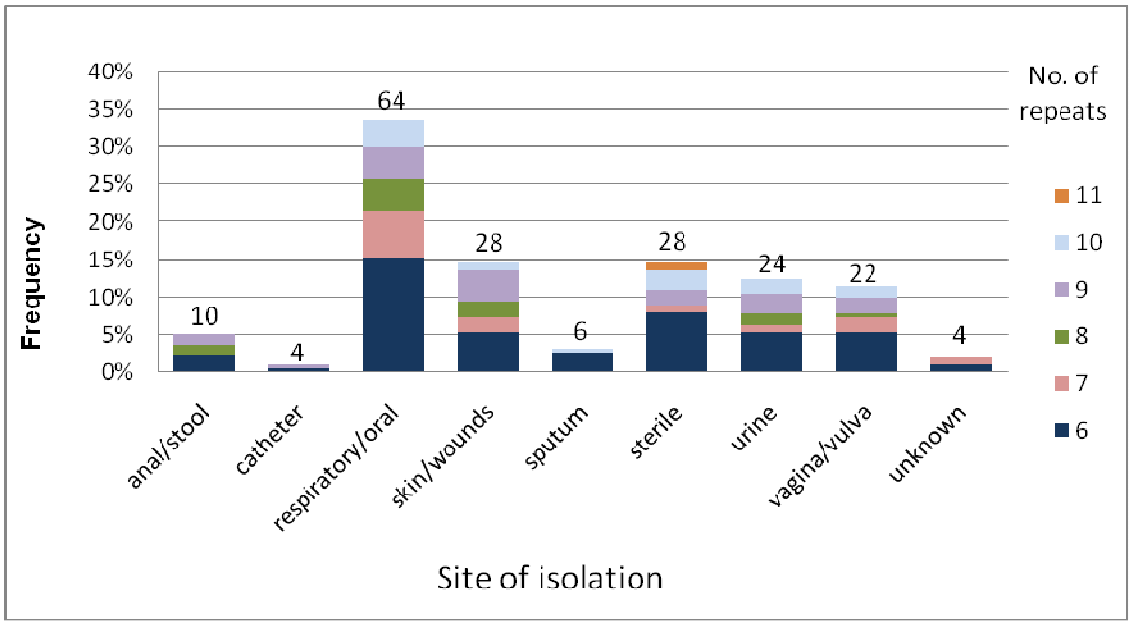


Figure 3.16 (a). Site of isolation of infection-causing *SSR1* repeat region 1. The numbers above the bars indicate number of the alleles analysed.

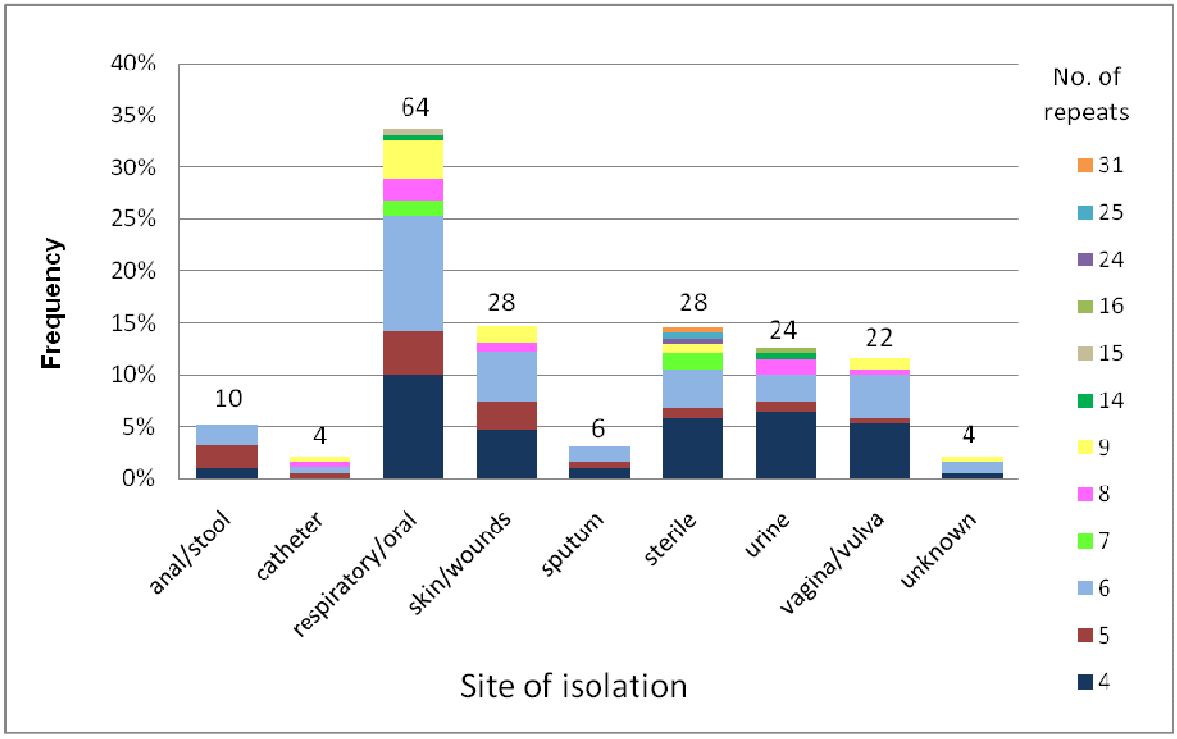


Figure 3.16 (b). Site of isolation of infection-causing *SSR1* repeat region 2. The numbers above the bars indicate number of the alleles analysed.

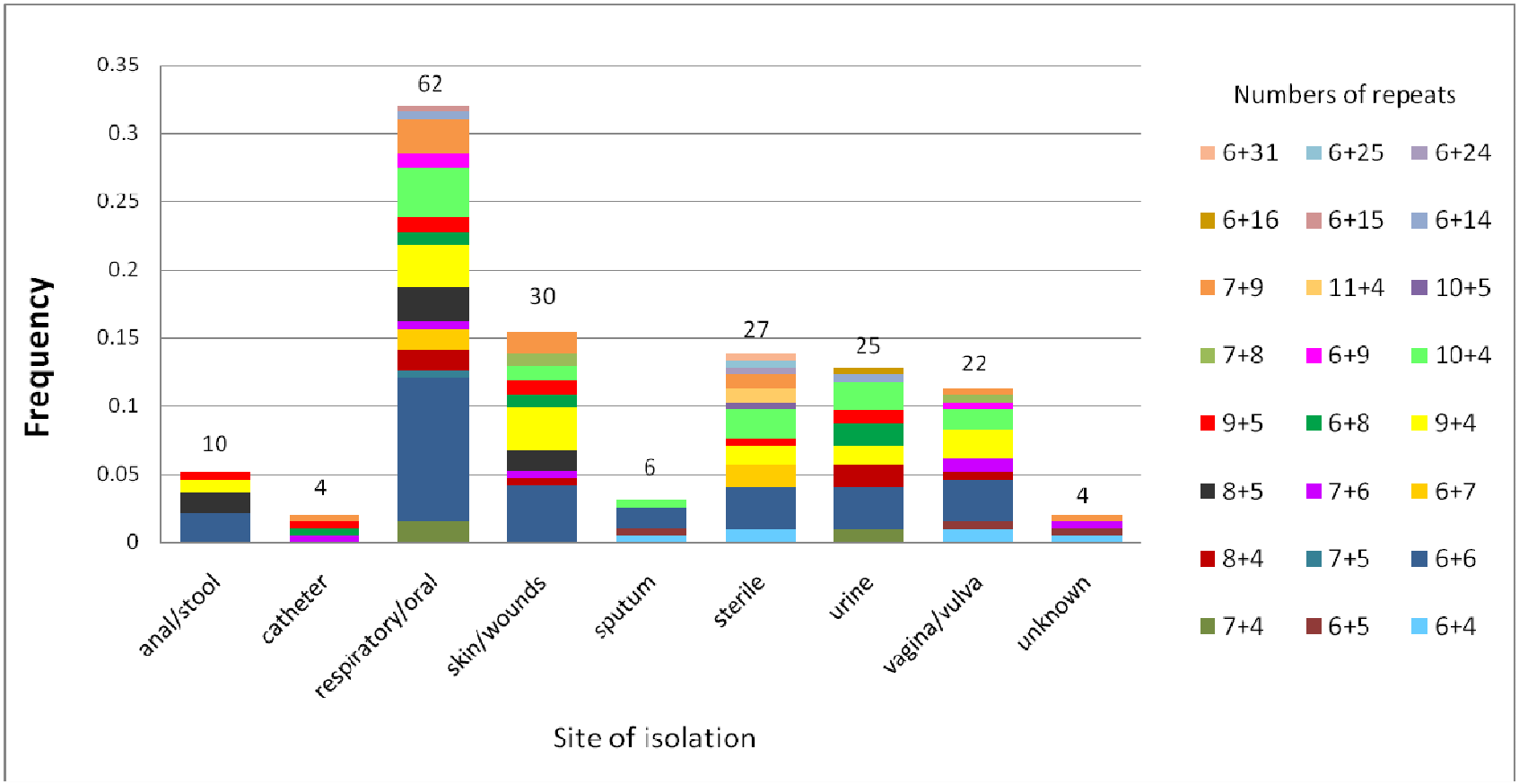


Figure 3.16 (c). Site of isolation of infection-causing *SSR1* alleles. Alleles are labeled as repeat numbers in fragment 1 + repeat numbers in fragment 2. The numbers above the bars indicate number of the alleles analysed.

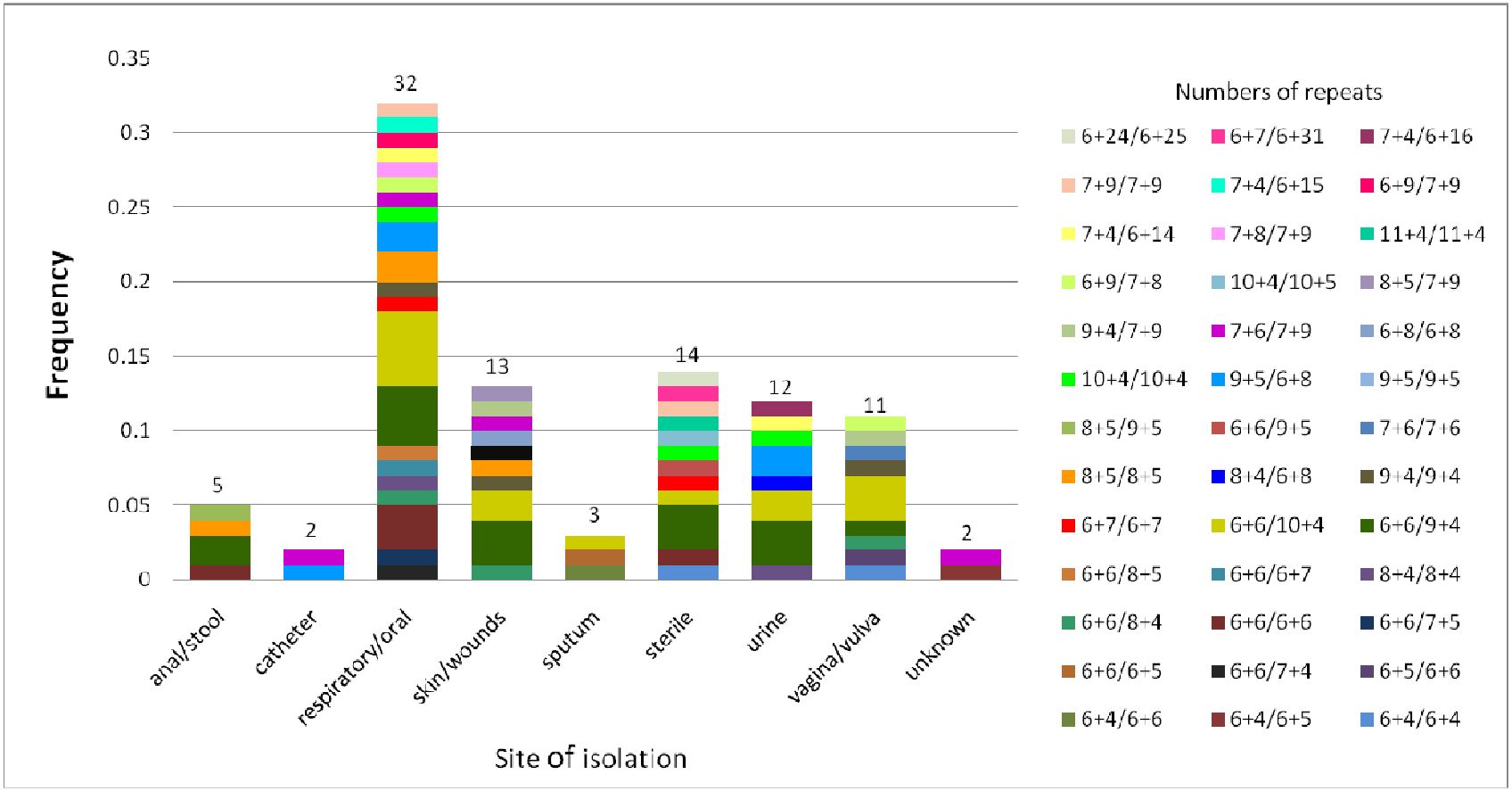


Figure 3.16 (d). Site of isolation of infection-causing *SSR1* allele combinations. Allele combinations are labeled with repeat numbers in allele 1 (fragment 1+fragment 2) followed by repeat numbers in allele 2 (fragment 1+fragment 2). The numbers labelled above the bars present numbers of the strains. The numbers above the bars indicate number of the strains analysed.

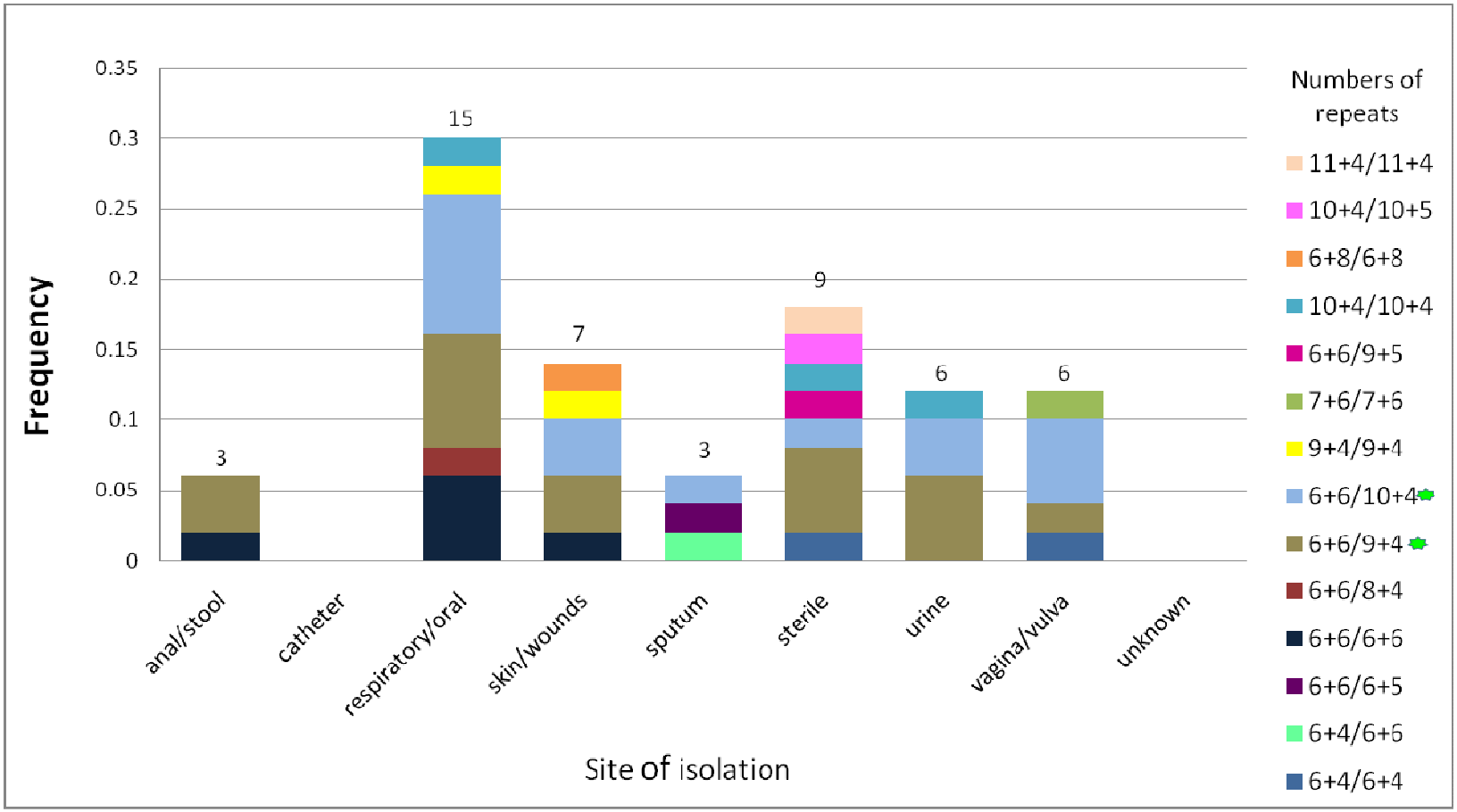


Figure 3.17 (a). Site of isolation of infection-causing GPG *SSR1* allele combinations. Allele combinations are labeled with repeat numbers in allele 1 (fragment 1+fragment 2) followed by repeat numbers in allele 2 (fragment 1+fragment 2). *Most prevalent combinations in GPG infection-causing strains. The numbers above the bars indicate number of the strains analysed.

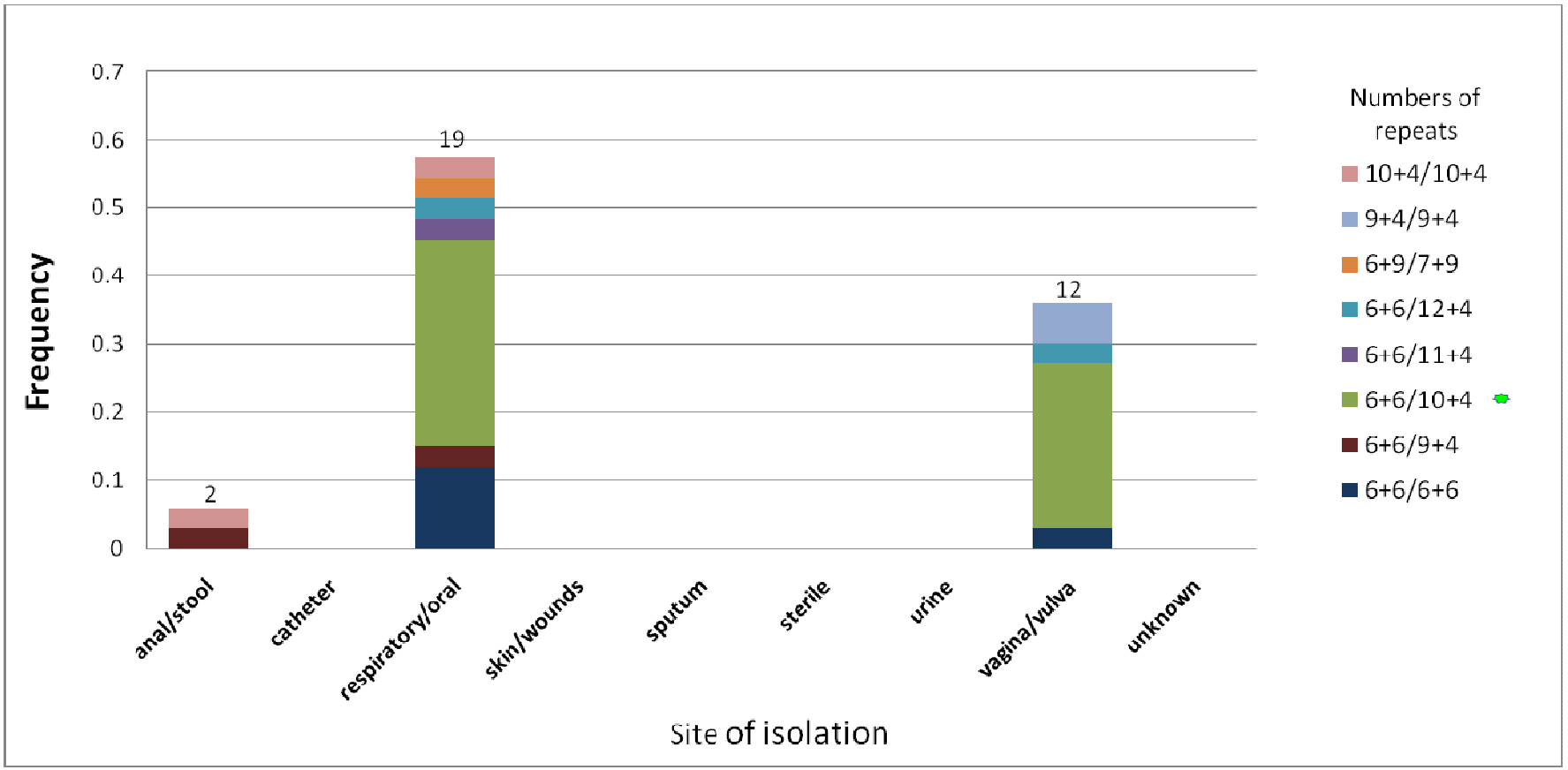


Figure 3.17 (b). Site of isolation of commensal GPG *SSR1* allele combinations. Allele combinations are labeled with repeat numbers in allele 1 (fragment 1+fragment 2) followed by repeat numbers in allele 2 (fragment 1+fragment 2).* Most prevalent combination in GPG commensal strain. The numbers above the bars indicate number of the strains analysed.

3.5. Generation of new alleles in laboratory cultures

3.5.1. Generation of new alleles from the longest *SSR1* allele using laboratory cultures

All previous experiments with repeat regions in this study used isolates from humans. As discussed in the introduction, the repeat regions that have been maintained may be those that are the most beneficial genotype in this environment. If this is the case, the variation of repeat regions between generations would be expected to be relatively rare. To be able to fully interpret the allele distribution data, it is essential to determine at what rates new alleles are generated by mutation in the absence of the human host environment as selection in the host may eliminate many such mutations (Zhang *et al.*, 2010). In laboratory culture, where nutrition is provided and the cells have less stress or selection pressure, clones with mutations among generations might be expected to survive and the frequency with which they can be detected can be used to estimate the rate of mutation.

The mutation rate of the *SSR1* gene was estimated for a strain cultured *in vitro* for 300 generations (see **2.12**). Alleles that contain more repeat units tend to have a higher frequency of mutation than other alleles (Legendre *et al.*, 2007). For this reason, the non-GPG strain (RIHO30) with the allele combination 6+7/6+31 was studied. Another strain, HUN91 was also subcultured for 300 generations but due to time constraints, colonies were stored but not investigated further.

After the 300 divisions, new alleles were detected in 2 out of 60 single colonies by genotyping (**Figure 3.18**). Repeat regions 2 in allele 2 of colonies #8 and #53 both had about a 6 bp (1 repeat unit) increase from 31 to 32 repeat units. This gives an estimate of the mutation rate for *SSR1* allele 2, in strain RIHO30, of 1.11×10^{-4} mutations (2 mutations in 60 colonies after 300 divisions) per cell division and 5.5×10^{-5} per allele and 2.7×10^{-5} per repeat region (binomial confidence: 0.0041 - 0.1153). However, this method probably underestimates the true rate of generation

of new *SSR1* alleles because it cannot detect changes in sequence, or any new alleles that arose and reverted within the 300 generations. Furthermore only one type of adaptive environment was provided.

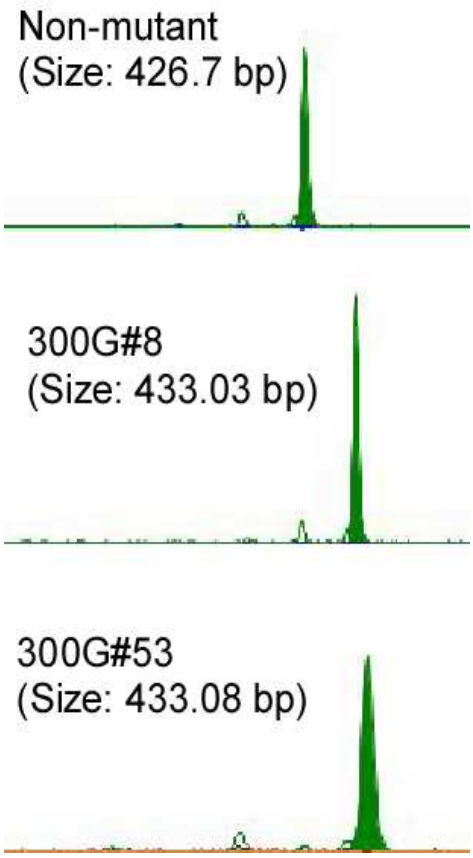


Figure 3.18. Genotyping of repeat region 2 in mutant alleles generated after 300 generations.

3.5.2. Estimated rates of mutations in other alleles

As mentioned in section **3.2**, the SERV software program can predict the mutability of repeat regions in *Saccharomyces cerevisiae*, based on the repetitiveness and length of a repeat region, expressed as a so-called VARscore (Legendre *et al.*, 2007). Although it is not known if these predictions also hold true for *C. albicans*, the above experimentally determined mutation rate appears to be near the range of rates that would be expected for such a repeat in *S. cerevisiae* (**Figure 3.19**). The data of Legendre and coworkers can be used to provide a rough estimate of the range of mutation rates expected in the various *SSR1* alleles. On this basis, the RIHO30 *SSR1* alleles with a VARscore of 0.328 (the average of the VAR scores of all 4 repeat regions, see **Table 3.6**) would be expected to have a mutation rate of ~1×10^{-6} (**Figure 3.19**). Compared with the mutation rate of the artificial repeat genes in *S. cerevisiae*, the observed mutation rate of 2.7×10^{-5} per repeat region of *SSR1* (RIHO30 allele) in *C.albicans* is appears higher (**Figure 3.19**). Also, compared with the mutation rate of the *C. albicans* repeats-containing gene *PNG2*, which has a mutation rate of 2.8×10^{-5} per cell division (Zhang *et al.*, 2010), the mutation rate of 1.11×10^{-4} per cell division of *SSR1* is appears higher.

Table 3.6. VARscores of 4 repeat regions in RIHO30 *SSR1* alleles.

Allele	Repeat region 1	VARscore	Repeat region 2	VARscore
RIHO30 (1)	6 units	**0.0596**	7 units	**0.353**
RIHO30 (2)	6 units	**-0.072**	31 units	**0.974**

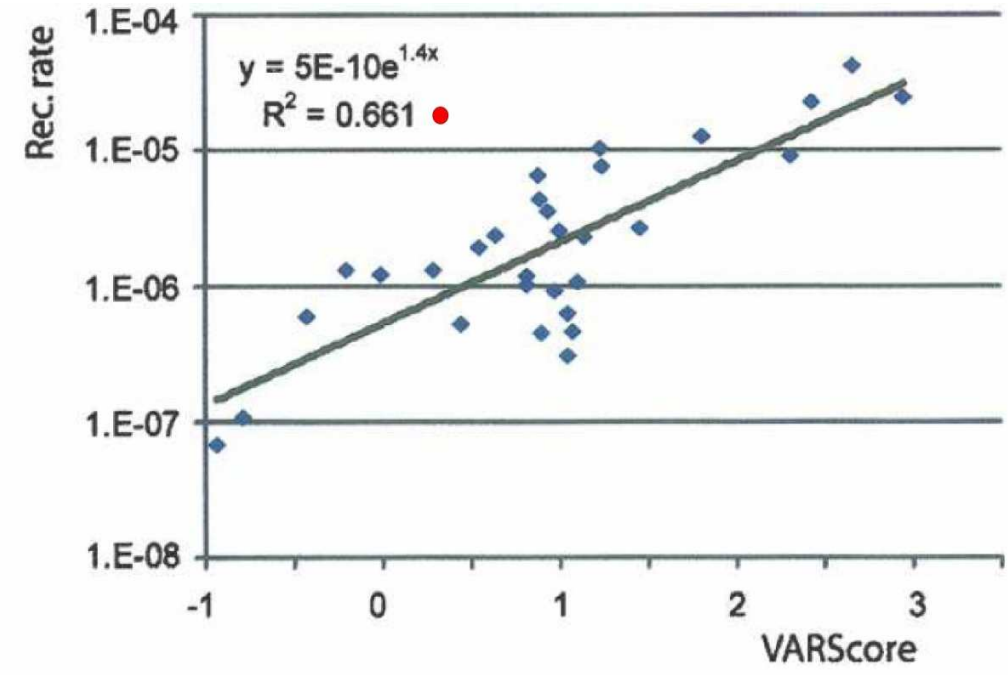


Figure 3.19. The correlation (binomial confidence: 0.0041- 0.1153) of VARscore and mutation rates (Legendre *et al.*, 2007). The gray dots = each repeats' VARscores and its experimental mutation rate from Legendre *et al.*, (2007). The VAR score and mutation rate determined in RIHO30 (6+31) is shown as a red data point. Because RIHO30 has a total of four repeat regions (2 per allele), the y value (**2.7×10^{-5}**) of the red point represents the number of mutations per generation per one repeat region (the rate of new alleles per cell per generation divided by 4) and the x value (**0.328**) represents the average of the VAR scores of all 4 repeat regions in the two alleles)

3.6. Rapid mutation is not detected in repeat regions of *SSR1* in rat models

Experiments at the University of Otago (with a primary aim unrelated to this thesis) have established long-term commensal gastrointestinal colonization in rats following incubation with 5 pairs of infection-causing human isolates (section **2.13**). This provided an opportunity to test if a change of host, a change of interaction from infection to commensalism, or a combination of both, might bring about detectable changes in *SSR1* alleles. If the new conditions very strongly select for alleles different from those of the original ones, detectable changes might occur, and this could directly show that *SSR1* can act as a contingency gene in short term adaptation.

Three groups of rats each received two *C.albicans* strains as parents in the mating experiment (**Table 3.7**). Samples were obtained 4 weeks after inoculation of the rats. This represented approximately 39 doublings in total as the *C. albicans* doubling time in the rat host is ~1.4 doublings per day (R. D. Cannon, unpublished observations). Colonies were mixed in each sample from each rat model, and the repeat regions were amplified by PCR with primers F1/SSR1-1 and SSR1-2/SSR1-3, and analysed by PAGE (see **2.6** and **2.13**).

If the variation of the tandem repeats of *SSR1* genes provide for rapid adaptation when the host environment is changed significantly, mutations would be expected to occur when the cells are growing in the new host environment. If this is the case, due to the potential range of different sizes of alleles, the PCR products of the repeat regions would appear as a fuzzy band on the gel. The results showed sharp bands in the polyacrylamide gel (**Figure 3.20**), indicating that there was no detectable changes in *SSR1* tandem repeat size in the rat host environment.

To identify the possible mutations in more detail, or mutation on a smaller scale than would be detected above, the two repeat regions of the *SSR1* gene in 60 single colonies from each rat model culture were amplified by PCR, and the products were analysed by PAGE (see **2.9**). If *SSR1* acts as a contingency gene in short term adaptation, the mutated alleles would appear as different size bands to

the original. Otherwise, if no mutation occurred, the PCR product size would remain the same for all the 60 colonies. All of the products presented as sharp DNA bands and no size-affecting mutations were detected. An example of the PCR results is shown in **Figure.3.21**. These results suggest that the *SSR1* genes maintain a fixed size in a particular strain in the rat host.

The *SSR1* alleles in all 60 colonies (as shown by PAGE) represented only one of the parents (**Figure.3.21**, **Table 3.7**), suggesting that one strain had disappeared. This is in agreement with the finding of the Otago researchers, who had established that one strain of each pair had largely disappeared using the two markers MPA^r (conferring resistance to mycophenolic acid) and NAT^r (conferring resistance to nourseothricin) (Beckerman *et al.*, 2001; Shen *et al.*, 2005).

The present study used 8% polyacrylamide gels. As mentioned in **3.4.2**, one unit (6 bp) difference between alleles cannot be detected with certainty in these gels. The results strongly suggested that mutations incurring changes of two repeat units (12bp) or more had not occurred.

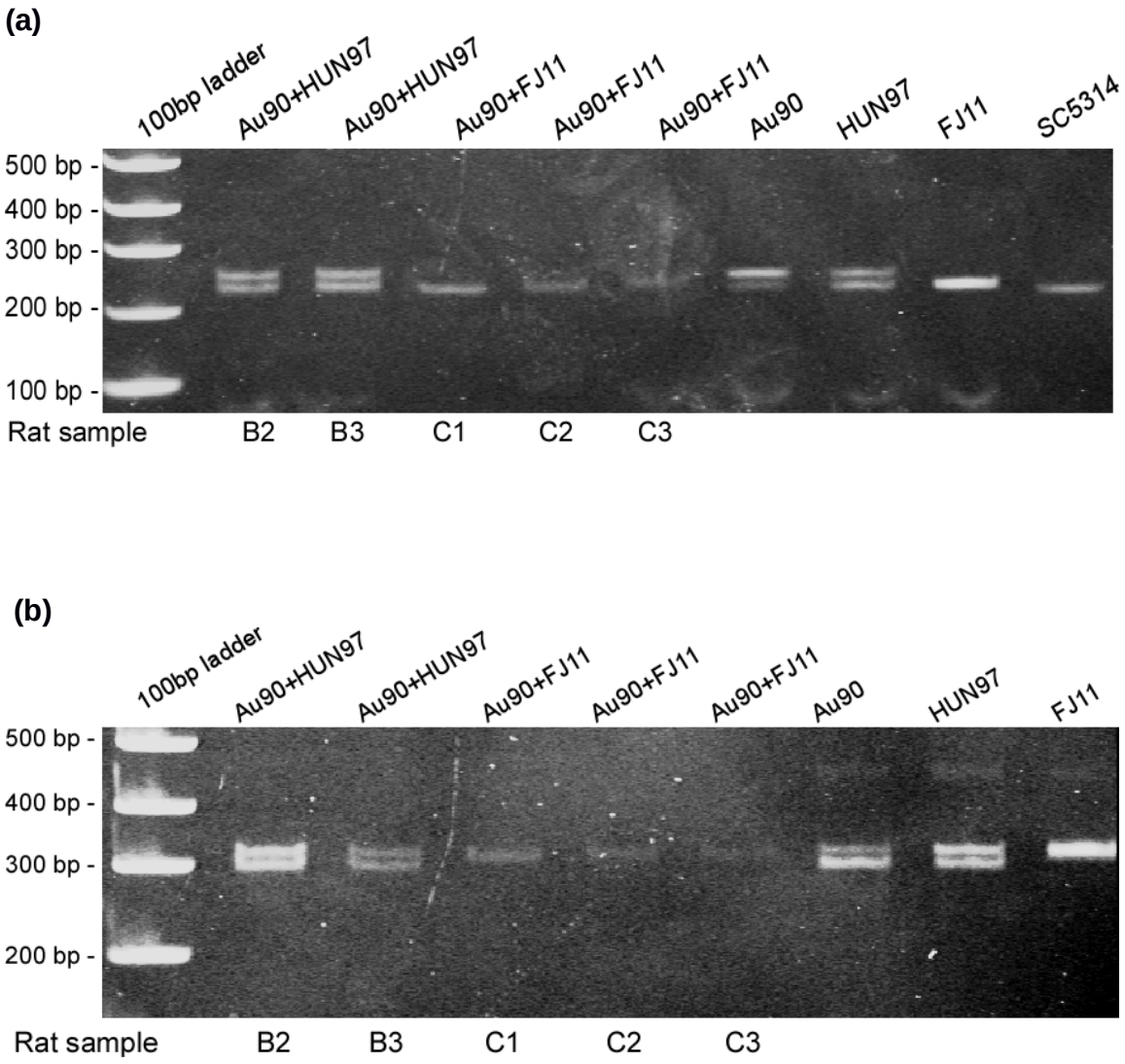


Figure 3.20. PCR results of colony mixtures from a culture from a rat model. (**a**) fragment 1 and (**b**) fragment 2. Lane headings show parent strains in each rat sample (B2, B3, C1, C2, C3), respectively, and strains used for controls.

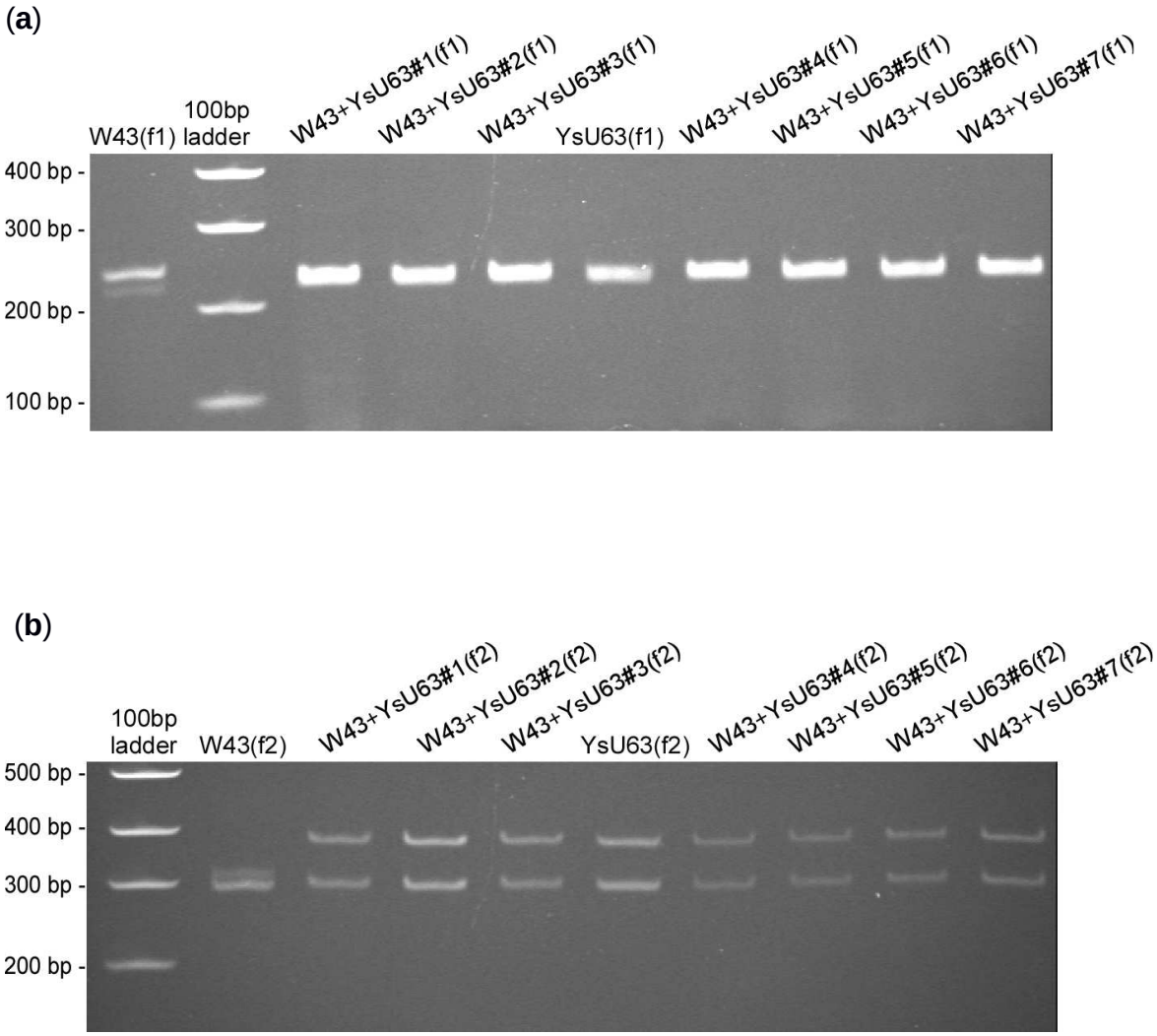


Figure 3.21. PCR results of 7 of the 60 colonies from the culture from a rat sample. (**a**) fragment 1 and (**b**) fragment 2. Lane headings show parent strains and colony number (#) for rat sample A1, strains used for controls, and fragment number (f).

Table 3.7. Comparison of parent strains and cultured strains in rat hosts.

Rat model samples	Parent strain 1	Parent strain 2	Fuzzy/Sharp PCR band	Dominant strain*
A1, A2, A3	YsU63 (non-GPG)	W43 (GPG)	Sharp band	YsU63
B2, B3	Au90 (GPG)	HUN97 (GPG)	Sharp band	?
C1, C2, C3	Au90 (GPG)	FJ11 (non-GPG)	Sharp band	FJ11

*The dominant strain was determined from the pattern of *SSR1* fragments in PAGE (**Figure 3.20** and **3.21**). In group B samples, *SSR1* alleles are the same length in the 2 parent strains AU90 and HUN97, therefore it was impossible to distinguish between these two strains based on the length of the PCR products.

3.7. Synonymous point mutations reduce the predicted mutability of repeat regions

According to the sequencing results (**Figure 3.5**), the differences in repeat regions among strains are not only due to the number of repeats. Although the repeat unit is primarily TCTGCT (encoding SerAla), a few synonymous point mutations occur, i.e. in some repeat regions TCC or TCA are used to encode Ser, and GCC is used to encode Alanine. Usually, but not always, the use of a single codon for each amino acid increases repetitiveness and thus mutability of a DNA repeat region encoding amino acid repeats, as predicted by the VAR score (Schmid and Wilkins, unpublished results). In the case of all sequenced *SSR1* alleles, use of only the predominant codons in all repeat units (TCT for serine and GCT for alanine) should increase predicted mutability over that of the existing regions (**Table 3.8**). Because point mutations typically arise at a rate much lower than that of repeat unit loss or gain (Legendre *et al.*, 2007), units with point mutations would be expected to be rapidly replaced by other repeat units if the maximum mutation rate was selected for in *SSR1*. The presence of these synonymous mutations indicates that this is not the case.

Table 3.8. Comparison of true VARscores with maximum VARscores for repeat regions.

Strains	GPG/ ngpg	Allele Repeats#	Region 1			Region 2		
			Actual VARscore	Uniform codon usage VARscore	**Uniform minus actual ***	Actual VARscore	Uniform codon usage VARscore	**Uniform minus actual ***
Au11	ngpg	9+4	0.725	0.725	**0**	-0.533	-0.491	**0.042**
Au19	GPG	10+4	0.746	0.746	**0**	-0.533	-0.491	**0.042**
CH42	GPG	9+4	0.572	0.572	**0**	-0.533	-0.491	**0.042**
OTG6	ngpg	9+5	0.725	0.725	**0**	-0.188	0.158	**0.346**
RIHO16	GPG	11+4	0.888	0.888	**0**	-0.533	-0.491	**0.042**
OD9014	GPG	6+6	-0.072	0.06	**0.132**	0.101	0.232	**0.131**
var1.4vag	GPG	9+4	0.418	0.418	**0**	-0.533	-0.491	**0.042**
var1.5vag	ngpg	6+8	-0.072	0.059	**0.131**	0.567	0.586	**0.019**
CLB49	ngpg	7+4	0.325	0.325	**0**	-0.533	-0.491	**0.042**
RIHO30	ngpg	6+7	0.0596	0.0596	**0**	0.353	0.375	**0.022**

***** These columns represent the difference of VARscores after the repeat units uniformed. These numbers are calculated by matching the difference between actual VARscore and the uniform codon usage VARscore.

3.8. Identification of a potential functional domain in Ssr1p using bioinformatics and literature search

3.8.1. Ssr1p contains a CFEM domain

After finding the variable repeats, bioinformatic analysis and literature searches were performed to see what function the repeats might be affecting. Analysis of the amino acid sequence of Ssr1p in Pfam (a website for protein family bioinformatics), a protein family was found that matched significantly the primary sequence between amino acids 20 and 82 of Ssr1p (**Figure 3.22**). It suggested that Ssr1p might contain an eight-cysteine-containing CFEM domain, common in several fungal extracellular membrane proteins (Kulkarni *et al.*, 2003). The location of this domain overlaps the Cys-rich region described by Garcera *et al.*, (2003) (**Section 1.5.2** and **Figure 3.1**). The relative positions of the CFEM domain and the two repeat regions in the *SSR1* gene is shown in **Figure 3.23**, based on strain SC5314.

3.8.2. What is CFEM domain and what does it do?

The CFEM domain was identified in 2003 during research into ACI1, an adenylate cyclase of the major rice pathogen *Magnaporthe grisea* (Kulkarni *et al.*, 2003). Protein ACI1 was found to share a domain of ~60 amino acids containing eight cysteine residues with other fungal membrane proteins (Kulkarni *et al.*, 2003). This CFEM domain is unique to fungal membrane proteins; no animal, plant or prokaryote sequences have been found to contain this domain (Kulkarni *et al.*, 2003). The CFEM-containing proteins do not show homology to any other proteins other than this group (Kulkarni *et al.*, 2003).

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

Sequence search results

Show the detailed description of this results page.

We found **5** Pfam-A matches to your search sequence (**1** significant and **4** insignificant) but we did not find any Pfam-B matches.

Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				
CFEM	CFEM domain	Domain	n/a	19	82	20	82	2	66	37.5	1.2e-09	n/a	Show

Insignificant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				
Defensin_2	Arthropod defensin	Domain	CL0054	23	51	25	50	10	33	15.0	0.012	n/a	Show
Mitofilin	Mitochondrial inner membrane protein	Family	n/a	47	221	52	213	27	186	12.3	0.038	n/a	Show
DUF2201	Predicted metallopeptidase (DUF2201)	Family	CL0128	16	196	53	171	127	226	9.4	0.3	n/a	Show
Macoilin	Transmembrane protein	Family	n/a	14	223	74	205	250	382	7.4	0.91	n/a	Show

Comments or questions on the site? Send a mail to **pfam-help@sanger.ac.uk**

Figure 3.22. A protein family that has significant homology with Ssr1p (http://pfam.sanger.ac.uk/). The blue oval shows the location of the matching domain (green) in the peptide: the first 80 N-terminal amino acid residues which is the Cys-rich region. The pink oval shows the name of the matched protein domain: CFEM.

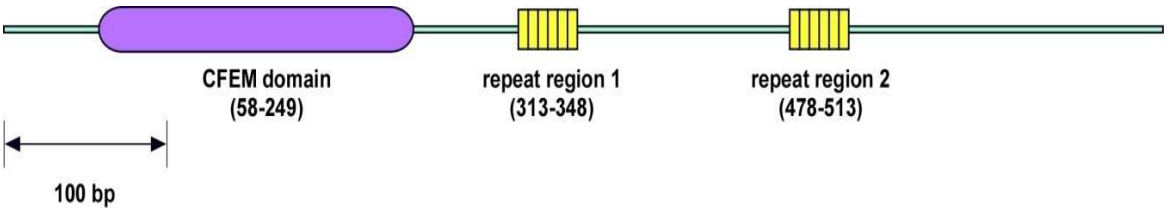


Figure 3.23. The CFEM domain and the two repeat regions in the *SSR1* gene (based on SC5314).

Multiple alignments of ACI1 and the other 29 CFEM-containing fungal proteins, showed that the CFEM domain was present in one or more copies in these fungal membrane proteins, usually with one copy close to the N-terminus of the polypeptide, with the GPI domain (if there was one) near the C-terminus (Kulkarni *et al.*, 2003). No other domain appears within the CFEM-containing domain (Kulkarni *et al.*, 2003). The amino acid sequence of this domain can be summarized as follows: PxC[A/G]x2Cx8–12Cx1–3[x/T]Dx2–5CxCx9–14Cx3–4Cx15–16C , where 'x' is any amino acid residue (Kulkarni *et al.*, 2003).

Kulkarni *et al* (2003) did not include Ssr1p as one of the alignments of 30 CFEM-containing proteins. However, the eight-cysteine-containing domain in the *SSR1* sequence (**Figure 3.1.E**) matches the general sequence given in Kulkarni *et al* (2003), with a few point mutations in residues other than cysteine. There are other signs that Ssr1p contains a CFEM domain. One of the 30 proteins aligned was *S. cerevisiae* protein CW14 (Kulkarni *et al.*, 2003) which is homologous to *C. albicans* Ssr1p (Garcera *et al.*, 2005). Interestingly, the *CSA1* gene of *C. albicans*, which is involved in pathogenicity and is highly expressed in mycelial stages, contains five CFEM domains (Lamarre *et al.*, 2000).

Other *C. albicans* cell wall proteins containing a CFEM domain are involved in biofilm formation, development, and/or preservation (Perez *et al.*, 2006). These proteins include Rbt5p, Pga10 (also called Rbt51), and Wap1p (also called Csa1p)

(Perez *et al.*, 2006). The null mutants of the genes encoding these proteins showed significantly reduced biofilm-forming ability (Perez *et al.*, 2006). Consistent with cell wall proteins containing CFEM domains playing roles in *C. albicans* biofilm formation (Perez *et al.*, 2006). Deletion of the CFEM domain-containing proteins makes the cell more sensitive to cell-wall-perturbing agents including calcofluor white, Congo red and SDS (Perez *et al.*, 2006), which is the same response observed in the *SSR1* null-mutant cells (Garcera *et al.*, 2003; Garcera *et al.*, 2005). These results further support the idea that Ssr1p carries a CFEM domain and that it may be involved in biofilm formation, although previous publications on the *SSR1* gene and Ssr1p do not mention CFEM domains.

As well as biofilm formation, CFEM domain-containing proteins Rbt5p and its homologues Rbt51p/Pga10p are involved in haemoglobin-iron utilization (Sosinska *et al.*, 2008; Weissman & Kornitzer, 2004). The ability to utilize iron from the host haemoglobin is considered a major factor in the pathogenicity of *C. albicans* (Weissman & Kornitzer, 2004). The expression of *RBT5* was significantly increased during iron starvation, and a null-mutant of *RBT5* had decreased haem-iron utilization ability (Weissman & Kornitzer, 2004). *S. cerevisiae* protein Ccw14, which is homologous to *C. albicans* Ssr1p, is also considered similar to Rbt5p (Weissman & Kornitzer, 2004). The Ccw14 of *S. cerevisiae* does not appear to be involved in iron utilisation from human haemoglobin, indicating that not every CFEM domain-containing protein is involved in haemoglobin-iron utilization. However, this does not exclude the possibility that Ssr1p may be involved in haemoglobin-iron utilization by *C. albicans*.

The bioinformatic analysis and literature search showed that Ssr1 protein contains a CFEM domain which may relate to its function and pathogenicity. Although the exact function of the CFEM domain remains to be elucidated the CFEM-containing Ssr1p may play a role in pathogenicity via processes such as adhesion, biofilm formation or haem-iron binding.

CHAPTER FOUR - CONCLUSIONS, DISCUSSIONS AND FUTURE WORK

4.1. Key findings of this project

4.1.1. *SSR1* is a hypermutable gene

Two repeat regions in *SSR1* were investigated by genotyping and shown to be highly variable among a collection of 132 strains of *C. albicans*. DNA sequencing of a selection of 14 alleles showed that variation in number of tandem repeats in repeat regions 1 and 2 (mainly TCTGCT, encoding amino acids Ser and Ala) caused the different allele sizes while the rest of the gene was constant (**Section 3.3**). There were found 7 types of repeat region one, 12 types of repeat region two, 35 different alleles and 42 allelic combinations, indicating high allelic diversity. After approximately 300 divisions *in vitro*, new alleles were detected in repeat region two in 2 out of 60 single colonies tested (**Section 3.5.1**). This gives an estimated mutation rate of 1.11×10^{-4} mutations per cell division (binomial confidence: 0.0041 - 0.1153). This is higher than the mutation rate of artificial repeat genes of *S. cerevisiae* with the same VARscore (**Section 3.5.2**), suggesting *SSR1* is highly mutable. It is concluded that *SSR1* is a hypermutable gene and the variability is caused by the repeat units in both repeat regions of the gene.

4.1.2. *SSR1* alleles show clade specificity

Comparison of different groups of infection-causing strains revealed that GPG strains were less variable in their *SSR1* genotype than non-GPG (ngpg) strains, with 2 similar repeat combinations accounting for 58% of GPG strains (30% 6+6/9+4 and 28% 6+6/10+4). The distributions of these two allele combinations in GPG strains and non-GPG strains were significantly different (Fisher's exact test, $P=0.0001$ for allele combination 6+6/9+4 and, $P<0.0001$ for 6+6/10+4) (Section

3.4.3.3).

All *C. albicans* tandem repeat (TR)-containing genes investigated to date, including *ALS7*, *PNG2*, *HWP1*, *YWP1*, *EAP1* and *SSR1*, have an allele combination specific to GPG strains (clade A), with different levels of clade specificity (Zhang *et al.*, 2003; Wattimena *et al.*, 2009; Zhang *et al.*, 2010). The clade specificity of *SSR1* in infection - causing isolates, compared with repeat-containing genes *PNG2* (Zhang *et al.*, 2010), *YWP1*, *EAP1* (Wattimena *et al.*, 2009) is shown in **Table 4.1**. Like the other genes, *SSR1* shows strong clade-specificity for both of its two predominant genotypes.

Table 4.1. Repeat-containing genes and their clade-specificity.

Gene*	Occurrence of predominant allele combination in GPG strains (%)	Occurrence of same allele combination in Non-GPG strains (%)	Significance	Statistical test
PNG2	80	8	P<0.001	z-test
YWP1	94	11	P<0.0001	chi-square test
EAP1	57	2	P<0.001	chi-square test
SSR1 (1)**	30	2.20	P<0.0001	Fisher's exact test
SSR1 (2)**	28	0	P=0.0000	Fisher's exact test

**ALS7* is not included because the data and statistical test of the gene in *ALS7* paper was only base on allele, rather than allele combination. The two repeat regions (tandem repeat domain & VASES domain) were analyzed separately; these data do not fit in this table as the others do.

** *SSR1* has two predominant allele combinations.

The clades of *C. albicans* diverged more than 3 million years ago (Lott *et al.*, 2005). As a commensal coloniser, the growth rate of *C. albicans* in a rat model is 1.4 divisions per day (R. D. Cannon, unpublished observations). Previous research has shown a mutation rate of 10^{-2} -10^{-5} per generation for fungal TRs (Verstrepen *et al.*, 2005). In this study, the mutation rate of allele 6+31 *in vitro* was detected as 1.1×10^{-4} per cell division (section **3.5.2**). With these generation times and mutation

rates, GPG strains could have generated new alleles and broken the clade-specificity within 3 million years (Zhang *et al.*, 2009; Zhang *et al.*, 2010). This suggests that the maintenance of clade-specific alleles required strong selection.

Selection could be related to function at the protein level. In infection-causing strains, the *SSR1* RR2 in non-GPG strains was highly variable in length, from 4 to 31 repeats, including a number of long runs of 14, 15, 16, 24, 25 and 31 repeats (**Figure 3.12.b**), while *SSR1* RR1 in non-GPG strains was no longer than 9 repeats (**Figure 3.12.a**). In GPG strains RR2 did not expand beyond 8 repeats (**Figure 3.12.b**) while RR1 varied between 6 and 11 repeats (**Figure 3.12.a**). The results suggest that there may be selection against RR2 repeats longer than 8 in GPG strains only and selection against RR1 repeats longer than 9 in non-GPG strains (or against RR1 repeats longer than 11 in both GPG and non-GPG infection-causing strains). The differences in expansion between the two regions may relate to changes in the structure of the protein, e.g. affecting a functional domain such as for example, the CFEM domain identified in this study (**3.8.1**). This may be an important feature in *SSR1* that separates non-GPG from GPG strains.

SSR1 is highly mutable, as demonstrated by *in vitro* culture (**3.5.1**), showing that it has the potential to be used in short term adaptation. However, many synonymous mutations exist (**Figure 3.5**), lowering the mutability of the regions as indicated in the VARscores (**Table 3.7**), and the gene would be more mutable if the repeat units with these mutations were removed by looping out. This suggests that the synonymous mutations are kept because selection favours a lower mutation rate. These point mutations occur in both GPG and non-GPG strains (**Figure 3.4**). It is believed there are genomic TR stabilization mechanisms which remain undiscovered (Astolfi *et al*., 2003). These mechanisms might be involved in the generational maintenance of the synonymous mutations seen in this study. The clade specificity suggests that two allele combinations may fit best in their genetic background and, in this case, clade selection would also work against cells with different alleles.

4.1.3. *SSR1* and its role in short-term adaptation

4.1.3.1. Massive changes in *SSR1* allele distribution are not associated with different hosts or body sites.

Changing the host environment from human to rat did not appear to result in rapid mutation in *SSR1* alleles. However, the rat experiment would only allow very massive changes to be detected and the *C .albicans* cells had only had about 39 divisions in the rats before being analysed. If selection had occurred in the rat and it favoured other alleles, these would be difficult to detect unless cells with some of these alleles grew very much faster than the original strain.

Evidence for selection of different alleles at different body sits could likewise not be obtained. Distributions of alleles and allele combinations tentatively suggest that such selection may occur, in particular selection for unusual alleles in sterile sites. However no definite conclusion could be reached, because the number of isolates from each site was comparatively small.

4.1.3.2. A change in alleles may be associated with the switch from commensal to pathogen in GPG strains

Investigation of the allele distribution showed that allele 10+4, which accounted for 21% of GPG infection-causing alleles, does not appear in non-GPG strains (**Figure 3.13**). The allele combination 6+6/10+4, which was found in 28% of GPG infection-causing strains and 50% of GPG commensal strains, does not appear in non-GPG strains (**Figure 3.14** and **Figure 3.15**). Thus two very similar allele combinations are overrepresented among GPG strains, and they are likely to be the most widespread *SSR1* genotypes of *C. albicans*. In both combinations one allele had 6+6 repeats, and the other had either 9+4 repeats or 10+4 repeats (**Figure 3.14**).

The clade specificity of *SSR1* indicated that, as in other hypermutable TR-containing genes (Zhang *et al.*, 2003; Wattimena *et al.*, 2009; Zhang *et al.*, 2009; Zhang *et al.*, 2010), hypermutation of *SSR1* may assist *C. albicans* in optimizing interaction with the host, with different allele combinations working best in specific genetic backgrounds as a long term adaptation mechanism (Zhang *et al.*, 2009; Zhang *et al.*, 2010). However, a change in a single repeat might alter *SSR1* on a short term basis. The predominant genotypes in infection-causing strains and commensal strains vary by only one repeat unit (**Figure 3.15**). It is possible that this change may be related to different functions of Ssr1p associated with a change in response to the new circumstances of the *C. albicans* and its host. This is only the second time evidence has been reported for the selection of different repeat lengths under different circumstances in *C. albicans*, indicating a possible role in short term adaptation. Al-Aidan *et al.* (2007) showed that repeat region W2 (in gene *Efg1*) varied during infection. Our results help define *SSR1* as a hypermutable gene that only acts as a contingency gene in some circumstances.

As discussed above, changing the host environment from human to rat did not appear to result in any rapid mutation in *SSR1*, and an association between alleles and human body sites could not be clearly demonstrated (although this could be due to small sample size). These results did not support the idea that major changes in *SSR1* alleles are used for short term adaptation to different environments. In humans, there were different alleles in infection-causing and commensal strains but there was no definite relationship with body location, suggesting that the key difference is between commensalism and infection. It is possible that the healthy host favours the selection of ‘commensal’ alleles, or selects against infection-related allele. When the host is not healthy, for example, immunocompromised, this selection changes. Where *C. albicans* acts as a pathogen, it often penetrates tissues and tends to form more germ tubes (Smail & Jones, 1984). It is possible that the varied repeat regions of *SSR1* affect the CFEM domain, which may play a role in *C.albicans* biofilm formation (Perez *et al.*, 2006). In this case, the mutability of *SSR1* may enable the short-term adaptation, or switching, between commensal and pathogen.

4.2. A possible history of *Candida albicans SSR1*

For millions of years, the variable DNA of TRs contributed to the high genomic plasticity of *Candida albicans*, and provided the large repertoire of *C. albicans* phenotypes that helped *C. albicans* survive in the changing microflora in host environments. This study showed the *SSR1* gene, which encodes a GPI cell wall protein that includes a novel CFEM domain, also contains 2 TR regions (RR1 and RR2) that provide a repertoire of variable alleles. The variation of RR1 and RR2 may change protein structure by the addition or deletion of Ser/Ala repeats, and thus change the position of the CFEM domain. The function of CFEM is not known but could be related to biofilm formation, or have other roles in disease, by affecting fungal survivability and pathogenicity.

It is likely the most significant selection event in the evolution of *SSR1* occurred more than 3 million years ago, when one group of *C. albicans* with a certain genetic background diverged from the remainder of the species. This background evolved further to make these strains highly successful in humans today. This special genome made this group of *C. albicans* - known as GPG – widespread in the human host and it causes disease at frequencies 10-100 times greater than the other 37 groups identified (Schmid *et al.*, 1999). Because the genetic background of the *C. albicans* fitted in the human host environment well, selection pressure was low, aiding the stability of the GPG cluster genome. It is likely the pathogenicity is caused by combinations and modulations of different virulence genes (Tavanti *et al.*, 2010). There are many different virulence factors of *C. albicans* and no single one plays a crucial role in disease. The stable genetic background of GPG strains provides stable interactions and combinations that may give *C. albicans* an advantage in survivability in the human host, while more unstable genetic combinations disadvantage the organism. In other words, the specific genotype for the whole genetic background in GPG strains was strongly selected. For the *SSR1* gene, the genotypes best suited to the GPG gene background are 6+6/9+4 and 6+6/10+4. These are both predominant genotypes in

GPG strains with the deletion of one repeat unit at RR1 of the 2nd allele (from 10 to 9) possibly being associated with the change from commensalism to infection.

The mutability of TRs in the *SSR1* gene might play roles in both long term survival and short term adaptation. The GPG clade-specificity of *SSR1*, and other TR-containing genes, suggests a role for specific alleles in long term survival in human hosts within the stable genetic background. However, for GPG strains, a switch between *SSR1* allele 9+4 and 10+4 might occur for short periods and be used for short term adaptation, possibly with a role in pathogenicity of *C. albicans*. This kind of pathogenicity-related adaptation could be extremely short-term, only existing until the patient is cured or dead (Schmid *et al.*, 1999). This might explain why *SSR1* has two similar genotypes predominant in infection-causing GPG strains. Therefore, with its high mutability and clade-specificity, *SSR1* can act as a hypermutable contingency gene.

The population switching to alleles might be the result of the selection of varying host environments. This short term adaptation could be an advantage for *C. albicans* because causing infection may help the microorganisms invade the host tissue and gain nutrition (in an immunocompromised host). It may also trigger some host defenses, and even host death, bringing only short term advantage to the microorganisms, whereas survival in a host as a commensal has long term advantages. The ability to live as a commensal can be considered a more successful way to live. The ability to switch between infectivity and commensalism, and vice versa (perhaps under starvation conditions to gain nutrients), by expanding or deleting one TR, might make *SSR1* a very useful survival tool for *C. albicans*.

4.4. Future Work

Genomic: collect more data on the mutation rate of *SSR1 in vitro*. Complete genotyping experiments for HUN91 grown for 300 generations in laboratory culture (colonies stored). Determine the DNA sequence of mutated RIHO30 observed after 300 generations *in vitro*.

Proteomics: Determine the tertiary structure of Ssr1p and determine the effects of changes in repeat unit lengths in both repeat regions on function of the CFEM domain. In particular, investigate the possible effects of allele combinations 6+6/9+4 and 6+6/10+4 on protein function as these may relate to the pathogenicity of *C. albicans*.

Appendix

Appendix I: Supplies of Materials and Equipment

Material or Equipment	Supplier
100bp DNA ladder	Biolabs
1Kb plus DNA ladder	Invitrogen
Agar	Oxoid LTD
Agarose	Roche
Ammonium persulfate	Sigma
Ampicillin	GmbH
Bacto®-tryptone	Becton, Dickinson and Company
Bacto®-yeast extract	Becton, Dickinson and Company
Gel Doc Quantity One software programme	Bio-Rad
Gel Documentation Systems	Bio-Rad
Gel casting units	Bio-Rad/ Hoefer/ Owl
Glucose	Asia Pacific Specialty Chemical Limit
High Pure PCR Product Purification Kit	Roche
*Hin*dIII Kit	Invitrogen
IPTG	PROJEN Industries. Limited.
Long ranger™	BMA
Mini Protean® 3 system	Bio-Rad
Peak Scanner software programme	Applied Biosystems
Quant-iT™ dsDNA BR Assay Kit	Invitrogen
QuBit® fluorometer	Invitrogen
pLUG®-Multi TA-Cloning Vector Kit	iNtRON
Spectrophotometer	NOVA TECH

Appendix II: Primers used in this project (M13 primers supplied by Gibco BRL and the others supplied by invitrogen)

Primer	Sequence
F1	5' TGATGCTGATACTGCCATTT 3'
B2	5' AGCACCAAGACCAACCTTAG 3'
SSR1-1	5' AGAAGAAGCCTTGGCTGAAC 3'
SSR1-2	5' CTAGTTCAGCCAAGGCTTCTTC 3'
SSR1-3	5' AGAAGAAGCCTTGGCTGAAC 3'
SSR1-6	5' GTCTATCGTTACTACTACTACC 3'
SSR1-7	5' AATAATTGCGAAGGGGGCAG 3'
M13 forward	5’ CCCAGTCACGACGTTGTAAAACG 3’
M13 reverse	5’ AGCGGATAACAATTTCACACAGG 3’
Fluorescence labeled primers	**Sequence**
6-FAM-F1 (blue)	5' 6-FAM-TGATGCTGATACTGCCATTT 3'
HEX-SSR1-2 (green)	5' HEX-CTAGTTCAGCCAAGGCTTCTTC 3'

Appendix III: Amino acid sequence of *SSR1* of SC5314 showing CFEM domain
http://www.candidagenome.org/

Two regions of 6 base-pair repeat units have been identified.
>orf19.7030|SSR1 Assembly 19, Contig19-10262:72194-71490C, translated using codon table 12 (234 residues)

MASFLKISLIAIVSTLQTTLAAPPACLLACVAKVEKGSKCSGLNDLSCICTTKNSDVEKCLKE
ICPNGDADTAISAFKSSCSGYSSQSSSSESESESASSEESSASASASASSSAGKSSNVEA
STTKESSSAKASSSAAGSSEAVSSATETASTEESSSAAASASASASATKESSSEAASSTSS
TLKESKTSTTAAASSSESTTATGVLTQSEGSAAKVGLGALVGLVGAVLL*

Blue highlight: CFEM domain

Yellow highlight: repeat regions

Appendix IV: Genotyping results

A. GPG infection-causing strains

Strains	Cluster	Fragment 1		Fragment 1		Fragment 2		Fragment 2		whole						
		allele 1		allele 2		allele 1		allele 2		allele1			allele2			Allele combination
		(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	F1+F2	(bp)	repeats	F1+F2	
AU1	GPG	206.28	6			278.95	6					6+6			6+6	6+6/6+6
AU19	GPG	231.59	10			266.97	4					10+4			10+4	10+4/10+4
Au27	GPG	206.33	6			278.94	6	273.21	5			6+6			6+5	6+6/6+5
Au39	GPG	206.03	6			290.05	8					6+8			6+8	6+8/6+8
Au90	GPG	206.1	6	231.86	10	267.26	4	278.65	6	459.71	12	6+6	471.48	14	10+4	6+6/10+4
CH14	GPG	206.26	6	225.7	9	267.44	4	278.94	6	459.79	12	6+6	465.78	13	9+4	6+6/9+4
CH35	GPG	205.26	6	232.1	10	267.44	4	278.94	6	459.75	12	6+6	471.55	14	10+4	6+6/10+4
CH42	GPG	225.27	9			266.89	4					9+4			9+4	9+4/9+4
CHB5	GPG	206.33	6	225.77	9	267.53	4	279.03	6	459.74	12	6+6	465.75	13	9+4	6+6/9+4
CLB42	GPG	206.23	6	232.02	10	267.32	4	278.79	6	459.65	12	6+6	471.41	14	10+4	6+6/10+4
CLB53	GPG	206.29	6			278.87	6					6+6			6+6	6+6/6+6
CLB56	GPG															
cour-c	GPG	206.16	6	231.96	10	267.26	4	278.82	6	459.7	12	6+6	471.51	14	10+4	6+6/10+4
FJ23	GPG	205.98	6	225.29	9	266.93	4	278.49	6	459.84	12	6+6	471.61	14	9+4	6+6/9+4
FJ26	GPG	206.09	6	231.8	10	267.13	4	278.68	6	459.59	12	6+6	471.46	14	10+4	6+6/10+4
FJ9*	GPG	206.33	6	219.43	8	267.49	4	278.98	6			6+6			8+4	6+6/8+4
HUN122	GPG	206.5	6	225.53	9	273.12	5	278.93	6	460.01	12	6+6	471.42	14	9+5	6+6/9+5
HUN127	GPG	206.16	6			267.28	4					6+4			6+4	6+4/6+4
HUN93	GPG	205.97	6	225.25	9	266.92	4	278.48	6	459.65	12	6+6	465.63	13	9+4	6+6/9+4
HUN95	GPG	206.21	6	225.79	9	267.53	4	279.01	6	459.82	12	6+6	465.78	13	9+4	6+6/9+4
HUN96	GPG	206.1	6	225.54	9	267.21	4	278.61	6	459.64	12	6+6	465.65	13	9+4	6+6/9+4
Jam-2c	GPG	206.1	6	225.44	9	267.08	4	278.66	6	459.7	12	6+6	465.62	13	9+4	6+6/9+4

Strains	Cluster	Fragment 1		Fragment 1		Fragment 2		Fragment 2		whole						
		allele 1		allele 2		allele 1		allele 2		allele1			allele2			Allele combination
		(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	F1+F2	(bp)	repeats	F1+F2	
ko-2c	GPG	206.47	6	226.13	9	267.71	4	279.12	6	459.62	12	6+6	465.55	13	9+4	6+6/9+4
OD8807	GPG	206.09	6	225.48	9	267.13	4	278.69	6	459.67	12	6+6	465.59	13	9+4	6+6/9+4
OD8826	GPG	206.25	6	225.73	9	267.5	4	278.96	6	459.85	12	6+6	465.75	13	9+4	6+6/9+4
OD8911	GPG	231.95	10			267.22	4					10+4			10+4	10+4/10+4
OD8916	GPG	206.17	6			278.53	6					6+6			6+6	6+6/6+6
OD9014	GPG	206.06	6	231.78	10	267.34	4	278.8	6	459.63	12	6+6	471.48	14	10+4	6+6/10+4
RIHO10	GPG	232.49	10			267.48	4	273.17	5			10+4			10+5	10+4/10+5
RIHO13	GPG	206.18	6	232.02	10	267.48	4	278.93	6	459.59	12	6+6	471.37	14	10+4	6+6/10+4
RIHO16	GPG	238.4	11			267.39	4					11+4			11+4	11+4/11+4
RIHO9	GPG	231.72	10			267	4					10+4			10+4	10+4/10+4
rolo-c	GPG	206.28	6	225.76	9	267.42	4	278.88	6	459.67	12	6+6	465.62	13	9+4	6+6/9+4
SC5314	GPG	205.92	6			278.46	6			459.85	12	6+6		12	6+6	6+6/6+6
sim-c	GPG	206.58	6	232.54	10	267.73	4	279.11	6	459.67	12	6+6	471.54	14	10+4	6+6/10+4
var1.10vag	GPG	211.14	7			278.48	6					7+6			7+6	7+6/7+6
var1.1vag	GPG	206.22	6	232.18	10	267.55	4	279.05	6	459.65	12	6+6	471.47	14	10+4	6+6/10+4
var1.3vag	GPG	206.23	6	232.03	10	267.32	4	278.8	6	459.66	12	6+6	471.47	14	10+4	6+6/10+4
var1.4vag	GPG	205.96?	6	225.29	9	266.93	4	278.52	6	459.87	12	6+6	465.79	13	9+4	6+6/9+4
var1.8vag	GPG	206.56	6			267.52	4					6+4			6+4	6+4/6+4
W132	GPG	225.79	9			267.6	4					9+4			9+4	9+4/9+4
W134	GPG	206.4	6	232.33	10	267.58	4	279.01	6	459.65	12	6+6	471.47	14	10+4	6+6/10+4
W26	GPG	205.97	6			278.45	6					6+6			6+6	6+6/6+6
W3	GPG	206.06	6	230.85	10	267.11	4	278.61	6	459.88	12	6+6	471.69	14	10+4	6+6/10+4
W43	GPG	206.21	6			278.96	6					6+6			6+6	6+6/6+6
W59	GPG	206.27	6			267.56	4	279.02	6			6+4			6+6	6+4/6+6
W68	GPG	206.12	6	231.99	10	267.18	4	278.63	6	459.62	12	6+6	471.49	14	10+4	6+6/10+4
YsM073	GPG	206.37	6	225.93	9	267.64	4	279.16	6	459.59	12	6+6	465.55	13	9+4	6+6/9+4

Strains	Cluster	Fragment 1		Fragment 1		Fragment 2		Fragment 2		whole						
		allele 1		allele 2		allele 1		allele 2		allele1			allele2			Allele combination
		(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	F1+F2	(bp)	repeats	F1+F2	
YsU568	GPG	206.31	6	232.09	10	267.46	4	278.96	6	459.77	12	6+6	471.53	14	10+4	6+6/10+4
YsU649	GPG	206.3	6	225.72	9	267.43	4	278.98	6	459.68	12	6+6	465.63	13	9+4	6+6/9+4
YsU751	GPG	206.16	6	225.69	9	267.44	4	278.96	6	459.89	12	6+6	465.82	13	9+4	6+6/9+4

B. Non-GPG infection-causing strains

Strains	Cluster	Fragment 1		Fragment 1		Fragment 2		Fragment 2		Whole						allele combination
		allele 1		allele 2		allele 1		allele 2		allele1			allele2			
		(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	F1+F2	(bp)	repeats	F1+F2	
Au11	ngpg	212.57	7	225.38	9	267.16	4	295.81	9	465.68	13	9+4	483.38	16	7+9	9+4/7+9
Au134	ngpg	206.34	6	225.71	9	273.25	5	290.49	8			9+5			6+8	9+5/6+8
Au2	ngpg	206.07	6			284.36	7					6+7			6+7	6+7/6+7
Au33	ngpg	206.28	6	219.16	8	267.33	4	278.72	6	459.65	12	6+6		12	8+4	6+6/8+4
AU36	ngpg	206.23	6	225.61	9	273.25	5	290.47	8			9+5			6+8	9+5/6+8
CH20	ngpg	206.27	6	219.14	8	267.29	4	278.85	6			6+6			8+4	6+6/8+4
CH3	ngpg	206.06	6	225.37	9	272.99	5	290.23	8	471.41	14	6+8		14	9+5	6+8/9+5
CH41.1	ngpg	219.03	8			267.3	4					8+4			8+4	8+4/8+4
CH9	ngpg	225.24	9			267.19	4					9+4			9+4	9+4/9+4
CLB44	ngpg	212.7	7			278.85	6	296.03	9			7+6			7+9	7+6/7+9
CLB45	ngpg	212.84	7	219.19	8	273.21	5	296.05	9	465.47	13	8+5	483.43	16	7+9	8+5/7+9
CLB49	ngpg	206.29	6	212.79	7	267.49	4	327.39	14	454.08	11	7+4	511.12	20	6+14	7+4/6+14
FJ12	ngpg	218.85	8			273.04	5					8+5			8+5	8+5/8+5
FJ27	ngpg	212.46	7			278.51	6	295.72	9			7+6			7+9	7+6/7+9
FJ3	ngpg	206.44	6			279.17	6	284.86	7			6+6			6+7	6+6/6+7
Gaymc-c	ngpg	206.11	6	212.55	7	266.9	4	278.5	6	454.39	11	7+4	460	12	6+6	7+4/6+6
HUN123	ngpg	212.34	7			295.7	9					7+9			7+9	7+9/7+9
HUN61	ngpg	212.95	7			296.13	9					7+9			7+9	7+9/7+9
HUN64	ngpg	218.84	8			272.76	5			465.58	13	8+5		13	8+5	8+5/8+5
HUN66	ngpg	206.38	6	225.98	9	267.59	4	279.09	6			6+6			9+4	6+6/9+4
HUN68	ngpg	206.22	6	225.66	9	273.3	5	290.44	8	471.47	14	6+8	471.47	14	9+5	6+8/9+5

Strains	Cluster	Fragment 1		Fragment 1		Fragment 2		Fragment 2		Whole						allele combination
		allele 1		allele 2		allele 1		allele 2		allele1			allele2			
		(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	F1+F2	(bp)	repeats	F1+F2	
HUN91	ngpg	206.1	6			383.94	24	389.83	25				6+24			6+25
HUN92	ngpg	206.13	6			278.78	6						6+6			6+6
OD8824	ngpg	219.06	8			267.1	4						8+4			8+4
OTG1	ngpg	206.05	6	212.65	7	290.33	8	295.93	9				6+9			7+8
OTG10	ngpg	219.1	8	225.54	9	273.19	5					8+5			9+5	8+5/9+5
OTG18	ngpg	206.44	6	213.03	7	296.08	9					6+9			7+9	6+9/7+9
OTG2	ngpg	206.29	6	212.64	7	273.02	5	278.74	6	459.67	12	6+6		12	7+5	6+6/7+5
OTG4	ngpg	212.98	7	225.98	9	267.55	4	296.19	9			9+4			7+9	9+4/7+9
OTG6	ngpg	224.36	9			272.96	5					9+5			9+5	9+5/9+5
RIHO2	ngpg	206.69	6			267.49	4	279.01	6			6+4			6+5	6+4/6+5
RIHO30	ngpg	206.53	6			284.71	7	426.91	31			6+7			6+31	6+7/6+31
RIHO5	ngpg	213.03	7			279.07	6	296.16	9			7+6			7+9	7+6/7+9
sw-17c	ngpg															
var1.5vag	ngpg	205.87	6	212.44	7	290.06	8	295.72	9	471.39	14	6+8	483.45	16	7+9	6+8/7+9
vare1.7vul	ngpg	206.25	6			273.01	5	278.76	6			6+5			6+6	6+5/6+6
W137b	ngpg	213.03	7			290.59	8	296.22	9			7+8			7+9	7+8/7+9
W142	ngpg	206.06	6	218.64	8	272.67	5	278.44	6	459.89	12	6+6	465.52	13	8+5	6+6/8+5
W17	ngpg	211.61	7			278.8	6	295.97	9			7+6			7+9	7+6/7+9
W53	ngpg	219.14	8			273.21	5					8+5			8+5	8+5/8+5
W55	ngpg	219.5	8			273.36	5					8+5			8+5	8+5/8+5
YasM1	ngpg	205.96	6			284.51	7					6+7			6+7	6+7/6+7
YASM42	ngpg	206.19	6	212.68	7	267.27	4	333.52	15	454.16	11	7+4	517.77	21	6+15	7+4/6+15
YasU709	ngpg	205.95	6	225.15	9	273	5	290.3	8			9+5			6+8	9+5/6+8
YsU123	ngpg	206.19	6	219.26	8	267.51	4	290.49	8	460.21	12	8+4	471.65	14	6+8	8+4/6+8

Strains	Cluster	Fragment 1		Fragment 1		Fragment 2		Fragment 2		Whole						allele combination
		allele 1		allele 2		allele 1		allele 2		allele1			allele2			
		(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	F1+F2	(bp)	repeats	F1+F2	
YsU363	ngpg	206.07	6	212.47	7	266.95	4	326.76	14	454.24	11	7+4	510.63	20	6+14	7+4/6+14
YsU63	ngpg	206.2	6	212.64	7	267.27	4	339.82	16	454.26	11	7+4	523.25	22	6+16	7+4/6+16

C. GPG commensal strains

Strains	Fragment 1				Fragment 2				Whole			Whole			Allele combination
Commensal GPG	allele 1		allele 2		allele 1		allele 2		Allele 1			Allele 2			
	(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	F1+F2	(bp)	repeats	F1+F2	
cfr2.10vul	206.36	6	232.41	10	267.53	4	279.01	6	459.77	12	6+6	471.53	14	10+4	6+6/10+4
cfr2.1vul	206.23	6	232.04	10	267.41	4	279.01	6	459.77	12	6+6	471.58	14	10+4	6+6/10+4
cfr2.3vul	206.21	6	244.79	12	267.5	4	278.94	6	459.79	12	6+6	483.57	16	12+4	6+6/12+4
cfr2.4vul	206.43	6		6	279.05	6		6			6+6			6+6	6+6/6+6
cfr2.8vag	206.32	6	232.24	10	267.59	4	279.08	6	459.71	12	6+6	471.58	14	10+4	6+6/10+4
cfr2.9vag	206.31	6	232.18	10	267.57	4	279.04	6	459.77	12	6+6	471.52	14	10+4	6+6/10+4
COD21	206.09	6		6	278.73	6		6			6+6			6+6	6+6/6+6
cpr2.2fec	206.45	6	226.01	9	267.69	4	279.18	6	459.7	12	6+6	465.59	13	9+4	6+6/9+4
HMHc1	231.86	10		10	267.25	4		4			10+4			10+4	10+4/10+4
HMHc2	206.15	6		6	278.71	6		6			6+6			6+6	6+6/6+6
HMHc4	206.01	6	244.28	12	267.08	4	278.64	6	459.67	12	6+6	483.54	16	12+4	6+6/12+4
HMHc5	205.93	6	212.47	7	295.82	9		9	477.24	15	6+9	483.37	16	7+9	6+9/7+9
HMHc6	206	6	225.28	9	266.92	4	278.53	6	459.8	12	6+6	465.75	13	9+4	6+6/9+4
HMHc9	205.94	6	231.5	10	266.94	4	278.47	6	459.71	12	6+6	471.64	14	10+4	6+6/10+4
hp10bt	206.03	6	231.73	10	267.08	4	278.61	6	459.67	12	6+6	471.46	14	10+4	6+6/10+4
hp11an	/		/		267.19?	4	278.73	6							
hp11vw	206.01	6	231.71	10	267.11	4	278.62	6	459.7	12	6+6	471.59		10+4	6+6/10+4
hp12bt	206.35	6		6	279.04	6		6			6+6			6+6	6+6/6+6
hp13vu	225.62	9		9	267.22	4		4			9+4			9+4	9+4/9+4
hp13vw	225.33	9		9	267.08	4		4			9+4			9+4	9+4/9+4
hp2bt	205.94	6	231.59	10	266.94	4	278.47	6	459.77	12	6+6	471.6	14	10+4	6+6/10+4
hp31an	232.12	10		10	267.45	4		4			10+4			10+4	10+4/10+4

Strains	Fragment 1				Fragment 2				Whole			Whole			Allele combination
Commensal GPG	allele 1		allele 2		allele 1		allele 2		Allele 1			Allele 2			
	(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	(bp)	repeats	F1+F2	(bp)	repeats	F1+F2	
hp33vu	206.39	6	232.22	10	267.57	4	279.03	6	459.74	12	6+6	471.48	14	10+4	6+6/10+4
hp33vw	206.46	6	232.37	10	267.67	4	279.12	6	459.71	12	6+6	471.56	14	10+4	6+6/10+4
hp3ch	/		/		/		/								
hp42bt	206.32	6	232.16	10	267.45	4	278.99	6	459.71	12	6+6	471.55	14	10+4	6+6/10+4
hp42vp	206.23	6	232.07	10	267.46	4	278.92	6	459.78	12	6+6	471.57	14	10+4	6+6/10+4
hp55bt	206.23	6		6	279.01	6		6			6+6			6+6	6+6/6+6
W104	206.4	6	232.39	10	267.69	4	279.15	6	459.76	12	6+6	471.6	14	10+4	6+6/10+4
W105	206.45	6	232.4	10	267.69	4	279.19	6	459.69	12	6+6	471.51	14	10+4	6+6/10+4
W106	206.35	6	232.35	10	267.72	4	279.17	6	459.69	12	6+6	471.58	14	10+4	6+6/10+4
W107	206.36	6	238.76	11	267.74	4	279.14	6	459.78	12	6+6	477.55	15	10+4	6+6/11+4
W108	206.37	6	232.46	10	267.83	4	279.23	6	459.7	12	6+6	471.53	14	10+4	6+6/10+4
W109	206.46	6	232.55	10	267.85	4	279.29	6	459.71	12	6+6	471.6	14	10+4	6+6/10+4
W111	206.21	6	231.95	10	267.27	4	278.72	6	459.76	12	6+6	471.67	14	10+4	6+6/10+4

References

Al-Aidan, A. W., Goessens, W., Lemmens-den Toom, N., Al-Ahdal, M. & van Belkum, A. (2007). Microevolution in genomic short sequence repeats of *Candida albicans* in non-neutropenic patients. *Yeast* **24**, 155-160.

Astolfi, P., Bellizzi, D. & Sgaramella, V. (2003). Frequency and coverage of trinucleotide repeats in eukaryotes. *Gene* **317**, 117-125.

Ausubel, F.M. *et al*., eds. (2008). Current Protocols in Molecular Biology. Wiley Interscience, New York.

Beckerman, J., Chibana, H., Turner, J. & Magee, P. T. (2001). Single-copy IMH3 allele is sufficient to confer resistance to mycophenolic acid in *Candida albicans* and to mediate transformation of clinical Candida species. *Infect Immun* **69**, 108-114.

Bates, S., de la Rosa, J. M., MacCallum, D. M., Brown, A. J., Gow, N. A. & Odds, F. C. (2007). *Candida albicans* Iff11, a secreted protein required for cell wall structure and virulence. *Infect Immun* **75**, 2922-2928.

Bayliss, C. D., Field, D. & Moxon, E. R. (2001). The simple sequence contingency loci of *Haemophilus influenzae* and *Neisseria meningitidis*. *J Clin Invest* **107**, 657-662.

Beaumont, H. J., Gallie, J., Kost, C., Ferguson, G. C. & Rainey, P. B. (2009). Experimental evolution of bet hedging. *Nature* **462**, 90-93.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580.

Braun, B. R., van Het Hoog, M., d'Enfert, C. & other authors (2005). A human-curated annotation of the *Candida albicans* genome. *PLoS Genet* **1**, 36-57.

Chaffin, W. L., Lopez-Ribot, J. L., Casanova, M., Gozalbo, D. & Martinez, J. P. (1998). Cell wall and secreted proteins of *Candida albicans*: identification, function, and expression. *Microbiol Mol Biol Rev* **62**, 130-180.

Chaffin, W. L. (2008). *Candida albicans* cell wall proteins. *Microbiol Mol Biol Rev* **72**, 495-544.

Chandra, J., Kuhn, D. M., Mukherjee, P. K., Hoyer, L. L., McCormick, T. & Ghannoum, M. A. (2001). Biofilm formation by the fungal pathogen *Candida albicans*: development, architecture, and drug resistance. *J Bacteriol* **183**, 5385-5394.

Chang, H. T. & Yeung, E. S. (1995). Poly(ethyleneoxide) for high-resolution and high-speed separation of DNA by capillary electrophoresis. *J Chromatogr B Biomed Appl* **669**, 113-123.

Cutler, J. E. (1991). Putative virulence factors of *Candida albicans*. *Annu Rev Microbiol* **45**, 187-218.

De Groot, P. W., Hellingwerf, K. J. & Klis, F. M. (2003). Genome-wide identification of fungal GPI proteins. *Yeast* **20**, 781-796.

De Groot, P. W., De Boer, A. D., Cunningham, J., Dekker, H. L., De Jong, L., Hellingwerf, K. J., De Koster, C. & Klis, F. M. (2004). Proteomic analysis of *Candida albicans* cell walls reveals covalently bound carbohydrate-active enzymes and adhesins. *Eukaryot Cell* **3**, 955-965.

Deitsch, K. W., Moxon, E. R. & Wellems, T. E. (1997). Shared themes of antigenic variation and virulence in bacterial, protozoal, and fungal infections. *Microbiol Mol Biol Rev* **61**, 281-293.

Douglas, L. J. (2003). *Candida* biofilms and their role in infection. *Trends Microbiol* **11**, 30-36.

Field, D., Eggert, L., Metzgar, D., Rose, R. & Wills, C. (1996). Use of polymorphic short and clustered coding-region microsatellites to distinguish strains of *Candida albicans*. *FEMS Immunol Med Microbiol* **15**, 73-79.

Forche, A., May, G., Beckerman, J., Kauffman, S., Becker, J. & Magee, P. T. (2003). A system for studying genetic changes in *Candida albicans* during infection. *Fungal Genet Biol* **39**, 38-50.

Fradin, C., De Groot, P., MacCallum, D., Schaller, M., Klis, F., Odds, F. C. & Hube, B. (2005). Granulocytes govern the transcriptional response, morphology and proliferation of *Candida albicans* in human blood. *Mol Microbiol* **56**, 397-415.

Garcera, A., Martinez, A. I., Castillo, L., Elorza, M. V., Sentandreu, R. & Valentin, E. (2003). Identification and study of a *Candida albicans* protein homologous to *Saccharomyces cerevisiae* Ssr1p, an internal cell-wall protein. *Microbiology* **149**, 2137-2145.

Garcera, A., Castillo, L., Martinez, A. I., Elorza, M. V., Valentin, E. & Sentandreu, R. (2005). Anchorage of *Candida albicans* Ssr1 to the cell wall, and transcript profiling of the null mutant. *Res Microbiol* **156**, 911-920.

Hobson, R. P. (2003). The global epidemiology of invasive *Candida* infections--is the tide turning? *J Hosp Infect* **55**, 159-168; quiz 233.

Holland, B. R., Huber, K. T., Dress, A. & Moulton, V. (2002). Delta plots: a tool for analyzing phylogenetic distance data. *Mol Biol Evol* **19**, 2051-2059.

Hoyer, L. L. (2001). The ALS gene family of *Candida albicans*. *Trends Microbiol* **9**, 176-180.

Hunter, P. R. (1991). A critical review of typing methods for *Candida albicans* and their applications. *Crit Rev Microbiol* **17**, 417-434.

Jackson, A. P., Gamble, J. A., Yeomans, T. & other authors (2009). Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. *Genome Res* **19**, 2231-2244.

Jones, T., Federspiel, N. A., Chibana, H. & other authors (2004). The diploid genome sequence of *Candida albicans*. *Proc Natl Acad Sci U S A* **101**, 7329-7334.

Klis, F. M., de Groot, P. & Hellingwerf, K. (2001). Molecular organization of the cell wall of *Candida albicans*. *Med Mycol* **39 Suppl 1**, 1-8.

Klis, F. M., Mol, P., Hellingwerf, K. & Brul, S. (2002). Dynamics of cell wall structure in *Saccharomyces cerevisiae*. *FEMS Microbiol Rev* **26**, 239-256.

Kolpakov, R., Bana, G. & Kucherov, G. (2003). mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res* **31**, 3672-3678.

Kulkarni, R. D., Kelkar, H. S. & Dean, R. A. (2003). An eight-cysteine-containing CFEM domain unique to a group of fungal membrane proteins. *Trends Biochem Sci* **28**, 118-121.

Lamarre, C., Deslauriers, N. & Bourbonnais, Y. (2000). Expression cloning of the *Candida albicans* CSA1 gene encoding a mycelial surface antigen by sorting of *Saccharomyces cerevisiae* transformants with monoclonal antibody-coated magnetic beads. *Mol Microbiol* **35**, 444-453.

Legendre, M., Pochet, N., Pak, T. & Verstrepen, K. J. (2007). Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res* **17**, 1787-1796.

Levdansky, E., Sharon, H. & Osherov, N. (2008). Coding fungal tandem repeats as generators of fungal diversity. *Fungal Biology Reviews* **22**, 85-96.

Li, F. & Palecek, S. P. (2008). Distinct domains of the *Candida albicans* adhesin Eap1p mediate cell-cell and cell-substrate interactions. *Microbiology* **154**, 1193-1203.

Linder, T. & Gustafsson, C. M. (2008). Molecular phylogenetics of ascomycotal adhesins--a novel family of putative cell-surface adhesive proteins in fission yeasts. *Fungal Genet Biol* **45**, 485-497.

Lott, T. J., Fundyga, R. E., Kuykendall, R. J. & Arnold, J. (2005). The human commensal yeast, *Candida albicans*, has an ancient origin. *Fungal Genet Biol* **42**, 444-451.

Lott, T. J., Holloway, B. P., Logan, D. A., Fundyga, R. & Arnold, J. (1999). Towards understanding the evolution of the human commensal yeast *Candida albicans*. *Microbiology* **145 (Pt 5)**, 1137-1143.

Lupetti, A., Guzzi, G., Paladini, A., Swart, K., Campa, M. & Senesi, S. (1995). Molecular typing of *Candida albicans* in oral candidiasis: karyotype epidemiology with human immunodeficiency virus-seropositive patients in comparison with that with healthy carriers. *J Clin Microbiol* **33**, 1238-1242.

MacCallum, D. M., Castillo, L., Nather, K., Munro, C. A., Brown, A. J., Gow, N. A. & Odds, F. C. (2009). Property differences among the four major *Candida albicans* strain clades. *Eukaryot Cell* **8**, 373-387.

Macia, M. D., Blanquer, D., Togores, B., Sauleda, J., Perez, J. L. & Oliver, A. (2005). Hypermutation is a key factor in development of multiple-antimicrobial resistance in *Pseudomonas aeruginosa* strains causing chronic lung infections. *Antimicrob Agents Chemother* **49**, 3382-3386.

McCullough, M. J., Ross, B. C. & Reade, P. C. (1996). *Candida albicans*: a review of its history, taxonomy, epidemiology, virulence attributes, and methods of strain differentiation. *Int J Oral Maxillofac Surg* **25**, 136-144.

Moxon, R., Bayliss, C. & Hood, D. (2006). Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu Rev Genet* **40**, 307-333.

Munro, C. A., Bates, S., Buurman, E. T. & other authors (2005). Mnt1p and Mnt2p of *Candida albicans* are partially redundant alpha-1,2-mannosyltransferases that participate in O-linked mannosylation and are required for adhesion and virulence. *J Biol Chem* **280**, 1051-1060.

Nather, K. & Munro, C. A. (2008). Generating cell surface diversity in *Candida albicans* and other fungal pathogens. *Fems Microbiology Letters* **285**, 137-145.

Netea, M. G., Gow, N. A., Munro, C. A. & other authors (2006). Immune sensing of *Candida albicans* requires cooperative recognition of mannans and glucans by lectin and Toll-like receptors. *J Clin Invest* **116**, 1642-1650.

Netea, M. G., Brown, G. D., Kullberg, B. J. & Gow, N. A. (2008). An integrated model of the recognition of *Candida albicans* by the innate immune system. *Nat Rev Microbiol* **6**, 67-78.

Odds, F. C., Brown, A. J. & Gow, N. A. (2004). *Candida albicans* genome sequence: a platform for genomics in the absence of genetics. *Genome Biol* **5**, 230.
Odds, F. C. (2010). Molecular phylogenetics and epidemiology of *Candida albicans*. *Future Microbiol* **5**, 67-79.

Oh, S. H., Cheng, G., Nuessen, J. A., Jajko, R., Yeater, K. M., Zhao, X., Pujol, C., Soll, D. R. & Hoyer, L. L. (2005). Functional specificity of *Candida albicans* Als3p proteins and clade specificity of *ALS3* alleles discriminated by the number of copies of the tandem repeat sequence in the central domain. *Microbiology* **151**, 673-681.

Padovan, A. C., Chaves, G. M., Colombo, A. L. & Briones, M. R. (2009). A novel allele of *HWP1*, isolated from a clinical strain of *Candida albicans* with defective hyphal growth and biofilm formation, has deletions of Gln/Pro and Ser/Thr repeats involved in cellular adhesion. *Med Mycol*, 1-13.

Panasiti, V., Devirgiliis, V., Borroni, R. G., Spataro, A., Melis, L., Petrella, M. C. & Pala, S. (2007). Atypical cutaneous manifestation of HSV-2 with *Candida albicans* co-infection in a patient with HIV-1. *J Infect* **54**, e55-57.

Pappas, P. G., Rex, J. H., Lee, J. & other authors (2003). A prospective observational study of candidemia: epidemiology, therapy, and influences on mortality in hospitalized adult and pediatric patients. *Clin Infect Dis* **37**, 634-643.

Perepnikhatka, V., Fischer, F. J., Niimi, M., Baker, R. A., Cannon, R. D., Wang, Y. K., Sherman, F. & Rustchenko, E. (1999). Specific chromosome alterations in fluconazole-resistant mutants of *Candida albicans*. *J Bacteriol* **181**, 4041-4049.

Perez, A., Pedros, B., Murgui, A., Casanova, M., Lopez-Ribot, J. L. & Martinez, J. P. (2006). Biofilm formation by *Candida albicans* mutants for genes coding fungal proteins exhibiting the eight-cysteine-containing CFEM domain. *FEMS Yeast Res* **6**, 1074-1084.

Pitarch, A., Nombela, C. & Gil, C. (2006). *Candida albicans* Biology and pthogencity: Insights from proteomics. In *Microbial Proteomics: Functional Biology of Whole Organisms* pp. 285-330. Edited by I. H. Smith & M. Hecker. Hoboken: John Wiley & sons, INC.

Plaine, A., Walker, L., Da Costa, G., Mora-Montes, H. M., McKinnon, A., Gow, N. A., Gaillardin, C., Munro, C. A. & Richard, M. L. (2008). Functional analysis of *Candida albicans* GPI-anchored proteins: roles in cell wall integrity and caspofungin sensitivity. *Fungal Genet Biol* **45**, 1404-1414.

Ramasamy, R. (1998). Molecular basis for evasion of host immunity and pathogenesis in malaria. *Biochim Biophys Acta* **1406**, 10-27.

Rauceo, J. M., Gaur, N. K., Lee, K. G., Edwards, J. E., Klotz, S. A. & Lipke, P. N. (2004). Global cell surface conformational shift mediated by a *Candida albicans adhesin. Infect Immun **72**, 4948-4955.*

Rex, J. H., Walsh, T. J., Sobel, J. D., Filler, S. G., Pappas, P. G., Dismukes, W. E. & Edwards, J. E. (2000). Practice guidelines for the treatment of candidiasis. Infectious Diseases Society of America. *Clin Infect Dis* **30**, 662-678.

Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-277.

Richard, M., De Groot, P., Courtin, O., Poulain, D., Klis, F. & Gaillardin, C. (2002). GPI7 affects cell-wall protein anchorage in *Saccharomyces cerevisiae* and *Candida albicans*. *Microbiology* **148**, 2125-2133.

Rinaldi, M. G., Luna, M. A., Torotoledo, M. E. & other authors (1992). *Candidiasis: pathogenesis, diagnosis, and tereatment*, 2 nd edn. New York: Raven Press.

Ritz, D., Lim, J., Reynolds, C. M., Poole, L. B. & Beckwith, J. (2001). Conversion of a peroxiredoxin into a disulfide reductase by a triplet repeat expansion. *Science* **294**, 158-160.

Roncero, C. & Duran, A. (1985). Effect of Calcofluor white and Congo red on fungal cell wall morphogenesis: in vivo activation of chitin polymerization. *J Bacteriol* **163**, 1180-1185.

Rouabhia, M., Schaller, M., Corbucci, C., Vecchiarelli, A., Prill, S. K., Giasson, L. & Ernst, J. F. (2005). Virulence of the fungal pathogen *Candida albicans* requires the five isoforms of protein mannosyltransferases. *Infect Immun* **73**, 4571-4580.

Rudd, J. J., Antoniw, J., Marshall, R., Motteram, J., Fraaije, B. & Hammond-Kosack, K. (2009). Identification and characterisation of *Mycosphaerella graminicola* secreted or surface-associated proteins with variable intragenic coding repeats. *Fungal Genet Biol* **47**, 19-32.

Ruiz-Herrera, J., Elorza, M. V., Valentin, E. & Sentandreu, R. (2006). Molecular organization of the cell wall of *Candida albicans* and its relation to pathogenicity. *FEMS Yeast Res* **6**, 14-29.

Schmid, J., Voss, E. & Soll, D. R. (1990). Computer-assisted methods for assessing strain relatedness in *Candida albicans* by fingerprinting with the moderately repetitive sequence Ca3. *J Clin Microbiol* **28**, 1236-1243.

Schmid, J., Odds, F. C., Wiselka, M. J., Nicholson, K. G. & Soll, D. R. (1992). Genetic similarity and maintenance of *Candida albicans* strains from a group of AIDS patients, demonstrated by DNA fingerprinting. *J Clin Microbiol* **30**, 935-941.

Schmid, J., Rotman, M., Reed, B., Pierson, C. L. & Soll, D. R. (1993). Genetic similarity of *Candida albicans* strains from vaginitis patients and their partners. *J Clin Microbiol* **31**, 39-46.

Schmid, J., Hunter, P. R., White, G. C., Nand, A. K. & Cannon, R. D. (1995a). Physiological traits associated with success of *Candida albicans* strains as commensal colonizers and pathogens. *J Clin Microbiol* **33**, 2920-2926.

Schmid, J., Tay, Y. P., Wan, L., Carr, M., Parr, D. & McKinney, W. (1995b). Evidence for nosocomial transmission of *Candida albicans* obtained by Ca3 fingerprinting. *J Clin Microbiol* **33**, 1223-1230.

Schmid, J., Herd, S., Hunter, P. R. & other authors (1999). Evidence for a general-purpose genotype in *Candida albicans*, highly prevalent in multiple geographical regions, patient types and types of infection. *Microbiology* **145 (Pt 9)**, 2405-2413.

Shen, J., Guo, W. & Kohler, J. R. (2005). CaNAT1, a heterologous dominant selectable marker for transformation of *Candida albicans* and other pathogenic Candida species. *Infect Immun* **73**, 1239-1242.

Shepherd, M. G., Poulter, R. T. M. & Sullivan, P. A. (1985). *Candida albicans*-Biology, genetics, and pathogenicity *Annu Rev Microbiol* **39**, 579-614.

Smail, E. H. & Jones, J. M. (1984). Demonstration and solubilization of antigens expressed primarily on the surfaces of *Candida albicans* germ tubes. *Infect Immun* **45**, 74-81.

Soll, D. R., Galask, R., Schmid, J., Hanna, C., Mac, K. & Morrow, B. (1991). Genetic dissimilarity of commensal strains of Candida spp. carried in different anatomical locations of the same healthy women. *J Clin Microbiol* **29**, 1702-1710.

Sonneborn, A., Tebarth, B. & Ernst, J. F. (1999). Control of white-opaque phenotypic switching in *Candida albicans* by the Efg1p morphogenetic regulator. *Infect Immun* **67**, 4655-4660.

Sosinska, G. J., de Groot, P. W., Teixeira de Mattos, M. J., Dekker, H. L., de Koster, C. G., Hellingwerf, K. J. & Klis, F. M. (2008). Hypoxic conditions and iron restriction affect the cell-wall proteome of *Candida albicans* grown under vagina-simulative conditions. *Microbiology* **154**, 510-520.

Srikantha, T., Tsai, L. K., Daniels, K. & Soll, D. R. (2000). EFG1 null mutants of *Candida albicans* switch but cannot express the complete phenotype of white-phase budding cells. *J Bacteriol* **182**, 1580-1591.

Staab, J. F., Bahn, Y. S., Tai, C. H., Cook, P. F. & Sundstrom, P. (2004). Expression of transglutaminase substrate activity on *Candida albicans* germ tubes through a coiled, disulfide-bonded N-terminal domain of Hwp1 requires C-terminal glycosylphosphatidylinositol modification. *J Biol Chem* **279**, 40737-40747.

Staab, J. F., Bradway, S. D., Fidel, P. L. & Sundstrom, P. (1999). Adhesive and mammalian transglutaminase substrate properties of *Candida albicans* Hwp1. *Science* **283**, 1535-1538.

Stoldt, V. R., Sonneborn, A., Leuker, C. E. & Ernst, J. F. (1997). Efg1p, an essential regulator of morphogenesis of the human pathogen *Candida albicans*, is a member of a conserved class of bHLH proteins regulating morphogenetic processes in fungi. *Embo J* **16**, 1982-1991.

Strand, M., Prolla, T. A., Liskay, R. M. & Petes, T. D. (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365**, 274-276.

Sumita, T., Yoko-o, T., Shimma, Y. & Jigami, Y. (2005). Comparison of cell wall localization among Pir family proteins and functional dissection of the region required for cell wall binding and bud scar recruitment of Pir1p. *Eukaryot Cell* **4**, 1872-1881.

Tavanti, A., Campa, D., Bertozzi, A., Pardini, G., Naglik, J. R., Barale, R. & Senesi, S. (2006). *Candida albicans* isolates with different genomic backgrounds display a differential response to macrophage infection. *Microbes Infect* **8**, 791-800.

Tavanti, A., Campa, D., Arancia, S., Hensgens, L. A., de Bernardis, F. & Senesi, S. (2010). Outcome of experimental rat vaginitis by *Candida albicans* isolates with different karyotypes. *Microb Pathog*.

Tsarfaty, I., Sandovsky-Losica, H., Mittelman, L., Berdicevsky, I. & Segal, E. (2000). Cellular actin is affected by interaction with *Candida albicans*. *FEMS Microbiology Letters* **189**, 225-232.

van Belkum, A. (1999). Short sequence repeats in microbial pathogenesis and evolution. *Cell Mol Life Sci* **56**, 729-734.

van het Hoog, M., Rast, T. J., Martchenko, M. & other authors (2007). Assembly of the *Candida albicans* genome into sixteen supercontigs aligned on the eight chromosomes. *Genome Biol* **8**, R52.

VazquezTorres, A. & Balish, E. (1997). Macrophages in resistance to candidiasis. *Microbiol Mol Biol Rev* **61**, 170-&.

Verstrepen, K. J., Jansen, A., Lewitter, F. & Fink, G. R. (2005). Intragenic tandem repeats generate functional variability. *Nat Genet* **37**, 986-990.

Wattimena, S., Patchett, M. & Schmid, J. (2009). Clade specific allele combination of *Candida albicans* contingency gene. In *Combio 2009*. Chritchurch.

Weissman, Z. & Kornitzer, D. (2004). A family of Candida cell surface haem-binding proteins involved in haemin and haemoglobin-iron utilization. *Mol Microbiol* **53**, 1209-1220.

Whiteway, M. & Bachewich, C. (2007). Morphogenesis in *Candida albicans*. *Annu Rev Microbiol* **61**, 529-553.

Xu, J. & Mitchell, T. G. (2003). Geographical differences in human oral yeast flora. *Clin Infect Dis* **36**, 221-224.

Zhang, N., Harrex, A. L., Holland, B. R., Fenton, L. E., Cannon, R. D. & Schmid, J. (2003). Sixty alleles of the ALS7 open reading frame in *Candida albicans*: ALS7 is a hypermutable contingency locus. *Genome Res* **13**, 2005-2017.

Zhang, N., Upritchard, J. E., Holland, B. R., Fenton, L. E., Ferguson, M. M., Cannon, R. D. & Schmid, J. (2009). Distribution of mutations distinguishing the most prevalent disease-causing *Candida albicans* genotype from other genotypes. *Infect Genet Evol* **9**, 493-500.

Zhang, N., Cannon, R. D., Holland, B. R., Patchett, M. L. & Schmid, J. (2010). Impact of genetic background on allele selection in a highly mutable *Candida albicans* gene, PNG2. *PLoS One* **5**, e9614.

Zhao, X., Oh, S. H., Jajko, R., Diekema, D. J., Pfaller, M. A., Pujol, C., Soll, D. R. & Hoyer, L. L. (2007). Analysis of ALS5 and ALS6 allelic variability in a geographically diverse collection of *Candida albicans* isolates. *Fungal Genet Biol* **44**, 1298-1309.